Check for updates

Original Article

Language *and* Speech

The Reliability of Individual Differences in VOT Imitation

Language *and* Speech
1–18
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0023830920947769
journals.sagepub.com/home/las
SAGE

Lacey Wade, Wei Lai and Meredith Tamminga
University of Pennsylvania, USA

Abstract
Recent work has shown that individuals vary in phonetic behaviors in ways that deviate from group norms and are not attributable to sociolinguistically relevant dimensions such as gender or social class. However, it is unknown whether these individual differences observed in the lab are stable characteristics of individuals or whether they simply reflect noise or sporadic fluctuations. This study investigates the individual-level stability in imitation of a model talker's artificially-lengthened VOT. We use a test–retest design in which the same set of participants perform the same lexical shadowing task on two separate occasions and find that degree of convergence or divergence is highly correlated on an individual basis across visits. Further, we find a strong correlation between individual VOT shifts toward a male model talker and shifts toward a female model talker. Findings contribute to a growing body of literature suggesting that averaging over groups of participants masks the complexity of phonetic behaviors, such as imitation, and suggest that individual differences in phonetic behavior are an area of promising future study.

Keywords
Phonetic imitation, individual differences, shadowing, test–retest reliability

Introduction

Experimental work in linguistics typically generalizes over the aggregate behavior of many participants. However, it is widely understood that such generalizations may obscure individual-level divergence from the overall group pattern. Individual differences are of theoretical interest because the ways in which speaker-hearers diverge can have implications for the structure of the phonetics-phonology interface and have been hypothesized to play a role in sound change. However, it is not well understood whether the individual differences we observe in the lab reflect stable, characteristic traits of particular individuals, or whether they simply reflect noise or chance fluctuations (Kingston et al., 2015). One domain in which individual differences have been

Corresponding author:
Lacey Wade, 3401 Walnut St., Suite C 300, Philadelphia, PA 19104-6243, USA.
Email: laceya@sas.upenn.edu

observed is phonetic imitation, also called convergence (Abrego-Collier et al., 2011; Babel, 2012; Yu et al., 2013). In this paper we adopt a widely-used experimental paradigm for eliciting imitation, speech shadowing, to ask whether individuals who participate in the same shadowing task on two different occasions respond in the same way both times. Our primary interest is in participants' imitation of voiceless stop VOTs that have been artificially lengthened. We show that individual differences on the VOT imitation task are quite reliable across two different occasions, bolstering the value of further inquiry into such differences.

2 Background

It has been well established that speakers tend to imitate various aspects of speech that they hear, including pitch (Babel & Bulatov, 2011; Gijssels et al., 2016), VOT (Nielsen, 2011; Shockley et al., 2004), speech rate (Schweitzer & Lewandowski, 2013; Staum Casasanto et al., 2010), vowel quality (Babel, 2009, 2012; Pardo et al., 2012), and coarticulatory nasalization (Zellou et al., 2016). Such imitation happens not only in socially-rich interactional contexts (Giles et al., 1991; Pardo, 2006), but also in relatively asocial laboratory settings, where participants who are exposed to audio clips of a model talker tend to become more like that talker over the course of an experiment (Goldinger, 1998; Nielsen, 2011; Shockley et al., 2004). While aggregate patterns of convergence are widely observed, it is also known to be the case that individuals sometimes differ substantially in whether or how much they converge toward an interlocutor or experimental model talker, a point which has attracted increasing attention in recent work (Babel, 2012; Pardo et al., 2018; Sonderegger et al., 2017; Zellou, 2017). Imitation is far from the only phenomenon for which such inter-individual variability can be observed; Yu and Zellou (2019) and Schertz and Clare (2019) provide valuable overviews of the many dimensions along which individuals may differ in their speech perception and production. While individual differences are frequently observed in the literature, there is still some question as to whether such differences observed on any given occasion represent stable, characteristic properties of those individuals and their linguistic systems (Kingston et al., 2015). A certain amount of chance fluctuation both within and across individuals is to be expected in empirical data on human behavior, so it is worth taking the time to establish the stability of apparent individual differences.

One approach to establishing that observed individual differences are characteristic of specific speaker-hearers, in imitation and in other behaviors, has been to correlate such differences with other individual-level properties. The reasoning is that if individual differences observed in linguistic performance simply reflected noise in the experimental results, they would not be expected to correlate with other individual-level, non-linguistic traits. In the domain of phonetic imitation, Yu et al. (2013) find that individuals with greater Openness scores on the Big Five Inventory (BFI) of personality traits (Goldberg, 1992) and higher Attention Switching scores on the Autism-spectrum Quotient (AQ, Baron-Cohen et al., 2001) imitate lengthened VOT to a greater extent, likely because greater openness and attention switching would indicate greater engagement with and focus on exposure materials. Lewandowski and Jilka (2019) also find greater convergence in a cooperative Diapix task for those with higher Openness scores on the BFI, as well as for those with greater attention switching capabilities, though this was measured more objectively using the Simon Test, rather than the self-assessed AQ survey. They further find that Neuroticism scores on the BFI and lower Behavior Inhibition Scale (BIS) scores (which measures motivation to avoid aversive outcomes) facilitated convergence and suggest this is because greater Neuroticism scores may indicate a higher need for social approval, which may motivate convergence, and those with lower behavior inhibition scores are more likely to try converging because they have a

lower punishment sensitivity. Relatedly, Aguilar et al. (2016) find that individuals with a high level of trait rejection sensitivity (the tendency to expect social rejection) exhibit greater convergence during conversation than those with low trait rejection sensitivity. Beyond imitation, individual-level correlations have been reported for many other combinations of linguistic behaviors and non-linguistic traits (Dimov et al., 2012; Kingston et al., 2015; Lev-Ari & Peperkamp, 2014; Morley, 2014; Perrachione et al., 2011; Stewart & Ota, 2008; Turnbull, 2015; Van Hedger et al., 2015; Yu, 2010, inter alia).

