Laboratory Phonology
Journal of the Association for Laboratory Phonology

Clopper, C. G., Dossey, E., & Gonzalez, R. (2024). Raw acoustic vs. normalized phonetic convergence: Imitation of the Northern Cities Shift in the American Midwest. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 15(1), pp. 1–34. DOI: https://doi.org/10.16995/labphon.10893

Open Library of Humanities

Raw acoustic vs. normalized phonetic convergence: Imitation of the Northern Cities Shift in the American Midwest

Cynthia G. Clopper*, Department of Linguistics, Ohio State University, Columbus, OH, USA, clopper.1@osu.edu

Ellen Dossey, Department of Linguistics, Ohio State University, Columbus, OH, USA, dossey.1@osu.edu

Roberto Gonzalez, Department of Linguistics, Ohio State University, Columbus, OH, USA, gonzalez.594@buckeyemail.osu.edu

*Corresponding author.

Word shadowing tasks elicit phonetic convergence to the stimulus model talkers, suggesting a tight perception-production link. The magnitude of this convergence is affected by linguistic and social factors, suggesting that the perception-production link is mediated by higher-level phonological and social structures. The current study explored the nature of the perception-production link in an explicit comparison of raw acoustic vs. normalized phonetic convergence in word shadowing. American Midwestern participants repeated words after a model talker with features of the Northern Cities Shift vowels in one of three instruction conditions, which varied in whether participants were primed with the regional background of the model talker and in whether they were asked explicitly to imitate her. The results revealed normalized phonetic convergence to the model talker's Northern Cities Shift vowels, even when this convergence entailed divergence from the raw acoustics, and token-by-token variability in her productions, consistent with a tight perception-production link that is mediated by linguistic structure. Modest effects of instruction condition on the magnitude of phonetic convergence were also observed, consistent with social information mediating this perception-production link. The results of this study provide converging evidence for phonetic convergence that is both phonetically-detailed and subject to constraint by higher-level representations.

1. Introduction

Phonetic convergence, in which talkers imitate features of their interlocutors' speech, has been taken as evidence of a tight connection between speech perception and production processes (Goldinger, 1998; Shockley et al., 2004). Goldinger's episodic model of phonetic convergence has typically been interpreted as involving shared perception-production exemplars in an acoustic space, whereas Shockley et al. proposed a shared articulatory (i.e., gestural) perception-production space, consistent with motor theory and direct realism models of speech perception (Fowler, 1996; Liberman & Mattingly, 1985). However, Goldinger explicitly acknowledged the possibility of an exemplar space involving gestures (vs. acoustic features), and Shockley et al. noted that the manipulation of voice-onset-time (VOT) in their shadowing task made it difficult to distinguish an acoustic from a gestural account. More recent work has further suggested a role for phonological structure (e.g., Mitterer & Ernestus, 2008; Nielsen, 2011) and social salience (e.g., Babel, 2010; Clopper & Dossey, 2020; Wade, 2022) in phonetic convergence. The goal of the current study was to further explore the linguistic structures that are imitated in a word shadowing task, with a particular focus on raw acoustic vs. normalized phonetic information. In particular, we examined convergence to a model talker with relatively high formant frequencies overall, consistent with a relatively short vocal tract, who produced vowel features consistent with the Northern Cities Shift in American English. This combination of high overall formants with the Northern Cities Shift pattern allowed us to explicitly compare convergence to the model talker's high formant values (i.e., raw acoustics) vs. the features of the Northern Cities Shift (i.e., normalized phonetics). Our results suggest a perception-production link that is both phonetically-detailed in acoustic or articulatory space and mediated by normalized phonetic representations.

1.1. What is imitated in a word shadowing task?

Phonetic convergence in word shadowing tasks is often assessed with acoustic measures (e.g., Babel, 2010; Nielsen, 2011; Pardo et al., 2013), which may leave the impression that what is imitated is raw acoustic features. Likewise, when perceptual measures of phonetic convergence are used instead of, or in addition to, acoustic measures, they are typically interpreted as reflecting the properties of the acoustic (or gestural) signal (Goldinger, 1998; Pardo et al., 2013, 2017; Shockley et al., 2004). However, the participants in a shadowing task do not produce utterances that acoustically match those of the model talker (Babel & Bulotov, 2012). Indeed, the characterization of the imitation observed in these tasks as phonetic convergence (vs. acoustic convergence) reflects an assumed mapping between the perception of the model talkers' utterances and the production of the shadowers' utterances. That is, shadowers are assumed to map the model talker's utterances onto their own production space, resulting in convergence to normalized phonetic patterns instead of raw acoustic values.

This assumption about the target of convergence as normalized phonetic information is implicit in much of the literature (Cohen Priva & Sanker, 2019). Indeed, phonetic convergence is variably defined as being about acoustic characteristics (e.g., Babel, 2012), acoustic-phonetic repertoire (Pardo et al., 2013), and speech features (Clopper & Dossey, 2020), leaving some ambiguity about the assumed target of convergence. In the temporal domain, the target of convergence could potentially be raw acoustic features. For example, the convergence observed on VOT (Nielsen, 2011; Shockley et al., 2004), word or vowel duration (Clopper & Dossey, 2020; Lewandowski & Nygaard, 2018; Pardo et al., 2017), and speaking rate (Gonzalez & Clopper, 2022; Jungers & Hupp, 2009) in shadowing tasks could be characterized as shadowers' directly imitating the model talkers. However, in the spectral domain, the target of convergence seems more likely to be either gestural or normalized phonetic features, such that the acoustic properties of the model talker are mapped onto the acoustic production space of the shadower (Shockley et al., 2004). Thus, for example, convergence on vowel spectra is typically assessed using normalized acoustic measures (e.g., Babel, 2010, 2012; Clopper & Dossey, 2020; Lewandowski & Nygaard, 2018; Pardo et al., 2017), especially with mixed-gender talker pairs, because the assumption is not that the shadowers are converging to the acoustic details of the voice of the model talker, but to the linguistic features of the model talker's speech. That is, for example, a female shadower may not converge to the lower overall F1 of a male model talker, but rather map the male model talker's F1 range onto her own F1 range and converge to the model talker's F1 scaling of a particular vowel within that range. Although studies of phonetic convergence on vowel spectra typically rely on normalized formant frequency estimates, reflecting this assumption about a normalized phonetic target of convergence, others use raw (hertz) or Bark-transformed acoustic values (e.g., Nguyen et al., 2012; Sato et al., 2013; Walker & Campbell-Kibler, 2015).[1] This use of raw acoustic values in the assessment of convergence makes it less clear whether the assumed target of convergence is raw acoustic values or normalized phonetic patterns.

Relatedly, phonetic convergence on f0 might also be assumed to involve phonetic normalization. That is, for example, a female shadower may not converge to the lower overall f0 of a male model talker, but rather map the male model talker's f0 range onto her own f0 range and converge to the model talker's f0 scaling within that range. However, the majority of studies of phonetic convergence on f0 use raw or ERB-transformed acoustic values (e.g., Babel & Bulatov, 2012; Pardo et al., 2013, 2017; Sato et al., 2013), although some rely on normalized f0 values (e.g., Lewandowski & Nygaard, 2018). Again, this use of raw acoustic values in the analysis makes it difficult to determine whether the target of convergence on f0 is assumed to be raw acoustic values or normalized phonetic values.

[1] Given that transformations to the Bark, ERB, and related scales capture human peripheral auditory processing and do not effectively normalize for physiological talker differences related to size (Adank et al., 2004), we consider these transformations to be comparable to raw acoustic values in hertz in the context of distinguishing between raw acoustic and normalized phonetic convergence.

