




2021

The Influence of Socioindexical Information on the Speech Perception-Production Link: Evidence from a Shadowing Task

Kyler B. Laycock

University of Kentucky, kyler.laycock@uky.edu

Author ORCID Identifier:

 <https://orcid.org/0000-0002-3731-0841>

Digital Object Identifier: <https://doi.org/10.13023/etd.2021.191>

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

Recommended Citation

Laycock, Kyler B., "The Influence of Socioindexical Information on the Speech Perception-Production Link: Evidence from a Shadowing Task" (2021). *Theses and Dissertations--Linguistics*. 41.
https://uknowledge.uky.edu/ltt_etds/41

This Master's Thesis is brought to you for free and open access by the Linguistics at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Linguistics by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Kyler B. Laycock, Student

Dr. Kevin B. McGowan, Major Professor

Dr. Allison Burkette, Director of Graduate Studies

THE INFLUENCE OF SOCIOINDEXICAL INFORMATION ON THE SPEECH
PERCEPTION-PRODUCTION LINK: EVIDENCE FROM A SHADOWING TASK

THESIS

A thesis submitted in partial
fulfillment of the requirements for
the degree of Master of Arts in
Linguistic Theory & Typology in
the College of Arts and Sciences at
the University of Kentucky

By
Kyler B. Laycock
Lexington, Kentucky

Director: Dr. Kevin B. McGowan, Professor of Linguistics
Lexington, Kentucky
2021

Copyright© Kyler B. Laycock 2021
<https://orcid.org/0000-0002-3731-0841>

ABSTRACT OF THESIS

THE INFLUENCE OF SOCIOINDEXICAL INFORMATION ON THE SPEECH PERCEPTION-PRODUCTION LINK: EVIDENCE FROM A SHADOWING TASK

The body of work on speech perception demonstrates the ability of listeners to utilize both visual and acoustic information in their processing of a given speech signal. More recent studies have established that listeners are sensitive to cues in both these modalities which inform their perception of a speaker's identity in parallel with the linguistic message, but the relationship between social information in perception and production together is unclear. This study reports the results of an experiment designed to test the hypothesis that expectations about a speaker's identity is able to influence a listener's perception and production of speech in tandem. The shadowing task addresses the degree to which listeners faithfully reproduce L2 accented English when presented with four ethnically distinct faces in congruent and incongruent auditory-visual pairs in a within-subject design. Analyses of the degree of acoustic similarity to model talkers in speakers' imitations revealed a slight average trend toward convergence on vowel spectra, vowel duration, and average fundamental frequency. Significant predictors of the degree of change in a speaker's production were shown to be the vowel quality measured and the voice presented, but these predictors were agnostic with respect to whether these changes represented phonetic convergence or divergence. The variance in degree of similarity suggests that speakers' convergence is subject to linguistic selectivity, but it is less clear the role social selectivity plays when presented with unfamiliar varieties. Overall these findings are consistent with exemplar models which consider the inherent coupling of individuals' speech perception and production, but that the visual stimuli had no significant effect on these analyses may be reflective of listeners' adaptive processes during perception of L2-accented speech

KEYWORDS: speech perception, speech production, social expectations, shadowing, phonetic convergence

Kyler B. Laycock

May 15, 2021

THE INFLUENCE OF SOCIOINDEXICAL INFORMATION ON THE SPEECH
PERCEPTION-PRODUCTION LINK: EVIDENCE FROM A SHADOWING TASK

By
Kyler B. Laycock

Dr. Kevin B. McGowan
Director of Thesis

Dr. Allison Burkette
Director of Graduate Studies

May 15, 2021
Date

Dedicated to my sister, Nayana, and to all good listeners.

ACKNOWLEDGMENTS

I'd first like to thank my advisor and thesis committee chair, Dr. Kevin McGowan, who has been a constant source of support and encouragement during my time at the University of Kentucky, and who first introduced me to the study of language. His help has been invaluable during every step of this project. For inspiring my research, for going above and beyond to help me succeed, and for always believing in my ability to do good work, I am so grateful. I don't think I could have asked for a better mentor (but even if I could, I probably wouldn't have).

I'd also like to thank the members of my thesis committee, Dr. Jennifer Cramer, Dr. Josef Fruehwald, for their guidance and support which were invaluable to the development of this thesis, and to my own development as a researcher. I've been incredibly fortunate to have their insight and assistance.

To Dr. Jennifer Cramer, thank you for everything you've taught me, for always making time to help with absolutely anything, and for never letting me get away with less than my best work. The background knowledge and feedback you've provided for this project led to a (hopefully) better informed and more holistic discussion.

To Dr. Josef Fruehwald, thank you for all your assistance with R, statistical modeling, FAVE, and the intel about Labovian transcriptions. I appreciate all the tips and insight you've given me on the tools and software used in my analyses.

To Dr. Fabiola Henri and Dr. Mark Lauersdorf thank you for every piece of professional advice, for the moral support you've provided over the years, and for showing me how much more there is to learn. To Dr. Allison Burkette thank you for helping to refine this project in its early stages, for your advice on writing, teaching, and academia in general, and for every discussion that made me consider things

from another perspective. To Katia Davis, thank you for all you do as office manager, and for all the hard work that keeps things operating smoothly. I'd also like to extend my gratitude to all the faculty members at UK who were essential to my development in the MALTT program. I've been profoundly influenced by the education I've had the privilege to receive, and want to further thank all the educators who have taught me throughout the years.

Additionally, I would like to say thank you to the people I've met in the MALTT program for all their assistance and friendship. To Avery, Ryan, Taha, Chris, Aleah, Steve, and the rest of my cohort, thank you for being such great friends, for all the laughs, all the help, and all the memories. To Xueying, Nour, Crissandra, Monica, Ty, and everyone else, thank you for all you've done to support and encourage me and one another. There are far too many times to list when y'all have shared your thoughts, expertise, time, and energy to help me in one way or another. I've been lucky to have colleagues like all of you. And to Monica Larcom, thank you for pushing back the boundaries of ignorance and reminding me to do the same.

I want to express my gratitude to the friends and loved ones who weren't much help with respect to the technical and academic matters mentioned above (although I'd be remiss to not mention Helen Mitchell's great advice on linear regression), but who gave me so much support and encouragement along the way. Lastly, to Gerald Bankes, I don't think I could ever be able to express how thankful I am for you and for all you do. While I've worked on this thesis, you've proofread, debugged, and on at least one occasion spent hours covering a whiteboard with equations. Thank you for all your hard work, thank you for taking out the trash (again), thank you for teaching me about R notebooks, and thank you for bringing so much joy to my life.

TABLE OF CONTENTS

Acknowledgments	iii
List of Tables	vi
List of Figures	vii
Chapter 1 Introduction	1
1.1 Awareness and Perception	2
1.2 Social Expectations	3
1.3 Linguistic and Social Selectivity in Imitation	4
1.4 The Current Study	4
Chapter 2 Methodology: Shadowing Task	7
2.1 Participants	7
2.2 Stimulus Materials	7
2.3 Procedure	8
2.4 Measurements	10
2.5 Assessing Changes in Production	15
Chapter 3 Results	17
3.1 Statistical Analysis	22
Chapter 4 Discussion	28
Chapter 5 Conclusions	30
Appendix	32
References	33
Vita	39

LIST OF TABLES

2.1 FAVE-extract Example Output	12
3.1 Summary Statistics	17
3.2 Summary Statistics: Vowel Spectra	19
3.3 Summary Statistics: Vowel Duration	20
3.4 Summary Statistics: F0	20
3.5 Model Results: Vowel Spectra	23
3.6 Model Results: Vowel Duration	25
3.7 Model Results: F0	25

LIST OF FIGURES

2.1	Visual Stimuli	9
2.2	Baseline Vowel Formant Frequencies	13
2.3	Baseline Vowel Durations	14
2.4	Baseline F0	15
3.1	Histograms: Difference in Distance	17
3.2	Faceted Ridgeline Plot: Vowel Spectra	19
3.3	Faceted Ridgeline Plot: Vowel Duration	20
3.4	Faceted Ridgeline Plot: F0	21
3.5	Model Predictions: Vowel Spectra	24
3.6	Model Predictions: Vowel Duration	26
3.7	Model Predictions: F0	27

It is a tendency of listeners to acquire the varieties of language in use around them. This fact is most obvious in the acquisition of a child's first language through exposure to the patterns of the language users in their environment (e.g. Saffran, 2001; Chambers et al., 2003). A similar phenomenon in which language users adapt their own speech relative to the speech they perceive has been well documented by the substantial body of work related to phonetic convergence. Phonetic convergence (alternately referred to as phonetic imitation or accommodation) describes these adaptive processes which result in speakers' adoption of features from the speech of interlocutors (Pardo, 2012). Effects of convergence have been documented in spontaneous interactions and pre-designed collaborative speech tasks (Giles et al., 1973; Natale, 1975; Pardo, 2007), as well as emerging during markedly non-social laboratory tasks such as word shadowing (Goldinger, 1998; Shockley et al., 2004). That convergence is observed to occur in these various circumstances, as well as in instances where speakers do not share a common dialect, has given rise to proposals which characterize these behaviors as the result of automatic cognitive processes (Babel, 2012; Delvaux & Soquet, 2007; Nguyen et al., 2012; Xie et al., 2021).

The ability of speakers to attend to the manifold acoustic variability in speech, as well as achieve a robust constancy of perception, is remarkable in and of itself, and even more so when considering that this variation is an important part of the speech signal to be processed as a source of acoustically encoded socioindexical information (Eckert, 2012; Foulkes & Hay, 2015; McGowan, 2015). Considering these aspects of the nature of speech perception alongside the characterization of phonetic convergence as automatic, the realization of phonetic accommodation in production necessarily depends upon a speaker's *perception* of the speech to which they are accommodating. The link between an individual's perception and production is further borne out from the results of accommodation studies, in which listeners-turned-speakers adapt their speech in relation to conversational partners and/or stimuli with the tendency to become more phonetically similar to the input speech (for review, see Pardo, 2012), as well as studies examining the role the relationship between perception and production plays in sound change (Lindblom et al., 1995; Pierrehumbert, 2001; Harrington et al., 2008a; Beddor, 2009; Beddor et al., 2018). Crucially, the effects of perception-production coupling in accommodation studies take place with respect to individuals. Subjects in these

studies display effects of convergence to different degrees, with the variance in behavior being explained by individual sensitivity to linguistic factors (such as sub-phonemic features, language-specific phonemic contrasts and allophonic variation, and suprasegmental features) (Mitterer & Ernestus, 2008; Nielsen, 2011; Babel, 2012; Walker & Campbell-Kibler, 2015; Clopper & Dossey, 2020), and/or social factors (such as attractiveness or salience of features indexed in a particular way) (Babel, 2012; Yu et al., 2013; Walker & Campbell-Kibler, 2015; Clopper & Dossey, 2020).

1.1 AWARENESS AND PERCEPTION

While it could feasibly be the case that as an automatic cognitive process, explanations of phonetic convergence need not consider a speaker's awareness of a given feature perceived with respect to linguistic selectivity, the fact that these processes are sensitive to speaker attractiveness (Babel, 2012), or salience of dialectal features (Walker & Campbell-Kibler, 2015; Clopper & Dossey, 2020) such that there is an effect on the degree of phonetic convergence may complicate models of accommodation that view the phenomenon as strictly automatic. For this reason, it is worth considering that there are also conscious cognitive processes which determine the realizations of speakers' accommodation and imitation.

