



University of Pennsylvania
ScholarlyCommons


Publicly Accessible Penn Dissertations

2021

The Online Adjustment Of Speaker-Specific Phonetic Beliefs In Multi-Speaker Speech Perception

Wei Lai
University of Pennsylvania

Follow this and additional works at: <https://repository.upenn.edu/edissertations>

 Part of the [Anthropological Linguistics and Sociolinguistics Commons](#), and the [Behavioral Neurobiology Commons](#)

Recommended Citation

Lai, Wei, "The Online Adjustment Of Speaker-Specific Phonetic Beliefs In Multi-Speaker Speech Perception" (2021). *Publicly Accessible Penn Dissertations*. 4110.
<https://repository.upenn.edu/edissertations/4110>

This paper is posted at ScholarlyCommons. <https://repository.upenn.edu/edissertations/4110>
For more information, please contact repository@pobox.upenn.edu.

The Online Adjustment Of Speaker-Specific Phonetic Beliefs In Multi-Speaker Speech Perception

Abstract

This dissertation examines how listeners' knowledge of interspeaker variability guides their generalization of perceptual learning in multi-talker listening. A series of perceptual learning experiments are conducted to evaluate whether listeners generalize what they have learned about a previous talker's production of sibilants and stop VOT to another speaker either of the same gender or a different gender. Experiment 1 and 2 finds that the perceptual learning of sibilants constantly generalizes across speakers of different genders under an acoustics-phonology mismatch constraint. The constraint states that perceptual learning fails to generalize if there is a mismatch between the directions of perceptual shifts intended by the raw acoustic distributions of stimuli and by their phonological distribution in the perceptual space. Experiment 3 reports evidence for the perceptual generalization of stop VOT across speakers of different genders. These results lend support to a cumulative update account, which suggests that perceptual learning updates across speakers in such a way where previous and current perceptual learning experiences are re-integrated to form a cumulative perceptual expectation that listeners use for upcoming perception events. Building on the above findings, Experiment 4 investigates the constraints of speaker identity and gender on the perceptual generalization of sibilants and stops by introducing and manipulating visual identity and voice gender cues. The results show reduced magnitude for perceptual generalization across genders than within gender, and, in the latter case, for perceptual generalization across speakers than within speaker. These results raise the possibility that socioindexical specificity imposes a constraint on perceptual learning by modulating the magnitude of perceptual generalization across social groups, instead of blocking its occurrence. They also suggest that listeners' knowledge of structure in talker variability may be more fine-grained than hard-and-fast bindings of social-demographic groups and lend support to the sophisticated interweaving of social information in the architecture of the phonetics-phonological mapping system.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Linguistics

First Advisor

Meredith M. Tamminga

Subject Categories

Anthropological Linguistics and Sociolinguistics | Behavioral Neurobiology

THE ONLINE ADJUSTMENT OF SPEAKER-SPECIFIC PHONETIC BELIEFS IN
MULTI-SPEAKER SPEECH PERCEPTION

Wei Lai

A DISSERTATION

in

Linguistics

Presented to the Faculties of the University of Pennsylvania

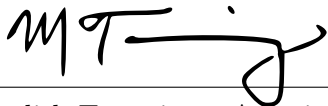
in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

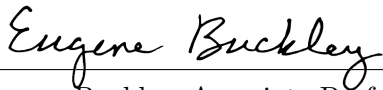
2021

Supervisor of Dissertation



Meredith Tamminga, Associate Professor of Linguistics

Graduate Group Chairperson



Eugene Buckley, Associate Professor of Linguistics

Dissertation Committee:

Jianjing Kuang, Associate Professor of Linguistics

Joseph Toscano, Assistant Professor of Psychological and Brain Sciences

THE ONLINE ADJUSTMENT OF SPEAKER-SPECIFIC PHONETIC BELIEFS IN
MULTI-SPEAKER SPEECH PERCEPTION

© COPYRIGHT

2021

Wei Lai

This work is licensed under the

Creative Commons

Attribution-NonCommercial-ShareAlike 4.0

International License

To view a copy of this license, visit

<http://creativecommons.org/licenses/by-nc-sa/4.0/>

*Dedicated to my parents, Weizhen Xie and Xunxu Lai,
whose love makes me fearless.*

Acknowledgements

At the very outset of this dissertation, I would like to take this space to express my sincere and heartfelt gratitude to all the personages who have helped me throughout my graduate program and academic career. Without their help, encouragement, and comfort, I would not have made it this far to earn my doctoral degree.

First and foremost, I would like to thank my advisor, Meredith Tamminga, who has always been a wise, devoted, and supportive mentor to me. Meredith is one of the earnest persons that I have ever met in my life. When we discuss research questions together, she always listens attentively and never easily lets go of a single underestimated opinion that deserves more elaboration. When I receive invitations for job interviews, she is always the first person to propose arrangement of a practise talk before I ask for anything. Meredith advises me smartly on each step I need to take to achieve my goal but seldom interferes with the big decisions that I am supposed to make by myself. She shows me how to simultaneously maintain the courage to stand up for one's belief and the wisdom to strategize around foreseeable obstacles. I benefit a lot from her advice and my time spent with her in general.

This dissertation has also significantly benefited from discussions with my committee members, Jianjing Kuang and Joe Toscano. I want to thank Jianjing Kuang for guiding me through the first few phonetics projects at UPenn and teaching me how to find a specific research question. I am grateful to Joe Toscano for providing insightful interpretations of the results I obtained as well as professional advice on the experimental design. I also want to thank Mark Liberman and David Embick for their countless feedback on the different sets of projects I pursued and on the data that I obtained at each step of the dissertation during the lab meeting. I am also especially grateful to Gareth Roberts for being a constant

source of mentorship to me during my Ph.D. career. I enjoyed every minute of the time we spent on our collaborative artificial-language-learning project.

During my stay in the program, I have spent most of my time in two labs – the Penn Phonetics lab and the Variation and Cognition Lab. These places are also where I made most of my friends at Penn. I want to thank my sociolinguist office mates, Lacey Wade, Ruairidh Purse, Yosiane White, and Aini Li, for all the moments of closeness and belonging that I have enjoyed in the Variation and Cognition lab. I also want to thank cohort phoneticians, Jia Tian, and Hong Zhang, for keeping me company through good times and bad times. I am also grateful to the other members of the phonetics lab, Alethia Cui, Nari Rhee, Caitlin Richter, and Ollie Sayeed, for always being there for discussion and help. Another big thank-you goes out to the phoneticians and sociolinguists in higher grades – Sunghye Cho, Mao-Hsu Chen, Jinjing Tan, Betsy Sneller, Gudrun (Duna) Gylfadottir, and Sabriya Fisher – for helping me navigate through the most challenging areas and setting role models of academic performance to me upon the first few years of my arrival. Finally, I also want to thank the first-year phoneticians, Pik Yu (May) Chan, Xin Gao, and Christine Soh, who made my last year at Penn truly enjoyable.

I have also enjoyed a lot of professional and mental support from my classmates outside these two labs. I would like to thank the members of my cohort: Spencer Caplan, Andrea Ceolin, Ava Creemers, Jordan Kodner, Caitlin Richter, Jia Tian, Lacey Wade, and Hong Zhang. Each one of you is so smart, professional, and helpful. Being your cohort makes me become a better person. I also want to acknowledge our department coordinator, Amy Forsyth, and a number of higher-grade students at my time, Einar Freyr Sigurdsson, Amy Goodwill, Robert Wilder, Haitao Cai, Luke Adamson, and Milena Šereikaitė. I want to thank them for their kind support and useful guidance when I was stressed out from time to time in my early years in the program. I also owe my thanks to Robert Wilder for generously sharing with me the Ibex scripts and R code for setting up and analyzing perceptual learning experiments, which I have been referring to throughout my dissertation. Lastly, I want to express special gratitude to my writing buddy, Nattanun (pleng) Chanchaochai. Thank

you for always booking the library room for each writing slot we scheduled, for setting up google docs to keep track of our writing progress, for trying out hot food with me in different places, and for skipping the small talks and going really deep in conversations with me.

I obtained a Master's degree from Beijing Normal University under the supervision of Xiaoying Xu. I want to thank her for her continuous encouragement and support in my pursuit of becoming a linguist. Another experience that I really appreciate is collaborating with Jiahong Yuan. My master dissertation was an acoustic study of the intonation of Chinese interrogative questions, and that was when I got to know Jiahong Yuan's research work. At that time, I would not believe that some day in the future, I could be fortunate enough to meet Jiahong in person and collaborate with him on research projects. Outside the Penn community, I am blessed to have encountered many other amazing personages who have inspired me, encouraged me, and provided insightful feedback on my work. In particular, I would like to thank Jie Zhang, Georgia Zellou, Abby Walker, Pam Beddor, Laura Dilley, Dave Kleinschmidt, and Charlotte Vaughn, for shaping some of my view on phonetics, phonology, language variation, and the linguistics field in general. I also want to thank Arthur Samuel attending my dissertation defense in response to my last-minute invitation, which makes me very honored.

Finally, I must thank my family for their continuous love, care and support. My deepest gratitude goes to my amazing parents, Weizhen Xie and Xunxu Lai, and my beloved husband, Desen Lin. You are more than I deserve.

ABSTRACT

THE ONLINE ADJUSTMENT OF SPEAKER-SPECIFIC PHONETIC BELIEFS IN MULTI-SPEAKER SPEECH PERCEPTION

Wei Lai

Meredith Tamminga

This dissertation examines how listeners' knowledge of interspeaker variability guides their generalization of perceptual learning in multi-talker listening. A series of perceptual learning experiments are conducted to evaluate whether listeners generalize what they have learned about a previous talker's production of sibilants and stop VOT to another speaker either of the same gender or a different gender. Experiment 1 and 2 finds that the perceptual learning of sibilants constantly generalizes across speakers of different genders under an acoustics-phonology mismatch constraint. The constraint states that perceptual learning fails to generalize if there is a mismatch between the directions of perceptual shifts intended by the raw acoustic distributions of stimuli and by their phonological distribution in the perceptual space. Experiment 3 reports evidence for the perceptual generalization of stop VOT across speakers of different genders. These results lend support to a cumulative update account, which suggests that perceptual learning updates across speakers in such a way where previous and current perceptual learning experiences are re-integrated to form a cumulative perceptual expectation that listeners use for upcoming perception events. Building on the above findings, Experiment 4 investigates the constraints of speaker identity and gender on the perceptual generalization of sibilants and stops by introducing and manipulating visual identity and voice gender cues. The results show reduced magnitude for perceptual generalization across genders than within gender, and, in the latter case, for perceptual generalization across speakers than within speaker. These results raise the possibility that socioindexical specificity imposes a constraint on perceptual learning by modulating the magnitude of perceptual generalization across social groups, instead of blocking its occurrence. They also suggest that listeners' knowledge of structure in talker variability may be more fine-grained than hard-and-fast bindings of social-demographic groups and lend

support to the sophisticated interweaving of social information in the architecture of the phonetics-phonological mapping system.

Table of Contents

Acknowledgements	iv
Abstract	vii
List of Tables	xiii
List of Figures	xv
Preface	1
1 Introduction	1
1.1 Theoretical and empirical origins	2
1.1.1 Abstract speech representations and speaker normalization	3
1.1.2 Episodic speech representations and exemplars	6
1.1.3 A shift of research foci in the literature	7
1.2 Perceptual learning as a testing ground	10
1.2.1 Flexibility vs. durability	13
1.2.2 Specificity vs. generalizability	16
1.2.3 Perceptual generalization by structures in talker variability	21
1.3 Research questions and goals	24
1.3.1 Is the perceptual learning of sibilant specific to talker or gender? . .	26
1.3.2 Is the perceptual learning of VOT specific to speaker or gender? . .	28
1.3.3 Is the perceptual learning of sibilant more susceptible to speaker and gender specificity than that of VOT?	29
2 General Method	31
2.1 Review on the methodology of perceptual learning	31
2.1.1 Context-guided vs. distribution-based learning	32
2.1.2 The influence of stimulus, paradigm, and task	35
2.2 Methodological decisions in this dissertation	42
2.3 General method of the dissertation	43
2.3.1 Block and trial	44
2.3.2 Recording	48

2.3.3	Manipulation and pilot	49
2.3.4	Subject recruitment	50
2.4	Planned analyses and result interpretation	51
3	Exp 1: Perceptual Learning of /s-f/ across Speaker Genders	57
3.1	Background and research question	57
3.1.1	The acoustic properties of /s-f/ and gender variation in production .	58
3.1.2	Perceptual correlates of /s-f/ and speaker gender effects on perception and learning	60
3.1.3	Research question and hypotheses	63
3.2	Method Overview	64
3.2.1	Experimental conditions	64
3.2.2	Word list and recording	67
3.2.3	Step selection and synthesis	68
3.2.4	Stimulus acoustics	70
3.3	Experiment and result	71
3.3.1	Pilot study: Learning Female A's /s-f/	71
3.3.2	Exp 1a: Previous /s/-favoring training with Female A	75
3.3.3	Exp 1b: Previous /f/-favoring training with Female A	81
3.3.4	Exp 1c: No previous training with Female A's /s-f/	88
3.4	Discussion	93
3.4.1	Cumulative update of perceptual expectations across speakers	94
3.4.2	Effect size, trial order, categorization slope, and transitional bias . .	96
3.4.3	Remaining questions	101
4	Exp 2: Acoustic Interference in the Perceptual Learning of /s-f/ across Speaker Genders	103
4.1	Background and research question	104
4.1.1	Acoustic overlap between the training and the test stimuli	104
4.1.2	Acoustic overlap and mismatch between training phases	106
4.2	Method overview	110
4.2.1	Experimental conditions	110
4.2.2	Word list and recording	111
4.2.3	Step selection and synthesis	112
4.2.4	The acoustics of synthesized stimuli	113
4.3	Experiment and result	114
4.3.1	Pilot study: Learning Female B's /s-f/	114
4.3.2	Exp 2a: Previous /s/-favoring training with Female B	117
4.3.3	Exp 2b: Previous /f/-favoring training with Female B	122
4.3.4	Exp 2c: No previous training with Female B's /s-f/	127
4.4	Discussion	131

4.4.1	Effect of the acoustic distributions of the training and test stimuli . . .	132
4.4.2	Effect of the acoustic distributions of different training phases	135
5	Exp 3: Perceptual Learning of /t-d/ across Speaker Genders	139
5.1	Background and research question	140
5.1.1	Voice onset time and other phonetic cues to English stop voicing . .	140
5.1.2	Talker and gender variability in VOT production and their influence on perceptual learning	144
5.1.3	Research questions	149
5.2	Method Overview	150
5.2.1	Experimental conditions	150
5.2.2	Word list and recording	152
5.2.3	Step selection and manipulation	154
5.3	Experiment and Result	158
5.3.1	Pilot study: Learning Female A's /t-d/	158
5.3.2	Exp 3a: Perceptual learning of /t-d/ with Female A and Male A . .	161
5.3.3	Exp 3b: No previous training with Female A's /t-d/	166
5.4	Discussion	172
6	Exp 4: Comparing Speaker Effects on the Perceptual Learning of /s-f/ and /t-d/ within and across Genders	175
6.1	Background and research question	176
6.1.1	Speaker, gender, and phoneme type	176
6.1.2	Qualitative vs. quantitative speaker effect	177
6.1.3	Research question and hypotheses	178
6.2	Method overview	179
6.2.1	Experimental conditions	179
6.2.2	Stimuli	182
6.2.3	Participant	183
6.2.4	Speaker perception	183
6.3	Experiment and Result	186
6.3.1	Exp 4a: Perceptual learning of /s-f/ with Female A and B	186
6.3.2	Exp 4b: Perceptual learning of /t d/ with Female A and B	193
6.4	Discussion	198
7	Discussion and Conclusion	203
7.1	Major findings of this dissertation	203
7.1.1	Perceptual generalization by cumulative update	203
7.1.2	Constraints on cross-talker perceptual generalization	206
7.2	Theoretical implications and future research directions	209
7.2.1	The role of listeners' knowledge of structure in talker variability . .	209

7.2.2	Perceptual generalization in the acoustic vs. phonological space . . .	212
7.2.3	Implications for the mental representation of speech variability in the phonetics-phonology interface	213
A	Results of Pilot Studies	215
B	Supplementary models	220
	Bibliography	225

List of Tables

1.1	Previous findings of the specificity versus generalization of perceptual learning across speakers and genders for fricatives and stops	26
2.1	Examples of training stimuli in different experimental conditions: IPA transcriptions and two options provided for identification	45
2.2	Word list of the training trials for the perceptual learning of /s-f/	46
2.3	Word list of the training trials for the perceptual learning of /t-d/	47
2.4	Word list of fillers and their foil options in perceptual learning	48
2.5	Word list of test trials in the form of minimal pairs	48
3.1	Predictions of the four possible mechanisms of perceptual generalization on the result of Exp 1	64
3.2	The proportion of [s] mixed in the most ambiguous Step of sibilant chosen for each word frame for Female A and Male A	69
3.3	The splicing of the test stimuli	70
3.4	The fixed effects of the logistic mixed-effects model in Exp 1 pilot	74
3.5	The fixed effects of the logistic mixed-effects model in Exp 1a	79
3.6	The fixed effects of the logistic mixed-effects model in Exp 1b	86
3.7	The fixed effects of the logistic mixed-effects model in Exp 1c	91
4.1	Summary of how experimental conditions in Exp 2 fit into the two hypothesized acoustic constraints	109
4.2	The proportion of [s] mixed in the most ambiguous step of sibilant chosen for each word frame for Female B	113
4.3	The fixed effects of the logistic mixed-effects model in Exp 2 pilot	116
4.4	The fixed effects of the logistic mixed-effects model in Exp 2a	120
4.5	The fixed effects of the logistic mixed-effects model in Exp 2b	125
4.6	The fixed effects of the logistic mixed-effects model in Exp 2c	129
5.1	The temporal compression factor chosen for each word of the training stimuli for Female A and Male A	157
5.2	The VOT duration at each step of the /t-d/ continuum embedded in different minimal-pair words in the test phase (ms)	158
5.3	The fixed effects of the logistic mixed-effects model in Exp 3 pilot	160
5.4	The fixed effects of the logistic mixed-effects model in Exp 3a	164
5.5	The fixed effects of the logistic mixed-effects model in Exp 3b	169
5.6	Condition estimates of the four logistic mixed-effects models for responses obtained with different lexical frames in Exp 3b	170

6.1	The participant information in each condition in Exp 4	183
6.2	Voice gender responses in each condition of Exp 4	184
6.3	The fixed effects of the logistic mixed-effects model in Exp 4a	189
6.4	The fixed effects of the logistic mixed-effects model in Exp 4b	195
B.1	The fixed effects of Model 1c-b	220
B.2	The fixed effects of Model 1c-c	220
B.3	The fixed effects of Model 2c-b	221
B.4	The fixed effects of Model 2c-c	221
B.5	The fixed effects of Model 3c- <i>tear</i>	222
B.6	The fixed effects of Model 3c- <i>time</i>	222
B.7	The fixed effects of Model 3c- <i>town</i>	223
B.8	The fixed effects of Model 3c- <i>touch</i>	223
B.9	The fixed effects of Model 4a- <i>relevel</i>	224
B.10	The fixed effects of Model 4b- <i>relevel</i>	224

List of Figures

2.1	A diagram of the basic experimental procedure in this dissertation	44
2.2	Potential outcomes of perceptual learning with speaker A and B successively	52
3.1	Spectral peak location of fricatives by place of articulation and by gender; figure adopted from Jongman et al. (2000)	59
3.2	The structure of sub-experiments and conditions in Exp 1	66
3.3	Mean and 95% confidence interval of the center of gravity of /s/ and /ʃ/ of Female A and Male A in natural speech production (Hz)	68
3.4	Mean and 95% confidence interval of the center of gravity of sibilants in different training phases and in the test phase in Exp 1 (Hz)	71
3.5	Exp 1 pilot: Boundary shift after exposure to Female A's /s/-favoring and /ʃ/-favoring speech compared to the categorization baseline (mean and standard error)	73
3.6	Exp 1a: /s/ response rate as a result of cross-gender perceptual learning with different sibilant distributions (mean and standard error)	78
3.7	The COG of sibilants in the Two genders - opposite conditions in Exp 1a and 1b (mean and 95% CI)	82
3.8	Exp1b: /s/ response rate as a result of cross-gender perceptual learning with different sibilant distributions (mean and standard error)	84
3.9	Exp 1c: /s/ response rate as a result of training with Male A compared to the baseline (mean and standard error)	90
3.10	Exp 1c: /s/ response rate as a result of training with Male A compared to two-genders training conditions (mean and standard error)	92
3.11	Asymmetric effect sizes between experimental conditions in Exp 1a and 1b	98
3.12	/s/ response rate in two-gender training conditions by phoneme and experiment (mean and standard error)	101
4.1	A schema of acoustic overlapping between fricatives under experimental conditions of different perceptual biases	108
4.2	The structure of experiments and conditions in Exp 2	111
4.3	The COG of sibilants of Female A, Female B and Male A in natural speech production (mean and 95% CI)	112
4.4	The COG of sibilants in different training phases and in the test phase in Exp 2 (mean and 95% CI)	114
4.5	Exp 2 pilot: Boundary shift after exposure to Female B's /s/-favoring and /ʃ/-favoring speech compared to the categorization baseline (mean and standard error)	116

4.6	Exp 2a: /s/ response rate as a result of cross-gender perceptual learning with different sibilant distributions (mean and standard error)	119
4.7	The COG of sibilants in the two - gender opposite conditions in Exp 2a and 2b (mean and 95% CI)	123
4.8	Exp 2b: /s/ response rate as a result of cross-gender perceptual learning with different sibilant distributions (mean and standard error)	124
4.9	Exp 2c: /s/ response rate as a result of training with Male A compared to the Female B baseline (mean and standard error)	128
4.10	Exp 2c: /s/ response rate as a result of training with Male A compared to two-genders training conditions (mean and standard error)	130
4.11	The COG of Male A's training stimuli of Male A and Female A's test stimuli	133
4.12	The COG of the training stimuli in Exp 2a and 2b	136
5.1	The structure of sub-experiments and conditions in Exp 3	151
5.2	The VOT length of /t d/ and the duration of /t d/-containing words from Female A and Male A (mean and 95% confidence interval)	153
5.3	A sound clip of the word <i>town</i> and its cue-bearing areas of relevance to the perception of the critical phoneme	155
5.4	Exp 3 pilot: Boundary shift after /t/-biased learning compared to the baseline categorization with Female A (mean and standard error)	160
5.5	Exp 3a: /t/ response rate as a result of cross-gender perceptual learning with different combinations of /t d/ production biases (mean and standard error)	163
5.6	Exp 3b: /t/ response rate as a result of training with Male A compared to the two-gender condition and the baseline condition (mean and standard error)	168
5.7	Exp 3b: /t/ response rate obtained with different minimal-pair test stimuli in the Female A baseline condition, the Male A /t/-favoring condition, and the Two genders - same condition (mean and standard error)	171
6.1	The structure of sub-experiments and conditions in Exp 4	180
6.2	Voice number responses in each condition of Exp 4a and 4b	185
6.3	Exp 4a: /s/ response rate as a result of opposite perceptual learning in different social-indexing conditions (mean and standard error)	187
6.4	Exp 4a: /s/ response rate by condition and phoneme (mean and 95% confidence interval)	190
6.5	Exp 4b: /t/ response rate as a result of opposite perceptual learning in different social-indexing conditions (mean and standard error)	194
6.6	Exp 4b: /t/ response rate by condition and word (mean and 95% confidence interval)	196
7.1	Potential outcomes of perceptual learning with speaker A and B successively	204
A.1	Lexical decision results for /s/-containing words from Female A (Ashley) with different proportions of /ʃ/ blended into the fricative	215
A.2	Lexical decision results for /ʃ/-containing words from Female B (Vicky) with different proportions of /ʃ/ blended into the fricative	216
A.3	Lexical decision results for /t/-containing words from Female A (Ashley) . .	217
A.4	Lexical decision results for /t/-containing words from Male A (Gabriel) . .	218

A.5 Lexical decision results for /d/-containing words from Male A (Gabriel) . . . 219

Chapter 1

Introduction

Human speech is highly variable. In recent decades, the role of subphonemic phonetic variability has been highlighted for its facilitating role in speech comprehension and processing (e.g., Goldinger, 1996, 1998). The kinds of variability in consideration include both phonetic idiosyncrasies that are specific to unique tokens or random contexts (Bradlow et al., 1996), and global phonetic attributes that remain relatively stable with a particular talker (Eisner and McQueen, 2005; Kraljic and Samuel, 2006; Norris et al., 2003) or a speech community (Clarke and Garrett, 2004; Maye et al., 2008). A growing body of experimental studies shows that listeners make use of these kinds of variability in speech perception and processing, raising the possibility that phonetic specificity might be more closely interwoven into our linguistic knowledge than previously considered (see Pisoni and Levi, 2007, for a review).

Situated in this research background, this dissertation seeks a more precise understanding of the involvement of phonetic details in the phonetics-phonology mapping through the window of perceptual learning, a process by which listeners adjust their phonetics-phonology mapping in perception to be better aligned with the production of a particular speaker. I investigate how listeners adjust their perceptual expectations as a response to *speaker-specific* phonetic characteristics in a multiple-speaker setting. In such a context, phonetic tokens of different speakers are interleaved, and listeners need to go back and forth between their perceptual beliefs of different talkers to cope with talker variability. This dissertation investigates whether listeners form speaker-specific perceptual expectations that they track separately for different speakers or speakers of different genders. Or, alternatively, they

may integrate the phonetic characteristics of different speakers to update their perceptual beliefs within a holistic phonetics-to-phonology mapping system. I also investigate whether the answer to this question varies with different natures of the phoneme under discussion. Answers to these questions are expected to shed light on how social identity information is interwoven in the architecture of the phonetics-phonology mapping system.

1.1 Theoretical and empirical origins

Theoretical elaborations on the involvement of intra-category phonetic information in speech perception and processing are often framed in terms of a debate between *abstract* and *episodic* views of the mental representations of phonology. These debates revolve around whether phonological units are represented with discrete and invariant linguistic symbols (e.g., phonemes, Halle, 1985), or with a collection of specific acoustic instances or perceptual episodes (Goldinger, 1996, 1998). The discrepancy between these two lines of models essentially reflects a trade-off between emphasises on the structural versus the variable aspects of speech in speech processing. Prioritizing variability over structure leads to a prototypical exemplar-based approach, in which the representation of phonemes emerges from accumulated acoustic distributions of speech inputs in the listeners' experience. In this case, however, the convenience of handling variability comes at the cost of stronger encoding effort of individual acoustic instances encountered, and this type of encoding results in more complicated rules for the generalization of this knowledge to new scenes. On the other hand, a prototypical abstractionist approach addresses the phonetics-phonology mapping problem by mapping variable acoustic signals onto a symbolic phoneme system that does not retain substantial surface acoustic details. Theories of this kind have the advantage of providing straight-forward underlying systems of linguistic units with lower needs for memory. However, these advantages come at the risk of losing explanatory power towards phonetic variability in concrete social contexts. This approach still requires a better delineated theory of those phonetics-to-phonology mappings.

This section provides a review of the basic mechanistic principles underneath the op-

erations of these two kinds of speech perception models, as well as the historical research background that they are rooted in. Section 1.1.1 explains the fundamental principles underpinning the operation of “talker normalization”, a critical mechanism delineated by abstractionist models of speech perception that maps variable acoustic signals to a fixed set of phonemic units. Section 1.1.2 reviews the configurations of exemplar models and their alternative approach to deal with acoustic instances and category labels. Section 1.1.3 presents a brief review on the historical origins of these two theoretical views and outlines the progress of empirical findings in the area of speech perception that crucially motivates a shift of research focus from phonemic to subphonemic units in speech perception and processing.

1.1.1 Abstract speech representations and speaker normalization

The mapping between acoustic signals and linguistic categories is a many-to-many problem (e.g., Liberman et al., 1967). A certain kind of linguistic category can be realized in various ways in the acoustic space, and a particular acoustic pattern may correspond to several different linguistic interpretations. One classic example for the flexibility of mapping between speech acoustics and abstract phonemic categories is Peterson and Barney (1952). They measured the formant frequencies of American English vowels produced by adult men, adult women, and children, and found considerable overlap between the categories /ɪ/ (as in *bit*) and /ɛ/ (as in *bet*) in the vowel space defined by the first and second resonance frequencies of the vocal tract. Each phoneme generates many different acoustic values, and for many of the acoustic values they could plausibly have been generated by either phoneme. In perception, the nondeterministic relationship between signals and representational units may impose a challenge for listeners to maintain “perceptual constancy” (Shankweiler et al., 1977), which refers to the “stable recognition of the phonetic structure of utterances” (Nusbaum and Schwab, 1986, page 2).

Various approaches have been proposed to address how listeners manage to maintain phonetic constancy across talkers (see Shankweiler et al., 1977, for a comprehensive discus-

sion). One of the most influential mechanisms proposed to cope with this issue is “speaker normalization”, which is hypothesized to transform acoustic signals into mental representations that characterize a stable system of linguistic categories. The term “normalization” was first used to deal with categorization problems in the visual modality. To visually recognize an object, one needs to mentally rotate the input patterns, to expand or compact their size, and to translate them across spatial patterns prior to pattern matching, such that the input patterns can be mapped onto the right prototype template without being affected by properties unrelated to the basic pattern (Robert, 1965). In this case, “normalization” involves at least three fundamental types of elements: representations of information (mental templates), transformations of the representations (rotation, size expansion and compaction, spatial orientation), and control structures that determine the sequencing of transformations (Nusbaum and Schwab, 1986). Normalization essentially renders an input pattern into a canonical form for pattern matching through the operation of a set of passive and simple transformations in the order specified by control structures. Extended to the auditory modality, speech normalization presupposes a stable-state system of linguistic representations characterized by features resembling their acoustic properties. To continue with the previous example of vowels, phoneticians have attempted to characterize vowels by the loci of their first and second formants scaled according to the vocal characteristics of different speakers. After normalization for variances in the shape and size of vocal tract, the relational values of vowel formants should be sufficient to characterize different vowels and should stay invariant across speakers. Following this idea, Nordström and Lindblom (1975) proposed that such a uniform scaling method is adopted in vocal tract length normalization. In their view, speech is mapped onto a reference vocal tract by estimating the length of the speaker’s vocal tract and then scaling the formant frequencies as if they had been produced by the reference tract with a single scale factor based on vocal tract length. The mechanism of normalization by the virtue of vocal characteristics encoded in the signals is referred to as *intrinsic* normalization (Nearey, 1989). An alternative mechanism, referred to as *extrinsic* normalization (Nearey, 1989), refers to normalization by cues

derived from preceding contexts to constrain the interpretation of a target utterance (e.g., Joos, 1948). For example, Ladefoged and Broadbent (1957) had listeners identify synthetic /bVt/ words presented at the end of the precursor sentence “please say what this word is.” Six versions of the precursor sentence were synthesized, each with a different range of F1 and F2 that corresponds to different talkers. Identification of the target vowel was found to vary with the formant values in the carrier sentence. For example, the same token was perceived as /bit/ most of the time with higher F1 of vowels in the precursor sentence and it was perceived as /bet/ when the F1 of the precursor sentence were manipulated to be lower. Essentially, both of these mechanisms suggest that the relational values characterizing vowels in a metric space can be fixed after filtering out variances associated with the vocal characters of specific talkers - they only differ in their preference of what measure to use as a proxy of speaker’s vocal characteristics.

Although the concept of speaker normalization has been widely accepted and used in many contexts in speech perception, researchers have not really succeeded in coming up with a general transformation algorithm that yields a close approximation to the relational invariance between linguistic units, or vowels, in this context. Peterson and Barney (1952) evaluated whether the invariant relation can be preserved by linear scaling by examining whether the formant frequencies of two speakers’ vowels are constant multiples of one another; however, they found that such relationship does not hold. Later, Ladefoged (1967) tried to reduce variability by employing phoneticians as talkers, and found that the separation of all vowels cannot be attained by scaling the F1 and F2 in linear or nonlinear fashions, neither as on a scale of mels. Finally, I want to note that it is commonly assumed that speaker normalization is a crucial process for understanding *new* talkers that one has not encountered before (e.g., Joos, 1948; Ladefoged, 1967). In this process, a stretch of a new talker’s speech will be required for deriving the calibration parameters for normalization, although it remains a mystery how short the minimally required speech samples are. The integration of adaptation-level perspectives in speaker normalization theories provides a natural transition to the main research question of this dissertation – perceptual learning,

which I am going to review in more detail in Section 1.2.

1.1.2 Episodic speech representations and exemplars

Different from abstractionist views that presuppose explicit phonological units, exemplar views usually consider phonetics-phonology mapping as a process where linguistic structures emerges based on episodic memories of acoustic instances. The basic operation of exemplar model as “to categorize perceptual objects by evaluating the similarity between the item to be categorized and a set of stored category exemplars” (Johnson, 1997, page 105). Johnson (1997) lays out the concepts of this approach and developed a computational implementation to apply it in the area of speech perception. According to Johnson’s model, speech waves are first converted into a sequence of auditory spectra. Then, each auditory spectrum is vector-quantized and compared with a set of stored spectra represented on an “exemplar covering map”. If the incoming spectrum is not similar to any of the existing ones, then the system creates a new representational localization for it. If the spectrum is similar to one of the stored spectra, then the system returns the index of the stored spectrum and the localization of the stored spectrum shifts slightly towards the new spectrum. The third stage compares. Finally, category nodes (which are words in Johnson’s model) are represented as labels over the map, with each label associated with a frequency distribution of stored spectra of that label. Word node activation is the product of similarity to the covering map location and the association between that location and word node weighted by the frequency of exemplars at that location.

Pierrehumbert (2001) provided another computational implementation of phonological activation using exemplar theory. In this model, linguistic categories (which are equivalent to phonemes instead of words as in Johnson’s model) are associated with a cloud of percepts or exemplars, and the strength of the activated exemplars cumulates in activating category labels. Therefore, more frequent categories have more exemplars and more highly activated exemplars than less frequent categories. Pierrehumbert’s model makes progress on the meticulous delineation of the cognitive processes involved in phonological activation. In her

model, the activation strength of exemplars depends on both their frequency and recency, such that more recent exemplar has higher strength than temporally remote ones, and this recency-associated activation gradually decays as time elapses. Her model also describes the competition and inhibition of linguistic labels in the classification stage. The activation level of a specific label for a newly entered token is derived from the similarity between the novel token and the exemplars associated with the target label as well as the aggregate activation strength of exemplars under that label. In the case of perceptual ambiguity where a token is equally similar to two categories of exemplars, the category with more numerous or more activated exemplars has advantage in the competition, or in other words, the model predicts a bias towards a high-frequency label. In addition, Pierrehumbert (2001) also sketches how the model would extend to speech production.

This approach provides a way to capture the influence of fine-grained phonetic instances on speech behaviors. For one thing, it accounts for findings of the role of frequency in linguistic activation relatively well. For another thing, it does not need to pre-specify a set of linguistic structures or units to start with, which provides explanation power for addressing acquisition problems. In first-language acquisition, children do not have explicit phonological knowledge demonstrated to them; but rather, they need to figure out the structure themselves with limited exposure to language input. Moreover, this approach provides a way to respond to typological results showing that languages differ in arbitrarily fine phonetic detail, because it assumes that a distribution can be learned at an arbitrary location on the phonetic map with a substantial amount of exposure.

1.1.3 A shift of research foci in the literature

The abstract view and the episodic view to the mental representation of speech are rooted in different historical backgrounds and motivated to address different sets of findings. Through a brief recap on the origins and motivations of these different views, I hope to show that the research endeavors to obtain and interpret a wider range of empirical findings have brought rapid progress to our understanding of speech perception, and the focus shift of the research

question is a natural result of this progress. In other words, the current exploration of the role of within-category phonetic details is an extension rather than a setback of the earlier explorations of the structure of language.

The theoretical background of the conventional symbolic views goes back to the middle and late twentieth century. The primary debates concerning sublexical units back then did not revolve around how phonetic details are incorporated in the mental representations, but whether word-internal units are necessary at all. Several theories have proposed approaches to directly access lexical items either based on either spectral envelopes of the acoustic signals (Klatt, 1980) or linguistic adjacency patterns such as neighborhoods (Luce et al., 1991) and cohorts (Marslen-Wilson, 1987), without an intervening layer of phonemes. “Phoneme” as a linguistic unit was mainly the concern of formal phonology, in which, discrete representational units are the nuts and bolts for systematic derivation of sound regularities (Chomsky and Halle, 1968). Other than that, intuitions about phonemes were rather limited, and primarily came from rhetoric and orthographic conventions such as the existence of spelling and writing systems and rhymes of poetry (e.g., Liberman et al., 1974).

In the late twentieth century, more empirical findings in favor of word-internal structures arose from the areas of speech production and experimental psycholinguistics. Studies in the former showed that speech errors frequently occur in forms of substitutions and exchanges of single segments (Fromkin, 1973; Garrett, 1976; Shattuck-Hufnagel and Klatt, 1979), and those in the latter showed that listeners are able to detect phonemes in words (Foss and Swinney, 1973) and nonwords (Foss and Blank, 1980; Foss and Gernsbacher, 1983), as well as insert, delete and move around sounds as discrete units to form novel words (Treiman, 1983, 1985). Together with insights from formal phonology, rhetoric and orthography, these findings motivate a symbolic view of speech units, according to which speech signals are converted into linear combinations of discrete, abstract, symbolic units like phonemes, and then these combinations feed forward to the activation of higher-level units such as words. Extreme symbolic views see abstract regularities as emerging from computational processes at the time of retrieval (Joos, 1948); phonetic details were discarded as soon as it has been

normalized out and were kept away from the core linguistic processing.

While empirical findings on language users' sensitivity to sublexical segments are well addressed under a symbolic view, experimental works have moved forward to examine the influence of sublexical phonetic realizations from a more fine-grained level, namely, the speaker-or-context-specific phonetic realizations of phonemes and words. In the 1990s, empirical research demonstrated a processing benefit from the recurrence of the same acoustic instances or speaker voice. Regarding the latter, the general pattern is that, compared with hearing speech from multiple speakers, hearing speech from a single talker benefits the listeners with faster processing speed (Church and Schacter, 1994; Mullennix et al., 1989), higher intelligibility (Nygaard and Pisoni, 1998) and longer durability in memory (Martin et al., 1989; Palmeri et al., 1993). This effect has been examined with experimental tasks including word list recall (Martin et al., 1989), recognition memory (Palmeri et al., 1993), auditory priming (Church and Schacter, 1994), perceptual identification (Goldinger, 1996; Mullennix et al., 1989) and intelligibility tests (Nygaard and Pisoni, 1998).

In the meantime, researchers began to realize that the conventional symbolic view was focused on the influence of the regularities rather than idiosyncrasies of the sound system. It was not meant for addressing processing consequences caused by nuanced acoustic variability. Episodic views were motivated to bridge the gap that the influence of nuanced phonetic variability in linguistic processing was becoming recognized but not theoretically established. Such approaches depart from traditional views to increase the encoding strength of acoustic instances in the representations of linguistic structures (Goldinger, 1996; Nygaard and Pisoni, 1998), in order to cope with consequences of variability.

The symbolic view is a step taken towards a more explanatory analysis of sublexical structures as representational units, whereas the episodic view is intended to obtain a more precise understanding of where phonetic details might also become relevant to linguistic representations. Logically, the two views are not mutually exclusive – abstract units and stored acoustic information can coexist. Although strong episodic theories have sometimes claimed that phonology directly emerges from episodic representations without intermediate

discrete units such as phonemes, that point of view is not automatically shared by all the studies investigating episodic subphonemic details. Phoneme is acknowledged in some of their frameworks (e.g., Pisoni and Levi, 2007) and exists in the form of exemplar clouds clustering around the most frequent tokens in episodic models (e.g., Pierrehumbert, 2001). Nowadays, most model usage-based approaches do involve phoneme-level representations (Beckner and Bybee, 2009; Harrington et al., 2018; Hay and Foulkes, 2016; Todd et al., 2019). Such approaches assume that the instances of the same phoneme in different words may have different but related representational bases, which increases their explaining power towards phoneme-centered linguistic phenomena such as regular sound change.

Along the exploration trajectory down from word to phoneme and to subphonemic variability, plenty of questions and concerns have arisen from inconsistency between experimental results and uncertainty of accurate interpretations of these results. In specific, the robustness of facilitating effects associated with phonetic details may vary with different kinds of the stimuli (González and McLennan, 2007; Luce and Lyons, 1998; McLennan and Luce, 2005), different hearing conditions (Jackson and Morton, 1984; Schacter and Church, 1992) and the examined time course (Wilder, 2018). This inconsistency suggests that more precise delineations are needed regarding the scope and conditions under which phonetic variability of particular kinds makes a difference. On the theoretical side, it remains unclear what kind of phonetic information is retained in representations, how much information is there, and whether there are different mechanisms responding to different contrasts, tasks and time-course. These questions are still to be resolved.

1.2 Perceptual learning as a testing ground

This dissertation takes “perceptual learning” (also called “perceptual recalibration”, or “perceptual retuning”) as a way to probe how listeners cope with the phonetic characteristics of idiosyncratic speakers in a multiple speaker setting. As a psychological term, “perceptual learning” was originally used to refer to the general adaptation mechanism to stimuli with different kinds of features represented at various levels of mental abstraction.

For example, Gibson and Gibson (1955) defines perceptual learning as the mental process of “making the perceiver more sensitive to the variables of the stimulus array” (page 40). Goldstone (1998) defined perceptual learning as “relatively long-lasting changes to an organism’s perceptual system that improve its ability to respond to its environment and are caused by this environment” (page 586).

In the last decade, perceptual learning studies have proliferated in the field of speech perception. Within the speech area, however, the term of “perceptual learning” is still used in different lines of literature that focus on different aspects of this issue. Samuel and Kraljic (2009) sorted studies on the perceptual learning of speech into two broad themes. One line of these studies focuses on the phenomenon that exposure to certain types of stimuli that the listeners are unfamiliar with (e.g., nonnative speech, accentual speech, degraded speech) leads to improvement in listeners’ ability to comprehend or identify speech stimuli of that type (e.g., Bradlow and Bent, 2008; Clarke and Garrett, 2004). The other line of studies examine how exposure to non-canonical speech make listeners more attuned to the acoustic distribution of tokens in the prevailing environment (e.g., Kraljic and Samuel, 2005, 2006; Norris et al., 2003). The basic procedure is to present listeners with phonetically ambiguous stimuli with contextual information to disambiguate the stimulus. Perceptual learning is then measured by a shift in the categorization boundary between the two phonemes such that the phonetic space assigned to the contextually favored phoneme is expanded. This dissertation investigates the second of these types of perceptual learning. Again, experimentally, this kind of perceptual learning is quantified by the amount of adjustment listeners make to their perceptual category boundary between two phonemes after exposure to ambiguous stimuli with contexts that favor the perception of one of those phonemes. For example, if we use contextual information to make listeners believe that a non-typical /s/-like sound is just an /s/, they may expand the range of acoustic inputs they accept as /s/ to account for the new sound. As discussed in Samuel and Kraljic (2009), a great virtue of this paradigm is that “the observed category boundary shifts provide a clear indication of exactly what is changing in perceptual processing as a function of experience” (page 1208).

One of the most important purposes of speech perceptual learning is to facilitate perception efficiency by adapting to speech variability. It is well established that there is a substantial amount of speech variability in the speech production of speakers of the same language. Some of these variations may reflect stratification in social identity, while others are idiosyncratic traits of individual speakers. Abundant studies have shown that listeners adapt rapidly to accommodate idiosyncratic speech properties of a particular talker. However, it remains a mystery what listeners do with this information after they encounter a different speaker. We come up with several possibilities: Either listeners toss out what they have learned and establish a new set of perceptual expectations for the current speaker from scratch, or they apply their retuned perception criterion immediately to the next speaker they encounter, or they store the outcome of perceptual learning for future use when they encounter somebody similar. These questions have been investigated by previous experimental studies with different stimulus materials and paradigms. However, their observations are not always consistent with each other, not to mention that there is still room for different interpretations of the same set of results.

The remainder of this section provides a review of what we already know about the *mechanistic* characteristics of the operation of perceptual learning from previous research, including how much input is needed to induce a boundary shift (flexibility), how long a perceptual learning effect lasts (durability), and whether the perception adjustment generalizes to other speakers and contexts (generalizability) or stays specific to the speaker whose speech triggers the shift (specificity). Section 1.2.1 reviews findings of different studies pointing to the flexibility and durability of perceptual learning, which are two seemingly contrastive properties. Section 1.2.2 reviews previous findings lending support to another pair of contrasting characteristics – specificity and generalizability – of perceptual learning. In both of these sections, I further discuss whether these discrepancies are really incompatible or can be reconciled at some level. Note that the theme of the current dissertation is in fact more directly linked to the second pair of contrasts, since I investigate how perceptual learning behaves in respond to switches between talkers. Finally, Section 1.2.3 discusses the

possibility of reconciling the generalization and talker-specific aspects of perceptual learning by resorting to listeners knowledge of structure in talker variability (see e.g., Kleinschmidt, 2019; Kleinschmidt and Jaeger, 2015), with explanations of how this idea works and how I plan to test this idea in the present dissertation with experimental approaches.

1.2.1 Flexibility vs. durability

It has been mysterious whether and how flexibility and durability simultaneously apply to perceptual learning. In terms of flexibility, researchers have found that listeners usually adjust their perceptual expectations fairly quickly in response to only a few acoustic tokens. In perceptual learning experiments, perceptual learning is typically induced with 40 critical word stimuli, namely, 20 for each of the two target phonemes (e.g., Eisner and McQueen, 2005; Kraljic and Samuel, 2005, 2006). More extreme cases show that perceptual shift can be caused by exposure to as few as 20 critical lexical stimuli, 10 for each phoneme (Kraljic and Samuel, 2007). Another piece of support is that listeners are willing to make a further perceptual adjustment after subsequent exposure to additional speech materials, even when the acoustic distributions of those materials are at odds with those in their previous training. Saltzman and Myers (2018) exposed listeners to four interleaved blocks of lexical decision that were designed to bias perception in opposite directions along a /s-/ continuum, and recorded the distribution shift with a phonetic categorization task after each lexical decision block. They found that, in each session, listeners' perceptual bias was consistent with the cue distributions in the immediately preceding lexical decision block. These results all suggest that the system of human speech perception is highly sensitive to newly encountered acoustic instances and can make rapid adjustments accordingly.

Meanwhile, findings pointing to the durability of perceptual learning show that listeners are able to store the effect of perceptual learning in their mind for a long time, such that they still maintain the perceptual shift on their second visit after hours and even days. Kraljic and Samuel (2005) have shown that learning effects are reliable after a 25-min interval, unless listeners are exposed to unambiguous tokens of the critical sound that come from

the voice of the exposure talker. Eisner and McQueen (2006) have subsequently shown that the retuning can persist for twelve hours, regardless of whether subjects sleep in the interim. Pushing the examination interval even longer, Zhang and Samuel (2014) had listeners participate in two perceptual learning sessions separated by a full week. Listeners were assigned to two groups who received perceptual learning towards opposite directions on the first and second visit, e.g., treating an ambiguous sound halfway between /s-f/ as /s/ on one visit and /f/ on the other visit. They found that only the first session, not the second, showed the expected difference between groups. Zhang and Samuel interpreted the absence of boundary shift in the second session as reflecting dissipation of the first shift after a week and inhibition of the expected second shift by previous exposure to good tokens one week ago. Therefore, they concluded that the time limit of the perceptual learning effect should fall into the interval between 12 hours and a week.

I want to briefly note here that the above discussion brings up several other dimensions affecting the encoding strength of different acoustic instances. One of these factors is the standardness or typicality of the instance compared to other exemplars within a phoneme category. The inhibition effect of exposure to standard tokens on later perceptual learning, mentioned in the above case of Zhang and Samuel (2014), essentially suggests that tokens at different typicality levels are associated with different encoding strengths. This point is also supported by several other studies. Hay et al. (2015) proposed a mechanism that discards tokens or prevents their encoding in memory when that recognition process involved excessive ambiguity, even in cases with ultimately accurate recognition. In other words, exposure to items of more extreme category atypicality is less likely to cause perceptual adjustment because they are not fully encoded in the first place. Babel et al. (2019) shows that there is still fine-grained covariation between the degree of typicality and the extent of perceptual learning in less extreme cases. They found that, among nonstandard instances, the perfect ambiguous instances cause greater perceptual learning than atypical instances, and the latter still causes greater perceptual learning than remapped instances (i.e., instances sounding like a different phoneme category).

Another hidden question relevant to the flexibility vs. durability of perceptual learning is “how does the encoding strength of recently encountered instances and temporally remote instances vary?” As discussed earlier in this chapter, exemplar models (e.g., Pierrehumbert, 2001) consider the encoding strength of acoustic instances to be determined by their frequency and recency¹ and predict higher strength associated with more recently encountered tokens. This view is supported by Saltzman and Myers (2018) (reviewed at the beginning of this subsection), where the authors found that the perceptual bias induced by the most recent training completely overrides the previous one. Saltzman and Myers (2018) therefore conclude that listeners rely more heavily upon the most recent information received and down-weight older, consolidated information. In the meantime, empirical findings by other studies point to slightly different points of view. Theodore and Monto (2019) conducted a similar study with *unsupervised* learning of /k-g/. Their experiment consists of two blocks of categorization tasks. Half of the listeners were exposed to the narrow distributions of VOT variances followed by the wide distributions, and the other half of them had the order reversed. The result of the earlier block shows a difference in identification slope between groups who encountered different VOT distributions, but this difference was attenuated at the end when everyone heard all of the stimuli. This result was interpreted to show that listeners did not disregard prior experience with a talker, but rather used cumulative statistics to guide phonetic decisions. However, the between-group difference in Theodore and Monto (2019) is not completely gone at the final stage, which means that the cumulative statistics listeners learned was still mediated by the sequential properties of the training instances.

Although the influence of the sequential/temporal aspects of the training stimuli is not the primary concern of this dissertation, a good understanding of this issue could help us make wise methodological decisions in experiment design and interpret the results more accurately. Other than these purposes, the tension between flexibility and durability of

¹Such theories also tend to acknowledge greater strength of tokens with higher salience caused by social, attentional, and experiential factors. See Sumner et al. (2014) for example of social salience and see Jaeger and Weatherholtz (2016) and Lai et al. (2020) for salience associated with unexpectedness or surprisal.

perceptual learning is an independently interesting question for many reasons. It is one of the key assumptions that Bayesian models of perceptual learning need to include in order to model the adjustment of perceptual expectations over time (e.g., Kleinschmidt and Jaeger, 2015). In addition, it is a crucial concern for models of diachronic sound change because they make different predictions about whether novel or nontypical linguistic variants can survive long enough to be re-used by the listeners and spread into the speech community (Tamminga et al., 2020). Through an overview of the two lines of findings reviewed in this section, it is unclear whether the properties of flexibility and durability necessarily contradict one another. Flexibility speaks to the operation of perceptual learning in response to acoustic *input* from a specific speaker and in a specific context, whereas durability refers to the state of perceptual learning *outcome* without discussing further exposure to additional training materials. These two properties delineated such a perceptual learning mechanism: It is flexible to the acoustic distributions of incoming speech materials when relevant speech input comes in; it stops operation and stores the output once relevant speech inputs are suspended and will not start operating unless relevant speech input becomes available again. By thinking that they are contradicting, one actually presupposes that the speech instances a listener encountered in any context and at any time point goes to a single set of training input that are essentially relevant to each other, which is not necessarily the case. The two properties could become compatible if we think of separate perceptual learning processes in charge of adaptation to a set of acoustic instances encountered in specific contexts or from specific speakers. Then, how do listeners divide up the acoustic instances they have encountered decide which ones go together? This question is to be further unpacked in the next two subsections.

1.2.2 Specificity vs. generalizability

In real-world perceptual learning, listeners do not cope with only one kind of speech variability at a time, and linguistic variants would not come from only one particular category of speakers or contexts. As more of these factors come into play, the perceptual learning

mechanism unavoidably needs to make decisions about whether it needs to extend linguistic distributions they have picked up in one situation to a different situation, which may involve different speakers, different phonemes, or different types of communication tasks. Whether and how broadly does perceptual learning generalize? This question has been attracting a substantial body of research, but the answers provided by different studies are somewhat debatable.

Some studies do find that perceptual learning may generalize to the perception of different speakers (Kraljic and Samuel, 2005, 2006, 2007; Reinisch and Holt, 2014; Xie et al., 2018) and different phonemes with similar contrast (Kraljic and Samuel, 2006; Weatherholtz, 2015). Regarding generalization across speakers, many studies have found that exposure to a single speaker's speech is sufficient to make listeners apply their knowledge of how that speaker sounds to other speakers (Kraljic and Samuel, 2005, 2006, 2007; Reinisch and Holt, 2014). Studies in this line have also examined the effect of training with multiple speakers, and they find that exposure to multiple talkers with the same kind of pronunciation characteristics can promote the perceptual learning generalization to other talkers (e.g., Bradlow and Bent, 2008). In a slightly different vein, studies have also examined whether listeners generalize their knowledge of the acoustic realizations of phonemic contrasts that listeners have learned for one pair of phonemes to a different pair of phonemes with the same contrast. Kraljic and Samuel (2006) find that listeners who were trained on the VOT of /t-d/ generalize what they learned to /p-b/, which are a pair of new phonemes that are previously unheard but share the same voicing feature distinction. Weatherholtz (2015) examined the generalization of perceptual learning of a vowel chain, which shifts /i/ to [ɪ], /I/ to [ɛ], /ɛ/ to [æ], and /ae/ to [ɑ]. Weatherholtz find that listeners are able to generalize learning to fill in incidental gaps in their experience, indicating that listeners learned the covariation pattern between vowel categories, rather than learning each constituent shift independently. However, this finding is not replicated in their experiment on the perceptual learning of back vowel raising, raising questions about the robustness of the systematic learning of the vowel space. These findings suggest that it is possible for perceptual learning to generalize,

as one would expect for real-world listening with complicated linguistic co-varying factors.

Nonetheless, cases of failures to induce perceptual learning generalization are not uncommon (and maybe even more frequent) in this body of literature. As it is implausible for listeners to acquire the distributional properties of all the linguistic variants they encounter from different speakers and contexts, it is equally implausible for listeners to generalize what they have learned without distinction. For example, it is not likely for one to generalize their perceptual learning of nonstandard (e.g., nonnative) accents to every other speaker they encounter, if the encountered speaker does not share the same pronunciation characteristics. Also, listeners rarely need to generalize their perceptual learning of newly encountered speakers to those they have known for a long time. This is because listeners have already accumulated sufficient linguistic exposure from those they know well and have established stable and reliable perceptual expectations for them. Papers that do not find evidence for the generalization of perceptual learning across speakers would argue for a “speaker specificity” view of perceptual learning. That is, listeners only apply the acquired perceptual adjustment to the perception of the particular speaker who that causes it. Note that in this context, experimental conditions on different speakers commonly use stimuli in a male voice and those in a female voice to differentiate between speakers. Therefore, many of the findings about the “speaker specificity” of perceptual learning essentially reflect the constraint of different speaker genders on the generalization of perceptual learning. One of the first studies taken to provide evidence for talker specificity in perceptual learning is Eisner and McQueen (2005). They find that perceptual learning of a fricative boundary does not arise when listeners are trained on stimuli from a female voice but tested on stimuli from a male voice. They conclude that “the perceptual adjustment investigated here does not generalize across talkers” Eisner and McQueen (2005). Another result that has been cited as evidence for talker specificity comes from Kraljic and Samuel (2005). This study adds an “unlearning” phase in between the training and test phases, in which listeners sometimes hear additional spoken input that either contains no cases of the critical phonemes or contains non-ambiguous instances of the critical phonemes as a form of corrected input. Kraljic

and Samuel found that only when the unlearning input is presented in the same talker's voice does the perceptual learning effect become attenuated. The learning is unaffected by the unlearning phase with speech of a different talker (of a different gender).

These findings seem to suggest that the specificity of perceptual learning is conditioned on speaker identity/gender as signaled by voice. However, this is not the case for the two studies above. In each of these studies, the authors simultaneously show that if a target pair of phonemes are sufficiently acoustically similar between two speakers, listeners are still able to generalize across speakers regardless of their different gender voices. In another experiment of the Eisner and McQueen (2005) study, they find that listeners generalize the perceptual learning of a female speaker's pronunciation of /s-ʃ/ to an /ɛs-ɛʃ/ continuum when the vowel /ɛ/ is spoken by either a male or a female novel speaker, as long as the fricatives were from the original talker's speech. When the continuum was created entirely from the speech of a novel talker, there was no perceptual learning, unless the novel talker's fricatives had been spliced into the original talker's speech during exposure. Although this result is sometimes taken as a support for the speaker specificity of perceptual learning, it essentially highlights the specificity of the productions of the target phonemes, rather than speaker identity as indexed by voice. Kraljic and Samuel (2005) find that perceptual learning with a female speaker's fricatives can be transferred to a male voice in the test phase, but this does not work backwards: listeners do not generalize their perceptual learning with the female speaker to the male speaker's fricatives. They interpret this result as caused by the presence vs. absence of acoustic overlapping between the two sets of fricatives in training and test. The female speaker's training fricatives have COGs that lie within the COG range of the male speaker's test fricatives, whereas the COGs of the male training fricatives are distinct from those of the female speaker's test fricatives. They thus conclude that listeners track the acoustic properties of each speaker's fricatives and apply generalization whenever there is a sufficient match. These findings raise the possibility that the previous "speaker specificity" effect is not essentially induced by different talker genders (or identities), but rather, it reflects the constraint of acoustic similarity between the phoneme productions of

different speakers on the generalization of perceptual learning.

The role of acoustic similarity is later echoed in Reinisch and Holt (2014). This paper differs from Kraljic and Samuel (2005) and Eisner and McQueen (2005) in that the ambiguous fricatives are embedded in Dutch-accented English and that they examined perceptual learning generalization both across and within genders. Instead of talker specificity vs. generalization, their question is about accent adaptation more broadly, namely, whether the presence of a foreign accent promotes generalization across talkers. Reinisch and Holt (2014) show that a lexically-guided /s-f/ boundary shift generalizes from a female training voice to a novel female voice, but does not generalize to a novel male voice without manipulating the acoustic similarity of the critical phonemes between the two speakers. This result not only supports the findings of Eisner and McQueen (2005) and Kraljic and Samuel (2005) in the role of acoustic similarity in cross-gender perceptual learning generalization. It also echoes another line of literature pointing to a promotion of the generalization of perceptual learning by the presence of an entirely different accent or cluster of features. (e.g., Baese-Berk et al., 2013; Bradlow and Bent, 2008; Weatherholtz, 2015). Perceptual learning studies from an accent-adaptation angle generally report quite robust generalization, although they sometimes differ in how consistent the accents of different speakers need to be for there to be generalization. Bradlow and Bent (2008) showed that adaptation to a novel talker of Mandarin-accented English but not to a talker of Slovakian-accented English following exposure to multiple talkers of Mandarin-accented English. This learning is not talker-independent but accent-dependent, for the multiple talkers from a single language background. However, Baese-Berk et al. (2013) examines training on talkers from five language backgrounds and finds that listeners generalized their learning to novel talkers from language backgrounds both included and not included in the training set. These findings suggest that generalization of foreign-accent adaptation is the result of exposure to systematic variability in accented speech that is similar across talkers from multiple language backgrounds.

Given the evidence amassed so far, it seems likely that generalization of perceptual

learning is itself flexible, occurring across some speakers under some circumstances. The factors determining when perceptual generalization occurs may include speaker gender, voice similarity, and accentedness. The last constraining factor on perceptual generalization that I review here is the type of phonological contrast. Up to this point, the phoneme in question is always a fricative. When the same question is tested on stops, different results are obtained. Kraljic and Samuel (2006) evaluate whether listeners generalize their perceptual learning with ambiguous stops between /t/ and /d/ to a novel speaker of a different gender. They find that listeners show equally robust perceptual learning effects for the same speaker and for the new speaker. Kraljic and Samuel (2007) expose listeners to blocks of words from a male voice and a female voice with opposite directions of potential retuning, to investigate whether listeners could simultaneously show learning for the pronunciations of more than one talker. Again, they find that the perceptual learning is speaker specific only when the critical phonemes are fricatives; for stops, the perceptual learning result reflects the most recent pronunciation heard, regardless of the speaker. In discussing these results, Kraljic and Samuel suggest that “when the to-be-learned phoneme highlights a temporal-voicing contrast that does not provide local, acoustic cues to speaker, as in our stop manipulations, learning will be speaker-independent. But when it highlights a spectral-place contrast that does acoustically distinguish one speaker from another, as in one of our fricative manipulations, learning is speaker-specific” (2007, 3).

1.2.3 Perceptual generalization by structures in talker variability

In the previous section, I have reviewed experimental findings showing that perceptual generalization is sometimes inhibited across certain kinds of speaker groups or on certain types of phonemes. Further, these findings shed light on the potential structures of speakers or phonemes underpinning the function of this process. Regarding the difference between phoneme classes in the generalization of perceptual learning, one of the most intriguing explanations is that fricatives contain more information about speaker identity than stops. I have reviewed Kraljic and Samuel (2007)’s point in the last section that fricative variability

allows them to learn to restrict the boundary shift to a particular speaker’s productions, whereas stop variability provides less information such that it is merely adequate to support learning at the broader featural level (see also Allen and Miller, 2004; Newman et al., 2001, for discussion about the informativity of fricatives and stops).

A related proposal is found in the “ideal-adapter” framework (Kleinschmidt, 2017, 2019; Kleinschmidt and Jaeger, 2015), which posits that listeners should generalize across speakers according to the social grouping that conditions variability in speech production. This proposal is dependent on the premise that listeners have good knowledge of the speech properties of speakers from different sociophonetic speaker groups. The speaker structure in listeners’ knowledge is not purely determined by social-demographic criteria. But rather, it may be a reflection of the more nuanced co-variation between speakers’ phonetic properties and social aspects in different possible dimensions. Then, listeners use the information of speakers’ sociophonetic group to make predictions about the characteristics of speakers’ productions², including which speakers sound alike. Their knowledge of structure in talker variability is obtained from listeners’ experience with the real-world sociophonetic speech variability. In the above cases with fricatives, since men and women on average produce fricatives (especially /s/) with different spectral peaks (e.g., Jongman et al., 2000), listeners should use information about speaker gender to categorize fricatives when they encounter new talkers, and therefore should not transfer what they learn about a male talker’s fricatives to a female talker or vice versa. Meanwhile, this model also predicts that if women tend to produce broadly similar fricatives, listeners might not maintain a separate mental model for the fricatives of each individual woman they encounter. This prediction also has some experimental supports (Reinisch and Holt, 2014; Tamminga et al., 2020), which will be reviewed in more detail in relevant chapters. The predictive power of the ideal-adapter framework can be used for many other types of phonemes in addition to fricatives and stops, as long as speech data are available and contain the productions of the phoneme in question produced by speakers of various backgrounds. Through computational modeling based on

²This framework also works to predict speakers’ social group from their phonetic productions, but this aspect is not directly relevant to the research question of the dissertation and is not reviewed here.

speech corpora, Kleinschmidt calculates informativity of speakers' gender and dialect on their productions of vowel formant and VOT length. He shows that gender and dialect are more informative about speakers' vowel formant frequencies than VOT length; among them, gender is more informative about the absolute vowel formant frequencies, whereas accent is more informative about the relative distribution of vowels in normalized space.