This distinction between raw acoustic and normalized phonetic convergence has been explicitly considered in the context of other kinds of spectral information, beyond vowel formants and f0. For example, Zellou et al. (2016) observed normalized phonetic convergence to both increased and decreased vowel nasality for a model talker with especially nasalized vowels overall. That is, the shadowers in Zellou et al.'s (2016) study produced increased nasality when the model talker's nasality was artificially increased through resynthesis and decreased nasality when the model talker's nasality was artificially decreased, but did not converge acoustically to the model talker's overall high degree of nasalization. In contrast, Hauser et al. (2023) observed raw acoustic convergence to both artificially increased and decreased /s/ spectral mean for a model talker with an especially high /s/ spectral mean overall. That is, the shadowers in Hauser et al.'s study converged to the raw acoustic /s/ spectral mean of the model talker, independent of the spectral mean manipulation. In particular, participants whose baseline /s/ spectral mean was lower than the /s/ spectral mean in the exposure materials increased their spectral mean towards that of the model talker, even when her /s/ spectral mean had been manipulated to be lower than her natural productions.

Hauser et al. (2023) proposed that fricative spectra may differ from the cues to vowel nasalization in Zellou et al.'s (2016) study in that the fricative cues serve as the primary cue to fricative place of articulation, undergo perceptual normalization, and can be directly compared across talkers. In contrast, they argued that the nasality cue that Zellou et al. (2016) examined serves as a secondary cue to nasalization, may not undergo perceptual normalization, and cannot be directly compared across talkers. Hauser et al. concluded that one or more of these differences may be a critical factor in predicting raw acoustic vs. normalized phonetic convergence. Like fricative spectra, vowel formant frequencies serve as the primary cues to vowel quality and undergo perceptual normalization, and may therefore exhibit raw acoustic convergence. However, like cues to nasality, vowel formant frequencies are not directly comparable across talkers, and may therefore exhibit normalized phonetic convergence. In the current study, we explicitly examined raw acoustic vs. normalized phonetic convergence to vowel formant frequencies, testing these competing predictions.

Previous research has demonstrated phonetic convergence in the absence of specific instructions to imitate (e.g., Goldinger, 1998; Shockley et al., 2004), and even with instructions to avoid imitation (Walker & Campbell-Kibler, 2015). These findings suggest that phonetic convergence is at least partially automatic. At the same time, greater convergence is typically observed with explicit instructions to imitate the model talker's productions (Clopper & Dossey, 2020; Delaney et al., 2010; Dufour & Nguyen, 2013; Sato et al., 2013; cf. Michelas & Nguyen, 2011). These findings suggest that phonetic convergence is at least partially under conscious control (e.g., Clopper & Dossey, 2020; German et al., 2013; Schertz et al., 2023). Dufour and

Nguyen (2013) argued that a single mechanism underlies both phonetic convergence during word shadowing in the absence of specific instructions to imitate and phonetic convergence with explicit instructions to imitate, with the difference in magnitude attributed to enhanced attention to indexical features of the model talker's voice under instructions to imitate. One possibility, following Dufour and Nguyen, is that explicit instructions to imitate would lead to more evidence of raw acoustic convergence, whereas the absence of instructions to imitate would lead to more evidence of normalized phonetic convergence. That is, whereas automatic convergence may reflect the mapping between the model talker's utterances and the shadowers' utterances in a normalized phonetic space, explicit instructions to imitate may lead to conscious imitation of raw acoustic properties of the model talker's voice. In the current study, we compared convergence in conditions with and without explicit instructions to imitate, testing this prediction.

The magnitude of phonetic convergence is also constrained by phonological structure. For example, Nielsen (2011) observed convergence to lengthened VOT in /p/, but not to shortened VOT in /p/, suggesting that contrast maintenance (here, between /p b/) can override phonetic convergence to VOT. Similarly, Nguyen et al. (2012) observed phonetic convergence by Northern French shadowers to the Southern French back mid-vowel merger. However, the Northern French shadowers did not collapse the two vowel categories and instead produced distinct vowels that were acoustically closer in shadowing than at baseline, consistent with contrast maintenance. Finally, Mitterer and Ernestus (2008) reported the results of a shadowing task involving the two rhotic variants in Dutch. They found virtually no imitation of the phonological variants because shadowers reliably produced their preferred variant throughout the task, as well as no difference in response time due to the mismatch between the model talker's and shadowers' variant productions (see also Mitterer & Müsseler, 2013). They interpreted these results as demonstrating a clear role for phonological abstraction in imitation. That is, since the two rhotic variants were phonologically equivalent, the shadowers neither imitated them, nor were slowed by mapping their non-preferred variant in perception to their preferred variant in production. Together, these results provide further evidence that phonetic convergence is mediated by higher-level phonological structure.

Moreover, convergence has also been observed at the level of phonological structure (Cole & Shattuck-Hufnagel, 2011; Nielsen, 2011). For example, Nielsen found that exposure to a model talker's lengthened VOT in /p/ led to lengthened VOT in shadowers' /k/ productions, suggesting imitation at the level of phonological features or the laryngeal gesture. Similarly, Kwon (2015) found that enhancing only the VOT cue to aspirated stops in Korean led shadowers to produce both longer VOT, consistent with acoustic or phonetic convergence to the model talker, and higher f0 on the following vowel, again suggesting imitation at a phonetic or phonological

level of representation.[2] Further, Cole and Shattuck-Hufnagel found more robust convergence to phonological prosodic features (i.e., location of pitch accents and boundary tones) than to phonetic detail (i.e., pause duration, voice quality) in an explicit imitation task, whereas Song and Clopper (2023) observed the opposite pattern in a shadowing task with implicit imitation. Thus, beyond any distinction between raw acoustic and normalized phonetic targets, convergence reflects more than simple acoustic imitation and is mediated by a range of phonetic and phonological features in both the segmental and prosodic domains.

The magnitude of phonetic convergence is also constrained by phonetic and social variation. For example, Babel (2012) argued that variation in convergence across vowel categories reflects differences in the phonetic variability (or phonetic space available) for different vowels. Relatedly, the baseline distance between the model talker and the shadower promotes convergence, with larger distances allowing greater convergence to be observed (Babel, 2010; Kim & Clayards, 2019; Ross et al., 2021; Walker & Campbell-Kibler, 2015). Phonetic convergence is also more likely for perceptually salient or linguistically marked features relative to less salient or unmarked features (Honorof et al., 2011; Podlipský & Šimáčková, 2015; Wade et al., 2023). However, across dialects, convergence may be blocked for socially salient or stereotyped features, such as Southern American English /aɪ/ monophthongization (Babel, 2010; Clopper & Dossey, 2020; Lee-Kim & Chou, 2024; Ross et al., 2021; Walker & Campbell-Kibler, 2015; cf. Wade, 2022). Although the constraints related to phonetic space, perceptual salience, and phonetic distance could arguably be attributed to limits on articulatory flexibility, perceptual acuity, and our ability to observe small changes, respectively, the constraints due to social salience parallel the phonological structure constraints and suggest that phonetic convergence is also mediated by higher-level social structure. Thus, independent of the target of convergence as raw acoustic or normalized phonetic information, the magnitude of convergence is affected by linguistic and social factors.

1.2. Assessing phonetic convergence from acoustic measures

Just as the acoustic measures used to assess convergence may reflect the underlying assumptions about what is imitated in word shadowing tasks, the analytical methods used to assess convergence likewise reflect these underlying assumptions. The most common approach to assessing phonetic convergence from acoustic measures is the difference-in-distance (DID) metric, in which the acoustic distance between the shadowers' baseline (or pre-exposure) productions and the model talker's productions is compared to the acoustic distance between the shadowers' post-exposure productions and the model talker's productions (Babel, 2012;

[2] An alternative interpretation of this link between imitation of VOT and f0 is that higher f0 is a physiological result of lengthening VOT (Hoole & Honda, 2011). To the extent that VOT and f0 are correlated for biomechanical reasons, this finding may not reflect mediation by higher-level representations on convergence.