In a study of cross-dialect imitation, Preston (1992) frames the linguistic features present in imitation as the clearest indicator that variation in speech is a salient marker of identity. The focus of Preston's study is on the imitation of Black speakers by white speakers and vice versa, and in particular on the nature of the features employed in these imitations as caricatures. It is of note that while frequency of the features in the imitated variety may predict their use as caricature features, it is not necessarily the case that caricature features were present in the imitated variety at all. Similarly, it was often the case that respondents in Preston's study were unable to comment on the features they had used in imitation. This seems to suggest that, while there is some level of perceptual awareness with respect to linguistic markers of identity, this awareness may or may not extend itself into metalinguistic awareness about these varieties, and may reflect a perception of these varieties that is influenced by individuals' *beliefs* about speakers and that is not solely grounded in linguistic reality. Crucially, becoming aware of a given object of perception¹ likely involves both conscious and unconscious processes which are

¹that is, *noticing* something (as discussed in Preston (2016))

informed by an individual's prior experiences, ideologies, and attitudes (Preston, 2016; Campbell-Kibler, 2010).

1.2 SOCIAL EXPECTATIONS

In the context of speech perception, an individual's attitudes, ideologies, and beliefs as constructed on the basis of said individual's experience can be conceptualized as *expectations*. Socioindexical variation in speech is meaningful as a source of information to listeners, but as demonstrated in Rubin (1992), McGowan (2011), and McGowan (2015), even the expectation of social information can affect listener's perceptions. Rubin (1992) presented students with recorded lectures, as well as photographs depicting individuals of different ethnicities for the purpose of establishing expectations about the speaker from the recording. Students in this study shown a photograph of an Asian person were found to perceive the voice from the recording to be more strongly accented than students shown a photograph of a white person.

McGowan (2015) employed an inverted matched guise design in an experiment in which participants transcribed speech from audio recordings of Chinese accented English. As in Rubin (1992), participants were shown a photograph of either a Chinese or white American person, but the results of this study show that congruity between face and voice (i.e., Chinese face paired with Chinese voice) improved the accuracy of their transcriptions compared to the incongruous stimuli pair of white face and Chinese voice. Together, these studies show the ability of social expectations to influence perception, even when the expectation is established by stimuli in a different perceptual modality (i.e. visual as opposed to auditory). That expectations may alternately complicate or facilitate perception is seemingly consistent with exemplar models of speech perception.

Exemplar theories of perception resolve many of the issues arising from multimodal perceptual sources (Johnson, 2005). Exemplar models of speech perception have been utilized to stand as an explanation for several aspects of speech perception and production (e.g. Johnson, 1997; Johnson, 2005). This present study operates from the assumption that exemplars are stored memories of a given event which contain all linguistic, situational, and socioindexical information as a unified percept (for review, see Pierrehumbert, 2001). In this framework, the task of speech perception is a processing task in which the continuous returning percepts of speech are compared to stored exemplars in the memory of the listener.

The exemplar² most similar to the perceptual properties of the current experience will serve as a reference point in the categorization of the new percept. Similarly, the task of speech production is conceptualized in this model as a processing task which returns speech based upon the averaging of relevant exemplars. One's stored exemplars are both created and refined by their experiences, and it is these products of experience which inform our continued perception.

1.3 LINGUISTIC AND SOCIAL SELECTIVITY IN IMITATION

Individual differences in reproductions of the same speech on the bases of perceived linguistic and social variables are described in the literature on accommodation as selectivity in imitation. Babel (2012) suggests both linguistic and social selectivity as an explanation for the degree of imitation for vowel formant frequencies varying between vowel qualities differently, and further relates the differences to participant dialect on the basis of whether there was sufficient phonetic space for accommodation to a greater magnitude. There is evidence for linguistic selectivity of convergence with respect to both abstract phonological categories of a language (phonemic contrasts, and allophonic variation, for example), and the acoustical and articulatory properties of speech (e.g. coarticulatory effects, or speaking rate).

Dialectal variation is often a focus of accommodation studies for the purposes of designing experiments sensitive to both linguistic and social selectivity, although it can be difficult to assess attitudes and beliefs for speakers and communities with respect to the variety in question, and whether the variety is perceived by participants as sufficiently prestigious to facilitate convergence, or carries a sufficiently negative bias such that it exhibits less convergence or even divergence (Clopper & Dossey, 2020; Babel, 2012; Walker & Campbell-Kibler, 2015).

1.4 THE CURRENT STUDY

There are three assumptions which most directly affect the design and interpretation of the present study:

1. Exemplars represent a unified percept that is stored in memory with all acoustic, social, and contextual information intact.
2. The sociodexical variation present in the speech signal- as informed by factors such as linguistic background, co-articulation, physiology, gender, and

²or set of exemplars dependent upon the present perceptual task

identity of the speaker-informs our perception of speech. Socioindexical information is inextricable from linguistic information in a given experience.

3. The variation present in an individual's speech corresponds to, interacts with, and/or is generated by the same cognitive processes as the variability in that individual's perception of speech.

At the most fundamental level, the questions which motivate this study stem from one of the issues central to the study of speech perception throughout the discipline's history. Namely: how do listeners handle the acoustic variation in speech that arises from both intra- and inter-speaker variability (Jusezyk & Luce, 2002)? However, the more specific question is: how do expectations set by socially relevant visual information about a speaker affect a listener's perception and production jointly?

A shadowing task (as a methodology implemented in the theoretical framework of phonetic convergence) was chosen as a means of seeking answers to this research question (Goldinger, 1998). Additionally, in order to assess effects of visual information about a speaker on speech perception and production, an inverted matched guise element was included in the shadowing task, so that speakers heard a voice to shadow, but also saw a face which may or may not have been congruous to that voice. The matched guise element of this experiment is considered inverted as it presents visual stimuli for the purpose of creating expectations about a speaker, such that measurements can reflect the extent to which these expectations are integrated in the perception and of speech (W. Lambert et al., 1960; Rubin, 1992; McGowan, 2015).

In the context of existing work on phonetic convergence, non-native (L2) accented speech is an area of interest in studying these phenomena. This is due to the unique relationship between L2-accented and native-accented speech in that the phonology of L2-accented speech is systematically distinct from the phonology of a native English speaker in ways that are less familiar to a greater degree than the speech of other native English speakers. The implication here for phonetic convergence is that speaker-specific processes of linguistic and social selectivity (with respect to the perception of linguistically encoded social information) will behave differently in the absence of more sufficiently familiar phonological contrasts and salient markers of identity for a native speaker of American English. The lack of familiarity is expected to increase participants' reliance on their expectations (with

respect to visual the inverted matched guise stimuli) to inform their perception and imitations.

CHAPTER 2. METHODOLOGY: SHADOWING TASK

2.1 PARTICIPANTS

Eighteen participants (7 male, 11 female) ranging in age from 18 to 22 years old were recruited from undergraduate introductory Linguistics classes at the University of Kentucky. 2 participants reported their race/ethnicity to be Asian, 2 as Black, and 14 as white. All participants self-reported as having no history of hearing, speech, or communication disorders. Participants were native speakers of English; 14 participants indicated no additional language proficiencies, while the remaining 4 participants reported native bilingualism in English and Punjabi, English and Hebrew, English and Korean, and English and Arabic. No participants indicated any knowledge of or familiarity with Mandarin, Lugisu, or Spanish. Participants completed a brief language history and demographic questionnaire following the shadowing task. Data were collected from an additional 7 participants, but these participants' data were excluded from analysis because they indicated proficiency in one or more of the languages listed above ($N=2$), did not provide a sufficient number of usable data due to mispronunciations or background noise on recordings ($N=2$), an excess of trials in which the participant had a large delay in producing shadowed speech ($N=1$), or due to experimenter error ($N=2$).

2.2 STIMULUS MATERIALS

Stimulus materials consisted of 25 sentences from the American English Hearing in Noise Test (HINT) (Nilsson et al., 1994; Vermiglio, 2008; Soli & Wong, 2008). Ten of these sentences were presented in standard English orthography on a computer screen in the first block of the shadowing task. These sentences as read aloud served as a baseline for each participants' baseline read speech. The remaining 15 sentences were presented auditorily over headphones for participants to reproduce aloud. HINT sentences are designed such that 2-4 target words are embedded within simple declarative sentences. Sentences were chosen to ensure target words containing each of nine stressed monophthongal vowel qualities /i, ɪ, ε, æ, α, o, ʊ, u, ʌ/ were present in both the baseline and shadowing blocks.

AUDITORY STIMULI

Audio recordings of the shadowing block sentences were retrieved from the ALLSSTAR (Archive of L1 and L2 Scripted and Spontaneous Transcripts And Recordings) corpus (A. R. Bradlow, n.d.). The ALLSSTAR database is comprised of recordings of speakers with varying linguistic backgrounds performing speech production tasks in English. The recordings selected for this experiment were of sentences from the HINT produced by four speakers whose first languages were Mandarin, Lugisu, Mexican Spanish, and American English respectively¹. All speakers were male and between the ages of 20-35. The 15 HINT sentences were recorded in English by each of the four speakers, resulting in a total of 60 unique auditory stimuli for the shadowing block.

VISUAL STIMULI

As discussed earlier, the experiment itself is an inverted matched guise task (W. E. Lambert et al., 1960; Rubin, 1992; McGowan, 2011; McGowan, 2015). In order to test the ways in which speech perception (and therefore production) is informed by social information, for the duration of each trial in the shadowing block, one of four images was presented to participants alongside one of the 60 audio stimuli. The faces in these images were collected from the Chicago Face Database (Ma et al., 2015), a resource containing high-resolution, standardized images of faces indexed by gender and ethnicity. The faces are normalized for both physical attributes (i.e., measurements of particular facial dimensions), and subjective ratings such as attractiveness. The faces were selected primarily based upon ethnicity and gender in order to match the reported ethnicity and gender of the speakers from the ALLSSTAR corpus such that a single face's ethnicity corresponded to a single voice's ethnicity, however the selection process was also sensitive to other available demographic information of the ALLSSTAR speakers and CFD participants. The four faces used in this experiment can be seen in Figure 2.1.

2.3 PROCEDURE

Participants were seated at a computer inside a Whisper Room sound-attenuated booth² in the University of Kentucky Phonetics Lab. Auditory stimuli were pre-

¹demographic information about each of the four speakers from the ALLSSTAR corpus is available in Appendix I

²The sound booth used in this study meet all ANSI 3.1 requirements of acceptable ambient noise levels for audiometric testing rooms as defined by the Acoustical Society of America (ANSI/ACA,

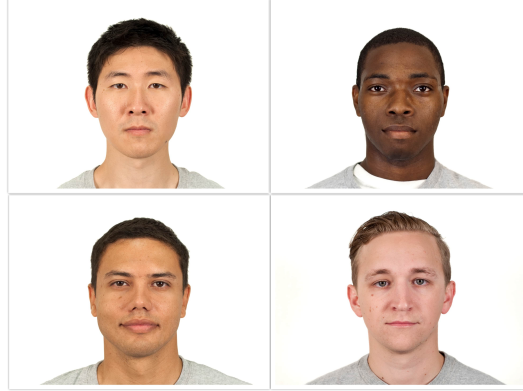


FIGURE 2.1: *The faces used as visual stimuli. Clockwise from the top left, they correspond to the racial/ethnic categories Asian, Black, white, and Hispanic*

sented over AKG K273 MK II headphones. Speech data were collected using a boom-mounted Electro-Voice RE20 microphone which recorded the data directly to the computer used for the experiment. Stimuli were presented using OpenSesame version 3.2.8 (Mathôt et al., 2012), an open-source graphical experiment builder³

In the first block of the experiment, participants were instructed to read aloud and record a series of 10 HINT sentences presented in standard orthography. The sentences appeared on the screen individually and persisted onscreen until the participant indicated they had finished the current recording by pressing an arbitrary key on the computer’s keyboard. All 10 sentences were presented to each subject in a random order. Upon completion of this portion of the experiment, participants were presented with instructions for the second block, in which they listened to each of the recordings from the ALLSTAR corpus and were asked to repeat each sentence they heard. Participants were also instructed at this point to mimic the voice they heard to the best of their ability in their reproduction of the stimuli. The recorded sentences were provided as audio played over headphones with no corresponding orthographic representation. It was not possible for participants to hear the recordings more than once. Participants were given twice the total duration of the target utterance plus two seconds $((2t)+2)$ in which to shadow each sentence and indicate that they had finished by pressing an arbitrary key on the keyboard to advancing to the next stimulus⁴. If the participant exceeded this time limit, the microphone recording stopped automatically, and the screen displayed a reminder (2018).