However, there is also some reason to be cautious about the extent to which individual differences are stable and systematic. For one thing, it is not always the case that the specific correlates of individual differences are consistent across studies of the same behavioral phenomenon, raising questions about their replicability and interpretation that are especially difficult to answer in the face of widespread structural issues such as the file drawer problem, which refers to the tendency for positive results to be published over null results (Rosenthal, 1979). For example, Tamminga et al. (2018) do not replicate Yu et al.'s (2013) result that BFI Openness and AQ Attention Switching correlate with VOT imitation. The fact that these studies used different imitation tasks raises another issue, which is that different types of tasks intended to gauge the same phonetic behavior in the same linguistic domain do not necessarily correlate on an individual basis (Shultz et al., 2012; Tilsen & Cohn, 2016; Yu & Zellou, 2019; Zellou, 2017). Yu and Lee (2014) ask whether listeners show consistent behavior on two different task types that nominally tap the same behavior, perceptual compensation for coarticulation. They find that the individual-level correlation is statistically significant but not very strong. Similarly, Lai and Tamminga (2019) find that listeners who exhibit greater perceptual compensation for lexical cues compensate less for coarticulatory cues. In the domain of imitation, Pardo et al. (2018) show that when the same group of participants perform both an isolated-word shadowing task and a conversational task, the overall degree of convergence is similar across tasks but the participant-level correlations are weak. Of course, different tasks may have different goals, recruit different peripheral skills, and ultimately vary in regard to whether idiosyncratic properties are relevant at all (Yu & Zellou, 2019). More broadly, general linguistic behaviors such as imitation do not necessarily correlate across different linguistic contexts where that behavior may arise; for example, Sanker (2015) finds that the degree of imitation exhibited by an individual varies across different phonetic features. For instance, degree of imitation among F1, F2, F0, vowel duration, and intensity showed no correlations on a individual speaker basis. This suggests that it may not be possible to classify an individual as an across-the-board imitator, although it could plausibly be the case that individual propensities toward imitation are feature-specific but nonetheless stable. Relatedly, Schertz and Clare (2019) survey a range of both positive and null results in studies testing individual-level perception–production correlations for various linguistic features. Again, this may or may not undermine our confidence in the stability of individual differences; as Schertz and Clare (2019) discuss in detail, there are many reasons why perception and production might dissociate.

We would argue that the strongest evidence that observed individual differences reflect characteristic traits of those individuals are studies that assess whether individuals perform similarly across different occasions. The small number of studies in this vein have, broadly speaking, yielded more positive results than the ones just discussed. Schertz et al. (2015) find a strong correlation between listeners' cue weighting in the discrimination of Korean stop contrasts across two laboratory visits. Relative perceptual weighting of VOT, F0 at vowel onset, and closure duration were stable for individuals across visits, but were not predicted by their own production. Similarly, Kong and Edwards (2016) find that individuals are consistent in their degree of categorical perception and cue weighting on two different occasions. Individuals who exhibited a more gradient response

pattern for the stop voicing contrast in English were also more sensitive to F0 as a perceptual cue, and this pattern was stable for individuals across visits. These studies are promising for at least a narrow view of the reliability of individual differences: that, having done a particular task with particular stimuli in a particular context on one day, an individual participant will generally perform similarly on a different day, faced again with the same task with the same stimuli in the same context. However, the study of individual differences in general would benefit greatly from a larger number of studies following this kind of design; in particular, both of these studies focus on cue-weighting behavior in speech perception. In this paper we report a simple test-retest reliability study of an imitation task modeled on a standard paper from a widely-used experimental paradigm: Shockley et al. (2004) and their use of a lexical repetition task, commonly called shadowing. We show that, as in Schertz et al. (2015) and Kong and Edwards (2016), our participants behave differently from each other, but each individual behaves similarly to their own previous performance when they revisit the lab. Establishing this narrow kind of reliability across a wider range of phenomena, in production as well as perception, is useful for contextualizing and interpreting the various dissociations, mismatches, and inconsistencies of the literature we have just discussed. It will allow us to scale up to questions of greater generality, such as whether participants who imitate one feature are likely to imitate another, or whether participants who imitate one voice are likely to do so with a novel voice.

3 Experimental design

The experiment we report here is based on the methods from Shockley et al. (2004), who found imitation of artificially extended VOT using a lexical shadowing design. In this experiment, participants read aloud words that begin with voiceless stops, then repeat the same words aloud after a model talker whose word-initial VOT has been artificially doubled. Shockley et al. (2004) compared these two conditions using both direct VOT measurements and a second experiment to elicit AXB perceptual assessments. Both measures find evidence of convergence toward the model talkers: the shadowed tokens were more likely than the baseline tokens to be judged a match for the model talker's tokens, and they also had longer VOT than the baseline tokens. Our experiment differs from Shockley et al. (2004) in a number of ways. We forego the AXB task and report only the difference in VOT between reading and shadowing conditions. After participants failed to converge in an initial pilot of the experiment, we made several design decisions in order to increase the motivation to converge. For instance, our stimuli contain both a male and a female voice, with accompanying facial images to increase the social motivation to converge. We also add a lexical decision phase between reading and shadowing to increase exposure to the lengthened-VOT stimuli. However, the lexical decision task data will not be analyzed here, as the task was designed solely to increase exposure to the VOT-lengthened stimuli. Most importantly, we have the same participants return to the lab to repeat the entire experiment on a second occasion, 7 to 14 days after their first lab visit.

3.1 Participants

We recruited 34 undergraduate students at the University of Pennsylvania who identified themselves as native speakers of English, and who received course credit for their participation. Eleven participants were excluded due to recording failures, not attending a second visit, or failing to complete more then half of the experiment. Twenty-three participants, 12 male and 11 female, are included for analysis here. The mean number of days elapsed between Visit 1 and Visit 2 was 7.9, with a range of 7–14.

Figure 1. Photos paired with female and male voices.

3.2 Materials

Stimuli for the shadowing task consisted of 84 words, divided evenly among /p/-initial, /t/-initial, and /k/-initial words, shown in Figure 5. All words were disyllabic and contained no other voiceless stops. Stimuli were recorded with a Yeti microphone at a sampling rate of 44,100 Hz. Half of the words were read by a male talker and the other half were read by a female talker, which allows us to additionally perform a post-hoc analysis of whether convergence is relatively consistent across model talkers. Words were balanced across the two talkers for frequency, using the Subtlex corpus (Brysbaert & New, 2009) LOG10 CD measure, and initial segment (/p t k/), as well as roughly balanced for following vowel. VOT of each stimulus was lengthened by doubling the duration of the aspiration from the stop burst to the onset of the vowel using the duration tier in Praat (Boersma & Weenick, 2018), which allows for lengthening or shortening of relative duration while maintaining pitch, using linear interpolation between duration points. VOT was lengthened to a mean of 221.7 ms for the female talker and 153.4 ms for the male talker. Peak amplitude was normalized across stimuli using the Amplify feature in Audacity (Audacity Team, 2020),. The same recordings were used in the lexical decision exposure phase, and the non-words used for that phase were recorded in the same way. An image of a female face was paired with all stimuli presented in the female voice, and an image of a male face was paired with all stimuli presented in the male voice. The facial images were both of white young adults smiling, as shown in Figure 1.