Pardo et al., 2013). If this difference-in-distance indicates a smaller post-exposure distance than pre-exposure distance, the results are interpreted as demonstrating convergence. When raw acoustic measures are used to calculate the DID, the resulting interpretation necessarily involves a raw acoustic target of convergence because the shadowers' productions are acoustically more similar to the model talker's productions after exposure than before exposure. In contrast, when normalized acoustic measures are used, the resulting interpretation involves a normalized phonetic target of convergence because the shadowers' productions are more similar in the normalized phonetic space to the model talker's productions after exposure than before exposure. Thus, raw acoustic DIDs are often used to assess convergence in the temporal domain (e.g., Clopper & Dossey, 2020; Lewandowski & Nygaard, 2018; Pardo et al., 2013) where raw acoustic imitation is examined, whereas normalized acoustic DIDs are often used to assess convergence in the spectral domain (e.g., Babel, 2012; Clopper & Dossey, 2020; Lewandowski & Nygaard, 2018; Pardo et al., 2013; Ross et al., 2021) where normalized phonetic convergence is examined. However, raw acoustic DIDs have also been used to assess convergence in the spectral domain (e.g., Babel & Bulatov, 2012; Walker & Campbell-Kibler, 2015; Zellou et al., 2016). Critically, as Zellou et al. (2016) observed, raw acoustic DIDs are difficult to interpret when spectral measures are used because, in the absence of normalization, DIDs that suggest raw acoustic divergence may actually be consistent with normalized phonetic convergence, depending on the baseline productions of the shadowers and the model talker.

An alternative to the DID approach is to directly compare the realization of the phonetic variable of interest in post-exposure productions to pre-exposure productions by including block (e.g., baseline reading vs. shadowing) as a factor in the analysis. This explicit comparison of performance across blocks is often used for categorical variables, such as allophonic alternations, cross-dialect variants, and intonation contours (e.g., Gessinger et al., 2021; Song & Clopper, 2023). In these analyses, convergence is defined as a significant increase from pre-exposure to post-exposure in the proportion of productions that match the target form. The by-block approach has also been used to assess phonetic convergence from raw acoustic values without normalization when the direction of the acoustic change from pre-exposure to post-exposure that is consistent with normalized phonetic convergence is known (Hauser et al., 2023; Nguyen et al., 2012; Sato et al., 2013). For example, Nguyen et al. used a by-block analysis to examine cross-dialect phonetic convergence to French mid-back vowels. In their study, Northern French participants produced /o/ with a higher F1 during a shadowing task relative to a baseline reading task, regardless of their baseline F1 values and consistent with normalized phonetic convergence to the Southern French model talker's merged mid-back vowel system. Hauser et al. (2023) explicitly highlighted the value of this by-block approach when raw acoustic and normalized phonetic convergence do not align in the spectral domain, such as for their model talker with an especially high /s/ spectral mean or Zellou et al.'s (2016) model talker with an especially high

degree of nasality. In particular, the by-block analysis returns both the magnitude and, critically, the direction of shifts in the shadowers' productions from pre-exposure to post-exposure. The direction of a significant shift can then be interpreted in the context of the model talker's and the shadowers' baseline raw acoustic productions. If the shift is in the direction of the model talker's acoustic values (Hauser et al., 2023), the results suggest raw acoustic convergence, whereas if the shift is consistent with a mapping between the model talker's and the shadowers' acoustic production spaces (Nguyen et al., 2012; Zellou et al., 2016), the results suggest normalized phonetic convergence. We therefore adopted the by-block approach in the current study because it can distinguish between raw acoustic and normalized phonetic convergence, allowing us to address our primary research question.

A second alternative to the DID approach is the linear combination approach (Cohen Priva & Sanker, 2019; see also Tobin et al., 2018, Tobin, 2022, Wade et al., 2020, for similar approaches). In this analysis, the acoustic properties of the shadowed utterance are predicted by the acoustic properties of the associated baseline and model talker utterances, as well as by the interactions between the model talker utterances and the relevant experimental design factors. The baseline utterance predictor provides a measure of within-shadower internal consistency, whereas the model talker utterance predictor provides a measure of phonetic convergence, beyond the overall similarity between the shadowers' baseline and shadowed utterances. The interactions between the model talker predictor and the experimental design factors capture variation in the magnitude of phonetic convergence as a function of the experimental manipulations. When this method is applied to shadowing tasks, in which triples of pre-exposure, model talker, and post-exposure tokens are available, phonetic convergence can be assessed on a token-by-token basis (Hauser et al., 2023; MacLeod, 2021; Tobin et al., 2018). That is, a significant effect of the model talker utterance predictor in this linear combination approach indicates that token-by-token variation in the phonetic space is imitated by the shadower, independently of their own pre-exposure token-by-token variation. Critically, as Hauser et al. (2023) argued, the linear combination approach cannot assess the overall direction of shifts in the shadowers' productions from pre-exposure to post-exposure and therefore cannot distinguish between overall raw acoustic and normalized phonetic convergence. Rather, this kind of token-by-token imitation is similar to the notion of synchrony in interactive tasks, in which participants adjust their productions in parallel, potentially without becoming more similar overall (e.g., Edlund et al., 2009; Marekova et al., 2023), and is therefore distinguished from phonetic convergence, or changes in proximity between talkers over time in either a raw acoustic or normalized phonetic space, as assessed by DIDs or by-block comparisons.

1.3. The current study

The existing literature on phonetic convergence in word shadowing tasks suggests both a tight perception-production link, allowing for imitation of fine-grained phonetic detail in acoustic or

articulatory space, as well as higher-level linguistic and social constraints on that link, allowing for imitation to be blocked when phonological contrasts or social stereotypes are at stake (e.g., Babel, 2012; Nielsen, 2011). In the current study, we explicitly considered the raw acoustic vs. normalized phonetic targets of convergence to the vowel variants of a non-stereotyped regional dialect of American English and how those targets might vary under different task instructions.

Our model talker was a young female adult from the Northern American English dialect region, who exhibited features of the Northern Cities Shift (NCS) in her vowel productions. The target words in the shadowing task contained stressed /ɪ ɛ æ ɑ/, which are involved in the NCS: /ɪ ɛ/ are lower (higher F1) and backer (lower F2), /æ/ is higher (lower F1) and fronter (higher F2), and /ɑ/ is lower (higher F1) and fronter (higher F2) than in non-NCS varieties of American English (Labov et al., 2006). The shadowers were male and female adults from the American Midwest, which includes both the Northern American English dialect region and the non-NCS Midland American English dialect region. An acoustic analysis of the model talker's vowels revealed high overall formant values relative to the shadowers' productions, consistent with a relatively short vocal tract (see Section 3.1). The combination of the features of the NCS with the model talker's high overall formant values provided an opportunity to explicitly examine whether the target of convergence for vowel quality is raw acoustic values or normalized phonetic patterns. In particular, given the model talker's high overall formant values, raw acoustic convergence would be realized as an overall increase in the shadowers' formant values for both F1 and F2 for all four target vowels from the baseline task to the shadowing task. However, for the backing of /ɪ ɛ/ and raising of /æ/ in the NCS, which involve lower formant frequencies relative to non-NCS vowels, normalized phonetic convergence to the NCS would result in a decrease in the shadowers' formant values for the F2 of /ɪ ɛ/ and the F1 of /æ/, respectively, from the baseline task to the shadowing task. A summary of the predicted changes in the shadowers' formants for raw acoustic convergence to the model talker's voice vs. normalized phonetic convergence to the NCS is shown in **Table 1**. Given previous research on phonetic convergence to vowel variation in word shadowing tasks (e.g., Babel, 2010, 2012; Clopper & Dossey, 2020; Lewandowski & Nygaard, 2018; Pardo et al., 2017; Sato et al., 2013), we expected to obtain evidence for normalized phonetic convergence to the NCS in the current study, in parallel to Zellou et al.'s (2016) results for vowel nasalization and in contrast to Hauser et al.'s (2023) results for fricative spectral mean.