In a nutshell, models of perceptual generalization based on structure in talker variability propose that listeners have knowledge of sociophonetic similarity between speakers and use it to direct the storage, retrieval, and update of the mental distributions of instances associated with a phoneme label. How well this kind of model captures human speech perception will depend in part on empirical findings of presented in this dissertation. Before diving into more details of the specific assumptions and principles underpinning such a model, I would like to first articulate why mechanistic properties of perceptual learning (flexibility vs. durability, specificity vs. generalization, etc.) are of interest to some of the theoretical questions outlined in Section 1.1. They have implications for a critical issue at the heart of speech perception and processing, namely, the architecture of the phonetics-phonology interface. Research in this area aims at identifying and fixing the missing links between language structure and use, by pushing on questions like whether and how many acoustic details are stored in listeners' memories, whether phonetic knowledge is stored and represented as episodic exemplars or abstract prototypes, and how these representations respond to typical and atypical acoustic instances. In our case, conclusions of the specificity versus generalization of perceptual learning may point to opposite views of how the phonetics-phonology mapping system works. Specificity seems to suggest that that phonetic details are dynamically incorporated into listeners' mental representations with rich, context-marked experiences, whereas generalization tends to implicate that phonetic instances feed into a unique representational system of phonological categories, one that tracks acoustic distributions of tokens under specified categories but leaves out the information of the speakers and contexts of those tokens. Based on the above findings and insights, the present dissertation seeks to provide a more accurate delineation of the mechanistic properties of perceptual

learning and further clarify how the amassed empirical findings both in previous literature and here speak to the involvement of knowledge about talker-variability structures in the operation of perceptual learning.

1.3 Research questions and goals

The broad goal of this dissertation is to investigate the intriguing possibility that the generalization of perceptual learning across speakers within and between social groups reflects listeners' sociophonetic knowledge that mirrors the structure of real-world speaker variability. However, as one of the first research endeavors to examine this proposal systematically, the specific goal of this dissertation is to evaluate one of the most critical empirical supports for talker-variability based generalization models, namely, that the perceptual learning of fricatives is more speaker-specific than that of stops. Although this claim has been implicated by the results of several separate studies reviewed earlier, these studies are not designed to evaluate such a proposal specifically. They are intended to either evaluate relevant questions of different scopes (e.g., "whether perceptual learning is talker-specific") or evaluate a different set of questions (e.g., accent adaptation). Therefore, the proposal of structuralized speaker variability is often drawn upon as a post-doc explanation rather than a research hypothesis. As a result, a number of missing links can be identified between the results of these studies and the ultimate claim about different levels of speaker specificity between fricatives and stops. For one thing, none but one of these studies includes a comparison between perceptual learning of different phoneme types. For another thing, most of them suffer from the confound between specificity to idiosyncratic speakers versus speaker genders. This dissertation is aimed at filling in these missing links and providing a more comprehensive picture of the generalization of perceptual learning across speakers within and between gender groups for sibilants and stops.

The research questions of this dissertation are threefold in general. The first line of inquiry regards the granularity of the social category that allows for generalization of the perceptual learning of sibilants across members within such a category. Three levels of

granularity are in consideration, namely, generalization across gender groups, generalization within gender group, and generalization within individual. Similarly, the second line of inquiry regards the granularity of the social category that allows for generalization of the perceptual learning of stop VOT across members within the category. In other words, is the perceptual learning of stop VOT talker-specific, gender-specific, or generalizable across speakers and genders? The third line of inquiry involves a comparison between these two types of phonemes, asking whether the perceptual learning of sibilant is more speaker- or gender-specific than that of stop VOT.

Table 1.1 provides a summary of previous perceptual-learning studies of relevance to the research questions outlined above. Because fricatives in general (instead of only sibilants) have been argued to be informative of speaker identity (Newman et al., 2001), I have included findings of fricatives in addition to those of sibilants in this review to gather a larger amount of evidence. Four possibilities are under consideration for each of the two phoneme types, namely, that the perceptual learning of phoneme type X is specific to individuals within gender/generalizable across individuals within gender/specific to individuals across genders/generalizable across individuals across genders. The enumerated studies may lend support to one or more of these four possibilities. Studies in different cells should not be simply considered as contradictory to each other, because these empirical findings need to be interpreted in combination with the specific contexts or conditions where they are observed. In our case, these conditions include whether perceptual learning is induced by distributional or lexical information, whether visual cues to speaker identity are presented, and whether the acoustic properties of the critical phonemes are similar between the training and the test phase.

The remainder of this section discusses the specific research questions to be addressed in this dissertation. Under the heading of each of these questions, I will walk through some of the puzzling areas and vulnerability of previous studies, and discuss where there is room for further investigation. I will also briefly introduce the corresponding experiments associated with each of these questions, as well as what I expect to learn from the results

		Individuals across genders	Individuals within gender
fricative	specific	Kraljic and Samuel (2005, 2007); Tamminga et al. (2020) [◦]	none
	generalize	Eisner and McQueen (2005) [†] ; Reinisch and Holt (2014) [†]	Reinisch and Holt (2014); Tamminga et al. (2020) [◦]
Stop VOT	specific	Munson (2011) [◦]	Allen and Miller (2004) [*] , Theodore and Miller (2010) [*] , Myers and Theodore (2017) [*]
	generalize	Kraljic and Samuel (2006, 2007)	none

* indicates the availability of visual cues to speaker identity. † indicates the dependence of results on acoustic manipulation of the target phonemes. ◦ indicates distributional learning.

Table 1.1: Previous findings of the specificity versus generalization of perceptual learning across speakers and genders for fricatives and stops

of this dissertation.

1.3.1 Is the perceptual learning of sibilant specific to talker or gender?

Under this line of inquiry, my first question regards whether the perceptual learning of sibilants generalizes across speakers of **different genders** (Q.1). Although Kraljic and Samuel (2005, 2007) have provided some evidence that perceptual learning of fricatives does not generalize across speakers of different genders, they still leave several questions unsolved. Kraljic and Samuel (2005, 2007) added an unlearning phase between the training phase and the test phase, which raised the confound of whether the absence of the learning effect is due to an offset between two perceptual learning shifts towards opposite directions or disposal of a previous perceptual shift after encountering a new voice. Tamminga et al. (2020) evaluate perceptual shift induced by distributional properties instead of lexical contexts, with the latter being more common in the context of the perceptual learning literature. This question will be investigated in Experiment 1 of the current dissertation. Based on Kraljic and Samuel (2005, 2007), I cope with the confound between offset and disposal by adding more conditions of the intermediate learning phase that bias perception towards different directions, either contradictory or consistent with the first phase of training. The results are expected to provide a clearer answer to the question of specificity versus generalization of sibilant perceptual learning across genders.

Table 1.1 also seems to present a discrepancy between findings of perceptual specificity (Kraljic and Samuel, 2005, 2007; Tamminga et al., 2020) and those of generalization (Eisner and McQueen, 2005; Reinisch and Holt, 2014) of fricatives. However, these two lines of findings are not really contradictory because they occur in different situations. In particular, findings of perceptual generalization of sibilants only occur when additional manipulation was implemented to ensure the acoustic similarity between fricatives in the training phase and those in the test phase. In Eisner and McQueen (2005), generalization only applies when the target phonemes involved in the training and test phases all come from a single speaker and are later spliced with different voices; when the fricative instances come from different speakers, the generalization does not occur. In Reinisch and Holt (2014), perceptual generalization does not occur until the acoustic range of the test continuum is tailored to be less extreme and more comparable with the acoustic distributions of the training stimuli. The second research question of this dissertation regards how the perceptual generalization versus specificity of fricatives is conditioned on **acoustic similarity** (Q.2). This question is investigated in Experiment 2 to obtain a clearer idea of the criterion for “acoustic similarity” between two sibilant instances: Do they need to be produced by the same speaker, or do they need to overlap with each other along some critical dimensions? If their similarity can be captured by the degree of acoustic overlap between these instances along specific acoustic dimensions, then what is minimal the degree of overlap that is required to allow for perceptual generalization? I expect the results of Experiment 2 to shed light on these questions.

The third question under this line of inquiry regards whether the perceptual learning of sibilants generalizes across speakers of the **same gender** (Q.3). Supporting evidence of this hypothesis has been reported by two studies, one based on the perceptual learning of foreign-accented speech (Reinisch and Holt, 2014) and the other based on distributional learning (Tamminga et al., 2020, but with some lexical input in the mix). There is little empirical evidence that lexically induced perceptual learning of fricative boundary is specific to speakers of the same gender. Moreover, in perceptual learning studies with speakers of the

same gender, it is unclear how similar the adopted voices are or whether listeners correctly perceive these voices as coming from different speakers. Experiment 4 of this dissertation evaluates this question, namely whether the presence of top-down cues to talker identity facilitates speaker-specific perceptual learning. The results are expected to tease apart the confound between talker-specificity and gender-specificity in previous studies that use voices of different genders to stand for two different individuals.

1.3.2 Is the perceptual learning of VOT specific to speaker or gender?

The second line of inquiry extends the above questions to a different type of phoneme contrast, namely, stop voicing indexed by the temporal cue of voice-onset-time (VOT). In Table 1.1, previous studies have reported talker-specific perceptual learning for speakers of the same gender (Allen and Miller, 2004; Myers and Theodore, 2017; Theodore and Miller, 2010) and perceptual generalization for speakers of different genders (Kraljic and Samuel, 2006, 2007). This is a somewhat unusual combination, because it is usually assumed that talker variability is larger between genders than within gender, and that perceptual learning generalizes among similar talkers but not dissimilar ones. This can be partially attributed to the different paradigms and contexts of these two lines of experiments. In the line of studies lending support to specificity (Allen and Miller, 2004; Myers and Theodore, 2017; Theodore and Miller, 2010), additional visual cues to speaker identity (including photos of speakers' faces patterned with their names) are presented along with the acoustic stimuli throughout the experiment. This information is expected to enhance listeners' awareness of individual speakers and help them distinguish between speakers who have similar voices.

As with sibilants, the second goal of this dissertation is to investigate whether the generalization of VOT perceptual learning can generalize across individuals of different gender groups, remain specific to different gender groups but generalize across individuals of the same gender group, or remain specific to individuals. These questions are evaluated in parallel to those for sibilants, except that acoustic similarity is not a concern for VOT because there has been no evidence that it matters to VOT adaptation. I expect results to these

questions to give us a better understanding of how perceptual generalization behaviors with multiple speakers vary as a function of phoneme type. In particular, the distribution of VOT production in natural speech seems to exhibit only a weak correlation with speaker gender (e.g., Kleinschmidt, 2019), which differs from spectral cues that exhibit strong correlations with speaker gender (e.g., Jongman et al., 2000). A better idea of how perceptual learning behaviors vary as a function of talker identity and phoneme type helps us understand the nature of the talker structure that constraints perceptual generalization, or in other words, what are listeners' arguments behind their perceptual generalization/specificity decisions. If this talker structure is more a reflection of listeners' perception and mental construction of speakers' social identity, then we should expect the perceptual learning of sibilants and stops to behave similarly with multiple speakers because the involved speakers maintain a stable set of social attributes (e.g., gender) across experiments. In contrast, if the talker structure involved in perceptual generalization reflects more of listeners' knowledge about sociophonetic variability in the acoustic realizations of phonemes, then we should expect different perceptual learning behaviours for sibilants and stops due to their different degrees of covariation with gender.

1.3.3 Is the perceptual learning of sibilant more susceptible to speaker and gender specificity than that of VOT?

The third line of inquiry extends questions along the previous two lines and asks whether sibilants' perceptual learning exhibits a higher level of specificity than stops, as predicted by models of structure in speaker variability. This dissertation approaches this broad inquiry from the following two perspectives of views:

From a qualitative view, I have divided the specificity of perceptual learning into three levels according to the granularity of the social structure that constrains their generalization. From finer to coarser granularity, the perceptual learning of sibilants and stops may be either "specific to talkers", "specific to talker gender", or "can be generalized across talkers and talker genders". As mentioned above, a between-gender multi-talker perceptual learning

experiment with various bias conditions of the intermediate learning phase will be conducted to examine the generalization of perceptual learning across genders for sibilants (Experiment 1) and stops (Experiment 3). A within-gender multi-talker perceptual learning experiment with conditions of auditory versus audiovisual cues to talker identity will be conducted to examine the generalization of perceptual learning across speakers (Experiment 4). Then, building on answers to these questions, we can finally ask whether the perceptual learning of stops generalizes more broadly with coarser-grained constraints of speaker identity, whereas the perceptual generalization of sibilants across speakers is more subject to finer-grained constraints of social identity (Q.6). If we found that perceptual learning is talker-specific for sibilants but generalizes across talkers for stops, or that it is gender-specific for sibilants but can be generalized across talker genders for stops, then that is the qualitative evidence for higher specificity of the perceptual learning of sibilants than stops.

If the perceptual generalization of sibilants and that of stop VOT exhibit sensitivity to the same level of speaker social category, for example, they both show generalization across speakers of different genders, the specificity level of these two mechanisms of perceptual learning can still differ in the dimension of quantity. This specificity is reflected by the extent to which the perceptual learning result reflects the distributional properties of the specific speaker being tested. We may imagine that, after exposure to the speech of two speakers, A and B, the listeners' categorization of speaker A's speech reflect a mixture of A's and B's distributional properties for both sibilants and stops. However, if the perceptual learning of sibilants reflect 80% influence of A's speech and 20% influence of B's speech, whereas the perceptual learning of stops reflect 50% influence of A's speech and 50% influence of B's speech, then the former still involves a higher level of talker specificity because the result is more dependent on the acoustic properties of the specific test speaker. This is the final specific research question that I plan to evaluate, namely, whether the perceptual learning result of sibilants show higher consistency with the distributional properties of the test speaker in training than that of stop VOT (Q.7).

Chapter 2

General Method

This chapter reviews some of the major methodological issues in perceptual learning that have been shown to affect the experiment outcome and provides a general description of the basic experimental paradigm adopted in this dissertation. Section 2.1.2 provides a brief review of the methodological differences between existing perceptual learning studies and how they may alter the observed outcomes, including tasks and instructions, stimulus properties, test paradigm, and result quantification. Building on these, I further discuss the pros and cons of adopting different designs and summarize their implications for the methodological decisions made in the current dissertation. Section 2.3 provides a general introduction to the principal methodological factors involved in this dissertation, including stimulus recording and manipulation, type of block, subject, and experimental task and procedure. Experiments presented in Chapter 3-5 are derived from the paradigm described in this chapter and are designed following a similar logic.

2.1 Review on the methodology of perceptual learning

This section reviews some of the principal methodological elements in the experimental paradigms of perceptual learning. Section 2.1.1 introduces the standard practice of a perceptual learning experiment in great detail, with a focus on two important learning mechanisms involved in perceptual learning, i.e., context-guided learning and distributional learning. Then, Section 2.1.2 reviews some of the more detailed methodological factors that may cause gradient differences in perceptual learning results. The findings reviewed

in this section are expected to inform some of the methodological decisions made in this dissertation, which I will discuss in a later section (Section 2.3).

2.1.1 Context-guided vs. distribution-based learning

Perceptual learning experiments typically involve using context to induce a degree of perceptual bias for an originally ambiguous sound. The experiments in this dissertation are largely modeled on the tradition of Norris et al. (2003), one of the most used techniques for eliciting “lexically-guided” perceptual learning. Given the importance and influence of this paradigm, it is necessary for us to go into details of Norris et al. (2003) before we extend it to a wider range of topics and scopes.

Norris et al. (2003)’s paradigm uses a training phase in which lexical cues suggest the categorical interpretation of an ambiguous sound to the learner. It takes advantage of the well-established phenomenon where listeners prefer to hear words whenever possible (Ganong, 1980). For example, if a listener hears [ɔ̃æ?], where [?] represents a fricative halfway between [f] and [s], the listener will be inclined to interpret the fricative as /f/ to form the word “giraffe,” because hearing the fricative as /s/ would produce only the nonword “girasse.” Hearing the same ambiguous fricative in [hæ?], on the other hand, might lead the listener to an /s/ interpretation. In Norris et al. (2003), participants were randomly assigned to either an /f/-biased or /s/-biased condition (with Dutch stimuli) and were trained on a lexical decision task that consistently signaled the phonemic identity of the ambiguous fricative according to whichever condition they were in. After training, participants were tested on categorization of the ambiguous fricative in the syllables [ɛf] and [ɛs]. Listeners who had been trained on /s/-biased stimuli were more likely to categorize [?] as /s/ in these syllables than those who had been trained on /f/-biased stimuli, suggesting that the training had shifted either or both groups’ perceptual boundaries between /f/ and /s/. An additional group of listeners was exposed to the ambiguous fricatives in nonwords, and they showed no perceptual learning effect in the post-test. In this paradigm, the perceptual learning effect is *quantified* by comparing /s-f/ perceptual boundaries exhibited

by listeners on different conditions in the post-test categorization task.

The lexically-guided perceptual learning paradigm has been exploited considerably (e.g., Kraljic and Samuel, 2005, 2006, 2007; Maye et al., 2008; Reinisch and Holt, 2014). In addition to lexical context, there are other contextual cues that have been shown to induce such a shift of phoneme boundary. Bertelson et al. (2003) and subsequent studies (e.g., van Linden and Vroomen, 2007; Vroomen and Baart, 2009a,b; Vroomen et al., 2007, 2004) have successfully induced perceptual learning with visual cues of place of articulation. Cutler et al. (2008) find that even phonotactic regularities can provide sufficient context to trigger perceptual learning. In these cases, perceptual learning either directly implicates or is closely related to a known bias in speech perception, be it a bias towards wordhood (Fox, 1984; Ganong, 1980), audio-visual integrity (McGurk and MacDonald, 1976), or phonological well-formedness (Hallé and Best, 2007; Massaro and Cohen, 1983), and listeners need this information to infer the intended phoneme. The contextual cues mentioned in the above cases provide crucial information that labels the intended category for the listener. In this sense, Kleinschmidt et al. (2015) classified the cases of perceptual learning under the guide of contextual information as problems of *supervised learning*, because such contexts use lexical context to provide labels for ambiguous phoneme instances and guide listeners' categorization decisions in training. In the meantime, there are situations where contextual labels are not as informative or approachable. These cases include when critical instances are embedded in minimal pairs where both directions of perception yield a real word, or in language acquisition where top-down linguistic labels have not been included in the knowledge of infant learners. This is when the adaptation mechanism needs to cope with *unsupervised learning* problems, that is, to update one's perceptual expectations based on the acoustic distribution of speech input alone without associative labels.

Therefore, in addition to seeking for contextual feedback, listeners also need to keep track of the bottom-up distributional information in continuous acoustic streams to optimize the mapping between acoustic statistics to speech sound categories. For example, VOT typically shows two clusters centered around 0 and 50 ms with a boundary at 25

ms (Lisker and Abramson, 1964). However, if a talker’s VOT clusters around 15 and 65 ms, listeners might reasonably learn a new boundary at around 35 ms (e.g., Kleinschmidt et al., 2015; Maye and Gerken, 2001). This process of learning from simply being exposed to frequency distributions of stimuli is often referred to as “statistical learning” or “distributional learning”. (e.g., Lacerda et al., 1995). These terms are sometimes used in a broader sense; regarding perceptual learning in particular, it means that listeners are sensitive to the distributional/statistical properties of the speech acoustics and use this information to categorize them and update their norms.

Distributional learning is often considered to be an important mechanism in infant speech category development (e.g., Maye et al., 2002), since infants lack lexical or phonotactic knowledge as sources for supervised learning. In adulthood perceptual adaptation, distributional learning has been attested experimentally with continua of ambiguous segments embedded in nonwords (Maye and Gerken, 2001) and, more recently, in words of minimal pairs (Kleinschmidt et al., 2015; Munson, 2011; Theodore and Monto, 2019). This mechanism is also described mathematically (often with Bayesian models) in computational frameworks of acquisition of phonetic categories (e.g., McMurray et al., 2009), spoken word recognition (Norris and McQueen, 2008) and perceptual learning (Kleinschmidt and Jaeger, 2015; Toscano and McMurray, 2010). In general, these frameworks model listeners’ categorization decisions with probability density functions derived from the distributional properties (e.g, mean and variance) of training acoustic values.

When sources for distributional learning and for contextual-guided learning coexist, the outcome perceptual boundary may reflect a mixture of the two effects. In light of this, in a perceptual learning experiment, special attention should be paid to avoid unintended distributional learning effect induced by irrelevant acoustic properties, which further interferes with the actual word-guided learning result. This point will be discussed in more detail in the next section.

2.1.2 The influence of stimulus, paradigm, and task

This section reviews more detailed information of methodological factors in perceptual learning experiments, especially those interfering with the magnitude of the learning outcome. The factors include different designs and tasks used to induce and quantify perceptual learning, the acoustic properties of the training and test stimuli, and the attentional factors associated with the nature of different stimuli and tasks.

2.1.2.0.1 The (a)typicality of critical instances in the training stimuli In the training stimuli of perceptual learning, one of the two critical segments usually deviates from the “canonical” form to some extent, such that listeners expand their boundary of the intended phoneme to include the atypical instances. Meanwhile, it is also understood that standard and nonstandard allophonic instances are assigned different weights in their encoding: Nonstandard instances usually have less leverage on the shift of perceptual boundary than standard ones (e.g., Sumner et al., 2014). As put by Kleinschmidt and Jaeger (2015, p. 13), “(when) listeners encounter odd-sounding, often synthesized speech in a laboratory setting, they may have little confidence ... that any of their previous experiences are directly informative”. Therefore, the degree of atypicality needs to be limited to a certain scope for perceptual learning to work.

The importance of the balance between deviation and typicality is also demonstrated by empirical results. On the one hand, studies show that perceptual learning is less likely to happen if listeners have prior exposure to *standard* instances of the to-be-expanded category (Kraljic and Samuel, 2011). This is sometimes referred to as the “inoculation effect”. In other words, once a speaker has mentally anchored a phoneme with standard instances of that phoneme from a particular speaker, they would be less willing to shift their perceptual boundary with further exposure to nonstandard instances of the same category. On the other hand, however, empirical studies have also reported that perceptual learning is undermined by anomalous acoustic instances, including those with extreme acoustic values (Kleinschmidt et al., 2015), those with a heavy nonnative accent (Witteman et al., 2013),

and those with low intelligibility because of signal degradation (Sohoglu and Davis, 2016). By contrast, a considerable number of lexically-guided perceptual learning studies have induced robust perceptual learning with training stimuli containing “maximally ambiguous” instances (Kraljic et al., 2008a,b; Norris et al., 2003; Reinisch and Holt, 2014). An instance is identified as “maximally ambiguous” if it is half the time categorized as phoneme A and half the time as phoneme B without lexical context.

In a nutshell, the above findings suggest that the most efficient training materials for perceptual learning should fall in an optimal range between ambiguity and typicality. The question becomes how “maximally ambiguous” training stimuli commonly used in perceptual learning studies fit into this range. Babel et al. (2019) examined the magnitude of perceptual learning effect induced by “typical”, “ambiguous”, “atypical” and “remapped” training materials. They found a maximal learning effect induced by typical but ambiguous pronunciations, an attenuated effect by atypical pronunciations, and no effect by typical and unambiguous pronunciations, even for the remapped case. These findings provide justifications for using “maximally ambiguous” stimuli as training materials.

2.1.2.0.2 The (in)variance of critical instances in the training stimuli In addition to the (a)typicality of critical instances, another dimension of the acoustic distribution of training materials that might affect the outcome is (in)variance. It has been well attested that listeners’ perception boundary of a phoneme will shrink after repetitive exposure to invariant instances of that phoneme. For example, in phoneme identification with a /ba-pa/ continuum, listeners with prior exposure to repeated good /ba/ are more likely to categorize ambiguous sounds as /pa/ but not /ba/. This phenomenon is called selective adaptation (Eimas and Corbit, 1973). Different from perceptual learning, the stimuli of selective adaptation are usually repetitive presentations of an identical sound. As a result of the repetition, listeners are less willing to include less prototypical acoustic instances in the listener’s category.

It is unclear whether selective adaptation shares the same mechanism with perceptual learning or operates separately (see e.g., Kleinschmidt and Jaeger, 2015, for an integrated

account for selective adaptation and perceptual learning). However, Kraljic and Samuel (2005) did raise the caveat that the interpretations of perceptual learning results are sometimes confounded by coexisting effects of selective adaptation. Imagine a situation where listeners who have heard a standard /ba/ and an ambiguous /pa/ repetitively end up perceiving more /pa/ for ambiguous sounds. It is unclear whether this is due to the exclusion of ambiguous tokens from the constricted /ba/ or the inclusion of those tokens into the expanded /pa/. The easiest way to avoid this confound is simply to ensure an amount of acoustic variability of the training stimuli.

2.1.2.0.3 The (a)symmetry of the continuum in the categorization test While training stimuli for perceptual learning are expected to induce perceptual shifts efficiently, test stimuli should be used to record the boundary location at the moment faithfully without introducing additional learning effects. In this sense, researchers should look out for any acoustic properties of the test stimuli that might induce unintended distributional learning, as described in Section 2.1.1. One of the acoustic properties that makes the categorization results susceptible to distributional learning is the asymmetry of the test continua. For example, Tamminga et al. (2020) adopted a pre-test/post-test phoneme categorization paradigm to quantify the shift of perceptual learning before and after exposure to atypical phonemes. Surprisingly, they found that listeners started to shift their perceptual boundary between /s-f/ over the ten repetitions of pre-test phoneme categorization even before they heard any lexically-guided training stimuli. They attributed this finding to the fact that the /s-f/ continuum for the test is not symmetric but leans towards /s/ acoustically and perceptually. Exposure to such a distribution induces listeners to take tokens at the endpoints of the continuum as exemplars and shift their boundary to anchor with the midpoint of the continuum. The finding implies that, for the test stimuli of perceptual learning, symmetric continua is preferred over asymmetric ones, because the perceptual learning results collected with asymmetric continua might induce additional statistical learning during the test phase, which is incorrectly taken to be top down.

2.1.2.0.4 Experimental tasks in lexically-guided learning The most common tasks used for lexically-guided perceptual learning include lexical decision and word identification. The two tasks share similar ways to integrate listeners' bias for wordhood and use it for perceptual learning. To induce a boundary shift between a pair of phonemes, lexical stimuli are designed to contain standard instances of one phoneme and nonstandard instances of the other phoneme. For example, to expand category X and shrink category Y, the stimuli are designed to contain nonstandard sounds of X and standard realizations of Y in their own lexical contexts. If listeners recognize those stimuli as words, they unconsciously accept the atypical sounds as realizations of the word-congruent phoneme.

The two tasks differs in how they lead listeners to recognize those stimuli as words. In lexical decision, listeners need to respond to both word and nonword stimuli and decide whether what they heard is an existing word or a nonword. With low transitional probabilities between phonemes or syllables in the language, the nonword stimuli are intended to be obvious to listeners. In word identification, listeners hear word stimuli only (with nonstandard X and standard Y embedded) and need to identify from two written options the one they hear. Both of the two options contain the intended phoneme but only one of them has the right word context (e.g., for rehear?al the answers would be "rehearsal" and "reversal"). In this example, listeners are "forced" to accept the ambiguous sound as /s/ because there are no other options available to them.

Although the two tasks share a similar rationale of lexically guided perceptual learning, they also have their own pros and cons as decided by the nature of the tasks. In lexical decision, there is no guarantee word stimuli with ambiguous sounds would be classified as "word" by participants. Although studies have shown that listeners' endorsement of lexical status has substantial impacts on the efficiency of perceptual learning (Scharenborg and Janse, 2013), the examination of lexical decision responses is sometimes overlooked in perceptual learning studies with this kind of paradigm. This is not a problem for similar studies adopting a word identification paradigm, because the information of lexicality is presupposed by the task. Therefore, adopting a word identification task has the advantage

of needing a shorter experiment and a simpler analysis due to the unnecessary of an equal number of nonword stimuli to word stimuli or assessment on listeners' lexicality endorsement. That being said, it is still important for researchers to have a clear idea of how convinced participants are of the lexical status of the stimuli, in order to prevent inefficient learning due to the inadequate activation of word-level processing.

2.1.2.0.5 Methodological factors related to attention orientation Experimental elements could also exert an influence on the outcome of perceptual learning through directing listeners' attention either to the informational contents or to the acoustic properties of the speech. Presumably, an attention-directing factor in such experiments is the nature of the tasks for training and test. Listeners are more likely to turn on a "how mode" (Lindblom et al., 1995) whereby they pay more attention to pronunciation over contents if the task focuses on speech signals at a sublexical level. These tasks include passive exposure to training stimuli (Tillery, 2015), phoneme categorization with nonword stimuli (Norris et al., 2003; Reinisch et al., 2014), and sound discrimination tasks (Clarke-Davidson et al., 2008). With comprehension-oriented tasks that involve lexical activation and semantic processing, however, listeners tend to adopt a "what-mode" (Lindblom et al., 1995) by attending more to what is said over pronunciation. Tasks of this kind include lexical decision (e.g., Kraljic and Samuel, 2005), identification of written words (McAuliffe, 2015) or pictures (Kleinschmidt et al., 2015), and comprehension of sentences (Davis et al., 2005; McAuliffe, 2015) and stories (Eisner and McQueen, 2006).

While the primary task plays an important role in attention orientation, other factors including certain properties of the stimuli and the wording of instructions can also manipulate listeners' attentional mode. Some of those factors are listed in the following.

- Stimulus monotony: Larger stimulus variation is shown to lead to a more comprehension-oriented attentional set (e.g., Cutler et al., 1987);
- Explicit instructions: Explicit instructions of the ambiguous phoneme promote a more perception-oriented attentional set (e.g., McAuliffe, 2015; Pitt and Szostak, 2012);

- Cognitive load: Listeners show the same and even increased lexical bias effects under a higher cognitive load associated with more complicated tasks (Mattys and Wiget, 2011) or attentional distractors (Zhang and Samuel, 2014);
- The location of the critical phoneme in a word: Listeners are less likely to have full activation of a lexical item at the onset of a word compared to later positions; (Jesse and McQueen, 2011);
- The predictability of the critical word in a sentence: Words of lower predictability induce more high-level processing (McAuliffe, 2015).

Although these factors are tested with various paradigms of perceptual learning to address different research questions, their impacts can be synthesized under the generalization that perceptual learning is enhanced by a comprehension-oriented mode and attenuated by a perception-oriented mode (McAuliffe, 2015). For example, comparison between perceptual learning results with and without explicit instructions about the ambiguity of the critical sounds show that explicit instructions would attenuate or totally inhibit the effect (McAuliffe, 2015; Pitt and Szostak, 2012). Regarding the position of the critical sound in a word, existing findings show that the perceptual learning effect is barely induced with stimuli where the ambiguous sounds are located in word-initial positions (Jesse and McQueen, 2011; McAuliffe, 2015; McAuliffe and Babel, 2016), and the effect is weaker for stimuli with ambiguous sounds in word-initial positions than those with ambiguous sounds later positions (McAuliffe, 2015; McAuliffe and Babel, 2016; Pitt and Samuel, 2006; Samuel, 1981). Studies involving different predictability level of the critical word based on sentential context show larger learning effect with unpredictable stimuli than predictable ones (McAuliffe, 2015). Put together, the above review all point to the importance of a comprehension-oriented mode in the elicitation of efficient perceptual learning.

2.1.2.0.6 Between-subject versus within-subject design The effect of perceptual learning can be measured either through a between-subject paradigm or a within-subject one. The standard practice in the literature is to compare the categorization results of

listeners who are in different prior training conditions. In Norris et al. (2003), for example, participants are randomly assigned to experimental conditions that get opposite treatments in the training stage; then they participate in a phoneme categorization task with the same sound continuum. The difference of the categorization results between groups is taken as an indicator of the perceptual learning effect. The downside of this paradigm, however, is one shared by between-subject designs in general. That is, the difference between the final categorization results can be attributed not only to the difference in training but also to a potential difference in the categorization baseline between different groups of participants.

One way to avoid this problem is to adopt a within-subject paradigm in perceptual learning studies, by adopting a pretest-exposure-posttest design. That is, listeners are first pretested on their categorization with stimuli along an acoustic continuum; after exposure to ambiguous segments in biased contexts, they are tested again on the same continuum. The difference between the pretest and posttest categorization results indexes the learning effect induced by the exposure phase. This is the standard setup of visually-guided recalibration experiments following Bertelson et al. (2003), and it has also been successfully used to evaluate lexically-guided perceptual learning (Eisner and McQueen, 2006). However, this approach suffers from another set of confound, that is the involvement of distributional learning during pre-test and post-test categorization. As discussed earlier in this section, when the continuum is asymmetric, the learning effect based on the acoustic distribution of the continuum in the pretest and posttest interferes with the lexically-guided learning effect during the exposure stage. Such an approach is also affected by the innocent constraint, which states that listeners are less likely to shift their boundary after learning if they have been exposed to the well-articulated tokens of the phoneme in question from the same speaker before. In other words, the involvement of standard tokens at the endpoints of the pretest continuum might also undermine the introduction of perceptual learning effects.

2.2 Methodological decisions in this dissertation

The findings reviewed above provide basis for some of the methodological decisions made in this dissertation. This dissertation uses the well established lexically-guided paradigm of perceptual learning as a testing ground to investigate the role of speaker identity and gender. Regarding the experimental task, this dissertation adopts a word identification task throughout the entire experiment, including both the training phase and the test phase. Again, the use of this task benefits from several advantages: Compared to a lexical decision task, word identification needs fewer stimuli for the omission of nonword ones. In addition, since the lexical status of stimuli is implied by the nature of the task, there is less concern of the lexical endorsement of stimuli by participants, which may affect the efficiency of perceptual learning.

Like the training phase, word identification is also used for the test phase. Steps on an acoustic continuum are spliced into minimal pairs, and listeners' perception boundary can be examined from their choice between two words contrasted by one phoneme. From an ecological validity perspective, hearing sounds embedded in words is more common in real life than hearing sounds in nonword syllables. Also, this makes the training and test phases more consistent in paradigm and so their division less obvious to participants. Listeners would not need to change their attentional set in order to cope with the transition from a comprehension-oriented task to a perception-oriented one. On one hand, a comprehension-oriented task plays a facilitating role in perceptual learning in the training phase. On the other hand, with less attention to the acoustic distributions of the stimuli, listeners are less likely to shift their perception boundary through distributional learning in the test phase.

The above review also shed light on the appropriate range of acoustic properties of stimuli. Following the Goldilocks zone of perceptual learning (McAuliffe and Babel, 2016), the training stimuli are made of maximally ambiguous segments between the two target phonemes embedded in lexical contexts. Previous findings also provide caveats on the importance of introducing an amount of variability in the acoustic distribution of the tokens,

in order to rule out the confound of selective adaptation. To introduce more variability, different critical segments will be chosen for each of the lexical items instead of splicing an identical sound into different word frames.

Regarding the test stimuli, pilot studies of phoneme categorization in the form of lexical decision are conducted in order to pick the right steps that are symmetric around the 50% categorization point. Endpoints of the continuum are not included in the test stimuli in order to prevent an “inoculation effect” (Kraljic and Samuel, 2011), which says that exposure to standard instances blocks perceptual learning. It is also ideal to reduce the number of steps and repetitions of the ambiguous stimuli, and to increase the variety of lexical guises of minimal pairs in the test, in order to avoid the elicitation of a perception-oriented attentional mode.

Lastly, the dissertation adopts a between-subject design instead of a within-subject one in order to prevent extra learning from the pretest categorization phase, or any blocking effect by standard instances in the pretest. In the meantime, training conditions towards opposite directions are adopted in addition to a baseline condition without training. This is done to ensure that the difference of perception boundary between conditions are consistent with the training conditions, instead of merely a baseline difference.

2.3 General method of the dissertation

This dissertation examines the perceptual learning of two pairs of phonemes (/t d/, /s f/) across different speakers and genders. Experiments on the perception of /s f/ are presented in Chapter 3-4 and those on the perception of /t d/ are presented in Chapter 5. The perceptual learning experiments in the two chapters normally contain one or several training blocks followed by a test block, both implemented in the form of a identification task. The basic experimental procedure is shown in Fig. 2.1.

Each block contains an identical amount of word identification trials. The training blocks may differ from each other in two dimensions, namely, the voice of the speaker (A, B), and the acoustic distribution of the target phoneme (X-favoring, Y-favoring, unbiased). The

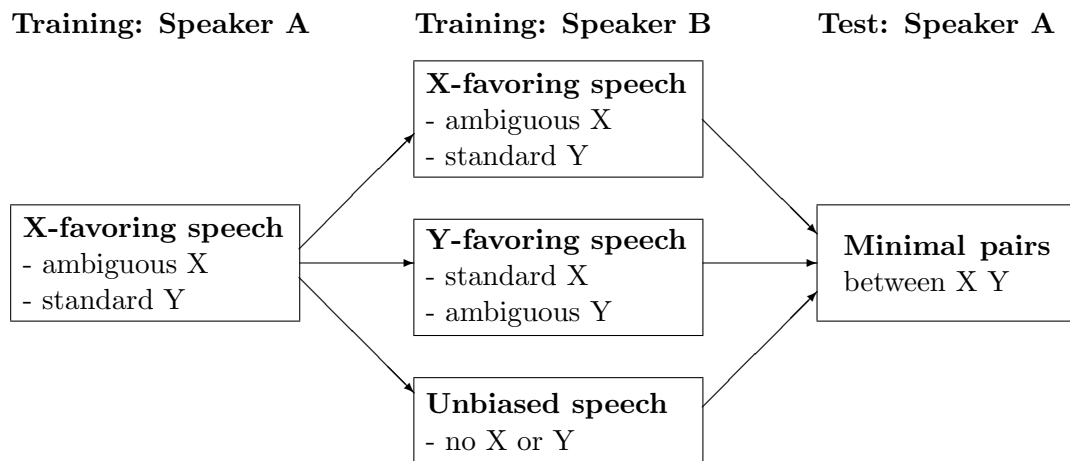


Figure 2.1: A diagram of the basic experimental procedure in this dissertation. X and Y stand for the two target phonemes involved in the experiment

training trials are responses to lexical stimuli with ambiguous instances for one phoneme (X) and standard instances for the other phoneme (Y). After exposure to these stimuli, listeners are expected to expand their category X and shrink category Y in order to account for these instances (i.e., X-favoring). Trials in test blocks were generated by embedding ambiguous segments in frames of minimal pairs contrasted by the two target phonemes, so that categorizing the segment as either of these phonemes would result in an English word. Filler trials without the target phonemes were also included in these types of blocks.

This paradigm allows for independent manipulations of speaker identity and phonetic distribution across experimental conditions. More details about the makeup of each type of block and trial are introduced in the following subsections.

2.3.1 Block and trial

An experimental block may either be a training block, an intervening block, or a test block, depending on the functions and proportional makeup of different types of trials. All these blocks contain the same number of trials (N=51), and the trial order is randomized for each participant.

A training block contains 34 training trials and 17 filler trials. The former is evenly split

between 17 trials containing instances of phoneme X and 17 containing Y, while the latter do not contain any X or Y. The training trials are designed to use lexical context to guide the perception of ambiguous sounds towards the intended direction of the experiment. As explained earlier, the crucial design that induces perceptual learning is that the instances of one of the critical phonemes are standard whereas those of the other critical phoneme are ambiguous acoustically. For example, an /s/-favoring training condition contains ambiguous /s/ sounds and standard /f/ sounds embedded in lexical contexts that favors /s/ and /f/ respectively (e.g., [i'niʃəl] and [ɪ'hæʔəl]). The provided options are the correct word and one foil of a phonetically similar word. Crucially, they are not contrasted on the critical sound. In the above example, the options for /i'niʃəl/ are *initiate* and *initial*, both supporting the perception of /f/, and options for /ɪ'hæʔəl/ are *rehearsal* and *reversal*, both pointing to /s/. See Table 2.1 for more examples of trials on other training conditions.

/s/-favoring	[ɪ'hæʔəl]	[i'niʃəl]
/f/-favoring	[ɪ'hæʃəl]	[i'niʃəl]
correct answer/foil	rehearsal/reversal	initial/initiate
/t/-favoring	[frʌn'ʔɪ]	[ə'ʃendə]
/d/-favoring	[frʌn'tʰɪ]	[ə'ʃenʔə]
correct answer/foil	frontier/frontal	agenda/Amanda

Table 2.1: Examples of training stimuli in different experimental conditions: IPA transcriptions and two options provided for identification

Table 2.2 and 2.3 provides the full lists of words that respectively contain /s-f/ and /t-d/ used for training in the experiments of this dissertation. The two word sets with /s/ and /f/ are balanced for lexical frequency, and so are those with /t/ and /d/. Word frequency is determined using the SUBTLEX corpus (Brysbaert and New, 2009) FREQcount measure. The information of word frequency and the foil option is also presented for each word. Note that I sometimes used low-frequency word such as names of specific places (e.g., Grayshott) as foils in order to maximize their phonetic similarity with the target word. Even if listeners do not recognize these words, it should not prevent them from deriving their pronunciations and choosing the correct word instead. The remaining 17 filler trials in the training block contain neither of the two critical phonemes. For example, for the stimulus /faʊl/ listeners

need to choose between *foul* and *vowel*, and no /s f/ or /t d/ are included in the stimulus. Their purpose is simply to make listeners attend more to the content of the word rather than the nuanced acoustic variability of the target phonemes.

	/f/-containing			/s/-containing		
	<i>target</i>	<i>freq</i>	<i>foil</i>	<i>target</i>	<i>freq</i>	<i>foil</i>
1	compensate	124	condensate	ambition	273	inhibition
2	dinosaur	203	dining-set	beneficial	40	artificial
3	embassy	397	embarrassing	brochure	97	butcher
4	episode	627	webisode	commercial	829	financial
5	eraser	51	harasser	crucial	234	cruel
6	falsetto	15	falsehood	efficient	253	effective
7	faucet	73	flawless	evaluation	225	valuation
8	hallucinate	15	deracinate	glacier	38	Grayscott
9	legacy	256	legally	graduation	500	substitution
10	medicine	1744	medical	impatient	206	impacted
11	obscene	176	obscuring	initial	325	essential
12	parasite	126	parasol	negotiate	342	negation
13	peninsula	70	Pennsylvania	official	1224	optimal
14	personally	1870	personality	parachute	162	paragon
15	pregnancy	334	presidency	publisher	230	publicly
16	reconcile	58	gracile	refreshing	187	infringing
17	rehearsal	635	reversal	vacation	1673	vocation
	Mean: 419.8			Mean: 402.2		

Table 2.2: Word list of the training trials for the perceptual learning of /s-f/

Filler words exist in every experimental block in this dissertation, except that different types of blocks have different proportions of fillers. In an intervening block, the whole 51 experimental trials are filler trials, which is intended to evaluate the persistence or change of perceptual learning effect without further exposure to critical phonemes in the speech. Table 2.4 provides a list of filler trials used in this dissertation.

A test block contains 35 test trials and 16 filler trials. The 35 test trials are generated by splicing 7 repetitions of 5 steps along a acoustic continuum into word frames of minimal pairs. Different perception of the critical sounds yields in two possible words, which are also the two choices on that particular test trial. For example, for the stimulus /?em/ listeners need to choose between *same* and *shame*). This is similar with /t d/, such that the options for /?ai/ may be *tie* and *die*. I did not include minimal pairs with coda contrasts (e.g.,

	/d/-containing			/t/-containing		
	<i>target</i>	<i>freq</i>	<i>foil</i>	<i>target</i>	<i>freq</i>	<i>foil</i>
1	academic	238	academia	cafeteria	289	criteria
2	accordion	67	according	casualty	138	casually
3	agenda	373	Amanda	cemetery	443	cemented
4	armadillo	8	armchair	consultation	64	consultative
5	coincidence	948	coincident	frontier	167	frontal
6	comedian	209	commission	hesitation	94	hesitate
7	confidence	993	confident	infantile	38	infantry
8	crocodile	115	crockery	lunatic	433	lunar
9	handy	623	handling	magnetism	49	mathematics
10	hazardous	94	hassling	military	2103	militant
11	iodine	64	idol	momentary	56	moment
12	kingdom	787	kindle	overtime	318	overtake
13	melody	337	metallic	relative	397	related
14	merchandise	245	merchandiser	royalty	200	loyalty
15	remedial	20	remediate	scientific	580	scientist
16	residence	421	resilience	voluntary	103	voluntarily
17	secondary	209	security	warranty	42	warranted
	Mean: 338.3			Mean: 324.4		

Table 2.3: Word list of the training trials for the perceptual learning of /t-d/

led vs. *let*) or containing more than one syllable (e.g., *writer* vs. *river*). This is because vowels preceding voiceless stops are shorter in duration than vowels preceding the voiced cognates, and this difference is meaningful in perception as well (House and Fairbanks, 1953; Klatt, 1976; Zimmerman and Sapon, 1958). Excluding minimal pairs contrasted on word-final /t/ and /d/ prevents integration of additional duration cues. However, there may still be influences of other secondary accompanying cues such as F_0 (Whalen et al., 1993), formant transition (Lieberman et al., 1958) and the energy distribution of the burst (Chodroff and Wilson, 2014), admittedly. In this chapter, the only acoustic cue of /t d/ to be manipulated is the VOT distribution. Therefore, in the stimuli manipulation process, I have ensured that the secondary cues reviewed above are ambiguous or not informative enough to signal a strong preference for the identification of one segment or the other. The final candidates included in the recording are shown in Table 2.5.

Filler	Foil	Filler	Foil	Filler	Foil
airline	alkaline	foul	vowel	nothing	nutting
amongst	among	framing	Fleming	raccoon	balloon
anvil	angle	gable	gamble	raven	ravel
average	advantage	gargoyle	char-broil	ribbon	region
banana	barbarous	honey	hobby	row	low
beloved	belated	iguana	Guana	runaway	takeaway
buffalo	buffer	January	February	thumbnail	toenail
dragonfly	dragon-fruit	jewelry	jealousy	verify	varified
earning	earring	journal	journey	village	voltage
eyebrow	eye-bolt	lonely	longing	volleyball	valleygirl
feeling	feeding	marina	Maria	vugar	Vodka
firefly	fairfax	enamel	enable	waffle	castile
follow	fallow	Nepal	Napoleon	wharf	wolf

Table 2.4: Word list of fillers and their foil options in perceptual learning

/s-f/	same-shame	sake-shake	seat-sheet	suit-shoot	sock-shock
	sip-ship	sign-shine	sigh-shy	sell-shell	self-shelf
/t-d/	done - ton	dime - time	deer - tear	dip - tip	dose - toes
	down - town	Dutch - touch	die - tie	dim - Tim	dean - teen

Table 2.5: Word list of test trials in the form of minimal pairs

2.3.2 Recording

Materials recorded for further manipulation in this dissertation contained 68 training words in Table 2.2 and 2.3, 39 filler words as shown in Table 2.4, and 40 test words in Table 2.5. All the words were first read by a female native English speaker, who is also a linguist at Penn. She was instructed to read each word with a falling intonation and a smooth speech rate in a consistent manner. The 68 training words were read once normally and a second time with the critical phoneme replaced by the opposite one. In other words, both [ɪˈnɪʃəl] and [ɪˈnɪsəl] were produced for *initial*, and both [ˈeɪvɪdəns] and [ˈeɪvɪtəns] are pronounced for *evidence*. Then, two female speakers and two male speakers of standard American English were recruited from the undergraduate subject pool at UPenn to read after the first speaker’s production, including all the training words with both the original the replaced phoneme, as well as the filler and test words. The four speakers were asked to

imitate the model talker’s pronunciation and intonation. The whole recording process was monitored by the author to make sure of the consistency of the prosodic patterns, including a falling intonation trend and consistent positions of lexical stress. All the stimuli were recorded in a sound-proofed recording booth at the University of Pennsylvania, with a Yeti microphone at a sampling rate of 44.1 kHz.

2.3.3 Manipulation and pilot

The two types of critical phonemes – sibilants and stops – are annotated and manipulated in different ways, which will be described in greater details in the chapter where they become relevant. Broadly, for sibilant stimuli, each pair of s-sh instances produced with the same word frame (e.g., compensate and compens~~h~~ate) are cut out and mixed with each other by five steps of proportions. Then the synthesized continuum is spliced back to the production instance of the correct phoneme (compensate).

For stops, however, the ambiguous sounds halfway between /t d/ are all manipulated from voiceless /t/ sounds by VOT compression. For each production with /t/, five sections of areas are annotated for each critical sound. They are obstruction, burst, aspiration, transition and vowel. The VOT coarsely corresponds to the “aspiration” part and is measured automatically based on the annotation, whereas the burst is not included in the VOT measurement. In addition to VOT time-compression, the manipulation may also include the reduction of the burst amplitude, as well as the splicing of the transition (and sometimes also the following vowel), if pure VOT manipulation does not make the phoneme ambiguous enough.

For the training stimuli, a pilot lexical decision experiment is conducted to select the most ambiguous step of sibilant and VOT for each lexical item. The synthesized acoustic steps on the continuum are each presented in their unambiguous word frames. Listeners need to judge whether these productions are English words or not. Finally, based on the results of this lexical decision task, the most ambiguous acoustic steps (in stimuli chosen as nonword 50% of the time) are selected to be further used to construct the training materials

in perceptual learning.

Like the training stimuli, perception experiments were also conducted to select the steps and word frames of minimal pairs to be used for stimuli in the test phase. The test trials are generated by splicing eleven acoustic steps on a /s-f/ or /t-d/ continuum into different word frames of minimal pairs. Participants need to choose between the two words in the minimal pair which one they hear. Based on their responses, the most ambiguous steps that triggers an equal number of the two words are selected as the center point of a new five-step continuum to be spliced into the test trials.

2.3.4 Subject recruitment

All the experiments are programmed and implemented through the Ibex online experimental platform, along with the PennController system (Zehr and Schwarz, 2018). The subjects are recruited either from the UPenn subject pool or from Prolific, a subject pool for online experiment (Palan and Schitter, 2018). Attention has been paid to ensure that participants on different conditions of the same experiment whose results are to be compared come from the same subject pool, since the demographic difference between these two groups' participants may affect their perceptual learning results.

Internet-based research is considered beneficial for several reasons. It allows for the recruitment of a large sample of participants coming from diverse backgrounds at a relatively low cost. Also, the lack of researcher presence in internet-based research helps prevent researcher bias and ensures procedure replicability (e.g., Birnbaum, 2004). Research on methodology comparison has been conducted to evaluate whether data collected through internet delivery and face-to-face contact produce comparison results both in the field of self-reporting surveys and questionnaires (e.g., Carlbring et al., 2007; Whitaker, 2007), and with experimental approaches (e.g., Birnbaum, 2001; Reips, 2002; Vellido and Matute, 2011). Mostly, these studies reported that internet-based studies tend to obtain their efficiency while producing results similar to traditional laboratory results, lending support to the integrity of results obtained through internet-based approaches (see Honing and Reips,

2008, for a review).

2.4 Planned analyses and result interpretation

The experimental design adopted in this dissertation (as elaborated with Fig. 2.1) provides a specific situation of perceptual learning with multiple talkers: If a listener encountered A who has high-frequency /s/ and /ʃ/ sounds and learned to adapt to it; after a while, she encountered B who has a low-frequency /s/ and /ʃ/ sounds (different speaker and acoustic distribution). Then by testing listeners' categorization behaviors with Speaker A's speech in the final phase, we want to figure out what kind of phonetic distribution listeners will adopt when they hear A's speech again. Assuming perceptual learning happens with the speech of both A and B, we may consider the above question to be a combination of two sub-questions: 1) how much of the perceptual learning of Speaker A's speech would be applied to the perception of Speaker A's speech in the final stage; and 2) how much of the perceptual learning of Speaker B's speech would be applied to the perception of Speaker A's speech in the final stage. For each of two questions, the answer could fall into a continuum ranging from "not at all" to "a hundred percent", and combinations of different parameters set for these two dimensions will result in qualitatively different patterns.

Figure 2.2 is plotted to help visualize what the possibility space looks like. As a first step, if we only consider the extreme situations (either 0% or 100%), four different patterns can be derived: *cumulative update*, *recency update*, *retention* and *reset*. In what follows, I will interpret these four possibilities, and present findings in favor of each of them in previous literature on perceptual learning.

2.4.0.1 Retention

Retention refers to the possibility that the established speaker-specific phonetic belief is retained in the listener's mind, and becomes activated again each time when the same speaker is encountered. In the case discussed above, *retention* would predict that listeners will only use what has been learned from speaker A to cope with A's speech, without

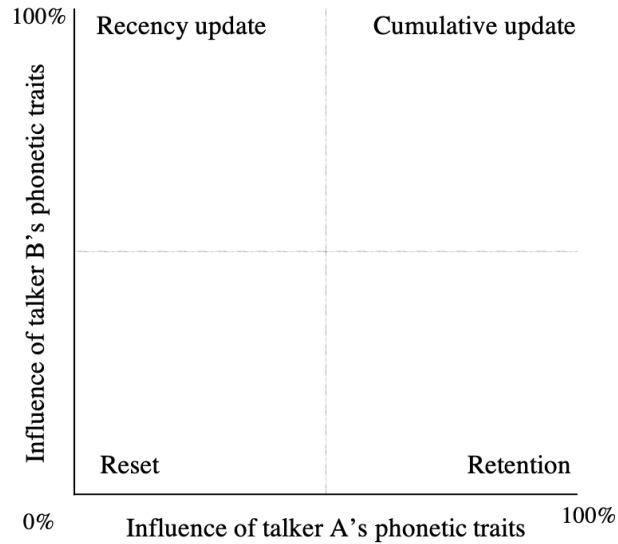


Figure 2.2: Potential outcomes of perceptual learning with speaker A and B successively interference from the intervention of B’s speech acoustics. Such a mechanism would require memory of phonetic distributions to be long-term and highly speaker-specific.

Kraljic and Samuel (2005) provided a thorough investigation on this question, by asking what kind of intervening materials can be used to “undo” the learning of /s-f/ distributions with a previous speaker. The intervening materials they examined include: a) speech from the same model talker or a different talker that contains no critical phonemes; b) speech from the same model talker or a different talker that contains good instances of /s/ and /f/; c) speech from the same model talker or a different talker good instances of /s/ or /f/ that is opposite to the training (e.g., if the training contains ambiguous /f/, then the intervening block contains non-ambiguous /f/); and d) a silent visual game with equivalent duration. The two speakers used in their experiment have different genders.

Their results showed robust durability (and sometimes a boost) of perceptual learning under several conditions. The learning effect was not canceled out by an intervening silent visual game or exposure to a different speaker with or without the critical phonemes. The boundary also remained where it was after encountering speech of the same speaker without instances of or with good instances of /s/ and /f/. The only intervening speech that made a difference is the speech of the same speaker with an opposite acoustic distribution of the

critical phoneme compared to the preceding block. This is expected because there were not multiple speakers; listeners were just updating their cue distributions for the same speaker. The results showing the retention of learning after exposure to a different voice with opposite cue distributions was replicated later in Kraljic and Samuel (2007). Based on these findings, it seems that perceptual learning is quite robust and is resistant to interference with different voices and variable critical instances.

Other studies lending support to the possibility of *retention* of perceptual learning include Allen and Miller (2004) and Theodore and Miller (2010). With a somewhat different experimental paradigm, they showed that perceptual beliefs about VOT distribution are robustly talker-specific. In specific, they trained listeners on the speech of two female talkers, one with short VOTs and the other with long VOTs, and evaluated whether listeners choose a short or long VOT variant to represent a given talker's speech. They found that listeners were able to select the variant consistent with a given talker's VOT characteristics. Moreover, they can generalize the learned feature to novel words with the same phoneme (Allen and Miller, 2004) and novel voiceless stops with the same contrast (Theodore and Miller, 2010). This set of experiments differ from Kraljic and Samuel (2005, 2007) in several aspects: The two speakers they used have the same gender, and stimuli were each paired with the name and photo of their own speakers. In addition, they used unsupervised learning in the training phase by embedding the critical sounds in lexically ambiguous contexts.

The above findings all indicate that listeners can maintain speaker-specific results of perceptual learning in a robust manner.

2.4.0.2 Reset

Reset refers to the process of wiping out previous perceptual outcomes and beginning to establish new phonetic beliefs from scratch. This possibility is diagnosed if we see a previously shifted perception boundary goes back to where it have been (the baseline) after exposure to a different talker's speech.

Findings in favor of *reset* also comes from Kraljic and Samuel (2007). While they

observed *retention* of talker-specific phonetic beliefs with s-f, they did not find the same thing with stops of /t-d/. The experimental paradigm was similar. Participants were first trained with speech of two talkers with different genders (A and B) in two sequential training blocks, with their VOT distribution biased toward different phonemes (either /t/ or /d/). Then participants were tested on the perception boundary of /t-d/ in talker A's voice. Surprisingly, the results showed no learning effect this time. Kraljic and Samuel (2007) proposed two possibilities that somewhat paralleled the possibilities of *cumulative update* and *reset*, namely, either that listeners had integrated the two halves of opposite VOT distribution together to form a new aggregate distribution that was not biased towards either end of the continuum (*cumulative update*), or that they had reset the phonetic belief about Speaker A's speech in order to orient themselves towards the speech of an upcoming speaker (*reset*). Then, they kept the training phases identical and changed the test stimuli to be speaker B's speech, and this change successfully induced an amount of boundary shift. Based on these results, Kraljic and Samuel (2007) argued for the explanation of *reset* against *cumulative update*. According to their interpretation, if distributions in the first block and in the second block canceled each other off by accumulation, then the second-phase training would not have the power to shift the perceptual boundary of either A's or B's speech any more with their local statistics. The result itself is clear, but corresponding interpretations were somewhat puzzling. However, a possible concern is that the relationship between two speakers need not be symmetric in principle. The reasoning they used to rule out *cumulative update* would no longer valid if, for example, the perceptual learning of male speech are more general and that of female speakers are more specific. In this case, *cumulative update* still may partially account for the observations in Kraljic and Samuel (2007). In addition, it is also unclear why the results were at odds with Allen and Miller (2004) and Theodore and Miller (2010), both of which observed the retention of talker-specific phonetic beliefs of talker VOT.

2.4.0.3 Recency update

Recency update refers to a process, according to which, phonetic beliefs are by and large shaped by the most recent acoustic instances of recently encountered speakers, while learning based on “old” acoustic instances fade out rapidly. This process is diagnosed if the perceptual learning result consistently mirrors the distribution of the recent acoustic instances.

Unlike studies showing that phonetic beliefs were long-lasting and specific to speakers in multispeaker speech perception (e.g., Allen and Miller, 2004; Kraljic and Samuel, 2005), very few studies argue that perceptual learning is simply tracking the most recent statistics from any talker. However, the phenomenon that listeners disregard prior experience with a talker and update to the most recent input is not uncommon in studies of perceptual learning with a single speaker. Saltzman and Myers (2018) had listeners exposed to four interleaved blocks of lexical decision that were designed to skew the perception boundary between /s/ and /ʃ/ towards opposite directions. They examined the perception shift after each lexical decision block with a phonetic categorization task. Their result showed that, in each session, listeners’ perceptual bias was consistent with the cue distributions in the immediately preceding lexical decision block. Saltzman and Myers (2018) therefore argue that, in perceptual learning, listeners rely more heavily upon the most recent information and down-weight older, consolidated information.

The influence of recent acoustic instances was commonly observed in empirical perceptual learning experiments, but it is challenged for making predictions about the short-term nature of perceptual learning. For example, Kleinschmidt and Jaeger (2015) posited that talker-specific distributions cannot be created or maintained if a listener simply tracks the recent statistics from a talker. Debates are still under way regarding the role of newly encountered talker-specific instances (“recent” or “local” statistics), as opposed to the cumulative distribution of a talker’s instances (“cumulative” or “global” statistics), which I am about to review in the next section.

2.4.0.4 Cumulative update

Cumulative update refers to the process of building up phonetic beliefs based on the cumulative acoustic distribution of a collection of “old” and “new” instances of a phoneme. As illustrated in Figure 2.2, this process is diagnosed if the outcome of perceptual learning reflects aggregation of partly speaker A’s distribution and partly speaker B’s distribution.

Empirical findings in favor of accumulation are numerous, both with speech of a single speaker, or with multiple speakers. Kraljic and Samuel (2005) showed that speech inputs from the same speaker with different acoustic distributions of /s/ and /ʃ/ canceled each other out in perceptual boundary shift. Van Linden and Vroomen (2007) induced perceptual learning along a /t-b/ continuum using both lexical biases and lip-reading cues, and examined the time-course of perceptual learning with sporadic trials distributed in different positions of a block. They found that, as more recent instances were integrated towards the end of the block, the learning effect became smaller and finally gone.

Theodore and Monto (2019) demonstrated a different manifestation of the build-up of the global statistical effect, by manipulating the range instead of the mean of VOT distribution. They adopted an unsupervised paradigm to induce perceptual learning with the /k-g/ contrast, by having half of the participants exposed to a narrow VOT distribution in one block followed by a wide distribution block, with the order reversed for the other half of the listeners. The result showed a steeper identification slope from the narrow-wide group compared to the wide-narrow group for earlier trials, with the difference attenuated towards the end of the experiment. This result was interpreted to show that listeners did not disregard prior experience within a talker, but rather used cumulative statistics to guide phonetic decisions.

Chapter 3

Exp 1: Perceptual Learning of /s-f/ across Speaker Genders

This chapter reports on Experiment 1, which evaluates how the perceptual learning of fricatives operates across speakers of different genders. Experiment 1 contains a series of sub-experiments that investigate how listeners integrate the acoustic distributions of /s/ and /f/ of speakers of different genders to adjust their perceptual expectations in multi-speaker listening. This chapter is organized into four sections. Section 3.1 summarises findings in the previous literature on the acoustic analysis and perceptual correlates of /s-f/ and outlines the research question and predictions of this experiment. Section 3.2 provides an overview of the methodology, including general design, experimental conditions, stimulus manipulation, and stimulus acoustics. Section 3.3 reports on results of a pilot study and a series of three sub-experiments. Section 3.4 discusses the implications of the main findings and concludes this chapter.

3.1 Background and research question

The background literature reviewed in this section are twofold: Section 3.1.1 summarizes findings on the production and acoustic properties of /s-f/ and how they vary with speaker gender in speech production; Section 3.1.2 summarizes findings on the perceptual correlates of /s-f/ and the influence of speaker gender on the categorization and perceptual learning of /s-f/. Building on the review, Section 3.1.3 articulates the research questions of this chapter and lay out the predicted patterns of the experiment results for each hypothesis as

a potential answer to the research question.

3.1.1 The acoustic properties of /s-f/ and gender variation in production

The acoustic properties of the sibilants /s/ and /ʃ/ have been well examined, both regarding how their acoustic realizations are different from other non-sibilant fricatives, and in terms of how they differ from each other. Jongman et al. (2000) examined the acoustic differences between sibilants with four places of articulation: labial (/f, v/), dental (/θ, ð/), alveolar (/s, z/) and post-alveolar (/ʃ, ʒ/) with a set of extensive spectrum measurements, and found /s, ʃ/ to be different from other fricatives in a number of ways.