³For the purposes of this experiment, a custom plug-in was written for OpenSesame to allow the audio of participants’ speech data to be recorded during the experiment.

⁴For example, if the stimulus duration was 1.5 seconds, participants were given a total of 5 seconds before the microphone automatically stopped recording.

to press a key when finished recording.

Each time a sentence to be shadowed was presented, one of the four faces was simultaneously displayed onscreen. The faces remained onscreen until the participant pressed a key, at which point a blank screen was displayed for 500ms to signal the start of the next trial, followed by a new audio-visual stimuli pair presented in the same manner. Based upon similar study designs (e.g. [Mcgowan, 2011](#)), a relatively small effect was expected and so a relatively large number of trials were given per participant. Each participant shadowed each of the 60 audio recordings in combination with each of the four faces exactly once for a total of 240 trials per participant in the shadowing block. These trials were presented in a random order for each participant.

2.4 MEASUREMENTS

The baseline reading task produced a total of 180 recordings (10 per participant) and the shadowing task produced a total of 4320 recordings (240 per participant). These 4500 recordings, as well as the 60 stimulus recordings from the ALLSSTAR corpus, were initially annotated and aligned with word and phoneme boundaries by the from the Forced Alignment and Vowel Extraction (FAVE) suite's FAVE-align implementation of the Penn Phonetics Lab Forced Aligner ([Rosenfelder et al., 2014](#); [Yuan and Liberman, 2008](#)). The aligner takes a sound file containing speech and matching orthographic transcript as input, and produces acoustic segmentation of the speech aligned with a Praat TextGrid file. The segmentation of words and unstressed monophthongal vowels were assessed visually through inspection of spectrograms in Praat ([Boersma & Weenink, 2021](#)), and hand-corrected for segmentation errors. As FAVE-align's acoustic models were trained only on American English from the SCOTUS corpus ([Yuan & Liberman, 2008](#)), special attention was given to the audio stimuli produced by L2 speakers of English, and the alignments produced by FAVE were compared to the alignments provided in the ALLSSTAR corpus for these speakers' productions of the HINT sentences⁵.

Following the initial alignment, 127 recordings were excluded from the analysis due to the absence of speech or presence of background noise in either of the two blocks, or for mispronunciation/reading the sentences incorrectly (i.e. substituting

⁵The ALLSSTAR corpus utilizes the Montreal Forced Aligner ([McAuliffe et al., 2017](#)) which utilizes acoustic models trained on a wide variety of languages for segmentation. Of note is that there were very few discrepancies between the two alignments with respect to word and segment boundaries.

a word present in the onscreen sentence with a different word by) in the baseline block. The exclusion of these recordings took place after the alignment, not as a result of the auto-segmentation process itself, but rather because FAVE provides additional error files in the event of a failed alignment, which informed and somewhat streamlined the process of identifying potentially problematic recordings.

The remaining 4373 recordings files containing their alignments were passed to FAVE-extract, the other program contained in the FAVE suite. FAVE-extract is a tool designed for automatic vowel measurement, taking as input both a sound file containing recorded speech data and the alignment of the same speech data as a Praat TextGrid. The program utilizes Praat's formant tracking algorithm and selects a set of formant tracks estimated from Praat's Linear Predictive Coding (LPC) analysis. These candidate measurements are compared to the distribution of measurements for the same vowel quality in the *Atlas of North American English (ANAE)*, and then re-evaluated based upon the distribution of the speakers' own productions of the vowel in question (Labov et al., 2013). For this reason, baseline and shadowing productions from the same participant were treated as being produced by distinct speakers. This was the case for **only** the extraction process to ensure that FAVE's selection process didn't result in a reduction of meaningful change between experiment blocks. One final important point with respect to the internal processes of FAVE-extract is that by default measurement points are selected differently for different vowels in the interest of better representing their central tendencies and reducing the likelihood of measurement errors (Labov et al., 2013). The majority of these specified measurement points pertain to diphthong vowels in varying phonological environments which were not considered in analysis of this data, however the measurement point set for the vowel /u/ is selected at the vowel onset if /u/ was immediately preceded by a coronal consonant. This particular measurement point is informed by Chapter 12 of Labov et al. (2006) in which the density of measurements for tokens of /u/ tended to be centered around a much higher F2 when occurring in post-coronal environments (i.e. the coarticulatory effects of consonants like /t/ result in a 'fronter' /u/). All other monophthongs are measured at one-third of the duration between their onset and offset. Unless otherwise stated, FAVE-extract's default measurement points for each vowel are the measurements reported here.

The final output of FAVE-extract for each pair of input files is a single file containing the vowel(s) measured, binary values representing stress (1= stressed, 0= unstressed), the frequencies in Hertz of F1, F2, and F3, the time point in the record-

ing at which the vowel was measured, the vowel’s duration in seconds, as well as several more columns pertaining to the word in which the vowel was located, and speaker demographics as specified in a separate input file for each speaker. A sample of the output from FAVE-extract can be seen in table 2.1

TABLE 2.1: The output of FAVE-extract for a single shadowing recording. The "Speaker" and "sex" column contain the participant’s identification number and self-reported sex, "face" and "voice" describe which stimuli were presented during the shadow, in this case the model talker was a Native speaker of Mandarin Chinese, and the face on screen was the Asian face from the CFD. Note that value of "vowel" has been converted from ARPABET transcription to IPA, and all Hz have been converted to Bark scale. (Not shown: Additional columns with formant measurements at different points in the vowel, bandwidths, environments, and other information.)

Speaker	sex	face	voice	vowel	stress	word	F1	F2	time	dur
01	F	A	CH	ɑ	1	HOT	7.036024	9.466382	1.639	0.139

As the focus of the first of three analyses, all first and second formant frequency estimates from the extraction were checked for outliers and converted to the Bark scale (Traunmüller, 1990). To allow for comparisons of both male and female participants’ productions to the four male model talkers, the formant frequency estimates from the baseline and shadowing blocks were normalized relative to the acoustic center of a two dimensional F1 × F2 space in Bark. The center was defined as the grand mean of all participants’ /i, α/⁶ measurements at one-half the duration between vowel onset and offset in the baseline reading block (A. Bradlow et al., 1996; Scarborough & Zellou, 2013; Harrington et al., 2008b; Clopper & Dossey, 2020). Only midpoint F1 and F2 estimates were used to define the center of the acoustic space in order to avoid differing coarticulatory effects between instances of a single target vowel. Only data from the baseline reading block were used in defining the acoustic center to avoid variation expected in the shadowing trials. Each participant’s midpoint F1 and F2 measurements for each instance of both the vowels /i/ and /α/ were used in four separate calculations of arithmetic mean per participant. The acoustic center of the F1 × F2 space in Bark was then defined by calculating the grand mean of F1 and F2 for /i/, and the grand mean of F1 and F2 for /α/. That is, the acoustic center can be conceptualized as the geometric midpoint of a line passing through points /i/ and /α/. If we assume that the vowel space can be modeled as roughly quadrilateral with /i/ and /α/ corresponding to opposite vertices, we can treat (i↔α) as a diagonal which bisects that quadrilateral such that

⁶/u/-fronting was observed for several speakers in the baseline block. Were /u/ measurements used in calculating the center of the acoustic space, the F2 estimate and subsequent normalization would result in fronting of all vowel qualities.

the midpoint between opposing vertices corresponds to the geometric center⁷. All estimates of F1 and F2 for each participant were normalized relative to the acoustic center by subtracting the F1 and F2 grand mean from each formant estimate in Bark. A summary of normalized F1 and F2 in baseline productions compared to the four model talker’s productions is shown in Figure 2.2. Additionally, to consider

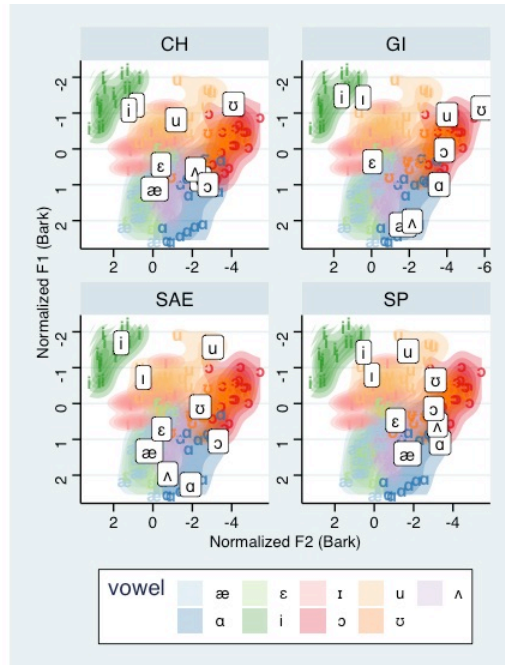


FIGURE 2.2: The four faceted panels show normalized baseline formant frequencies for each vowel category against each of the four model talkers' formant frequencies. Colors represent density of participant baseline F1 and F2 for each vowel category, and vowel symbols represent the average for each speaker. White square labels indicate mean formant frequencies for a given model talker's vowels.

multiple dimensions of phonetic change between baseline and shadowing block productions, vowel duration and fundamental frequency (F0) data were examined (Babel, 2012; Nycz & Hall-Lew, 2013). Vowel duration as reported by FAVE-extract was used as another means of measuring the degree of change between baseline and shadowed productions. A summary of vowel duration for participant baseline production compared to the four model talkers is shown in Figure 2.3. F0 was estimated for all recordings by the Robust Epoch And Pitch Estimator (REAPER) program (Talkin, 2015). Reaper defines the local F0 estimates of a speech signal as

⁷While, in reality the notion of the 'vowel space' is only an abstraction with respect to articulatory and acoustic properties of the speech stream itself, it may be important to clarify: if we consider the vowel space as the conventional roughly trapezoidal, finding the acoustic center would not be possible with a single diagonal as diagonals of trapezoids are neither mutually bisecting nor perpendicular. However, the same process of /u/-fronting mentioned above has resulted in the shape of the vowel space being a parallelogram which allows the geometric midpoint of either diagonal alone to satisfactorily coincide with the acoustic center