The competing predictions for raw acoustic and normalized phonetic convergence shown in **Table 1** make the interpretation of an analysis of difference-in-distance scores fairly complex, especially in the absence of acoustic normalization (see Zellou et al., 2016). Human listeners, including the shadowers in the current study, almost certainly perceptually normalize for physiological talker differences in size that lead to overall shifts in raw formant frequencies (Johnson & Sjerps, 2021). This perceptual normalization requires very little input and listeners can successfully normalize for talker variability based on single word utterances to achieve

very high word recognition accuracy (Mullennix & Pisoni, 1990). However, although numerous quantitative transformations have been proposed for acoustic measures to capture this perceptual normalization, the most effective transformations require a relatively large sample of the vowel space to both scale and center individual talkers' formant frequencies (Adank et al., 2004; Johnson, 2020). Moreover, even the most successful of the proposed normalization transformations often fail when the vowel spaces of the talkers differ in shape (Adank et al., 2004; Clopper et al., 2005). For example, the raising of /æ/ in the Northern Cities Shift leads to a more triangular vowel space relative to non-NCS spaces, which tend to be more trapezoidal, whereas back-vowel fronting, which is widespread in many regions of the United States, including the Midwest (Clopper et al., 2019), leads to a parallelogram-shaped vowel space. Thus, although acoustic normalization for shadower vocal tract length would be ideal, the sample of the vowel space included in the stimulus materials in our study is insufficient for robust normalization. Although we have previously normalized a different dataset involving the same stimulus materials based on the realization of /i ɑ/ (Clopper & Dossey, 2020), that approach would be inappropriate for the current data, given that /ɑ/ is implicated in the Northern Cities Shift and is therefore not stable across shadowers and the model talker. Indeed, both variability in the NCS, as well as variability in back-vowel fronting, across the shadowers and the model talker in the current study, lead to a variety of overall vowel space shapes that are not captured well by the most common normalization transformations in the literature, including z-scoring (Lobanov, 1971) and Nearey's uniform scaling (Nearey, 1978). A comparison of normalization transformations of our data is available on the Open Science Framework repository for this project: https://osf.io/ye7nw/. The primary effects of interest, related to the target of convergence as raw acoustic values vs. normalized phonetic patterns and the effect of instruction condition, are qualitatively similar in analyses using Bark-transformed, z-scored, and Nearey uniform scaling measures. We therefore rely on raw (Bark-transformed) acoustic measures for the current analysis.

In contrast to the DID approach, a comparison of the shadower productions from the baseline task to the shadowing task (Hauser et al., 2023; Nguyen et al., 2012) allows for a straightforward interpretation of the results, even without normalization: If all of the formants increase from baseline to shadowing, we have evidence of raw acoustic convergence, whereas if the direction of the shifts from baseline to shadowing differs in accordance with the NCS predictions in **Table 1**, we have evidence of normalized phonetic convergence to the NCS. We therefore first assessed overall convergence in our data using an explicit by-block comparison of baseline and shadowing tokens to determine whether the target of convergence in the shadowing task was raw acoustics or normalized phonetic patterns. As noted above, the linear combination approach provides a complementary perspective to this overall comparison by assessing token-by-token phonetic convergence or synchrony. We therefore also assessed convergence within the phonetic

vowel spaces of our shadowers using the linear combination approach (Cohen Priva & Sanker, 2019; Hauser et al., 2023; MacLeod, 2021; Tobin et al., 2018) to further assess the nature of convergence to vowel quality variation in the word shadowing task.

Vowel formant	Acoustic convergence	NCS convergence
/ɪ/ F1	increase	increase
/ɪ/ F2	**increase**	**decrease**
/ɛ/ F1	increase	increase
/ɛ/ F2	**increase**	**decrease**
/æ/ F1	**increase**	**decrease**
/æ/ F2	increase	increase
/ɑ/ F1	increase	increase
/ɑ/ F2	increase	increase

Table 1: Summary of the predictions for raw acoustic convergence vs. normalized phonetic convergence to the Northern Cities Shift (NCS) for F1 and F2 of the target vowels /ɪ ɛ æ ɑ/. Bolded rows show competing predictions.

In the current study, we manipulated the shadowing task instructions to explore how implicit vs. explicit imitation might impact the target of phonetic convergence (Dufour & Nguyen, 2013; Schertz et al., 2023). In a between-subject design, we instructed participants either (1) to simply repeat the words that they heard in the shadowing block; (2) to repeat the words after the talker from Chicago, Illinois; or (3) to imitate the way the talker from Chicago, Illinois, said the words. The first condition is the typical method used in word shadowing tasks (e.g., Goldinger, 1998; Shockley et al., 2004), where imitation is assumed to arise implicitly as a result of the perception-production link. The second condition allowed us to confirm that the Northern dialect of American English, represented by Chicago, Illinois, is not sufficiently socially salient to affect implicit imitation. That is, we expected that listeners would not be able to identify the dialect of the model talker from the stimulus materials alone (Dailey, 2018), so in the second condition, we explicitly named the regional background of the model talker in the instructions to introduce the potential for social constraints on phonetic convergence to emerge. However, given that the Northern dialect is not socially stereotyped (Campbell-Kibler, 2012; Niedzielski, 1999), we did not expect the addition of regional background information about the model talker to constrain convergence and therefore did not expect to observe differences in phonetic

convergence between the first and second conditions. The third condition allowed us to examine acoustic convergence to the model talker's voice when the shadowers were explicitly asked to imitate her. Consistent with previous work (Clopper & Dossey, 2020; Delaney et al., 2010; Dufour & Nguyen, 2013; Sato et al., 2013; cf. Michelas & Nguyen, 2011), we expected to observe greater convergence overall in the explicit imitation condition than in the other two conditions. As in the second condition, we explicitly named the regional background of the model talker in the instructions, following Clopper and Dossey (2020), although we did not expect the addition of regional background information about the model talker to affect overall convergence in either condition.

2. Methods

2.1. Participants

Sixty-nine adults participated as shadowers in the current study. They were recruited from among the visitors to the Center of Science and Industry (COSI) museum in Columbus, Ohio. The participants were asked to self-report their age, gender, native (first) language, other languages they speak, and their residential history. The participants included 32 females and 37 males, who ranged in age from 19–74 years old (M = 38 years, SD = 14 years). They were all native speakers of American English. Thirty-four participants reported speaking one or more other languages, including Arabic (N = 1), American Sign Language (N = 3), Chinese (N = 1), French (N = 12), German (N = 4), Italian (N = 2), Japanese (N = 1), Korean (N = 1), Mandarin (N = 2), Polish (N = 1), Portuguese (N = 1), Russian (N = 2), and Spanish (N = 15). The participants had varied residential histories, but all had lived in the American Midwest, which includes both the Midland and Northern dialect regions. Thirty-four participants had lived exclusively in the American Midwest and the remaining 35 participants had lived in the American Midwest and at least one other dialect region in the United States or abroad. This inclusion of participants who had lived in the American Midwest was intended to increase the likelihood of participant exposure to the Northern Cities Shift. Given that the Northern dialect is not socially stereotyped (Campbell-Kibler, 2012; Niedzielski, 1999), previous exposure to the variety was desirable to maximize the likelihood of observing social constraints on phonetic convergence. The limited residential history we were able to obtain from participants in the science museum setting (i.e., a list of all of the states where they had lived) made a more fine-grained analysis of likely exposure to the Northern vs. Midland dialects impossible because the dialect boundary bisects several Midwestern states. Data from an additional 16 participants were excluded prior to the analysis because the participants had never lived in the Midwest or did not report their residential history (N = 11), because of excessive mispronunciations or mishearing of the stimulus targets (N = 1), or because the responses were not recorded due to experimenter error (N = 4). All participants provided informed consent prior to beginning the

study. The experiment was reviewed and approved by the Institutional Review Board at Ohio State University (protocol #2012B0213).