The first measure that captures their differences is the *root-mean-square (RMS) amplitude*. Consistent with previous research (Behrens and Blumstein, 1988; Strevens, 1960), Jongman et al. (2000) found the highest amount of RMS amplitude normalized by vowel amplitude with /s z ʒ ʒ/: their RMS amplitude is higher by around 10-15 dB than /f v θ ð/. The second parameter is the F2 transition. /s z ʒ ʒ/ have lower F2 transition than /f v θ ð/, and /ʃ/ has a still lower F2 transition than /s/. This is because the place of articulation of /s-ʃ/ is backer than the labial fricatives examined. Thirdly, /s ʃ/ has the longest noise duration after normalization among all the examined fricatives, and the noise duration of /s/ is still longer than that of /ʃ/. Lastly, sibilants are also found to have the lowest frequencies of spectral peaks and the smallest spectral variances. This property is also well-defined in previous literature (Behrens and Blumstein, 1988; Heinz and Stevens, 1961; Hughes and Halle, 1956; Strevens, 1960, etc.). In general, these studies show that the spectral shapes of sibilants are more distinct, while labiodental and interdental fricatives display a relatively flat spectrum.

In the meantime, /s/ and /ʃ/ also differ dramatically from each other in several acoustic dimensions. Among the examined fricatives in Jongman et al. (2000), /s/ has the highest spectral mean and kurtosis, and the lowest skewness and relative amplitude by discrete Fourier transform (DFT), whereas /ʃ/ has the lowest spectral mean and kurtosis, and the highest skewness and relative amplitude by DFT. These findings indicate /ʃ/ has the

strongest energy concentrated in the low frequencies whereas /s/ has it in the high frequencies, and that /f/ has a flatter spectrum than /s/ (McFarland et al., 1996; Nittrouer, 2002; Tomiak, 1990). In vocalic contexts, /s/ shows a very different peak location after lip-rounding vowels /o u/, whereas the peak of /f/ does not deviate from its original location because of lip rounding.

It is well documented that the acoustic realizations of sibilants vary systematically between gender. Jongman et al. (2000) presented a comparison of the acoustics of sibilants between genders, which is adopted in Figure 3.1.

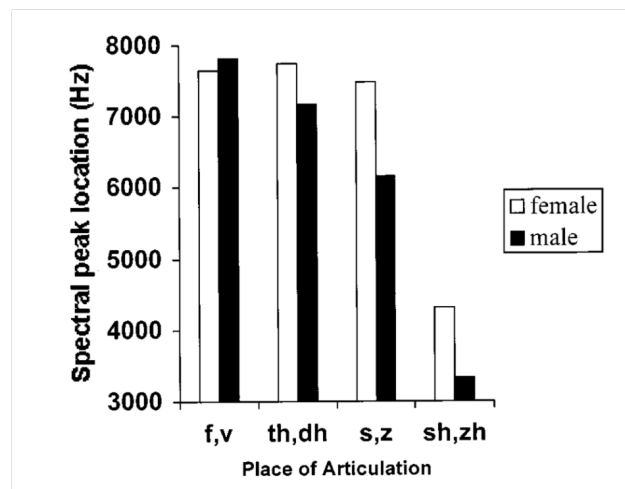


Figure 3.1: Spectral peak location of fricatives by place of articulation and by gender; figure adopted from Jongman et al. (2000)

As shown in Fig. 3.1, fricatives of females generally have higher spectral peak locations than those of males (except for /f, v/). Similarly, the sibilants of females have higher values for spectral mean and kurtosis and lower values for skewness compared to those of males, indicating that the spectra of female sibilants had clearer peaks and a concentration of energy towards higher frequencies. Timing-related measures reveal that sibilants produced by females had a slightly smaller normalized duration than those by male speakers. No gender difference is found in amplitude-related measures.

The covariation between sibilant properties and speaker gender is more than anatom-

ical; it also has socially grounded motivations. In the sociolinguistic literature, sibilant consonants are well-established as resources for the negotiation of gender and sexuality (see Zimman, 2017, for a review). For example, Flipsen Jr et al. (1999) show that children begin displaying gender differences in /s/ at a young age before they start to show any sex differences of the vocal tract. Zimman (2017) finds that English-speaking men modulate their /s/ production in a way that is consistent with their self-identified gender category. In specific, non-binary male speakers produce /s/ sounds with a higher center of gravity (COG) than trans-gender speakers, whose COG is still higher than those of the straight speakers.

3.1.2 Perceptual correlates of /s-ʃ/ and speaker gender effects on perception and learning

A number of studies have examined the potential perceptual correlates of /s-ʃ/, and their findings show that relevant cues are distributed across the frication and vocalic portions. Primarily, the perception of these sounds is affected by the frication spectrum, especially the location of spectral peak (Heinz and Stevens, 1961; Hughes and Halle, 1956, among others). Formant transition at the onset of voicing (especially F2) is also shown to play a substantial role (Heinz and Stevens, 1961; Nittrouer and Studdert-Kennedy, 1987; Stevens and House, 1956; Whalen, 1991, among others). Other supporting cues that have been shown to play a role include the duration of frication (Jongman, 1989) and consonant amplitude relative to vowel (Hedrick and Ohde, 1993).

The gender variation of sibilants in speech production reviewed above has substantial consequences on perception. One of the consequences is that listeners use the socially constructed concept of gender as a source of speaker normalization in the categorization of /s ʃ/. Strand and Johnson (1996) reported a shift of the perception boundary between /s/ and /ʃ/ as a function of the voice gender on the rest of the syllable – an /i/ vowel. In specific, when listeners heard a vowel spoken by a female voice, they would expect sibilants of higher frequency based on their experience that female speakers' sibilants are centered

around a higher frequency than male speakers'. Listeners thus ended up shifting their perception boundary towards a higher frequency for tokens in female-sounding voices, and towards a lower frequency for those in male-sounding voices. This sibilant boundary shift can also be induced by visual gender stereotypes, e.g., by identical auditory tokens in a gender-ambiguous voice patterned with different gender faces.

Due to the robust covariation and mutual-indexical relationship between sibilants and speaker gender, relevant studies have also proposed a potential gender interference in the perceptual learning of /s ʃ/. Proposals along this line first came up in Kraljic and Samuel (2007), where they suggest that listeners learn talker-specific representations for a fricative contrast (/s ʃ/) but do not do the same for a stop contrast (/t d/). However, the relevant experiments in Kraljic and Samuel (2007) does not efficiently tease apart the possibilities of perceptual reset and distributional offset. In other words, when exposure to two speakers' speech with acoustic biases at odds with each other ends up with no perceptual shift, it is unclear whether it is because the opposite acoustic distributions cancel each other out, or because listeners set aside the previous perceptual learning outcomes as they encounter a new speaker. It still remains mysterious whether the mechanistic difference leading to asymmetric behaviors between /s ʃ/ and /t d/ is an *update* one or a *reset* one.

The idea that speaker identity or gender may impose different constraints on the perceptual learning of stops and fricatives is later integrated into an "ideal adapter" framework (Kleinschmidt, 2017). According to this framework, gender may serve as an information-based sociophonetic speaker structure for the perceptual learning of phonemes that are contrasted in frequency distributions. Through computational modeling, Kleinschmidt showed that gender and dialect are more informative about speakers' vowel formant frequencies than VOT length. Specifically, gender is more informative about the absolute vowel formant frequencies, whereas accent is more informative about the normalized vowel space. This finding provides potential motivation for listeners to build separate mental representations of frequency-related phonemes for speakers of different genders. Although Kleinschmidt did not directly compare the informativeness of gender for fricatives and stops, the result

revealed that speaker gender is useful for anticipating frequency distributions but not voice onset time. In this sense, it also aligns with the earlier experimental finding (Kraljic and Samuel, 2007) that perceptual learning of /s f/ does not seem to generalize across genders, because it predicts that the storage of retrieval of relevant speech representations operates for male speakers and female speakers separately.

More broadly, the general idea behind an “ideal adaptor” framework is that an ideal listener would represent information about speakers according to an information-based sociophonetic speaker structure: This structure would have minimal speech difference within sociophonetic speaker groups to ensure the accuracy of representational generalizations among group members, and maximal speech differences between sociophonetic speaker groups. Such a speaker variability structure would require listeners to have sophisticated knowledge of the informativeness of sociophonetic speaker conditions for different types of phonemes, which emerges from listeners’ experience with the real-world sociophonetic structures of speech variability. With such structures of speaker variability accessible to listeners mentally, it is rational for them to manage a coherent representational system for speakers of the same sociophonetic group while building separate representational systems for speakers from different sociophonetic groups. Even though such representations may require additional processing cost to maintain separate phonetics-phonology mapping, they still better satisfy the needs of efficiency and accuracy of the speech processing system than managing every bit of speech episodes separately or encoding them all together with one representational system.

Recently, Tamminga et al. (2020) reports on a set of experimental data that explicitly compares cross-talker generalization of fricative boundary perceptual learning in same-gender and different-gender pairs. They adopted a pretest-training-posttest paradigm and unexpectedly found that listeners shift their perceptual boundary in the categorization phase by anchoring the perceptual boundary with the center of the continuum. Crucially, this shift appears to be reset at the beginning of the post-test if the intervening training phase comes with a speaker of a different gender from the speaker in the pre- and post-

test. In contrast, the shift remains where it is if the intervening training speaker shares the gender of the test speaker. Although the boundary shift observed in Tamminga et al. (2020) is not induced by lexical bias as in previous studies and in this dissertation, it does raise the intriguing possibility that exposure to the speech of different speakers of the same gender does not hinder the retention of the previous perceptual learning outcome. This lends further support to the role of speaker gender as a sociophonetically informative social speaker structure in the generalization of perceptual learning for fricatives.

3.1.3 Research question and hypotheses

Under the broad inquiry of this dissertation about whether the cross-speaker generalization of perceptual learning is conditioned on a specific sociophonetic speaker structure, this chapter asks how listeners integrate the speech properties of speakers of different genders into their perception expectations for an upcoming speaker. Experiment 1 provides a concrete example of this situation: If the listener has had exposure to the speech of a female speaker and a male speaker successively, then whose phonetic distribution will they draw upon to cope with the categorization of the previous female speaker’s speech? The two speakers whose speech is used in this chapter will be referred to as Female A and Male A throughout this dissertation. Broadly, a *specificity* hypothesis would predict that only the knowledge of Female A’s acoustic distribution would become relevant in this case, whereas a *generalization* hypothesis would expect otherwise.

Still, various situations can happen under the umbrella of “otherwise”. Recall that we have discussed what the possibility space looks like in Section 2.4. The nature of speaker-specific knowledge listeners use in the test phase can be captured by a two-dimension possibility space, which describes a) whether or not it reflects the distribution specific to Female A, and b) whether or not it reflects the distribution specific to Male A. This combination gives rise to four possibilities, as briefly summarized in Table 3.1. Note that each of these four possibilities could further be refined in terms of the relative strength of the contributions of the different distributions, but for now, I will be focusing on the four

broad kinds of possibilities for the sake of simplicity.

Hypotheses	Influence of Female A	Influence of Male A
cumulative update	+	+
recency update	-	+
retention	+	-
reset	-	-

Table 3.1: Predictions of the four possible mechanisms of perceptual generalization on the result of Exp 1

A *retention* hypothesis, corresponding to the possibility of *specificity* mentioned above, predicts that the outcome perceptual shift aligns with Female A’s acoustic distribution while remains unaffected by Male A’s acoustic distribution. In other words, when listeners re-encounter Female A for a second time, they return to the expectations they had formed about her speech from their earlier exposure. If Female A’s distribution is not reflected in the result of the final test, then the possibility lies with either *recency update* or *reset* depending on whether or not the test result reflects Male A’s distribution. Both of these situations indicate suggest that the perceptual learning of sibilants is not speaker-specific. Finally, if both of Female A’s and Male A’s distributions have laid an influence on the final test, then it suggests that perceptual learning is not strictly speaker-specific given that it updates in response to acoustic exposures from other speakers as well. This kind of result lends support to a *cumulative update* account.

3.2 Method Overview

3.2.1 Experimental conditions

Experiment 1 contains four parts – a pilot study and three sub-experiments (Exp 1a-1c). All of these experiments end with a categorization test on the same /s-f/ continuum of Female A’s speech, but they differ in the speakers and the acoustic conditions of these speakers that participants have had exposure to before the categorization test. In what follows, I provide a brief overview of the design and purpose of each of these experiments.

The pilot study reports the /s-f/ categorization results of three conditions: a baseline

condition where participants have not received any prior training before the test, and two training conditions where participants have received either an /s/-favoring or an /ʃ/-favoring training phase with Female A’s speech before the categorization test, depending on the specific condition they are in. The goal of the pilot study is twofold. The first goal is to show that the categorization boundary (represented by the 50% probability point) is aligned with the center of the continuum, and the second goal is to show that the perceptual learning design works with the stimuli of Female A, as evidenced by a resulting boost either in /s/-equivalent or /ʃ/-equivalent responses depending on the training condition.

Exp 1a and 1b each contain three experimental conditions. Conditions in Exp 1a consist of a training block with Female A’s /s/-favoring speech, a consecutive training block with Male A’s speech, and a final categorization test with Female A’s speech. The three conditions differ in whether the intermediate training phase with Male A’s speech is /s/-favoring (in the *same* direction with the first training phase), /ʃ/-favoring (in the *opposite* direction to the first training phase), or containing no /s ʃ/ (a *neutral* condition). Exp 1b differs from 1a in having a Female /ʃ/-favoring training phase instead of an /s/-favoring one as the first training block. It is also followed by a consecutive training block with Male A’s speech, whose sibilant characteristics are manipulated to be /s/-favoring (*opposite*), /ʃ/-favoring (*same*), or /s ʃ/-free (*neutral*) depending on specific conditions. By comparing the results of the three conditions within each sub-experiment, we are able to know to what extent the exposure to Male A’s speech matters for the categorization of Female A’s speech in the test phase.

The results from Exp 1a and 1b will show that listeners’ exposure to Male A’s speech influences their perception of Female A’s speech in the final test phase. Building on that, Exp 1c is designed to follow up on the question how much of Female A’s speech distribution is maintained in the final categorization phase. In other words, Exp 1c is aimed to tease apart the possibilities of *cumulative update* and *recency update*. The two *update* mechanisms share the property that they predict the acoustic distribution of Male A’s sibilants has been integrated and reflected in the final categorization result. However, they make different

predictions about whether Female A’s sibilant distribution is also retained to some extent with Male A’s distribution. *Recency update* refers to the possibility that listeners have forgotten their earlier training with Female A’s speech and only use what they have learned from Male A to apply to the categorization test. In contrast, *cumulative update* predicts that both the training with Female A and with Male A are exerting an influence on the final categorization result. By comparing the categorization results of participants who have received training only with Male A in 1c and those who have received training with Female A and Male A in Exp 1a and 1b, we are able to know how much the earlier training phase still matters to the categorization result.

Fig. 3.2 shows a summary of the experimental designs and procedures in each condition of the different sub-experiments in Experiment 1.

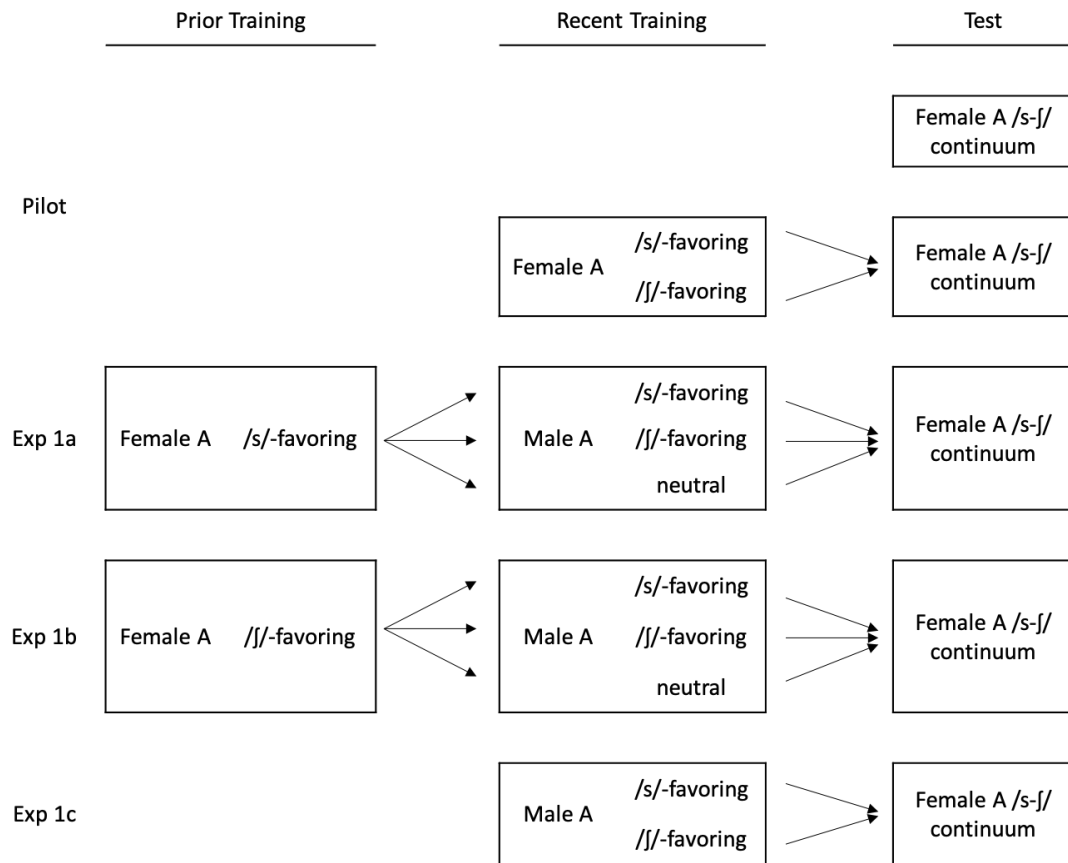


Figure 3.2: The structure of sub-experiments and conditions in Exp 1

When there are two training phases, I refer to them as the “prior” training phase and the “recent” training phase according to the order by which they take place. It is clear from Fig. 3.2 that the pilot study and Exp 1c differ from Exp 1a and 1b in the number of training phases involved. Pilot and Exp 1c only include no more than one training phase, and therefore the training participants have received can only be recent rather than prior. Pilot and Exp 1c differ in the speaker used in the training phase (if any), and Exp 1a and 1b differ in the acoustic condition of the first training phase with Female A.

3.2.2 Word list and recording

The stimuli used in Experiment 1 are manipulated from recordings of spoken words from Female A and Male A obtained following the procedure described in Chapter 2. For each speaker, the spoken words used in Experiment 1 consist of 17 /s/-containing words and 17 /f/-containing words (Table 2.2) each produced once with /s/ and once with /f/, 51 words without /s f/ (with 39 words selected from Table 2.4 and the remaining 12 words selected from Table 2.3 and 2.5), and 7 minimal pairs of words contrasted by word-initial /s f/ (selected from the 10 pairs in Table 2.5). These words are listed in the following:

- */s/-containing words*: compensate, democracy, dinosaur, embassy, episode, eraser, falsetto, faucet, hallucinate, legacy, medicine, obscene, parasite, peninsula, pregnancy, reconcile, rehearsal (N=17);
- */f/-containing words*: ambition, beneficial, brochure, commercial, crucial, efficient, evaluation, glacier, graduation, impatient, initial, negotiate, official, parachute, publisher, refreshing, vacation (N=17);
- */s-f/ minimal pairs*: same-shame, sake-shake, seat-sheet, sign-shine, sigh-shy, sell-shell, self-shelf (N=14);
- *Words without /s f/*: airline, among, anvil, average, banana, beloved, deer, earning, eyebrow, feeling, firefly, follow, foul, framing, gable, gargoyle, gavel, honey, iguana, moreover, Nepal, raccoon, raven, ribbon, row, runaway, thumbnail, town, verify, vol-

leyball, vulgar, waffle, wharf, time, tie, dragonfly, nothing, tag, village, Tim, down, tip, tear, marina, buffalo, dim, dime, lonely, journal, jewelry, January (N=51);

The critical sound of each word is annotated in Praat by hand and the annotations are saved as TextGrids. The center of gravity (COG) is measured for the 17 /s/ sounds in /s/-containing words and the 17 /ʃ/ sounds in /ʃ/-containing words from the two speakers. Their COG distributions are presented in Fig. 3.3 to represent the original acoustic properties of the two target phonemes from Female A and Male A. The error bars represent the means and the 95% confidence intervals of the COG measures for each phoneme of each speaker. We can see that the COG measures of the two speakers' sibilants are consistent with the general trend of gender variation. In other words, Female A's /s/ and /ʃ/ sounds are distributed in higher spectral frequencies than Male A's, although the difference is not as large as expected from the average acoustic distribution of sibilants from male and female speakers (e.g., as in Fig. 3.1).

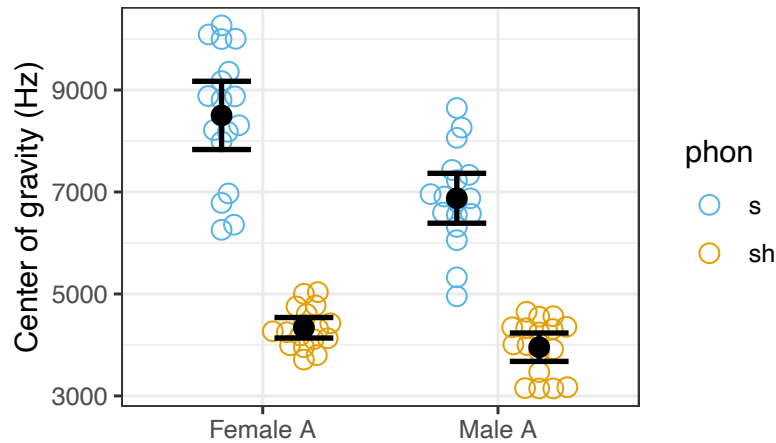


Figure 3.3: Mean and 95% confidence interval of the center of gravity of /s/ and /ʃ/ of Female A and Male A in natural speech production (Hz)

3.2.3 Step selection and synthesis

For the synthesis of the training stimuli, the critical proportion of sibilants sharing the same word frame (e.g., *compensate* and *compenshate*) are cut out and mixed with each other by

five steps of proportions. The five steps of sibilants for each word frame by each speaker vary from 0.3[s]0.7[f] to 0.7[s]0.3[f] with an increase of 0.1[s] and a decrease of 0.1[f] by each interval, and then they are spliced back to the lexical frame with the correct phoneme (e.g., the frame of compensate in the previous example). Finally, all the synthesized stimuli are normalized to 70 dB.

A lexical decision task is conducted to select the most ambiguous step of sibilant for each word frame to be used in the training phase. Participants needed to judge, for lexical frames spliced with each of the five sibilant steps, whether they are an English word or not. The results of this lexical decision task are shown in Section A. The mixture proportion that provides the most ambiguous (50%) categorization result is selected to be further used to construct the training materials in perceptual learning. If two sibilant steps are approximately equally far from the 50% point, then the mean of the two steps is used as the blending ratio for stimulus construction. Table 3.2 showed the most ambiguous steps chosen for each critical lexical frame for Female A and Male A through lexical decision.

ID	/f/-word	Female A	Male A	/s/-word	Female A	Male A
1	ambition	0.3	0.3	compensate	0.5	0.4
2	beneficial	0.25	0.35	democracy	0.5	0.35
3	brochure	0.35	0.35	dinosaur	0.6	0.35
4	commercial	0.35	0.3	embassy	0.55	0.4
5	negotiate	0.25	0.35	episode	0.45	0.4
6	crucial	0.25	0.25	eraser	0.6	0.4
7	official	0.25	0.4	falsetto	0.6	0.4
8	parachute	0.35	0.35	faucet	0.6	0.35
9	efficient	0.3	0.45	hallucinate	0.6	0.35
10	impatient	0.35	0.45	legacy	0.5	0.35
11	initial	0.25	0.5	medicine	0.4	0.5
12	vacation	0.35	0.5	obscene	0.5	0.35
13	evaluation	0.35	0.45	parasite	0.4	0.55
14	publisher	0.25	0.35	peninsula	0.6	0.45
15	refreshing	0.25	0.4	pregnancy	0.6	0.35
16	glacier	0.3	0.35	rehearsal	0.65	0.4
17	graduation	0.35	0.4	reconcile	0.4	0.35

Table 3.2: The proportion of [s] mixed in the most ambiguous Step of sibilant chosen for each word frame for Female A and Male A

Like the training stimuli, perception experiments are also conducted to select the steps and word frames of minimal pairs to be used for stimuli in the test phase. The test trials are generated by splicing 5 steps on a /s-ʃ/ continuum into 7 word frames of minimal pairs. Each set of lexical contexts contains interleaving /s/-adjacent and /ʃ/-adjacent lexical frames along the continuum, as shown in Table 3.3. For example, according to Table 3.3, the lexical frame of ?ake spliced onto Step 1, 3, and 5 are originally produced after /ʃ/, and ?ake spliced onto Step 2 and 4 are originally produced after /s/.

	Step 1 0.35[s]0.65[ʃ]	Step 2 0.45[s]0.55[ʃ]	Step 3 0.55[s]0.45[ʃ]	Step 4 0.65[s]0.35[ʃ]	Step 5 0.75[s]0.25[ʃ]
1. ?ake	shake	sake	shake	sake	shake
2. ?ame	same	shame	same	shame	same
3. ?elf	shelf	self	shelf	self	shelf
4. ?eat	seat	sheet	seat	sheet	seat
5. ?ell	shell	sell	shell	sell	shell
6. ?ign	sign	shine	sign	shine	sign
7. ?igh	shy	sigh	shy	sigh	shy

Table 3.3: The splicing of the test stimuli. Words in row 1-7 indicate the original word where the remaining of the lexical context spliced onto each sibilant step comes from.

3.2.4 Stimulus acoustics

Fig. 3.4 shows the relationship between the acoustic distributions of sibilants in different training conditions and those of the test phase after manipulation. It shows the means and 95% confidence intervals of the center of gravity for sibilants in the training stimuli (with points and error bars), and marks the COG values for the five steps of ambiguous sibilants in the test continuum (in red triangles).

From Fig. 3.4, we first see that the sibilants of Female A have higher COG values than those of Male A after manipulation. We can also see that sibilants in /ʃ/-favoring conditions generally have higher COG values than those in /s/-favoring conditions. However, given the relatively low frequencies of Female A’s sibilants compared to the average level of female sibilants, the COG of sibilants in the four conditions still have substantial overlap with one another in the acoustic space. In addition, the resulting continuum in the test phase does

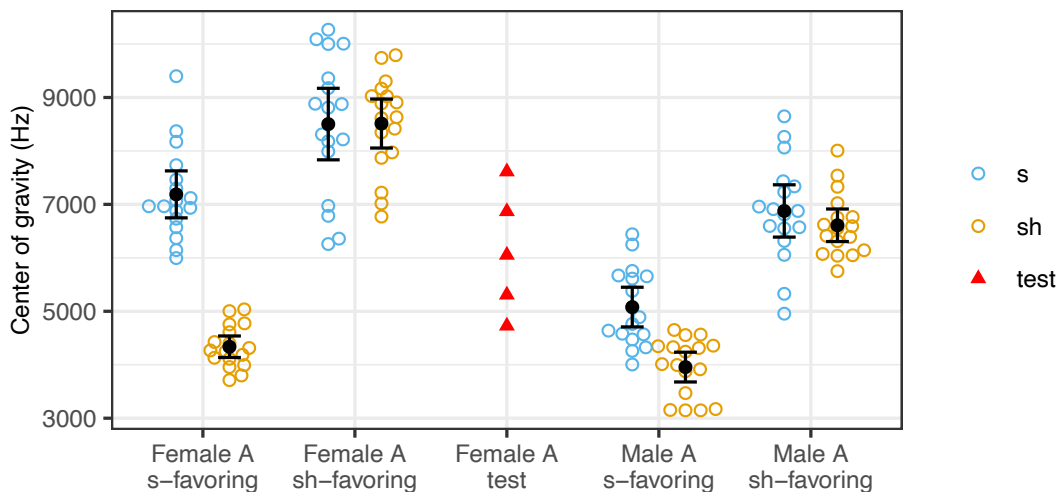


Figure 3.4: Mean and 95% confidence interval of the center of gravity of sibilants in different training phases and in the test phase in Exp 1 (Hz)

not seem to stay distinct from the /s f/ sounds in any of the training conditions. Maybe they are somewhat distinct from sibilants in the Male /s/-favoring condition, but there is still some overlap between the low end of the test continuum and the high end of the /s/ sounds in the Male /s/-favoring condition (see also Calder, 2019a,b; Hall et al., 2020, etc., for examples of the fierceness of /s/ as an indexical source of speaker gender).

3.3 Experiment and result

3.3.1 Pilot study: Learning Female A’s /s-f/

3.3.1.1 Experimental conditions and goals

The pilot study contains three experimental conditions, namely, *baseline*, */s/-favoring* learning with Female A, and */f/-favoring* learning with Female A. Participants in the baseline conditions complete a single test block, containing 35 test trials with ambiguous sibilants embedded in minimal pairs and 17 filler words without sibilants in Female A’s voice. The result of this condition is taken as a reference of the default /s-f/ perceptual boundary for Female A. participants in the two learning conditions first complete either an /s/-favoring

training block or an /f/-favoring training block with Female A’s speech before they proceed to complete the same test block as in the baseline condition.

The goal of the pilot study is twofold. The first goal is to demonstrate that the /s/-favoring and /f/-favoring perceptual learning effects have been successfully elicited with the speech of Female A. The second goal is to demonstrate that the 50% perceptual boundary between /s-f/ has been successfully aligned with the center of the continuum by default without prior training.

3.3.1.2 Participant

31 participants are recruited from Prolific to participate in the baseline condition. They are 18 female and 13 male, aged 18 to 58 years old ($Mean = 28.5, SD = 9.0$). Participants in the Female A /s/-favoring condition are recruited from Prolific. They are 15 male and 15 female, aging from 21 to 72 years old ($Mean = 35.6, SD = 12.7$). The participants in the Female /f/-favoring condition are recruited from the UPenn subject pool. They are 29 participants (6 male and 23 female), aging from 18 to 22 years old ($Mean = 19.9, SD = 1.3$).

3.3.1.3 Result

Fig. 3.5 shows the results of phoneme categorization by participants in the baseline conditions (in grey), Female A /s/-favoring condition (in yellow), and Female /f/-favoring condition (blue). We can see that firstly, the 50% point of the categorization boundary in the baseline condition aligns with the middle step of the continuum (Step 55). Secondly, the /s/-favoring training and the /f/-favoring training seem to have worked in inducing a perceptual bias towards the expected direction compared to the baseline condition: Participants in the /s/-favoring condition show more /s/-equivalent responses on every step of the categorization continuum than those in the baseline condition. Similarly, participants in the /f/-favoring condition show fewer /s/-equivalent responses on Step 45-75 than those in the baseline condition. Also, it seems that the /s/-favoring perceptual learning effect is larger than /f/-favoring effect with this set of stimuli of Female A’s speech.

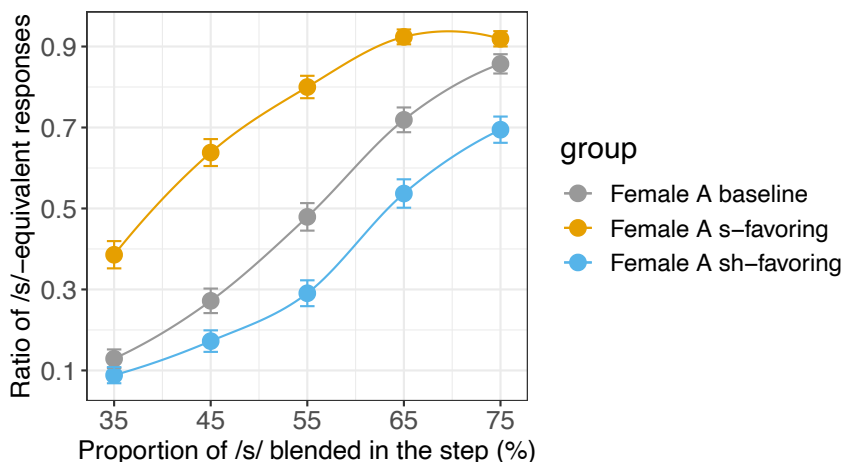


Figure 3.5: Exp 1 pilot: Boundary shift after exposure to Female A’s /s/-favoring and /ʃ/-favoring speech compared to the categorization baseline (mean and standard error)

I then run a logistic mixed-effects regression model to evaluate whether the two learning effects we see in Fig. 3.5 are statistically significant. A mixed-effects model is conducted to predict the Response of each trial ($S=0$, $SH=1$), with Step (35-75, scaled and centered), Trial (1-51, scaled and centered), Condition (treatment coded, reference: baseline), and Phoneme (the original phoneme associated to each auditory frame, sum-coded, reference: SH) as the fixed effects, Condition:Step and Condition:Trial as the interaction items, and Step by Subject, Phoneme by Subject, and Step by Frame as random slopes. The factor Phoneme turns out to exhibit co-linearity with the random intercept of Subject and the random slope of Step. I then took phoneme out from the random slopes and changed them to Step by Subject and Step by Frame. The result of the model (Model-pilot1) is shown in Table 3.4.

According to the statistical result, the main effect of Step is significant, suggesting that in the baseline condition, the larger the proportion of [s] is mixed in the stimulus, the less likely that stimulus is perceived as /ʃ/-equivalent ($\beta = -2.23, p < 0.001$). This trend also holds for the other two training conditions, as evidenced by the lack of significant interaction between Step and Condition for the /s/-favoring condition ($\beta = 0.47, p = 0.10$) and for the /ʃ/-favoring condition ($\beta = 0.40, p = 0.15$). It also suggests that the slopes of the categorization boundaries along the continuum steps are not different between the three

Fixed effects	Estimate	Std. Err.	z value	Pr(> z)
(Intercept)	0.13	0.38	0.34	0.74
Step	-2.23	0.21	-10.69	< 0.001***
Condition Female A s-favoring	-2.15	0.46	-4.69	< 0.001***
Condition Female A sh-favoring	1.07	0.45	2.36	0.02*
Trial	-0.47	0.10	-4.82	< 0.001***
PhonemeS	-0.57	0.06	-9.90	< 0.001***
Step:Condition Female A s-favoring	0.47	0.28	1.65	0.10
Step:Condition Female A sh-favoring	0.40	0.28	1.46	0.15
Condition Female A s-favoring:Trial	0.64	0.14	4.54	< 0.001***
Condition Female A sh-favoring:Trial	-0.02	0.14	-0.12	0.90

Model-pilot1: Response~Step*Condition+Condition*Trial+Phoneme+(Step|Subj)+(Step|Frame)

Table 3.4: The fixed effects of the logistic mixed-effects model in Exp 1 pilot

conditions.

The effect of Condition also turns out to be significant. It suggests that participants with /s/-favoring training experience are less likely to show /f/-equivalent responses than the baseline condition ($\beta = -2.15, p < 0.001$). Similarly, participants with /f/-training experience are more likely to show /f/-equivalent responses than the baseline condition ($\beta = 1.07, p = 0.02$). The effect sizes revealed by these coefficients are in line with our impression that the /s/-favoring shift is larger in magnitude than the /f/-favoring shift as a result of perceptual learning.

The model also reveals a main effect of Trial ($\beta = -0.47, p < 0.001$), meaning that test trials coming up at a later point are less likely to be perceived as /f/-containing in the baseline condition. The influence of Trial also applies to the /f/-favoring condition, as evidenced by the lack of significant interaction between Condition and Trial for this condition ($\beta = -0.02, p = 0.90$). This may be because the responses in these conditions consist of more /f/ than /s/ in general (although the difference seems small for the baseline condition), and listeners are trying to balance their responses towards the end of the block by reducing the number of /f/ responses. In contrast, this interaction is significant for the /s/-favoring condition ($\beta = 0.64, p < 0.001$), indicating that listeners are actually more likely to perceive a /f/ at a later point of the block. This is expected, because the responses of the /s/-favoring condition contain far more /s/ responses than /f/ ones, and the significant

positive effect of Trial implicates listeners’ tendency to balance their responses by reporting more /j/-equivalent responses towards the end of the test. Lastly, the effect of Phoneme is also significant, implying that the /s/-initial word frames are more likely to be perceived of bearing an /s/ than average even after splicing ($\beta = -0.57, p < 0.001$).

3.3.1.4 Summary

The results in Fig. 3.5 and Table 3.4 both suggest that the design of training phases with Female A’s speech works to achieve the goals laid out in the beginning of this section, namely, to align the 50% response point of the categorization boundary with the center of the continuum, and to induce a significant amount of boundary shift towards the expected directions.

3.3.2 Exp 1a: Previous /s/-favoring training with Female A

3.3.2.1 Experimental conditions and goals

The goal of Exp 1a is to tease apart the possibilities of *retention*, *reset*, and *update* as alternative possibilities of the mechanism involved in sibilant perceptual learning with multiple speakers. As a brief reminder of these concepts (described in Section 7.1), *retention* means that perceptual learning operates in a speaker-specific way; *reset* means that listeners reset their perceptual expectation to the default each time they encounter a new speaker; *update* refers to the possibility that listeners integrate the acoustic distributions they have learned to update their perceptual expectations and generalize this knowledge across speakers. *Update* may take the form of either cumulative update or recency update. The two mechanisms both predict an amount of integration of Male A’s speech distribution, but they differ in their predictions regarding how much of the learning with Female A’s speech is still maintained. This two possibilities cannot be teased apart in the current sub-experiment; instead this question will be addressed in Exp 1c.

To distinguish between the above three possibilities, I designed three experimental conditions in Exp 1a with different combinations of speaker and acoustic distribution for their

training phases. These three conditions are referred to as Two genders - *opposite*, Two genders - *same*, and Two genders - *neutral*. Participants in all three conditions first complete /s/-favoring training phase with Female A and then a training phase with Male A's speech with different sibilant manipulations depending on the condition they are assigned to. Male A's speech in the second learning phase is either /s/-favoring (*same* as Female A's speech), /ʃ/-favoring (*opposite* to Female A's speech), or sibilant-free (the neutral condition). In the end, participants are tested on Female A's sibilants on a /s-ʃ/ continuum spliced into minimal pairs.

Crucially, the three possibilities of cross-speaker perceptual learning behaviors each make different predictions about the outcomes of the three training conditions. If listeners *retain* the knowledge of acoustic distribution that is specific to Female A and do not apply what they have learned from Male A's speech to the categorization of Female A's speech, then we expect the categorization results of all the three conditions to show a similar shift towards /s/. If listeners *reset* their perceptual expectation for every new talker they encountered, then we expect that the categorization results reflect neither Female A's nor Male A's acoustic distribution. Instead, all the three conditions would have categorization boundaries overlapping with the boundary of the baseline condition. If listeners integrate the acoustic distributions of Male A to *update* their perceptual expectations and generalize this knowledge across speakers, we expect to see more /s/-equivalent responses in the Two genders - *same* condition and more /ʃ/-equivalent responses in the Two genders - *opposite* condition, and the categorization boundary of the Two genders - *neutral* condition lying between the above two conditions. In other words, if the perceptual boundary in the test consistently patterns with the acoustic distributions of the second learning phase, then it supports an *update* account.

3.3.2.2 Participant

Participants in the three conditions of Exp 1a are all recruited from Prolific. There are 32 participants in the Two genders - *same* condition. They are 10 male, 21 female, and

one non-binary in gender, aging from 18 to 68 years old ($Mean = 29, SD = 12$). 33 participants are recruited for the Two genders - *opposite* condition, including 20 female and 13 male, aging from 19-66 years old ($Mean = 33, SD = 12$). Lastly, 30 participants are recruited for the Two genders - *neutral* condition. They include 10 male listeners, 19 female listeners, and one non-binary gender listener. Their ages range from 19 to 57 ($Mean = 30, SD = 10$). Along with the data of the above participants, I have also plotted the results of participants in the baseline condition and the Female A /s/-favoring condition as a reference (see Section 3.3.1.2 for the information on those participants).

3.3.2.3 Result

Fig. 3.6 shows the means and standard errors of the categorization result at each fricative step in different experimental conditions, along with the results of the baseline condition and the Female A /s/-favoring conditions represented by the grey lines (same as the grey line and the yellow line in Fig. 3.5). The blue lines indicates the percentage of /s/-equivalent responses in the Two genders - *same* condition (dashed line) and in the Two genders - *opposite* condition (solid line). The yellow line lying in between is the average /s/ responses at each step in the Two genders - *neutral* condition, where the speech of the intervening male speaker does not contain any /s/ or /ʃ/.

Recall that participants in all the three conditions have had identical exposure to Female A's /s/-favoring speech in the first training phase. Therefore, we can attribute any differences observed between these conditions to their training in the second phase with Male A, under the assumption that participants on different conditions have the same baseline perception of /s ʃ/. Indeed, the overall /s/ responses of the three two-gender learning conditions are consistent with the sibilant properties of Male A's speech in their second training phase: The most /s/-equivalent responses are exhibited in the Two genders - *same* condition where Male A's sibilants are /s/-favoring, and the least /s/-equivalent responses are exhibited in the Two genders - *opposite* condition where Male A's sibilants are /ʃ/-favoring, with the results of the Two genders - *neutral* condition where there are no /s/ or /ʃ/ in

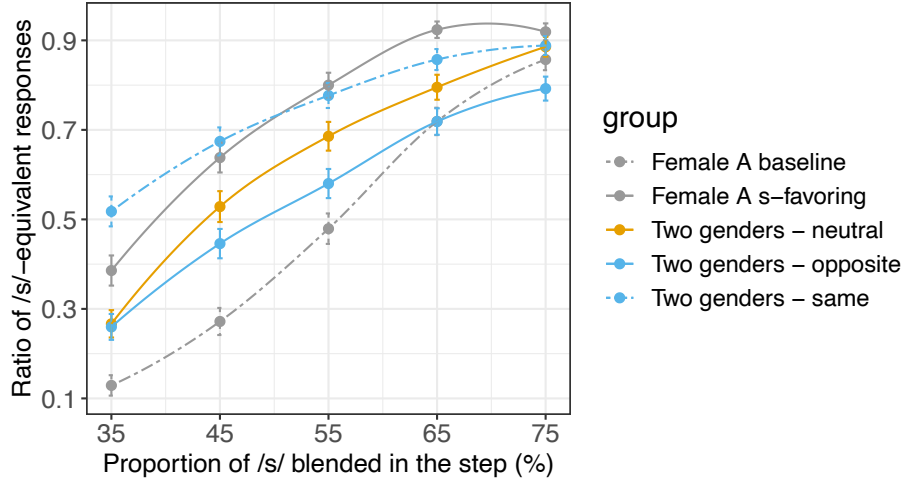


Figure 3.6: Exp 1a: /s/ response rate as a result of cross-gender perceptual learning with different sibilant distributions (mean and standard error)

Male A’s speech lying in between. Such a pattern seems to lend support to an *update* mechanism, which predicts that the acoustic properties of Male A has been integrated to update further perceptual expectations. In addition, the three two-phase learning conditions seem to have categorization boundaries of shallower slope than either the baseline condition and the Female A /s/-favoring condition, which might also be a result of exposure to Male A’s speech. In other words, simply being exposed to Male A’s speech, even if it contains no fricatives at all, changes the way people categorize Female A’s fricatives.

We can also see that results of the three two-phase learning conditions approximately fall between the *baseline* condition and the Female /s/-favoring condition, with an approximately equal amount of /s/ responses in the Female /s/-favoring condition and the Two genders - *same* condition. There are several possible reasons for this decay. One is that the decay in the /s/-equivalent rates of the three two-gender conditions compared to the Female A /s/-favoring condition is due to a lapse in time during the second-training phase. Another possibility is that listeners are still updating to Male A’s /s f/ boundary they have conceived of even though they have never heard any actual sibilants from Male A. These possibilities will be further discussed in the Discussion section (Section 3.4).

A mixed-effects model (Model-1a) is fitted to the categorization data of the three two-

gender conditions in this sub-experiment and the baseline and Female A /s/-favoring condition from what is presented in the pilot study. The model uses Step (scaled and centered), Trial (scaled and centered), and Condition (treatment coded, ref: baseline), and Phoneme (the original phoneme associated to each auditory frame, sum-coded, reference: SH) as the fixed effects, Condition:Step and Condition:Trial as the interaction items, and Step by Subject and by Frame as random slopes, to predict the Response of each trial (S=0, SH=1). The result is shown in Table 3.5.

Fixed Effects	Estimate	SE	z value	Pr(> z)
(Intercept)	0.11	0.41	0.27	0.79
Step	-2.21	0.18	-12.25	< 0.001***
Condition Female A s-favoring	-2.20	0.53	-4.17	< 0.001***
Condition Two genders - same	-2.13	0.52	-4.11	< 0.001***
Condition Two genders - neutral	-1.29	0.52	-2.48	0.01*
Condition Two genders - opposite	-0.55	0.50	-1.09	0.28
Trial	-0.47	0.10	-4.84	< 0.001***
PhonemeS	-0.57	0.04	-13.10	< 0.001***
Step:Condition Female A s-favoring	0.42	0.25	1.68	0.09
Step:Condition Two genders - same	0.91	0.25	3.69	< 0.001***
Step:Condition Two genders - neutral	0.39	0.25	1.59	0.11
Step:Condition Two genders - opposite	0.83	0.23	3.57	< 0.001***
Condition Female A s-favoring:Trial	0.62	0.14	4.49	< 0.001***
Condition Two genders - same:Trial	0.44	0.14	3.17	0.001**
Condition Two genders - neutral:Trial	0.57	0.14	4.22	< 0.001***
Condition Two genders - opposite:Trial	0.25	0.13	1.94	0.053

Model-1a: Response~Step*Condition+Condition*Trial+Phoneme+(Step|Subj)+(Step|Frame)

Table 3.5: The fixed effects of the logistic mixed-effects model in Exp 1a

In Table 3.5, again, we see significant main effects of Step, Condition, and Phoneme. The significantly negative Step effect indicates that, with a larger proportion of [s] is mixed in the stimulus, the probability of that stimulus being perceived as /j/-equivalent is lower ($\beta = -2.21, p < 0.001$). The model also reveals a significant interaction between Step and Condition for the Two genders - *same* condition ($\beta = 0.91, p < 0.001$) and the Two genders - *opposite* condition ($\beta = 0.83, p < 0.001$). Since the proportion of /j/-equivalent responses should be decreasing with the increase of step, a positive value for the interaction item would indicate that the decrease becomes shallower instead of becoming sharper. Therefore,

the significant interaction between Step and Condition means that listeners who have had exposure to Male A's /s/-favoring or /j/-favoring speech show a shallower categorization boundary than those in the baseline condition. In contrast, this interaction is not significant for the Two genders - *neutral* condition ($\beta = 0.39, p = 0.11$).

The main effect of Condition is significant for all experimental conditions except for the Two genders - *opposite* condition. The other two conditions newly introduced in this experiment, namely, Two genders - *same* ($\beta = -2.13, p < 0.001$), and Two genders - *neutral* ($\beta = -1.29, p = 0.01$), both show significantly more /s/-equivalent responses (and fewer /j/-equivalent responses) than the baseline condition. This is expected because stimuli on the training phases in the three conditions are either /s/-favoring or /s j/-free, giving rise to an overall boost in /s/. The categorization result in the Two genders - *opposite* condition is not essentially different from the baseline condition ($\beta = -0.55, p = 0.28$). This suggests that listeners in the Two genders - *opposite* condition have integrated the /j/-favoring distribution during their training with Male A and used this knowledge to cope with the categorization task with Female A. This cancels out the influence of the earlier /s/-favoring training with Female A, leading the final result to return to the baseline again.

In addition, the effects of Trial ($\beta = -0.47, p < 0.001$) and Phoneme ($\beta = -0.57, p < 0.001$) are also significant on the baseline condition, as we have already explained for Table 3.4 in the previous section. Again, this means that listeners are less likely to report on /j/ for later trials and for auditory frames original produced with /s/. As with the Female A /s/-favoring condition where the interaction between Condition and Trial is significant ($\beta = 0.62, p < 0.001$), Condition: Trial is also significant for the Two genders - *same* condition ($\beta = 0.44, p = 0.001$) and the Two genders - *neutral* condition ($\beta = 0.57, p < 0.001$), and it is marginally significant for the Two genders - *opposite* condition ($\beta = 0.25, p = 0.053$). This implies that listeners on the three conditions are reporting more /j/-equivalent responses at a later point of the test phase. Again, might be driven by the response distribution that the three conditions overall have fewer /j/-equivalent responses than the baseline condition.

To further check whether the second-phase exposure to Male A’s speech has shifted listener perceptual boundary further away from the Female A /s/-favoring condition, I relevelled the Condition factor with “Female A s-favoring” as the baseline and re-ran the model. The result shows that among the three two-gender conditions, only the Two genders - *opposite* condition exhibits a significant difference from the Female A /s/-favoring condition ($\beta_{opposite} = 1.64, p = 0.001$), while the other two conditions do not ($\beta_{neutral} = -0.9, p = 0.09$; $\beta_{same} = 0.07, p = 0.89$). This suggests a significant influence of the Male A /f/-favoring training but none of the Male A /s/-favoring training on top of the Female A /s/-favoring training in the two-gender perceptual learning in Exp 1a.

3.3.2.4 Summary

The goal of Exp 1a is to tease apart the possibilities of *retention*, *reset*, and *update*. The categorization results differ among the three two-gender conditions in a consistent way with their exposure to Male A’s speech. This helps us rule out the possibility of *retain* which predicts no difference between these conditions and that they all show a perceptual shift towards /s/. Also, the boundary reveals a reset to baseline in the Two genders - *opposite* condition but not in the Two genders - *neutral* condition and in the Two genders - *same* condition. These results rule out the possibility that hearing a new voice will make listeners *reset* their perceptual expectations. The result lends support to an *update* process that reflects the joined effects of perceptual learning with Female A and Male A.

3.3.3 Exp 1b: Previous /f/-favoring training with Female A

3.3.3.1 Experimental conditions and goals

The goal of Exp 1b is to evaluate whether the findings of Exp 1a can be replicated after the acoustic properties of Female A have been changed. Like in Exp 1a, Exp 1b also has three experimental conditions where listeners accept two sequential training phases with Female A and Male A, and the three conditions are contrasted by the acoustic distributions of Male A’s sibilants. Participants in all three conditions first complete /f/-favoring training phase

with Female A and then a training phase with Male A’s speech with different sibilant manipulations depending on the condition they are assigned to. Male A’s speech in the second learning phase is either /f/-favoring (*same* as Female A’s speech), /s/-favoring (*opposite* to Female A’s speech), or sibilant-free (the neutral condition). In the end, participants are tested on Female A’s sibilants on a /s-f/continuum spliced into words of minimal pairs. The only difference between Exp 1a and 1b lies in that the first training phase with Female is /s/-favoring in Exp 1a and /f/-favoring in Exp 1b.

There are some consequences of manipulating the acoustic bias of Female A’s speech towards a different direction. One of them is that the acoustic distributions of Female A’s sibilants and Male A’s sibilants become more distinctive in the acoustic space. This is especially reflected by the opposite conditions. Fig. 3.7 shows the COG measures of the sibilants of Male A and Female A in the Two genders - *opposite* condition. We can see that the sibilants of Female A and Male A show similar mean COGs for each of the two phonemes in Exp 1a (left), whereas the mean COGs of the same phoneme are very different between speakers, as shown in Exp 1b (right).

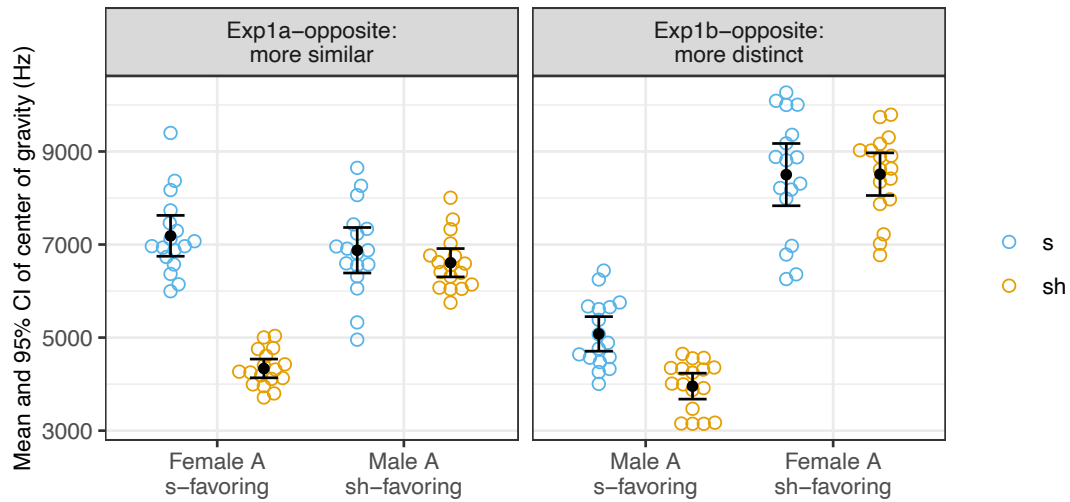


Figure 3.7: The COG of sibilants in the Two genders - opposite conditions in Exp 1a and 1b (mean and 95% CI)

It is unclear whether the relationship between the acoustic distributions of the two

speakers in training would affect the conclusion of cross-speaker perceptual learning obtained in Exp 1a. If our previous finding about the *update* of perceptual expectation across talkers is not constrained on acoustic dissimilarity, then we expect to see that the results of the three conditions in Exp 1b show similar patterns to those in Exp 1a. Otherwise, if the update of perceptual expectations across talkers is hindered by a change in the acoustic properties the two speakers' sibilants, then we would expect that the categorization boundary either stays at the baseline position or remains a shift towards /f/ as a result of previous training with Female A.

3.3.3.2 Participant

Participants in this experiment are all recruited from the UPenn subject pool. There are 29 participants in the same condition. They are 20 female and 9 male, aging from 19-26 years old ($Mean = 20, SD = 1.6$). The opposite condition has 33 participants, including 19 female and 14 male, aging from 18-22 years old ($Mean = 19.6, SD = 1.2$). The neutral condition has 30 participants. Among them there are 22 female and 8 male, aging from 16-29 years old ($Mean = 19.8, SD = 2.1$). Along with the responses of these participants, I have also plotted the results of participants in the baseline condition and the Female A /f/-favoring condition to provide visual references (see Section 3.3.1.2 for the information of those participants).

3.3.3.3 Result

Fig. 3.8 shows the means and standard errors of /s/-equivalent response rates at each step in the three two-phase learning conditions, along with the results of the baseline and pilot training conditions represented by the grey lines. The blue lines indicates the percentage of /s/-equivalent responses in the Two genders - *same* condition (dashed line) and in the Two genders - *opposite* condition (solid line). The yellow line in between is the result of the Two genders - *neutral* condition, where the speech of the intervening male speaker does not contain any /s/ or /f/.

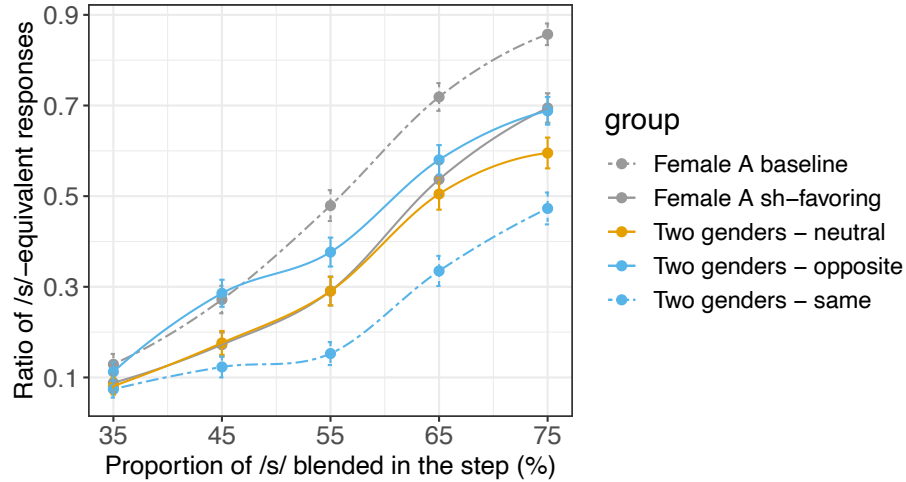


Figure 3.8: Exp1b: /s/ response rate as a result of cross-gender perceptual learning with different sibilant distributions (mean and standard error)

Recall that participants in all the three conditions have had identical exposure to Female A’s /f/-favoring speech in the first training phase, which has induced a shift towards /f/ as represented by the grey solid line. On top of that, a second-phase exposure to Male A’s /f/-favoring speech in the Two genders - *same* condition causes further shift towards /f/, whereas a second-phase exposure to Male A’s /s/-favoring speech in the Two genders - *opposite* condition shift the categorization boundary backwards but it has not returned to the baseline. The results of the Two genders - *neutral* condition where there are no /s/ or /f/ in Male A’s speech lies in between. Like in Exp 1a, the overall /s/ responses of the three two-gender learning conditions are consistent with the sibilant properties of Male A’s speech in their second training phase, suggesting that listeners have updated their perceptual expectations through exposure to Male A’s speech and applied their updated expectations to the phoneme categorization of Female A’s speech. These findings are consistent with an *update* account.

One way the result of Exp 1b differ from that of Exp 1a is that the perception boundary of the Two genders - *neutral* condition roughly overlaps with the boundary in the Female A /f/-favoring condition, meaning that a second-phase exposure to Male A’s speech without /s f/ does not diminish the learning effect established in the first phase. This is not the case

of Exp 1a, where we can observe a decay of the boundary shift towards /s/ compared to the Female /s/-favoring condition. Recall that I have laid out several possibilities for this decay. One possibility is that this is caused by a lapse in time during the second-training phase. The other possibility is that listeners are adapting to Male A's boundary that they have conceived of without actually hearing the actual production of Male A's sibilants. If the patterns we observed for Exp 1b is statistically validated, then it might suggest that the learning decay of the neutral condition in Exp 1a is more likely caused by adaptation towards the expected boundary of the speaker rather than memory decay. In other words, listeners might have inferred the distributional properties of Male A's sibilants from his neutral speech and adapted to that distribution. Moreover, the categorization boundaries in the genders - *same* and *opposite* conditions become obviously shallower in slope with a concavity at Step 55. This can be explained by the experimental design of the test phase shown in Table 3.3, where the split of /s/-embedding frames and /ʃ/-embedding ones is not even across steps. I will elaborate on these points in more details in the section of general discussion.

Similarly, a mixed-effects model (Model-1b) is conducted with Step (35-75, scaled and centered), Trial (1-51, scaled and centered), and Condition (treatment coded, reference: baseline), and Phoneme (the original phoneme associated to each auditory frame, sum-coded, reference: SH) as the fixed effects, Condition:Step and Condition:Trial as the interaction items, and Step by Subject and by Frame as random slopes, to predict the Response of each trial (S=0, SH=1). The result is shown in Table 3.6.

According to Table 3.6, the main effect of Step is significant, suggesting that the larger the proportion of [s] is mixed in the stimulus, the less likely that stimulus is perceived as /ʃ/-equivalent ($\beta = -2.08, p < 0.001$). The magnitude of the Step effect is found to vary with Condition: The model reveals a significant interaction between them for the two conditions that involves training with Male A's sibilants ($\beta = 0.83, p < 0.001$ for the Two genders - *same* condition; $\beta = 0.71, p < 0.001$ for the Two genders - *opposite* condition; $\beta = 0.64, p = 0.003$ for the Two genders - *neutral* condition). Since the proportion of /ʃ/

Fixed effects	Estimate	SE	z value	Pr(> z)
(Intercept)	0.05	0.33	0.14	0.89
Step	-2.08	0.17	-12.44	< 0.001***
Condition Female A sh-favoring	1.03	0.41	2.53	0.01*
Condition Two genders - same	1.96	0.41	4.73	< 0.001***
Condition Two genders - neutral	1.09	0.40	2.71	0.007**
Condition Two genders - opposite	0.58	0.39	1.48	0.14
Trial	-0.45	0.09	-4.78	< 0.001***
PhonemeS	-0.52	0.04	-12.54	< 0.001***
Step:Condition Female A sh-favoring	0.40	0.22	1.78	0.08
Step:Condition Two genders - same	0.83	0.23	3.60	< 0.001***
Step:Condition Two genders - neutral	0.64	0.22	2.92	0.003**
Step:Condition Two genders - opposite	0.71	0.21	3.32	< 0.001***
Condition Female A sh-favoring: Trial	0.00	0.13	-0.04	0.97
Condition Two genders - same: Trial	0.34	0.14	2.47	0.01*
Condition Two genders - neutral: Trial	0.16	0.13	1.23	0.22
Condition Two genders - opposite: Trial	0.47	0.12	3.74	< 0.001***

Model-1b: Response~Step*Condition+Condition*Trial+Phoneme+(Step|Subj)+(Step|Frame)

Table 3.6: The fixed effects of the logistic mixed-effects model in Exp 1b

responses should be decreasing with the increase of Step, a positive value for the interaction item would indicate that the decrease becomes shallower instead of sharper. Therefore, the significant interaction between Step and Condition means that listeners who have had exposure to Male A’s speech show a shallower categorization boundary than the baseline condition. In contrast, the difference in categorization slope between the baseline condition and the Female A sh-favoring condition is only marginally significant ($\beta = 0.40, p = 0.08$).

The estimates of Condition in all conditions are positive, meaning that each of them has /f/ equivalent responses than the baseline condition quantitatively. The effect of Condition is significant for the Two genders - *same* condition ($\beta = 1.96, p < 0.001$), the Two genders - *neutral* condition ($\beta = 1.09, p = 0.007$), and the Female A sh-favoring condition ($\beta = 1.03, p = 0.01$), but it is not significant for the Two genders - *opposite* condition ($\beta = 0.58, p = 0.14$). These results means that the former three conditions are significantly different from the baseline condition in that they exhibit more /f/-equivalent responses, whereas the Two genders - *opposite* condition does not differ significantly from the baseline condition. Basically, the Condition differences revealed by Table 3.6 have replicated the

result of Exp 1a, lending support to a *update* account where the speech of Male A has been integrated and reflected in the final categorization stage, leading to the categorization boundary in the Two genders - *opposite* condition to return to the baseline condition.

The model also reveals main effects of Trial ($\beta = -0.45, p < 0.001$) and Phoneme ($\beta = -0.52, p < 0.001$). The Trial effect implies that trials coming up at a later point are less likely to be perceived as /j/-containing. The Phoneme effect implies that the /j/-initial word frames are more likely to be perceived as /j/-initial. The Trial effect interacts with Condition for the Two genders - *same* condition ($\beta = 0.34, p = 0.01$) and the Two genders - *opposite* condition ($\beta = 0.47, p < 0.001$), but not for the Female A sh-favoring condition ($\beta < 0.001, p = 0.97$) or the Two genders - *neutral* condition ($\beta = 0.16, p = 0.22$). Simply put, the Female A sh-favoring condition and the neutral condition share the Trial effect in the baseline condition such that later trials are more likely to be perceived as /s/. However, this effect is significantly smaller for the two conditions involving training with Male A's sibilants (*same* and *opposite*).