3.3 Procedure

Participants completed the imitation experiment in the Language Variation and Cognition Lab at the University of Pennsylvania. The experiment consisted of three phases: a read-aloud phase, a lexical decision phase, and a shadowing phase, all administered using PsychoPy (Peirce et al., 2019). In the read-aloud phase, participants were instructed to "Identify the word you see by speaking it into the microphone quickly but clearly." Participants read 84 target words as they appeared on the screen at 2-second intervals, and participant responses were recorded with a Yeti microphone. Next, participants completed an auditory lexical decision task, where they were instructed to "Decide as quickly as you can whether what you hear is a word or a non-word." Lexical decision stimuli consisted of the same 84 words with artificially doubled VOT, plus 40 non-words (shown in Figure 6) which did not contain a voiceless stop. These words were presented over headphones, and 'Word' versus 'Non-word' judgments were made using the computer keyboard. This phase was

included in order to increase exposure to the manipulated stimuli under conditions that demand some degree of listener attention; because it was not designed to generate usable data, we do not report any lexical decision results. During the lexical decision task, the male and female voices alternated and a picture of either a young man or young woman appeared on the screen for the corresponding talker, as shown in Figure 1. The latency between the presentation of the image of the talker and the corresponding sound clip was .25 seconds. Participants then completed the shadowing task. The same 84 stimuli were presented auditorily over headphones and participants repeated each word aloud. Following Shockley et al., participants were not asked to imitate the talker they heard but were simply instructed to "Identify the word you hear by speaking it into the microphone quickly but clearly" (2004, 424). Just as with the lexical decision task, male and female voices alternated and were accompanied by the corresponding picture.

Words were elicited in the same order for all participants so that performance on the task would be strictly comparable across individuals. The order that was held constant across participants was pseudo-randomized across phases. Stimuli were divided into four blocks, balanced for talker, initial consonant, and lexical frequency, using the Subtlex corpus LOG10 CD measure. Blocks were presented in the same order in the read-aloud, lexical decision, and shadowing phases. Words always occurred in the same block, though their order within the block varied across the three phases. This was done so that items elicited early on in the read-aloud phase would also be elicited early on in the shadowing phase, which ensures that the distance between a lexical item being read aloud and that same lexical item produced in shadowing is roughly constant.

Recordings were forced-aligned using the Penn Phonetics Lab Forced Aligner (Yuan & Liberman, 2008), which takes an audio recording and a transcript and automatically identifies the beginning and end points in the audio for each corresponding word in the transcript. The output of forced alignment is a Praat TextGrid with a word tier containing the identified word boundaries, which was checked by hand for accuracy. Next, the AutoVOT script from Keshet et al. (2014) was used to automatically identify the VOT portion of each word-initial voiceless stop and add it to a new tier in the TextGrid. These measurements were also checked and hand-corrected as necessary. For hand-adjustments, starting boundaries were placed at the beginning of the release burst and ending boundaries were placed at the onset of following vowel voicing, indicated by regular pulses in the waveform.

The VOT measurements and word durations were then extracted from the TextGrid using a Praat script. Items that participants skipped or mispronounced were excluded. Outliers 2.5 standard deviations beyond the mean were omitted on a by-participant and by-word basis. By-participant means and standard deviations were calculated over all tokens and any outliers 2.5 standard deviations beyond the by-participant means were marked, then the same procedure was done on a by-word basis. All tokens marked as outliers (i.e., tokens that were 2.5 standard deviations beyond by-participant means, by-word means, or both) were omitted from the analysis. If a word was excluded from one phase (either as an outlier or due to mispronunciation or participant omission), we excluded that participant's other token of the same word from the opposite phase in our analyses. In total, 568 tokens were omitted (7%), leaving 7,160 tokens for analysis. For the remaining tokens, participants' baseline VOT during the read-aloud phase was compared to post-exposure VOT during the shadowing phase.

4 Results

A mixed effects linear regression model was fit to the phonetic imitation data using the lmerTest package (Kuznetsova et al., 2017) in R (R Core Team, 2015). VOT is the dependent variable, and

is log-transformed for the model for a more normal distribution of model residuals. The fixed predictors are as follows:

- Condition: categorical predictor with levels baseline (reading) or post-exposure (shadowing), treatment coded with baseline as reference level.
- Gender: participant gender, categorical predictor with levels male and female, sum coded.
- Talker: model talker label, categorical predictor with levels male and female, sum coded.
- Visit: categorical predictor with levels Visit 1 and Visit 2, sum coded.
- RestDur: duration of the rest of the word (minus VOT), continuous predictor, log-transformed. This predictor is included to account for global shifts in word duration (due to factors such as shifts in speech rate) that may influence VOT production.
- Phoneme: categorical predictor with levels /p/, /t/, or /k/, treatment coded with /t/ as the reference level.
- Trial: order within the condition that the word appeared.
- Frequency: lexical frequency, continuous predictor, *z*-scored using the *scale* function in R.
- ModelVOT: continuous predictor of model talker VOT of each item, *z*-scored within each model talker to avoid multicollinearity with the Talker predictor.

The model also includes several interactions. A two-way interaction between Condition and Gender was included because previous studies have found gender to be a relevant predictor of convergence, though the direction of this influence has been somewhat inconsistent. For instance, some studies have found that female participants converge more than males (Babel et al., 2014b; Miller et al., 2010; Namy et al., 2002; Pardo et al., 2016), while others have found that males converge more than females (Pardo, 2006; Pardo et al., 2010). Others still have observed no gender differences in convergence rates (Lewandowski & Jilka, 2019; Pardo et al., 2013; Yu et al., 2013). A two-way interaction between Condition and ModelVOT (*z*-scored within talker) was also included to determine whether participants target individual token values when converging. Finally, a three-way interaction between Condition, Talker, and Visit was included to assess our main question of whether participants converge similarly on two separate occasions. This interaction also allows us to ask whether participants consistently converge toward the two model talkers and whether this differs across visits. Random effects were included based on likelihood ratio test significance. The Condition*Visit random slope was tested first, as motivated by our main research question of whether participants shift across conditions similarly on two different visits. Then all fixed predictors and interactions that were significant in the model were added to the model as by-participant random slopes, then by-word random slopes (where appropriate). The *rand* function in the lmerTest (Kuznetsova et al., 2017) package in R was used to perform likelihood ratio tests on all random effects. Those that significantly improved the model ($p < 0.05$) were included in the final model. None of the by-word random slopes significantly improved the model, so only random intercepts are included for Word. Random slopes were included for Condition*Visit+Phone+RestDur by Participant. We include the full model in Table 1.