2.2. Stimulus materials

The stimulus materials comprised 24 multisyllabic target words, with six words for each of the stressed vowels /ɪ ɛ æ ɑ/. These vowels are implicated in the Northern Cities Shift (Labov et al., 2006) and were selected for this study so that phonetic convergence to the Northern Cities Shift could be assessed. As shown in **Table 1**, the Northern Cities Shift involves the backing and lowering of /ɪ ɛ/, raising and fronting of /æ/ and fronting and lowering of /ɑ/. Phonetic convergence to the Northern Cities Shift would therefore be realized as shifts in these same directions from the baseline to the shadowing block of the experiment. An additional 24 multisyllabic filler words were selected, with six words for each of the stressed vowels /i aɪ oʊ u/. These vowels are not implicated in the Northern Cities Shift and therefore served as distractors in the current study. The full set of 48 stimulus words is shown in **Table 2**.

The model talker was selected from the Indiana Speech Project corpus (Clopper et al., 2002) and was a female white native speaker of American English. She had lived exclusively in the Northern dialect region until age 18 years and was 22 years old at the time of the recording. Her productions of the 48 target words were used as the stimulus materials in the shadowing block of the current study.

Target words		**Filler words**	
/ɪ/	amphibian, aristocrat, conspicuous, imposition, liberator, precipitate	/i/	delete, equalizer, misconceive, obedient, readable, sweeten
/ɛ/	clarinet, embezzle, epilepsy, legendary, obsession, silhouette	/aɪ/	comply, mighty, sizeable, spider, tiger, unadvised
/æ/	appetizer, caterpillar, deactivate, evaporate, procrastinate, spatula	/oʊ/	afloat, commotion, coyote, rodeo, tapioca, unbroken
/ɑ/	hypnotic, octopus, rhinoceros, robin, roster, slobber	/u/	accuser, bazooka, commute, disapproval, kangaroo, kazoo

Table 2: Stimulus words in the shadowing task.

2.3. Procedure

The participants were seated at a personal computer equipped with a headset microphone in a glass-enclosed laboratory space in the science museum. The experiment involved two blocks: a baseline reading block and a shadowing block. In the baseline reading block, the stimulus words were presented on the screen one at a time and the participants were asked to read each word

aloud. On each trial, one word was presented on the screen for 3500 ms, followed by a blank screen for 500 ms. The order of presentation of the words was randomized separately for each participant. In the shadowing block, the stimulus materials produced by the model talker were presented over the headset one at a time at a comfortable listening level and the participants were asked to repeat each word aloud. On each trial, a fixation cross was presented on the screen for 4500 ms, followed by a blank screen for 1000 ms. One of the stimulus words was presented auditorily 500 ms after the onset of the fixation cross. The order of presentation of the words was randomized separately for each participant. The participants' responses were recorded directly to the computer with a sampling rate of 44.1 kHz and 16-bit quantization.

Three between-subject conditions differed in the instructions provided to the participants in the shadowing block. In the first condition, the participants were simply asked to repeat the words after the model talker. This "implicit imitation" condition was designed to assess automatic phonetic convergence to the Northern Cities Shift, in the absence of explicit instructions to imitate. This condition was expected to elicit phonetic convergence, as in previous studies (Goldinger, 1998; Shockley et al., 2004). In the second condition, the participants were asked to repeat the words after the model talker and were told that she is from Chicago, Illinois. This "primed imitation" condition was designed to assess the effects of explicit social information about the talker on the magnitude of implicit phonetic convergence. This condition was not expected to differ from the implicit imitation condition because the Northern dialect is not strongly stereotyped and phonetic convergence to the Northern dialect was therefore not expected to be inhibited by the regional background information (Babel, 2010; Clopper & Dossey, 2020; Ross et al., 2021; Walker & Campbell-Kibler, 2015). In the third condition, the participants were asked to imitate the way the model talker produced the words and were told that she is from Chicago, Illinois. This condition was designed to assess explicit imitation of the Northern Cities Shift and was expected to elicit greater evidence of convergence overall, as in previous studies (Clopper & Dossey, 2020; Delaney et al., 2010; Dufour & Nguyen, 2013; Sato et al., 2013; cf. Michelas & Nguyen, 2011). As in the second condition, the inclusion of explicit social information was not expected to affect the overall magnitude of phonetic convergence, but this social information was provided in parallel to Clopper and Dossey's design. The participants were randomly assigned to one of the three experimental conditions, with 23 participants per condition.

2.4. Acoustic analysis

The shadowing task elicited a total of 3312 target word tokens (69 shadowers $\times$ 24 target words $\times$ 2 blocks). The participants' recordings were first automatically segmented at the word and phone levels using the Penn Phonetics Lab Forced Aligner (Yuan & Liberman, 2008). The segmentation boundaries for the stressed vowel in each target word were then hand-corrected,

following segmentation conventions described by Peterson and Lehiste (1960). Estimates of the first and second formant frequencies were obtained for each stressed vowel at five temporal points: 20%, 35%, 50%, 65%, 80% of the vowel duration using a 12th order LPC analysis (Burg method) in the frequency range of 0–5500 Hz in Praat (Boersma & Weenink, 2023). Formant estimate outliers, defined as values outside of 3 SDs of the mean including all five temporal points across both blocks for each talker for each vowel, were hand-checked and corrected by adjusting the LPC order and/or frequency range of the analysis to visually align the formant estimates with the spectrogram. The formant estimates were converted to the Bark scale for analysis to capture automatic human auditory peripheral processing of frequency information and therefore align the formant estimates with human sensory processing (Traunmüller, 1990).

Prior to the statistical analysis, 138 tokens were excluded (69 pairs of baseline-shadowed tokens) due to mispronunciations or excessive background noise. The final analysis included 3174 word tokens (1587 pairs of baseline-shadowed tokens).

2.5. Statistical analysis

The first analysis compared shadowers' vowel durations and formant frequencies in the two blocks of the experiment (Gessinger et al., 2021; Hauser et al., 2023; Nguyen et al., 2012; Song & Clopper, 2023). In this analysis, we treated vowel duration (seconds), F1 (Bark) at temporal midpoint, and F2 (Bark) at temporal midpoint from the baseline and shadowing blocks as the dependent variables in a series of linear mixed effects models. The fixed effects were the target vowel category (/ɪ ɛ æ ɑ/), shadower gender (female, male), instruction condition (implicit, primed, explicit), and block (baseline, shadowing). All interactions were included as fixed effects. To control for vowel duration effects on formant frequencies (Moon & Lindblom, 1994), vowel duration and its interaction with vowel category were included as covariates in the F1 and F2 models. The results of these models were then examined in relation to the model talker and shadowers' baseline productions to assess whether shifts in production between the baseline and shadowing blocks were consistent with raw acoustic or normalized phonetic convergence to the model talker.