Finally, I re-evaluated the model with “Female A sh-favoring” as the baseline to examine whether the second-phase exposure to Male A's speech has shifted listener perceptual boundary further away from the Female A /j/-favoring condition. This time, only the Two genders - *same* condition exhibits significant difference from the Female A /j/-favoring condition ($\beta_{same} = 0.93, p = 0.03$), while the other two conditions do not ($\beta_{neutral} = 0.06, p = 0.88$; $\beta_{opposite} = -0.45, p = 0.26$).

3.3.3.4 Summary

In Exp 1b, again, I evaluate whether listeners use their knowledge of the acoustic distributions of a recently encountered male talker to categorize the speech of a previously encountered female talker. Although we have already seen some evidence from Exp 1a that this might be the case, we still wonder whether this finding applies to Exp 1b where the acoustic distributions of the two speakers are farther apart from each other. The result of Exp 1b shows that it does. On one hand, a significant difference between the Two genders

- *opposite* condition and the Female A /f/-favoring condition helps rule out the possibility of *retention*, which predicts that exposure to Male A's speech would not cause perceptual shifts to the categorization of Female A's speech. On the other hand, the lack of difference between the Female A /f/-favoring condition and two of the two-gender conditions, *same* and *neutral*, helps rule out the possibility of *reset*, which predicts that simply exposure to Male A's speech is sufficient to make listeners reset their perceptual learning boundaries, no matter what Male A's acoustic distribution looks like. Up to this point, the result of Exp 1a and 1b both lend support to the account of *update*, which predicts that the acoustic distribution of Male A is also integrated to cope with categorization with Female A's speech.

3.3.4 Exp 1c: No previous training with Female A's /s-ʃ

3.3.4.1 Experimental conditions and goals

In Exp 1a and 1b, listeners have been trained on the sibilants of Female A and Male A in two sequential training phases, and then they complete a categorization test that evaluates which perceptual expectation(s) they would use for the identification of Female A's sibilants. Results obtained so far consistently indicate that listeners have used their knowledge about Male A's sibilants in the categorization of Female A's sibilants. In Exp 1c, I ask to what extent the results reflect training with Male A's speech alone (where training with Female A's speech has been forgotten) instead of the joint training with Female A's and Male A's speech. Will training with Male A's speech on its own give rise to the same categorization boundary as induced by exposure to both Female A's and Male A's speech?

The two possible answers to the above question correspond to two possibilities laid out in Section 2.4: The situation where listeners have integrated both Female A's and Male A's sibilant properties to cope with the categorization is called *cumulative update*, and the situation where listeners have integrated Male A's sibilant properties only and have left behind Female A's sibilant properties is called *recency update*. These two situations share the similarity that the training with the recent speaker (Male A in this case) is exerting an influence on the final categorization result, but they make different predictions about

whether the outcome of earlier learning experience is still retained.

The research goal of Exp 1c is to evaluate which of these situations is a more accurate description of the categorization results we have observed. Participants in this sub-experiment only receive training on Male A’s speech in a single phase, which is manipulated to be either /s/-favoring or /ʃ/-favoring. Then they are tested with /s-ʃ/ minimal pairs of Female A. I compare the categorization result of training with Male A only (in 1c) and the result of training with two speakers in 1a and 1b, where Male A’s speech shares the same distribution with 1c. If the results turn out to be similar, then it suggests that the shift observed in the two-gender training conditions in Exp 1a and 1b can be largely attributed to the training with Male A in the second phase alone. If they turn out to be different, then it suggests that the training with Female A in the first stage also exerts an influence on the categorization results we have observed in 1a and 1b.

3.3.4.2 Participant

34 participants are recruited to participate in the Male A /ʃ/-favoring condition from the UPenn subject pool. They are 14 male and 20 female, aging from 18 to 31 years old ($Mean = 20.3, SD = 2.5$). 27 participants are recruited for the Male A /s/-favoring condition. Among them, 17 are recruited from Prolific and the remaining 10 are recruited from the UPenn subject pool. They are 10 male and 17 female, aging from 18 to 56 years old ($Mean = 24.5, SD = 8$).

3.3.4.3 Result

The first analysis I did is to compare the categorization boundaries of the Female A baseline condition and the Male-only conditions. This comparison gives us an idea about whether training with Male A’s speech successfully induced an amount of shift to the intended direction. Fig. 3.9 shows the results of the two experimental conditions in Exp 1c, namely, training with Male A’s speech only, in either the s-favoring or sh-favoring directions, along with the baseline categorization results along the test continuum in Female A’s voice.

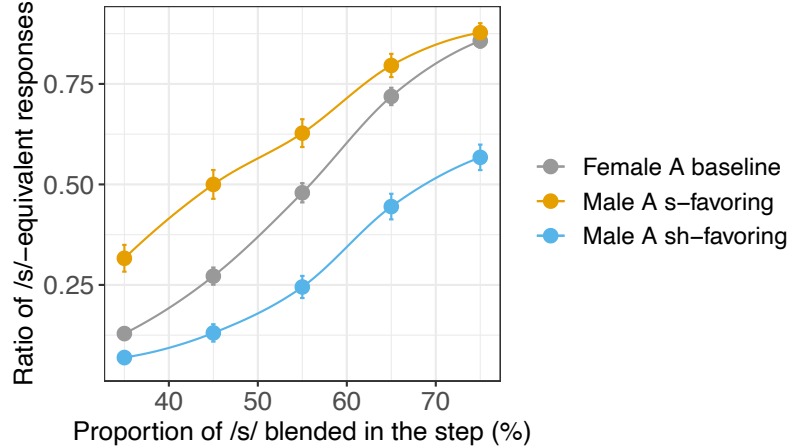


Figure 3.9: Exp 1c: /s/ response rate as a result of training with Male A compared to the baseline (mean and standard error)

Fig. 3.9 show that the training stimuli of Male A successfully induced an amount of perceptual shift to the intended direction compared to the baseline categorization. A logistic mixed-effects model (Model-1c-a) is fitted to examine whether the categorization results of the Male s-favoring condition and the Male /j/-favoring condition are significantly different from the Female A baseline. The dependent variable is the response of each trial (S:0, SH:1). The main effects are Condition (Male A s-favoring/Male A sh-favoring/Female A baseline, treatment coded, baseline: Female A baseline), Phoneme (the original phoneme associated to each auditory frame, sum-coded, baseline: SH), and Step and Trial (both scaled and centered). The models also include Condition:Step and Condition:Trial as the interaction items, by-Subject Step as the random slope, and Frame as a random intercept (the random slope of Frame is highly correlated with that of Step and is therefore dropped). The result of this model is shown in Table 3.7.

Table 3.7 reveals a significant difference between the Female A baseline condition and the Male A s-favoring condition ($\beta = -0.8, p = 0.043$), as well as a significant difference between the Female A baseline condition and the Male A sh-favoring condition ($\beta = 1.44, p < 0.001$). The model also reveals significant main effects of Trial ($\beta = -0.44, p < 0.001$) and Phoneme ($\beta = -0.49, p < 0.001$), as well as significant interactions of Step:Condition for both conditions ($\beta_{MaleS} = 0.62, p = 0.002; \beta_{MaleSH} = 0.55, p = 0.006$) and Condition:Trial

Fixed effects	Estimate	Std. Err.	z value	Pr(> z)
(Intercept)	0.17	0.32	0.52	0.60
Step	-2.04	0.15	-13.31	< 0.001***
Condition Male A s-favoring	-0.80	0.39	-2.03	0.043*
Condition Male A sh-favoring	1.44	0.39	3.68	< 0.001***
Trial	-0.44	0.09	-4.81	< 0.001***
PhonemeS	-0.49	0.05	-9.64	< 0.001***
Step:Condition Male A s-favoring	0.62	0.20	3.09	0.002**
Step:Condition Male A sh-favoring	0.55	0.20	2.76	0.006**
Condition Male A s-favoring:Trial	0.44	0.13	3.47	< 0.001***
Condition Male A sh-favoring:Trial	0.17	0.12	1.38	0.17

Model-1c-a: Response~Step*Condition+Condition*Trial+Phoneme+(Step|Subj)+(1|Frame)

Table 3.7: The fixed effects of the logistic mixed-effects model evaluating the effect of training with Male compared to the Female A baseline in Exp 1c

for the Male A s-favoring condition ($\beta = 0.44, p < 0.001$). Crucially, this model suggests that the training materials have successfully triggered an amount of perceptual boundary shift.

The second set of comparisons I made is between the Male-only training conditions and the Two-gender conditions that contain the same Male training phase. Fig. 3.10 shows the results of Exp 1c along with Two-gender conditions sharing the same manipulation of Male A’s speech in Exp 1a and 1b. The left facet shows the results of training with Male A’s /s/-favoring alone speech (grey) and training with Male A’s /s/-favoring speech preceded by either Female A’s /s/-favoring speech (blue) or Female A’s /j/-favoring speech (yellow). Similarly, the right facet shows the results of training with Male A’s /j/-favoring speech alone (grey) and training with Male A’s /j/-favoring speech preceded by either Female A’s /j/ speech (blue) or Female A’s /s/-favoring speech (yellow).

Crucially, in both of the two facets, we can see that the grey line lies between the yellow line and the blue line. This means that a preceding training phase with Female A’s /s/-favoring speech results in a higher amount of /s/-equivalent responses (represented by the blue line in the left facet and yellow line in the right facet), whereas a preceding training phase with Female A’s /j/-favoring speech results in a lower amount of /s/-equivalent responses (represented by the yellow line in the left facet and the blue line in the right

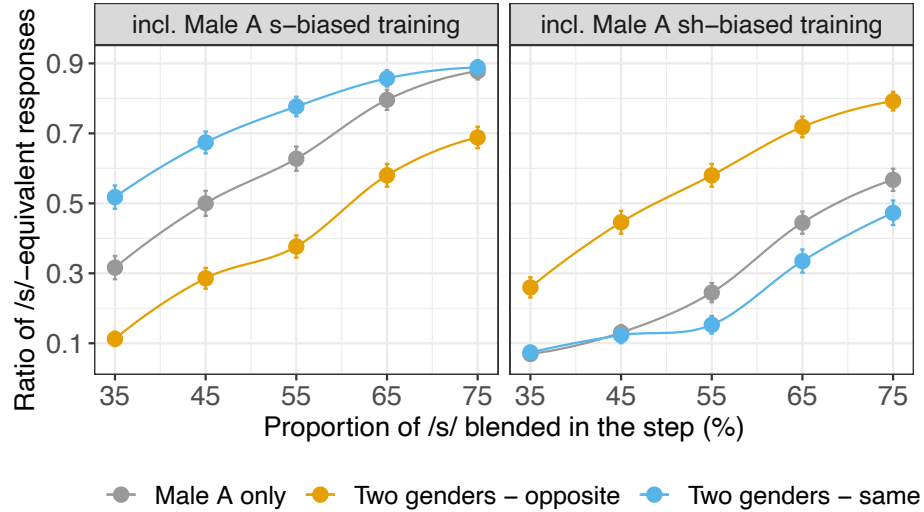


Figure 3.10: Exp 1c: /s/ response rate as a result of training with Male A compared to two-genders training conditions (mean and standard error)

facet), compared to the result of training with Male A’s speech alone.

To evaluate whether the difference between the grey line and the other two lines in each of the facet is significant, I evaluated another two logistic mixed-effects models (Model 1c-b and 1c-c) respectively for the data of experimental conditions including Male A /s/-favoring training and those including Male A /f/-favoring training. The dependent variable is SH responses in the test. The main effects are Condition (Male only, Two genders - opposite, Two genders - same; treatment coded, baseline: Male only), Step and Trial (both scaled and centered), and Phoneme (the original phoneme associated to each auditory frame, sum-coded, baseline: SH). The models also include Condition:Step and Condition:Trial as the interaction items, Subject as a random slope, and Frame as a random intercept.

A full list of the estimates in the two models are presented in Table B.1 and Table B.2 in Appendix B. For conditions including Male s-favoring training (i.e., the left facet in Fig. 3.9), a significant Condition difference is found between the Two genders - *opposite* condition and the Male only condition ($\beta = 1.54, p = 0.002$), as well as between the Two genders - *same* condition and the Male only condition ($\beta = -1.12, p = 0.03$). This means that compared to exposure to Male A’s /s/-favoring speech alone, a preceding training phase of

Female A’s /s/-favoring speech results in significantly fewer /ʃ/ responses (the blue line), while a preceding training phase of Female A’s /ʃ/-favoring speech results in significantly more /ʃ/ responses (the yellow line). For conditions including Male sh-favoring training (i.e., the right facet in Fig. 3.9), however, only the Two genders - *opposite* condition shows a significant difference from the Male only condition ($\beta = -2.16, p < 0.001$), while the Two genders - *same* condition does not ($\beta = 0.23, p = 0.23$). Overall, these results show the results of “Male A only” conditions differ from the results of two-gender training conditions that involve the same training condition with Male A, except for the difference between the Male A /ʃ/-favoring condition and the Female A /ʃ/-favoring Male A /ʃ/-favoring condition, which we will discuss more in the Discussion section.

3.3.4.4 Summary

Exp 1c evaluates the results of training with Male A’s /s/-favoring or /ʃ/-favoring speech only, and compares them with the two-gender training conditions that involve the same training condition with Male A. The results show a clear difference between two-gender training conditions and Male A only conditions, except for the difference between the Male A /ʃ/-favoring condition and the Two genders - /ʃ/-favoring condition. This might have been caused by a ceiling effect since Male A’s voice tends to induce more /ʃ/-equivalent responses overall. This point will be discussed further in the next section. In general, these results lend support to an account of *cumulative update*, where the speech of both Female A and Male A exerts an influence in the test phase. They are not consistent with the other possibility of *recency update*, which would otherwise predict little difference between the male-only conditions and their corresponding two-gender conditions.

3.4 Discussion

Experiment 1 aims at teasing apart the four possible mechanisms (*retention*, *reset*, *cumulative update*, and *recency update*) involved in the perceptual learning of multiple speakers. To achieve this goal, I separately manipulated the phonetic characteristics of the two talk-

ers in the training phase and evaluated the perceptual consequences caused by each step of manipulation with the same test continuum. This section provides a summary of the main findings of Exp 1 in response to this question, and discusses the implications of other empirical observations and remaining puzzles.

3.4.1 Cumulative update of perceptual expectations across speakers

One of the major findings of Exp 1 is that perceptual learning *updates* across speakers of different genders for fricatives, when the sibilants of different speakers are *similar enough* acoustically. This is reflected by the results of Exp 1a and 1b. Exp 1a compares the influences of three different training conditions on the categorization of a female speaker's sibilant continuum coming afterwards. The three training conditions each contain a training phase with the target female speaker's speech and a second training phase with a male speaker's speech. The critical design is that the conditions are contrasted by the acoustic distributions of the male speaker, such that they are intended to either favor /s/ or /ʃ/, or being neutral. The result of Exp 1a shows that the categorization boundaries vary consistently with the training input of Male A's speech in the second training phase: Participants who have received /s/-favoring training with Male A show the highest /s/ response rate among the three conditions; participants trained with /ʃ/-favoring Male A show the lowest /s/ response rate, and participants who have had no exposure to Male A's sibilants exhibit an /s/ response rate between the former two conditions.

This result is replicated by Exp 1b, which replaces Female A's /s/-favoring speech with her /ʃ/-favoring speech as the training materials of the first training phase. Again, the results show that exposure to a conflicting sibilant distribution from Male A successfully canceled out the earlier /ʃ/-favoring training with Female A, meaning that perceptual learning generalization still applies in the condition where the acoustic distributions of Female A's sibilants are manipulated in an opposite direction. Although the above trends are based on empirical observations rather than statistical analysis, they indeed show an amount of consistency between Male A's speech properties and the direction of boundary shift in each

condition indicates that the differences between these conditions are not simply a reflection of the between-group differences inherently associated with different groups of participants. Instead, it suggests that the acoustic distribution of the intervening speaker Male A is learned and applied to the final test phase when listeners perform the final categorization task with the voice of Female A.

When the statistical results are taken into consideration, the above picture is further complicated but not essentially changed. For each of Exp 1a and 1b, two kinds of statistical comparisons are conducted. One is the comparison between the results of the Female A baseline condition and those of the two-gender experimental conditions. The results consistently show that, in Exp 1a and 1b, exposure to Male A’s speech either with the *same* acoustic bias or with no /s f/ does not set back the existing perceptual shift induced by the first-phase training with Female A, whereas exposure to Male A’s speech biased towards the *opposite* direction would cancel out the previous perceptual shift such that the categorization result is not statistically different from the baseline. The other is the comparison between the results of the female-only single-phase training condition and those of the two-gender conditions. The results of this comparison are more mixed, and I will discuss them in more details in the next section (Section 3.4.2.0.1).

Among the hypotheses about how perceptual learning operates with multiple talkers, as laid out in Section 2.4, the current set of results lends support to the *update* account, which claims that listeners update their phonetic expectations in response to the recent acoustic input from different speakers. It is not consistent with other possibilities such as *reset*, which suggests that the boundary goes back to the baseline each time after encountering a different talker’s voice, or *retention*, which suggests that the perceptual learning functions in an absolute speaker-specific way. The question remains, though, is whether listeners set aside what they learned with Female A while integrating Male A’s acoustic distributions into their perceptual expectations. To what extent do the categorization patterns in the results of Fig. 3.6 and 3.8 maintain an influence of training with Female A’s speech?

Exp 1c is designed to evaluate this question by investigating the categorization boundary

as a result of exposure to Male A's training speech alone. Through a comparison between a specific male-only condition and the two-gender conditions that contain it, we have seen that their categorization results are not identical. This suggests that the results of the two-gender conditions cannot be attributed to training with Male A alone. In the Two genders - *same* condition, participants who heard only Male A's /s/-favoring stimuli gave fewer /s/ responses than participants who heard Female A's /s/-favoring stimuli prior to Male A /s/-favoring stimuli; similarly, participants who heard only Male A's /ʃ/-favoring stimuli gave fewer /ʃ/ responses than participants who heard Female A's /ʃ/-favoring stimuli prior to Male A /ʃ/-favoring stimuli. These patterns indicate that the results of the two-gender conditions reflect the integration of the acoustic distributions of both Female A and Male A. The same logic applies to the comparison between the Male only conditions and the Two genders - *opposite* conditions. For example, participants who heard only Male A's /s/-favoring stimuli gave more /s/ responses than participants who heard Female A's /ʃ/-favoring stimuli prior to Male A /s/-favoring stimuli. The above findings are represented by the result pattern in Fig. 3.10, where the results of the Male-only conditions lie between those of the Two genders - *opposite* condition and the Two genders - *same* condition that contain the same training stage with Male A.

Taken together, results of Exp 1 lend support to an account of *cumulative update* for perceptual learning, which predicts the integration of perceptual learning of both Female A's and Male A's acoustic distributions, instead of *recency update*, which predicts that listeners mainly integrate Male A's acoustic distributions and toss off Female A's to cope with the final categorization test.

3.4.2 Effect size, trial order, categorization slope, and transitional bias

After going through the main findings that shed light on the research question raised at the beginning of this chapter, I now turn to discuss some other findings of Exp 1 that are less relevant to the research question but are independently interesting. These factors include asymmetric effect sizes, trial order effect, categorization boundary slope, and transitional

bias of the test stimuli.

3.4.2.0.1 Asymmetric effect sizes between conditions Although the main patterns described above are statistically supported, one concern that comes up from the statistical results of the models is the occasional lack of a significant difference between the results of the two-gender conditions and the female-only condition. In Exp 1a and 1b, I compared the categorization results of the two-gender conditions with their corresponding female-only conditions by re-coding the “Female A training” condition as the baseline. This comparison is intended to evaluate how much change is caused by the second-stage training on top of the first-stage perceptual learning. The results, as reproduced in Fig. 3.11, shows an asymmetry regarding the effect size of training with Male A with different directions of biases: In Exp 1a, exposure to Male A’s neutral speech and speech favoring perceptual shift in the *same* direction did not induce an extra amount of shift compared to exposure to Female A’s training stimuli alone, whereas exposure to Male A’s speech favoring *opposite* perceptual learning has induced a significant setback. However, what we observe in Exp 1b is the reverse: Exposure to Male A’s neutral speech and speech favoring perceptual shift in the *opposite* direction did not induce a significant amount of additional shift compared to exposure to Female A’s training stimuli alone, whereas exposure to Male A’s speech favoring perceptual learning in the *same* direction with Female A’s speech has induced an additional shift towards /ʃ/ on top of the previous shift induced by Female A’s speech.

The lack of significant difference in the above two experimental conditions does not mean that the perceptual learning effect of the second phase in those conditions is absent. The second-phase training effect is captured by the model when “Female A baseline” is coded as the reference level. In Exp 1a, for example, although the model does not reveal a difference between the Two genders - *opposite* condition and the Female A /s/-favoring condition, a comparison with the baseline condition suggests that the result of the Two genders - *opposite* condition has been set back to the baseline and the significant shift towards /s/ as induced by training with Female A has been canceled out by exposure to Male A’s speech.

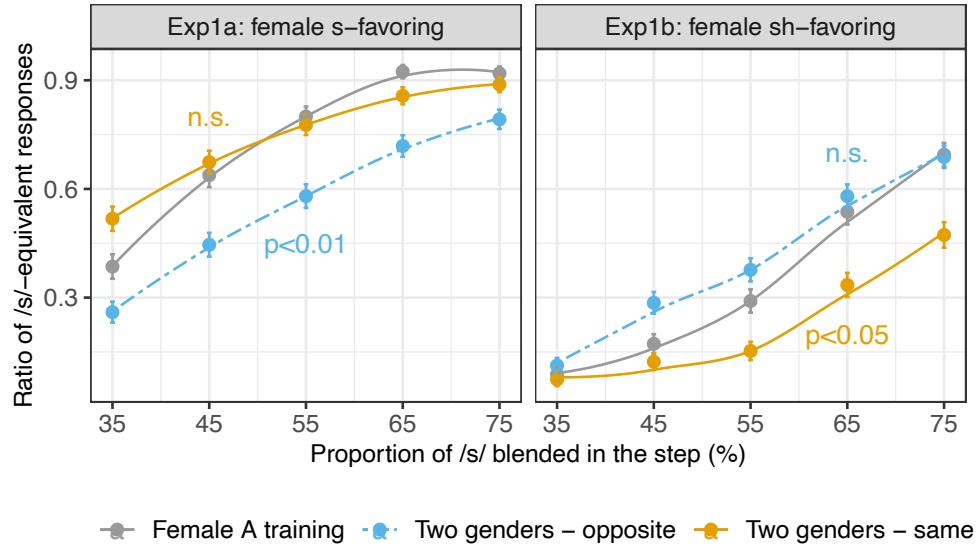


Figure 3.11: Asymmetric effect sizes between experimental conditions in Exp 1a and 1b

Instead, I propose that the asymmetric perceptual shift in different directions reflects the inherent acoustic biases of the training speech. Recall that the training stimuli of Female A result inducing a larger perceptual shift towards /s/ ($\beta = -2.15, p < 0.001$) than towards /j/ ($\beta = 1.07, p = 0.02$), despite the identical methodology adopted for the manipulation of the two sets of training stimuli (Fig. 3.5, Table 3.4). The situation of the Male A training speech is backward: it induces a larger shift towards the /j/ end ($\beta = 1.62, p < 0.001$) than the /s/ end ($\beta = -0.92, p = 0.07$) (Fig. 3.9, Table 3.7). These properties may explain why the female-only condition does not significantly differ from the two-gender (*same* condition in Exp 1a and the two-gender *opposite* condition in Exp 1b – because the latter two conditions both contain training with Male A’s /s/-favoring speech in their second phase, and /s/-favoring training with Male A does not induce robust perceptual shifts by itself. Moreover, the fact that Female A’s training speech induces a larger perceptual shift towards /s/ than /j/ makes the Two genders - *same* condition in Exp 1a more vulnerable to a ceiling effect than in Exp 1b, which contributes to the insignificance of the difference between the *same* condition and the Female A s-favoring condition in Exp 1a.

3.4.2.0.2 Trial order Another factor that consistently turns out to be significant in the model statistics is trial order, namely, the order in which test stimuli are presented. In the experiment, the stimuli are randomized within block for each participant, and as a result, each test stimulus is assigned number from 1 to 51, with larger numbers standing for occurrence at a later time point. Intriguingly, I find that the Trial effect in Exp 1 usually reflects listeners’ tendency to balance the numbers of their /s/ and /ʃ/ responses through the time span of the test phase. This is regardless of the fact that there are only 35 critical trials in the test phase. For example, /s/-favoring training usually comes with a positive coefficient of Trial (for SH), because of listeners’ tendency to remedy for their bias towards choosing /s/ at the beginning of the test phase. Similarly, /ʃ/-favoring training usually comes with a negative coefficient of Trial (for SH), because of listeners’ tendency to remedy for their bias towards choosing /ʃ/ at the beginning of the test phase. This pattern can be observed from the estimates of Condition: Trial interaction in Table 3.4, 3.5, and 3.6). This seems to reflect a degree of distributional learning of the listeners from the acoustic distributions of the continuum. In other words, these task-specific “balancing” effects actually work against the perceptual learning effects of interest, making them appear smaller. Similar behaviors have also been reported in other studies. In specific, listeners tend to balance the identification choices distributed along the provided continuum in categorization tests. Tamminga et al. (2020) attribute this phenomenon to either a range effect (Brady and Darwin, 1978; Keating et al., 1981; Rosen, 1979) that involves interpreting the continuum endpoints as phonemic anchors, or a frequency effect that reflects a bias toward hearing each option an equal number of times in a two-alternative forced choice task.

3.4.2.0.3 Categorization slope In phoneme categorization or perceptual learning studies, the perceptual shift is usually quantified by the aggregate change in the identification rate of a certain phoneme. Although changes in the categorization slope along the acoustic continuum are not uncommon, the implication of this factor is sometimes neglected. Statistical models in this chapter evaluated how the slopes of categorization boundaries vary

between conditions, by including the interaction between Condition and Step. A robust pattern consistently reflected by models in Exp 1a-1c is that exposure to Male A’s training stimuli usually results in a shallower slope of the categorization boundary, in addition to the increase or decrease in the mean “s” or “sh” response rates (see Table 3.5, 3.6, and 3.7)). In contrast, exposure to Female A’s training speech does not change the slope of the categorization boundary (see Table 3.4). The consistent behavior of boundary slope as a response to different speakers’ speech across speakers might indicate that the categorization slope reflects nuanced dimensions inherent to the acoustic distribution of Male A’s speech. This idea is also embodied in some of the previous studies on the modeling of perceptual learning. For example, Kleinschmidt and Jaeger (2015) derived the slope of the categorization boundary from the variance of the underlying acoustic distribution of related phonemes in his model. Further investigations on the behavior of categorization slope as a response to training speech with different distributions and speakers is a promising direction for this type of research.

3.4.2.0.4 Transitional bias The lexical frames used for test stimuli are manipulated from /s/-containing and /ʃ/-containing minimal-pair words. To account for any influence of the transitional cues associated with the original phoneme produced with the minimal pairs, I have included Phoneme as a fixed effect in the model. It turns out that Phoneme has a significant effect on the categorization results in all of the models in this chapter. Fig. 3.12 demonstrates the /s/ response rates of Exp 1a and 1b by Phoneme in the original productions of minimal pairs. By comparing the left and right facets of each row, we can see that /s/-transitional frames (left) consistently end up with more /s/ responses than /ʃ/-transitional ones (right). In other words, although the sibilants of the minimal pairs have been sliced out and replaced with ambiguous ones along the continuum, the transitional cues remaining in the lexical contexts still have a significant influence on the categorization of the embedded phoneme.

This design leads to an unusual curvy shape of the categorization boundary in some of the two-gender conditions. This is because the /s/- and /ʃ/- transitional lexical frames are

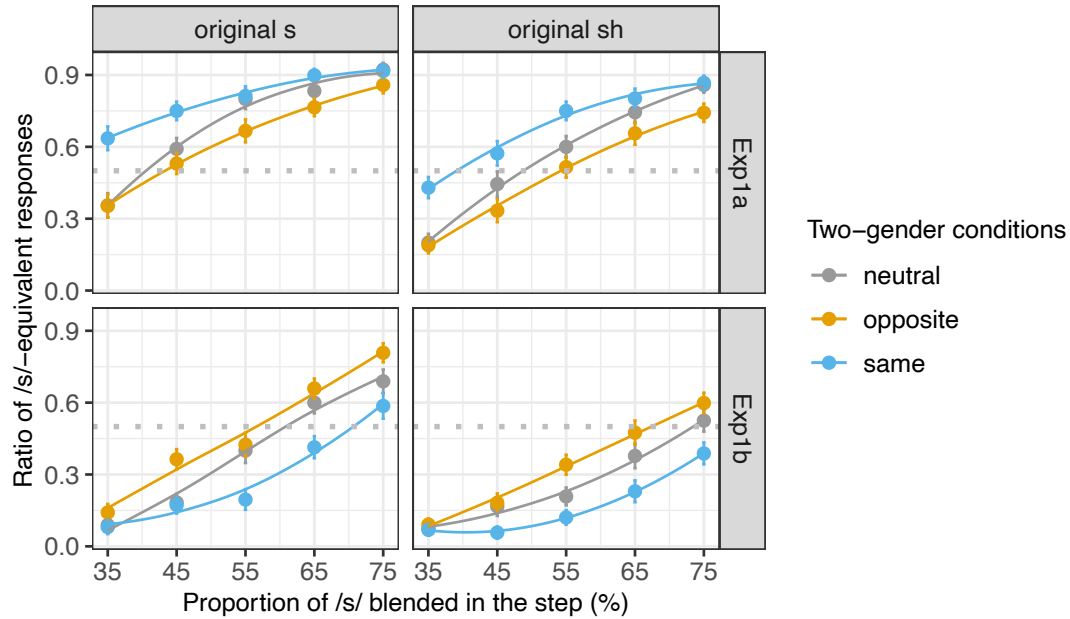


Figure 3.12: /s/ response rate in two-gender training conditions by phoneme and experiment (mean and standard error)

organized in an interleaving way along the steps of the continuum (see Table 3.3): In the seven repetitions of the continuum, /s/-transitional frames have occurred three times at Step 35, 55 and 75 and four times at Step 45 and 65, whereas /ʃ/-transitional frames have occurred four times at Step 45 and 65 and three times at Step 35, 55 and 75. As a result, the categorization boundaries in some of the two-gender conditions are curved toward the floor at Step 55 and curved towards the ceiling at Step 45 and 65 (e.g., in Fig. 3.8).

In order to obtain smoother categorization results in the next chapter, I have changed the adoption of interleaving transitional lexical frames in the test phase and used /ʃ/-containing productions for test trials at all steps.

3.4.3 Remaining questions

The final issue related to different directions of perceptual learning is the design of opposite manipulations of Female A's speech in the first learning phase between Exp 1a and 1b. This is intended to evaluate whether the update of perceptual expectations across speakers is conditioned on a degree of acoustic overlap between sibilants in different experimental phases.

In the beginning of Section 3.3.3, we have seen that, compared to a female /s/-favoring distribution, a female /ʃ/-favoring distribution is farther apart from a male speaker's sibilant distributions in the acoustic space. This is because /s/-favoring manipulation keeps /ʃ/ sounds unaltered while reducing the frequencies of /s/ sounds, leading to lower sibilant frequencies overall and making them closer to the male speakers' sibilants. In contrast, /ʃ/-favoring manipulation keeps /s/ sounds unchanged while increasing the frequency of /ʃ/ sounds, making the overall frequency distribution higher and thus farther from a male speaker's sibilants.

A constraint of acoustic overlap on perceptual learning is claimed in previous perceptual learning literature (e.g., Kraljic and Samuel, 2006), although they mainly discussed the acoustic overlap between the training stimuli and the test stimuli. In the current experiment, the lack of acoustic overlap occurs between training phases in the Two genders - *different* condition in Exp 1b (Fig. 3.4). This lack of acoustic overlap between training phases does not seem to make a difference to the result of perceptual learning. Exposure to a conflicting sibilant distribution of Male A successfully canceled out the earlier /ʃ/-favoring training with Female A, meaning that perceptual learning generalization still applies in the condition where the acoustic distributions for sibilants are the most far apart.¹

The current results does not show evidence for constraints of acoustic overlap between training phases on perceptual learning across speakers. However the result does not directly contradict the findings of an acoustic overlap effect reported in Kraljic and Samuel (2006) at this point, because there is no lack of acoustic overlap between the training and test stimuli (Fig. 3.3) in all sub-experiments. It remains unclear whether the acoustic overlap between the training stimuli and the test stimuli would impose such a constrain in this set of experiments. The question of the acoustic constraints on perceptual learning will be examined in more details in the next chapter.

¹One may argue that the result of Exp 1b where the Two-gender *opposite* condition is not significantly different from the female-only condition in Exp 1b is a consequence of the lack of acoustic overlap. However, as I discussed in the earlier part of this section, this cannot be the full story since we see that /s/-favoring training with Male A does not induce a robust perceptual shift in other sub-experiments (Exp 1a, 1c).

Chapter 4

Exp 2: Acoustic Interference in the Perceptual Learning of /s-ʃ/ across Speaker Genders

This chapter reports on Experiment 2, which investigates whether there is an interaction between the effects of speaker gender and the acoustic properties of /s ʃ/ in cross-speaker perceptual learning. Exp 2 shares similar experimental structures with Exp 1 but replaces the previous female speaker, Female A, with a different one, Female B. Compared to Female A, Female B has higher frequencies of energy distribution for sibilants, which causes more dramatic acoustic asymmetry between the sibilants of the female speaker and the male speaker in training and test. This chapter contains four sections. Section 4.1 reviews previous findings of the effect of acoustic properties on perceptual learning and explains how replacing Female A with Female B provides a suitable testing ground for the investigation of this question. Section 4.2 provides an overview of the experimental design, stimuli, and conditions of Exp 2. Section 4.3 reports on a pilot and a series of three sub-experiments and their main results. Finally, Section 4.4 discusses the implications of the main findings in Exp 2 and concludes this chapter. The result of this experiment is expected to shed light on whether and how the acoustic properties of sibilants constrain perceptual learning across speakers of different genders.

4.1 Background and research question

The previous chapter (Section 3.1) provides a review of the acoustic variation of /s-f/ with speaker gender and the influence of this covariation on perception. In this section, I further demonstrate that the gender variation in the natural production of /s f/ gives rise to predictable patterns of acoustic properties of the training and test stimuli in perceptual learning experiments. This is determined by the typical stimulus manipulation method adopted by perceptual learning studies. Usually, clear sibilants in the stimuli are copied from the original /s/ and /f/ pronunciations of the training speaker, and ambiguous sibilants are generated by blending the spectra of clear speech production of different phonemes. As a result, the acoustic distributions of critical phonemes in the training speech are correlated with the acoustic properties of their original production of those phonemes. In the meantime, as sibilants also exhibit a robust gender variation in the naturalistic production, complicated relationships between the acoustic distributions of the training and test stimuli are generated in different conditions.

The remainder of this section delineates two potential ways in which the acoustic properties of experimental stimuli may constrain cross-speaker perceptual learning, and summarizes the predictions of the experiment result under each hypothesized constraint.

4.1.1 Acoustic overlap between the training and the test stimuli

Empirical findings have suggested that the acoustic alignment or overlap between different speakers' critical phonemes makes a difference to the generalization of perceptual learning. Eisner and McQueen (2005) raised the possibility that the relationship between acoustic distributions might be an even more reliable predictor of whether perceptual learning generalizes than speaker gender. Eisner and McQueen showed that listeners would not generalize what they had learned from a female training speaker to a male test speaker unless the female training speaker's fricatives are replaced with the male test speaker's fricatives, which are then spliced into the female training speaker's word frames. They interpret this result

as evidence that listeners attend to the acoustic properties of the phonemes of different speakers and only generalize what they learned from previous instances to upcoming ones when the acoustic properties of these instances are similar enough.

This result is echoed by Kraljic and Samuel (2005) if we consider the overlap between two acoustic distributions as a specific aspect of acoustic similarity. Kraljic and Samuel (2005) found that the outcome of training with a female speaker can be transferred to a male voice in the test phase; however, training with a male speaker does not generalize to a female test speaker. They suggested that the decisive factor in cross-speaker generalization for fricatives is the acoustic similarity between the fricatives heard during exposure and those categorized at test. In their case, acoustic measurements of the fricatives revealed that the female speaker's fricatives during exposure fell within the range of the male speaker's test continuum, whereas the male speaker's fricatives during exposure are acoustically distinct from the female speaker's test continuum. Listeners thus seem to track the acoustic properties of each speaker's fricative productions and apply generalization whenever there is a sufficient match. They put it this way: "perhaps unintended acoustic differences between our Female and Male training and test items led to different training-test mappings for these voices... Female training transfers to the Male voice presumably because the Female training stimuli are spectrally relatively close to the Male testing stimuli, but the Male training does not transfer to the Female voice because Male training and test items are virtually identical in average spectral mean, and relatively distant from the Female test stimuli." (2005, pp. 166).

Reinisch et al. (2014) investigated the cross-talker generalization of sibilant perceptual learning with foreign-accented speech. They found that the retuning of fricative perception is not speaker-specific, but generalization depends on how two speakers' test continua are sampled across perceptual space. In specific, Reinisch et al. (2014) evaluated whether the /s-f/ boundary learned from a Dutch-accented female speaker can be generalized to a different female speaker and a male speaker with the same Dutch-accented speech. They found generalization across the female speakers, although the listeners reported some confusion

about speaker identity. As with the male speaker, they found that generalization depended on sampling of his fricatives to match or mismatch the perceptual space of the female exposure speaker's fricatives. It remains unclear whether this result would be expected to generalize when not using L2-accented speech. In a nutshell, the above studies suggest that the generalization of perceptual learning does require a certain amount of acoustic overlapping between the sibilants of the training speaker and the test speaker.

As the first research of this chapter, I investigate whether the influence of acoustic overlap between the training and the test stimuli on multi-speaker perceptual learning can be observed with lexically-guided paradigm. Exp 1 does not provide a testing ground for us to evaluate how perceptual learning across speakers is affected by the lack of acoustic overlap between the training and the test stimuli, because the test continuum overlaps more or less with sibilants in each experimental condition in the acoustic space. The female speech adopted in this chapter has higher frequencies of energy distribution for sibilants, which leads to a lack of acoustic overlap between the male /s/-favoring speech and the female test speech (to be explained with Fig. 4.4). If there turns out to be a constraint of acoustic similarity between the training and test phase, then we should expect that Male A's /s/-favoring speech induces no perceptual shift on Female B's continuum, regardless of the fact that it has induced a significant amount of shift on Female A's continuum in the previous chapter.

4.1.2 Acoustic overlap and mismatch between training phases

Differing from the design of Kraljic and Samuel (2005) and Reinisch and Holt (2014) described in the previous subsection, other perceptual learning studies reported in previous literature and in this dissertation involve training with multiple speakers instead of a single speaker. Therefore, in addition to the relationship between the acoustic distributions of the training phase and the test phase, relevant issues may also exist between different training phases. At the beginning of Exp 1b in the previous chapter, I have demonstrated that the relationship between the acoustic distributions of different training phases varies depending

on the specific directions of the imposed perceptual biases. Recall that different amount of acoustic overlap of Female A's and Male A's speech is derived in the Two-gender *opposite* conditions between Exp 1a and 1b (see Fig. 3.7). The two speaker's acoustic distributions are closer to one another when the materials consist of Female A's /s/-favoring speech and Male A's /j/-favoring, and they become farther apart when the materials consist of Female A's /j/-favoring speech and Male A's /s/-favoring speech.

This situation is not limited to experiments in this dissertation but also applies to similar studies on sibilant perceptual learning with manipulations of perceptual biases towards different directions. To provide a more general schema for similar issues in experiments with such a design, I demonstrate how different combinations of distributional parameters may cause different relationships between acoustic distributions through simulation, which is presented in Fig. 4.1. The left sub-figure demonstrates the simulated distribution of the center of gravity (COG) of a female speaker's clear /s/ (solid), clear /j/ (dashed), and ambiguous (dotted) sibilants between these two phonemes, as well as the COG of a male speaker's speech of the three kinds. Overall, the female speaker's sibilants have higher COG than the male speaker's, consistent with the real-world gender variation in spectral frequency. The two sub-figures on the right demonstrate what the acoustic distribution looks like under different training conditions depending on the specific perceptual biases associated with each of the two speaker's speech: The training condition with /s/-favoring male speech and /j/-favoring female is shown in the top of the right figure. It consists of the female speaker's clear /s/ and ambiguous /j/ as well as the male speaker's clear /j/ and ambiguous /s/. The training condition with male /j/-favoring male speech and female /s/-favoring speech is shown in the bottom of the right. It consist of the female speaker's clear /j/ and ambiguous /s/ and the male speaker's clear /s/ and ambiguous /j/.

Although the above two designs in Fig. 4.1 both represent a training condition where the speech of speakers of different genders induces perceptual biases towards different directions, the specific decisions about the directions of perceptual biases still make a difference to the relationship between the acoustic distributions of different speakers. Two issues emerge

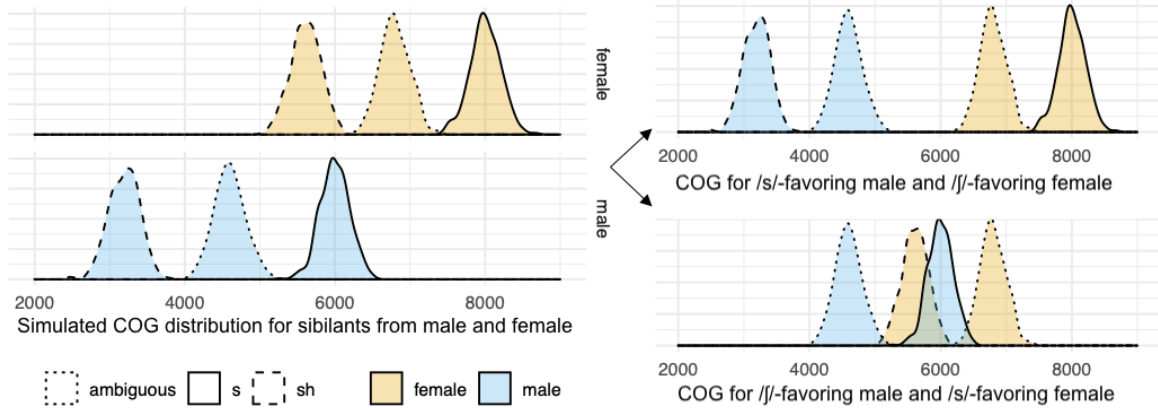


Figure 4.1: A schema of acoustic overlapping between fricatives under experimental conditions of different perceptual biases

from Fig. 4.1 regarding the potential role of acoustic properties in constraining multispeaker perceptual learning.

The first issue lies in the acoustic *dissimilarity* between the speech of the two speakers in training. According to Fig. 4.1, the resulting acoustic distributions of different speakers are distinct from one another in the experimental condition on top, whereas in the bottom, the COG of the two speakers' sibilants overlaps substantially with each other. This asymmetry raises the issue whether the lack of acoustic overlap blocks the perceptual learning effect across speakers. To the best of my knowledge, experiments reporting on acoustic constraints of cross-speaker perceptual learning are primarily concerned about the degree of acoustic overlap between the training stimuli and the test stimuli. Few studies have looked into the relationship between the acoustic distributions of different training phases and asked how it affects the perceptual learning results. Although the result of Exp 1 does not lend support to this possibility, I still want to check whether the finding of perceptual generalization can be replicated when the female speaker has a higher frequency range. If so, we should see that perceptual learning fails to work for the condition in the right top facet, where participants received /j/-favoring training with Female B and /s/-favoring training with Male A:

The second issue is a potential *mismatch* between the acoustic place and the phonologi-

cal space between experimental conditions. This situation applies to the condition described in the right bottom facet in Fig. 4.1. The manipulation of the experimental stimuli is intended to shift the categorization boundary between two phonemes by anchoring phoneme representation with acoustic distribution of a less typical range. In specific, a /s/-favoring manipulation is achieved by lowering down the general frequencies of /s ʃ/ in the perceptual expectation and anchoring the ambiguous sibilants with /s/. Similarly, a /ʃ/-favoring distribution raises the sibilant frequencies in the perceptual expectation and anchors the ambiguous sibilants with /ʃ/. However, in the situation of the right bottom facet, an ambiguous sound (which corresponds to the COG of around 6000 Hz) acoustically equals to /s/ in the training condition that favors /ʃ/ perception (the blue distributions) and equals to /ʃ/ in the condition that favors /s/ perception (the yellow distributions). It is unclear how the mismatch between the two speakers' raw acoustic distributions and their distributions in the phonological space (namely, distribution after normalization by speaker) blocks perceptual learning. If so, we should see that perceptual learning fails to work for the condition in the right bottom facet, where participants received /s/-favoring training with Female B and /ʃ/-favoring training with Male A.

The second research question of Exp 2 is how the multitalker perceptual learning of sibilants is constrained on the relationship between the acoustic distributions of different speakers in the training phase. Two constraints of acoustic distributions have been formed, which I refer to as *acoustic dissimilarity* and *acoustics-phonology mismatch*. They make different predictions about the results of Exp 2. The predictions of these two hypothesized constraints are summarized as Table 4.1.

	<i>Acoustic Dissimilarity</i>	<i>Acoustics-phonology Mismatch</i>
Female /ʃ/-favoring & Male /s/-favoring	Yes (no overlap)	No (no mismatch)
Female /s/-favoring & Male /ʃ/-favoring	No (acoustic overlap)	Yes (/s/-favoring sounds have higher frequencies than /ʃ/-favoring ones)

Table 4.1: Summary of how experimental conditions in Exp 2 fit into the two hypothesized acoustic constraints

4.2 Method overview

4.2.1 Experimental conditions

Exp 2 contains a pilot study and three sub-experiments. The pilot study reports the /s-ʃ/ categorization results of three conditions – a baseline condition where participants have not receive any prior training before the test, and two training conditions where participants have received either a /s/-favoring or a /ʃ/-favoring training phase with Female B’s speech, before they finally complete an identical categorization test on a /s-ʃ/ continuum of Female B’s speech spliced into minimal pairs. Like in Exp 1, the pilot study is intended to show that the categorization boundary is aligned with the center of the continuum, and that the design successfully induces the expected perceptual shift in both directions with Female B’s speech.

Exp 2a and 2b each contain three experimental conditions. The two experiments begin with a training phase of Female B’s speech. The training speech is manipulated to be /s/-favoring in Exp 2a and /ʃ/-favoring in Exp 2b, which corresponds to the two female training conditions in Fig. 4.1. Then listeners proceed to the second training phase with Male A’s speech, which is manipulated to be /s/-favoring, /ʃ/-favoring, or containing no /s ʃ/ depending on the specific condition listeners are assigned to. Finally, the learning outcome is evaluated by a categorization test phase with Female B’s speech. By comparing the results of the three conditions within each sub-experiment, we are able to know to what extent the exposure to Male A’s speech matters for the categorization of Female B’s speech. By comparing the results of Exp 2a and 2b, we are able to detect any effects associated with acoustic overlapping situations of stimuli in different training phases.

Fig. 4.2 shows a summary of the experimental designs and procedures in each condition in Exp 2a and 2b. In a nutshell, the pilot study differs from Exp 2a and 2b in the number of training phases involved. Exp 2a and 2b differ in the acoustic condition of the first training phase with Female B.

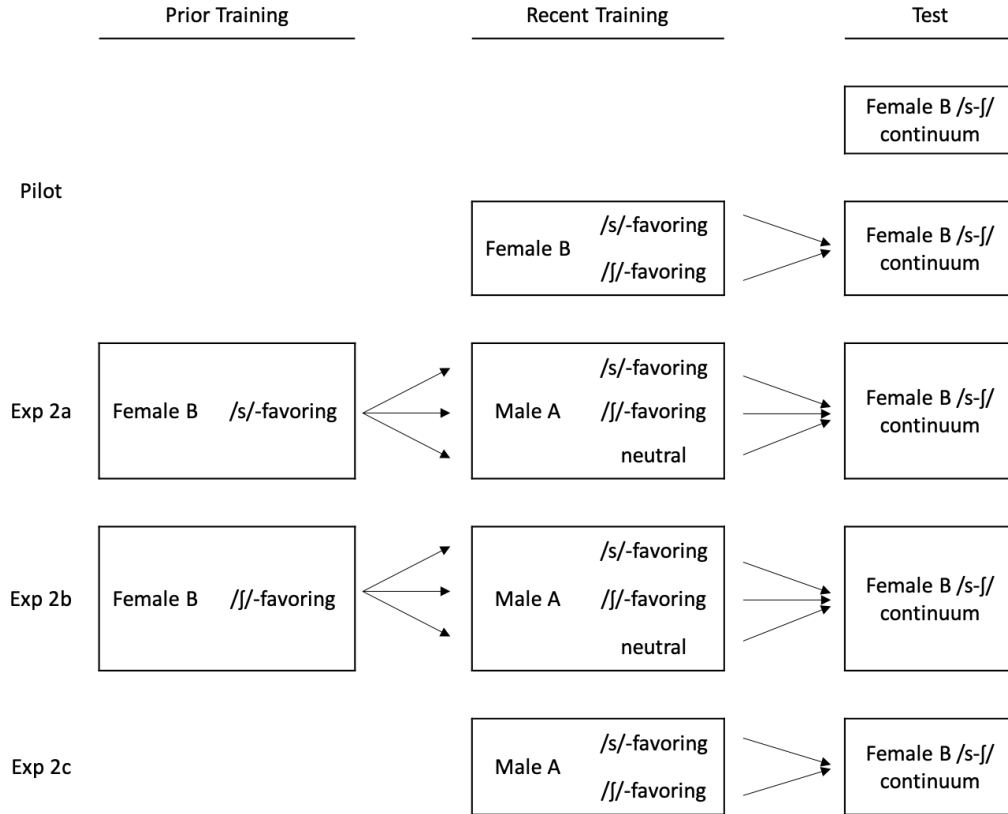


Figure 4.2: The structure of experiments and conditions in Exp 2

4.2.2 Word list and recording

Stimuli used in Exp 2 are manipulated from recordings of spoken words from Female B and Male A obtained according to the procedure described in Section 2.3.2. For each speaker, the spoken words used for training contain 17 /s/-containing words and 17 /j/-containing words. The word list used in this experiment is identical to that of Exp 1 (provided in Section 3.2.2). For each word, the critical sibilant is annotated by hand in the TextGrid layer in Praat and measured for their the center of gravity (COG) value. Fig. 4.3 compares the raw means and the 95% confidence intervals of the 34 sibilants for Female B, Male A and Female A. We can see that Female B has a mean COG of around 10000 Hz for /s/, which is considerably higher than the mean COG of /s/ of Female A (8500 Hz) and Male A (7000 Hz). The COG of Female B's /s/ also has a narrower frequency range, reflected by the smaller 95% confidence interval. The /j/ sound of Female B does not seem to differ

much from Female A in terms of mean COG (above 4000 Hz), both of which is higher than the mean COG of Male A's /f/ (below 4000 Hz).

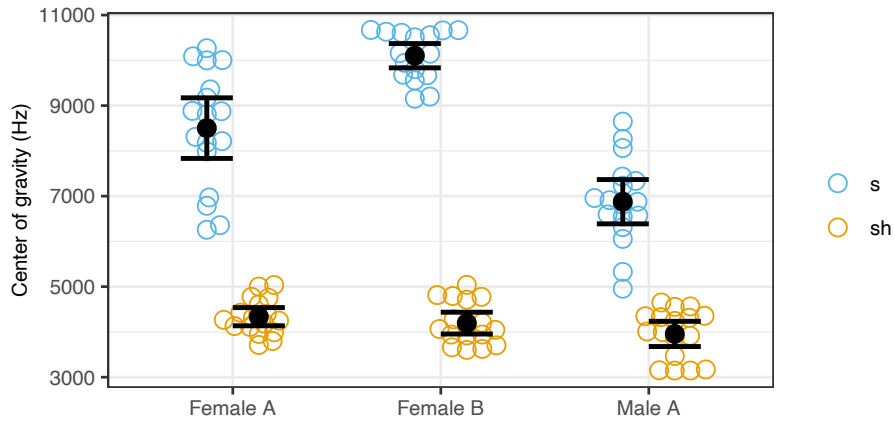


Figure 4.3: The COG of sibilants of Female A, Female B and Male A in natural speech production (mean and 95% CI)

4.2.3 Step selection and synthesis

Again, for the synthesis of the training stimuli, the critical proportion of sibilants sharing the same word frame (e.g., compensate and compensshate) are cut out and mixed with each other by five steps of proportions. The five steps of sibilants for each word frame by each speaker vary from 0.3[s]0.7[ʃ] to 0.7[s]0.3[ʃ] with an increase of 0.1[s] and a decrease of 0.1[ʃ] by each interval, and then they are spliced back to the lexical frame. All the synthesized stimuli are normalized to 70 dB. A lexical decision task is conducted to select the most ambiguous step of sibilant for each word frame to be used in the training phase. Participants needed to judge for lexical frames spliced with each of the five sibilant steps, whether they are an English word or not. The results of lexical decision is shown in Section A. The mixture proportion that provides the most ambiguous (50%) categorization result is selected to be further used to construct the training materials in perceptual learning. If two steps bear approximately similar distance to the 50%, then their average is used for stimulus construction. Table 4.2 showed the ambiguous steps chosen for each critical lexical frame for Female B.

Perception experiments are also conducted to select the steps and word frames of min-

ID	/ʃ/-words	Step	/s/-words	Step
1	ambition	0.35	compensate	0.4
2	beneficial	0.55	democracy	0.4
3	brochure	0.25	dinosaur	0.4
4	commercial	0.3	embassy	0.4
5	negotiate	0.35	episode	0.3
6	crucial	0.35	eraser	0.65
7	official	0.4	falsetto	0.4
8	parachute	0.35	faucet	0.5
9	efficient	0.4	legacy	0.25
10	impatient	0.3	medicine	0.4
11	initial	0.35	obscene	0.4
12	vacation	0.35	personal	0.35
13	evaluation	0.3	parasite	0.5
14	publisher	0.35	peninsula	0.55
15	refreshing	0.25	pregnancy	0.4
16	glacier	0.25	rehearsal	0.3
17	graduation	0.25	reconcile	0.3

Table 4.2: The proportion of [s] mixed in the most ambiguous step of sibilant chosen for each word frame for Female B

imal pairs to be used for stimuli in the test phase. The test trials are generated by splicing 5 steps on a /s-ʃ/ continuum into 7 word frames of minimal pairs. Different from Exp 1, However, the categorization test with Female B’s speech does not involve interleaving /s/ and /ʃ/ lexical frames across the continuum, as shown in Table 3.3. Instead, all of the frames are /ʃ/-containing (*shake, shame, sheet, shelf, shell, shine, shy*). Accordingly, the statistical models evaluated for Exp 2 will not have Phoneme as a fixed effect.

4.2.4 The acoustics of synthesized stimuli

Fig. 4.3 compares the means and 95% confidence intervals of /s ʃ/ for each speaker in each experimental condition between Exp 1 and 2. Recall that the two experiments differ in the speech of the female speaker used. Exp 2 uses the speech of Female B, who has higher frequencies of energy distribution for sibilants than Female A and Male A. The two experiments each have two sub-experiments that differ in the sibilant frequencies of the female speaker in training: female speakers exhibit /s/-favoring distributions in Exp 1a and

2a while exhibit /f/-favoring distributions in Exp 1b and 2b. Subjects within each sub-experiment received the same training with the female speaker but different training with the male speaker. Depending on the condition they are assigned to, they may be exposed to either /s/-favoring or /f/-favoring speech of Male A.

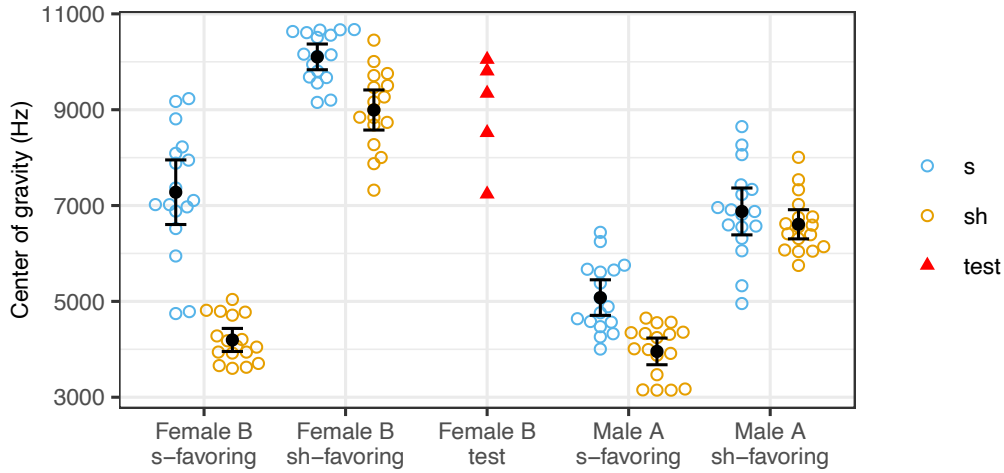


Figure 4.4: The COG of sibilants in different training phases and in the test phase in Exp 2 (mean and 95% CI)

4.3 Experiment and result

4.3.1 Pilot study: Learning Female B's /s-f/

4.3.1.1 Experimental conditions and goals

Like the pilot study of Female A, the pilot study of Female B also contains three conditions – baseline, /s/-favoring learning, and /f/-favoring learning. Participants in the baseline conditions completed a single test block, containing 35 test trials with ambiguous sibilants embedded in minimal pairs and 17 filler words without sibilants in Female B's voice. The result of this condition is taken as a reference of the default /s-f/ perceptual boundary for Female B. Participants in the two learning conditions first completed either an /s/-favoring training block or an /f/-favoring training block with Female B's speech before they

proceeded to complete the same test block as in the baseline condition. The goal of the pilot study is twofold. The first goal is to demonstrate that the /s/-favoring and /j/-favoring perceptual learning effects have been successfully elicited with the speech of Female B. The second goal is to demonstrate that the 50% perceptual boundary between /s-f/ has been successfully aligned with the center of the continuum by default without prior training.

4.3.1.2 Participant

Participants in the three conditions of the pilot with Female B are all recruited from the UPenn undergrad subject pool. 31 participants are recruited to attend the baseline condition, including 23 female and 8 male, aged from 18-22 years old ($Mean = 19.7, SD = 1.1$). 32 participants are recruited for the Female B /j/-favoring condition, including 24 female and 8 male, aged from 18-22 years old ($Mean = 19.8, SD = 1.2$). 30 participants are recruited for the female B /s/-favoring condition, including 14 female and 16 male, aged from 18-22 years old ($Mean = 20.2, SD = 1.2$).

4.3.1.3 Result

Fig. 4.5 shows the results of phoneme categorization by participants in the baseline conditions (in grey), Female B /s/-favoring condition (in yellow), and Female /j/-favoring condition (blue). We can see that firstly, the 50% point of the categorization boundary in the baseline condition aligns with the middle step of the continuum (Step 55). Secondly, the /s/-favoring training and the /j/-favoring training seem to have worked in inducing a perceptual bias towards the expected direction compared to the baseline condition: Participants in the /s/-favoring condition show more /s/-equivalent responses on Step 35-55 than those in the baseline condition. Similarly, participants in the /j/-favoring condition show fewer /s/-equivalent responses on Step 45-75 than those in the baseline condition.

We then ran a logistic mixed-effects regression model (Model-pilot2) to evaluate whether the two learning effects we see in Fig. 4.5 are statistically significant. A mixed-effects model is conducted to predict the Response of each trial ($S=0, SH=1$), with Step (35-75, scaled

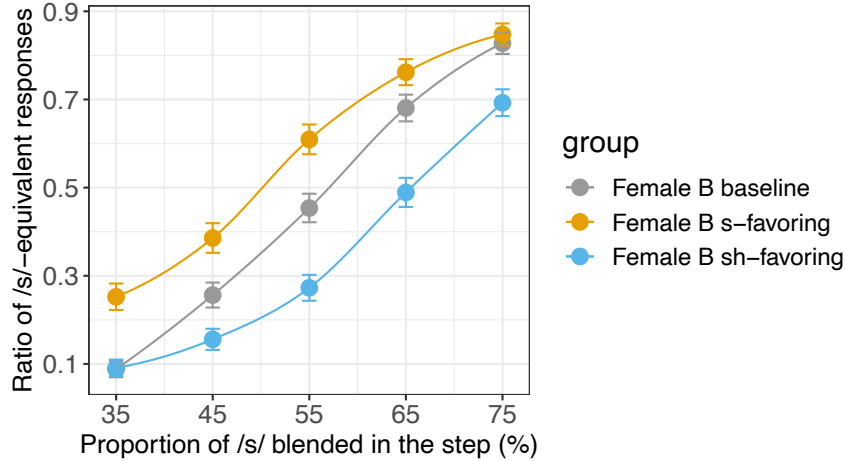


Figure 4.5: Exp 2 pilot: Boundary shift after exposure to Female B’s /s/-favoring and /ʃ/-favoring speech compared to the categorization baseline (mean and standard error)

and centered), Trial (1-51, scaled and centered), and Condition(treatment coded, reference: baseline) as the fixed effects, Condition:Step and Condition:Trial as the interaction item, Step by Subject and Step by Frame as random slopes. The result is shown in Table 4.3.

Fixed effects	Estimate	SE	z value	Pr(> z)
(Intercept)	0.31	0.42	0.74	0.46
Step	-2.25	0.17	-13.14	< 0.001***
Condition Female B s-favoring	-0.93	0.51	-1.81	0.07
Condition Female B sh-favoring	1.05	0.51	2.06	0.04*
Trial	0.25	0.09	2.77	0.006**
Step:Condition Female B s-favoring	0.62	0.23	2.73	0.006**
Step:Condition Female B sh-favoring	0.35	0.23	1.53	0.13
Condition Female B s-favoring: Trial	-0.03	0.13	-0.22	0.83
Condition Female B sh-favoring: Trial	-0.21	0.13	-1.66	0.10

Model-pilot2: Response~Step*Condition+Condition*Trial+(Step|Subj)+(Step|Frame)

Table 4.3: The fixed effects of the logistic mixed-effects model in Exp 2 pilot

The model shows a main effect of Step ($\beta = -2.25, p < 0.001$), suggesting that in the baseline condition, the larger the proportion of [s] is mixed in the stimulus, the less likely that stimulus is perceived as /ʃ/-equivalent. This trend also holds for the sh-favoring condition, as reflected by the lack of significant interaction between Step and Condition ($\beta = 0.35, p = 0.13$). The slope of the categorization boundaries along the continuum step becomes significantly shallower for the Female s-favoring condition, as indexed

by the interaction item ($\beta = 0.62, p = 0.006$). The effect of Condition is significant for the sh-favoring condition ($\beta = 1.05, p = 0.04$) and marginally significant for the s-favoring condition ($\beta = -0.93, p = 0.07$), suggesting that participants with /s/-favoring training experience are less likely to show /j/-equivalent responses than the baseline condition while those /j/-training experience are more likely to show /j/-equivalent responses than the baseline condition. Lastly, the interaction between Condition and Trial is not significant for both of the conditions (s-favoring: $\beta = -0.03$, sh-favoring: $\beta = -0.21$).