In the aggregate, participants produce words with slightly longer VOTs after exposure to the model talkers' artificially doubled VOT. During Visit 1, average baseline VOT is 76.4 ms, which increases to 80.5 ms after exposure. For Visit 2, average baseline VOT is 74.8 ms, which increases to 80 ms after exposure. Though these increases are small in absolute terms, the positive main effect of Condition, indicating the shift from baseline to shadowing, is significant in the model ($Est. = 0.063$, $p < 0.05$). Also note that RestDur, referring to the duration of the rest of the word minus VOT, does not reach significance in the model ($Est. = 0.062$, $p = 0.064$), and even with this predictor included in the model, we still see a main effect of Condition. This suggests that the

Table 1. Linear mixed effects regression model predicting usage of log-transformed raw VOT.

Scaled residuals	Min	1Q	Median	3Q	Max			
	−4.771	−0.6	0.028	0.652	3.837			
Random effects:								
Groups	Name	Variance	Std.Dev.	Corr				
word	(Intercept)	0.003	0.055					
participant	(Intercept)	0.55	0.742					
	Shadowing	0.021	0.145	−0.09				
	Visit2	0.016	0.128	−0.37	0.52			
	PhoneP	0.002	0.05	−0.2	−0.23	−0.08		
	PhoneT	0.003	0.055	−0.11	−0.5	0.04	0.64	
	RestDur	0.015	0.122	−0.96	−0.09	0.26	0.24	0.16
	Shad:Vis2	0.013	0.113	0.23	−0.59	−0.92	0.28	0.06 −0.14
	Residual	0.034	0.185					

Fixed effects	Estimate	Std.Error	df	*t*-value	*p*-value	
(Intercept)	3.934	0.193	38.56	20.385	<2e-16	***
ConditionShadowing	0.063	0.026	22.67	2.474	0.021	*
TalkerFemale	−0.008	0.007	93.81	−1.129	0.262	
GenderFemale	0.001	0.041	22.64	0.015	0.988	
PhoneK	0.033	0.020	74.66	1.681	0.097	
PhoneP	−0.122	0.019	81.59	−6.481	6.5e-09	***
log(RestDur)	0.062	0.033	37.8	1.911	0.064	
scale(Freq)	−0.006	0.007	75.57	−0.84	0.403	
Order	3.3e-04	2.5e-04	113	1.326	0.188	
scale(ModelVOT)	0.025	0.007	92.89	3.454	0.001	***
Visit1	0.007	0.014	23	0.534	0.598	
ConditionShad:TalkerF	0.01	0.004	6925	2.201	0.028	*
ConditionShad:GenderF	0.025	0.02	22.21	1.304	0.205	
ConditionShad:scale(ModVOT)	0.005	0.004	6952	1.221	0.222	
ConditionShad:Visit1	−0.004	0.013	22.82	−0.332	0.743	
TalkerF:Visit1	−0.003	0.003	6922	−0.811	0.417	
ConditionShad:TalkerF:Visit1	0.001	0.004	6919	0.142	0.887	

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

shift in VOT from baseline to shadowing cannot be accounted for solely by global shifts in overall word duration, which may simply reflect speech rate. The model also shows a main effect of Phoneme: /p/ is produced with a significantly shorter VOT than /t/ (*Est.* = −0.122, $p < 0.001$), and /k/ is produced with a slightly longer VOT than /t/ (but the /t/-/k/ difference is not significant at the 0.05 level). This is the expected result for the effect of place of articulation on VOT (Lisker & Abramson, 1967). There is also a significant main effect of ModelVOT for each item, *z*-scored within each of the two model talkers (*Est.* = 0.025, $p < 0.001$). Since the female talker produced generally longer VOTs than the male talker, *z*-scoring *within* talker allows us to isolate the influence of individual token length from model talker average VOT (which is captured by the Talker predictor). However, a lack of significant interaction between Condition and ModelVOT (*Est.* = 0.005, $p = 0.222$) suggests that participants do not *shift* more toward words heard with longer VOTs on an item-by-item basis. Rather, the effect of ModelVOT seems to

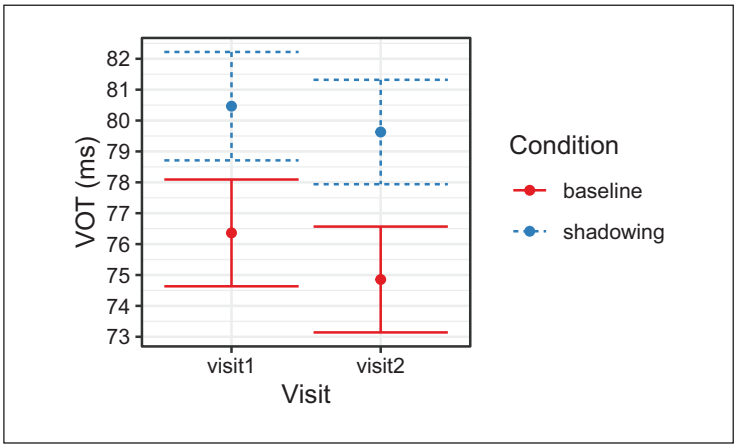

Figure 2. Mean VOT by visit with 95% confidence intervals based on by-word means.

capture a general tendency for some words to have longer VOTs than others, which independently influences participants' and the model talkers' productions of these words. The model does not, therefore, offer evidence that participants target precise VOT values for individual tokens in convergence. Finally, there is a significant interaction between Talker and Condition, such that participants produce greater VOT shifts for words spoken by the female talker ($Est. = 0.001$, $p < 0.05$). This effect is likely because the female model talker produced naturally longer VOT (mean of 110.8 ms) than the male talker (mean of 76.7 ms). Longer VOT may promote a greater degree of convergence for several reasons. For one, tokens with relatively longer VOT are more likely to be outside of participants' normal range of VOT and thus would require an observable production shift in order for convergence to occur. However, for tokens with relatively shorter VOT (such as some of those from the male model talker), more of these are likely to fall within participants' normal range of VOT production, so convergence toward (or matching of) the model talker's productions could occur with little to no observable production shift. Further, lengthened VOT may be more novel and perceptually salient to listeners, which has been shown to facilitate convergence (e.g., Babel et al., 2014b; Walker & Campbell-Kibler, 2015). However, the effect may also encompass differing attitudes toward the model talkers and therefore differing willingness to converge to each voice. Despite its general interest for theories of convergence, the present study's design does not allow for teasing apart these possibilities.