The second analysis examined token-by-token phonetic convergence using the linear combination method (Cohen Priva & Sanker, 2019; Hauser et al., 2023; MacLeod, 2021; Tobin et al., 2018). In this analysis, we treated F1 (Bark) at temporal midpoint and F2 (Bark) at temporal midpoint from the shadowing block as the dependent variables in a pair of linear mixed effects models. The fixed effects were the relevant acoustic measure (F1 at temporal midpoint, F2 at temporal midpoint) from the baseline block and from the model talker, as well as target vowel category (/ɪ ɛ æ ɑ/), shadower gender (female, male), and instruction condition (implicit, primed, explicit). All interactions involving the three experimental design factors (vowel category, shadower gender, and instruction condition) with one another and with the

model talker acoustic predictor were included. No interactions involving the shadower baseline predictor were considered. As in the by-block models, vowel duration and its interaction with vowel category were included as covariates.

In both analyses, the categorical variables were sum-contrast coded and, in the linear combination model, the shadower baseline and model talker predictors were scaled and centered using z-scores. The maximal random effect structure that converged was used for each model (Bates et al., 2015). Significant main effects and interactions were assessed using the Satterthwaite approximation of degrees of freedom, as implemented in the *lmerTest* package in R (Kuznetsova et al., 2017). Post-hoc comparisons were conducted using estimated marginal means and slopes comparisons, as implemented in the *emmeans* package in R (Lenth, 2024). The data and analysis code for this study are available on the Open Science Framework repository for this project: https://osf.io/ye7nw/.

3. Results

3.1. Shadower baseline vs. model talker utterances

Before considering phonetic convergence during shadowing, we first needed to understand the relationship between the model talker's utterances and the shadowers' baseline utterances. A summary of the target /ɪ ɛ æ ɑ/ and distractor /i oʊ u/ vowel productions for the model talker and the shadowers is shown in **Figure 1**, separately for the female and male shadowers.

Since we were unable to normalize the formant frequencies effectively (see Section 1.3), we relied on a visual inspection of the vowel spaces in **Figure 1** as a starting point for understanding phonetic convergence in our data. As shown in the left panel of **Figure 1**, the model talker (light symbols) has higher formant values overall than most of the female shadowers (dark symbols), so that her vowel space is shifted down and to the left relative to the female shadowers' baseline. Consistent with the Northern Cities Shift, the model talker's /ɛ æ/ are closer in F1 and reversed along F2 (i.e., /æ/ is fronter than /ɛ/) relative to the mean of the female shadowers' utterances. Evidence for the Northern Cities Shift in the model talker's /ɪ ɑ/ is somewhat harder to assess, given the overall shift in her vowel space relative to the female shadowers. However, the model talker's /ɑ/ is lower than /æ/, unlike the female shadowers' mean /æ ɑ/, which are more similar along F1. The model talker's relatively lower /ɑ/ could be the result of either /æ/ raising or /ɑ/ lowering, consistent with the NCS. The model talker's /ɑ/ is also fronter than /oʊ/, consistent with the NCS, although both /oʊ u/ are backed for the model talker relative to the female shadowers' mean /oʊ u/, consistent with variable back-vowel fronting among American Midwesterners (Clopper et al., 2019). Finally, the model talker's /ɪ/ aligns in F1 with /oʊ/ and in F2 with /ɛ æ/, whereas the female shadowers' mean /ɪ/ is slightly higher than /oʊ/ and slightly fronter than /ɛ æ/. This pattern is also consistent with the NCS in the model talker's vowel space, although the differences are quite modest.

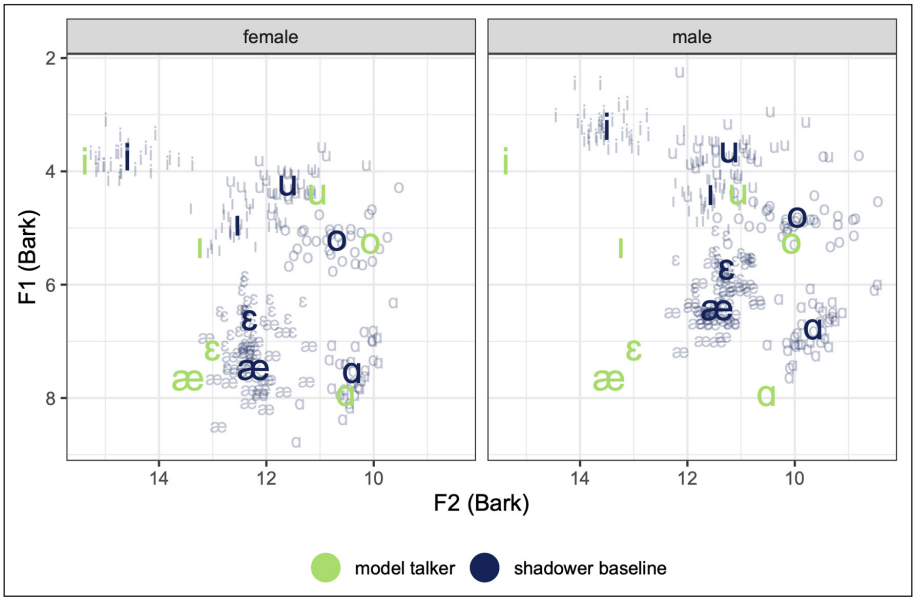

Figure 1: Summary vowel spaces of the model talker's utterances and the shadowers' baseline utterances for female (left) and male (right) shadowers. Small symbols show individual shadower means. Large symbols show model talker means and shadower grand means.

As shown in the right panel of **Figure 1**, the model talker (light symbols) also has higher formant values overall than all of the male shadowers (dark symbols), as expected. Consistent with the Northern Cities Shift, the model talker's /ɛ æ/ are closer in F1 and more distant in F2 relative to the male shadowers' utterances. As in the comparison to the female shadowers, evidence for the NCS in the model talker's /ɪ ɑ/ is harder to assess, given the overall shift in her vowel space relative to the male shadowers. A visual comparison of the female and male shadowers' vowel spaces suggests greater fronting of /æ/ and/or backing of /ɛ/, consistent with the NCS, among the male shadowers' baseline than the female shadowers' baseline overall. In comparison to the female shadowers, the male shadowers also show a relatively lower /ɑ/ than /æ/ and an alignment in F2 of /ɪ/ with /ɛ æ/, similar to the model talker and consistent with the NCS, although these differences from the female shadowers are again quite modest.

A visual inspection of the individual shadowers' vowel spaces revealed modest evidence of the Northern Cities Shift in their baseline productions. Ten shadowers (4 female, 6 male) produced lowering or backing of /ɛ/ and/or fronting or raising of /æ/ in the baseline block, consistent with the NCS. Three shadowers (1 female, 2 male) produced /ɑ/ fronting relative to /oʊ/ in the baseline block, consistent with the NCS, although /oʊ/ fronting was highly variable across shadowers as shown in **Figure 1** and therefore is not a highly reliable benchmark for assessing /ɑ/ fronting. The shadowers varied little in their production of /ɪ/ in the baseline block, providing limited evidence for the NCS. Thus, a small number of shadowers produced

vowels with evidence of the NCS in the baseline block, and the NCS was present among more male shadowers than female shadowers, consistent with the overall shadower means shown in **Figure 1**. Individual vowel spaces for each shadower are available on the Open Science Framework repository for this project: https://osf.io/ye7nw/.