While it is less ideal that the difference between the baseline condition and the s-favoring condition is only marginally significant, this result verified that they have induced a perceptual bias towards the expected direction.

4.3.2 Exp 2a: Previous /s/-favoring training with Female B

4.3.2.1 Experimental conditions and goals

The goal of Exp 2a is to verify that what we found in Exp 1a, namely, the update of perceptual expectations for fricatives across speakers of different genders, also applies to a different set of speakers. Three experimental conditions are formed with the manipulated speech of Female B and Male A as the training stimuli and the fricative continuum of Female B spliced into minimal pairs as the test stimuli. Participants in all three conditions first complete /s/-favoring training phase with Female B, followed by a training phase with Male A's speech with different sibilant manipulations. Depending on the condition. Male A's speech is either manipulated to be /s/-favoring (*same*), /j/-favoring (*opposite*), or sibilant-free (*neutral*). In the end, participants are tested on Female B's sibilants on an /s-j/continuum spliced into minimal pairs.

If the result replicates what we see in Exp 1a, then it predicts that listeners integrate the acoustic distributions of Female A to update their perceptual expectations and apply this knowledge to the final categorization phase with Female B. As a result, we should see, among the three conditions, the most /s/-equivalent responses in the same condition and the most /j/-equivalent responses in the opposite condition, with the results of the neutral

condition lying in between.

4.3.2.2 Participant

Participants in the three conditions of Exp 2a are all recruited from Prolific, with 29 of them in the opposite condition, 29 in the same condition, and 23 in the neutral condition. Participants in the opposite condition are 17 male and 12 female, aged from 17 to 21 years old ($Mean = 19.7, SD = 1.4$). Participants in the same condition include 11 female and 18 male, aged from 18 to 29 years old ($Mean = 20.5, SD = 2.1$). Participants in the *neutral* condition are 13 female and 10 male, aged from 18 to 22 years old ($Mean = 20.2, SD = 1.1$).

Along with the data of the above participants, I have also plotted the data of participants in the baseline condition and the Female B /s/-favoring condition as a reference (see Section 4.3.1.2 for the information of those participants).

4.3.2.3 Results

Fig 4.6 shows the means and standard errors of the categorization result at each fricative step in different conditions, along with the results of the baseline condition and the Female B /s/-favoring conditions represented by the grey lines. The blue lines indicates the percentage of /s/-equivalent responses in the same condition (dashed line) and opposite condition (solid line). The yellow line lying in between is the average /s/ responses at each step in the neutral condition where the male speech does not contain any /s/ or /ʃ/.

In Fig. 4.6, the most /s/-equivalent responses are exhibited in the same condition where Male A's sibilants are /s/-favoring, and the least /s/-equivalent responses are exhibited in the opposite condition where Male A's sibilants are /ʃ/-favoring. The results of the neutral condition lie in between. Since the identical training with Female B in the first phase is supposed to provide these participants with a similar perceptual expectation for Female B, the differences we observed between conditions should be attributed to their training in the second phase with Male A. Indeed, the overall /s/ responses of the three two-gender learning conditions are consistent with the sibilant properties of Male A's speech in their

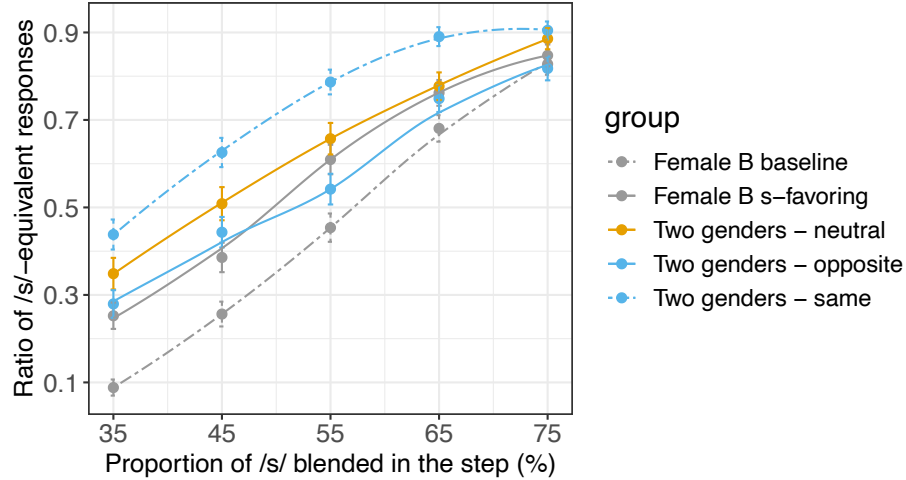


Figure 4.6: Exp 2a: /s/ response rate as a result of cross-gender perceptual learning with different sibilant distributions (mean and standard error)

second training phase. This pattern consistent with what we have seen in Exp 1 and lend support to an account of *update* where listeners generalize their perceptual learning across genders.

A mixed-effects model (Model-2a) is evaluated to predict the Response of each trial ($S=0, SH=1$) in Exp 2a, with Step (scaled and centered), Trial(scaled and centered), and Condition(treatment coded, ref: baseline) as the fixed effects, Condition:Step and Condition:Trial as the interaction items, and by-Subject Step and Frame as random slopes. The model reports the *isSingular* error. An examination of correlation between dependent variables suggest substantial colinearity between Step and Frame. Then, the random factors of the model are changed to include random slopes of Step by Subject and a random intercept of Frame. The model compiles and the result is shown in Table 4.4

Table 4.4 reveals significant main effects of Step, Condition, and Trial. The significantly negative Step effect indicates that, with a larger proportion of [s] is mixed in the stimulus, the probability of that stimulus being perceived as /f/-equivalent is lower ($\beta = -2.20, p < 0.001$) in the baseline condition. The model also reveals a significant interaction between Step and Condition for all the non-baseline conditions with positive coefficients. Since the proportion of /f/-equivalent responses should be decreasing with the

Fixed Effects	Estimate	SE	z value	Pr(> z)
(Intercept)	0.29	0.36	0.81	0.42
Step	-2.20	0.18	-12.27	< 0.001***
Condition Female B s-favoring	-0.84	0.47	-1.78	0.07
Condition Two genders - same	-1.95	0.48	-4.06	< 0.001***
Condition Two genders - neutral	-1.20	0.50	-2.40	0.02*
Condition Two genders - opposite	-0.68	0.47	-1.43	0.15
Trial	0.22	0.09	2.59	0.009**
Step:Condition Female B s-favoring	0.60	0.24	2.47	0.01*
Step:Condition Two genders - same	0.78	0.25	3.13	0.002**
Step:Condition Two genders - neutral	0.79	0.26	3.09	0.002**
Step:Condition Two genders - opposite	0.76	0.24	3.14	0.002**
Condition Female B s-favoring:Trial	0.03	0.14	0.25	0.80
Condition Two genders - same:Trial	-0.26	0.13	-2.03	0.04*
Condition Two genders - neutral:Trial	-0.38	0.14	-2.70	0.007**
Condition Two genders - opposite:Trial	-0.09	0.14	-0.65	0.52

Model-2a: Response~Step*Condition+Condition*Trial+(Step|Subj)+(1|Frame)

Table 4.4: The fixed effects of the logistic mixed-effects model in Exp 2a

increase of step, a positive value for the interaction item would indicate that the decrease becomes shallower instead of becoming sharper. Therefore, the slope of categorization boundary along the fricative continuum is sharper in the baseline condition than all the other conditions ($\beta_{same} = 0.78, p = 0.002$; $\beta_{neutral} = 0.79, p = 0.002$; $\beta_{opposite} = 0.76, p = 0.002$). In other words, participants who had exposure to the training stimuli in any condition show a shallower categorization boundary than those who had no prior exposure to training stimuli.

The Condition effect is marginally significant the Female B s-favoring condition ($\beta = -0.84, p = 0.007$), which is consistent with the result of the pretest. For experimental conditions newly introduced in this experiment, the Condition effect is significant for the *same* ($\beta = -1.95, p < 0.001$) condition and the *neutral* ($\beta = -1.20, p = 0.02$) condition, both of which show significantly more /s/-equivalent responses (and fewer /ʃ/-equivalent responses) than the baseline condition. This is expected because stimuli on the training phases in the three conditions are either /s/-favoring or /s ʃ/-free, giving rise to an overall boost in /s/. The categorization result in the opposite condition is not essentially different from the baseline condition ($\beta = -0.68, p = 0.15$). This suggests that listeners in the

opposite condition have integrated the /j/-favoring distribution during their training with Male A and used this knowledge to cancel out the influence of the earlier /s/-favoring training with Female B. This is not surprising especially when the /s/-favoring training effect is only marginally significant in the first place.

The effect of Trial is also significant ($\beta = 0.22, p = 0.009$), suggesting that listeners are more likely to report on /j/ for later trials. The interaction between Trial and Condition is significant for the same condition ($\beta = -0.26, p = 0.04$) and the neutral condition ($\beta = -0.38, p = 0.007$), but not for the Female B s-favoring condition ($\beta = 0.03, p = 0.8$) or the opposite condition ($\beta = -0.09, p = 0.52$). This implies that listeners tend to report fewer /j/-equivalent responses at a later point of the test phase in the *same* and the *neutral* conditions than the baseline condition. This pattern is different from what we have seen in Exp 1a, and we do not know for sure why there is such a discrepancy.

To further check whether the second-phase exposure to Male A's speech has shifted listener perceptual boundary further away from the Female B /s/-favoring condition, I relevelled the Condition factor with Female B s-favoring as the baseline and re-ran the model. The result shows that among the three two-gender conditions, only the *same* condition exhibits a significant difference from the Female B /s/-favoring condition ($\beta = -1.1, p = 0.02$), while the other two conditions do not (*neutral*: $\beta = -0.36, p = 0.47$; *opposite*: $\beta = 0.16, p = 0.73$). The results with both kinds of model leveling point to an influence of the Male A /s/-favoring training but none of the Male A /j/-one in two-gender perceptual learning in Exp 2a.

4.3.2.4 Summary

Exp 2a is designed to examine whether the conclusions of Exp 1a can be replicated when Female A's speech is replaced by Female B's. The sibilants of Female B has a more gender-typical frequency distribution than Female A. The result echoes the conclusion of Exp 1a, showing that the final categorization result reflects the joint perceptual learning outcomes with the female speaker and the male speaker. In specific, like in Exp 1a, the final catego-

rization results differ among the three two-gender conditions in consistent ways with their exposure Male A’s speech. Again, this lends support to an *update* account suggesting that the perceptual learning with Male A’s sibilants is generalized to the perception of Female A’s speech.

4.3.3 Exp 2b: Previous /f/-favoring training with Female B

4.3.3.1 Experimental conditions and goals

Exp 2b evaluates whether the results of Exp 2a can be replicated when /f/-favoring sibilant manipulation is adopted instead of an /s/-favoring one. Just like Exp 2a, Exp 2b also contains three conditions that differ in Male A’s sibilant distribution in the second training phase. Participants first completed a learning phase with Female B’s /f/-favoring speech, then proceeded to a second learning phase with Male A’s speech that is either /s/-favoring (*opposite*), /f/-favoring (*same*) or sibilant-free (*neutral*). In the end, participants are tested on Female A’s sibilants along an /s-f/ continuum spliced into words of minimal pairs. The only difference between Exp 2a and 2b lies in the first training condition with Female B: It is /s/-favoring in Exp 2a but /f/-favoring in Exp 2b. As elaborated in Section 4.1, an /s/-favoring manipulation would result in lower center of gravity and an /f/-favoring manipulation would result in higher center of gravity in general. Therefore, we expect that Female B’s and Male A’s speech in the Two genders - *opposite* condition to have more distinct spectral distributions in Exp 2b than in Exp 2a.

This is shown by Fig. 4.7, which demonstrates the distribution of COG measures of the sibilants of different speakers in Exp 2a and 2b. We can see that the COG distributions of the sibilants of Female B and Male A are more similar in the left facet (Exp 2a) than in the right facet (Exp 2b). Based on the previous finding that the generalization requires a certain amount of acoustic overlapping between sibilants of the different speakers (e.g., Eisner and McQueen, 2005; Kraljic and Samuel, 2006), I expect that the update of perceptual learning across the two training phases might fail to occur in Exp 2b, at least in the Two genders - *opposite* condition.

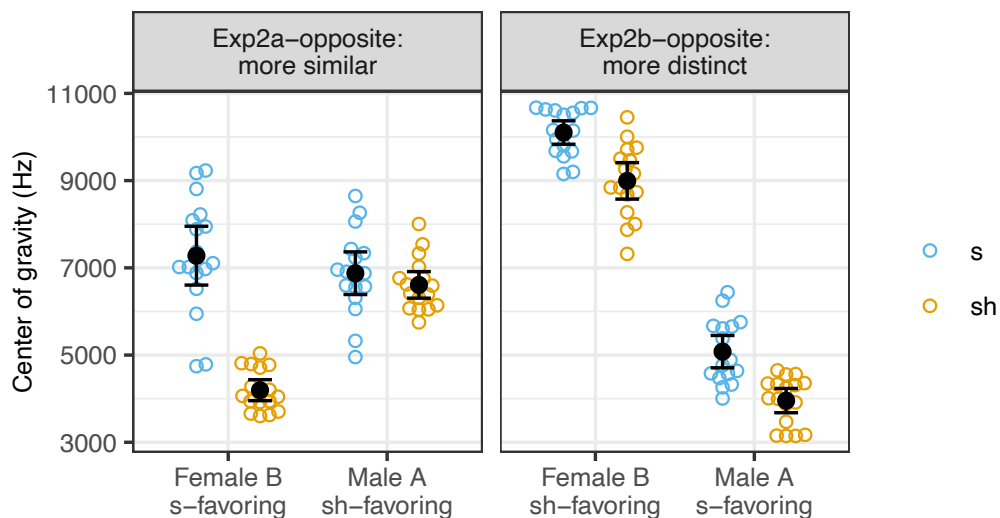


Figure 4.7: The COG of sibilants in the two - gender opposite conditions in Exp 2a and 2b (mean and 95% CI)

4.3.3.2 Participant

Participants in the three conditions of Exp 2b are all recruited from Prolific, with 31 of them in the *same* condition, 30 of them in the *opposite* condition, and 31 in the *neutral* condition. Among participants in the *same* condition, 12 identified themselves as female, 17 identified as male, and 2 chose not to identify. Their ages vary from 18 to 45 ($Mean = 26.5, SD = 9.0$). Participants in the *opposite* condition include 21 female, 7 male, and 2 persons who chose not to self identify. They age from 18 to 48 years old ($Mean = 26.2, SD = 8.4$). Participants in the *neutral* condition are 14 female, 15 male and 2 persons who preferred not to identify. Their age vary from 18 to 62 years old ($Mean = 30.4, SD = 14.0$). Along with the data of the above participants, I have also plotted the data of participants in the baseline condition and the Female A /f/-favoring condition as a reference (see the “Participant” subsection under Section 4.3.1.2 for the information of those participants).

4.3.3.3 Results

Fig. 4.8 shows the result of Exp 2b. The grey lines indicate categorization without training or with one-phase training of Female B’s /f/-favoring speech. The blue lines indicate the categorization results after two phases of perceptual learning with Female B and Male A sequentially, with either the same or opposite directions. The yellow line indicates a two-phase learning condition where the speech of the intervening male speaker does not contain /s/ or /f/. We can see that the results of the two-gender non-neutral conditions (marked by the two blue lines) lie roughly above the dashed grey line (Female B baseline). This indicates that, no matter whether the training speech in the second phase is /s/-favoring or /f/-favoring, they both end up boosting the response of /s/-equivalent responses after the perceptual shift towards /f/ in the first training phase.

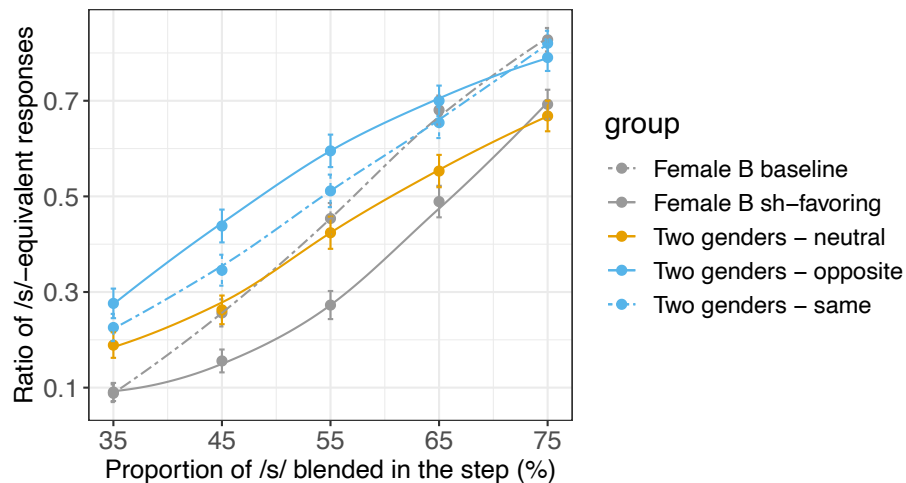


Figure 4.8: Exp 2b: /s/ response rate as a result of cross-gender perceptual learning with different sibilant distributions (mean and standard error)

The pattern shown in Fig. 4.8 is different from the results of previous experiments in several regards. The most obvious discrepancy is that the result of the Two genders - *same* condition (blue dashed line) does not shift the perceptual boundary to the /f/ end compared to the baseline (grey dashed line). This happens regardless of the fact that /f/-favoring training with Female B has already shifted the boundary substantially to have fewer /s/ and more /f/. In other words, a second-phase exposure to /f/-favoring stimuli

with Male A actually causes participants to hear more /s/ in the test phase, compared to /f/-favoring training with Female B alone. The second discrepancy is that listeners who had exposure to Male A’s neutral training speech actually reported fewer /s/ sound compared to the Two genders - *same* and *opposite* conditions, instead of falling between these two conditions.

I then evaluated a logistic mixed effects model to predict the response in the test (S=0, SH=1), with the main effects of Condition (baseline: Female B baseline), Step, and Trial (continuous, both scaled and centered) as the fixed effects, Subject as the random slope, and Frame as the random intercept. The main effects are presented in Table 4.5. The result revealed significant effects of Step ($\beta = -2.31, p < 0.001$) and Trial ($\beta = 0.24, p = 0.005$), as well as significant interactions between Step and Condition for the three two-gender experimental conditions ($\beta_{same} = 0.53, p = 0.03; \beta_{neutral} = 1.01, p < 0.001; \beta_{opposite} = 0.75, p = 0.002$). However, neither of the three two-gender conditions shows a significant shift from the baseline condition.

Fixed Effects	Estimate	SE	z value	Pr(> z)
(Intercept)	0.25	0.45	0.55	0.58
Step	-2.31	0.18	-13.04	< 0.001***
Condition Female B sh-favoring	1.10	0.54	2.06	0.04*
Condition Two genders - same	-0.38	0.54	-0.71	0.48
Condition Two genders - neutral	0.34	0.54	0.62	0.53
Condition Two genders - opposite	-0.74	0.54	-1.37	0.17
Trial	0.24	0.09	2.79	0.005**
Step:Condition Female B sh-favoring	0.38	0.24	1.57	0.12
Step:Condition Two genders - same	0.53	0.24	2.20	0.03*
Step:Condition Two genders - neutral	1.01	0.23	4.35	< 0.001***
Step:Condition Two genders - opposite	0.75	0.24	3.13	0.002**
Condition Female B sh-favoring: Trial	-0.15	0.13	-1.17	0.24
Condition Two genders - same: Trial	-0.18	0.13	-1.36	0.18
Condition Two genders - neutral: Trial	-0.30	0.13	-2.30	0.02*
Condition Two genders - opposite: Trial	-0.35	0.13	-2.60	0.009**

Model-2b: Response~ Step*Condition+Condition* Trial+(Step|Subj)+(1|Frame)

Table 4.5: The fixed effects of the logistic mixed-effects model in Exp 2b

To evaluate whether the two-gender conditions significantly diverge from the Female B /f/-favoring condition, I re-ran the model with the “Female B sh-favoring” condition as the

baseline, to evaluate how the second-phase training changed the categorization boundary of the outcome of /j/-favoring training with Female B. The result shows that both exposure to Male A /s/-favoring speech and exposure to Male A /j/-favoring speech significantly boosted /s/-equivalent responses compared to the Female B /j/-favoring condition (same: $\beta = -1.49, p = 0.006$; opposite: $\beta = -1.85, p = 0.0007$). Exposure to the neutral speech of Male A does not make any difference to the final responses ($\beta = -0.77, p = 0.15$).

4.3.3.4 Summary

Exp 2b shares the same structure with Exp 2a, with Female B's /s/-favoring speech replaced by her /j/-favoring speech. The hypothesis to be examined is whether the acoustic dissimilarity between sibilants of different speakers affects the update of perceptual learning. If it does not make a difference, we should expect the same pattern as shown in previous experiments, namely, that the results of different experimental conditions show consistency with their training with Male A at the second phase. If it does make a difference, then this pattern will not occur. Instead, the result of the Two genders - *opposite* condition will not differ significantly from the Female B j-favoring condition due to the failure of integrating Male A's /s/-favoring training into the final result.

However, the result of Exp 2b is not consistent with either of the above predictions. Surprisingly, Male A's /s/-favoring and /j/-favoring distributions both turn out to exert an /s/-favoring influence on the perception of Female B's speech. This set of results does not replicate the update of perceptual expectations across speakers as found in Exp 1a, 2b and 2a, because the perceptual boundaries in different conditions are not consistent with the sibilant distributions of Male A's speech in the second training phase. It is also different from what the "acoustic dissimilarity" account predicts, i.e., the update of perceptual expectations will be blocked for the Two genders - *opposite* condition. In fact, the result of the *opposite* condition is consistent with its second-phase training condition, but the result of the *same* condition is not, unexpectedly. There must be other factors playing a role that leads to such a pattern. Before we jump into possibilities accounting for the unexpected

result of the Two genders - *same* condition, another crucial questions we want to figure out first is whether training with Male A's speech alone would successfully trigger a perceptual shift in the test with Female B's sibilant continuum. This question will be examined in Exp 2c. After this observation is obtained from Exp 2c, I will then come back to discuss the potential factors that lead to the results of Exp 2b.

4.3.4 Exp 2c: No previous training with Female B's /s-f/

4.3.4.1 Experimental conditions and goals

In Exp 2a and 2b, listeners have been trained on the sibilants of Female B and Male A in two sequential training phases, and then they are tested regarding which piece of knowledge to use for the categorization of Female B's sibilants. Results obtained so far are not consistent between the two sub-experiments. Most surprisingly, Male A's /f/-favoring training leads to a significant boost of /s/-equivalent responses compared to the result of its previous /f/-training phase with Female B. Exp 2c is intended to evaluate whether training with Male A's speech alone successfully induces the intended perceptual shift on the sibilant continuum of Female B's speech.

4.3.4.2 Participant

30 participants are recruited for the Male A /s/-favoring condition and 26 are recruited for the Male A /f/-favoring condition from the UPenn subject pool. Participants in the Male A /s/-favoring condition are 10 male and 20 female, aging from 18 to 22 years old ($Mean = 19.5, SD = 1.20$). Participants in the Male A /f/-favoring condition are 15 female and 11 male, aging from 18 to 22 years old ($Mean = 19.8, SD = 1.0$).

4.3.4.3 Result

Fig. 4.9 compares the categorization boundaries of the Female B baseline condition and the Male-only conditions. It shows that, while Male A's /s/-favoring training stimuli induces a certain amount of perceptual shift to the intended direction, Male A's /f/-favoring training

stimuli does not induce any perceptual shift, as indicated by the overlapping categorization results between the Female B baseline condition and the Male A /ʃ/-favoring condition.

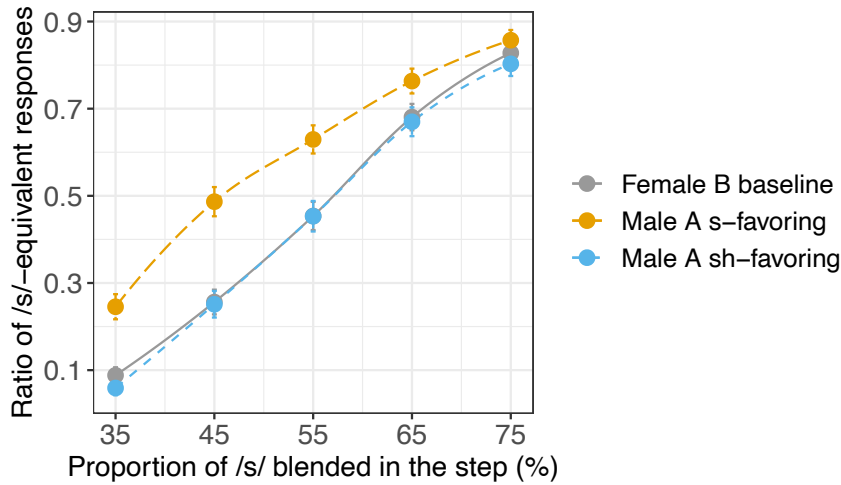


Figure 4.9: Exp 2c: /s/ response rate as a result of training with Male A compared to the Female B baseline (mean and standard error)

A logistic mixed-effects model (Model-2c-a) are fitted to examine whether the categorization results of the Male s-favoring condition and the Male /ʃ/-favoring condition are significantly different from the Female B baseline. The dependent variable is the response of each trial (S:0, SH:1). The main effects are Condition(Male A s-favoring/Male A sh-favoring/Female B baseline, treatment coded, baseline: Female B baseline), Phoneme (the original phoneme associated to each auditory frame, sum-coded, baseline: SH), and Step and Trial (both scaled and centered). The models also include Condition:Step and Condition:Trial as the interaction items, by-Subject Step as the random slope, and Frame as a random intercept. The result of this model is shown in Table 4.6.

Table 4.6 reveals a marginally significant difference between the Female B baseline condition and the Male A s-favoring condition ($\beta = -1.02, p = 0.056$) but no difference between the Female B baseline condition and the Male A sh-favoring condition ($\beta = 0.31, p = 0.58$). The model also reveals significant main effects of Step ($\beta = -2.29, p < 0.001$) and Trial ($\beta = -0.44, p < 0.001$), as well as significant interactions of Condition:Trial for both conditions ($\beta_{MaleS} = -0.26, p = 0.04; \beta_{MaleSH} = -0.36, p = 0.008$) and Step:Condition for the

Fixed effects	Estimate	Std. Err.	z value	Pr(> z)
(Intercept)	0.16	0.43	0.37	0.71
Step	-2.29	0.19	-12.17	< 0.001***
Condition Male A s-favoring	-1.02	0.53	-1.91	0.056
Condition Male A sh-favoring	0.31	0.55	0.55	0.58
Trial	0.25	0.09	2.71	0.007**
Step:Condition Male A s-favoring	0.56	0.25	2.21	0.03*
Step:Condition Male A sh-favoring	0.05	0.27	0.18	0.86
Condition Male A s-favoring:Trial	-0.26	0.13	-2.07	0.04*
Condition Male A sh-favoring:Trial	-0.36	0.14	-2.65	0.008**

Model-2c-a: Response~Step*Condition+Condition*Trial+(Step|Subj)+(1|Frame)

Table 4.6: The fixed effects of the logistic mixed-effects model evaluating the effect of training with Male compared to the Female B baseline in Exp 2c

Male A s-favoring condition ($\beta = 0.56, p = 0.03$). In a nutshell, the model suggests that training with Male A managed to induce a certain amount of perceptual shift towards /s/ but completely failed to induce any shift towards /j/ on Female B’s sibilant continuum.

The second set of comparisons I made is between the Male-only training conditions and the Two-gender conditions that contain the same Male training phase. Fig. 4.10 shows the results of Exp 2c along with Two-gender conditions sharing the same manipulation of Male A’s speech in Exp 2a and 2b. The left facet shows the results of training with Male A’s /s/-favoring alone speech (grey) and training with Male A’s /s/-favoring speech preceded by either Female B’s /s/-favoring speech (blue) or Female B’s /j/-favoring speech (yellow). Similarly, the right facet shows the results of training with Male A’s /j/-favoring speech alone (grey) and training with Male A’s /j/-favoring speech preceded by either Female B’s /j/ speech (blue) or Female B’s /s/-favoring speech (yellow).

Two logistic mixed-effects models (Model 2c-b and 2c-c) respectively for the data of experimental conditions including Male A /s/-favoring training and those including Male A /j/-favoring training. The dependent variable is the SH response in the test. The main effects are Condition(Male only, Two genders - *opposite*, Two genders - *same*; treatment coded, reference level: Male only), and Step and Trial (both scaled and centered). The models also include Condition:Step and Condition:Trial as the interaction items, Subject as a random slope, and Frame as a random intercept. Full list of the estimates in the two

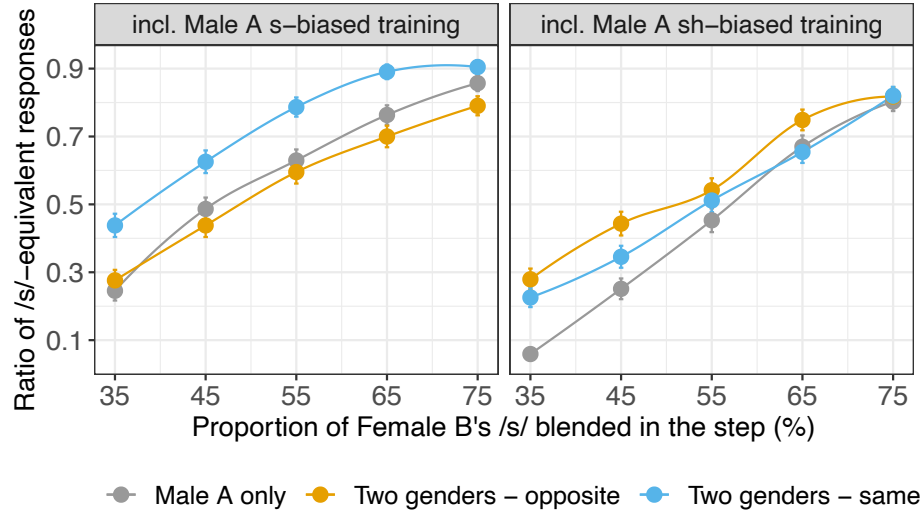


Figure 4.10: Exp 2c: /s/ response rate as a result of training with Male A compared to two-genders training conditions (mean and standard error)

models are provided in Table B.3 and Table B.4 in Appendix B. Model 2c-b (Table B.3) shows that a preceding training phase with Female B's /s/-favoring speech results in a marginally significant boost in /s/ responses compared to the result of training with Male A's s/-favoring speech alone indexed by the grey line and the blue line ($\beta = -0.93, p = 0.07$). However, a preceding training phase with Female B's /f/-favoring speech results in an insignificant reduction of /s/-equivalent responses compared to the Male A only training indexed by the grey line and the yellow line ($\beta = 0.27, p = 0.6$). In the right facet, the three lines stay together closely, with the categorization boundaries of the two two-gender conditions lying above the Male /f/-favoring condition. Model 2c-c (Table B.4) shows a marginally significant boost in /s/ responses after a previous Female B /s/-favoring training phase and an insignificant boost in /s/ responses after a previous Female B /s/-favoring training phase compared to exposure to Male A /f/-favoring training alone.

This is consistent with both Exp 2a and the first analysis of Exp 2c. All of them suggest that exposure to Male A's /s/-favoring speech leads to a certain amount of boost in /s/ responses, whereas exposure to Male A's /f/-favoring speech does not successfully trigger a significant amount of reduction in /s/ response rate.

4.3.4.4 Summary

Exp 2c evaluates how training with Male A's /s/-favoring or /ʃ/-favoring speech generalizes to the categorization of Female B's continuum. The results show that exposure to Male A's /s/-favoring speech has successfully elicited a perceptual shift on Female B's /s-ʃ/ continuum whereas exposure to Male A's /ʃ/-favoring speech does not. This is different from what happens in Exp 1c when the same question is tested with Female A's continuum, meaning that the generalization of perceptual learning is conditioned on the raw frequencies of sibilants in the test phase. I further compare the results of Male only training conditions with the two-gender training conditions that involve the same training with Male A. The results

4.4 Discussion

In Exp 2, I evaluated whether the findings of Exp 1 can be replicated when the training and test speech of Female A is replaced with the speech of a different female speaker, Female B, who has higher sibilant frequencies. The most straightforward consequence of this change is that the acoustic properties of Female B's sibilants after manipulation are more distinct from the male speaker's sibilants, especially when Female B's speech is manipulated to favor the perception of /ʃ/ for ambiguous sibilants. Another less direct consequence is that sometimes there is a mismatch between the raw acoustic values of sibilants and their phonological anchors within the same training phase. The result of Exp 2 (Exp 2b and 2c in particular) shows that the replacement of female speaker indeed make a difference to the categorization results through perceptual learning. This section goes through the findings we have so far from Exp 2 and discusses their implications on the constraints of acoustic distributions on sibilant perceptual learning.

4.4.1 Effect of the acoustic distributions of the training and test stimuli

By comparing the results of Exp 1c and 2c, we see a discrepancy regarding whether exposure to Male A’s training speech could induce an amount of shift to the perception of a female speaker’s sibilant continuum: In Exp 1c, training with Male A’s /f/-favoring speech successfully generalized to Female A’s sibilant continuum (as indicated by Fig. 3.9 and Table 3.7), whereas in Exp 2, training with Male A’s /f/-favoring speech failed to induce a perceptual shift to the perception of Female B’s sibilant continuum (as indicated by Fig. 4.9 and Table 4.6). As the training speech of Male A is identical across the two experiments, the obviously differentiating factor between Exp 1c and Exp 2c lies in the acoustic property of the sibilant continua. Given that acoustic energy of Female B’s sibilants are distributed at higher frequency areas than that of Female A’s in natural production (Fig. 4.3), we shall expect the same for the sibilants on the test continua since they are synthesized following the same procedure.

Fig. 4.11 shows the COG values of the sibilants of Male A’s training stimuli in two conditions as well as Female A’s and Female B’s test stimuli on a five-step continuum. The female speakers’ COG measures are indicated with red circles, with Female A’s in the left facet (same as in Fig. 3.4) and Female B’s in the right facet (same as in Fig. 4.4). As expected, Female B’s sibilants have higher energy frequencies (with COG ranging from 7000-10000 Hz) than Female A’s overall (with COG ranging from 5000-9000 Hz). The circles in blue and yellow represent the COG values of the 17 sibilants in /s/-words and the other 17 in /f/-words in either a /s/-favoring training condition or a /f/-favoring one. The black diamond stands for the mean value of the two kinds of sibilants in each training condition, which can be interpreted as the perceptual boundary predicted from the 34 sibilants in training. The COG values of Male A’s sibilants are identical across the two facets, and they are presented twice to facilitate the comparison between Male A’s COGs and those of different female speakers in different experiments.

We first discuss whether the effect of *acoustic dissimilarity* provides an explanation for the discrepancy we have observed, namely, perceptual learning with Male A successfully

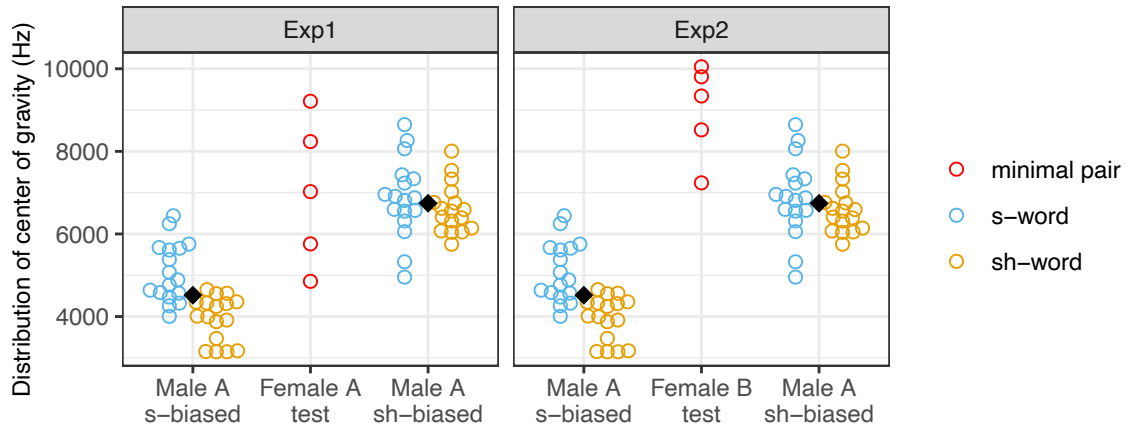


Figure 4.11: COG measures of the training stimuli of Male A and the test stimuli of Female A and Female B. Black diamonds indicate the mean COG values of all the sibilants in an experimental condition.

generalizes to the female speaker’s speech except that /*f*/-favoring perceptual learning with Male A failed for the test with Female B’s speech. According to previous studies (e.g., Kraljic and Samuel, 2006), the lack of acoustic similarity or overlap between the training stimuli and the test stimuli may block the generalization of perceptual learning across speakers. In our case, this account predicts the absence of perceptual learning in the Male A /*s*/-favoring condition in Exp 2c compared to Female B’s baseline, because it is the only condition with no overlap between the COGs of Male A’s sibilants and the female speaker’s sibilants. On the contrary, if the acoustic similarity between Female B’s test continuum and Male A’s training input does not matter, then we should expect the Male A training condition to work in identical ways in Exp 1c and Exp 2c. However, none of the two predictions are consistent with what I have found. In Exp 2c, while the /*s*/-favoring training with Male A has induced a marginally significant amount of perceptual shift towards /*s*/, the /*f*/-favoring training with Male A, instead failed to induce a perception shift with Female B’s sibilants. If this lack of perceptual learning effect is due to the little acoustic overlap between Male A’s /*f*/-favoring speech and Female B’s sibilant continuum, then we should also expect to see no effect of Male A /*s*/-favoring training because Male A’s COGs on the /*s*/-favoring condition is even more distinct from Female B’s sibilant continuum (Fig. 4.4). However, this

is not what we see. The above analysis suggests that the lack of acoustic overlap between the training and the test phases does not block the generalization of perceptual learning (as shown in Male A /s/-favoring condition), and that an amount of acoustic overlap does not entail the generalization of perceptual learning (as shown in the Male A's /j/-favoring condition).

Then, why did training with Male A's /j/-favoring speech fail to induce a perceptual shift on the continuum of Female B's continuum? I propose that this can be explained by the account of *acoustics-phonology mismatch*, which states that the adaptation to a speaker's speech involves learning the raw acoustic distributions of their phonemes, and the generalization of perceptual learning involves evaluating the relative acoustics between the acoustic targets of different speakers. In Fig. 4.11, the relative position of the acoustic boundaries predicted from Male A's speech (as indexed by the black diamonds) and the acoustic values of the female speakers' test trials is different between Exp 1c and 2c. Particularly in Exp 2c, the acoustic boundaries predicted from Male A's /s/-favoring and /j/-favoring speech are both lower in COG than Female B's test continuum. Even though Male A's /j/-favoring speech contains a number of /s/-sounding sibilant tokens according to the pilot perception study, the predicted acoustic boundary of the Male A /j/-favoring condition is still lower in COG than the lowest end of Female B's continuum. In other words, training Male A's /j/-favoring speech does not boost the number of /j/-equivalent responses on Female B's continuum, because Female B's sibilants are all acoustically /s/ according to the boundary predicted from Male A's /j/-favoring speech.

However, this is not the case with Exp 1c, where the COGs of some of Female A's test trials lie between the acoustic boundaries indicated by the Male A /s/-favoring and /j/-favoring conditions. Some of the trials of Female A's test continuum (at least for the last two test trials of Female A counting from the top of the left facet) can be interpreted as /s/ compared to the boundary set in the Male A /s/-favoring condition and /j/ compared to that set in the Male A /j/-favoring condition. This leaves some room for perceptual learning to successfully introduce both acoustically and perceptually plausible interpretations in

opposite directions. In comparison, in Exp 2c, the Male A /ʃ/-favoring training is designed to set a high bar for the perception of /s/ sounds perceptually whereas the bar is not sufficiently high in the acoustic space. The training fails to make it acoustically authentic for listeners to interpret Female B’s sibilants as /ʃ/. This explanation successfully resolves the discrepancy between Exp 1c and 2c.

To conclude, I did not find evidence that acoustic overlap is required for the generalization of perceptual learning to work in Exp 2c. Instead, the results show that the generalization of sibilant perceptual learning across speakers is subjected to the constraint of raw acoustic distributions of the sibilants in addition to their relative distributions in the phonological space.

4.4.2 Effect of the acoustic distributions of different training phases

This section evaluates the two hypothesized acoustic constraints, *acoustic dissimilarity* and the *acoustics-phonology mismatch*, regarding how well they account for the results of the two-gender conditions in Exp 2a and 2b. The results of these two sub-experiments do not show the same pattern. Particularly, in Exp 2b, which starts with Female A /ʃ/-favoring training, successive exposure to Male A’s /s/-favoring and /ʃ/-favoring speech both lead to more /s/-equivalent responses than the baseline condition. This is different from what we have observed in Exp 1a, 1b and 2a, where the two Male A conditions always result in different patterns, such that the perceptual learning effect with the female is maintained in one condition and is set back in the other.

I now discuss whether the unexpected pattern boundary shift we observed in Exp 2b can be well addressed by either the *acoustic dissimilarity* account or the *acoustics-phonology mismatch* account. Fig. 4.12 shows the COG values of the training stimuli in different training phases in Exp 2a (left) and 2b (right). The circles in blue and yellow respectively represent the COG values of the sibilants in /s/-words and in /ʃ/-words in each training phase. The black diamonds stand for the predicted perceptual boundary between /s-ʃ/ based on the training stimuli and are computed by averaging across the COGs of all the

sibilants within the same training phase. The COG values of Male A’s production are identical across facets, and they are presented twice because the two training phases with Male A occurred in each of the sub-experiment.

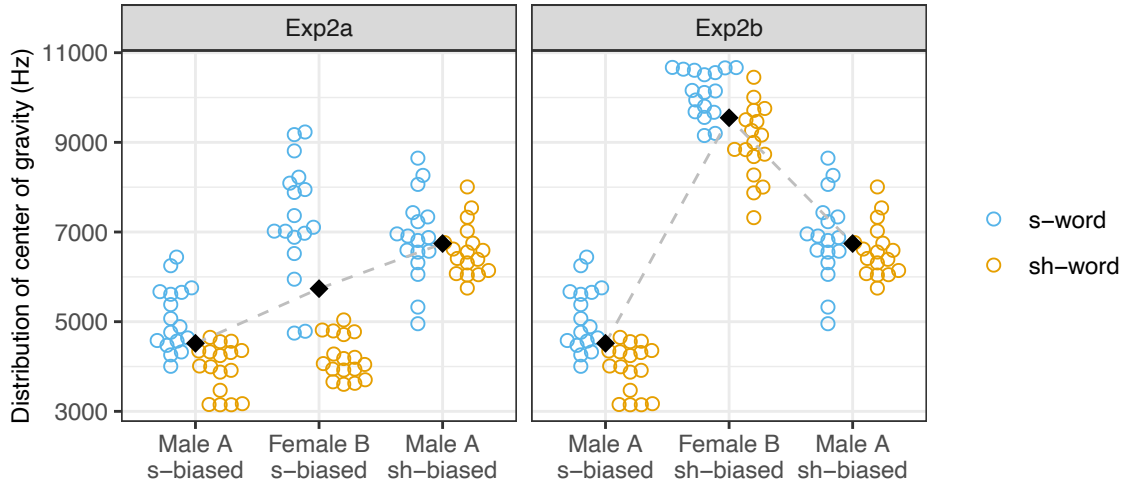


Figure 4.12: COG measures of the training stimuli in Exp 2a and 2b. Black diamonds indicate the mean COG values of all the sibilants in a training phase.

Let us first consider how the *acoustic dissimilarity* account works for the speech of different speakers in the training phases. By comparing the COGs distributions of different speakers between Exp 2a and 2b, we see that the two sub-experiments indeed involve different amounts of acoustic overlap between their training stimuli. In Exp 2a, substantial acoustic overlap can be observed between Female B’s sibilants and each set of the Male A’s sibilants with opposite biases. In Exp 2b, however, the acoustic overlap is sparser. The sibilants in the Male A /s/-favoring condition (< 7000 Hz) and those in the Female B /f/-favoring condition (> 7000 Hz) do not overlap at all in the COG dimension. The /f/-words in the Female B /f/-favoring condition overlaps partially with sibilants in the Male A /f/-favoring condition within the range from 7000 to 9000 Hz, but Female B’s /s/-words have no overlap with words in the Male /f/-favoring condition either.

The *acoustic dissimilarity* account predicts that the influence of Male A /s/-favoring training would be hindered since the acoustic distribution of sibilants in that phase is too different from the first phase. The effect of Male A /s/-favoring training would be more

likely to occur than the /j/-favoring training. In other words, the experimental condition that is the most vulnerable to such an acoustic effect is the Two-genders *opposite* condition of Exp 2b, where listeners received Female B /j/-favoring training and Male A /s/-favoring training successively. However, the results of Exp 2b (Section 4.3.3) are backwards to the prediction of the *acoustic dissimilarity* account. We keep seeing a robust influence of Male A /s/-favoring on the final categorization phase consistently in Exp 2a, 2b and 2c, which conflicts with the previous predictions stating the lack of effect of the Male A /s/-favoring training phase because of acoustic dissimilarity. Therefore, as with the results of Exp 2c, the result of Exp 2b does not lend support to an *acoustic dissimilarity* account either for speech materials in the training phases.

Moreover, Fig. 4.8 shows an unexpected pattern of results in the Two-genders *same* condition. In specific, the condition with both /j/-favoring training with Female B and /j/-favoring training with Male A end up giving rise to a boost in /s/-equivalent responses compared to the baseline condition and the Female B /j/-favoring condition. This suggests that instead of leading to fewer /s/-equivalent responses, a second phase exposure to Male A /j/-favoring speech actually boosted the amount of /s/-equivalent responses compared to participants who only had the first-phase training. This can be only addressed under an *acoustics-phonology mismatch* account where the operation of perceptual learning depends on the raw acoustic distributions to some extent. In Fig. 4.12, we can see that the perceptual boundary indicated by Male A's /s/-favoring speech is around 4500 Hz, and the boundary indicated by Male A's /j/-favoring speech is around 7000 Hz. According to the perceptual boundaries set by Male A's speech in both conditions, all tokens in Female B's /s/-favoring speech all of Female B's speech should be categorized as /s/ because all of them are above 7000 Hz. In other words, Male A's /s/-favoring and /j/-favoring speech are both /s/-favoring compared to Female B's speech in the first case in terms of the absolute COG values.

To conclude, in Exp 2, I investigate two potential acoustic constraint on the multispeaker perceptual learning. An *acoustic dissimilarity* constraint states that perceptual learning

fails to generalize if there is a lack of acoustic overlap or similarity between the acoustic distributions of sibilants in different phases or speakers. An *acoustics-phonology mismatch* constraint states that perceptual learning fails to generalize if there is a mismatch between the directions of perceptual shifts intended by the raw acoustic distributions of stimuli and by their phonological distribution in the perceptual space. The experiment results lend support to the latter account. Exp 2c provides a case where perceptual learning fails to induce the intended shift when there is an acoustics-phonology mismatch between the training phase and the test phase. Exp 2b shows that the stimuli of different training phases are also subject to such a constraint. The results of Exp 2 do not show evidence for an influence of acoustic dissimilarity between stimuli of different speakers. Further investigations are needed to evaluate whether results in previous studies that indicate the influence of acoustic overlap can be accounted for under alternative explanations.

Chapter 5

Exp 3: Perceptual Learning of /t-d/ across Speaker Genders

This chapter reports an experiment investigating how listeners adjust their perceptual beliefs in multi-speaker listening by integrate the VOT distributions of /t d/ from speakers with different genders. As a temporal cue, VOT is considered to differ from spectral cues in containing less information about the vocal track of the speaker, which is an indexical source of the speaker's gender identity. This chapter contains four sections. Section 5.1 summarises the relevant findings in the previous literature on the acoustic parameters and perceptual correlates of /t d/, focusing on how they vary with social or contextual factors and play a role in the perceptual learning of talker-specific productions of /t d/. Building on the review, I articulate the main research question and hypotheses of this chapter at the end of Section 5.1. Section 5.2 and 5.3 report on Experiment 3, which investigates the perceptual learning of /t-d/ across speakers of different genders. Section 5.2 provides an overview of the methodology, including experiment design and procedure, experimental conditions, and the makeup and manipulation of the stimuli; Section 5.3 reports on results of a pilot study and two sub-experiments. Finally, Section 5.4 summarises the main findings and their implications and concludes this chapter.

5.1 Background and research question

5.1.1 Voice onset time and other phonetic cues to English stop voicing

Stop voicing is signaled and identified by multiple acoustic cues in production and perception. Among these cues, the voice onset time (VOT) is generally considered the primary cue for English stop voicing. VOT refers to the time between the burst of a stop and the onset of vocal cord vibration in the following segment. This parameter was first proposed by Lisker and Abramson (1964) as a universally available parameter for obstruent contrast based on a cross-linguistic study. In Lisker and Abramson (1964)'s proposal, the voicing contrast of world languages is cued by three acoustic categories, namely, voicing lead, short lag (voiceless unaspirated), and long lag (voiceless aspirated). Although the three acoustic categories were later shown to be insufficient to categorize different voicing contrasts in the world (e.g., Cho and Ladefoged, 1999; Docherty, 2011), they identify American English stop consonants reliably. Voiced stop consonants (/b d g/) have short-lag VOTs whereas voiceless stop consonants (/p t k/) have relatively long-lag VOTs (e.g., Lisker and Abramson, 1964). Although voiced stops can be produced as fully voiced between voiced sounds, they are generally produced with lag VOT values in other contexts (e.g., Davidson, 2016).

VOT varies between and within phonemic categories. Regarding between-category differences, VOT not only varies between voiced and voiceless stops; it also varies also between stops sharing with the same voicing value. There is a general increase in VOT with more posterior places of articulation (i.e., /p/ < /t/ < /k/, Cho and Ladefoged, 1999; Chodroff and Wilson, 2017, *inter alia*). In the meantime, the VOTs of voiceless stops from the same talker robustly co-vary with one another in American English, as evidenced by strong correlations among the mean VOT of different voiceless stops at the individual level (Chodroff and Wilson, 2017). In other words, a talker's /p/ sound can reveal information about the talker's VOT of /t/ and /k/.

Regarding within-category variability, the VOT of a specific segment can be affected by speech rate, vocalic context, prosodic position, and lexical properties. Speech rate is a

natural correlate of VOT. Intuitively, faster speaking rates result in a significant decrease in VOT (e.g., Allen and Miller, 1999; Allen et al., 2003; Kessinger and Blumstein, 1997, 1998; Miller et al., 1986). VOT length is also subject to prosodic factors including prosodic edges (Cho and Keating, 2009), lexical stress (Byrd et al., 2006), phrasal accent (Cole et al., 2007), word length (Flege et al., 1998; Klatt, 1975), and whether the syllable is embedded in a sentence (Morris et al., 2008). Cho and Keating (2009) report that utterance-initial /t/ sounds are realized with a slightly longer VOT than utterance-medial ones with matched accent status, as a result of domain-initial strengthening effects. Cole et al. (2007), among others, further report an interaction between the domain-initial strengthening effect and accentual VOT lengthening, such that the domain-initial strengthening effect often disappears when the word bears phrasal accent. Klatt (1975) and Flege et al. (1998) find that the VOT of voiceless stops is longer in monosyllabic words than in polysyllabic words. In parallel, Morris et al. (2008) find longer VOT of stops in utterances of isolated syllables than in sentences. Regarding lexical properties, VOT is also subject to word frequency (Yao, 2009) and neighborhood density (Baese-Berk and Goldrick, 2009; Buz et al., 2016; Kirov and Wilson, 2012). Yao (2009) reports that frequent words tend to have shorter VOTs than infrequent ones. Baese-Berk and Goldrick (2009), among others, report that the VOT of word-initial voiceless stops is longer in words with a voiced-initial neighbor. Finally, regarding the influence of vocalic context, a substantial body of studies have reported longer VOTs before the vowel /i/ for voiceless stops (Flege et al., 1998; Klatt, 1975; Port and Rotunno, 1979; Weismer, 1979).

In this chapter, the only acoustic cue of /t d/ to be manipulated is the VOT distribution. This is a temporal cue that involves less of the spectral information associated with speaker identity, differing from sibilant contrasts that I investigated in the last two chapters. Other than VOT, a number of secondary cues are also identified to influence the perception of stop voicing (e.g., Lisker, 1986), which I am to review in the remainder of this section. The review of the secondary cues in this section is intended to inform the stimuli manipulation decisions in this chapter. With the primary goal of the experiments in this chapter to be

inducing perceptual shifts of voicing boundary along the VOT dimension, I will make other cues ambiguous or uninformative, such that they do not signal a strong preference for the identification of /t d/.

Following VOT, the second most well-studied acoustic cue associated with voicing contrasts of stops concerns perturbations to F_0 in the adjacent areas. Haggard et al. (1970) observed that listeners were more likely to perceive a synthesized stop consonant as being voiceless when it precedes a high F_0 than a low F_0 . Whalen et al. (1993) reported that listeners use the F_0 cue even when voice-onset time is unambiguous. Research has also indicated that F_0 trajectory shapes differ depending on the voicing of the stops. Whalen et al. (1993) and Shultz et al. (2012) based their perception studies on the assumption that the F_0 exhibits flat or rising contours into the following vowel after the voiced stops while displaying falling contours after the voiceless stops. Haggard et al. (1970) accounted for pitch change at the onset of voicing after the articulatory closure of a consonant as a reflection of glottis status during that closure. For a preceding stop consonant in English, a low rising pitch indicates a closed glottis, and a high falling pitch indicates a glottis that is still partly open. They also found that listeners would integrate the association between pitch change and glottis status in the perception of voicing for an onset stop.

In addition to VOT and F_0 , the effects of the first formant transition and frequency at the voicing onset are also shown to affect the voicing contrast in English. Liberman et al. (1958) found that voiced stops in initial position could sound like their voiceless counterparts with the beginning of the first-formant transition cut off (except for the combination between /t d/ and /o/). Liberman et al. (1958) attributed this to a delay in the excitation of F1 in voiceless stops in production, where the frequency of F1 transition at the voicing onset is much higher for English voiceless stops than for voiced stops. For voiced stops where voicing begins simultaneously with the release, acoustic energy from vocal fold vibration excites the first formant during its rise from the consonantal release to the steady-state vowel frequency. In contrast, for voiceless stops, since voicing onset much later than the release, F1 is not excited until late in the CV transition when the vocal tract is close to the steady-state

vowel configuration. Perceptual studies show that stops with a longer F1 transition and/or lower F1 frequency at voicing onset are more likely to be classified as voiced than stops with shorter F1 transitions and/or higher F1 frequency at voicing onset, all other things being equal (Kluender, 1991; Lisker, 1977; Stevens and Klatt, 1974; Summerfield and Haggard, 1977). The direction of the effect of the F1 transition pattern on voicing classification is thus consistent with the observed production data for English and other languages with aspiration in the voiceless stops.

The next phonetic correlate of voicing to be reviewed is the spectrum distribution of burst. Acoustically, voiceless labial and coronal stops are found to have greater energy at higher frequencies in comparison to homorganic voiced stops (Chodroff and Wilson, 2014; Halle et al., 1957; Parikh and Loizou, 2005; Sundara, 2005; Zue, 1976), as indexed by a number of measures. For example, Zue (1976) measured the mean spectral peak of the initial 10-15 ms of coronal stop bursts and found that /t/ had a mean spectral peak of 3600 Hz, higher than the mean of 3300 Hz for /d/. Perceptual studies evaluated the spectral cue with continua created by crossing burst shape and VOT and found that voiceless identifications were more likely for tokens with higher frequency bursts (e.g., Chodroff and Wilson, 2014; Keating, 1979; Nittrouer, 1999). Chodroff and Wilson (2014) also conducted tasks of goodness ratings, which showed that stops identified as voiceless labials and coronals are better members of their respective phonetic categories when COG is higher; in contrast, this effect is not as straightforward for stops categorized as voiced. They suspected that burst spectrum has a substantial impact on goodness only when VOT is non-prototypical for a phonetic category.

The last two cues documented in the literature that proved to be perceptual relevant to the voicing contrast are aspiration amplitude and vowel duration. Repp (1979) found that the perception of voicing for syllable-initial stops was affected by the amplitude of aspiration noise before voicing (relative to the following periodic portion of the vowel). Vowel duration can be a salient cue for the voicing perception of the stop following that vowel (De Jong, 2004; Klatt, 1976), but it becomes a less reliable cue of voicing in onset stops (e.g., De Jong,

2004; Peterson and Lehiste, 1960).

5.1.2 Talker and gender variability in VOT production and their influence on perceptual learning

From an indexical point of view, VOT is found to covary with a number of social factors including speaker dialect (Lipani et al., 2019; Scobbie, 2006), gender (Byrd, 1992; Smith, 1978; Swartz, 1992; Whiteside and Irving, 1998), and age (Benjamin, 1982; Morris and Brown Jr, 1994; Torre III and Barlow, 2009). Since a significant motivation of this dissertation is to compare speaker gender effects on the perceptual learning of sibilants and stops, I primarily focus on findings of the *gender* variation of VOT measures. Previous studies regarding the effects of speaker gender on VOT in English generally find that females produce longer average VOT values than males for voiceless stops (Morris et al., 2008; Swartz, 1992; Sweeting and Baken, 1982; Whiteside and Irving, 1997, 1998). In contrast, findings of VOT for voiced stops are more mixed. While some studies have reported longer mean VOT values for females than males (Morris et al., 2008; Swartz, 1992; Whiteside and Irving, 1997), other studies have reported the opposite (Sweeting and Baken, 1982; Whiteside and Irving, 1998). Several possible explanations have been proposed for the gender difference in VOT. A physiological account arises in the light of findings on a correlation between VOT length and lung volume (Hoit et al., 1993). According to this view, it is relatively easier for male speakers with generally larger supraglottal cavities to form sub and supraglottal air pressure differences for vocal cord vibration, which shortens their VOTs (Koenig, 2000; Smith, 1978; Swartz, 1992; Whiteside et al., 2004; Whiteside and Irving, 1997, 1998). Note that this explanation mainly accounts for VOT variation in voiceless stops rather than voiced ones. Meanwhile, a stylistic account proposes that the gender difference of VOT can also stem from different styles between gender. In this view, females tend to speak more carefully than males and therefore produce shorter VOTs for voiced stops and longer VOTs for voiceless stops to secure sufficient phonological contrasts between the two stops (Whiteside and Irving, 1997, 1998).

Substantial VOT variability has also been identified across specific talkers within relatively homogeneous groups, even after controlling for differences in speaking rate (Allen et al., 2003; Theodore et al., 2009). The cross-talker variability is particularly large among the voiceless categories. Although the VOTs of the voiceless stops observed in these studies generally belong to what is usually called long lag, they can span tens of milliseconds, making this source one of the larger factors of VOT variation. These findings are interpreted to imply that VOT may function as an indexical source for speaker specificity (Allen et al., 2003). More recently, Kleinschmidt (2019) quantifies the informativity of VOT distributions about speakers' social-indexical variables (gender, age, and speaker specificity) using the VOT data of stops from the Buckeye corpus (Pitt et al., 2007). Specifically, he evaluates the Kullback-Leibler (KL) divergence of the group-level VOT distribution from the overall cue distribution from all groups. He finds that VOT distributions grouped by individual speakers diverge more from the aggregate distribution than those grouped by age and gender. This finding seems to be consistent with Allen et al. (2003)'s claim about VOT being indexically useful for identifying idiosyncratic speakers. However, when comparing VOT distributions' informativity measures with the vowel formant distributions extracted from another corpus, Kleinschmidt (2019) shows that the divergence of by-speaker VOT from the aggregate VOT distribution is still much smaller than the divergence of by-speaker vowel formants from the aggregate formant distribution. These findings indicate that, although VOT might be more useful for the index of idiosyncratic speakers than speaker gender and age, the efficiency of VOT for identifying individuals may not be as high as other phonetic properties associated with other segments.

The influence of speaker specificity and gender on the perceptual learning of stop VOT has been examined with various training and test paradigms. However, the results are mixed, and the question of how much speaker specificity matters to the perceptual learning of VOT distributions remains unsolved. One line of relevant experiments uses a talker-speech matching task to evaluate listeners' sensitivity to speaker-specific VOT properties. This method was first developed by Allen and Miller (2004). In their study, speakers were

trained on the speech of two female talkers, Annie and Laura, with manipulated VOTs for word-initial voiceless stops. According to the experimental design, one speaker always had short VOTs, and the other always had long VOTs, with the association between VOT and speaker counterbalanced between groups. Allen and Miller (2004) found that, after exposure to these speech stimuli patterned with the name of the speaker, listeners were able to select the VOT variant consistent with their experience of Annie's and Laura's speech. Follow-up experiments in this line further showed that listeners were able to generalize their knowledge about talker-specific VOT distributions to the identification of novel words (Allen and Miller, 2004) and novel voiceless stops (e.g., training with /b-p/ but test with /g-k/, Theodore and Miller, 2010) of the speaker they were trained with. Based on these findings, the authors proposed that such sensitivity to talker-specific VOT properties makes it plausible for listeners to customize the mapping between the acoustic signal and speech sound for individual talkers.

However, studies evaluating the perceptual learning of talker-specific VOT distributions with phoneme identification tasks do not arrive at similar conclusions. These experiments evaluate listeners' sensitivity to talker specificity by quantifying their tendency to generalize the categorization boundary established based on one speaker's speech to the perception of a different speaker's speech. The more likely listeners are to generalize, the less they pay attention to speaker specificity. As mentioned in Section 1.2.3, one of the findings motivating this dissertation is that perceptual learning of stop consonant boundaries is more prone to generalize across talkers than that of fricative boundaries (Kraljic and Samuel, 2006, 2007). Kraljic and Samuel (2006) find generalization of perceptual learning across male and female talkers on a /t-d/ continuum. The perceptual shift in the voicing distinction also transfers to a /p-b/ continuum. They develop this point further in Kraljic and Samuel (2007), where they suggest that listeners learn talker-specific representations for a fricative contrast /s-ʃ/ but do not do the same for a stop contrast /t-d/.