Figure 2 displays mean VOT with confidence intervals based on by-word means for both visits. The degree of VOT shift is similar across visits. Indeed, we find no significant interaction between Condition and Visit ($Est. = -0.004$, $p = 0.743$). Also note that the shift toward longer VOT is not as long lasting as previous studies have found (Goldinger, 1998; Goldinger & Azuma, 2004); rather, participants appear revert back to their baseline VOT production in the interim between visit, as the model, which treats treatment-coded "baseline" speech as the reference level, shows no main effect of Visit for just the baseline measurements ($Est. = 0.007$, $p = 0.598$).

Figure 3 (left) plots the correlation in participants' percentage VOT shift from baseline between Visit 1 and Visit 2. We calculated percentage VOT shift as each participant's mean baseline VOT subtracted from their mean shadowed VOT, divided by the mean baseline VOT. Despite considerable individual variation, phonetic imitation in both measures appears to be quite stable across visits. There is a strong correlation in VOT shift (Pearson's $R = 0.681$, $p < 0.001$) across visits. Not only is overall imitation stable across visits, but it also appears to be

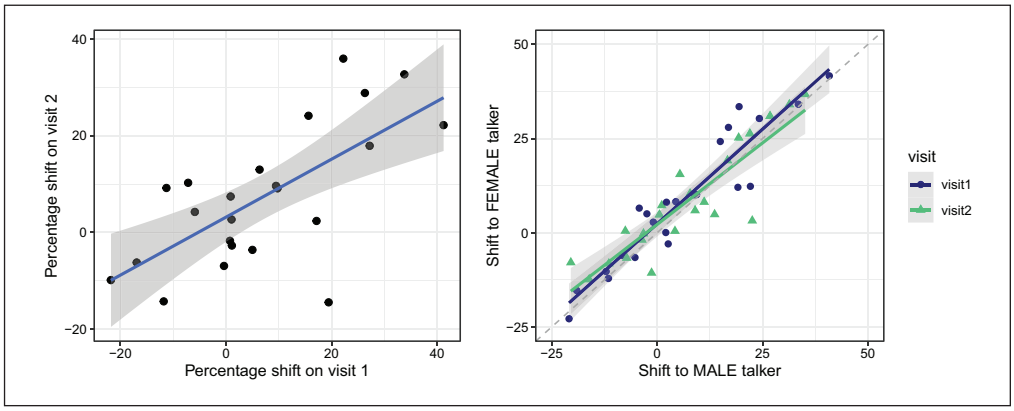

Figure 3. Percentage VOT shift from baseline by visit (left) and to each model talker (right).

stable across talkers. Figure 3 (right) shows a strong correlation between convergence to the male talker and convergence toward the female talker (Pearson's $R = 0.917$, $p < 0.001$), suggesting either that participants tend to converge toward various talkers equally (i.e., a high converger toward Talker A will be a high converger toward Talker B), or perhaps that participants exhibited a general shift toward lengthened VOT for the duration of the shadowing task that did not target individual talkers or items. Still, despite this high correlation, participants did converge somewhat more to the female talker in the aggregate, suggesting *some* tracking of talker-specific VOT.

While the aggregate shifts in VOT toward the model talkers are small (but statistically significant), they stem from considerable inter-speaker variation: many speakers show much greater increases in VOT after exposure, some speakers diverge from the model talkers and *decrease* their VOT production, and other speakers show little change between baseline and shadowing. Participants' change in VOT from baseline ranges from +41% to −20%, with the mean change being +7.1%. While convergence rates are greater, and the aggregate tendency is toward convergence, some participants do exhibit considerable divergence as well.

Figure 4 plots the percentage shift in VOT from the baseline to the shadowing condition for each individual participant, broken down by model talker. We can group individuals into categories based on their qualitative patterns. Some are convergers, others are divergers, and some show little to no shift (i.e., maintainers). Importantly, these patters are relatively stable across visits and model talkers. Participants 2, 9, 14, 19, 23, 25, 26, and 31 show relatively large shifts toward lengthened VOT, and these shifts are similar across talkers and visits. Others, such as participants 20, 22, and 27 diverge consistently across visits. While several others appear not to shift at all, with confidence intervals mostly overlapping with zero (e.g., 3, 6, 7, 29, 30, and 32), only two participants perform clearly differently on different occasions: Participant 13 diverges toward both talkers on Visit 1, then converges on Visit 2, while Participant 21 exhibits significant convergence toward both talkers during Visit 1, then diverges during Visit 2. For the most part, though, participants behave similarly across visits. If individual differences reflected merely noise, we would not expect to see such high correlations between Visit 1 and Visit 2.

To summarize, we found that the small but significant VOT imitation effect is the product of a mix of individual behaviors, including some participants who imitate to a much larger degree and some who diverge by shortening their VOT. Participants perform very similarly on the imitation task on different days.

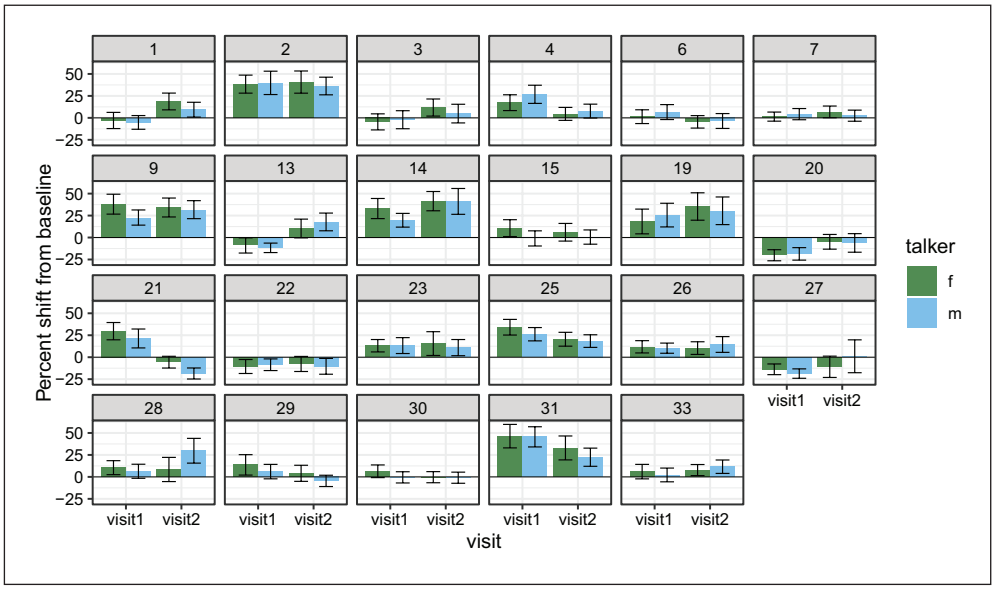

Figure 4. Participants' average percentage VOT shift from baseline, by visit and model talker. Facet labels are participant ID numbers.