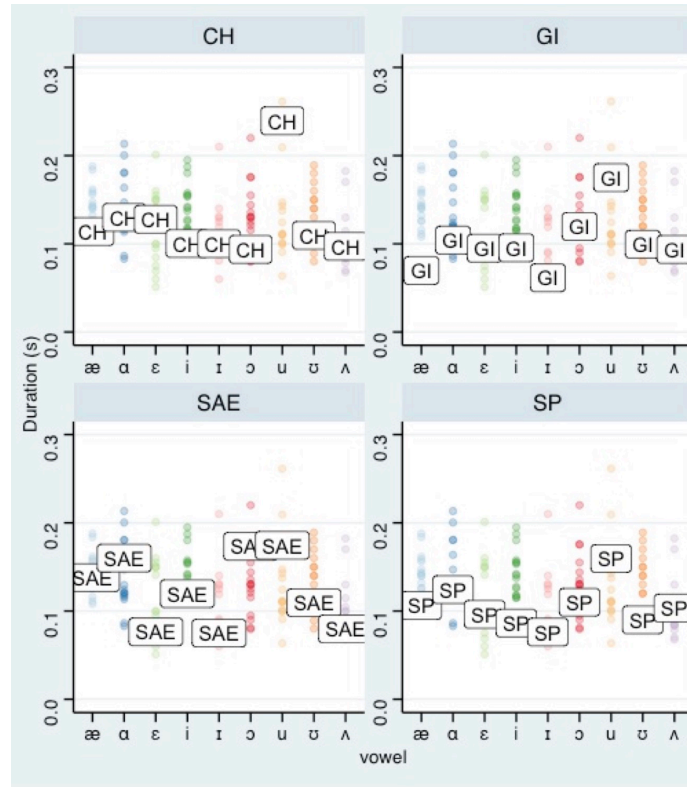


FIGURE 2.3: The four faceted panels show participants' mean duration for each vowel quality compared to each of the four model talkers. Each point represents a single participant's average duration for a given vowel. Square labels indicate the model talker's L1 (Mandarin Chinese, Lugisu, American English, or Spanish) and mean duration for a given vowel.

the inverse of time between consecutive glottal closure instants (GCI)⁸. REAPER has been shown to be accurate in its estimation of F0, and more effective at tracking F0 across very low pitch ranges than Praat with meaningful estimates being returned in ranges as low as 20 Hz (Lieberman, 2015; Dallaston & Docherty, 2017; Szakay & Torgersen, 2019). This is particularly important to consider given that creaky phonation⁹ was expected to have some presence among participant baseline recordings (Lee, 2016; Cantor-Cutiva et al., 2018; Dallaston & Docherty, 2020), and the inability to estimate F0 in the presence of creak would result in loss of valuable data. The output files from REAPER for each recording (text files containing three columns corresponding to: time of estimation, whether voicing was present, and the F0 estimation in Hertz) had every estimation associated with voiceless segments removed. Mean F0 was calculated across all recordings from the base-

⁸In reality, the REAPER refers to a separately derived normalized cross-correlation function (NCCF) for each GCI, selecting the location of the best candidate NCCF maximum relative to the GCI-estimated period as opposed to calculating F0 estimate based on the period between GCI estimates themselves.

⁹as characterized by compressed and thick vocal folds, resulting in slow vibration, hence low fundamental frequency (f0), and low air flow rates (Podesva, 2013).

line reading block by participant, across all stimuli by model talker and sentence, and across shadowing recordings by participant, and stimuli presented during each recording (i.e. each participant’s mean shadowing block F0 was calculated for each for the 16 possible face × voice stimuli pairs). A comparison of average participant F0 in baseline productions to the average F0 of each model talker can be seen in Figure 2.4.

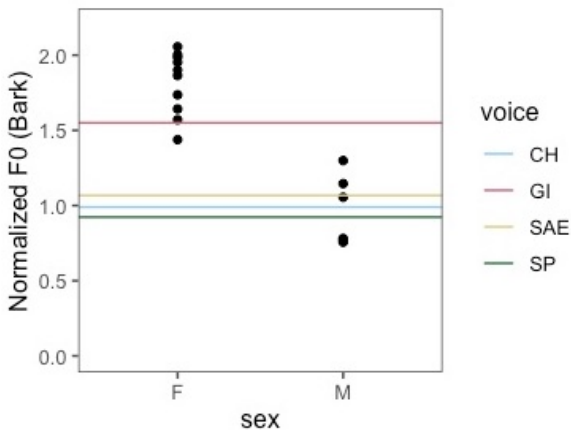


FIGURE 2.4: Mean normalized F0 for each participant plotted by participant sex. The four horizontal lines represent the mean F0 of each of the four stimuli voices (Mandarin Chinese, Lugisu, American English, and Mexican Spanish).

2.5 ASSESSING CHANGES IN PRODUCTION

As a metric of change in participants’ productions after exposure to the auditory and visual stimuli pairs, Euclidean distances were calculated from each participant’s baseline productions to the productions of each the four model talkers, and one from each participant’s shadowing productions to the voice of the model talker which they were shadowing in a given trial (Babel, 2012). Treating F1 and F2 estimates for a given token as a Cartesian Coordinate pair in a two-dimensional Euclidean space, the distance between two tokens in the F1 × F2 space allows the application of the general formula for calculating Euclidean distance between two such points:

$$d(p, q) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

where p and q are two points in Euclidean space, p_1 and q_1 are Euclidean vectors, and n is the number of dimensions of the space.

The measures of acoustic distance to the model talker productions from both the baseline reading and shadowing blocks were used to calculate the *difference in*

distance which was defined as the the shadowed distance subtracted by the baseline distance:

$$d_{shadow} = \sqrt{(F1_{stimuli} - F1_{participant_{shadow}})^2 + (F2_{stimuli} - F2_{participant_{shadow}})^2}$$

$$d_{baseline} = \sqrt{(F1_{stimuli} - F1_{participant_{baseline}})^2 + (F2_{stimuli} - F2_{participant_{baseline}})^2}$$

$$\Delta d = d_{shadow} - d_{baseline}$$

Difference in Euclidean distance was calculated in this same way¹⁰ for the F0 and vowel duration measurements. Because the calculation for difference in distance is calculated by subtracting baseline distance from shadowed distance, a negative difference in distance indicates a greater degree of similarity to the model talker in a given trial, while a positive difference in distance indicates change from baseline production that is more dissimilar to the model talker, and a distance of zero would indicate no change for a given measurement from the baseline (Babel, 2012). Because the calculations for these difference in distance measures include the baseline distances from the reading block, the baseline measurements themselves are not used in the analyses described below.

¹⁰F0 and duration are measurements of a single dimension, therefore both distances used in the difference calculation are equivalent to $|p-q|$

CHAPTER 3. RESULTS

On average across all shadowing trials, the majority of participants displayed a general tendency of convergence with the model talker. Figure 3.1 below presents the average difference in distance of participants with respect to each of the three metrics of acoustic similarity. The dashed lines designate a difference in distance of zero and correspond to no change in production between the two experiment blocks. As seen in panels 3.1a and 3.1c, a majority of participants' data are located on the negative side of the scale, indicating a slight trend toward convergence with respect to the vowel spectra and fundamental frequency of the model talker. Conversely, panel 3.1b shows most participants producing vowels with durations that *less* similar to those of the model talkers, though it should be noted that the range of peaks in the distributions shown in in panels 3.1b and 3.1c are considerably smaller than panel 3.1a. Summary statistics for all shadowed productions are shown in 3.1.

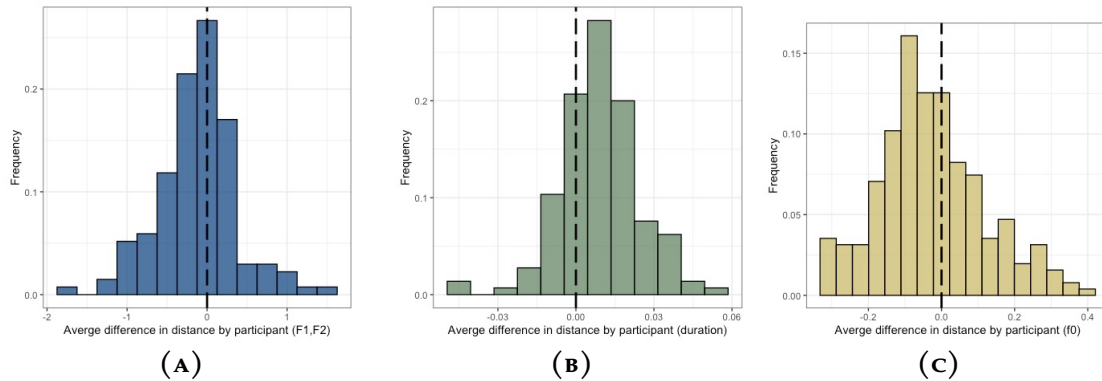


FIGURE 3.1: Left to right are three histograms, A, B & C of average difference in distance values for each of the three metrics respectively. Vertical lines at 0 denote no change, positive values indicate divergence, and negative values indicate convergence.

TABLE 3.1: summary statistics for distance in shadowing block

	Distance		
	Vowel Spectra	Duration	F0
<i>mean</i>	-0.131	0.009	-0.300
<i>median</i>	0.150	0.003	-0.030
<i>range</i>	9.300	1.030	1.790

Figures 3.2, 3.3, and 3.4 present these same data with distributions categorized by the 16 visual and auditory stimuli pairs presented during the shadowing task.

Each panel corresponds to the average difference in distance for each of the three metrics mentioned above. Within a given panel, there are four plots faceted by the L1 of each model talker, and within each plot are four density plots for each of the four faces presented alongside the speech to be shadowed. This results in 16 distributions per panel, one for each possible stimuli pair. As in Figure 3.1, the dashed lines denote a difference in distance of zero. While overall trends in the distribution of the data are less immediately clear from this second set of plots, it is possible to see how different pairings of face and voice seem to affect participants' productions differently. In 3.2, across most face voice pairings the distribution of difference in distance derived from the first and second formant estimates does seem to peak around zero, and the majority of average differences in distances again fall below zero indicating convergence toward convergence. This panel also shows that while there is little change from baseline distance given the English model talker, there seem to be stronger tendencies toward a negative difference in distance for the Lugisu and – to a lesser extent – Mandarin and Spanish voices. Figure 3.3 shows the distribution of difference in distance for duration. While English and Mandarin facets show a slight skew toward dissimilarity in production, there is still a clear peak in distribution near zero. The Lugisu and Spanish voice facets however, show a much clearer shift toward an overall positive distribution, and therefore even less similarity in duration. The F0 differences shown in Figure 3.4 tend toward a somewhat more normal distribution centered on zero for English, Mandarin, and Spanish, but the Lugisu facet shows a consistent peak in distribution on the negative side of the line as well. Interpretations of the effects of face on the distribution are less clear from these plots. In Figure 3.4, for example, distribution appears to differ only with respect to voice. In Figures 3.2 and 3.3, there do however, seem to be distributions which differ depending upon the face and voice pair presented for a given shadowing trial. Summary statistics for shadowed productions grouped by face and voice are presented alongside each plot in Tables 3.2, 3.3, and 3.4.

A series of 16 t-tests were performed for each of the three measures of acoustic similarity to determine whether the difference in distance values were significantly nonzero, as a change in either direction is equally illustrative of how participants react to the stimuli. A Bonferroni correction adjusted the significant levels for these tests to [$p=.0006$]. Therefore we can interpret the mean difference in distance in each condition to be statistically significant, which establishes that across all conditions there was evidence of a meaningful change from participant baseline distances.