The discrepancy between the above two lines of studies may be explained by the many differences between the two paradigms. For one thing, Allen and Miller (2004) and Theodore

and Miller (2010) have used photos to provide an additional enhancing cue for talker specificity, which might have made it easier for listeners to encode the VOT distributions associated with specific talkers. However, this may not be the essential factor leading to the discrepancy between studies, given that the talker voices used in Kraljic and Samuel (2006, 2007) are of different genders, which should have led to a more enhanced perception of talker differences than the voices in Allen and Miller (2004) and Theodore and Miller (2010). For another thing, the two paradigms also potentially differ in the involvement of explicit versus implicit memory in the experimental tasks, and the ambiguous versus well-defined exemplars adopted as stimuli in the training phases. These differences were addressed in Theodore et al. (2015), who introduced linguistic perception tasks at the end of their previous talker-speech matching task with the same two female speakers' speech that listeners had shown sensitivity to in Theodore and Miller (2010). This time, Theodore et al. collected listeners' responses of phoneme categorization and goodness ratings of the speakers' voiceless stops after training with their speech with manipulated VOTs patterned with their names. The results showed that although listeners dynamically adjusted internal category structure to be centered on experience with the talker's voice, the category boundary did not reflect speaker-specific VOT distributions. These results are consistent with Kraljic and Samuel (2006, 2007) in pointing to a null-effect of speaker specificity on the perceptual learning of VOT-based stop voicing.

Up to this point, it seems like evaluations of VOT perceptual learning with phoneme categorization tasks tend to find no effects of talker specificity, but this is not always the case. Munson (2011) evaluated whether listeners generalized what they have learned about one talker's VOT to a different talker on two separate days, with counterbalanced orders between the training talker and the test talker. If a participant received training with talker A and test with talker B on Day 1, they would receive training with talker B and test with talker A on Day 2. She found that perceptual learning generalized across talkers on Day 1 but not Day 2. This implies that cross-talker generalization when listeners did not have prior knowledge about the test talker's VOT distribution on Day 1; once they

obtained the knowledge from previous exposure, then the generalization was blocked on Day 2 because they still maintained what they had learned from the training speaker on Day 1. The results lend support to the role of talker specificity in the perceptual learning of VOT distributions by showing that the more recent training on the second talker did not overwrite the initial training for the original talker. Munson also finds that listeners learned talker-specific boundaries when the female talker had VOTs shifted towards the voiced end of the continuum, but not when the male talker was shifted in that direction. She attributed this asymmetry to the involvement of other cues to voicing such as F1 and F₀, which made it difficult to shift the male continua to the left and female continua to the voiceless end.

Munson (2011) shares a commonality with Kraljic and Samuel (2006) in that they both adopt a linguistic categorization task and identify perceptual learning through phoneme boundary shift on a voicing continuum. A methodological difference worth noting between these two studies is that the former uses unsupervised learning to induce the perceptual shift whereas the latter lexically-guided learning. It remains unclear to what degree their discrepant conclusions can be attributed to this methodological difference. A final piece of evidence to be reviewed for talker effects on the perceptual learning of VOT comes from neuroimaging studies. In Myers and Theodore (2017), listeners heard two talkers produce characteristically different VOTs for word-initial voiceless stops during a brief exposure phase. Following exposure, neural activation was measured using fMRI while listeners completed a phonetic categorization task for VOTs that was either consistent or inconsistent with their exposure. Right temporoparietal regions previously implicated in talker identification showed sensitivity to the match between VOT variant and talker, whereas left posterior temporal regions showed sensitivity to the typicality of phonetic exemplars, regardless of talker typicality. These results suggest that talker-specific VOT characteristics can be exploited for voice processing.

In summary, we have mixed findings regarding whether the perceptual learning of VOT distributions is talker-specific. This mixture can be partially attributed to the differences

in technical details between studies, including experimental procedure, tasks, and stimuli. Moreover, a deeper reason behind this discrepancy that sometimes gets ignored is that investigators differ in their expectations about the linguistic processing levels where speaker information makes a difference and what a talker effect should look like according to the results of different tasks. This discussion can benefit from more uniform criteria of the identification and quantification of a talker effect.

5.1.3 Research questions

The research question of this chapter is whether the perceptual learning of the VOT distributions of a specific speaker's /t d/ productions can be generalized to the perception of a different speaker's /t d/, who has a different gender. As with the experiments in Chapter 3, the experiment in this chapter is also set up to provide a concrete situation for the generalization of perceptual learning across talkers: If the listener has had exposure to the /t d/ productions of Female A and Male A, successively, then whose VOT boundary will listeners use to cope with the /t d/ categorization with Female A's speech?

Again, I expect the results of Experiment 3 to help us tease apart four hypotheses about the cross-gender generalization of the perceptual learning of VOT, which I have laid out in Section 2.4 (and in Table 3.1): A *retention* hypothesis states that perceptual learning is talker-specific, and only the knowledge of Female A's acoustic distribution is relevant to the perception of A's test speech. It predicts that the outcome perceptual shift aligns with Female A's acoustic distribution while remains unaffected by Male A's acoustic distribution. In contrast, if Female A's distribution is not reflected in the result of the final test, then the possibility lies with either *recency update* or *reset* depending on whether or not the test result reflects Male A's distribution. Both of these situations indicate suggest that the perceptual learning of sibilants is not speaker-specific. Finally, if both of Female A's and Male A's distributions have laid an influence on the final test, then it suggests that perceptual learning is not strictly speaker-specific given that it updates in response to acoustic exposures from other speakers as well. This kind of result lends support to a

cumulative update account.

5.2 Method Overview

5.2.1 Experimental conditions

This section reports Experiment 3, which makes the first attempt in this dissertation to evaluate the role of talker specificity in the perceptual learning of stop VOT. The stimuli adopted in this experiment are wordlist productions of Female A and Male A, who are the same speakers used in Experiment 1-2. Experiment 3 contains a pilot study and two sub-experiments (Exp 3a, 3b), all of which end with a categorization test on the same /t-d/ continuum of Female A's speech.

These sub-experiments differ in the speakers and acoustic conditions of the training phases that participants have received before the categorization test. Fig. 5.1 shows a summary of the experimental designs and procedures in each condition of the different sub-experiments in Experiment 3. We can see that the pilot study and Exp 3b differ from Exp 3a in the number of training phases involved. Pilot and Exp 1c only include no more than one training phase, and therefore the training participants have received can only be recent rather than prior. Pilot and Exp 3b differ in the speaker used in the training phase (if any).

The pilot study reports the /t-d/ categorization results of the baseline condition and the Female A /t/-favoring condition. The goal of the pilot study is to make sure that the 50% point at the categorization curve is anchored with the center of the continuum, and to make sure that the current design and stimuli works to induce a boundary shift as a result of perceptual learning.

Experiment 3a contains three experimental conditions. They each contain a training block with Female A's /t/-favoring speech, a consecutive training block with Male A's speech, and a final categorization test with Female A's speech. The three conditions differ regarding whether the intermediate training phase with Male A's speech is /t/-favoring (in the *same* direction with the first training phase), /d/-favoring (in an *opposite* direction

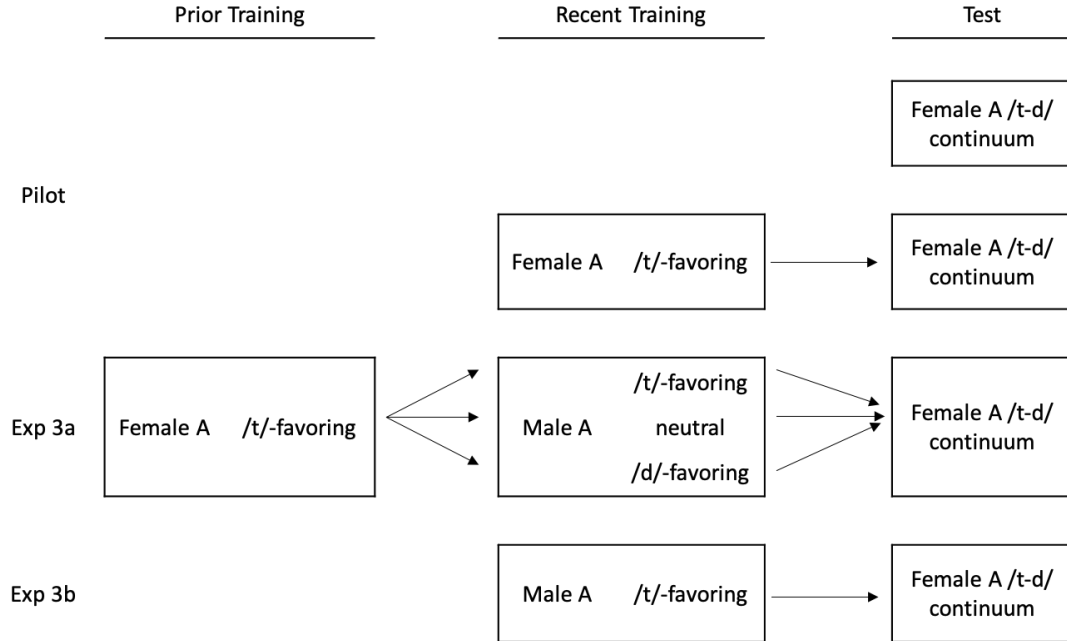


Figure 5.1: The structure of sub-experiments and conditions in Exp 3

to the first training phase), or containing no /t d/ (a *neutral* condition). By comparing the results of these three conditions, we are able to know to what extent the exposure to Male A’s speech matters for the categorization of Female A’s speech. If the outcome categorization boundary reflects the acoustic condition of the second training phase with Male A, then that lends support to the generalization of perceptual learning across genders. More detailed explanations will be presented at the beginning of experiment.

On the premise that Experiment 3a has lent some support to the generalization of perceptual learning across speakers, Experiment 3b is aimed at further teasing apart two possibilities under the generalization condition, i.e., *cumulative update* and *recency update*. The two possibilities both predict Male A’s VOT distributions has been integrated and reflected in the final categorization result, but they make different predictions about whether Female A’s VOT distribution is involved. The setup of Exp 3b parallels that of Exp 1c and 2c and follows the same logic to tease apart the two possibilities. By comparing the categorization results of participants who have received /t/-favoring training with Male A only in 3b and those who have received /t/-favoring training with both of the two speakers

in experiment 3a and 3b, we are able to know to what extent the earlier training phase still matter to the categorization result.

5.2.2 Word list and recording

The stimuli used in Experiment 3 are manipulated from recordings of Female A and Male A obtained according to the procedure described in Section 2.3.2. Each word with /t/ or /d/ is produced twice by the speakers, once with /t/ and the other time with /d/ (e.g., *cafeteria* and *cafederia*), in order to provide transitional signals to both /t/ and /d/ in the same lexical context. Words selected for stimulus construction in Experiment 3 are listed in the following. They include 17 /t/-containing words, 17 /d/-containing words, 51 words with no /t d/, and 8 words in minimal pairs that contrast each other by word-initial /t/ and /d/ segments.

- */t/-containing words*: authentic, cafeteria, cemetery, consultation, frontier, hesitation, infantile, lunatic, magnetism, military, momentary, novelty, overtime, relative, scientific, voluntary, warranty (count: 17)
- */d/-containing words*: academic, agenda, armadillo, legendary, comedy, avocado, crocodile, evidence, handy, hazardous, iodine, kingdom, melody, merchandise, remedial, secondary, residence (count: 17)
- */t-d/ minimal pairs*: down-town, Dutch-touch, deer-tear, dime-time (count: 8)
- *words without /t d/*: airline, among, anvil, average, banana, buffalo, village, waffle, wharf, earning, eyebrow, feeling, firefly, follow, foul, framing, gable, gravel, verify, raven, honey, iguana, January, jewelry, journal, row, marina, enamel, Nepal, nothing, sigh, shine, shame, same, obscene, runaway, crucial, flourishing, thumbnail, lonely, legacy, ribbon, gargoyle, volleyball, vulgar, initial, official, evaluation, rehearsal, eraser (count: 51)

The 34 words containing /t/ or /d/ are selected from Table 2.3, which are mostly adopted from Kraljic and Samuel (2006). In these words, /t/ and /d/ only occur in word-

medial positions. Among the 17 /t/-containing words, the /t/-initial syllable bears primary stress in 5 words, secondary stress in 7 words, and no stress in 5 words. Among the 17 /d/-containing words, the /d/-initial syllable bears primary stress in 2 words, secondary stress in 8 words, and no stress in 7 words.

The 51 words without /t d/ consist of 37 filler words with no /t d s ʃ/ from Table 2.4 and 14 words with either /s/ or /ʃ/ but no /t d/ from Table 2.2 and 2.5. Among the 51 filler words, six of them contain another minimal pair of stops – /p b/ (11%), but those words only occur in a neutral training block. In other words, they are not included as fillers in a /t/- or /d/- favoring training phase.

The word boundaries and VOT proportions are manually annotated for the 34 /t d/-containing words in Praat. The word lengths and VOT lengths are extracted and presented in Fig. 5.2. Each circle represents a unique word spoken by one of the two speakers. The black points represent the group means, and the error bars represent the 95% confidence intervals of the VOT lengths and word lengths.

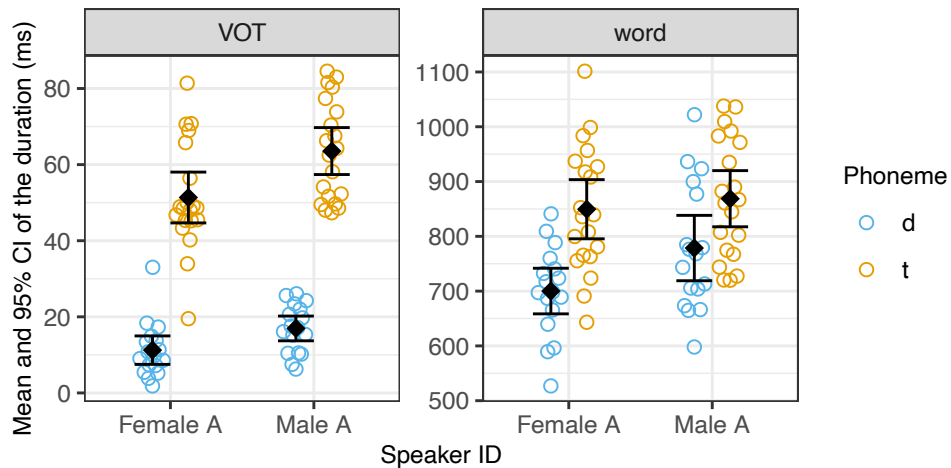


Figure 5.2: The VOT length of /t d/ and the duration of /t d/-containing words from Female A and Male A (mean and 95% confidence interval)

The left facet shows a bimodal distribution of positive VOT values for voiced and voiceless stops from both Male A and Female A. The VOT of /d/ varies from 0 to 30 ms, and the VOT of /t/ varies approximately from 40 to 80 ms. Male A’s VOT lengths of /t/

and /d/ are distributed far apart with no overlap VOT lengths between these two phonemes. Female A's VOT values are also distinct between phonemes except for two words, namely, a /t/-containing word with a 20-ms-long VOT (word *authentic*) and a /d/-containing word with a 35-ms-long VOT (word *armadillo*). In general, Female A's VOTs are shorter than Male A's, and this between-talker difference in VOT is larger for /t/ (Female A: 50 ms; Male A: 65 ms) than for /d/ (Female A: 12 ms; Male A: 18 ms). The right facet shows the word duration of each speaker's lexical tokens, which gives us an idea of the speech rates of the two speakers since the 34 /t d/ words they produced are identical. It appears that the distributions of word lengths are similar between Female A and Male A, except that Male A's /d/-containing words have longer duration than Female A's. Note that this difference is in the opposite direction of the average gender difference in the literature, which reports that female speakers have longer VOT than male speakers (Morris et al., 2008; Swartz, 1992; Sweeting and Baken, 1982; Whiteside and Irving, 1997, 1998).

5.2.3 Step selection and manipulation

As elaborated in Section 5.1.1, the voicing cue intended for perceptual learning and examination in Experiment 3 is the length of VOT. The secondary cues are neutralized to different extents for each word such that they are less informative of the original phoneme of the word frame. In the light of this goal, five sections of areas with cues to the critical phoneme are identified for the /t/- and /d/- containing tokens. They are obstruction, burst, aspiration, transition, and vowel. These five sections are exemplified in Fig. 5.3 by displaying the beginning of the spoken word "town". Area *a* corresponds to the obstruction part where the airflow is completely stopped while the pressure inside the vocal track builds up. Area *b* is where the obstruction is released in a burst of air. Area *c* corresponds to aspiration. Area *d* is the formant transition that links up the aspiration and the full vowel production. It is indexed by smaller waveform amplitude and lower energy distributed in high areas of the spectrum. Area *e* is the following vowel.

By definition, VOT corresponds to the section of area *b* (after the release of closure) and

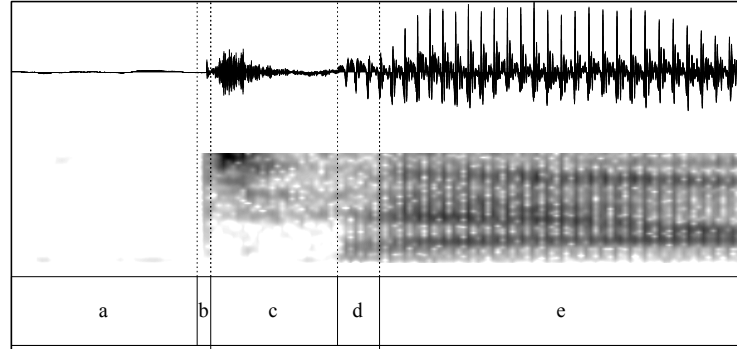


Figure 5.3: A sound clip of the word *town* and its cue-bearing areas of relevance to the perception of the critical phoneme: a - obstruction, b - burst, c - aspiration, d - transition, e - vowel.

c (before the beginning of vocal fold vibration) in Fig. 5.3. However, since the manipulation process would involve temporal compression or extension of the annotated section, I decided to exclude area *b* from the annotated period of VOT, because it is less natural to lengthen or shorten the duration of the burst. Therefore, among the five specified areas, the aspiration part (area *c* in the figure) is explicitly annotated for each token with /t d/, and their lengths are to be varied at different steps of the continuum

The 34 words used for training are each manipulated to embed a five-step /t-d/ continuum. The five steps of stimuli are manipulated from a /t/-substitution token of a corresponding word. For example, the /t-d/ continuum embedded in the word *cafeteria* is manipulated from the auditory word token *cafeteria*, whereas the /t-d/ continuum embedded in *academic* is manipulated from the auditory non-word token *acatemic*. Among the five steps of stimuli, the longest VOT step is a re-synthesis of the /t/-substitution production itself, and the remaining four steps are generated by temporally compressing area *c* in Fig. 5.3 to the 20%, 40%, 60%, and 80% of its original length. Temporal compression is conducted using the PSOLA algorithm in Praat and its graphical user interface for duration interpolation. I set aside a region of 1 ms both before and after area *c* as transitions where the speech rate slows down or speeds up gradually. In cases where tokens with shortened VOTs still introduce a perceptual bias towards /t/, I have to neutralize some of the secondary cues to the /t/ sound in these tokens in different extents depending on how salient

those cues are. These manipulations may include reducing the amplitude of the burst (area *b*), attenuating or cutting off the transitional area (area *d*), and cross-splicing the whole following vowel (area *d* and *e*) from the corresponding lexical frames produced with /d/. The secondary cues, once manipulated, are kept constant for each word throughout the continuum, rather than covarying with VOT at each step of the continuum. All the synthesized stimuli are normalized to 70 dB.

A lexical decision task is conducted to select the most ambiguous step of VOT for each training word to be used in the training phase. Since Exp 3 contains training with Female A towards a /t/-favoring direction and training with Male A towards both /t/-favoring and /d/-favoring directions (see Section 5.2.1), lexical decision tasks are conducted with Female A’s 17 /t/-containing words and Male A’s 34 /t d/-containing words. The above 51 words are each spliced with stops in five VOT steps. Then they are presented in a single block in a randomized order and participants needed to judge whether they are an English word. The results of this lexical decision task are shown in Appendix A. Among the five compression factors ranging from 0.2 to 0.8, the one yielding the most ambiguous lexical decision result is selected to be used for the stimulus construction of a particular lexical token. If two VOT steps are approximately equally far from the 50% point, then the mean of the two factors is used as the temporal compression ratio for stimulus construction (e.g., 0.3). Finally, if a /t/-containing token with a VOT of 0.2 of its original length still receives a high rate of /t/-equivalent responses (i.e., “Word” responses), then a compression factor of 0.1 will be adopted in addition to manipulations of secondary cues for this particular lexical item.

Table 5.1 shows the temporal compression step chosen for each critical lexical frame for Female A and Male A based on the lexical decision results. † indicates that the token has also received manipulations of its secondary cues to stop voicing.

Minimal-pair words produced by Female A that I use for testing are also manipulated to bear stops from a /t-d/ continuum. Like the training stimuli, continua of the minimal-pair words are also all manipulated from /t/-containing tokens (i.e., *touch*, *tear*, *town*, and *time*) by temporal compression of the aspiration area. 10 VOT steps are generated for each

ID	/d/ words	Male A	/t/ words	Female A	Male A
1	academic	0.4	authentic	0.1 [†]	0.1 [†]
2	according	0.2	cafeteria	0.2	0.5
3	agenda	0.2	cemetery	0.4	0.2
4	armadillo	0.4	consultation	0.2	0.1 [†]
5	avocado	0.2	frontier	0.2	0.2
6	comedy	0.4	hesitation	0.1 [†]	0.1 [†]
7	crocodile	0.3	infantile	0.1 [†]	0.2 [†]
8	evidence	0.4	lunatic	0.4	0.2
9	handy	0.4	magnetism	0.2	0.2
10	hazardous	0.3	military	0.2	0.2 [†]
11	iodine	0.6	momentary	0.2	0.2
12	kingdom	0.2	novelty	0.2	0.2 [†]
13	legendary	0.4	overtime	0.1 [†]	0.2
14	melody	0.2	royalty	0.2	0.2
15	merchandise	0.3	scientific	0.2 [†]	0.2
16	remedial	0.4	voluntary	0.2	0.2
17	secondary	0.3	warranty	0.2 [†]	0.2

[†] indicates simultaneous manipulations of secondary voicing cues.

Table 5.1: The temporal compression factor chosen for each word of the training stimuli for Female A and Male A

lexical token by varying the length of the aspiration area from the 10% to the 100% of its original length at an interval of 10% between adjacent steps. The temporal compression is also implemented using the PSOLA algorithm as described above and manipulations on the secondary cues of the burst, stop-to-vowel transition and the following vowel are manipulated for each of the four tokens to ensure ambiguity.

A word identification task is conducted to select five temporal steps out of ten for each of the four /t/-containing tokens of Female A. Listeners need to choose, for the four tokens spliced with stops at ten VOT steps, whether it is a /t/-initial word or /d/-initial one (e.g., “tear” or “dear” for tokens manipulated from *tear*). The step selection process follows such a procedure based on the categorization results of the word identification task. First, a most ambiguous step is chosen for each word frame by identifying the step at which the response rate for /t/ or /d/ is the closest to 50%. Then, two more steps are taken respectively on the left and the right of the continuum to make up a five-VOT continuum. All the synthesized stimuli are normalized to 70 dB.

Table 5.2 presents the the final VOT length at each of the five chosen steps for the four lexical token to be used in the test phase of Exp 3.

lexical frame	step 1	step 2	step 3	step 4	step 5
touch	9.79	19.58	29.37	39.16	48.95
tear	15.36	30.72	46.08	61.44	76.89
town	10.92	21.84	32.76	43.68	54.60
time	9.34	18.68	28.02	37.36	46.70

Table 5.2: The VOT duration at each step of the /t-d/ continuum embedded in different minimal-pair words in the test phase (ms)

Note that the interval between adjacent steps for the test continuum is 10% of the original VOT length of a /t/-word (around 10-15 ms), which makes a relatively fine-grained change for tokens on the continuum, instead of fully covering the VOT range between a typical /d/ and a typical /t/. This step selection procedure is designed to fulfill the following to requirements. One is that the center step of the continuum should anchor closely enough with the location of the /t-d/ categorization boundary along the continuum, to avoid any additional distributional learning induced by the continuum itself (Tamminga et al., 2020). The other is that stops at the five VOT steps are essentially ambiguous in nature, in order to avoid including any standard instances that are reported to prevents perceptual shift from happening (Zhang and Samuel, 2014).

5.3 Experiment and Result

5.3.1 Pilot study: Learning Female A’s /t-d/

5.3.1.1 Experimental conditions and goals

The goal of this pilot study is to verify that perceptual learning can be successfully elicited with the speech of Female A after manipulation. The result is also intended to demonstrate that the 50% perceptual boundary between Female A’s /t-d/ has been anchored with the center of the test continuum by default without prior training. The pilot study contains a *baseline* condition and a */t/-favoring learning* condition. Participants in the baseline

condition completed a single test block of Female A’s spoken words, which consist of 35 test trials with ambiguous dental stops embedded in /t-d/ minimal pairs and 17 filler words that do not contain /t/ or /d/. The result of this condition is taken as a reference of the default /t-d/ perceptual boundary for Female A’s speech. Participants in the /t/-favoring learning condition first completed a /t/-favoring training block with Female A’s speech before they proceeded to complete the same test block as in the baseline condition. The difference in the perceptual boundary between these two conditions is taken as an indicator of the perceptual learning effect.

5.3.1.2 Participant

All the participants in the pilot study are recruited from Prolific. Thirty participants are in the baseline condition. They are 17 female, 12 male, and a remaining participant who preferred not to identify their gender. Their age varies from 18 to 67 years old ($Mean = 29.4, SD = 12.4$). Another group of 27 participants is assigned to the Female A /t/-favoring condition. They are 16 female and 13 male, aged 19-57 ($Mean = 29, SD = 10.4$).

5.3.1.3 Result

Fig 5.4 shows the results of phone categorization in the baseline condition (the dashed line) and in the /t/-favoring condition (the solid line). Points and error bars indicate the mean and standard error of the /t/-equivalent response rate across the group.

As expected, Fig 5.4 shows that participants in the /t/-favoring condition report more /t/-equivalent responses on each step of the categorization continuum than those in the baseline condition. In other words, after exposure to stimuli with standard /d/ and non-standard /t/, listeners become more likely to categorize the ambiguous stops between /t-d/ as /t/ rather than /d/, compared to the baseline condition. In Fig. 5.4, we can also see that the 50% point of the categorization boundary in the baseline condition aligns with the middle step of the continuum (Step 3). The response rates at the endpoints have only reached 20% and 75% in the baseline condition, meaning that the test stimuli on the

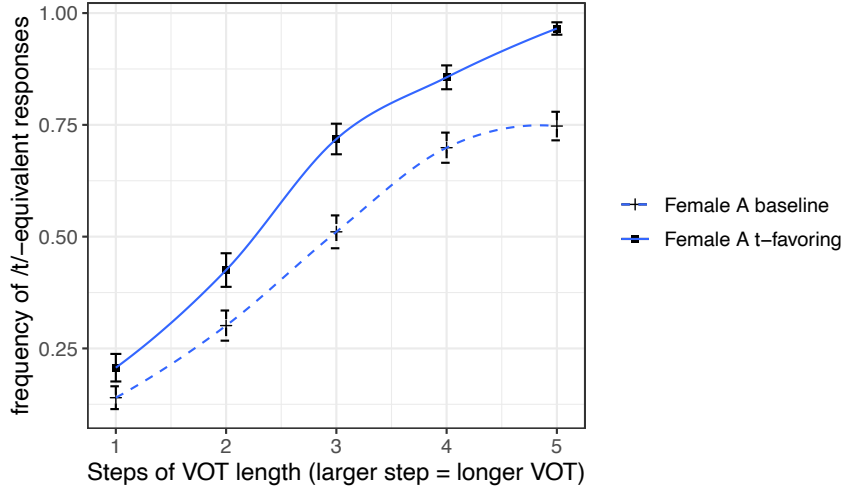


Figure 5.4: Exp 3 pilot: Boundary shift after /t/-biased learning compared to the baseline categorization with Female A (mean and standard error)

continuum are essentially ambiguous in nature, even for the ones at the endpoints.

A logistic mixed-effects regression model (Model-pilot3) is evaluated to predict the Response of each trial (T=0, D=1), with Step (1-5, scaled and centered), Trial (1-51, scaled and centered), and Condition (treatment coded, reference level: the baseline condition) as the fixed effects, Condition:Step and Condition:Trial as the interaction items, and Step by Subject and by Word as the random slopes. The result of the model (Model-pilot3) is shown in Table 5.3.

Fixed effects	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.04	0.44	0.09	0.93
Step	-1.39	0.27	-5.24	< 0.001***
Condition Female A t-favoring	-1.20	0.36	-3.38	< 0.001***
Trial	-0.15	0.09	-1.74	0.08
Step:Condition Female A t-favoring	-0.67	0.23	-2.94	0.003**
Condition Female A t-favoring:Trial	0.14	0.13	1.07	0.28

Model-pilot3: Response~Step*Condition+Condition*Trial+(Step|Subj)+(Step|Word)

Table 5.3: The fixed effects of the logistic mixed-effects model in Exp 3 pilot

The result reveals a significant main effect of Step ($\beta = -1.39, p < 0.001$), suggesting that in the baseline condition, longer VOTs at larger steps lead to fewer perception of /d/. Crucially, the effect of Condition also turns out to be significant ($\beta = -1.20, p < 0.001$),

suggesting that participants with /t/-favoring training are less likely to show /d/-equivalent responses than the baseline condition. A significant interaction between Step and Condition indicates that the slopes of the categorization boundary along the continuum becomes sharper in the Female A /t/-favoring condition ($\beta = -0.67, p = 0.003$). In other words, one step of VOT increase would lead to fewer /d/ perception in the Female A /t/-favoring condition than in the baseline condition. The effect of Trial is insignificant; neither is its interaction with condition.

5.3.1.4 Summary

The above results suggest that I have successfully aligned the 50% response point of the categorization boundary with the center of the continuum, and that the Female A /t/-favoring training works to induce a significant amount of boundary shift towards the expected direction. This pilot or the remaining of the experiment does not contain training with Female A /d/-favoring speech, because there is no hypothesis about how perceptual shift towards different directions might interact with our observations.

5.3.2 Exp 3a: Perceptual learning of /t-d/ with Female A and Male A

5.3.2.1 Experimental conditions and goals

The goal of Exp 3a is similar to that of Exp 1a, namely, to evaluate what listeners do with their existing perceptual expectation established for a previous speaker after encountering another speaker of a different gender. Possibilities of involved multi-talker perceptual learning mechanisms include *retention*, *reset*, and *update*, as described in Section 7.1. Three experimental conditions are designed in Exp 3a to tease apart the above possibilities. All participants in the three conditions of this experiment completed two phases of perceptual learning, one /t/-favoring training with Female A and the other training phase with Male A. Male A's speech in the second training phase has been manipulated to be either /t/-favoring, /d/-favoring, or /t d/-free depending on the experimental condition that each participant was assigned to. In the end, participants completed a word identification task

on Female A's /t d/ minimal pair words with VOT continua spliced into the minimal pairs.

Crucially, different possibilities of cross-speaker perceptual learning behaviors make different predictions about the outcomes of the three training conditions. If the final categorization is always consistent with the /t/-favoring distribution of Female A's speech across the three conditions, then indicates that listeners *retained* the phonetic information of Female A regardless of the talker switch. If the final results in the three experimental conditions are always consistent with the perceptual bias induced by their exposure to Male A's speech and thus differ from one another, it indicates that listeners have *updated* their beliefs as a function of the recent input regardless of speaker specificity. Finally, if the categorization results across conditions show a similar reset to the baseline condition, then it indicates that listeners *reset* their phonetic beliefs in response to the switch of talker.

5.3.2.2 Participant

90 participants are recruited from Prolific to participate in Experiment 3a, with 30 participants in each experimental condition. The responses of these participants are analyzed along with the data of the 30 participants in the baseline condition and the 27 participants in the Female A /t/-favoring condition, as introduced in Section 5.3.1. Participants in the Two genders - *neutral* condition are 20 female and 10 male, aged 18-65 years old ($Mean = 32.2, SD = 13.5$). Participants in the Two genders - *same* condition are 14 female and 16 male, aged 19 to 52 years old ($Mean = 32.4, SD = 9.0$). Participants in the Two genders - *opposite* condition are also 14 female and 16 male, aged 19 to 54 years old ($Mean = 30.97, SD = 10.23$).

5.3.2.3 Result

Fig. 5.5 shows the means and standard errors of the categorization result at each step in the three two-gender learning conditions. The blue lines indicates the perceptual learning outputs after two phases of perceptual learning sequentially with a male speaker and a female speaker, where Male A either have the same (dashed) or an opposite (solid) pronunciation

characteristics of /t d/ with Female A. The yellow line indicates a two-phase perceptual learning condition, where no /t d/ is involved in the intervening male speaker’s speech. The grey lines stand for the results of the baseline and the Female /t/-favoring conditions, which I have already reported in the pilot study.

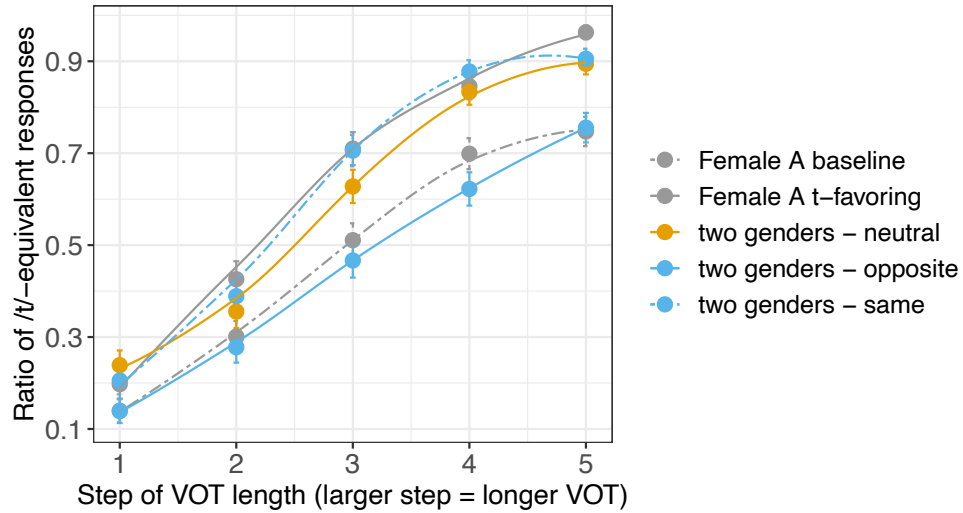


Figure 5.5: Exp 3a: /t/ response rate as a result of cross-gender perceptual learning with different combinations of /t d/ production biases (mean and standard error)

Among the three two-phase learning conditions in Fig. 5.5, the Two genders - *same* condition exhibits the most /t/-equivalent responses and the Two genders - *opposite* condition exhibits the fewest /t/-equivalent responses. The /t/-equivalent responses of the Two genders - *neutral* condition lie between the above two conditions, meanwhile it stays closer to results of the Two genders - *same* condition and farther apart from those of the *opposite* condition. Since participants in all the three conditions have had identical exposure to Female A’s /t/-favoring speech in the first training phase, we can attribute the differences between experimental conditions to their training in the second phase with Male A. Indeed, the overall /t/ responses of the three two-gender learning conditions are consistent with the perceptual biases associated with Male A’s /t d/ sounds. The highest /t/ response rate occurs in the condition where Male A’s speech favors the perception of /t/, and the lowest /t/ response rate occurs where Male A’s speech favors the perception of /d/, and the results of the condition with no /t d/ from Male A’s lie in between. Such a pattern

essentially suggests that the VOT properties of Male A has been acquired and integrated in the perception of Female A's /t d/ sounds. This is consistent with an *update* hypothesis for multi-talker perceptual learning, which claims that the perceptual learning outcome of one speaker is integrated to update listeners' perceptual expectations of other speakers.

A logistic mixed effects model was conducted to predict the response in the test (T=0, D=1), with Step (1-5, scaled and centered), Trial (1-51, scaled and centered), and Condition (treatment coded, reference level: the baseline condition), as the fixed effects, Condition:Step and Condition:Trial as the interaction items, and Step by Subject and by Word as the random slopes. The fixed effects of the model are shown in Table 5.4.

Fixed Effects	Estimate	SE	z value	Pr(> z)
(Intercept)	0.09	0.39	0.22	0.83
Step	-1.39	0.27	-5.12	< 0.001***
Condition Female A t-favoring	-1.17	0.34	-3.44	< 0.001***
Condition Two genders - <i>neutral</i>	-0.75	0.33	-2.28	0.02*
Condition Two genders - <i>opposite</i>	0.22	0.33	0.67	0.51
Condition Two genders - <i>same</i>	-1.06	0.33	-3.20	0.001**
Trial	-0.16	0.08	-1.85	0.06
Step:ConditionFemale A t-favoring	-0.67	0.24	-2.73	0.006**
Step:Condition Two genders - <i>neutral</i>	-0.30	0.23	-1.30	0.19
Step:Condition Two genders - <i>opposite</i>	0.02	0.23	0.09	0.93
Step:Condition Two genders - <i>same</i>	-0.55	0.24	-2.31	0.02*
ConditionFemale A t-favoring:Trial	0.15	0.13	1.14	0.25
Condition Two genders - <i>neutral</i> :Trial	0.08	0.12	0.64	0.52
Condition Two genders - <i>opposite</i> :Trial	-0.12	0.12	-0.97	0.33
Condition Two genders - <i>same</i> :Trial	-0.01	0.13	-0.07	0.95

Model-3a: Response~Step*Group+Group*Trial+(Step|Subj)+(Step|Word)

Table 5.4: The fixed effects of the logistic mixed-effects model in Exp 3a

Table 5.4 reveals a significant main effect of Step ($\beta = -1.39, p < 0.001$). This means that, in the baseline condition, segments with longer VOT steps are less likely to be perceived as /d/. The Condition effect is significant for the conditions of Female A /t/-favoring ($\beta = -1.17, p < 0.001$), Two genders - *neutral* ($\beta = -0.75, p = 0.02$), and Two genders - *same* ($\beta = -1.06, p = 0.001$), but not for the Two genders - *opposite* condition ($\beta = 0.11, p = 0.51$). In other words, compared to the baseline condition, the Two genders - *same* condition and the Two genders - *neutral* condition show significantly fewer /d/

responses and more /t/ responses. In contrast, the difference between the baseline condition and the Two genders - *opposite* condition is not significant, suggesting that the opposite perceptual learning with different speakers have cancelled each other out for listeners in the Two genders - *opposite* condition so that they are not different from the baseline condition anymore.

The interaction between Step and Condition is significant for the Female A /t/-favoring condition ($\beta = -0.67, p = 0.006$) and the Two genders - *same* condition ($\beta = -0.55, p = 0.02$), but not for the conditions of Two genders - *opposite* ($\beta = 0.02, p = 0.93$) or *neutral* ($\beta = -0.30, p = 0.19$). Again, this means that each step of increase in VOT causes a larger amount of increase in /t/ responses in the Female A /t/-favoring condition and the Two genders - *same* condition than in the baseline condition, whereas such a difference from the baseline condition does not occur to the Two genders - *opposite* or *neutral* conditions. Regarding effects relevant to Trial, no significant main effect or significant Trial:Condition interactions has been found in Exp 3a, which is different from what we have observed in Exp 1 and 2.

To further check whether the second-phase exposure to Male A's speech has shifted listeners' perceptual boundaries further away from the Female A /t/-favoring condition, I reran the model with the Female A /t/-favoring condition as the reference level of Condition. The result shows that among the three two-gender conditions, only the *opposite* condition exhibits a significant difference from the Female A /t/-favoring condition ($\beta = 1.39, p = 0.007$). A second-phase training with Male A's /t/-favoring speech or /t d/-free speech does not seem to induce any additional perceptual shifts on top of the shift towards /t/ induced by the first-phase training ($\beta_{same} = 0.11, p = 0.75; \beta_{neutral} = 0.42, p = 0.22$).

5.3.2.4 Summary

The three conditions of Exp 3a essentially differs in the directions of perceptual biases associated with the speech of a male speaker in an intervening training phase. The question is whether the perceptual learning outcome of this male speaker would be applied to the

perceptual categorization of a female speaker’s speech. The final categorization results of the three experimental conditions differ from one another in a consistent way with the male speaker’s production characteristics in their conditions. These results are not consistent with a *reset* account claiming that encountering a different speaker’s voice makes listeners set aside any perceptual expectations they have established previously, because this is not observed in the Two genders - *same* and *neutral* conditions. Instead, these results lend support to an *update* hypothesis that integrates the perceptual learning outcomes of both Female A and Male A into the update of perceptual expectations.

So far, we have demonstrated the influence of perceptual learning with Male A in the second phase on the final categorization results. The next question that comes up is whether or to what degree the perceptual learning with Female A in the first stage matters to the final categorization results. This question is to be evaluated in Exp 3b in the next section.

5.3.3 Exp 3b: No previous training with Female A’s /t-d/

5.3.3.1 Experimental conditions and goals

In Exp 3a, listeners have been trained on Female A’s and Male A’s /t d/ productions in two sequential training phases, and then they complete a categorization test that evaluates which perceptual expectation(s) they would apply to the identification of Female A’s /t d/. Results obtained so far indicate that listeners have generalized their knowledge about Male A’s speech properties to the categorization of Female A’s speech. In Exp 3b, I ask whether the results of the final categorization stage can be attributed to perceptual learning with Male A alone instead of the combination of perceptual learning outcomes with Female A and Male A. These two possibilities correspond to the two kinds of *update* mechanisms that I discussed in Section 7.1, namely, *recency* update and *cumulative* update. These two possibilities can be teased apart by comparing the perceptual shift in the two-gender conditions to the shift induced by the perceptual learning of Male A’s speech alone.

Among the three two-gender conditions in Exp 3a, the *neutral* condition lends the strongest support to the integration of perceptual learning with Female A into the final

categorization results; otherwise, it is hard to explain where the /t/-biased perceptual shift comes from. In contrast, the results of the Two genders - *same* condition and those of the *opposite* condition can be reasonably addressed either under a *recency* update account or under a *cumulative* update account. Exp 3b focuses on unpacking the result of the Two genders - *same* condition, asking how much of this result depends on the prior learning of Female A's speech. Participants first complete a /t/-favoring perceptual learning phase with male A and are then tested with /t d/ minimal pairs from Female A. By a comparison between the Male A /t/-favoring and the Two genders - *same* condition in Exp 3a, we will be able to tell whether whether training with Male A's /t/-favoring speech on its own gives rise to the same categorization boundary as induced by the /t/-favoring speech of both Female A and Male A.

5.3.3.2 Participants

An additional group of 29 participants are recruited from Prolific to participate in the Male A /t/-favoring condition in Exp 3b. They are 16 female and 13 male, aging from 18 to 67 years old ($Mean = 31.9, SD = 12.3$).

5.3.3.3 Result

Fig. 5.6 shows the planned comparison of the results in the Male A /t/-favoring condition (yellow) and the Two genders - *same* condition (blue). Crucially, a gap is shown between the Male A /t/-favoring condition and the Two genders - *same* condition. This means that, without previous training with Female A's /t/-favoring speech, training with Male A's /t/-favoring speech alone cannot induce as large a perceptual shift as shown in the Two genders - *same* condition.

The results of Fig. 5.6 suggest that, without the /t/-favoring training with Female A in the first place, the perceptual shift induced by Male A end up being much smaller in magnitude. If the difference between the Male A /t/-favoring condition and the Two genders - *same* condition is significant, then this pattern lends support to the account of

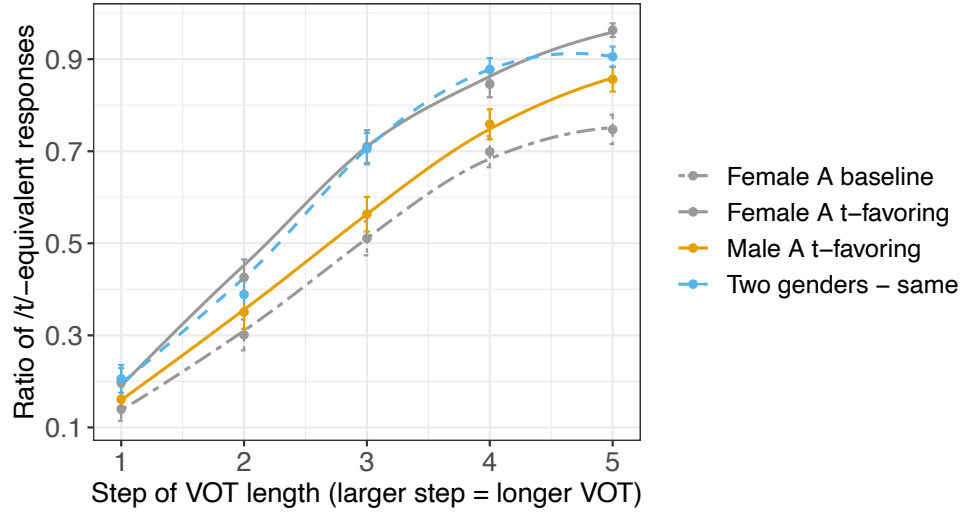


Figure 5.6: Exp 3b: /t/ response rate as a result of training with Male A compared to the two-gender condition and the baseline condition (mean and standard error)

cumulative update where the distribution of both Female A and Male A were exerting an influence on the test phase. A logistic mixed-effects model (Model-3b) is evaluated to predict the response of the test ($T=0$, $D=1$). It includes Step (1-5, scaled and centered), Trial (1-51, scaled and centered), and Condition (treatment coded, reference level: Male A /t/-favoring), as the fixed effects, Condition:Step and Condition:Trial as the interaction items, and Step by Subject and by Word as the random slopes. To evaluate whether the result of the Male A /t/-favoring condition is significantly different from those of the other three conditions, the model takes the Male A /t/-favoring condition as the reference level of the Condition variable. The fixed effects of the model are shown in Table 5.5.

The model reveals a significant Step effect ($\beta = -1.65, p < 0.001$), meaning that the number of /d/-equivalent responses decreases as the VOT of the stop becomes longer in the Male A /t/-favoring condition. The model also revealed a significant Condition effect for the Female A /t/-favoring condition ($\beta = -0.78, p = 0.03$) and a marginal significant Condition effect for the Two genders - *same* condition ($\beta = -0.67, p = 0.05$). The near-significant difference between the results of the Two genders - *same* condition and those of the Male A /t/-favoring condition lends some support to cumulative influences of perceptual learning from the two training stages in the Two genders - *same* condition.

Fixed Effects	Estimate	SE	z value	Pr(> z)
(Intercept)	-0.35	0.41	-0.84	0.40
Step	-1.65	0.29	-5.73	< 0.001***
Condition Female A baseline	0.42	0.34	1.23	0.22
Condition Female A t-favoring	-0.78	0.35	-2.20	0.03*
Condition Two genders - same	-0.67	0.34	-1.94	0.05
Trial	-0.05	0.10	-0.50	0.62
Step:Condition Female A baseline	0.26	0.22	1.15	0.25
Step:Condition Female A t-favoring	-0.39	0.24	-1.61	0.11
Step:Condition Two genders - same	-0.29	0.23	-1.25	0.21
Condition Female A baseline:Trial	-0.14	0.13	-1.12	0.26
Condition Female A t-favoring:Trial	0.01	0.14	0.10	0.92
Condition Two genders - same:Trial	-0.13	0.14	-0.91	0.36

Model-3b: Response~Step*Condition+Condition*Trial+(Step|Subj)+(Step|Word)

Table 5.5: The fixed effects of the logistic mixed-effects model in Exp 3b

An unexpected result that we can see in Table 5.5 is that the Condition effect does not turn out to be significant for the Female A baseline condition ($\beta = 0.42, p = 0.22$). In other words, training with Male A’s /t/-favoring speech does not seem to successfully induce a significant perceptual shift compared to the baseline condition. This raises some question regarding whether the Two genders - *same* condition is essentially identical to the Two genders - *neutral* condition in this experiment, in that the training with Male A’s speech only means additional exposure to a different speaker’s voice without introducing additional perceptual shifts. To figure out this question, I then divided the result up into four sets based on the lexical frame of the test phase, and reran Model-3b (with the random slope of Word taken out) again for each of the four sets of results. This time, I coded “Female A baseline” as the reference level for Condition, because it could tell us not only whether training with Male A’s /t/-favoring speech has induced a significant shift from the baseline, but also whether shifts have been induced by training with Female A’s /t/-favoring speech as well as by training with both of them, when we look at individual lexical frames of the test stimuli. Table 5.6 provides a summary of the Condition estimates in the four models. A full list of the fixed effects of each of these models (Model-3b-*tear*, Model-3b-*time*, Model-3b-*touch*, and Model-3b-*town*) can be found in Table B.5, B.6, B.8, and B.7 in

Appendix B.

(Condition ref: Female A baseline)	tear-dear	time-dime	town-down	touch-Dutch
Condition Male A t-favoring	-1.42*	-0.33	0.12	-0.52
Condition Female A t-favoring	-3.45***	-0.93	-1.27**	-1.00*
Condition Two genders - same	-1.48**	-0.64	-0.90*	-1.61***

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5.6: Condition estimates of the four logistic mixed-effects models for responses obtained with different lexical frames in Exp 3b

The results in Table 5.6 suggest that the occurrences and significance levels of perceptual shifts induced in different conditions also vary with the test stimuli. Based on the responses to “tear-dear” identification questions, a significant shift towards the intended direction occurs both in the Male A /t/-favoring condition ($\beta = -1.42, p < 0.05$) and in the Female A /t/-favoring condition ($\beta = -3.45, p < 0.001$). In contrast, responses to “town-down” or “touch-Dutch” identification questions only show a significant perceptual shift in conditions involving training with Female A’s speech, but not in the Male A /t/-favoring condition ($\beta_{town} = 0.12, \beta_{touch} = -0.52, p > 0.05$ in both cases). Finally, no perceptual shift shows up at all in identification responses of “time-dime” in any of the three training conditions under question. To better visualize the difference between words, I separately present the results obtained with “tear-dear” alternatives and those with alternatives of “town-down” and “touch-Dutch”, which is shown in Fig 5.7. I do not include a visualization of responses to the “time-dime” identification, because no effect has been found in any perceptual learning conditions with this set of test stimuli according to Table 5.6.

In the left facet of Fig 5.7, which presents identification responses of “tear-dear” identification, the results of the Male A /t/-favoring condition almost overlaps with those of the Two genders - *same* condition. At this point, the results of “tear-dear” identification cannot help us disentangle the possibilities of *cumulative* update and *recency* update. One may either argue that the /t/-favoring perceptual learning with Female A decays and the results of the Two genders - *same* condition essentially reflect that of the perceptual learning of Male A /t/-favoring speech. Alternatively, one may argue that the perceptual

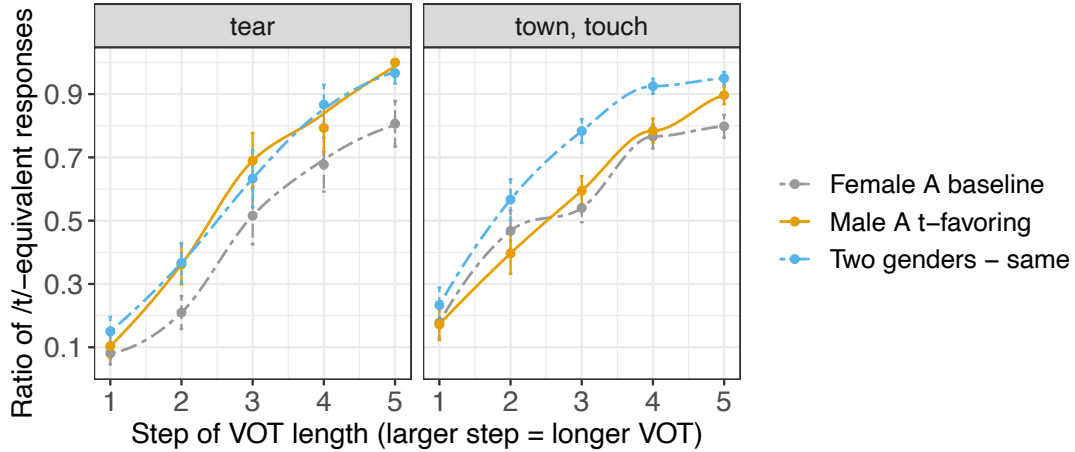


Figure 5.7: Exp 3b: /t/ response rate obtained with different minimal-pair test stimuli in the Female A baseline condition, the Male A /t/-favoring condition, and the Two genders - same condition (mean and standard error)

learning effect of Female A and Male A coexist, but they do not induce an additive amount of perceptual shift in the Two genders - *same* condition because of a ceiling constraint on the perceptual shift towards the /t/ end. Crucially, the results of “touch-Dutch” and “town-down” identification in the right facet shows a different pattern, where the results of the Male A /t/-favoring condition by and large overlaps with those of the Female A baseline condition rather than the Two genders - *same* condition. Since the training with Male A’s /t/-favoring speech by itself does not induce perceptual shifts to the perception of “town-down” and “touch-Dutch”, it is unlikely that it suddenly yields a significant perceptual shift in the Two genders - *same* condition. Like the results of the Two genders - *neutral* condition, the “town-down” and “touch-Dutch” identification results in the Male A /t/-favoring condition lend support to the persistence of the perceptual learning effect with Female A’s /t/-favoring speech.

5.3.3.4 Summary

Exp 3c evaluates the perceptual learning of Male A’s /t/-favoring speech alone and compares it with the results of the Two genders - *same* condition that involve /t/-favoring training with both Female A’s and Male A’s speech. The results show a clear difference between

two-gender training conditions and Male A only conditions. These results lend support to an account of *cumulative update*, where the speech of both Female A and Male A exerts an influence in the test phase. They are not consistent with the other possibility of *recency update*, which would otherwise predict little difference between the male-only conditions and their corresponding two-gender conditions.

5.4 Discussion

Experiment 3 is aimed at teasing apart the four possible mechanisms (*retention*, *reset*, *cumulative update*, and *recency update*) underpinning the perceptual learning of /t-d/ with multiple speakers. To address this question, I manipulated the phonetic characteristics of the /t d/ productions of two talkers – Female A and Male A – in the training phase. I then evaluated the perceptual consequences caused by each condition of manipulation with the same test phase. Experiment 3a compares the influences of three different perceptual learning conditions on the categorization of Female A’s /t-d/ continuum. The three training conditions each contain a training phase with Female A’s speech and a second training phase with Male A’s speech. The critical design lies in the acoustic properties of Male A’s /t d/ productions in each condition: These conditions are designed to induce a perceptual bias towards either /t/ or /d/, or they contain no /t d/ sounds and cause no perceptual bias. A comparison between the outcomes of the three two-gender perceptual learning conditions suggests that manipulating Male A’s /t d/ productions makes a difference to the final categorization results of Female A’s /t d/. Moreover, the directions of the perceptual shift exhibited in different conditions are consistent with Male A’s acoustic properties in the second training stage of those conditions: Participants who received /t/-favoring training with Male A show the highest /t/ response rate among the three conditions, and participants trained with /d/-favoring Male A showed the lowest /t/ response rate. These observations are not consistent with a *retention* (or speaker specificity) account that says listeners update their perceptual expectations in a speaker-specific way. Besides, participants who have not had exposure to Male A’s /t/ or /d/ sounds show a similar pattern with those in the Female

A /t/-favoring condition, meaning that additional exposure to 51 spoken words from Male A does not wipe out the outcome of previous perceptual learning with Female A. The results of the Two genders - *neutral* condition and the Two genders - *same* condition both reject a *reset* hypothesis that says that listeners set aside existing perceptual expectations they have established once they encounter the speech of a different speaker. This result lends support to the *update* account, according to which, listeners almost always update their phonetic expectations in response to the recent acoustic input, regardless of speaker specificity or gender.

Exp 3b is aimed at disentangling a confound involved in the interpretation of this result, namely, whether the categorization results can be largely attributed to the perceptual learning with Male A's speech alone, as opposed to a combination of the perceptual learning outcomes of the two training phases. Put differently, how much of the perceptual learning outcome of the first training phase with Female A has been maintained and reflected in the final result? Exp 3b addresses this question by taking out the first /t/-favoring learning phase with Female A and comparing the result of a Male A /t/-favoring condition with that of the Two genders - *same* condition. The results show that the shift induced by Male A's /t/-favoring stimuli alone is not as large as the shift induced by the /t/-favoring stimuli of both Female A and Male A. Moreover, the shift in the Male A /t/-favoring condition does not even turn out to be statistically significant compared to the baseline condition. I further break down the results according to the word frame of the test stimuli that participants responded to in their final categorization test. The results by test word suggest that the occurrence of perceptual shifts also depends on the lexical frame of the test stimuli. This is unexpected but not surprising, because the five steps of test stimuli in different lexical frames are manipulated to have different VOT values and span across different ranges (see Table 5.1 in Section 5.2). The five steps of stimuli that have the longest VOT and span across the widest VOT range have shown perceptual shifts in response to the training of the Female A /t/-favoring condition, the Male /t/-favoring condition, and the Two genders - *same* condition. For stimuli in the lexical frame of "town" and "touch"; however, a

perceptual shift only occurs in the Female A /t/-favoring condition and the Two genders - *same* condition, but not in the Male A /t/-favoring condition. We conclude that the null-effect in the latter case lends support to the critical role of exposure to Female A's /t/-favoring stimuli in the first phase, without which, the perceptual shift in the Two genders - *same* condition would not be significant either. Taken together, these findings support an *cumulative update* account of multi-speaker perceptual learning, which predicts that the experimental outcomes in the two-phase learning conditions indeed reflect the integration of the speech properties of both Female A and Male A.

Chapter 6

Exp 4: Comparing Speaker Effects on the Perceptual Learning of /s-f/ and /t-d/ within and across Genders

This chapter brings together the investigations of effects of speaker, gender and phoneme type in multi-speaker perceptual learning. I will be reporting on Experiment 4, which introduces visual cues to index speaker identity and compares potential constraints of speaker identity and speaker gender on the magnitude of perceptual generalization for both stops and sibilants. This chapter contains four sections. Section 6.1 recapitulate the effects of speaker specificity, speaker gender, and phoneme type on multi-talker perceptual learning in previous literature and lays out the fundamental research questions and hypotheses of this experiment. Section 6.2 provides an overview of the methodology, including experimental conditions, subjects, stimulus manipulation, and brief analysis of subjects' perception of speakers' identity and gender involved in the experiment. Section 6.3 reports on the result of two sub-experiments, which respectively evaluate the speaker and gender effects on the perceptual learning of /s-f/ and /t d/, by comparing the perceptual learning outcomes in different social-indexing experimental conditions. Section 6.4 summarizes the main findings of this experiment, compares the two sets of results obtained with different types of phonemes, and discusses their implications for the sociophonetic talker structure that underpins the tracking and generalization of perceptual learning across speakers.

6.1 Background and research question

6.1.1 Speaker, gender, and phoneme type

Previous findings about the effects of speaker identity, speaker gender, and phoneme type on multi-talker perceptual learning has been described in great detail in previous chapters (see, e.g., Sec. 1.3 in Ch. 1, Sec. 3.1 in Ch. 3, Sec. 5.1 in Ch. 5). This section brings them all together and highlights the missing links between these lines of studies that motivate the current experiment.

As a brief recap, previous studies have claimed that listeners make speaker-specific perceptual adjustments based on the unique speaker’s speech that triggers the adjustment (e.g., Eisner and McQueen, 2005; Kraljic and Samuel, 2005). However, they have also found evidence that listeners generalize their adjusted perceptual criteria to novel speakers that they have never heard (e.g., Kraljic and Samuel, 2006, 2007; Reinisch and Holt, 2014; Xie et al., 2018). Furthermore, the generalization of perceptual learning across speakers seems to vary with different types of phonemes. For example, it is reported that perceptual generalization across speakers of different genders is allowed on stop voicing indexed by VOT but inhibited on the place of articulation of fricatives as signaled by spectrum energy (Kraljic and Samuel, 2007). The different perceptual generalization behaviors of stops and sibilants raise an intriguing possibility that perceptual generalization behaviors across speakers reflect listeners’ sociophonetic knowledge that mirrors the structure of real-world speaker variability, where fricatives contain more information about speaker identity than stops (e.g., Kleinschmidt, 2017; Kraljic and Samuel, 2007).

The above proposal, interesting as it is, still has several weak ties of empirical evidence. One constraint lies in that studies of multi-talker perceptual learning make prevailing use of voices of different genders to represent different speakers (Eisner and McQueen, 2005; Kraljic and Samuel, 2005; Munson, 2011, etc.). Moreover, the evidence for speaker-specificity in the perceptual learning literature comes exclusively from cross-gender pairings. To the best of my knowledge, the literature does not contain a single case where generalization failed

between talkers of the same gender with the paradigm that examines perceptual generalization through perceptual shifts. One of the crucial questions remaining unclear here, then, is whether the difference between speakers needs to be salient enough to introduce speaker-specific perceptual learning. Another weak link in the argument for different speaker effects on the perceptual learning of fricatives and stops is that the comparison between phoneme types usually comes up as a post-hoc explanation rather than a hypothesis in relevant experiments. Except for Kraljic and Samuel (2007), no study has included the factor of phoneme types as a deliberate design of the experiment. In a nutshell, perceptual learning with conditions of different speaker pairings and phoneme pairings needs more comprehensive investigation.

6.1.2 Qualitative vs. quantitative speaker effect

In Section 1.3 of Ch 1, I have proposed two ways in which speaker specificity and gender may constrain the generalization of perceptual learning across speakers, which I refer to as qualitative and quantitative speaker effects. Qualitative speaker effects impose a categorical constraint on the occurrence of perceptual generalization across speakers, and quantitative speaker effects impose a gradient constraint on the magnitude of perceptual learning generalization. Following these two dimensions, a comparison between sibilants and stops regarding their susceptibility to speaker effects can be achieved by asking two kinds of questions.

One kind of question to investigate is the *categorical* constraints imposed by speaker structure on the perceptual generalization across talkers. Suppose we found that perceptual learning is talker-specific for sibilants but generalizes across talkers for stops, or that it is gender-specific for sibilants but can be generalized across talker genders for stops. In that case, the findings lend support to the phoneme type difference that sibilants are more affected by the social aspects of speakers than stops. This is also the kind of evidence that previous studies observed to suggest that sibilants are more sensitive to speaker properties than stops in terms of perceptual generalization.

However, the empirical results of Exp 1 and Exp 3 in this dissertation raise an issue with this kind of evaluation, because we have not observed much of a categorical alternation between generalization and failure to generalize either with stops or with fricatives. Instead, what we observed is that perceptual learning constantly generalizes across speakers of different gender groups for both sibilants and stops. If the situation of constant generalization also occurs in Exp 4, then we need to investigate a different kind of question to compare the susceptibility of stops and sibilants to speaker effects in perceptual learning, that is, whether the magnitude of perceptual generalization differ across social indexing conditions. The influence of speaker specificity is reflected by the extent to which the perceptual learning result reflects the speech properties of the specific speaker being tested instead of a cumulative learning output of all speakers in general.

6.1.3 Research question and hypotheses

In the light of the broad goal of investigating the constraints of speaker identity and speaker gender on the generalization of perceptual learning with different types of phonemes, Exp 4 aims at examining two hypotheses with respect to the perceptual learning of /s-ʃ/ and /t-d/. The first hypothesis is that listeners use voice gender cues to inhibit perceptual generalization across speakers of different genders compared to speakers of the same gender. The second hypothesis is that listeners use speaker identity cues to inhibit perceptual generalization across different speakers within the same gender. Note that these two hypotheses are not competing, but rather logically independent. The difference between these two possibilities requires comparing within-gender and cross-gender conditions.