5 Discussion

This experiment evaluated the reliability of individuals' imitation behavior on a single task repeated across two different occasions. The results show that individuals tend to produce a similar degree of imitation toward lengthened VOT during Visit 1 and Visit 2. Because the task was identical on both visits, this provides a usefully-narrow view of whether observed individual differences in imitation are actually characteristic of individuals. In studies that investigate whether imitation performance is correlated across different features, measures, model talkers, or task types, null results could indicate either that individual differences are merely chance fluctuations, or that the different stimuli or tasks modulated the individual responses. Here we learn that given the same input and task, participants will respond in a similar way (at least for VOT imitation). That stability does not tell us whether some individuals can be classified as "imitators" in general: we do not know if the participants who strongly imitated these particular talkers' lengthened VOT would also strongly imitate other phonetic features or other talkers, or if imitation varies based on various factors such as the feature's salience or an individual's personal experience with or sensitivity to a particular variable. In fact, there is some evidence that individuals do vary in their sensitivity to different phonetic features (Sanker, 2015; Schertz & Clare, 2019). However, having greater confidence in the reliability of any particular task does suggest that a lack of cross-feature or cross-task correlations in other studies reflects genuine differences in the individual differences utilized by different features or tasks.

The stable individual differences we observe are not merely quantitative differences in the magnitude of imitation: we observe that some individuals *diverge* from the model talker in their VOT production, and those who diverge on Visit 1 tend do so again on Visit 2. This might reflect either socially mediated divergence or divergence induced by the artificially

Male Talker					Female Talker				
word	**phone**	**frequency**	**talker**	**block**	**word**	**phone**	**frequency**	**talker**	**block**
collar	k	2.5944	m	1	cannon	k	2.4082	f	1
keyboard	k	1.8261	m	1	camel	k	2.143	f	1
kernel	k	1.2041	m	1	kennel	k	1.6628	f	1
caution	k	2.3541	m	1	kidney	k	2.3766	f	2
coffin	k	2.4082	m	2	kingdom	k	2.6474	f	2
cabbage	k	2.0414	m	2	cavern	k	1.5185	f	2
callous	k	1.699	m	2	canine	k	1.8573	f	2
comma	k	1.5441	m	3	cashmere	k	1.699	f	3
cable	k	2.7292	m	3	curly	k	2.1761	f	3
carriage	k	2.3838	m	3	colleague	k	2.4425	f	3
caddy	k	1.9085	m	3	cordial	k	1.6128	f	4
cobra	k	1.7924	m	4	cushion	k	1.9956	f	4
castle	k	2.5832	m	4	cabin	k	2.6628	f	4
canyon	k	2.3464	m	4	cargo	k	2.3784	f	4
avg = **2.1010643**					avg = **2.1129286**				

word	**phone**	**freq**	**talker**	**block**	**word**	**phone**	**freq**	**talker**	**block**
timber	t	1.8808	m	1	torso	t	1.8865	f	1
tailor	t	2.1584	m	1	tardy	t	1.6812	f	1
tangle	t	1.6721	m	1	tonsil	t	1.1761	f	1
terrace	t	2.0828	m	2	tissue	t	2.5888	f	1
tunnel	t	2.6493	m	2	timid	t	1.7993	f	2
turbo	t	1.7782	m	2	tender	t	2.5809	f	2
tinsel	t	1.2788	m	2	tendon	t	1.415	f	2
tenure	t	1.716	m	3	tango	t	2.0719	f	3
tennis	t	2.6064	m	3	tandem	t	1.2553	f	3
tavern	t	1.9294	m	3	tiger	t	2.6454	f	3
taffy	t	1.5798	m	4	tally	t	1.7782	f	3
tofu	t	1.7993	m	4	tuba	t	1.5185	f	4
towel	t	2.721	m	4	tumble	t	1.8129	f	4
tidy	t	2.2068	m	4	terror	t	2.5051	f	4
avg = **2.0042214**					avg = **1.9082214**				

Figure 5. *(Continued)*

word	phone	freq	talker	block	word	phone	freq	talker	block
pebble	p	1.699	m	1	pony	p	2.4183	f	1
poodle	p	1.9912	m	1	python	p	1.6232	f	1
pencil	p	2.5933	m	1	paddle	p	2.0334	f	1
puzzle	p	2.3962	m	2	penny	p	2.7767	f	1
parsley	p	1.5315	m	2	pigeon	p	2.2601	f	2
pedal	p	1.9345	m	2	parcel	p	1.7559	f	2
powder	p	2.7185	m	2	palace	p	2.6425	f	2
pilgrim	p	1.8865	m	3	publish	p	2.2304	f	3
punish	p	2.6042	m	3	panther	p	1.6721	f	3
panel	p	2.4048	m	3	perfume	p	2.5933	f	3
pollen	p	1.6435	m	4	poison	p	2.8432	f	3
passion	p	2.8603	m	4	partial	p	2.3181	f	4
purchase	p	2.4298	m	4	pillow	p	2.6571	f	4
puddle	p	1.9542	m	4	pelvis	p	1.8865	f	4
avg = **2.1891071**					avg = **2.2650571**				

Figure 5. List of elicited words.

manipulated stimuli of the tokens. Regardless, this stability sheds light on the central role that divergence plays in understanding the cognitive mechanisms of imitation. While imitation is sometimes attributed to the automatic effects of a mechanical perception-production feedback loop akin to priming (e.g., Goldinger, 1998; Pickering & Garrod, 2004), such theories do not as easily account for the possibility of divergence as an outcome. Demonstrating that divergence can be a consistent and reliable behavior and not a sporadic finding highlights the need for theories of phonetic imitation to account for such behavior. An alternative explanation may be that participants did not diverge at all, but that apparent divergers started to hypo-articulate in the second phase as they spent more time or became more familiar with the task. These possibilities cannot be ruled out with our current data.