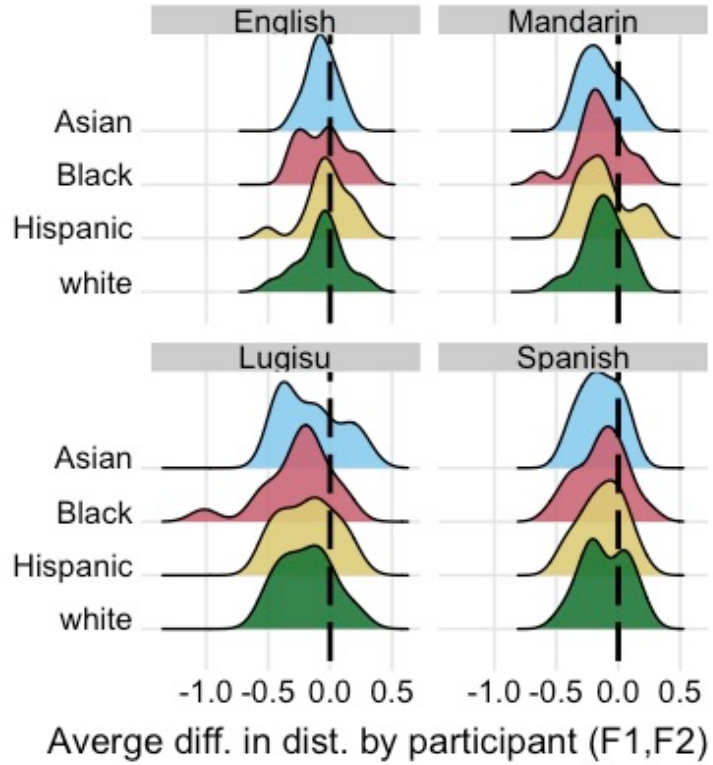


FIGURE 3.2: Average Difference in Distance by Participant F1,F2 spectra. Faceted Panels show distributions by stimuli voice. Ridgelines in each panel show distributions by face for a given voice

TABLE 3.2: Summary statistics- Distance for vowel spectra by condition

	English				Mandarin				Lugisu				Spanish			
	A	B	H	W	A	B	H	W	A	B	H	W	A	B	H	W
<i>mean</i>	-0.07	-0.05	-0.01	-0.07	-0.15	-0.13	-0.13	-0.12	-0.18	-0.24	-0.20	-0.21	-0.15	-0.15	-0.14	-0.11
<i>median</i>	-0.07	-0.05	-0.04	-0.05	-0.16	-0.13	-0.15	-0.18	-0.20	-0.25	-0.24	-0.24	-0.21	-0.22	-0.15	-0.15
<i>min</i>	-2.79	-2.56	-2.47	-2.68	-2.40	-3.11	-2.50	-2.70	-3.71	-3.63	-3.86	-3.86	-2.74	-2.40	-2.49	-2.69
<i>max</i>	2.46	4.70	2.87	2.83	4.49	3.60	3.32	4.22	4.87	4.30	3.77	3.98	5.44	3.70	4.34	4.09
<i>range</i>	5.26	7.26	5.34	5.51	6.89	6.71	5.82	6.91	8.58	7.93	7.63	7.84	8.18	6.10	6.83	6.78

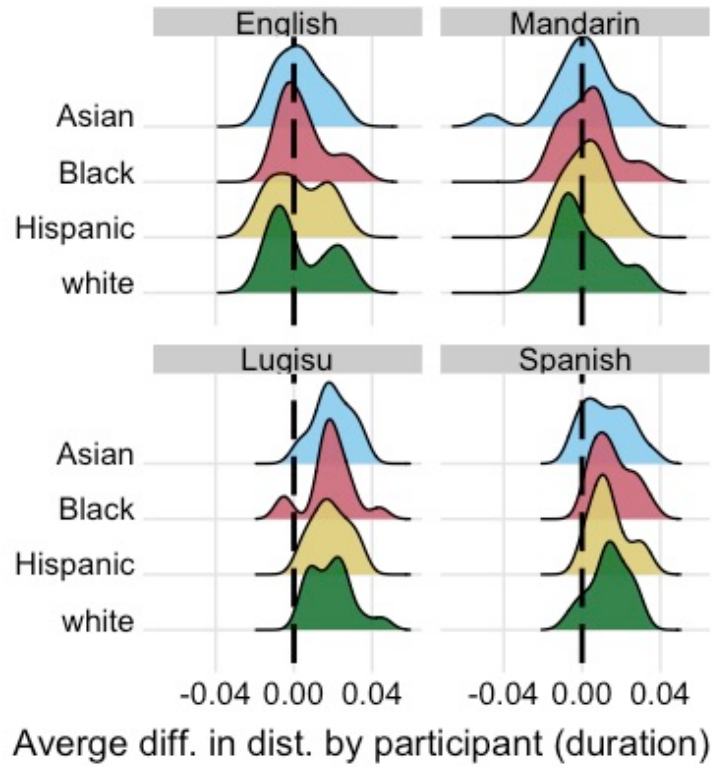


FIGURE 3.3: Average Difference in Distance by Participant vowel duration. Faceted Panels show distributions by stimuli voice. Ridgelines in each panel show distributions by face for a given voice

Table 3.3: Summary statistics- Distance for duration by condition

	English				Mandarin				Lugisu				Spanish			
	A	B	H	W	A	B	H	W	A	B	H	W	A	B	H	W
<i>mean</i>	0.003	0.005	0.003	0.003	0.004	0.004	0.003	0.001	0.020	0.019	0.019	0.020	0.013	0.016	0.014	0.014
<i>median</i>	-0.005	0.000	0.000	-0.001	-0.006	-0.004	-0.006	-0.009	0.011	0.009	0.011	0.010	0.005	0.006	0.005	0.007
<i>min</i>	-0.110	-0.114	-0.123	-0.113	-0.170	-0.091	-0.120	-0.130	-0.105	-0.095	-0.116	-0.085	-0.080	-0.100	-0.110	-0.074
<i>max</i>	0.243	0.191	0.260	0.211	0.302	0.249	0.253	0.231	0.210	0.255	0.179	0.856	0.254	0.496	0.488	0.197
<i>range</i>	0.353	0.305	0.384	0.325	0.471	0.339	0.372	0.360	0.315	0.350	0.295	0.941	0.333	0.595	0.597	0.271

TABLE 3.4: Summary statistics- Distance for F0 by condition

	English				Mandarin				Lugisu				Spanish			
	A	B	H	W	A	B	H	W	A	B	H	W	A	B	H	W
<i>mean</i>	-0.008	-0.000	-0.007	-0.003	-0.016	-0.018	-0.023	-0.029	-0.018	-0.021	-0.030	-0.025	-0.049	-0.063	-0.065	-0.044
<i>median</i>	-0.005	-0.001	-0.021	-0.016	-0.030	-0.030	-0.027	-0.026	-0.019	-0.010	-0.004	-0.028	-0.046	-0.035	-0.063	-0.045
<i>min</i>	-0.603	-0.475	-0.537	-0.675	-0.638	-0.662	-0.617	-0.941	-0.774	-0.770	-0.768	-0.777	-0.748	-0.787	-0.748	-0.662
<i>max</i>	0.605	0.614	0.506	0.696	0.571	0.771	0.542	0.634	0.787	0.700	0.678	0.850	0.645	0.571	0.787	0.781
<i>range</i>	1.208	1.089	1.043	1.371	1.209	1.433	1.159	1.574	1.561	1.470	1.446	1.627	1.393	1.358	1.535	1.443

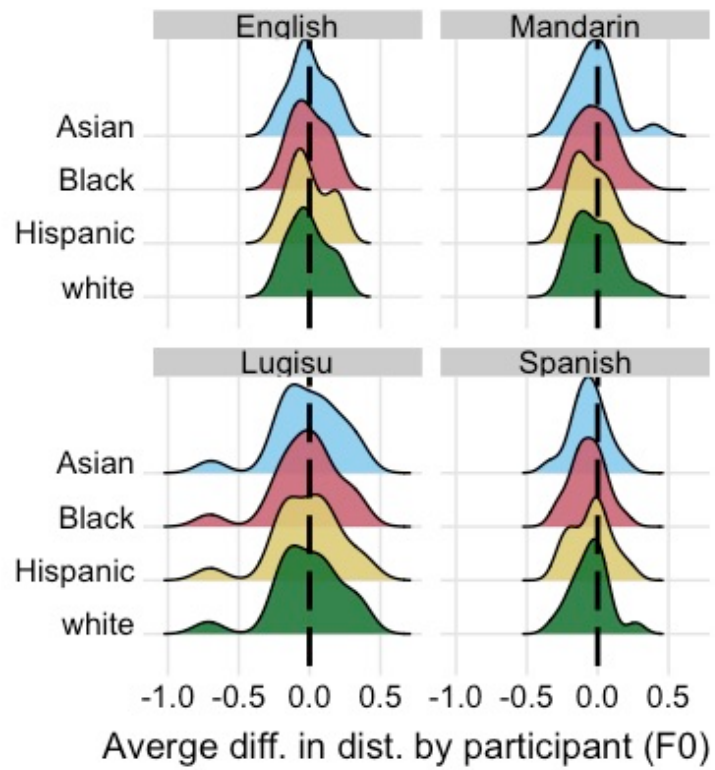


FIGURE 3.4: Average Difference in Distance by Participant F1,F2 spectra. Faceted Panels show distributions by stimuli voice. Ridgelines in each panel show distributions by face for a given voice

3.1 STATISTICAL ANALYSIS

To test predictions that the faces presented alongside speech would influence participants' productions, three linear mixed-effects models were designed to predict the difference in baseline and shadowed distances as a continuous dependent variable for each of the three acoustic measures. Vowel category, face, voice, and participant sex – multilevel, categorical, predictor variables – were included as sum-contrast coded fixed effects in each of these statistical models. To avoid over-fitting of any particular model, the maximal random effects structure justified by the experimental design and the resulting acoustic data were used (Barr, 2013; Bates et al., 2018), resulting in random intercepts fit for *participant* and *word* (or *sentence number* for the F0 model), as well as a random slope for the effect of sex within each level of *participant* (i.e. the correlation between intercept deviations and sex affect deviations between individual participants)¹. These regression models² were implemented in the *lme4* package and assessed for statistical significance in the *lmerTest* package (Bates et al., 2015; Kuznetsova et al., 2017)

MODEL 1

The first linear mixed model was fit to predict vowel spectra difference in distance with vowel quality, stimulus voice, stimulus face, and participant sex (*formula: vowel * voice + face + sex*) as sum-contrast coded fixed effects, and maximal data-driven random effects by participant, word, and a random slope for the effects of sex by participant (*formula: (1|ID) + (1|word) + (0 + sex | ID)*). Model comparison was used to justify the structure of effects for this and subsequent linear models. Of particular note is the interaction between vowel and voice in the fixed effects model as the only interaction which provide a better fit for this set of data. A likelihood ratio test between this and a simplified model without the vowel*voice interaction term (i.e. *vowel + voice + face + sex*) showed a significant interdependence of vowel and voice as fixed effects [$2(24)=364p < .001$]. Statistical significance of this and subsequent linear models' predictions was determined using Satterthwaite's approximation of degrees of freedom for F and t-statistics, as implemented in the *lmerTest* R package. The interaction between vowel and stimuli voice was significant as a predictor of difference in distance [$F(24) = 15.54, p < .001$] as well as significant main effects of vowel [$F(8) = 51.94, p < .001$] and voice [$F(3) = 13.07, p < .001$] as

¹The model failed to converge when using a more complex random effects structure such as a *Vowel*voice* interaction

²See appendix for more information about the statistical models implemented in this analysis

individual predictors. No other main effects were significant, contrary to predictions about the integration of social information and participant sex; neither face nor sex were significant predictors in this model (see table 3.5 for summary). The model's predictions for difference in distance based on the voice and face stimuli are shown in Figure 3.5a and Figure 3.5b respectively.