As with Exp 1-3, Exp 4 also contains two training phases each with a different speaker (A, B) and a test phase of phoneme categorization with the speaker in the first training phase (A). In this experiment, Speaker A and B always exhibit opposite acoustic characteristics for the phonemes in question. Three experimental conditions are designed to contrast in the available social indexical cues to speaker identity and gender. They are a *female-auditory* condition, where listeners hear female-sounding voices of both speakers, a *gender-auditory*

condition, where listeners hear a female-sounding voice and a male-sounding voice that distinguish the two speakers, and a female - visual condition, where listeners hear female-sounding voices of both speakers with each voice associated with a unique photo of that speaker.

The two comparison analyses are planned to evaluate the two hypotheses of the constraints of speaker gender and speaker identity. To evaluate the constraint of speaker gender, I plan to compare the perceptual learning results of the *female-auditory* condition and the *gender-auditory* condition. If the hypothesis is true, then it predicts that the categorization results of the *gender-auditory* condition reflect more of speaker A's speech characteristics and less of speaker B's compared to the *female-auditory* condition, because A and B are of different genders in the *gender-auditory* condition and of the same gender in the *female-auditory* condition. To evaluate the constraint of speaker identity, I plan to compare the perceptual learning results of the *female-auditory* condition and the *female-audiovisual* condition. If the hypothesis is true, it predicts that the categorization results of the *female-audiovisual* condition reflect more of speaker A's speech characteristics and less of speaker B's compared to the *female-auditory* condition, because the enhanced cues to speaker identity in the *female-audiovisual* condition inhibits the generalization of the perceptual learning with a different speaker in the second phase and enhances the retention of perceptual learning in the first phase.

6.2 Method overview

6.2.1 Experimental conditions

Experiment 4 contains two sub-experiments that work in parallel to each other to investigate the perceptual generalization of /s-f/ (Exp 4a) and /t-d/ (Exp 4b) with multiple speakers either from the same gender group or from different gender groups. Fig. 6.1 shows a summary of the experimental designs and procedures in each condition of the different sub-experiments in Experiment 4. We can see that the two sub-experiments of Exp 4 each

have three conditions, and participants in each of these conditions need to complete two training phases respectively with Female A’s and Female B’s speech and a test phase of phoneme identification with Female A’s speech. The two training phases, which I refer to as the “prior” training phase and the “recent” training phase according to the order by which they take place, are always designed to induce perceptual biases that work against one another. In Exp 4a, Female A’s speech is manipulated to favor the perception of /s/ in the prior training phase, whereas Female B’s speech favors the perception of /j/ in the recent training phase. Similarly, in Exp 4b, Female A’s speech is perceptually /t/-favoring while Female B’s speech is perceptually /d/-favoring in the two training phases. Finally, the test phase evaluates the perceptual categorization of the two phonemes under investigation (/s-j/ in Exp 4a and /t-d/ in Exp 4b) on a synthesized continuum spliced into Female A’s spoken words.

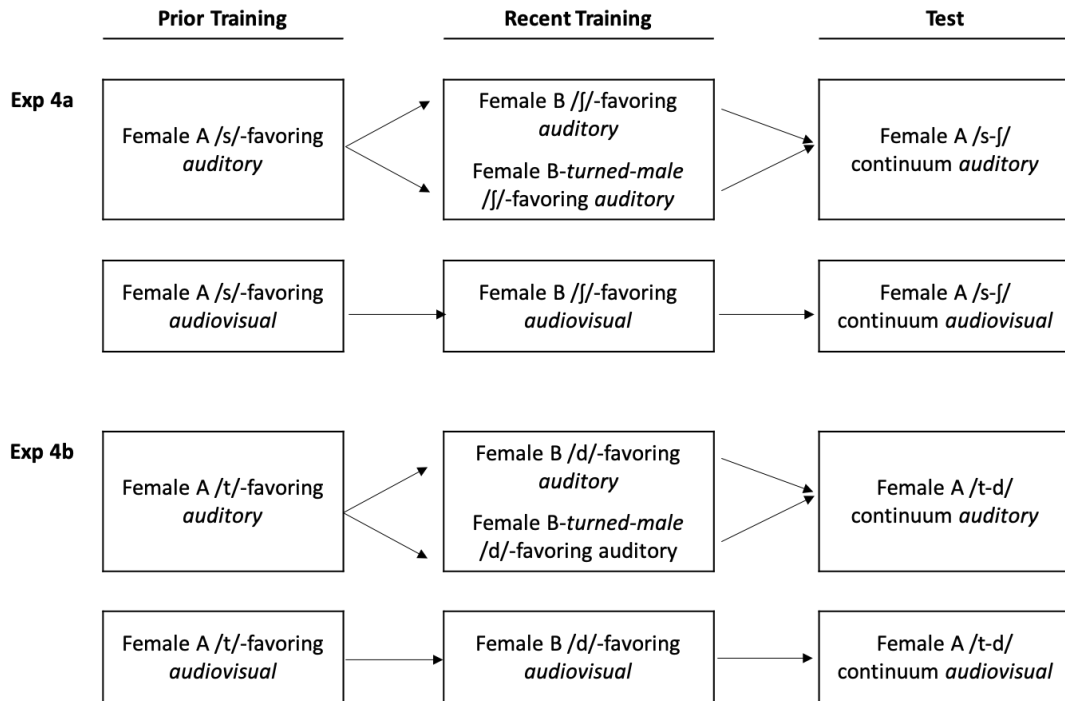


Figure 6.1: The structure of sub-experiments and conditions in Exp 4

The three conditions in each sub-experiment include a *female-auditory* condition, a *female-audiovisual* condition, and a *gender-auditory* condition. They all follow the setup

of two training phases with Female A and B that point to opposite perceptual biases and a test phase with Female A as described above. Since the perceptual biases triggered by the two training phases are at odds with one another, the degree to which listeners generalize their learning of Female B’s speech to Female A can be quantified by how much the final categorization output patterns with or deviates from the result of training with Female A in the first phase alone. The more that the categorization results reflect the bias of the first learning phase, the less that listeners generalize their perceptual learning with Female B to the categorization of Female A’s speech.

Meanwhile, different social indexing cues are available in each condition provides participants with different amount of information about the speakers. Listeners in the *female-auditory* condition (the uppermost one in each sub-experiment in Fig. 6.1) hear Female A’s voice in two phases (first, third) and Female B’s voice in one phase (second). No other cues or descriptions are presented to inform listeners of the number of speakers involved or their social characteristics. Listeners need to infer these pieces of information from the voices alone, and thus their inference may or may not be accurate. By contrast, participants in the *female-audiovisual* condition (the bottom one in each sub-experiment in Fig. 6.1) are exposed to the same set of spoken words from Female A and Female B as those in the *female-auditory* condition, but each voice is associated with a photo of a unique speaker, which co-occurs with the individual spoken words in each trial. I expect this design to implicate that there are two female speakers involved in the experiment, and that speakers in the first and third phases are the same person. Finally, participants in the *gender-auditory* condition (the middle one in each sub-experiment in Fig. 6.1) are also exposed to training stimuli of both Female A and Female B, and test stimuli of Female A; however, stimuli of Female B are manipulated to be male-sounding in this condition. With successful voice gender manipulation, participants in the *gender-auditory* condition are supposed to hear a female voice in the first phase and the third phase and a male voice in the second phase during the experiment.

6.2.2 Stimuli

The stimuli used in Experiment 4 are manipulated from recordings of spoken words from Female A and Female B obtained following the procedure described in Chapter 2.

In Exp 4a, each training phase consists of 17 /s/-containing words, 17 /ʃ/-containing words, and 17 filler words without /s ʃ/. The test phase consists of 35 test trials synthesized by splicing five steps of /s-ʃ/ continuum into 7 lexical frames of minimal pairs with word-initial /s ʃ/. See Sec 3.2.2 for a full list of words used in the experiment. See Sec 3.2.3 for more details about the step selection and stimulus synthesis of Female A’s speech tokens and Sec 4.2.3 for more details about the step selection and synthesis of Female B’s tokens. In Exp 4b, each training phase consists of 17 /t/-containing words, 17 /d/-containing words, and 17 filler words without /t d/.

Tokens of Female B are manipulated to be male-sounding for the *gender-auditory* condition via a series of manipulation with Praat. These manipulations include scaling the vowel formant by 0.8 of its original values, lowering the F_0 median of Female B to 110 Hz, and compressing the F_0 range to 0.9 of its original range. The manipulation is implemented to the whole word for /t-d/ stimuli in Exp 4b, whereas it is implemented to the remainder of the word excluding the sibilants for /s-ʃ/ stimuli. This is to ensure the comparability of the crucial acoustic dimensions to be evaluated as a result of perceptual learning. In Exp 4a, the distinction between /s-ʃ/ is largely signaled by the spectrum energy of the sibilants, and therefore I excluded this proportion from the above manipulations such that the spectral properties of the sibilants are not changed in vowel formant scaling¹. In Exp 4b, however, the distinction between /t-d/ is largely signaled by the temporal feature of VOT length, which does not change after the stimulus manipulations. Therefore the stops are not spliced out and assembled back before and after the voice gender manipulation.

¹I also tried implementing voice gender manipulation to the whole /s-ʃ/ words including the sibilant proportions. This manipulation indeed ends lowering the COG values of the sibilants according to acoustic measures

6.2.3 Participant

178 participants are recruited from Prolific to attend Experiment 4. All of them are self-reported to be native American English speakers with normal hearing. Table 6.1 shows a breakdown of participants in each of the three conditions of the two sub-experiments, including the number of participants, their gender distribution based on self report, and the range, mean and standard deviation of the age of participants in each condition. In this dataset, we can see that participants in different conditions are fairly comparable in terms of gender and age.

	Condition	N	Gender			Age	
			F	M	else	range	mean (sd)
Exp. 4a	Female-auditory	28	10	18	0	18-66	30.2 (10.6)
	Gender-auditory	30	9	21	0	18-61	34.4 (11.8)
	Female-audiovisual	30	9	19	2	19-56	30.5 (10.6)
Exp. 4b	Female-auditory	32	14	18	0	18-67	36.9 (14.4)
	Gender-auditory	32	11	17	2	19-61	30.0 (10.2)
	Female-audiovisual	26	11	15	0	18-67	38.5 (15.6)

Table 6.1: The participant information in each condition in Exp 4

6.2.4 Speaker perception

At the end of the experiment, participants reported their perception of speaker identity and gender involved in the experiment by responding to two questions in a final survey. One question asks “How many voices have you heard in this experiment?” Participants are expected to type in an integer, but input of characters are also acceptable with the setting of the survey. The other question asks “What is the gender/are the genders of that voice/those voices?” Listeners respond by choosing from three alternatives – “female only”, “male only”, and “both female and male”.

Listeners responses to these two questions are reported in this section to provide a clear idea of what listeners think they have heard in different experimental conditions, which is crucial for accurately interpreting any potential differences observed across these conditions. Table 6.2 shows the breakdown of their responses to the question about speaker gender.

We can see that listeners reported “female only” 100% of the time in the *female-auditory* and *female-audiovisual* conditions in both Exp 4a and 4b. This suggests that Female A’s and Female B’s voices are generally associated with correct gender perception with their original F_0 and vowel formants. For the *gender-auditory* condition where Female B’s voice are manipulated to be male-sounding, listeners tended to report that they had heard voices of both genders. The only exception (marked with a star) is one participant in the *gender-auditory* condition in Exp 4a, who perceived the original gender of Female B’s voice instead of the intended gender by manipulation and chose “female only” for the two voices.

	Experiment 4a			Experiment 4b		
	female only	male only	both	female only	male only	both
Female-auditory	28	0	0	32	0	0
Female-audiovisual	30	0	0	26	0	0
Gender-auditory	1*	0	29	0	0	32

Table 6.2: Voice gender responses in each condition of Exp 4

In general, listeners’ perception of voice gender is quite consistent with the intended voice gender(s) in each experimental condition. Especially, these results have confirmed that Female B’s voice becomes fairly male-sounding after F_0 and vowel manipulation. The responses of the starred participant is removed in further analysis, so that we can interpret our results based on the premise that listeners in all conditions have successfully perceived the intended gender of the speaker by voice cues.

Regarding the perception of voice identity, Fig. 6.2 shows the distribution of the number of voices reported by listeners in different conditions in Exp 4a and 4b. We can see that in each experimental condition, the majority of participants (57%-92%) correctly reported that they had heard 2 voices throughout the experiment. The second most common category of answer is 3, which include not only straightforward integer responses but also variations such as “2 or 3”, “3?”, “2-3 depending on whether the speakers in the first phase and the third phase are the same person”, and the like. In general, choosing 3 generally reflects listeners’ uncertainty about whether the speech in the first phase and in the third phase comes from the same person. The categories of “4-6” and “50+” are also combinations of

answers falling into that range.

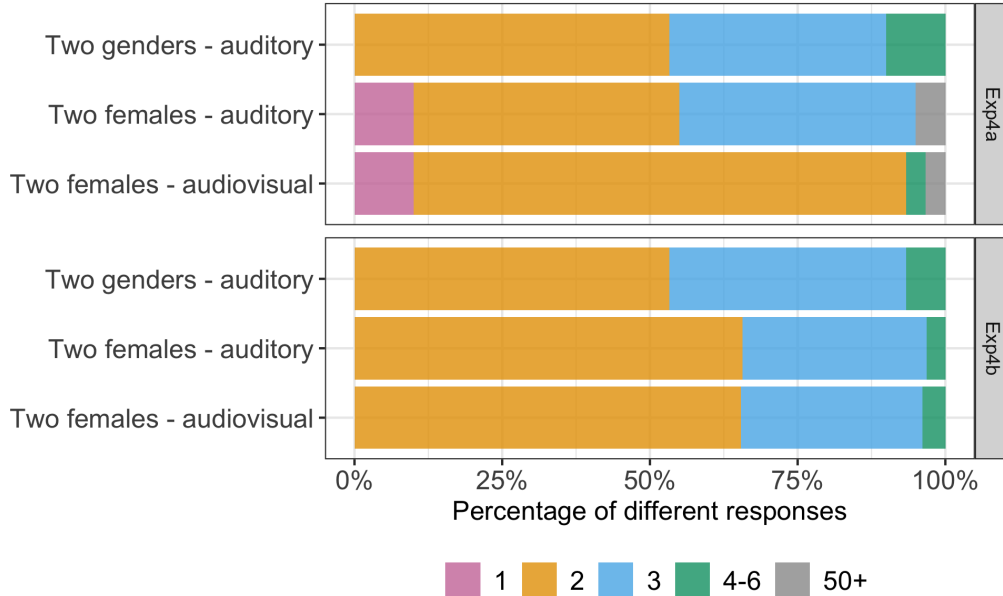


Figure 6.2: Voice number responses in each condition of Exp 4a and 4b

Compared to the *female-auditory* condition, additional manipulations of social indexical cues does make a difference to listeners' perception of speaker identity in the *female-audiovisual* condition and the *gender-auditory* condition. The influence of visual cues to speaker identity can be observed in Exp 4a, which shows that listeners who are presented with speaker photos are more likely to respond that they have heard two voices instead of three, presumably because seeing the same photo occurring in the first and third phase has convinced them that speech materials in these phases are produced by the same speaker. Oddly, this pattern does not come out in Exp 4b, where quite a few participants in the *female-audiovisual* condition still responded that they had heard three voices even though they had only seen two faces patterned with those voices throughout the experiment. Another effect of the social indexing manipulation is reflected by the difference between the *female-auditory* condition and the *gender-auditory* condition. In general, voices of different genders lead listeners to believe that there are more speakers in the experiment than an equal number of voices of the same gender. Two voice genders also help wipe out listeners' misperception that there is only a single speaker throughout the experiment, which does

occur in the *female-auditory* and *female-audiovisual* conditions in Exp 4a.

6.3 Experiment and Result

6.3.1 Exp 4a: Perceptual learning of /s-f/ with Female A and B

Exp 4a evaluates the perception learning of /s-f/ with two speakers in the three different social indexing conditions, namely, *female-auditory*, *female-audiovisual*, and *gender-auditory*. All participants in these conditions are exposed to Female A's /s/-favoring speech and Female B's /f/-favoring speech sequentially in the training phase, and they then complete a phoneme identification task with Female A's /s-f/ continuum. Therefore, more /s/-equivalent responses in the final phase would reflect more retention of the specific speaker's speech characteristics, whereas more /f/-equivalent responses would reflect more generalization of a different speakers' speech characters to the speech perception of the test speaker. I focus on two comparison analyses that respectively evaluate the hypotheses of an identity constraint and a gender constraint on the perceptual generalization across speakers, which are laid out in Section 6.1.3.

To evaluate whether listeners use voice gender cues to inhibit perceptual generalization across genders as opposed to across speakers within gender, I compare the categorization results of the *female-auditory* condition and the *gender-auditory* condition. If the *gender-auditory* condition has a higher amount of /s/-equivalent responses than the *female-auditory* condition, which suggests a higher level of retention of talker-specific characteristic traits, then it lends support to the hypothesis of a gender group constraint.

Similarly, to evaluate whether listeners use identity cues to inhibit perceptual generalization across different speakers within the same gender, I compare the result of the *female-auditory* condition and that of the *female-audiovisual* condition. If the *female-audiovisual* condition ends up having more /s/-equivalent responses, which mirrors the result of training with Female A's /s/-favoring speech alone, than the *female-auditory* condition, then that means the availability of visual cues to talker identity has inhibited perceptual generalization

across speakers.

6.3.1.1 Aggregate analysis

Fig. 6.3 shows the means and standard errors of the categorization result at each /s-ʃ/ step in the three two-phase perceptual learning conditions, along with the results of the baseline condition and the Female A /s/-favoring conditions represented by the grey lines.

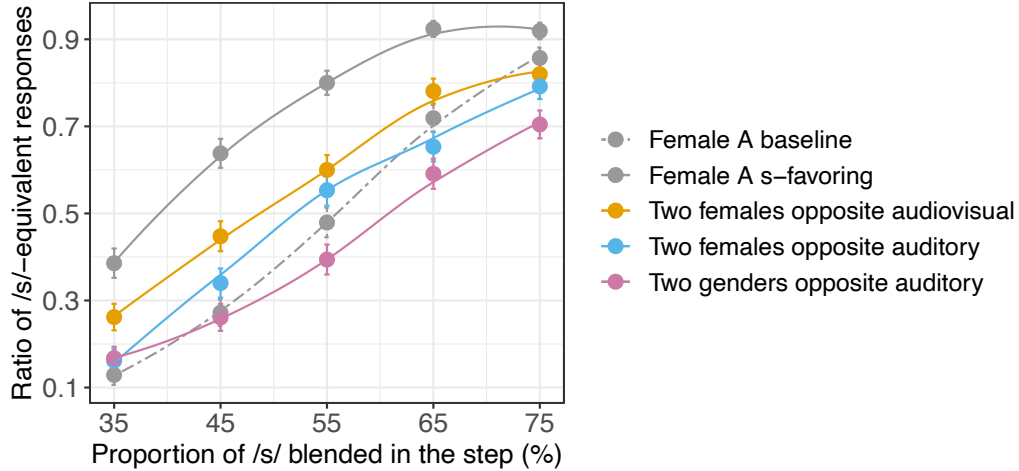


Figure 6.3: Exp 4a: /s/ response rate as a result of opposite perceptual learning in different social-indexing conditions (mean and standard error)

Recall that listeners in Exp 4a had received /s/-favoring training with Female A and /ʃ/-favoring training with Female B before the categorization task. Therefore, their categorization curves diverge from that of the Female A /s/-training condition and stay close to the baseline condition because the two opposite distributions of Female A's and Female B's sibilants canceled each other out to some extent. Other than that, we can see that the two-female audiovisual condition seems to have retained more /s/-equivalent responses than the *female-auditory* condition, revealed by the yellow line placed on top of the blue line. This is consistent with our earlier prediction that the photos of different talkers would provide enhancing cues to speaker specificity and allow listeners to retain more speaker-specific perceptual learning. In this case, listeners are informed by photo cues that the /ʃ/-favoring distribution in the second phase is associated with a different speaker. Therefore, in the

final categorization stage, they retain more of the /s/-favoring perceptual expectations induced by Female A. In contrast, we can see that the results of the *gender-auditory* condition show fewer /s/-equivalent responses than the Two females auditory condition, revealed by the plum line placed below the blue line. This is not expected because we predict that the different gender voices would serve as an additional cue to speaker specificity and allow listeners to retain more speaker-specific perceptual learning of Female A. I discuss the potential reasons for this pattern in the summary section in Sec 6.3.1.3.

A logistic mixed-effects regression model (Model-4a) is evaluated to predict the Response of each trial (S=0, SH=1). Since we are specifically wondering about the comparisons between *female-auditory* and *female-audiovisual*, as well as between *female-auditory* and *gender-auditory*, I code *female-auditory* as the reference level of the variable Condition. Model-4a includes Step (35-75, scaled and centered), Trial (1-51, scaled and centered), Condition (treatment coded, reference: *female-auditory*), and Phoneme (the original phoneme associated to each auditory frame, sum-coded, reference-level: SH) as the fixed effects, Condition:Step and Condition:Trial as the interaction items, Step by Subject, Phoneme by Subject, and Step by Word as random slopes. The model estimates of the fixed effects are shown in Table 6.3.

In Table 6.3, the Condition effect is only significant for the Female A s-favoring condition ($\beta = -1.98, p < 0.001$) but no other conditions. Especially, the Condition effect does not turn out significant for either the Two females opposite audiovisual condition ($\beta = -0.51, p = 0.29$) or the Two genders opposite auditory condition ($\beta = 0.75, p = 0.13$). Although the resulting pattern of the Two females audiovisual condition is consistent with my prediction, the availability of the visual cues to different talkers does not seem to make a statistical difference. I have also evaluated whether the results of the three two-phase perceptual learning conditions deviate significantly from the baseline condition, by re-running the model with “Female A baseline” as the reference level of the Condition variable. The results show that the *female-auditory* condition, the *female-audiovisual* condition, and the *gender-auditory* condition all pattern with the baseline condition as a result of two training

Fixed Effects	Est.	SE	z value	Pr(> z)
(Intercept)	-0.08	0.39	-0.20	0.85
Condition Female A baseline	0.19	0.48	0.39	0.70
Condition Female A s-favoring	-1.98	0.49	-4.02	< 0.001***
Condition Two females opposite audiovisual	-0.51	0.48	-1.05	0.29
Condition Two genders opposite auditory	0.75	0.49	1.52	0.13
Step	-1.65	0.20	-8.07	< 0.001***
Trial	-0.16	0.08	-1.89	0.06
PhonemeS	-0.54	0.04	-12.93	< 0.001***
Condition Female A baseline:Step	-0.58	0.28	-2.06	0.04*
Condition Female A s-favoring:Step	-0.13	0.29	-0.47	0.64
Condition Two females opposite audiovisual:Step	0.05	0.28	0.20	0.84
Condition Two genders opposite auditory:Step	0.01	0.28	0.03	0.98
Condition Female A baseline:Trial	-0.32	0.13	-2.51	0.01*
Condition Female A s-favoring:Trial	0.32	0.13	2.50	0.01*
Condition Two females opposite audiovisual:Trial	-0.13	0.12	-1.07	0.28
Condition Two genders opposite auditory:Trial	-0.26	0.12	-2.07	0.04*

Model-4a: Response~Step*Condition+Condition*Trial+Phoneme+(Step|Subj)+(Step|Word)

Table 6.3: The fixed effects of the logistic mixed-effects model in Exp 4a

stages inducing opposite perceptual biases².

In addition to Condition, Model-4a also reveals a number of other effects. As always, the main effect of Step is significant ($\beta = -1.65, p < 0.001$), meaning that each step of increase in mixed [s] results in a lower likelihood of perceiving /j/ by 1.65 in log scale for the Two Females audiovisual condition. Step shows significant interactions with the baseline condition ($\beta = -0.58, p = 0.04$) but not the other two two-phase training conditions in this experiment (*female-audiovisual*: $\beta = 0.05$, *gender-auditory*: $\beta = 0.01$, n.s.). This means that exposure to Female B’s sibilants gives rise to a shallower categorization line than the Female A baseline condition. The effects of Phoneme ($\beta = -0.54, p < 0.001$) is significant, meaning that listeners are less likely to report on /j/ for later trials in the *female-auditory* condition. The effect of Trial is marginally significant ($\beta = -0.16, p = 0.06$) for the *female-auditory* condition and it shows significant interaction with the Female A baseline condition ($\beta = -0.32, p = 0.01$), the Female A s-favoring condition ($\beta = 0.32, p = 0.01$), and the *gender-auditory* condition ($\beta = -0.26, p = 0.04$). These suggest that participants in the

²See a full list of the fixed effects of Model 4a-*relevel* in Table B.9 in Appendix B.

Female A s-favoring condition are more likely to report an /ʃ/ for test stimuli coming up later. By contrast, participants in the other four conditions are more likely to report an /s/, to different degrees, for test stimuli occurring in a later time phase.

6.3.1.2 Analysis by phoneme

As I have demonstrated in previous chapters, the test stimuli may differ regarding whether they are /s/- initial or /ʃ/-initial in their original productions (i.e., the variable of Phoneme), and the sibilant-associated transitional cues also give rise to perceptual biases that may interact with the results of perceptual learning. To examine whether such interaction exists in Exp 4a, I examined the results of the three experimental conditions for test stimuli produced with /s/ and /ʃ/ separately. Given the smaller number of observations after being split by Phoneme, I did not further distribute the observations over five sibilant steps; but rather, I examined the overall result of each experimental condition averaged across sibilant steps. Fig. 6.4 shows the means and 95% confidence intervals of the aggregate /s/ response rates by phoneme and experiment condition.

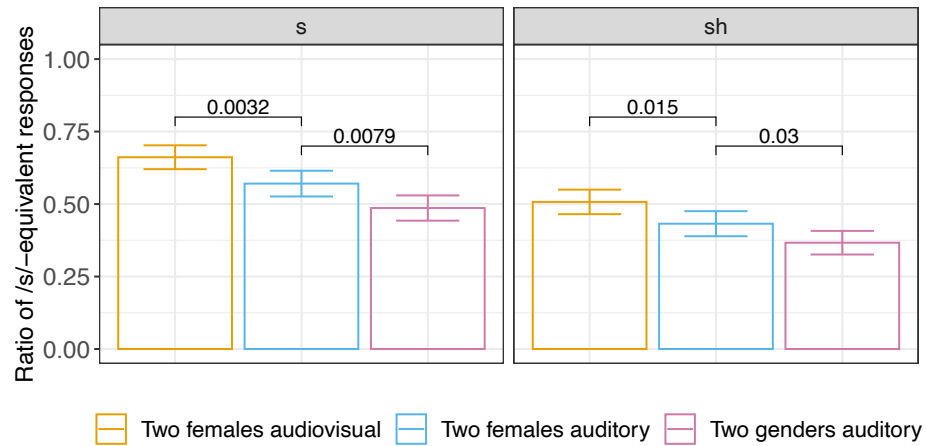


Figure 6.4: Exp 4a: /s/ response rate by condition and phoneme (mean and 95% confidence interval). Values above the comparison bars indicate the p-values of Wilcoxon rank sum tests.

The pattern exhibited in both of the two facets in Fig. 6.4 is consistent with the pattern in Fig. 6.3. All of them reveal the highest /s/ response rate in the *female-audiovisual* con-

dition, followed by lower rates in the *female-auditory* condition and the lowest /s/ response rate in the *gender-auditory* condition. A series of Wilcoxon rank sum tests are conducted to evaluate whether the /s/ response rate of the *female-auditory* condition is significantly different from those of the other two conditions respectively when the test stimuli start with an /s/ or an /ʃ/ before manipulation. As noted in Fig. 6.4 already, the between-condition differences under inquiry are each significant. In both of the two Phoneme conditions, the *female-audiovisual* condition has a significantly higher /s/ response rate than the Two females auditory condition significant (original /s/-initial stimuli: $W = 134346$, $p = 0.0032$; original /ʃ/-initial stimuli: $W = 147763$, $p = 0.015$), and the *gender-auditory* condition has a significantly lower /s/ response rate than the Two females auditory condition (original /s/-initial stimuli: $W = 112553$, $p = 0.0079$; original /ʃ/-initial stimuli: $W = 128421$, $p = 0.03$).

6.3.1.3 Summary

In a nutshell, through the planned comparisons between the *female-auditory* condition and the *female-audiovisual* condition and between the *female-auditory* and *gender-auditory* condition, we can see that the availability of different social indexical cues in these conditions makes a difference to the outcomes of multi-speaker perceptual learning. The availability of additional visual cues to the identity of the two female speakers shows enhanced perceptual generalization within the same speaker and weakened perceptual generalization across speakers, compared to the *female-auditory* condition. This is reflected by the result that the *gender-audiovisual* condition exhibits more /s/-equivalent responses that are consistent with the speech characteristics of the speaker in question. However, this difference is not statistically significant in the aggregate logistic regression model, which also accounts for other fixed effects such as Step and random effects such as Subject and Word. This difference does turn out significant according to the Wilcoxon rank-sum test, in which the results are grouped by experimental condition and the original phoneme of the test stimuli. This result has lent some weak evidence to an identity constraint of multi-talker perceptual

learning, which predicts that listeners make use of social cues of speaker identity to inhibit perceptual generalization across different speakers.

The second comparison to make is between the *female-auditory* condition and the *gender-auditory* condition, intended to evaluate whether listeners generalize more across speakers of the same gender and less across speakers of different genders. The results show that participants in the *gender-auditory* condition actually show more /ʃ/-biased shift that is consistent with Speaker B's speech characteristics instead of Speaker A's, who is also the speaker of the test phase. This is not consistent with our prediction that turning Speaker B's voice into a different gender will undermine the perceptual generalization of B's speech to the test speaker. Moreover, this pattern is quite robust because the difference between these two auditory conditions is statistically significant according to both the logistic mixed effect model and the Wilcoxon rank-sum tests. I argue that this unexpected pattern is caused by the interference of the speaker normalization mechanism, according to which altering the acoustic properties of the contextual materials causes different perceptions of the same piece of acoustic signals. In this experiment, I intentionally excluded the sibilant proportions from manipulation in order to keep them comparable across different conditions – because those sibilants in the *female-auditory* or audiovisual conditions have not undergone such manipulations. However, this decision gives rise to another problem. Since a /ʃ/-favoring training phase already involves standard /s/ pronunciations and ambiguous /ʃ/ pronunciations, the sibilants within that training phase are already /s/-sounding in general. In addition, I also lowered the formant frequencies of the contextual vowels, which also contributes to the /s/-sounding properties of sibilants in the second training phase of the *gender-auditory* condition. These additive /s/-sounding (/ʃ-favoring) effects might have given rise to a stronger perceptual learning effect to shift the perceptual boundary towards /ʃ/. That is why the *gender-auditory* condition exhibits the least /s/-equivalent responses in the test phase among all the conditions. More discussion about this will be provided in the Discussion section. Up to this point, I did not find evidence for the speaker gender constraint on the multi-talker perceptual learning of sibilants.

6.3.2 Exp 4b: Perceptual learning of /t d/ with Female A and B

6.3.2.1 Experimental conditions and goals

Experiment 4b shares the experimental structure of Experiment 4a and evaluates how talker identity and gender affect the magnitude of perceptual generalization across speakers for /t d/. Like Exp 4a, Exp 4b also has three experimental conditions: *female-auditory*, *female-audiovisual*, and *gender-auditory*. All participants in these conditions are exposed to Female A's /t/-favoring speech and Female B's /d/-favoring speech in two sequential training phases. In the end, they complete a phoneme categorization task along Female A's /t-d/ continuum. Different social indexing cues are presented in each condition to give participants different information about the speakers.

As with the results of Exp 4b, I also focus on two comparison analyses in order to evaluate the hypotheses of an identity constraint and a gender constraint on the perceptual generalization of /t d/ across talkers. To evaluate the identity constraint hypothesis, I compare the result of the *female-auditory* condition and that of the *female-audiovisual* condition. If listeners use identity cues to inhibit perceptual generalization across different speakers within gender, then it predicts that the *female-audiovisual* condition will show a higher /t/ response rate, in line with Female A's speech characteristics, than the *female-auditory* condition. Similarly, to evaluate the gender constraint hypothesis, I compare the categorization results of the *female-auditory* condition and the *gender-auditory* condition. If the perceptual generalization across speakers from the same gender group is enhanced compared to that across speakers from different gender groups, then we should expect to see a higher amount of /t/-equivalent responses in the *gender-auditory* condition than the *female-auditory* condition. This is because the influence of training with Female B is weakened for it is manipulated to have a different voice gender in the *gender-auditory* condition.

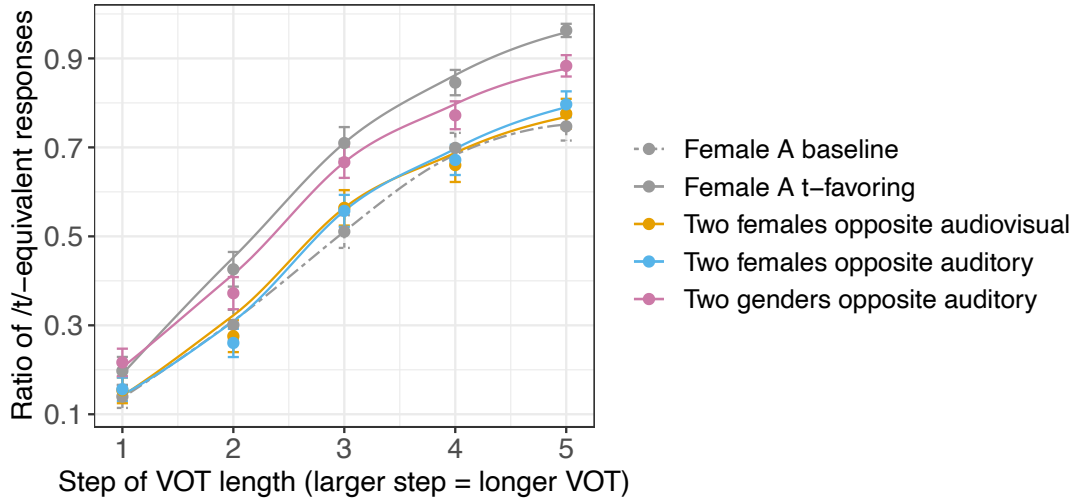


Figure 6.5: Exp 4b: /t/ response rate as a result of opposite perceptual learning in different social-indexing conditions (mean and standard error)

6.3.2.2 Aggregate analysis

Fig. 6.5 shows the means and standard errors of the /t-d/ categorization results at each VOT step in the three experimental conditions of Exp 4b, along with the results of the baseline condition and the Female A /t/-favoring conditions represented by the grey lines. Recall that listeners in the two Two Females conditions had received /t/-favoring training with Female A and /d/-favoring training with Female B before the categorization task. According to the figure, the second training phase with Female B has canceled out the learning outcome in the first perceptual learning phase with Female A, such that the categorization boundary goes back to the baseline condition for these conditions. In addition, Fig. 6.5 does not seem to reveal any difference between the Two Females auditory condition and the Two Females audiovisual condition, as reflected by the overlapping blue line and the yellow line. Listeners in the *female-audiovisual* condition did not show more perceptual bias towards /t/ than those in the *female-auditory* condition to reflect a heavier influence of the /t/-favoring training with Female A. In other words, adding additional visual cues of speaker faces does not seem to help listeners establish more speaker-specific categorization boundaries.

Similarly, a mixed-effects model (Model-4b) is conducted to predict the Response of

each trial (T=0, D=1), with Step (scaled and centered), Trial (scaled and centered), and Group (treatment coded, baseline: *female-auditory*) as the fixed effects, Group:Step and Group:Trial as the interaction items, and Step by Subject and by Word as random slopes. Again, the *female-auditory* condition is coded as the reference level in order to provide straightforward comparisons between these condition and each of the other two conditions. The fixed effects of Model-4b are shown in Table 6.4.

Fixed Effect	Est.	SE	z	Pr(> z)
(Intercept)	0.02	0.35	0.05	0.96
Step	-1.45	0.29	-4.98	< 0.001***
Condition Female A t-favoring	-1.13	0.34	-3.34	< 0.001***
Condition Female A baseline	0.04	0.32	0.13	0.90
Condition Two females opposite audiovisual	0.02	0.34	0.06	0.95
Condition Two genders opposite auditory	-0.69	0.32	-2.13	0.03*
Trial	-0.11	0.09	-1.21	0.23
Step:Condition Female A t-favoring	-0.54	0.20	-2.64	0.008**
Step:Condition Female A baseline	0.08	0.18	0.43	0.67
Step:Condition Two females opposite audiovisual	0.07	0.19	0.34	0.74
Step:Condition Two genders opposite auditory	-0.14	0.19	-0.73	0.47
Condition Female A t-favoring:Trial	0.07	0.14	0.49	0.63
Condition Female A baseline:Trial	-0.09	0.12	-0.72	0.47
Condition Two females opposite audiovisual:Trial	0.01	0.13	0.11	0.91
Condition Two genders opposite auditory:Trial	-0.16	0.13	-1.23	0.22

Model-4b: Response~Step*Condition+Condition*Trial+(Step|Subj)+(Step|Word)

Table 6.4: The fixed effects of the logistic mixed-effects model in Exp 4b

Table 6.4 reveals a significant main effect of Step ($\beta = -1.45, p < 0.001$), indicating that segments with longer VOT at larger Steps are less likely to be perceived as /d/ in the *female-auditory* condition. Step interacts with Group in the Female A /t/-favoring condition ($\beta = -0.54, p < 0.001$), meaning that the slope of the categorization line becomes sharper in the Female A /t/-favoring condition. The interaction is not significant for the baseline condition ($\beta = 0.08, p = 0.67$), the *female-audiovisual* condition ($\beta = 0.07, p = 0.74$), or the *gender-auditory* condition ($\beta = -0.14, p = 0.47$).

The Group effect is only significant for the conditions of Female A /t/-favoring ($\beta = -1.13, p < 0.001$) and the *gender-auditory* condition ($\beta = -0.69, p = 0.03$), but not for any of the other conditions ($\beta_{baseline} = -0.04, p = 0.90$; $\beta_{audiovisual} = 0.02, p = 0.95$). This

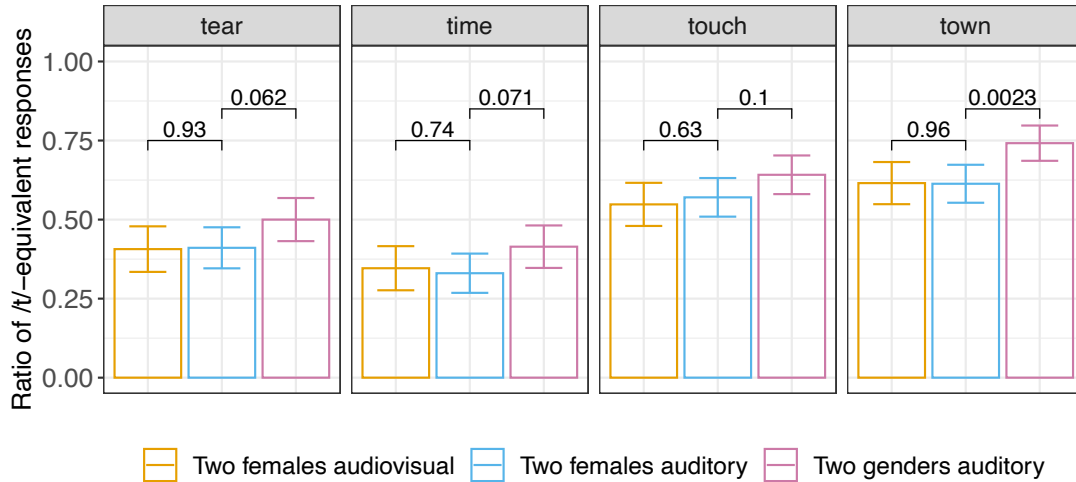


Figure 6.6: Exp 4b: /t/ response rate by condition and word (mean and 95% confidence interval). Values above the comparison bars indicate the p-values of Wilcoxon rank sum tests.

is consistent with the pattern revealed in Fig. 6.5 showing that the Female A t-favoring condition and the *gender-auditory* condition both show a much higher /t/-response rate than the other three conditions. The effect of Trial ($\beta = -0.11, p = 0.23$) is not significant, nor is any of its interactions with Group.

6.3.2.3 Analysis by lexical frame

As demonstrated in Sec 5.3.3.3 of Ch 5, the perceptual learning results of /t-d/ may vary substantially with the lexical frames of the test stimuli, because each lexical frame is spliced with /t-d/ sounds from a unique VOT continuum whose range is determined specifically for that frame. To examine whether such lexical influences exist in Exp 4a, I separately examined the results obtained with different lexical frames in Exp 4b. Fig. 6.6 shows the means and 95% confidence intervals of the aggregate /t/ response rates by word and experimental condition. The four facets show a similar pattern that a higher /t/ response rate is exhibited in the *gender-auditory* condition than the other two female conditions, with no obvious difference exhibited between the latter two conditions.

A series of Wilcoxon rank sum tests are conducted to evaluate whether the differences

between the *female-auditory* condition and each of the other conditions are statistically significant for different lexical frames. As shown in Fig. 6.6, the difference between the *female-audiovisual* condition and the *female-auditory* condition does not turn out significant in any of the four lexical conditions ($W_{tear} = 20300, p = 0.93$; $W_{time} = 20706, p = 0.74$; $W_{touch} = 26032, p = 0.63$; $W_{town} = 26680, p = 0.96$). This constant lack of difference between the Two-female conditions are consistent with the results in Fig. 6.5 and Table 6.4, indicating the absence of interactions between lexical frame and the difference between these two experimental condition. Regarding the comparison between the *female-auditory* condition and the *gender-auditory* condition, the significance levels of their difference vary with the lexical frame of the test stimuli: The difference is significant for responses obtained with “town-down” ($W = 34664, p = 0.002$) and is marginal significant for responses obtained with “tear-dear” ($W = 25620, p = 0.06$) and “time-dime” ($W = 25494, p = 0.07$). It is not significant with “touch” ($W = 32912, p = 0.10$). The existence of a significant or marginally significant difference between the *gender-auditory* and the *female-auditory* conditions points to a somewhat constant constraint of speaker gender on the perceptual generalization of /t d/.

6.3.2.4 Summary

In contrast to Exp 4a, Exp 4b shows evidence for speaker gender constraints on the perceptual learning of /t-d/ in multi-speaker listening. By comparing the results of the *female-auditory* condition and the *gender-auditory* condition in Exp 4b, we see that the influence of the /d/-favoring training with Female B diminished when Female B’s voice is manipulated to be male-sounding, compared to the condition where her voice is female-sounding. As a result, the categorization result of the *gender-auditory* condition still retains much of Female A’s speech properties and exhibits a distinctive perceptual shift towards /t/. In contrast, in the *female-auditory* condition, the perceptual bias induced by exposure to Female B’s speech is strong enough to cancel out the prior perceptual learning effect with Female A’s speech, leading to no perceptual shift from the baseline condition in the final categorization

stage. The difference between the *gender-auditory* condition and the *female-auditory* conditions in Exp 4b essentially reflects that listeners in the *gender-auditory* condition have generalized less of Female B's /f/-favoring speech bias to the final test phase due to a change of voice gender. This is regardless of the fact that all the speech tokens in the two conditions share the same set of temporal properties. Meanwhile, I did not find evidence for the constraint of speaker identity on the perceptual learning of /t-d/ with multiple speakers. This is reflected by the lack of difference between the *female-auditory* condition and the *female-audiovisual* condition.

6.4 Discussion

In Exp 4, I investigated the potential constraints of speaker identity and gender on the perceptual learning of /s-f/ and /t-d/ in a multi-talker listening setting. As a result, I find some evidence for the speaker gender constraint on the perceptual generalization of /t-d/ and some weak evidence for the speaker identity constraint on the perceptual generalization of /s-f/. I did not find evidence for the remaining combinations, namely, a gender constraint for /s-f/ and an identity constraint for /t-d/. However, the lack of evidence can be partially attributed to issues with experimental design and participant sampling and should be investigated further before conclusions are made. The remainder of this section goes over the findings or issues I encountered in Exp 4, respectively for the two types of phonemes and the two kinds of constraints in question.

Speaker gender and /s f/

The perceptual generalization of /s f/ across multiple speakers has been reported to depend on the gender of involved speakers (Kraljic and Samuel, 2007), and this interesting finding is a crucial motivation of this dissertation. However, in Exp 4a, I did not find evidence for the constraint of speaker gender on the perceptual learning of /s-f/. To be specific, when the stimuli of Female B are manipulated to be male-sounding, the phoneme categorization results actually reflect a bigger influence of the /f/-favoring training with Female B com-

pared to the condition where they are female-sounding. Nonetheless, this result is not a piece of powerful evidence against a speaker gender constraint on /s ʃ/ perceptual learning either because it can be interpreted in alternative ways. Especially, the observed perceptual shift towards /ʃ/ in the *gender-auditory* condition might be caused by the interference of a speaker normalization mechanism, as I have explained in Sec 6.3.1.3. The normalization mechanism may cause the sibilants of Female B to be even more /s/-sounding in the *gender-auditory* condition than they are intended to be, which adds up to the strength of the /ʃ/-favoring training with Female-turned-male B in the second phase.

Future studies are needed to follow up on this issue. However, this is not an easy question to cope with because one needs to make a compromise between the comparability of the acoustic properties of sibilants and the comparability of their phonological or perceptual locations across gender-pairing conditions. How do we ensure that one set of sibilants embedded in female-sounding speech would be strictly perceptually equivalent to another set of sibilants embedded in male-sounding speech? This poses a unique challenge for follow-up studies along this line.

Speaker gender and /t d/

Previous studies have reported that the perceptual generalization of /t-d/ across speakers is not constrained by different genders of the speakers (Kraljic and Samuel, 2007). Again, however, this is not what we found in Exp 4b. By comparing the results of the *female-auditory* condition and the *gender-auditory* condition in Exp 4b, we see that the /d/-favoring effect of Female B's speech on the perception of Female A's /t-d/ perception diminished when Female B's voice is manipulated to be male-sounding, compared to the condition where the voice is female-sounding.

A typical confound involved in the interpretation of the comparison between a same-gender condition and a different-gender condition is whether the difference is caused by speaker identity or speaker gender. Given that there is no guarantee that participants would always perceive the identity of voices of the same gender accurately, an alternative

explanation could always be that listeners might have treated Female A and B as the same person in the *female-auditory* condition while they are less likely to do so in the *gender-auditory* condition. Fortunately, this possibility can be ruled out for this set of data based on the results of speaker perception in Fig. 6.2. Although this kind of misperception could have happened and is captured from the responses of Exp 4a’s participants, it does not really show up with the responses of participants in Exp 4b.

The results in Fig. 6.2 also allow us to think a little bit deeper about the consequences of different voice gender conditions on speaker identity perception. A comparison between the speaker perception results in the *female-auditory* condition and the *gender-auditory* condition suggests that the voice gender manipulation has caused listeners to believe that they have heard more voices than they actually did. A conceivable possibility behind this result is that more listeners can identify that the speakers in the first and third phase are the same person in the *female-auditory* condition than in the *gender-auditory* condition. If this factor has affected the result, then we should expect that listeners in the *female-auditory* condition are more likely to retain Female A’s /s/-favoring perceptual expectations than those in the *gender-auditory* condition, because they are more aware of the fact that the speaker in the test phase is exactly the one who induced an /s/-favoring perceptual expectation in the first training phase. However, this is not the pattern we observed. Instead, we have seen that listeners in the two-gender condition actually managed to reflect more of Female A’s /s/-biasing speech properties in the test phase without a clearer realization that they have the same speaker. Therefore, this difference is caused by constraints of speaker gender instead of its side effects on speaker identity perception. Taken together, the difference between the *gender-auditory* and *female-auditory* conditions in Exp 4b lend support to a speaker gender constraint on the generalization of perceptual learning, which claims that listeners are less likely to generalize what they have learned from a previous speaker to another speaker of a different gender group than one from the same gender group.

Speaker identity and /s ʃ/

Experiment 4a shows that the /s/-equivalent response rates in the *female-auditory* condition is higher than that of the *female-audiovisual* condition. This is consistent with the prediction of speaker identity constraints on the perceptual generalization of Speaker identity and /s ʃ/, namely, that listeners are more likely to generalize their perceptual learning of the same speaker rather than that of a different speaker when they have accurate access to the information of speaker identity.

One concern with this set of result is that the difference between the above two conditions are not robust enough to show up in the logistic mixed effect model, although it does turn out to be significant in the Wilcoxon rank-sum tests and in both of the two phoneme conditions. Given that the realizations of sibilants are not only known to covary with gender but also other identity-related aspects such as gender orientation, this result opens up a new possibility to be explored, that is, the perceptual learning of /s-ʃ/ might also be sensitive to speaker specificity, and this sensitivity is an influence in a more gradient and less detectable way.

Speaker identity and /t d/

In Exp 4b, I did not find evidence for speaker identity constraints on the perceptual learning of /t-d/ with multiple speakers. By comparing the *female-auditory* condition and the *female-audiovisual* condition, we can see that the availability of additional visual cues to speaker identity does not make a difference to the phoneme categorization results. Listeners did not show more /t/-favoring patterns in line with Female A's speech properties after they had realized that the test speaker is Female A.

However, a worrisome in the interpretation of this set of results is that the pattern might be associated with the unusual sampling of participants in that particular condition. Recall that in Fig. 6.2, we have seen that the availability of visual cues makes a difference to the reported number of voices in Exp 4a but not in Exp 4b. In other words, participants with access to the visual cues should be less likely to report that they had heard three speakers

because only two speaker photos are presented throughout the experiment. This difference has come up in Exp 4a but not Exp 4b according to Fig. 6.2. Follow-up studies are needed to rule out the possibility that participants in the *female-audiovisual* condition in Exp 4b just happened not to have noticed that there are only two photos of unique female speakers presented in the experiment.

Chapter 7

Discussion and Conclusion

This dissertation's broad goal was to investigate the possibility that the generalization of perceptual learning across speakers within and between social groups reflects listeners' sociophonetic knowledge that mirrors the structure of real-world speaker variability. In this dissertation, I look into this possibility starting with the question of whether the perceptual learning of /s-f/ and /t-d/ generalizes across different speakers and genders. In Chapter 1, I have broken down the research question of the present dissertation into three lines. Exp 1 and 2 investigates the first line of inquiry regarding the perceptual generalization of sibilants in multi-talker listening and their interaction with acoustic constraints. Exp 3 looks into the second line of inquiry regarding perceptual generalization of stops across speakers and genders. Exp 4 provides a comparison of the social constraints on the perceptual learning of different types of phonemes, asking how they differ in qualitative and quantitative ways. In this chapter, I synthesize the results of these experiments and discuss their implications for the operation of perceptual learning in multi-talker listening, listeners' knowledge of structure in talker variability, and the interweaving of social information in the architecture of the phonetics-phonological mapping system.

7.1 Major findings of this dissertation

7.1.1 Perceptual generalization by cumulative update

In Exp 1 and 3, I asked how the perceptual learning of /s-f/ and /t-d/ generalizes across multiple speakers of different genders. The experiments and conditions generally share a similar

design and consist of one or two training blocks followed by a test block, all implemented in the form of an identification task. In the experimental conditions with two training phases, participants were exposed to the speech of Female A and Male A sequentially before they were tested with Female A’s speech again. Through independent manipulation of the presence or absence of exposure to different speakers’ speech and phonetic characteristics, I examined how listeners’ categorization responses were affected by each of those training phases.

In Chapter 2, I have proposed four hypotheses for how perceptual learning works with multiple speakers, which I reproduce here in Fig. 7.1.

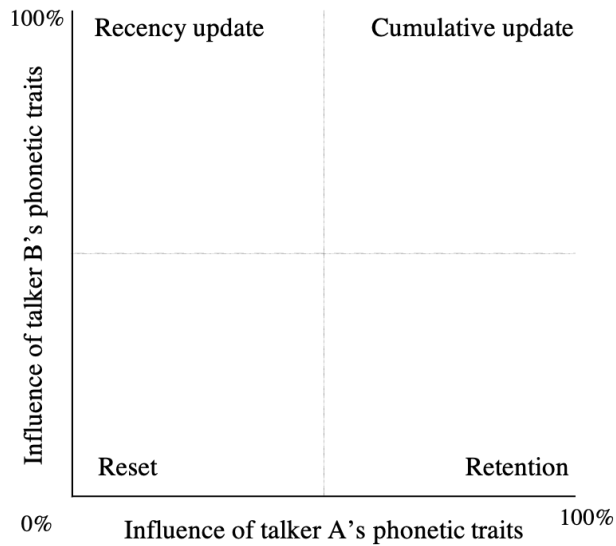


Figure 7.1: Potential outcomes of perceptual learning with speaker A and B successively

The first hypothesis to be examined is that perceptual learning is talker-specific. In the case of our experiments, if we only found an influence of the training with Female A but not with Male A on the final categorization results with Female A’s speech (100% talker A and 0% talker B), then it suggests that listeners retain speaker-specific perceptual learning outcomes (*retention*). In Exp 1a, 1b, and 3a, I have kept the speech characteristics of Female A constant and those of Male A varied across conditions. The hypothesis of *retention* predicts that participants in different conditions of these sub-experiments would not show different speech categorization patterns with Female A’s speech because they were

not expected to integrate the perceptual learning with a different speaker – Male A – in the final categorization task. However, the results of these sub-experiments are against this hypothesis. Instead, we keep seeing that the results of different conditions constantly deviate from each other in ways consistent with their second-phase training. This means that after listeners have already established their perceptual expectations with Female A’s speech in the first training phase, further exposure to a different speakers’ speech in the second phase can still change the perceptual expectations they established previously for Female A.

The second hypothesis I evaluated is that listeners set aside their previous perceptual learning and *reset* their perceptual boundary to the baseline each time they encounter a different speaker. A *reset* hypothesis predicts that listeners in different experimental conditions within Exp 1a, 1b, and 2a would exhibit the same categorization results aligned with the baseline condition, where participants have not received any perceptual training (0% talker A and 0% talker B). This is because participants in all the two-talker learning conditions have experienced a switch of talker’s voices that is supposed to re-initialize their perceptual learning process, according to a *reset* account, regardless of the previous learning treatment they have received. Again, our results do not support such a hypothesis. We have only observed a return to the baseline when the perceptual biases associated with training with Female A and Male A are in opposite directions. Other than that, when Male A’s speech introduces perceptual biases towards the same direction as Female A’s speech does, or when it contains no critical phonemes, the perceptual boundary does not return to the baseline.

The results of Exp 1a, 1b, and 3a all suggest that the perceptual learning with Male A generalizes to the perception of Female A’s speech and affects the categorization results in the test phase. At this point, the remaining hypotheses are either that listeners update their perceptual expectations to reflect the phonetic characteristics of the most recent speech input they are exposed to (*recency update*), or that listeners take all of the learning experience into account to update their perceptual expectations (*cumulative update*). Exp

1c and 3b are conducted to evaluate these two alternatives for sibilants and stops, asking to what extent the categorization results in Exp 1a, 1b, and 3a are also affected by the previous training experience with Female A. In the experimental conditions of Exp 1c and 3b, the training phase with Female A was taken out, and listeners only received training with Male A’s speech before they were tested with the same continuum of Female A’s speech. By comparing the results of the two-speaker training conditions and the results of training with Male A only, I find that training with Male A alone does not introduce perceptual shifts with comparable magnitudes to shifts induced by two phases of perceptual learning with Female A and Male A.

Taken together, we have seen frequent occurrences of perceptual generalization across speakers of different genders for both sibilants and stops, which is much more frequent than what previous claims in the literature would predict. The results of Exp 1 and 3 lend support to a *cumulative update* account, which suggests that perceptual learning updates across speakers in such a way where previous and current perceptual learning experiences are re-integrated to form a cumulative perceptual expectation that listeners use for upcoming perception events. Nonetheless, generalization is not universal or limitless. Next I will discuss some of the situations where the generalization of perceptual learning *is* inhibited.

7.1.2 Constraints on cross-talker perceptual generalization

Building on the previous finding that perceptual learning generalizes by cumulative update across speakers and genders, Exp 2 and 4 demonstrate how the occurrence and magnitude of perceptual generalization across talkers can be constrained by the phonetic properties of the phoneme realizations and the social condition of talkers involved in training.

In Exp 2, I evaluate two hypothesized phonetic constraints on the perceptual generalization of sibilants across speakers. One is the *acoustic dissimilarity* constraint (Kraljic and Samuel, 2005), which states that the perceptual generalization of sibilants across speakers will be blocked when the distributions of sibilant spectral energy are distinct between speakers. The other is a *acoustics-phonology mismatch* constraint that I proposed. It claims that

when the direction of the intended perceptual bias is at odds with the inferred perceptual shift from the sibilant spectral distributions, perceptual generalization does not occur.

By replacing Female A in Exp 1 with another female talker with higher sibilant frequencies (Female B), Exp 2 provides a testing ground for evaluating these hypotheses. In Exp 2b, the Two genders - same condition and the Two genders - opposite condition form a contrast in exhibiting acoustic dissimilarity versus acoustics-phonology mismatch (Fig. 4.12, right facet). In the Two females - opposite condition, sibilants of Male A and Female B in training do not overlap in the COG dimension, but this lack of acoustic similarity does not block the perceptual generalization of sibilants across speakers: listeners who had received training with Female B's /f/-favoring speech and Male A's /s/-favoring speech set back their perceptual boundary to the baseline position. In contrast, sibilants of the two speakers in the Two females - same condition do exhibit an amount of overlap in the COG dimension. However, the second /f/-favoring training phase with Male A contains lower-frequency sibilants that make them /f/-sounding compared to sibilants in first-phase training with Female B. This property sets a high bar for the perception of /f/ and therefore favors /s/ perception, which raises an acoustics-phonology mismatch. Due to the acoustics-phonology mismatch, the result shows that, surprisingly, listeners who had received two phases of /f/-favoring training exhibited a boost in the number of /s/ responses. An *acoustics-phonology mismatch* account also works to explain the results of Exp 2c, where the perceptual learning of Male A's /f/-favoring speech fails to generalize to Female B's sibilant continuum. This is because Male A's /f/-favoring sibilants have lower COG values than the lowest end of Female B's sibilant continuum, making it difficult for listeners to reconcile that the intended /s/-like sounds actually have a low frequency of spectral energy distribution. In a nutshell, the above results show evidence for the constraint of acoustic-phonology mismatch but not the constraint of acoustic similarity on the perceptual generalization of sibilants.

In Exp 4, I evaluated a different kind of constraint, namely, speaker identity and gender, on the perceptual generalization of /s-f/ and /t-d/ across speakers of either different genders or the same gender. For the perceptual generalization across speakers of different genders, I

ask whether generalization would be inhibited for learning with speakers of a different gender compared to learning with speakers of the same gender. The hypothesis is supported by the results of Exp 4b, where participants received exposure to Female A's /t/-favoring speech and Female B's /d/-favoring speech sequentially before they were tested with Female A's /t-d/ continuum. I find that participants set back their perceptual shifts to the baseline position when the two voices are both female-sounding, whereas they exhibited a higher number of /t/-equivalent responses than the baseline level in their categorization of Female A's /t-d/ when Female B's voice was turned to be male-sounding. This means that the /d/-favoring learning with Female B does not fully generalize to the final categorization when it becomes male-sounding. Therefore, the categorization results still maintain a shift as established in the first learning phase.

Similarly, for the perceptual generalization across speakers of the same gender, I ask whether the generalization would be inhibited for perceptual learning of a different speaker compared to learning of the same speaker. The hypothesis of a speaker identity constraint gains some support from the results of Exp 4a, where listeners were exposed to Female A's /s/-favoring speech and Female B's /ʃ/-favoring speech before they completed sibilant categorization with Female A's /s-ʃ/ continuum. When listeners are provided with visual cues to the identity of Female A and B, again, they show less integration of the perceptual learning of Female B's /ʃ/-favoring speech in the categorization phase. As a result, they exhibit more /s/-equivalent responses than participants who were not exposed to the visual cues. Note, however, that compared to participants who only received /s/-favoring training with Female A, participants in the audiovisual condition still have a much lower /s/ response rate. In other words, although participants in the audiovisual condition did not fully generalize their learning with Female B, they still did it to some extent. This result suggests that listeners generalize less for perceptual learning with a different speaker compared to learning with the same speaker.

The above results add to previous findings where the generalization of perceptual learning becomes inhibited in certain situations, such as when listeners see that speakers hold a

pen in their mouth (Kraljic et al., 2008b), or when the disambiguating text shows up at a later time point than the audio (Caplan et al., 2021). They also suggest that constraints on perceptual generalization may take place by modulating the magnitude of the consequential perceptual shift rather than forbidding it. These findings add to the full picture of how perceptual learning ceases, resumes, and modulates in feeding into upcoming perception behaviors within the bounds set by relevant constraints.

7.2 Theoretical implications and future research directions

7.2.1 The role of listeners' knowledge of structure in talker variability

Recall that the broad goal of this dissertation is to evaluate whether perceptual generalization behaviors with multiple talkers within and between social groups reflect listeners' knowledge of the real-world structure in talker variability. This proposal first came up in Kraljic and Samuel (2007) and was later adopted in an "ideal-adapter" framework by Kleinschmidt (2019). According to this proposal, listeners have good knowledge of the speech properties of speakers from different sociophonetic speaker groups; they use this information to predict which speakers share similar phonetic characteristics and to guide their perceptual generalization behaviors. One of the most well-known empirical observations in favor of this proposal is that Kraljic and Samuel (2007) reported that the perceptual learning of fricatives is speaker- or gender- specific, whereas the perceptual learning of stop VOT is generalizable across different speakers and genders. As one of the first few research endeavors to systematically evaluate this proposal, I ask whether the perceptual learning of sibilants remains more specific to the speaker or the type of speakers that trigger the learning effect than that of stop VOT. Two types of comparisons are considered between the perceptual generalization of /s-f/ and /t-d/ regarding their susceptibility to speaker identity and gender.

The first type of comparison regards perceptual generalization as a categorical behavior and characterizes their presence versus absence with different speaker groups and phoneme

types. Under this perspective, I evaluated the hypothesis that perceptual learning generalizes across speakers of different genders for stops (Exp 3) but not for fricatives (Exp 1-2). As we have already seen, the results do *not* lend support to such a hypothesis. Instead, they show that perceptual learning robustly generalizes across speakers of different genders with both sibilants (Exp 1, 2a) and stops (Exp 3). Later, we also see that the results of perceptual generalization are replicated with speech stimuli of the same voice gender in Exp 4 (the Two females - auditory condition). The prevailing occurrence of perceptual generalization across speakers is one of the most major findings of this dissertation, which has been replicated with speakers across and within gender groups, phoneme contrasts of sibilants and stops, and different combinations of acoustic properties of the target phonemes, in four experiments, ten sub-experiments, and thirty out of thirty-four experimental conditions. In contrast to the evidence for perceptual generalization, which shows up from the very first experiment to the very last experiment, evidence of categorical absence of perceptual generalization due to speaker- or gender- specificity is seldom observed in this dissertation.