This visual inspection of the shadowers' baseline utterances relative to the model talker's utterances leads to the competing predictions shown in **Table 1**. Whereas raw acoustic convergence to the model talker would result in increased formant values from the baseline to the shadowing block for all four vowels, normalized phonetic convergence to the Northern Cities Shift would result instead in at least lowering and backing of /ɛ/ and fronting and raising of /æ/, if not also backing and lowering of /ɪ/ and fronting and lowering of /ɑ/. The shadowers who produced vowels consistent with the NCS in the baseline block might be expected to exhibit a different pattern of normalized phonetic convergence than the shadowers who did not. In particular, shadowers who produced features of the NCS in the baseline block might phonetically converge less overall to the model talker because their baseline productions are more similar to the model talker's to begin with (Babel, 2012; Walker & Campbell-Kibler, 2015). Any normalized phonetic convergence would therefore be smaller in magnitude and/or harder to empirically observe, relative to shadowers without features of the NCS in the baseline block. Alternatively, shadowers who are more advanced in the NCS in the baseline block than the model talker might exhibit normalized phonetic convergence in the opposite direction of the predictions in **Table 1** if they shifted towards a less-advanced pattern in the shadowing block. In both cases, the shadowers who produced features of the NCS in the baseline block would weaken any evidence we obtain for normalized phonetic convergence to the NCS in our by-block analysis.

3.2. Overall phonetic convergence in the shadowing block

A summary of the target vowel durations for the model talker, the shadowers in the baseline block, and the shadowers in the shadowing block are shown in **Figure 2**, separately by vowel category. The overall patterns suggest convergence towards the model talker's longer /æ ɑ/, divergence from the model talker's shorter /ɪ/, and little change from the baseline block to the shadowing block for the duration of /ɛ/. There is no clear evidence of overall vowel reduction in the shadowing block relative to the baseline block due to second mention reduction (Fowler & Housum, 1987).

The linear mixed-effects model predicting vowel duration across both blocks revealed significant main effects of vowel category ($F(3, 20.0) = 14.91$, $p < .001$) and block ($F(1, 3043.0) = 54.96$, $p < .001$), as well as interactions between vowel category and instruction condition ($F(6, 3043.9) = 10.12$, $p < .001$), shadower gender and instruction condition

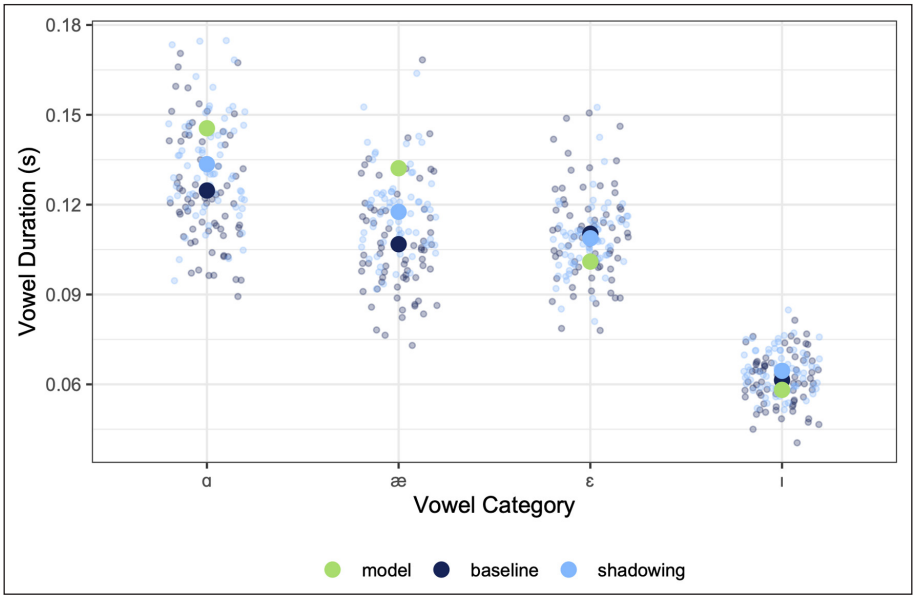

Figure 2: Summary of vowel durations of the model talker's utterances, the shadowers' baseline utterances, and the shadowers' shadowed utterances for each vowel category. Small symbols show individual shadower means. Large symbols show model talker means and shadower grand means.

($F(2, 63.0) = 4.29$, $p = .018$), vowel category and block ($F(3, 3043.0) = 15.44$, $p < .001$), and vowel category, shadower gender, and instruction condition ($F(6, 3043.9) = 4.56$, $p < .001$). None of the other main effects or interactions were significant. The maximal random effect structure that converged included by-shadower and by-word intercepts. The main effect of vowel category reflects vowel-intrinsic variation in duration overall, as expected. Post-hoc pairwise comparisons of estimated marginal means by vowel category revealed significantly (all $p < .001$) shorter durations for /ɪ/ than for the other three vowels for both shadower genders in all three conditions, as expected. For the male shadowers in the explicit condition, /ɛ/ was also significantly shorter than /ɑ/ ($\beta = .031$, $z = 2.86$, $p = .022$). The significant three-way interaction therefore reflects modest variation across participant groups in overall vowel durations. The main effect of block reflects an overall increase in duration from baseline to shadowing. However, this main effect is mediated by the significant interaction with vowel category. Post-hoc pairwise comparisons of estimated marginal means by block for each vowel category revealed significant effects of block for the duration of /ɑ/ ($\beta = –.008$, $z = –6.03$, $p < .001$), /æ/ ($\beta = –.010$, $z = –7.67$, $p < .001$), and /ɪ/ ($\beta = –.003$, $z = –2.14$, $p = .032$). The shadowers lengthened all three vowels in the shadowing block relative to the baseline block, reflecting convergence on /æ ɑ/ and divergence on /ɪ/.

A summary of the target vowel formant estimates for the model talker, the shadowers in the baseline block, and the shadowers in the shadowing block are shown in **Figure 3**, separately for

the female and male shadowers in each of the three instruction conditions. The overall patterns suggest convergence towards the Northern Cities Shift across conditions, including especially raising of /æ/ and fronting of /ɑ/.

The linear mixed-effects model predicting F1 across both blocks revealed significant main effects of vowel category ($F(3, 57.1) = 119.65$, $p < .001$) and shadower gender ($F(1, 63.0) = 138.40$, $p < .001$), as well as interactions between vowel category and shadower gender ($F(3, 62.5) = 14.44$, $p < .001$) and between vowel category and block ($F(3, 2770.7) = 11.75$, $p < .001$). The vowel duration covariate ($F(1, 2883.1) = 24.86$, $p < .001$) and its interaction with vowel category ($F(3, 2436.2) = 7.72$, $p < .001$) were also significant. None of the other main effects or interactions were significant. The maximal random effect structure that converged included by-shadower and by-word intercepts, by-shadower slopes for vowel category and block, and by-word slopes for instruction condition. The main effect of vowel category reflects vowel-intrinsic variation in F1 overall, as expected, and the main effect of shadower gender reflects higher overall F1 estimates for female shadowers than male shadowers, as expected. Post-hoc pairwise comparisons of estimated marginal means by vowel category, collapsed across blocks, for the female shadowers revealed significant (all $p < .001$) pairwise vowel differences in F1 for all vowel pairs except /æ ɑ/. For the male shadowers, all pairwise vowel comparisons were significant (/æ ɑ/: $p = .012$, all others: $p < .001$). The significant difference in F1 for /æ ɑ/ for the male shadowers is consistent with the visual impression of raised /æ/ in both blocks among the male shadowers relative to the female shadowers in **Figure 3**. The vowel duration covariate is positive ($\beta = 1.816$, $t(2883) = 4.99$, $p < .001$), suggesting lowering of longer vowels. However, this main effect is mediated by the significant interaction with vowel category. Post-hoc examination of estimated marginal trends reveals that this effect of vowel duration on F1 is significant only for /ɪ/ ($\beta = 4.442$, $z = 4.59$, $p < .001$) and /æ/ ($\beta = 2.281$, $z = 3.585$, $p < .001$). Post-hoc pairwise comparisons of estimated marginal means by block for each vowel category reveal a significant effect of block only for the F1 of /æ/ ($\beta = .129, z = 4.70, p < .001$). Consistent with the NCS predictions in **Table 1**, the F1 of /æ/ decreases from the baseline block to the shadowing block, after controlling for vowel duration.