Table 3.5

	Estimate	Std. Error	t value	Pr(> t)
voice CH	-0.09	0.03	-3.28	< .001
voice GI	-0.15	0.03	-5.61	< .001
voice SP	-0.09	0.03	-3.45	< .001
face A	-0.01	0.03	-0.25	0.756
face B	-0.01	0.03	-0.51	0.362
face H	0.01	0.03	0.42	0.812
sex M	0.09	0.06	1.49	0.361
vowel α	-0.12	0.03	-3.43	< .001
vowel ε	0.37	0.03	12.14	< .001
vowel i	0.33	0.04	7.98	< .001
vowel ɪ	0.48	0.04	12.12	< .001
vowel ɔ	0.27	0.04	6.60	< .001
vowel u	0.24	0.05	5.39	< .001
vowel ʊ	0.17	0.05	3.44	< .001
vowel ʌ	0.22	0.03	6.53	< .001

MODEL 2

A second linear mixed model was fit to predict vowel duration with face, voice, vowel and sex (formula: face + voice * vowel + sex). The model included only the intercept by participant as a random effect (*formula: 1 | ID*), as any more complex random effects structure resulted in a failure for the model to converge. Within this model there is again significant interaction between voice and vowel, again estimated by performing a likelihood ratio test with a simplified model. The main effect of face is not statistically significant ($F(3) = 0.47, p = 0.705$). The main effect of voice is statistically significant ($F(3) = 65.75, p < .001$). The main effect of vowel is statistically significant and ($F(8) = 10.34, p < .001$). The main effect of sex is statistically not significant ($F(1) = 0.23, p = 0.640$) (see table 3.6 for summary). The model's predictions for difference in distance based on the voice and face stimuli are shown in Figure 3.6a and Figure 3.6b respectively.

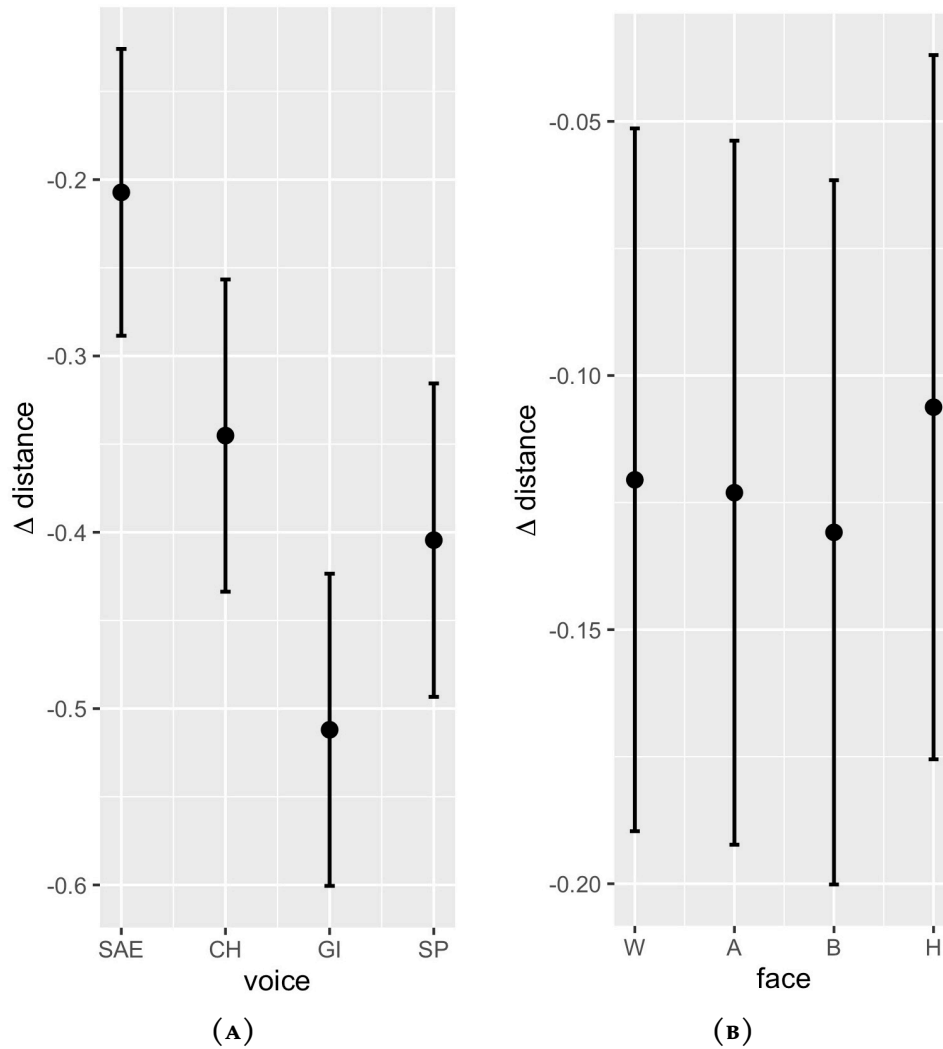


FIGURE 3.5: Linear mixed-effects model predicted values for difference in distance for vowel spectra. The left panel shows predictions based on voice (English, Mandarin, Lugisu, and Spanish left to right), and the right panel shows predictions based on face (white, Asian, Black, and Hispanic left to right)

MODEL 3

A third linear model was fit to predict F0 with face, voice, sentence and sex (formula: face + voice + sentence + sex). Notably, as F0 estimates were averaged by recording, vowel was not included as a fixed effect, and instead the sentence shadowed in the recording was included as a fixed effect. There were no interaction terms in this model as determined by model comparison. The model included only the intercept by participant as a random effect (formula: 1 | ID), as any more complex random effects structure resulted in a failure for the model to converge. Within the model, the main effect of face is statistically not significant ($F(3) = 0.30$, $p = 0.824$). The main effect of voice is statistically significant ($F(3) = 12.57$, $p < .001$).

Table 3.6

	Estimate	Std. Error	df	t value	Pr(> t)
face A	-0.00	0.00	9082.47	-0.18	0.85
face B	0.00	0.00	9082.28	1.16	0.25
face H	-0.00	0.00	9082.14	-0.58	0.56
voice CH	-0.01	0.00	9084.39	-7.29	0.00
voice GI	-0.01	0.00	9087.64	-8.38	0.00
voice SP	0.01	0.00	9082.82	10.72	0.00
vowel ɑ	0.00	0.00	9082.91	3.48	0.00
vowel ɛ	0.00	0.00	9082.41	0.66	0.51
vowel i	-0.01	0.00	9082.75	-6.44	0.00
vowel ɪ	0.01	0.00	9082.24	3.27	0.00
vowel ɔ	0.00	0.00	9083.03	3.11	0.00
vowel u	-0.00	0.00	9083.04	-0.01	1.00
vowel ʊ	-0.01	0.00	9082.17	-3.07	0.00
vowel ʌ	-0.00	0.00	9082.48	-0.90	0.37
sex M	-0.00	0.00	16.04	-0.48	0.64

The main effect of sentence is statistically significant ($F(14) = 6.55, p < .001$) The main effect of sex is statistically not significant ($F(1) = 4.08, p = 0.059$) (see table 3.7 for summary). The model's predictions for difference in distance based on the voice and face stimuli are shown in Figure 3.7a and Figure 3.7b respectively.

Table 3.7

	Estimate	Std. Error	df	t value	Pr(> t)
face A	0.00	0.01	3997.00	0.60	0.55
face B	0.00	0.01	3997.00	0.09	0.93
face H	-0.00	0.01	3997.01	-0.89	0.37
voice CH	0.02	0.01	3997.00	4.16	0.00
voice GI	0.00	0.01	3997.00	0.91	0.36
voice SP	0.00	0.01	3997.00	0.57	0.57
sex M	-0.12	0.06	17.00	-2.02	0.06

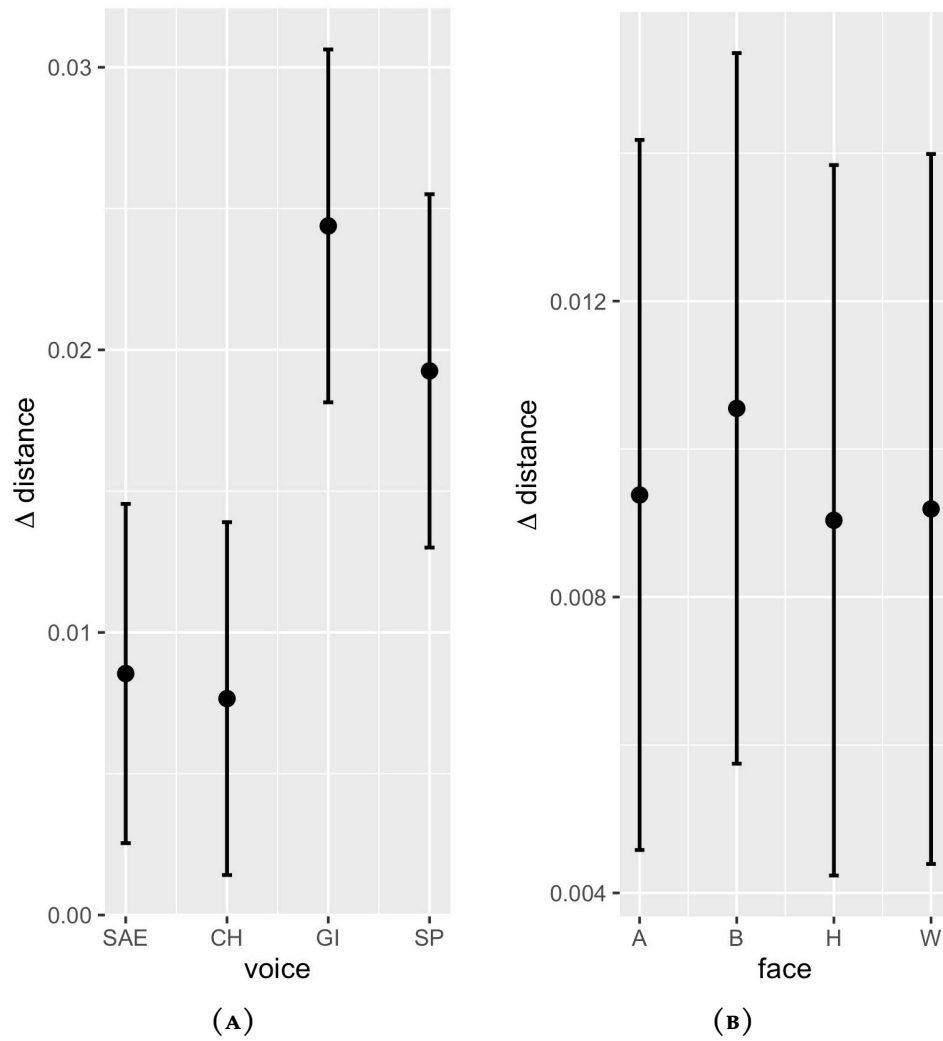


FIGURE 3.6: Linear mixed-effects model predicted values for difference in distance for vowel duration. The left panel shows predictions based on voice (English, Mandarin, Lugisu, and Spanish left to right), and the right panel shows predictions based on face (white, Asian, Black, and Hispanic left to right).