With the prevailing findings of perceptual generalization, I then asked whether the magnitude of boundary shifts as a result of perceptual generalization vary in different conditions of speaker identity and speaker gender. In Exp 4, I kept the acoustic properties of the target phonemes constant while manipulated the availability of visual cues to speaker identity and voice cues to speaker gender across experimental conditions. As described in the earlier section 7.1.2, a comparison between the categorization results of different conditions indicates that speaker social information does impose some gradient constraints on the magnitude of perceptual generalization. Results of /s-f/ learning in Exp 4a lend support to a speaker gender constraint by showing that listeners integrate less of the sibilant properties of Female B and generalize it to the phoneme categorization of Female A's speech when Female B's voice is manipulated to be male-sounding. Results of /t-d/ learning in Exp 4b support a speaker identity constraint by showing that listeners integrate less of the VOT properties of Female B and generalize it to /t-d/ categorization of Female A's speech when visual cues of speakers' faces were presented to inform them that Female B is a different speaker. I

then asked how the constraints of the same type of social information on perceptual generalization might differ in magnitude between sibilants and stops. However, since the results of Exp 4 only review evidence for constraints of speaker gender on the generalization of stops and constraints of speaker identity on the generalization of sibilants, it is impossible to compare the magnitude of the same kind of constraint on the generalization of different phonemes with the current result data.

In a word, the results of this dissertation are at odds with the previous empirical finding that perceptual learning is gender-specific for fricatives but not stops. Nonetheless, the results also provided some evidence that speaker social information imposes certain constraints on the generalization of perceptual learning across speakers. Put together, the results of this dissertation are still in line with proposals suggesting the involvement of listeners' sociophonetic knowledge in their perceptual generalization behaviors. However, without a chance to compare the influence of the same kind of constraint on the perceptual generalization of different types of phonemes, it remains a mystery how listeners' sociophonetic knowledge in relevant aspects looks like and how closely they pattern with the talker variability in the real world. Also, it is unclear whether the absence of gender constraint effects in Exp 4a and the absence of identity constraint effects in Exp 4b reflect the truth of reality or consequences of confounding factors. The interpretability of the null gender effect in Exp 4a suffers from a potential interference of the speaker normalization mechanism. Even though I have kept the speech signals of sibilants comparable across gender conditions, the same sibilants would still end up perceptually different when embedded in different gender voices. The null effect of speaker identity in Exp 4b is weakened by the unexpected talker perception behaviors of participants in the Two-females visual condition, namely, that exposure to visual cues of speaker identity does not help inform the listeners of the correct number of speakers in the experiments. Follow-up studies will be in need to rule out these confounds in order to obtain a more comprehensive understanding of the nature of speakers' knowledge of talker variability structures and the involvement of this knowledge in perceptual generalization behaviors.

7.2.2 Perceptual generalization in the acoustic vs. phonological space

In addition to the question of listeners' knowledge of talker variability, this dissertation raises another fundamental but understudied question, namely, what is the nature of the target of perceptual generalization? Are the perceptual beliefs updated by tracking the distributions of relevant phonemes' raw acoustic values or their perceptual locations relative to the standard phoneme instances in the phonological space? This question mainly comes up in Exp 4a, when we find that embedding a set of /j/-favoring sibilants in male-sounding lexical contexts boosts the effect of /j/-favoring perceptual learning. One way to account for this finding is to introduce a level of interference of the perceptual normalization mechanism. By this account, the perceptual normalization of sibilants against vowel contexts of lower formant frequencies makes the sibilants more /s/-sounding and therefore more /j/-favoring for the identification of upcoming instances. The additive /j/-favoring effects give rise to the boost in /j/-equivalent responses in the categorization phase. This is regardless of the fact that they are produced by speakers of different gender, which is expected to inhibit their effect sizes. In other words, the same acoustic pieces of signals embedded in two sets of vowels with different levels of vowel formant frequencies may induce perceptual learning at different strengths. This strength resulting from perceptual normalization is then entered into the perceptual learning process as a piece of input information, instead of/in addition to the raw acoustic values of the sibilants.

Existing computational models of perceptual adaptation based on a mixture of Gaussians (Toscano and McMurray, 2010) or Bayesian belief update (Kleinschmidt and Jaeger, 2015) normally take raw acoustic parameters of VOT lengths or formant frequencies as the model input to predict the location of the outcome perceptual boundary in the acoustic space. With these models, speech normalization is not a problem because they usually deal with the perceptual learning of a single speaker. However, with multiple voices from different speakers coming into play, the spectral frequencies of different voices naturally raise the issue of potential interference of speech normalization in the perceptual learning of sibilant, vowel, and pitch. In addition to spectral contrasts, the speaker normalization might also

interfere with the perceptual learning of temporal contrasts such as stop VOT. Speakers may also differ in their speech rates and other temporal dimensions of speech productions, which introduces variability into the normalization context for the perception of temporal contrasts. How does the perceptual generalization of VOT affected by the different speech rates of speakers? Do listeners generalize their learning of the raw duration of VOTs to distinguish stop voicing, or do they also need to account for speech rate and generalize the VOT target located in a phonological space? These questions are yet to be solved by future research.

7.2.3 Implications for the mental representation of speech variability in the phonetics-phonology interface

In the last section, I discuss the theoretical implications of this dissertation for the mental representations of speech in phonetics-phonology mapping. This dissertation is situated in the background when there is a shift of research foci from the fundamental units of human speech to the role of intra-category phonetic details in speech perception and processing. Debates have arisen regarding whether the mental representation of phonological units is better modeled as discrete linguistic symbols or a collection of perceptual episodes. Nowadays, there is a trend in the literature to acknowledge both phoneme-level and exemplar-level representations and integrate them through explorations of “hybrid” models (e.g., Wilder, 2018). Most of the current usage-based models have also integrated phoneme-level representations to account for phoneme-driven linguistic phenomena such as regular sound change (e.g., Beckner and Bybee, 2009; Harrington et al., 2018; Hay and Foulkes, 2016; Todd et al., 2019).

The results of this dissertation reinforce the development of the hybrid view. It is normally assumed that perceptual adaptation behaviors occur at the processing level and respond to random speaker variability on the fly. In contrast, perceptual normalization is a more regular part of linguistic knowledge and is essential to the access and comprehension of linguistic units. In our result, we can see that the input of the former mechanism is

dependent on the output of the latter. This at least suggests that the two mechanisms need to be activated in adjacent time windows and coordinate to feed into the operation of each other. The results of this dissertation lend support to such a representational system, where phonemes and exemplars should be both represented cognitively with associative links between them to allow for interactions on the fly. The question remains at which processing levels perceptual normalization and perceptual generalization each come into play and what kind of specific procedures they follow to interact. Future studies are still in need to delineate the cooperation of these two mechanisms in more detail.

Appendix A

Results of Pilot Studies

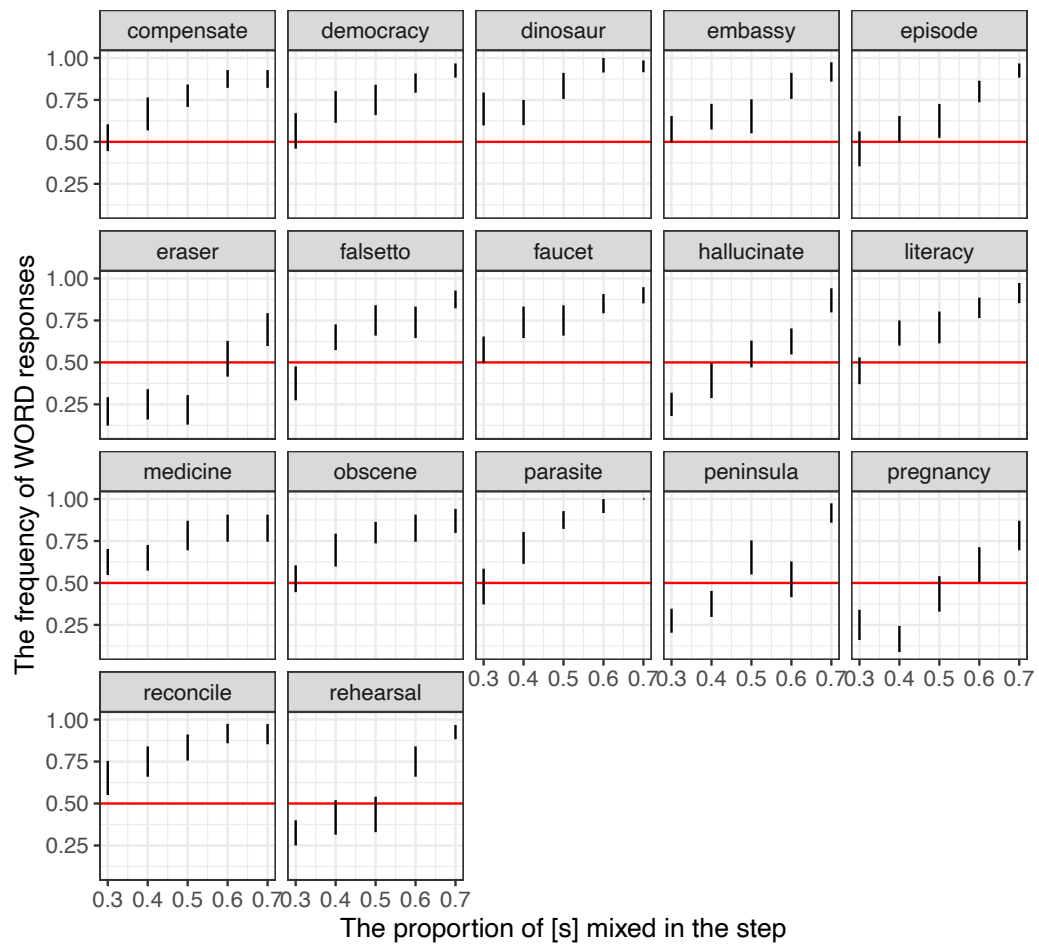


Figure A.1: Lexical decision results for /s/-containing words from Female A (Ashley) with different proportions of /f/ blended into the fricative

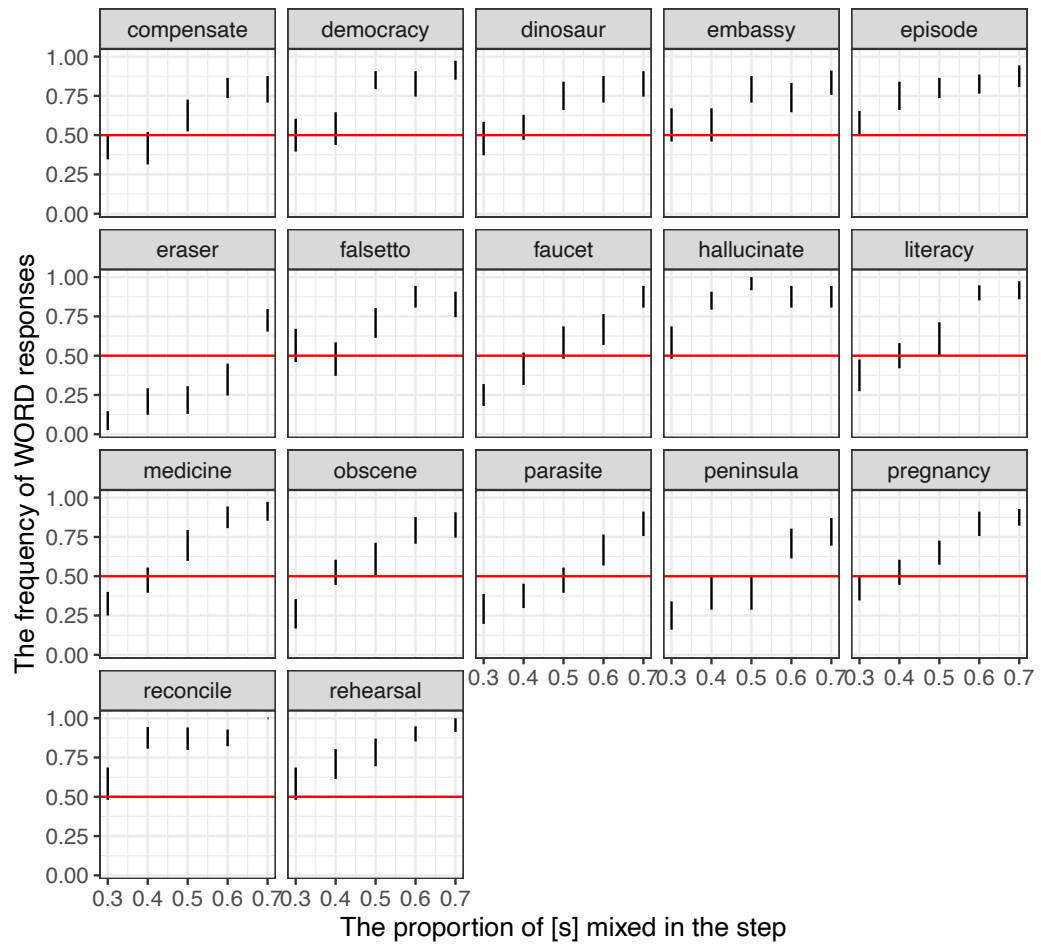


Figure A.2: Lexical decision results for /f/-containing words from Female B (Vicky) with different proportions of /f/ blended into the fricative

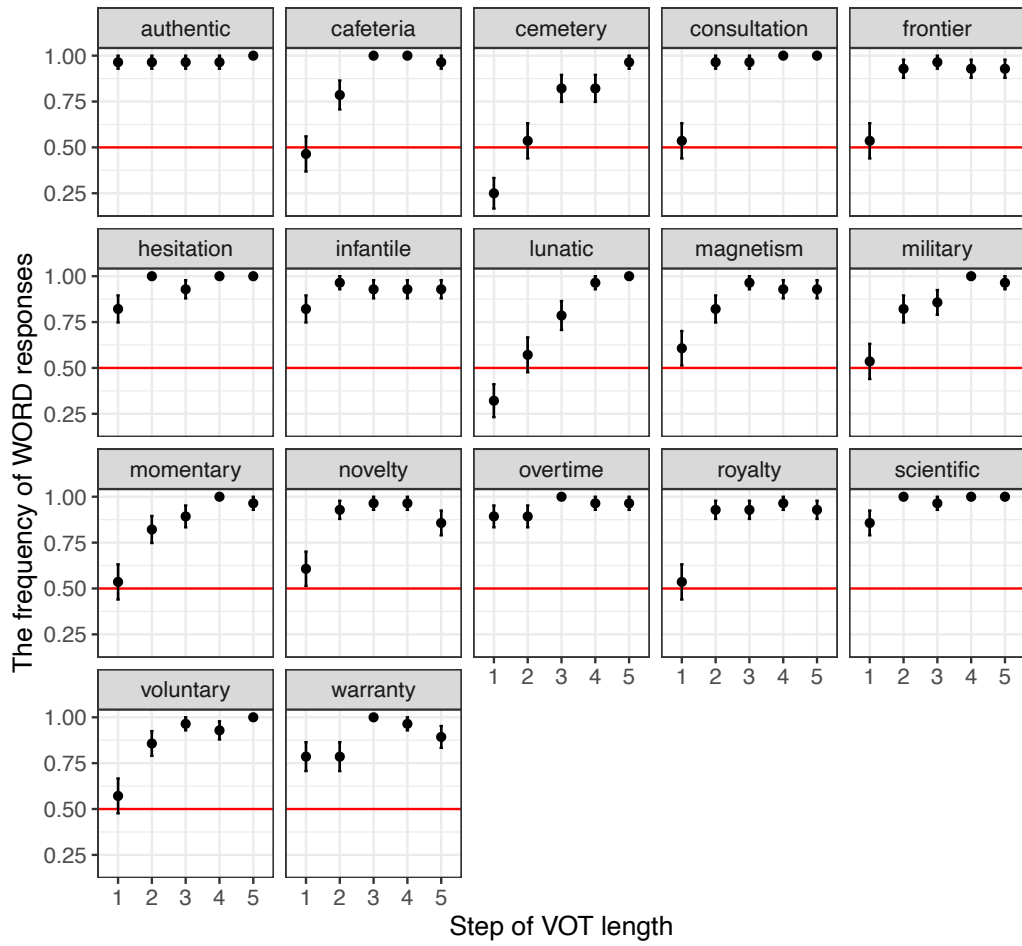


Figure A.3: Lexical decision results for /t/-containing words from Female A (Ashley)

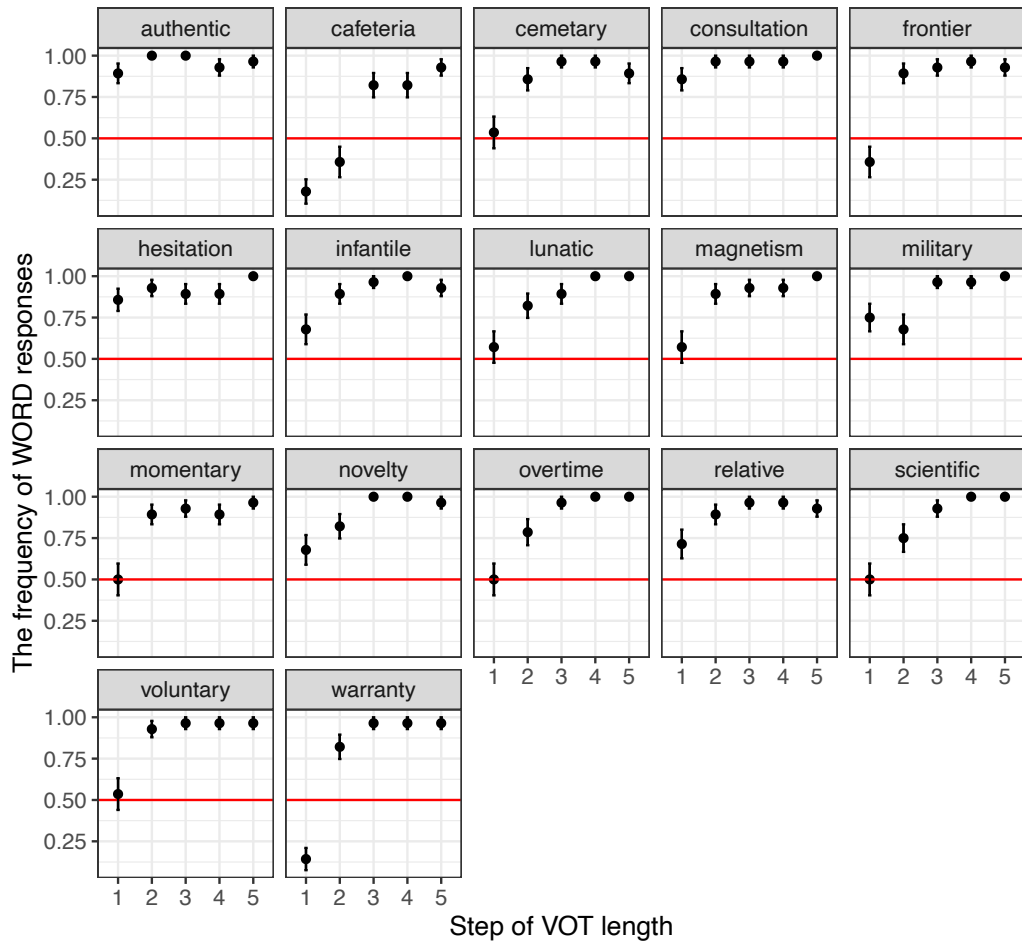


Figure A.4: Lexical decision results for /t/-containing words from Male A (Gabriel)

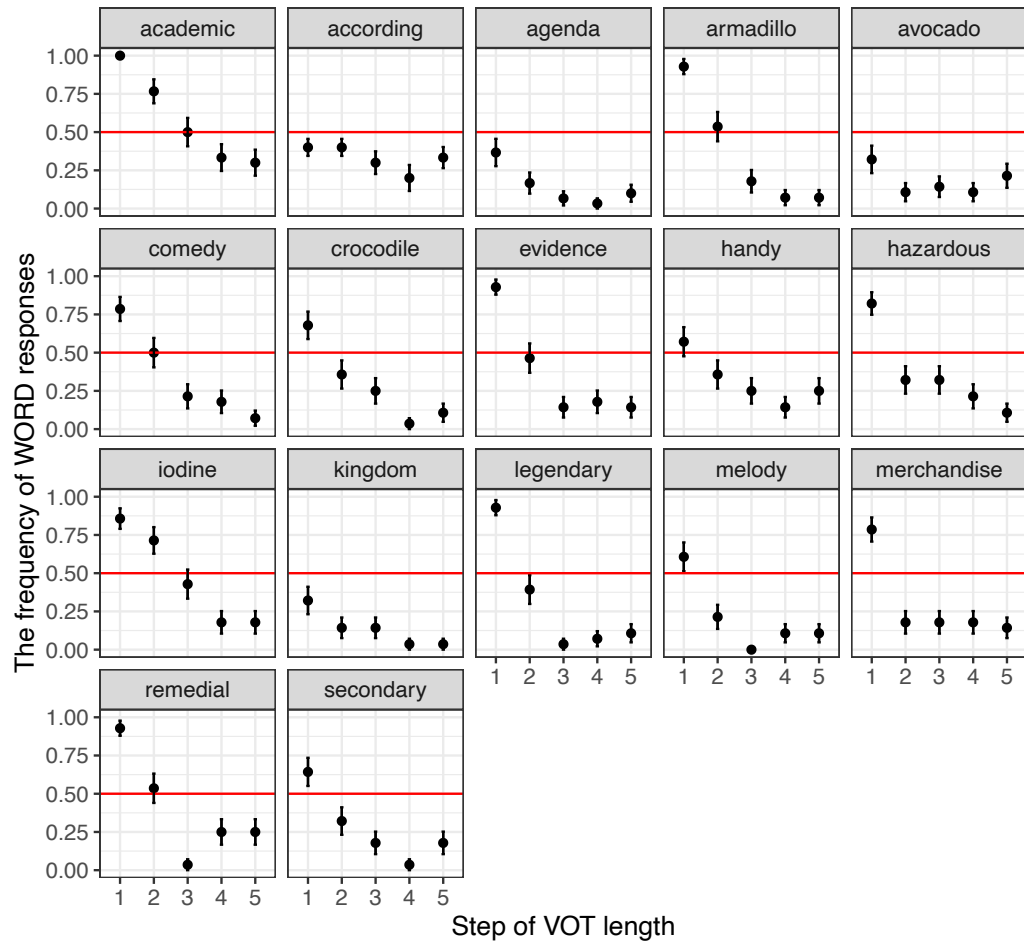


Figure A.5: Lexical decision results for /d/-containing words from Male A (Gabriel)

Appendix B

Supplementary models

Fixed Effect	Est.	SE	z	Pr(> z)
(Intercept)	-0.81	0.41	-1.97	0.049*
Step	-1.68	0.15	-11.22	< 0.001***
Condition Two genders - opposite	1.54	0.51	3.04	0.002**
Condition Two genders - same	-1.12	0.52	-2.16	0.03*
Trial	-0.04	0.09	-0.47	0.64
PhonemeS	-0.60	0.05	-11.21	< 0.001***
Step:Condition Two genders - opposite	0.24	0.19	1.28	0.20
Step:Condition Two genders - same	0.46	0.20	2.28	0.02*
Condition Two genders - opposite:Trial	0.06	0.12	0.45	0.65
Condition Two genders - same:Trial	0.02	0.13	0.13	0.90

Model-1c-b: Response~Step*Group+Group*Trial+Phoneme+(Step|Subj)+(1|Word)

Table B.1: The fixed effects of Model 1c-b for conditions including Male s-favoring training in Exp 1c

Fixed Effect	Est.	SE	z	Pr(> z)
(Intercept)	1.72	0.37	4.70	< 0.001***
Step	-1.54	0.13	-11.46	< 0.001***
Condition Two genders - opposite	-2.16	0.45	-4.80	< 0.001***
Condition Two genders - same	0.37	0.47	0.79	0.43
Trial	-0.31	0.09	-3.66	< 0.001***
PhonemeS	-0.55	0.05	-10.50	< 0.001***
Step:Condition Two genders - opposite	0.22	0.18	1.21	0.23
Step:Condition Two genders - same	0.23	0.19	1.20	0.23
Condition Two genders - opposite:Trial	0.09	0.12	0.79	0.43
Condition Two genders - same:Trial	0.20	0.13	1.57	0.12

Model-1c-c: Response~Step*Group+Group*Trial+Phoneme+(Step|Subj)+(1|Word)

Table B.2: The fixed effects of Model 1c-c for conditions including Male sh-favoring training in Exp 1c

Fixed Effect	Est.	SE	z	Pr(> z)
(Intercept)	-0.84	0.40	-2.12	0.03*
Step	-1.73	0.19	-9.17	< 0.001***
Condition Two genders - opposite	0.27	0.51	0.53	0.60
Condition Two genders - same	-0.93	0.52	-1.79	0.07
Trial	-0.01	0.08	-0.14	0.89
Step:Condition Two genders - opposite	0.18	0.26	0.67	0.50
Step:Condition Two genders - same	0.21	0.27	0.77	0.44
Condition Two genders - opposite:Trial	-0.08	0.12	-0.65	0.51
Condition Two genders - same:Trial	-0.02	0.12	-0.18	0.86

Model-2c-b: Response~Step*Group+Group*Trial+Phoneme+(Step|Subj)+(1|Word)

Table B.3: The fixed effects of Model 2c-b for conditions including Male s-favoring training in Exp 2c

Fixed Effect	Est.	SE	z	Pr(> z)
(Intercept)	0.55	0.47	1.18	0.24
Step	-2.27	0.20	-11.21	< 0.001***
Condition Two genders - opposite	-1.00	0.58	-1.72	0.09
Condition Two genders - same	-0.69	0.57	-1.20	0.23
Trial	-0.12	0.10	-1.21	0.23
Step:Condition Two genders - opposite	0.78	0.26	2.98	0.003**
Step:Condition Two genders - same	0.50	0.26	1.88	0.06
Condition Two genders - opposite:Trial	0.23	0.13	1.74	0.08
Condition Two genders - same:Trial	0.16	0.13	1.22	0.22

Model-2c-c: Response~Step*Group+Group*Trial+Phoneme+(Step|Subj)+(1|Word)

Table B.4: The fixed effects of Model 2c-c for conditions including Male sh-favoring training in Exp 2c

Fixed Effect	Est.	SE	z	Pr(> z)
(Intercept)	0.35	0.39	0.91	0.37
Step	-2.52	0.42	-5.98	< 0.001***
Condition Male A t-favoring	-1.42	0.56	-2.53	0.01*
Condition Female A t-favoring	-3.45	0.77	-4.51	< 0.001***
Condition Two genders - same	-1.48	0.56	-2.64	0.008**
Trial	-0.56	0.23	-2.43	0.01514 *
Step:Condition Male A t-favoring	-0.42	0.54	-0.77	0.44
Step:Condition Female A t-favoring	-2.57	0.82	-3.13	0.002**
Step:Condition Two genders - same	-0.31	0.54	-0.58	0.56
Condition Male A t-favoring: Trial	0.24	0.33	0.73	0.47
Condition Female A t-favoring: Trial	0.34	0.41	0.83	0.41
Condition Two genders - same: Trial	0.31	0.33	0.92	0.36

Model 3c-tear: Response~Step*Group+Group*Trial+Phoneme+(Step|Subj)

Table B.5: The fixed effects of Model 3c-tear for responses to stimuli in the word frame of “tear” in Exp 3c

Fixed Effect	Est.	SE	z	Pr(> z)
(Intercept)	1.02	0.35	2.93	0.003**
Step	-0.88	0.20	-4.32	< 0.001***
Condition Male A t-favoring	-0.33	0.49	-0.68	0.50
Condition Female A t-favoring	-0.93	0.50	-1.86	0.06
Condition Two genders - same	-0.64	0.49	-1.31	0.19
Trial	-0.06	0.18	-0.35	0.73
Step:Condition Male A t-favoring	-0.05	0.28	-0.18	0.85
Step:Condition Female A t-favoring	-0.58	0.31	-1.87	0.06
Step:Condition Two genders - same	-0.44	0.30	-1.48	0.14
Condition Male A t-favoring: Trial	0.43	0.28	1.55	0.12
Condition Female A t-favoring: Trial	-0.02	0.27	-0.07	0.95
Condition Two genders - same: Trial	-0.08	0.28	-0.28	0.78

Model 3c-time: Response~Step*Group+Group*Trial+Phoneme+(Step|Subj)

Table B.6: The fixed effects of Model 3c-time for responses to stimuli in the word frame of “time” in Exp 3c

Fixed Effect	Est.	SE	z	Pr(> z)
(Intercept)	-0.94	0.30	-3.09	0.002**
Step	-1.72	0.31	-5.62	1.9e-08 ***
Condition Male A t-favoring	0.12	0.43	0.28	0.78
Condition Female A t-favoring	-1.27	0.47	-2.72	0.006**
Condition Two genders - same	-0.90	0.44	-2.03	0.04*
Trial	-0.55	0.20	-2.79	0.005**
Step:Condition Male A t-favoring	-0.41	0.41	-1.00	0.32
Step:Condition Female A t-favoring	-0.16	0.44	-0.37	0.72
Step:Condition Two genders - same	-0.47	0.43	-1.11	0.27
Condition Male A t-favoring: Trial	0.28	0.30	0.94	0.35
Condition Female A t-favoring: Trial	0.31	0.34	0.90	0.37
Condition Two genders - same: Trial	-0.05	0.31	-0.16	0.87

Model 3c-town: Response~Step*Group+Group*Trial+Phoneme+(Step|Subj)

Table B.7: The fixed effects of Model 3c-town for responses to stimuli in the word frame of “town” in Exp 3c

Fixed Effect	Est.	SE	z	Pr(> z)
(Intercept)	0.05	0.30	0.16	0.87
Step	-1.28	0.27	-4.78	< 0.001***
Condition Male A t-favoring	-0.52	0.42	-1.24	0.22
Condition Female A t-favoring	-1.00	0.44	-2.28	0.02*
Condition Two genders - same	-1.61	0.44	-3.65	< 0.001***
Trial	0.07	0.16	0.43	0.67
Step:Condition Male A t-favoring	-0.43	0.37	-1.15	0.25
Step:Condition Female A t-favoring	-0.82	0.40	-2.07	0.04*
Step:Condition Two genders - same	-0.83	0.39	-2.12	0.03*
Condition Male A t-favoring: Trial	0.22	0.25	0.90	0.37
Condition Female A t-favoring: Trial	0.08	0.26	0.31	0.76
Condition Two genders - same: Trial	0.12	0.27	0.45	0.65

Model 3c-touch: Response~Step*Group+Group*Trial+Phoneme+(Step|Subj)

Table B.8: The fixed effects of Model 3c-touch for responses to stimuli in the word frame of “touch” in Exp 3c

Fixed Effects	Est.	SE	z value	Pr(> z)
(Intercept)	0.11	0.38	0.29	0.77
Condition Two females opposite auditory	-0.19	0.48	-0.39	0.70
Condition Female A s-favoring	-2.16	0.48	-4.49	< 0.001***
Condition Two females opposite audiovisual	-0.69	0.47	-1.47	0.14
Condition Two genders opposite auditory	0.55	0.47	1.16	0.25
Step	-2.22	0.21	-10.73	< 0.001***
Trial	-0.48	0.10	-5.00	< 0.001***
PhonemeS	-0.54	0.04	-13.03	< 0.001***
Condition Two females opposite auditory:Step	0.58	0.28	2.07	0.04*
Condition Female A s-favoring:Step	0.45	0.29	1.55	0.12
Condition Two females opposite audiovisual:Step	0.63	0.28	2.29	0.02*
Condition Two genders opposite auditory:Step	0.56	0.28	2.01	0.04*
Condition Two females opposite auditory:Trial	0.32	0.13	2.51	0.01*
Condition Female A s-favoring:Trial	0.64	0.14	4.68	< 0.001***
Condition Two females opposite audiovisual:Trial	0.19	0.13	1.44	0.15
Condition Two genders opposite auditory:Trial	0.07	0.13	0.51	0.61

Table B.9: The fixed effects of Model 4a-*relevel*

Fixed Effects	Est.	SE	z value	Pr(> z)
(Intercept)	0.06	0.36	0.16	0.87
Step	-1.37	0.29	-4.69	< 0.001***
Condition Female A t-favoring	-1.17	0.34	-3.42	< 0.001***
Condition Two females opposite audiovisual	-0.02	0.34	-0.06	0.95
Condition Two females opposite auditory	-0.04	0.32	-0.13	0.90
Condition Two genders opposite auditory	-0.73	0.33	-2.23	0.03*
Trial	-0.19	0.08	-2.42	0.02*
Step:Condition Female A t-favoring	-0.62	0.21	-3.01	0.003**
Step:Condition Two females opposite audiovisual	-0.01	0.19	-0.07	0.94
Step:Condition Two females opposite auditory	-0.08	0.18	-0.43	0.67
Step:Condition Two genders opposite auditory	-0.22	0.19	-1.14	0.25
Condition Female A t-favoring:Trial	0.15	0.13	1.16	0.25
Condition Two females opposite audiovisual:Trial	0.10	0.13	0.80	0.43
Condition Two females opposite auditory:Trial	0.09	0.12	0.72	0.47
Condition Two genders opposite auditory:Trial	-0.07	0.12	-0.59	0.55

Model 4b-*relevel*: Response~Step*Group+Group*Trial+(Step|Subj)+(Step|Word)

Table B.10: The fixed effects of Model 4b-*relevel*

Bibliography

- Allen, J. S. and Miller, J. L. (1999). Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *The Journal of the Acoustical Society of America*, 106(4):2031–2039.
- Allen, J. S. and Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, 115(6):3171–3183.
- Allen, J. S., Miller, J. L., and DeSteno, D. (2003). Individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America*, 113(1):544–552.
- Babel, M., McAuliffe, M., Norton, C., Senior, B., and Vaughn, C. (2019). The Goldilocks zone of perceptual learning. *Phonetica*, 76(2-3):179–200.
- Baese-Berk, M. and Goldrick, M. (2009). Mechanisms of interaction in speech production. *Language and cognitive processes*, 24(4):527–554.
- Baese-Berk, M. M., Bradlow, A. R., and Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America*, 133(3):EL174–EL180.
- Beckner, C. and Bybee, J. (2009). A usage-based account of constituency and reanalysis. *Language Learning*, 59:27–46.
- Behrens, S. J. and Blumstein, S. E. (1988). Acoustic characteristics of English voiceless fricatives: A descriptive analysis. *Journal of Phonetics*, 16(3):295–298.
- Benjamin, B. J. (1982). Phonological performance in gerontological speech. *Journal of Psycholinguistic Research*, 11(2):159–167.

- Bertelson, P., Vroomen, J., and de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk after-effect. *Psychological Science*, 14(6):592–597.
- Birnbaum, M. H. (2001). *Introduction to behavioral research on the Internet*. Pearson College Division.
- Birnbaum, M. H. (2004). Human research and data collection via the internet. *Annu. Rev. Psychol.*, 55:803–832.
- Bradlow, A. R. and Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106:707–729.
- Bradlow, A. R., Torretta, G. M., and Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech communication*, 20(3-4):255–272.
- Brady, S. A. and Darwin, C. J. (1978). Range effects in the perception of voicing. *Journal of the Acoustical Society of America*, 63(5):1556–1558.
- Brysbaert, M. and New, B. (2009). Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavioral Research Methods*, 41(4):977–90.
- Buz, E., Tanenhaus, M. K., and Jaeger, T. F. (2016). Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers’ subsequent pronunciations. *Journal of memory and language*, 89:68–86.
- Byrd, D. (1992). Preliminary results on speaker-dependent variation in the timit database. *The Journal of the Acoustical Society of America*, 92(1):593–596.
- Byrd, D., Krivokapić, J., and Lee, S. (2006). How far, how long: On the temporal scope of prosodic boundary effects. *Journal of the Acoustical Society of America*, 120:1589–1599.
- Calder, J. (2019a). The fierceness of fronted /s/: Linguistic rhematization through visual transformation. *Language in Society*, 48(1):31–64.

- Calder, J. (2019b). From sissy to sickening: The indexical landscape of /s/ in SoMa, San Francisco. *Journal of Linguistic Anthropology*, 29(3):332–358.
- Caplan, S., Hafri, A., and Trueswell, J. C. (2021). Now you hear me, later you don't: The immediacy of linguistic computation and the representation of speech. *Psychological Science*, page 0956797620968787.
- Carlbring, P., Brunt, S., Bohman, S., Austin, D., Richards, J., Öst, L.-G., and Andersson, G. (2007). Internet vs. paper and pencil administration of questionnaires commonly used in panic/agoraphobia research. *Computers in Human Behavior*, 23(3):1421–1434.
- Cho, T. and Keating, P. (2009). Effects of initial position versus prominence in English. *Journal of Phonetics*, 37(4):466–485.
- Cho, T. and Ladefoged, P. (1999). Variation and universals in VOT: evidence from 18 languages. *Journal of phonetics*, 27(2):207–229.
- Chodroff, E. and Wilson, C. (2014). Burst spectrum as a cue for the stop voicing contrast in American English. *The Journal of the Acoustical Society of America*, 136(5):2762–2772.
- Chodroff, E. and Wilson, C. (2017). Structure in talker-specific phonetic realization: Co-variation of stop consonant VOT in American English. *Journal of Phonetics*, 61:30–47.
- Chomsky, N. and Halle, M. (1968). *The sound pattern of English*. Harper & Row, New York.
- Church, B. A. and Schacter, D. L. (1994). Perceptual specificity of auditory priming: implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(3):521.
- Clarke, C. M. and Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, 116(6):3647–3658.
- Clarke-Davidson, C. M., Luce, P. A., and Sawusch, J. R. (2008). Does perceptual learning in

- speech reflect changes in phonetic category representation or decision bias? *Perception & psychophysics*, 70(4):604–618.
- Cole, J., Kim, H., Choi, H., and Hasegawa-Johnson, M. (2007). Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from radio news speech. *Journal of Phonetics*, 35(2):180–209.
- Cutler, A., McQueen, K., Butterfield, S., and Norris, D. (2008). Prelexically-driven perceptual retuning of phoneme boundaries. In *Proceedings of Interspeech 2008*, Brisbane, Australia.
- Cutler, A., Mehler, J., Norris, D., and Segui, J. (1987). Phoneme identification and the lexicon. *Cognitive Psychology*, 19(2):141–177.
- Davidson, L. (2016). Variability in the implementation of voicing in American English obstruents. *Journal of Phonetics*, 54:35–50.
- Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, 134(2):222.
- De Jong, K. (2004). Stress, lexical focus, and segmental focus in English: patterns of variation in vowel duration. *Journal of Phonetics*, 32(4):493–516.
- Docherty, G. J. (2011). *The timing of voicing in British English obstruents*. De Gruyter Mouton.
- Eimas, P. D. and Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, 4(1):99–109.
- Eisner, F. and McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & psychophysics*, 67(2):224–238.

- Eisner, F. and McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America*, 119(4):1950–1953.
- Flege, J. E., Frieda, E. M., Walley, A. C., and Randazza, L. A. (1998). Lexical factors and segmental accuracy in second language speech production. *Studies in Second Language Acquisition*, pages 155–187.
- Flipsen Jr, P., Shriberg, L., Weismer, G., Karlsson, H., and McSweeney, J. (1999). Acoustic characteristics of /s/ in adolescents. *Journal of Speech, Language, and Hearing Research*, 42(3):663–677.
- Foss, D. J. and Blank, M. A. (1980). Identifying the speech codes. *Cognitive Psychology*, 12(1):1–31.
- Foss, D. J. and Gernsbacher, M. A. (1983). Cracking the dual code: Toward a unitary model of phoneme identification. *Journal of Verbal Learning and Verbal Behavior*, 22(6):609–632.
- Foss, D. J. and Swinney, D. A. (1973). On the psychological reality of the phoneme: Perception, identification, and consciousness. *Journal of Verbal Learning and Verbal Behavior*, 12(3):246–257.
- Fox, R. A. (1984). Effect of lexical status on phonetic categorization. *Journal of Experimental Psychology: Human perception and performance*, 10(4):526.
- Fromkin, V. (1973). *Slips of the tongue*. WH Freeman San Francisco, CA.
- Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of experimental psychology: Human perception and performance*, 6(1):110.
- Garrett, M. F. (1976). Syntactic processes in sentence production. *New approaches to language mechanisms*, 30:231–256.
- Gibson, J. and Gibson, E. (1955). Perceptual learning: Differentiation or enrichment? *Psychological Review*, 62:32–41.

- Goldinger, S. D. (1996). Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of experimental psychology: Learning, memory, and cognition*, 22(5):1166.
- Goldinger, S. D. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological review*, 105(2):251.
- Goldstone, R. (1998). Perceptual learning. *Annual Review of Psychology*, 49:585–612.
- González, J. and McLennan, C. T. (2007). Hemispheric differences in indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 33(2):410.
- Haggard, M., Ambler, S., and Callow, M. (1970). Pitch as a voicing cue. *The Journal of the Acoustical Society of America*, 47(2B):613–617.
- Hall, K., Borba, R., and Hiramoto, M. (2020). Language and gender. *The International Encyclopedia of Linguistic Anthropology*, pages 1–22.
- Halle, M. (1985). Speculations about the representation of words in memory. *Phonetic linguistics*, pages 101–114.
- Halle, M., Hughes, G. W., and Radley, J.-P. (1957). Acoustic properties of stop consonants. *The Journal of the Acoustical Society of America*, 29(1):107–116.
- Hallé, P. A. and Best, C. T. (2007). Dental-to-velar perceptual assimilation: A cross-linguistic study of the perception of dental stop+/l/clusters. *The Journal of the Acoustical Society of America*, 121(5):2899–2914.
- Harrington, J., Kleber, F., Reubold, U., Schiel, F., and Stevens, M. (2018). Linking cognitive and social aspects of sound change using agent-based modeling. *Topics in cognitive science*, 10(4):707–728.
- Hay, J. and Foulkes, P. (2016). The evolution of medial /t/ over real and remembered time. *Language*, 92(2):298–330.

- Hay, J. B., Pierrehumbert, J. B., Walker, A. J., and LaShell, P. (2015). Tracking word frequency effects through 130 years of sound change. *Cognition*, 139:83–91.
- Hedrick, M. S. and Ohde, R. N. (1993). Effect of relative amplitude of frication on perception of place of articulation. *The Journal of the Acoustical Society of America*, 94(4):2005–2026.
- Heinz, J. M. and Stevens, K. N. (1961). On the properties of voiceless fricative consonants. *The Journal of the Acoustical Society of America*, 33(5):589–596.
- Hoit, J. D., Solomon, N. P., and Hixon, T. J. (1993). Effect of lung volume on voice onset time (VOT). *Journal of Speech, Language, and Hearing Research*, 36(3):516–520.
- Honing, H. and Reips, U.-D. (2008). Web-based versus lab-based studies: A response to kendall (2008). *Empirical Musicology Review*, 3:73–77.
- House, A. S. and Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Journal of the Acoustical Society of America*, 25(1):105–113.
- Hughes, G. W. and Halle, M. (1956). Spectral properties of fricative consonants. *The journal of the acoustical society of America*, 28(2):303–310.
- Jackson, A. and Morton, J. (1984). Facilitation of auditory word recognition. *Memory & Cognition*, 12(6):568–574.
- Jaeger, T. F. and Weatherholtz, K. (2016). What the heck is salience? How predictive language processing contributes to sociolinguistic perception. *Frontiers in Psychology*, 7:1115.
- Jesse, A. and McQueen, J. M. (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic bulletin & review*, 18(5):943–950.
- Johnson, K. (1997). The auditory/perceptual basis for speech segmentation. *OSU Working Papers in Linguistics*, pages 101–103.

- Jongman, A. (1989). Duration of frication noise required for identification of English fricatives. *The Journal of the Acoustical Society of America*, 85(4):1718–1725.
- Jongman, A., Wayland, R., and Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, 108(3):1252–1263.
- Joos, M. (1948). Acoustic phonetics. *Language*, 24(2):5–136.
- Keating, P. (1979). Differences in production and perception of VOT in Polish and English. *The Journal of the Acoustical Society of America*, 66(S1):S87–S87.
- Keating, P. A., Mikós, M. J., and Ganong, III, W. F. (1981). A cross-language study of range of voice onset time in the perception of initial stop voicing. *Journal of the Acoustical Society of America*, 70(5):1261–1271.
- Kessinger, R. H. and Blumstein, S. E. (1997). Effects of speaking rate on voice-onset time in Thai, French, and English. *Journal of Phonetics*, 25(2):143–168.
- Kessinger, R. H. and Blumstein, S. E. (1998). Effects of speaking rate on voice-onset time and vowel production: Some implications for perception studies. *Journal of Phonetics*, 26(2):117–128.
- Kirov, C. and Wilson, C. (2012). The specificity of online variation in speech production. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34.
- Klatt, D. H. (1975). Voice onset time, frication, and aspiration in word-initial consonant clusters. *Journal of Speech and Hearing Research*, 18(4):686–706.
- Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, 59(5):1208–1221.
- Klatt, D. H. (1980). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Perception and production of fluent speech*, pages 243–288.
- Kleinschmidt, D. (2017). *Perception in a variable but structured world: The case of speech perception*. PhD thesis, University of Rochester.

- 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.
- Kleinschmidt, D. F. and Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological review*, 122(2):148.
- Kleinschmidt, D. F., Raizada, R. D., and Jaeger, T. F. (2015). Supervised and unsupervised learning in phonetic adaptation. In *CogSci*.
- Kluender, K. R. (1991). Effects of first formant onset properties on voicing judgments result from processes not specific to humans. *The Journal of the Acoustical Society of America*, 90(1):83–96.
- Koenig, L. L. (2000). Laryngeal factors in voiceless consonant production in men, women, and 5-year-olds. *Journal of Speech, Language, and Hearing Research*, 43(5):1211–1228.
- Kraljic, T., Brennan, S. E., and Samuel, A. G. (2008a). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, 107(1):54–81.
- Kraljic, T. and Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive psychology*, 51(2):141–178.
- Kraljic, T. and Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic bulletin & review*, 13(2):262–268.
- Kraljic, T. and Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, 56(1):1–15.
- Kraljic, T. and Samuel, A. G. (2011). Perceptual learning evidence for contextually-specific representations. *Cognition*, 121(3):459–465.
- Kraljic, T., Samuel, A. G., and Brennan, S. E. (2008b). First impressions and last resorts: How listeners adjust to speaker variability. *Psychological science*, 19(4):332–338.

- Lacerda, F. et al. (1995). The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory. In *Proceedings of the XIIIth international congress of phonetic sciences*, volume 2, pages 140–147. Stockholm University Stockholm.
- Ladefoged, P. (1967). *Three areas of experimental phonetics: Stress and respiratory activity, the nature of vowel quality, units in the perception and production of speech*. Oxford University Press Oxford.
- Ladefoged, P. and Broadbent, D. (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America*, 29(1):98–104.
- Lai, W., Rácz, P., and Roberts, G. (2020). Experience with a linguistic variant affects the acquisition of its sociolinguistic meaning: An alien-language-learning experiment. *Cognitive science*, 44(4):e12832.
- Liberman, A., Cooper, F., Shankweiler, D., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6):431–461.
- Liberman, A. M., Delattre, P. C., and Cooper, F. S. (1958). Some cues for the distinction between voiced and voiceless stops in initial position. *Language and speech*, 1(3):153–167.
- Liberman, I. Y., Shankweiler, D., Fischer, F. W., and Carter, B. (1974). Explicit syllable and phoneme segmentation in the young child. *Journal of experimental child psychology*, 18(2):201–212.
- Lindblom, B., Guion, S., Hura, S., Moon, S.-J., and Willerman, R. (1995). Is sound change adaptive? *Rivista di linguistica*, 7:5–36.
- Lipani, L., Olsen, M., and Olsen, R. M. (2019). Voice onset time variation in natural southern speech. *The Journal of the Acoustical Society of America*, 146(4):3011–3011.
- Lisker, L. (1977). Rapid versus rabid: A catalogue of acoustic features that may cue the distinction. *The Journal of the Acoustical Society of America*, 62(S1):S77–S78.

- Lisker, L. (1986). “Voicing” in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and speech*, 29(1):3–11.
- Lisker, L. and Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20(3):384–422.
- Luce, P. A. and Lyons, E. A. (1998). Specificity of memory representations for spoken words. *Memory & Cognition*, 26(4):708–715.
- Luce, P. A., Pisoni, D. B., and Goldinger, S. D. (1990-01-01). *Cognitive models of speech processing: psycholinguistic and computational perspectives*, chapter 6. Similarity neighborhoods of spoken words. The MIT Press.
- Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, 25(1-2):71–102.
- Martin, C. S., Mullennix, J. W., Pisoni, D. B., and Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(4):676.
- Massaro, D. W. and Cohen, M. M. (1983). Phonological context in speech perception. *Attention, Perception, & Psychophysics*, 34(4):338–348.
- Mattys, S. L. and Wiget, L. (2011). Effects of cognitive load on speech recognition. *Journal of memory and Language*, 65(2):145–160.
- Maye, J., Aslin, R. N., and Tanenhaus, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science*, 32(3):543–562.
- Maye, J. and Gerken, L. (2001). Learning phonemes: How far can the input take us. In *Proceedings of the 25th annual Boston University conference on language development*, volume 1, page 480. Citeseer.
- Maye, J., Werker, J. F., and Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3):B101–B111.

- McAuliffe, M. (2015). *Attention and salience in lexically-guided perceptual learning*. PhD thesis, University of British Columbia.
- McAuliffe, M. and Babel, M. (2016). Stimulus-directed attention attenuates lexically-guided perceptual learning. *The Journal of the Acoustical Society of America*, 140(3):1727–1738.
- McFarland, D. H., Baum, S. R., and Chabot, C. (1996). Speech compensation to structural modifications of the oral cavity. *The Journal of the Acoustical Society of America*, 100(2):1093–1104.
- McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588):746–748.
- McLennan, C. T. and Luce, P. A. (2005). Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(2):306.
- McMurray, B., Tanenhaus, M. K., and Aslin, R. N. (2009). Within-category VOT affects recovery from "lexical" garden paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language*, 60(1):65–91.
- Miller, J. L., Green, K. P., and Reeves, A. (1986). Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica*, 43(1-3):106–115.
- Morris, R. J. and Brown Jr, W. (1994). Age-related differences in speech variability among women. *Journal of Communication Disorders*, 27(1):49–64.
- Morris, R. J., McCrea, C. R., and Herring, K. D. (2008). Voice onset time differences between adult males and females: Isolated syllables. *Journal of Phonetics*, 36(2):308–317.

- Mullennix, J. W., Pisoni, D. B., and 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.
- Munson, C. M. (2011). *Perceptual learning in speech reveals pathways of processing*. PhD thesis, University of Iowa.
- Myers, E. B. and Theodore, R. M. (2017). Voice-sensitive brain networks encode talker-specific phonetic detail. *Brain and language*, 165:33–44.
- Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America*, 85(5):2088–2113.
- Newman, R. S., Clouse, S. A., and Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *J Acoust Soc Am*, 109(3):1181–1196.
- Nittrouer, S. (1999). Do temporal processing deficits cause phonological processing problems? *Journal of Speech, Language, and Hearing Research*, 42(4):925–942.
- Nittrouer, S. (2002). Learning to perceive speech: How fricative perception changes, and how it stays the same. *The Journal of the Acoustical Society of America*, 112(2):711–719.
- Nittrouer, S. and Studdert-Kennedy, M. (1987). The role of coarticulatory effects in the perception of fricatives by children and adults. *Journal of Speech, Language, and Hearing Research*, 30(3):319–329.
- Nordström, P.-E. and Lindblom, B. (1975). *A normalization procedure for vowel formant data*. Univ.
- Norris, D. and McQueen, J. M. (2008). Shortlist b: a Bayesian model of continuous speech recognition. *Psychological review*, 115(2):357.
- Norris, D., McQueen, J. M., and Cutler, A. (2003). Perceptual learning in speech. *Cognitive psychology*, 47(2):204–238.

- Nusbaum, H. C. and Schwab, E. C. (1986). The role of attention and active processing in speech perception. In *Pattern recognition by humans and machines*, pages 113–157. Elsevier.
- Nygaard, L. C. and Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & psychophysics*, 60(3):355–376.
- Palan, S. and Schitter, C. (2018). Prolific. ac — a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.
- Palmeri, T. J., Goldinger, S. D., and Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(2):309.
- Parikh, G. and Loizou, P. C. (2005). The influence of noise on vowel and consonant cues. *The Journal of the Acoustical Society of America*, 118(6):3874–3888.
- Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the acoustical society of America*, 24(2):175–184.
- Peterson, G. E. and Lehiste, I. (1960). Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America*, 32(6):693–703.
- Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. *Typological studies in language*, 45:137–158.
- Pisoni, D. B. and Levi, S. V. (2007). Some observations on representations and representational specificity in speech perception and spoken word recognition. *The Oxford handbook of psycholinguistics*, pages 3–18.
- Pitt, M., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., and Fosler-Lussier, E. (2007). *Buckeye Corpus of Conversational Speech (2nd release)*. Columbus, OH: Department of Psychology, Ohio State University (Distributor).

- Pitt, M. A. and Samuel, A. G. (2006). Word length and lexical activation: Longer is better. *Journal of Experimental Psychology: Human Perception and Performance*, 32(5):1120.
- Pitt, M. A. and Szostak, C. M. (2012). A lexically biased attentional set compensates for variable speech quality caused by pronunciation variation. *Language and Cognitive Processes*, 27(7-8):1225–1239.
- Port, R. F. and Rotunno, R. (1979). Relation between voice-onset time and vowel duration. *The Journal of the Acoustical Society of America*, 66(3):654–662.
- Reinisch, E. and 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.
- Reinisch, E., Wozny, D. R., Mitterer, H., and Holt, L. L. (2014). Phonetic category recalibration: What are the categories? *Journal of phonetics*, 45:91–105.
- Reips, U.-D. (2002). Standards for internet-based experimenting. *Experimental psychology*, 49(4):243.
- Repp, B. H. (1979). Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. *Language and Speech*, 22(2):173–189.
- Robert, L. (1965). Chapter 9: Machine perception of three-dimensional solids. *Optical and Electro-Optical Information processing. The MIT Press, Cambridge, Massachusetts and London, England*.
- Rosen, S. M. (1979). Range and frequency effects in consonant categorization. *Journal of Phonetics*, 7:393–402.
- Saltzman, D. and Myers, E. (2018). Listeners are maximally flexible in updating phonetic beliefs over time. *Psychonomic bulletin & review*, 25(2):718–724.
- Samuel, A. G. (1981). Phonemic restoration: insights from a new methodology. *Journal of Experimental Psychology: General*, 110(4):474.

- Samuel, A. G. and Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, 71(6):1207–1218.
- Schacter, D. L. and Church, B. A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(5):915.
- Scharenborg, O. and Janse, E. (2013). Comparing lexically guided perceptual learning in younger and older listeners. *Attention, Perception, & Psychophysics*, 75(3):525–536.
- Scobbie, J. M. (2006). Flexibility in the face of incompatible English VOT systems. *Laboratory Phonology 8 Varieties of Phonological Competence*.
- Shankweiler, D., Strange, W., and Verbrugge, R. (1977). Speech and the problem of perceptual constancy. *Perceiving, acting, and knowing: Toward an ecological psychology*, pages 315–345.
- Shattuck-Hufnagel, S. and Klatt, D. H. (1979). The limited use of distinctive features and markedness in speech production: Evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior*, 18(1):41–55.
- Shultz, A. A., Francis, A. L., and Llanos, F. (2012). Differential cue weighting in perception and production of consonant voicing. *The Journal of the Acoustical Society of America*, 132(2):EL95–EL101.
- Smith, B. L. (1978). Temporal aspects of English speech production: A developmental perspective. *Journal of Phonetics*, 6(1):37–67.
- Sohoglu, E. and Davis, M. H. (2016). Perceptual learning of degraded speech by minimizing prediction error. *Proceedings of the National Academy of Sciences*, 113(12):E1747–E1756.
- Stevens, K. N. and House, A. S. (1956). Studies of formant transitions using a vocal tract analog. *The Journal of the Acoustical Society of America*, 28(4):578–585.

- Stevens, K. N. and Klatt, D. H. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. *The Journal of the Acoustical Society of America*, 55(3):653–659.
- Strand, E. A. and Johnson, K. (1996). Gradient and visual speaker normalization in the perception of fricatives. In *KONVENS*, pages 14–26.
- Stevens, P. (1960). Spectra of fricative noise in human speech. *Language and speech*, 3(1):32–49.
- Summerfield, Q. and Haggard, M. (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *The Journal of the Acoustical Society of America*, 62(2):435–448.
- Sumner, M., Kim, S. K., King, E., and McGowan, K. B. (2014). The socially weighted encoding of spoken words: a dual-route approach to speech perception. *Frontiers in Psychology*, 4.
- Sundara, M. (2005). Acoustic-phonetics of coronal stops: A cross-language study of Canadian English and Canadian French. *The Journal of the Acoustical Society of America*, 118(2):1026–1037.
- Swartz, B. L. (1992). Gender difference in voice onset time. *Perceptual and motor skills*, 75(3):983–992.
- Sweeting, P. M. and Baken, R. J. (1982). Voice onset time in a normal-aged population. *Journal of Speech, Language, and Hearing Research*, 25(1):129–134.
- Tamminga, M., Wilder, R., Lai, W., and Wade, L. (2020). Perceptual learning, talker specificity, and sound change. *Papers in Historical Phonology*, 5:90–122.
- Theodore, R. M. and Miller, J. L. (2010). Characteristics of listener sensitivity to talker-specific phonetic detail. *The Journal of the Acoustical Society of America*, 128(4):2090–2099.

- Theodore, R. M., Miller, J. L., and DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *The Journal of the Acoustical Society of America*, 125(6):3974–3982.
- Theodore, R. M. and Monto, N. R. (2019). Distributional learning for speech reflects cumulative exposure to a talker’s phonetic distributions. *Psychonomic bulletin & review*, pages 1–8.
- Theodore, R. M., Myers, E. B., and Lomibao, J. A. (2015). Talker-specific influences on phonetic category structure. *Journal of the Acoustical Society of America*, 138(2):1068–1078.
- Tillery, A. K. H. (2015). *Individual Differences in the Perceptual Learning of Degraded Speech: Implications for Cochlear Implant Aural Rehabilitation*. Arizona State University.
- Todd, S., Pierrehumbert, J. B., and Hay, J. (2019). Word frequency effects in sound change as a consequence of perceptual asymmetries: An exemplar-based model. *Cognition*, 185:1–20.
- Tomiak, G. R. (1990). An evaluation of a spectral moments metric with voiceless fricative obstruents. *The Journal of the Acoustical Society of America*, 87(S1):S106–S107.
- Torre III, P. and Barlow, J. A. (2009). Age-related changes in acoustic characteristics of adult speech. *Journal of communication disorders*, 42(5):324–333.
- Toscano, J. C. and McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive science*, 34(3):434–464.
- Treiman, R. (1983). The structure of spoken syllables: Evidence from novel word games. *Cognition*, 15(1-3):49–74.

- Treiman, R. (1985). Onsets and rimes as units of spoken syllables: Evidence from children. *Journal of experimental child psychology*, 39(1):161–181.
- Vadillo, M. A. and Matute, H. (2011). Further evidence on the validity of web-based research on associative learning: Augmentation in a predictive learning task. *Computers in Human Behavior*, 27(2):750–754.
- van Linden, S. and Vroomen, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception and Performance*, 33:1483–1494.
- Van Linden, S. and Vroomen, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception and Performance*, 33(6):1483.
- Vroomen, J. and Baart, M. (2009a). Phonetic recalibration only occurs in speech mode. *Cognition*, 110:254–259.
- Vroomen, J. and Baart, M. (2009b). Recalibration of phonetic categories by lipread speech: Measuring aftereffects after a 24-hour delay. *Language and Speech*, 52:341–350.
- Vroomen, J., van Linden, S., de Gelder, B., and Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory-visual speech perception: Contrasting build-up courses. *Neuropsychologia*, 45(572–577).
- Vroomen, J., van Linden, S., Keetels, M., de Gelder, B., and Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: Dissipation. *Speech Communication*, 44:55–61.
- Weatherholtz, K. (2015). *Perceptual learning of systemic cross-category vowel variation*. PhD thesis, The Ohio State University.
- Weismer, G. (1979). Sensitivity of voice-onset time (VOT) measures to certain segmental features in speech production. *Journal of Phonetics*, 7(2):197–204.

- Whalen, D. H. (1991). Perception of the English /s-/ʃ/ distinction relies on fricative noises and transitions, not on brief spectral slices. *The Journal of the Acoustical Society of America*, 90(4):1776–1785.
- Whalen, D. H., Abramson, A. S., Lisker, L., and Mody, M. (1993). F0 gives voicing information even with unambiguous voice onset times. *The Journal of the Acoustical Society of America*, 93(4):2152–2159.
- Whitaker, B. G. (2007). Internet-based attitude assessment: does gender affect measurement equivalence? *Computers in Human Behavior*, 23(3):1183–1194.
- Whiteside, S. P., Henry, L., and Dobbin, R. (2004). Sex differences in voice onset time: A developmental study of phonetic context effects in British English. *The Journal of the Acoustical Society of America*, 116(2):1179–1183.
- Whiteside, S. P. and Irving, C. J. (1997). Speakers’ sex differences in voice onset time: Some preliminary findings. *Perceptual and motor skills*, 85(2):459–463E.
- Whiteside, S. P. and Irving, C. J. (1998). Speakers’ sex differences in voice onset time: a study of isolated word production. *Perceptual and motor skills*, 86(2):651–654.
- Wilder, R. (2018). *Investigating Hybrid Models of Speech Perception*. PhD thesis, University of Pennsylvania.
- Witteman, M. J., Weber, A., and McQueen, J. M. (2013). Foreign accent strength and listener familiarity with an accent codetermine speed of perceptual adaptation. *Attention, Perception, & Psychophysics*, 75(3):537–556.
- Xie, X., Weatherholtz, K., Bainton, L., Rowe, E., Burchill, Z., Liu, L., and Jaeger, T. F. (2018). Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker. *The Journal of the Acoustical Society of America*, 143(4):2013–2031.
- Yao, Y. (2009). Understanding VOT variation in spontaneous speech. *UC Berkeley PhonLab Annual Report*, 5(5).

- Zehr, J. and Schwarz, F. (2018). Penncontroller for internet based experiments (Ibex). *URL* <https://doi.org/10.17605/OSF.IO/MD832>.
- Zhang, X. and 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.
- Zimman, L. (2017). Variability in /s/ among transgender speakers: Evidence for a socially grounded account of gender and sibilants. *Linguistics*, 55(5):993–1019.
- Zimmerman, S. A. and Sapon, S. M. (1958). Note on vowel duration seen cross-linguistically. *The Journal of the Acoustical Society of America*, 30(2):152–153.
- Zue, V. W. (1976). Acoustic characteristics of stop consonants: A controlled study. Technical report, Massachusetts Inst of Tech Lexington Lincoln Lab.