We also find that the imitation effect observed in our test-retest experiment does not last until a second visit, and the predictor of Visit was not significant in the model. When participants return one to two weeks later, they have reset their baseline VOT production. While this result contradicts previous findings (Goldinger, 1998; Goldinger & Azuma, 2004), it is unsurprising: a learned *perceptual* adjustment to a particular talker's voice is limited in scope and could in principle last indefinitely without otherwise impacting the participant's linguistic behavior, but it would be surprising indeed if we managed to shift a participant's own speech production behavior so strongly that they were still talking differently a week later.

Additionally, we find more convergence to the female model talker; this is perhaps because the female talker's original VOT (and artificially doubled VOT) was longer than the male talker's. The female talker's longer VOT may promote convergence either because longer VOT is more more perceptually salient, or because it is less likely to be within participants' normal baseline productions, requiring greater production shifts for convergence to occur. However, we find no significant interaction between Condition and ModelVOT when *z*-scored within talker, suggesting that

Lexical Decision Task Non-words

word	talker	block	word	talker	block
ralace	f	1	lobee	m	3
winsa	f	1	larso	f	3
larno	m	1	shasoo	m	3
jumid	f	1	embolf	f	3
grushle	f	1	irzo	f	3
diffone	m	1	yenoo	f	3
gerswo	m	1	isaff	m	3
reenion	m	1	yebro	m	3
ragive	f	1	erfess	m	3
wooneal	m	1	sahoss	f	3
erwume	f	2	flobine	f	4
hinralf	m	2	frebbo	m	4
adlear	f	2	umboo	m	4
memso	f	2	nerror	f	4
amnoy	m	2	jesmey	m	4
thywin	m	2	javid	f	4
finone	m	2	ulerd	f	4
harfeen	f	2	chirash	m	4
silor	f	2	arrall	f	4
mimmon	m	2	ogry	m	4

Figure 6. List of elicited nonwords.

participants are not tracking and targeting item-specific VOT during imitation. The slight increase in convergence for words spoken by the female talker may also stem from tracking talker-specific VOT (i.e., recognizing that the female talker produces longer VOT on average without targeting individual tokens) or it may indicate a social preference for the female talker over the male.

Finally, the strong correlation between shifts toward the male talker and shifts toward the female talker may suggest that convergence patterns are stable across talkers; that is, high imitators for one talker are likely to be high imitators for other talkers. However, it may instead indicate a global shift in lengthened VOT production that lasts for the duration of the shadowing phase and is only minimally influenced by talker-specific values. Still, the greater convergence toward the female talker indicates *some* degree of differentiation between the two talkers in convergence. We do not,

however, find evidence for an average difference between male and female participants in how much they imitate, which is consistent with recent work from Pardo et al. (2016) but different from earlier results suggesting that women may be more prone to imitation than men (Babel et al., 2014a; Namy et al., 2002).

These results have provided evidence that individual differences in phonetic imitation are real and reliable rather than merely noise or chance fluctuations. Participants not only showed similar quantitative shifts across visits, but they were also consistent in the type of shifts exhibited (convergence, divergence, or maintenance). Findings further contribute to a growing body of literature suggesting that averaging over groups of participants masks the complexity of phonetic behaviors, such as imitation. Future research examining reliability of individual differences across different phonetic features, tasks, and model talkers will shed light on the role individual differences may play in explaining sound change or theories of the mental representations of phonetic and phonological information. Taken together, these findings suggest that individual differences in phonetic behavior are an area of promising future study.

Acknowledgements

Thanks to Elisha Cooper for her contributions in experimental design and implementation and to our many wonderful undergraduate research assistants. The ideas in this paper have been improved by feedback from audiences at Ludwig Maximilian University of Munich, UC Davis, Stanford, CUNY Graduate Center, and the LSA annual meeting.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This material is based on work supported by the National Science Foundation under Grant Number 1627972, "Cognitive characteristics of the leaders of language change."

ORCID iD

Lacey Wade https://orcid.org/0000-0002-9382-7191

References

Abrego-Collier, C., Grove, J., Sonderegger, M., & Yu, A. C. (2011). Effects of speaker evaluation on phonetic convergence. In *Proceedings of the 17th International Congress of the Phonetic Sciences*. International Phonetics Association.

Aguilar, L., Downey, G., Krauss, R., Pardo, J., Lane, S., & Bolger, N. (2016). A dyadic perspective on speech accommodation and social connection: Both partners' rejection sensitivity matters. *Journal of Personality*, *84*(2), 165–177.

Audacity Team (2020). *Audacity(R): Free Audio Editor and Recorder* [Computer application]. Version 2.4.2. https://audacityteam.org/

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, *40*, 177–189.

Babel, M. E. (2009). *Phonetic and social selectivity in speech accommodation*. University of California, Berkeley.

Babel, M., & Bulatov, D. (2011). The role of phonetic frequency in imitation. *Language and Speech*, *55*(2), 231–248.

Babel, M., McGuire, G., & Russell, J. (2014a). Whose sound changes do we follow? Selective attention to ingroups as a mechanism for sound change. *Paper presented at Sound Change in Interacting Human Systems*, May 29.

Babel, M., McGuire, G., Walters, S., & Nicholls, A. (2014b). Novelty and social preference in phonetic accommodation. *Laboratory Phonology*, *5*(1), 123–150.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males, females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, *31*, 5–17.

Boersma, P., & Weenink, D. (2018). *Praat: Doing phonetics by computer* [Computer program]. Version 6.0.56. http://www.praat.org/

Brysbaert, M., & New, B. (2009). Moving beyond Kuera 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.

Dimov, S., Katseff, S., & Johnson, K. (2012). Social and personality variables in compensation for altered auditory feedback. In M. Josep-Solé & D. Recasens (Eds.), *The initiation of sound change: Perception, production, and social factors* (pp. 185–210). John Benjamins.

Gijssels, T., Casasanto, L. S., Jasmin, K., Hagoort, P., & Casasanto, D. (2016). Speech accommodation without priming: The case of pitch. *Discourse Processes*, *53*(4), 233–251.

Giles, H., Coupland, N., & Coupland, I. (1991). Accommodation theory: Communication, context, and consequence. In H. Giles, J. Coupland, & N. Coupland (Eds.), *Contexts of Accommodation: Developments in Applied Sociolinguistics* (pp. 1–68). Cambridge University Press.