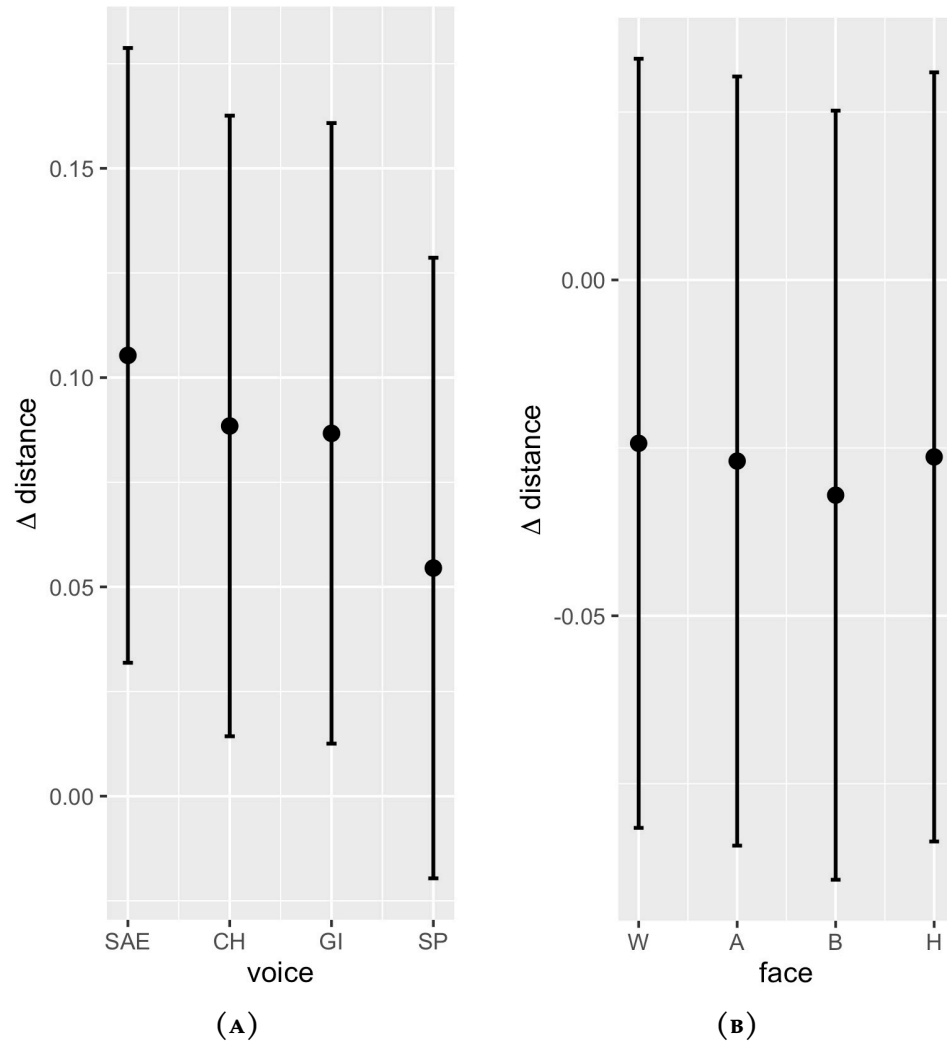


FIGURE 3.7: Linear mixed-effects model predicted values for difference in distance for fundamental frequency (F0). The left panel shows predictions based on voice (English, Mandarin, Lugisu, and Spanish left to right), and the right panel shows predictions based on face (white, Asian, Black, and Hispanic left to right).

CHAPTER 4. DISCUSSION

While there were clear changes to participants' productions between the baselines as compared to the shadowing block with reference to the stimuli, and there is evidence that participants converged toward the four model talkers with respect to three measures to varying degrees, the models fit to the production data were unable to determine significant effects of any independent variables aside from vowel quality and the voice that was shadowed. Although sex approached significance as a predictor of F0, this otherwise there seemed to be no reliable effect or interaction with speaker sex in contrast to the findings of previous studies that female participants converged more toward the model talker.

Potential failures of this experiment to replicate the results of similar production studies measuring phonetic convergence are perhaps explained in part by the lack of power resulting from the size of the participant pool, but also may be explained by the nature of the experimental design as fundamentally seeking answers to a question over and above convergence. Additionally, this study is lacking an additional perceptual experiment to supplement the limited sets of data in the acoustic analyses of convergence as is typical in the design of shadowing tasks. That three variables may not adequately capture the effects present in the data was planned for initially, although experiment two was unable to be implemented in a time frame that would allow data to be collected and analyzed.

Overall, my alternative hypothesis that face would be a predictor of change in production was not supported by these models. This is inconsistent with other the findings of other matched guise and inverted matched guise studies. There are several potential interpretations for this result, most simply that participants simply didn't attend to the visual stimuli, or didn't find it to be meaningfully related to the audio given that every possible face-voice pair was displayed in random order for each participant during the shadowing block. Alternatively, the lack of evidence that visual information interacted with linguistic information in this experiment may result from the explicit instructions given to participants to repeat the voice exactly as they heard it. Assuming that participants performed the trials to the best of their ability based on this instruction, the instructions themselves may have encouraged participants *not* to attend to the visual information, and prevented effects of social expectations as primed by face to have a strong influence on their imitations. Note that though this speculation may seem to be in contrast to find-

ings from [Dufour & Nguyen \(2013\)](#), [Pardo et al. \(2010\)](#), and [Sato et al. \(2013\)](#) as recently reported and confirmed by [Clopper & Dossey \(2020\)](#) that explicit instruction to imitate the model talker enhances, the claim here is only about expectations (as set by the integration of multiple perceptual modalities) as opposed to degree of convergence.

Similarly, and in line with discussions of non-native speech perception studies such as [Baese-Berk \(2019\)](#), wherein speakers reacting to sufficiently unfamiliar speech are better understood as learning to adapt to the novel stimuli. [Ferreira & Pashler \(2002\)](#) propose a central bottleneck effect in speech processing- that is, if two processes share resources (as perception and production would under the assumption that they are closely linked) then trying to simultaneously or in quick succession perform these relatively complex processes of perception and production would result in slow-down or hindrance of one or the other of these processes. Were there any statistical support for a third modality in which social information is integrated visually into language processing, there is an even greater potentiality that the cognitive bottleneck produced in a timed, non-native sentence shadowing task caused subjects to attend to only the most relevant (to the speaker) information present in a given modality such that the a parsimonious, but potentially less-accurate-than-possible imitation is produced. This would effectively reduce the complexity and phonetic detail available to speakers to attend to, and produce (cf. Discussion of interpretations of [Niedzielski \(1999\)](#) and [Schulman \(1983\)](#) in [McGowan \(2015\)](#)).

CHAPTER 5. CONCLUSIONS

The observed results of difference in distance as a metric for phonetic convergence are somewhat in line with the findings of accommodation studies in that similarity to model talker or interlocutor emerge as significant global trends, but that the magnitude of the convergence (and in this case divergence as well) is often quite varied among participants based upon social and linguistic predictor variables. I found little evidence in the data presented here that might provide insight into a linked model of perception and production without a way of further quantifying the change observed in imitations, and similarly have no robust evidence that visual information was at all considered by participants in performing processes of perception or production individually.

As in [Preston \(1992\)](#), that there were somewhat systematic patterns (in this case, a tendency toward acoustic similarity) among participants' imitations does suggest a level of awareness of the features of the model talkers' speech. Unlike Preston's observations of caricature features present in imitation that were not present in the imitated variety, it seems that the shadowing productions in this study were more often faithful to the stimuli across the metrics examined. That convergence to the auditory stimuli remained the overall tendency across participants and across conditions is particularly interesting given that the participants in this study reported having no familiarity with the three non-English language varieties spoken by the model talkers, and were given no training or instruction on the L2 accented speech beyond being asked to reproduce it. Based on earlier discussions of awareness and linguistic and social selectivity, it would seem more likely that unfamiliar linguistic features would rely more heavily on expectations set by visual stimuli, but the opposite seems to be the case for the participants in this study.

The results of this experiment are also in contrast to the findings of inverted matched guise studies such as [Rubin \(1992\)](#) and [McGowan \(2015\)](#). Though Rubin's study shows social expectations having a negative effect on perception, and McGowan's shows that perceptual accuracy improves when linguistic features are congruent with the social expectations about a speaker, the two both find that manipulation of listeners' social expectations has a significant effect on perception. In contrast, the experiment reported here shows no evidence of expectations established by visual stimuli impacting perception either positively or negatively. Participants' imitations of these voices didn't seem to be the product of perceptions

based on anything other than the features present in the stimuli itself. That is to say, these participants did exactly what was asked of them instead of what was expected based on previous studies.

To arrive at more meaningful conclusions, it seems feasible that a similar experimental design (integrating matched guise and shadowing paradigms) would remain a suitable way to address these research questions, but future work necessitates perceptual analyses, as well as more robust acoustic measurements, and how the relationships between these phenomena adapt over the course of exposure to unfamiliar linguistic and social categories. The lack of evidence for participants' expectations from visual stimuli in combination with what seems to be fairly consistent awareness of the features of unfamiliar accents and speakers is an area that warrants further investigation.

One possible direction for future work may be to examine these or similar data with reference to whether shadowed speech increases in similarity to the model talkers during the time-course of the experiment itself in order to understand if repeated exposure to the speech results in emergent awareness of the phonetic features present in an unfamiliar voice, and whether social expectations do play a role in these imitations before participants develop strategies for determining which features are most salient indices of a particular voice. If it is the case that participants behave differently as speakers become more familiar, such an analysis may help to explain the current study's unexpected results and to give more insight into the subjective nature of cognitive processes which shape perceptions of speech.

APPENDIX

Below is demographic information for each of the four model talkers as reported in the ALLSTAR corpus:

ID	ALL_037_M_CMN_ENG
Sex	male
Ethnicity/race	Asian
Age	22
Birthplace	Wuhan, Hubei Province, People's Republic of China
Language(s)	Mandarin, Wuhan Chinese, English, French
ID	ALL_024_M_GIS_ENG
Sex	male
Ethnicity/race	African American
Age	29
Birthplace	Mbale, Republic of Uganda
Language(s)	Gishu(Lugisu), Luganda, Swahili, English
ID	ALL_128_M_SPA_ENG_HT1
Sex	male
Ethnicity/race	Hispanic; white
Age	24
Birthplace	State of Tamaulipas United States of Mexico
Language(s)	Spanish, English, French
ID	ALL_052_M_ENG_ENG_HT1
Sex	male
Ethnicity/race	white
Age	19
Birthplace	New York City, New York, United States of America
Language(s)	English

REFERENCES

- ANSI/ACA. (2018). *Maximum permissible ambient noise levels for audiometric test rooms* (Tech. Rep. No. S3.1-1999 (R2018)). American National Standard Institute and Acoustical Society of America.
- Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics - J PHONETICS*, 40. doi: 10.1016/j.wocn.2011.09.001
- Baese-Berk, M. (2019). Interactions between speech perception and production during learning of novel phonemic categories. *Attention, Perception, & Psychophysics*, 81. doi: 10.3758/s13414-019-01725-4
- Barr, D. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, 4, 328. Retrieved from www.frontiersin.org/Article/10.3389/fpsyg.2013.00328
- Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). *Parsimonious mixed models*.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. doi: 10.18637/jss.v067.i01
- Beddor, P. (2009). A coarticulatory path to sound change. *Language*, 85, 785-821. doi: 10.1353/lan.0.0165
- Beddor, P., Coetzee, A., Styler, W., McGowan, K., & Boland, J. (2018). The time course of individuals perception of coarticulatory information is linked to their production: Implications for sound change. *Language*, 94, 931-968. doi: 10.1353/lan.2018.0051
- Boersma, P., & Weenink, D. (2021). *Praat: doing phonetics by computer [computer program]*. Retrieved from <http://www.praat.org/>
- Bradlow, A., Torretta, G., & Pisoni, D. (1996). Intelligibility of normal speech i: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Comm.*, 20, 255-272. doi: 10.1016/S0167-6393(96)00063-5
- Bradlow, A. R. (n.d.). *Oscar: The online speech/corpora archive and analysis resource*. Retrieved from <https://oscar3.ling.northwestern.edu>
- Campbell-Kibler, K. (2010). Sociolinguistics and perception. *Language and Linguistics Compass*, 4, 377-389. doi: 10.1111/j.1749-818X.2010.00201.x