Goldberg, L. (1992). The development of markers for the Big-Five factor structure. *Personality Assessment*, *4*(1), 26–42.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

Goldinger, S. D., & Azuma, T. (2004). Episodic memory reflected in printed word naming. *Psychonomic Bulletin and Review*, *11*(4), 716–722.

Keshet, J., Sonderegger, M., & Knowles, T. (2014). AutoVOT: A tool for automatic measurement of voice onset time using discriminative structured prediction [computer program].

Kingston, J., Rich, S., Shen, A., & Sered, S. (2015). Is perception personal? In *Proceedings of the 18th International Congress of the Phonetic Sciences*. International Phonetics Association.

Kong, E. J., & Edwards, J. (2016). Individual differences in categorical perception of speech: Cue weighting and executive function. *Journal of Phonetics*, *59*, 40–57.

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.

Lai, W., & Tamminga, M. (2019). Individual differences in simultaneous compensation for coarticulatory and lexical cues. Presented at the 5th Workshop on Sound Change, University of California, Davis, CA.

Lev-Ari, S., & Peperkamp, S. (2014). The influence of inhibitory skill on phonological representations in production and perception. *Journal of Phonetics*, *47*, 36–46.

Lewandowski, N., & Jilka, M. (2019). Phonetic convergence, language talent, personality, and attention. *Frontiers in Communication*, *4*(18), 1–19.

Lisker, L., & Abramson, A. S. (1967). Some effects of context on voice onset time in English stops. *Language and Speech*, *10*(1), 1–28.

Miller, R. M., Sanchez, K., & Rosenblum, L. D. (2010). Alignment to visual speech information. *Attention, Perception and Psychophysics*, *72*(6), 1614–25.

Morley, R. L. (2014). (non-) phonologization: Individual variation in an artificial grammar learning task. *Paper presented at Sound Change in Interacting Human Systems*, *Berkeley*, May 31.

Namy, L., Nygaard, L., & Sauerteig, D. (2002). Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology*, *21*(4), 422–432.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*(2), 132–142.

Pardo, J., Gibbons, R., Suppes, A., & Krauss, R. (2012). Phonetic convergence in college roommates. *Journal of Phonetics*, *40*(1), 190–197.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*, *119*(4), 2382–2393.

Pardo, J. S., Jay, I. C., & Krauss, R. M. (2010). Conversational role influences speech imitation. *Attention, Perception, & Psychophysics*, *72*(8), 2254–2264.

Pardo, J. S., Jordan, K., Mallari, R., Scanlon, C., & Lewandowski, E. (2013). Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *Journal of Memory and Language*, *69*, 183–195.

Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2016). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, Psychophysics*, *79*(2), 637–659.

Pardo, J. S., Urmanche, A., Wilman, S., Wiener, J., Mason, N., Francis, K., & Ward, M. (2018). A comparison of phonetic convergence in conversational interaction and speech shadowing. *Journal of Phonetics*, *69*, 1–11.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203.

Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *Journal of the Acoustical Society of America*, *130*(1), 461–472.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*, 169–226.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641.

Sanker, C. (2015). Comparison of phonetic convergence in multiple measures. *Cornell Working Papers in Phonetics and Phonology*, *2015*, 60–75.

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of Phonetics*, *52*, 183–204.

Schertz, J., & Clare, E. J. (2019). Phonetic cue weighting in perception and production. *Wiley Interdisciplinary Reviews: Cognitive Science*, 11(2), e1521.

Schweitzer, A., & Lewandowski, N. (2013). Convergence of articulation rate in spontaneous speech. In *Proceedings of Interspeech 2013* (pp. 525–529). International Speech Communication Association.

Shockley, K., Sabadini, L., & Fowler, C. (2004). Imitation in shadowing words. *Attention, Perception, and Psychophysics*, *66*(3), 422–429.

Shultz, A., Francis, A., & Llanos, F. (2012). Differential cue weighting in perception and production of consonant voicing. *The Journal of the Acoustical Society of America*, *132*, 95–101.

Sonderegger, M., Bane, M., & Graff, P. (2017). The medium-term dynamics of accents on reality television. *Language*, *93*(3), 598–640.

Staum Casasanto, L., Jasmin, K., & Casasanto, D. (2010). Virtually accommodating: Speech rate accommodation to a virtual interlocutor. In *32nd Annual Meeting of the Cognitive Science Society (CogSci 2010)* (pp. 127–132). Cognitive Science Society.

Stewart, M. E., & Ota, M. (2008). Lexical effects on speech perception in individuals with "autistic" traits. *Cognition*, *109*, 157–162.

Tamminga, M., Wade, L., & Lai, W. (2018, January 6). Stability and variability in phonetic flexibility. Presented at the LSA Annual Meeting, Salt Lake City, UT, USA.

Tilsen, S., & Cohn, A. (2016). Shared representations underlie metaphonological judgments and speech motor control. *Laboratory Phonology*, *7*, 14.

Turnbull, R. (2015). Patterns of individual differences in reduction: Implications for listener-oriented theories. In *Proceedings of the 18th International Congress of the Phonetic Sciences*. *International Phonetics Association*.

Van Hedger, S. C., Heald, S. L., Koch, R., & Nusbaum, H. C. (2015). Auditory working memory predicts individual differences in absolute pitch learning. *Cognition*, *140*, 95–110.

Walker, A., & Campbell-Kibler, K. (2015). Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task. *Frontiers in Psychology*, *6*.

Yu, A. C., & Zellou, G. (2019). Individual differences in language processing: Phonology. *Annual Review of Linguistics*, *5*(1), 131–150.

Yu, A. C. L. (2010). Perceptual compensation is correlated with individuals' "autistic" traits: Implications for models of sound change. *PLoS One*, *5*(8), e11950

Yu, A. C. L., Abrego-Collier, C., & Sonderegger, M. (2013). Phonetic imitation from an individual difference perspective: Subjective attitude, personality and "autistic" traits. *PLoS ONE*, *8*(9), e74746.

Yu, A. C. L., & Lee, H. (2014). The stability of perceptual compensation for coarticulation within and across individuals: A cross-validation study. *The Journal of the Acoustical Society of America*, *136*(1), 382–388.

Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics 2008* (pp. 5687–5690).

Zellou, G. (2017). Individual differences in the production of nasal coarticulation and perceptual compensation. *Journal of Phonetics*, *61*, 13–29.

Zellou, G., Scarborough, R., & Nielsen, K. (2016). Phonetic imitation of coarticulatory vowel nasalization. *The Journal of the Acoustical Society of America*, *140*(5), 3560–3575.