- Cantor-Cutiva, L. C., Bottalico, P., & Hunter, E. (2018, Jul). Factors associated with vocal fry among college students. *Logopedics, phoniatrics, vocology*, 43(2), 73-79. Retrieved from <https://doi.org/10.1080/14015439.2017.1362468> doi: 10.1080/14015439.2017.1362468
- Chambers, K., Onishi, K., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, 87, B69-77. doi: 10.1016/s0010-0277(02)00233-0
- Clopper, C. G., & Dossey, E. (2020). Phonetic convergence to southern american english: Acoustics and perception. *The Journal of the Acoustical Society of America*, 147(1), 671-683.
- Dallaston, K., & Docherty, G. (2017). Sociophonetic variation in the prevalence of creaky phonation in the speech of young adults from two capital cities in australia. *Paper presented at the Language Variation and Change (LVC-A3) Workshop*.
- Dallaston, K., & Docherty, G. (2020). The quantitative prevalence of creaky voice (vocal fry) in varieties of english. a systematic review of the literature. *PLoS ONE*, 15. doi: 10.1371/journal.pone.0229960
- Delvaux, V., & Soquet, A. (2007). The influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica*, 64, 145-73. doi: 10.1159/0000107914
- Dufour, S., & Nguyen, N. (2013). How much imitation is there in a shadowing task? *Frontiers in psychology*, 4, 346. doi: 10.3389/fpsyg.2013.00346
- Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, 41, 87-100. doi: 10.1146/annurev-anthro-092611-145828
- Ferreira, V., & Pashler, H. (2002). Central bottleneck influences on the processing stages of word production. *Journal of Experimental Psychology: Learning Memory and Cognition*, 28, 1187-1199. doi: 10.1037/0278-7393.28.6.1187
- Foulkes, P., & Hay, J. (2015). The emergence of sociophonetic structure. In (p. 292-313). doi: 10.1002/9781118346136.ch13
- Giles, H., Taylor, D., & Bourhis, R. (1973). Towards a theory of interpersonal accommodation through language: Some canadian data. *Language in Society*, 2, 177 - 192. doi: 10.1017/S0047404500000701
- Goldinger, S. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological review*, 105, 251-79. doi: 10.1037/0033-295X.105.2.251
- Harrington, J., Kleber, F., & Reubold, U. (2008a). Compensation for coarticulation, /u/-fronting, and sound change in standard southern british: an acoustic and perceptual study. *The Journal of the Acoustical Society of America*, 123, 2825-35. doi: 10.1121/1.2897042

- Harrington, J., Kleber, F., & Reubold, U. (2008b). Compensation for coarticulation, /u/-fronting, and sound change in standard southern british: An acoustic and perceptual study. *The Journal of the Acoustical Society of America*, 123(5), 2825-2835. Retrieved from <https://doi.org/10.1121/1.2897042> doi: 10.1121/1.2897042
- Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. , 145-165.
- Johnson, K. (2005). Speaker normalization in speech perception. *The handbook of speech perception*, 363-389.
- Jusezyk, P., & Luce, P. (2002). Speech perception and spoken word recognition: Past and present. *Ear and hearing*, 23, 2-40. doi: 10.1097/00003446-200202000-00002
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. doi: 10.18637/jss.v082.i13
- Labov, W., Ash, S., & Boberg, C. (2006). *The atlas of north american english Phonetics, phonology, and sound change* (American Dialect Society, Ed.). Mouton de Gruyter.
- Labov, W., Rosenfelder, I., & Fruehwald, J. (2013). One hundred years of sound change in philadelphia: Linear incrementaion, reversal, and reanalysis. *Language*, 89(1), 30-65.
- Lambert, W., Hodgson, R., Gardner, R., & Fillenbaum, S. (1960). Evaluational reactions to spoken languages. *Journal of abnormal and social psychology*, 60. doi: 10.1037/h0044430
- Lambert, W. E., Hodgson, R., Gardner, R. C., & Fillenbaum, S. (1960). Evaluational reactions to spoken language. *Journal of Abnormal and Social Psychology*, 60, 44-51.
- Lee, K. E. (2016). *The perception of creaky voice: Does speaker gender affect our judgments* (Master's thesis, University of Kentucky Theses and Dissertations–Linguistics). Retrieved from https://uknowledge.uky.edu/ltt_etds/17
- Lieberman, M. (2015). *Language log- reaper*. Retrieved from www.languagelog ldc.upenn.edu/n11/?p=17590
- Lindblom, B., Guion, S., Hura, S., Moon, S.-J., & Willerman, R. (1995). Is sound change adaptive? *Rivista di Linguistica*, 7, 5-37.
- Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, 47(4), 1122-1135. Retrieved from <https://doi.org/10.3758/s13428-014-0532-5> (Database can be accessed at <https://chicagofaces.org/>)
- Mathôt, S., Schreij, D., & Theeuwes, J. (2012). Opensesame: An open-source, graphical experiment builder for the social sciences. *Behavior research methods*, 44(2), 314-324.

- McAuliffe, M., Socolof, M., Stengel-Eskin, E., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). *Montreal forced aligner* [computer program]. Retrieved from <http://montrealcorpus tools.github.io/Montreal-Forced-Aligner/>
- McGowan, K. B. (2011). *The role of socioindexical expectation in speech perception* (Unpublished doctoral dissertation). University of Michigan.
- McGowan, K. B. (2015). Social expectation improves speech perception in noise. *Language and Speech, 58*(4), 502-521. Retrieved from <https://doi.org/10.1177/0023830914565191> doi: 10.1177/0023830914565191
- Mitterer, H., & Ernestus, M. (2008). The link between speech perception and production is phonological and abstract: Evidence from a shadowing task. *Cognition, 109*, 168-73. doi: 10.1016/j.cognition.2008.08.002
- Natale, M. (1975). Social desirability as related to convergence of temporal speech patterns. *Perceptual and Motor Skills, 40*, 827-830. doi: 10.2466/pms.1975.40.3.827
- Nguyen, N., Dufour, S., & Brunellière, A. (2012). Does imitation facilitate word recognition in a non-native regional accent? *Frontiers in psychology, 3*, 480. doi: 10.3389/fpsyg.2012.00480
- Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology, 18*. doi: 10.1177/0261927X99018001005
- Nielsen, K. (2011). Specificity and abstractness of vowel imitation. *Journal of Phonetics - J PHONETICS, 39*, 132-142. doi: 10.1016/j.wocn.2010.12.007
- Nilsson, M., Soli, S., & Sullivan, J. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and noise. *Journal of the Acoustical Society of America, 95*(2), 1085-1099.
- Nycz, J., & Hall-Lew, L. (2013). Best practices in measuring vowel merger. *The Journal of the Acoustical Society of America, 134*, 4198. doi: 10.1121/1.4831400
- Pardo, J. (2007). On phonetic convergence during conversational interaction..
- Pardo, J. (2012). Reflections on phonetic convergence: Speech perception does not mirror speech production. *Language and Linguistics Compass, 6*, 753-767. doi: 10.1002/lnc3.367
- Pardo, J., Jay, I., & Krauss, R. (2010). Conversational role influences speech imitation. *Attention, perception & psychophysics, 72*, 2254-64. doi: 10.3758/APP.72.8.2254
- Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. *Typological Studies in Language, 45*, 1-11. doi: 10.1075/tsl.45.08pie
- Podesva, R. (2013). Gender and the social meaning of non-modal phonation types..

- Preston, D. R. (1992). Talking black and talking white: A study in variety imitation. *Old English and New: Studies in Language and Linguistics in Honor of Frederic G. Cassidy*, 327-355.
- Preston, D. R. (2016). Whaddayaknow now? In A. M. Babel (Ed.), *Awareness and control in sociolinguistic research* (p. 177199). Cambridge University Press. doi: 10.1017/CBO9781139680448.010
- Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H., & Yuan, J. (2014). *Fave (forced alignment and vowel extraction) program suite*. doi: 10.5281/zenodo.22281
- Rubin, D. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative english-speaking teaching assistants. *Research in Higher Education*, 33, 511-531. doi: 10.1007/BF00973770
- Saffran, J. (2001). Words in a sea of sounds: The output of infant statistical learning. *Cognition*, 81, 149-69. doi: 10.1016/S0010-0277(01)00132-9
- Sato, M., Grabski, K., Garnier, M., Granjon, L., Schwartz, j.-l., & Nguyen, N. (2013). Converging toward a common speech code: Imitative and perceptuo-motor recalibration processes in speech production. *Frontiers in Psychology*, 4, 422. doi: 10.3389/fpsyg.2013.00422
- Scarborough, R., & Zellou, G. (2013). Clarity in communication: 'clear' speech authenticity and lexical neighborhood density effects in speech production and perception. *Journal of the Acoustical Society of America*, 134, 3793-3807.
- Schulman, R. (1983). Vowel categorization by the bilingual listener. *PERILUS Working Papers*, III, 8199.
- Shockley, K., Sabadini, L., & Fowler, C. (2004). Imitation in shadowing words. *Perception & psychophysics*, 66, 422-9. doi: 10.3758/BF03194890
- Soli, S. D., & Wong, L. L. (2008). Assessment of speech intelligibility in noise with the hearing in noise test. *International Journal of Audiology*, 47(6), 356-361.
- Szakay, A., & Torgersen, E. (2019). A re-analysis of f0 in ethnic varieties of london english using reaper..
- Talkin, D. (2015). *Robust epoch and pitch estimator*. Retrieved from <https://github.com/google/REAPER>
- Trautmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *Acoustical Society of America*, 89, 97100.
- Vermiglio, A. J. (2008). The american english hearing in noise test. *International Journal of Audiology*, 47(6), 386-387. (PMID: 18569120) doi: 10.1080/14992020801908251

- Walker, A., & Campbell-Kibler, K. (2015). Repeat what after whom? exploring variable selectivity in a cross-dialectal shadowing task. *Front. Psychol.*, 6(546), 1-18.
- Xie, X., Liu, L., & Jaeger, T. F. (2021). Cross-talker generalization in the perception of non-native speech: a large-scale replication. *Journal of Experimental Psychology General*. doi: 10.1037/xge0001039
- Yu, A., Abrego-Collier, C., & Sonderegger, M. (2013). Phonetic imitation from an individual-difference perspective: Subjective attitude, personality and "autistic" traits. *PloS one*, 8, e74746. doi: 10.1371/journal.pone.0074746
- Yuan, J., & Liberman, M. (2008). Speaker identification on the scotus corpus. In *Proceedings of Acoustics, '08*. Retrieved from <http://www.praat.org/>

VITA

Kyler B. Laycock was born in Central City, KY. He graduated from Muhlenberg County High School in spring 2015, and received a Bachelors of Arts in Linguistics from the University of Kentucky College of Arts and Sciences in the spring of 2019. Upon publication of this thesis, he will have completed an M.A. in Linguistic Theory and Typology from the University of Kentucky.