The linear mixed-effects model predicting F2 across both blocks revealed significant main effects of vowel category ($F(3, 29.1) = 21.98, p < .001$), shadower gender ($F(1, 63.0) = 159.47$, $p < .001$), and block ($F(1, 3041.1) = 8.72$, $p < .001$), as well as interactions between vowel category and shadower gender ($F(3, 3039.7) = 18.56, p < .001$), between vowel category and instruction condition ($F(6, 3040.0) = 10.48, p < .001$), between vowel category and block ($F(3, 3040.0) = 4.66$, $p = .003$), and between vowel category, shadower gender, and instruction condition ($F(6, 3040.2) = 8.64, p < .001$). The vowel duration covariate ($F(1, 3103.6) = 8.26$, $p = .004$) and its interaction with vowel category ($F(3, 3063.9) = 8.13, p < .001$) were also significant. None of the other main effects or interactions were significant. The maximal random effect structure that converged included by-shadower and by-word intercepts. The main effect

of vowel category reflects vowel-intrinsic variation in F2 overall, as expected, and the main effect of shadower gender reflects higher overall F2 estimates for female shadowers than male shadowers, as expected. Post-hoc pairwise comparisons of estimated marginal means by vowel category revealed significant (all $p < .001$) differences between /ɑ/ and the three front vowels for both shadower genders in all three conditions, as expected. For the female shadowers in the explicit condition, the /ɪ æ/ difference was also significant ($\beta = -.624$, $z = -2.91$, $p = .019$), consistent with the visual impression of fronted /æ/ in both blocks among the male shadowers relative to the female shadowers in **Figure 3**. The vowel duration covariate is positive ($\beta = 1.131$, $t(3104) = 2.74$, $p = .004$), suggesting fronting of longer vowels. However, this main effect is mediated by the significant interaction with vowel category. Post-hoc examination of estimated marginal trends reveals a positive effect of vowel duration on F2 for /ɪ/ ($\beta = 2.84$, $z = 2.76$, $p = .006$) and /ɛ/ ($\beta = 1.90$, $z = 3.194$, $p = .001$), as well as a negative effect of vowel duration on F2 for /ɑ/ ($\beta = -1.41$, $z = -2.477$, $p = .013$), consistent with spectral reduction of shorter vowels (Moon & Lindblom, 1994). The main effect of block reflects an overall increase in formant values from baseline to shadowing, consistent with the acoustic predictions in **Table 1**. However, this main effect is mediated by the significant interaction with vowel category. Post-hoc

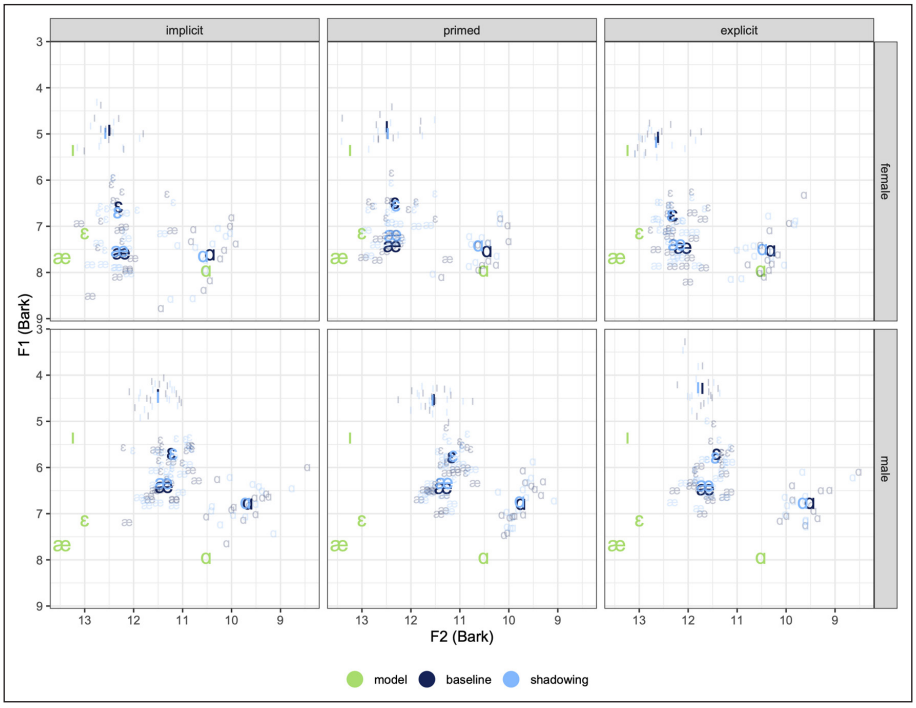

Figure 3: Summary vowel spaces of the model talker's utterances, the shadowers' baseline utterances, and the shadowers' shadowed utterances for female (top) and male (bottom) shadowers in each of the three instruction conditions (implicit, primed, explicit). Small symbols show individual shadower means. Large symbols show model talker means and shadower grand means.

pairwise comparisons of estimated marginal means by block for each vowel category revealed a significant effect of block only for the F2 of /ɑ/ (β = –.125, z = –4.64, p < .001). As shown in **Table 1**, this result is consistent with both raw acoustic convergence to the model talker's high overall formants and normalized phonetic convergence to the NCS, after controlling for vowel duration.

Taken together, the results of this analysis reveal overall shifts in production from baseline to shadowing for F1 of /æ/ and for F2 of /ɑ/. Although the shift in production for F2 of /ɑ/ could be the result of either raw acoustic or normalized phonetic convergence (see **Table 1**), the observed shift in production for F1 of /æ/ is consistent only with normalized phonetic convergence to the NCS. Critically, these shifts are independent of overall effects of vowel duration on formant frequencies.

3.3. Token-by-token phonetic convergence in the shadowing block

The relationship between the model talker utterances and the shadowing utterances is shown in **Figure 4**, separately by formant (F1, F2) and instruction condition. The overall patterns suggest token-by-token convergence to the model talker for both formants in all three instruction conditions.

The linear mixed-effects model predicting F1 in the shadowing block revealed significant main effects of the baseline utterance (F(1, 1363.0) = 296.13, p < .001), the model talker utterance (F(1, 15.2) = 5.10, p = .039), vowel category (F(3, 25.8) = 9.56, p < .001), and shadower gender (F(1, 528.9) = 43.38, p < .001). The vowel duration covariate (F(1, 923.1) = 5.07, p = .025) and its interaction with vowel category (F(3, 998.2) = 7.39, p < .001) were also significant. None of the other main effects or interactions were significant. The maximal random effect structure that converged included by-shadower and by-word intercepts and a by-shadower slope for model talker utterance. The main effect of the baseline utterance confirms within-shadower consistency from the baseline to the shadowing block, as expected, and the main effect of model talker utterance confirms token-by-token phonetic convergence to the model talker on F1. The main effect of vowel category reflects vowel-intrinsic variation in F1 overall, even after controlling for the shadower baseline, and the main effect of shadower gender reflects higher overall F1 estimates for female shadowers than male shadowers, as expected. The effect of the vowel duration covariate and its interaction with vowel category reflect lowering of longer vowels for /ɪ/ (β = 3.83, t(691) = 3.03, p = .003) and raising of longer vowels for /ɑ/ (β = –2.04, t(1407) = –2.77, p = .006). The lack of significant interactions involving model talker utterance suggests that token-by-token phonetic convergence on F1 was robust across vowel category, shadower gender, and condition.

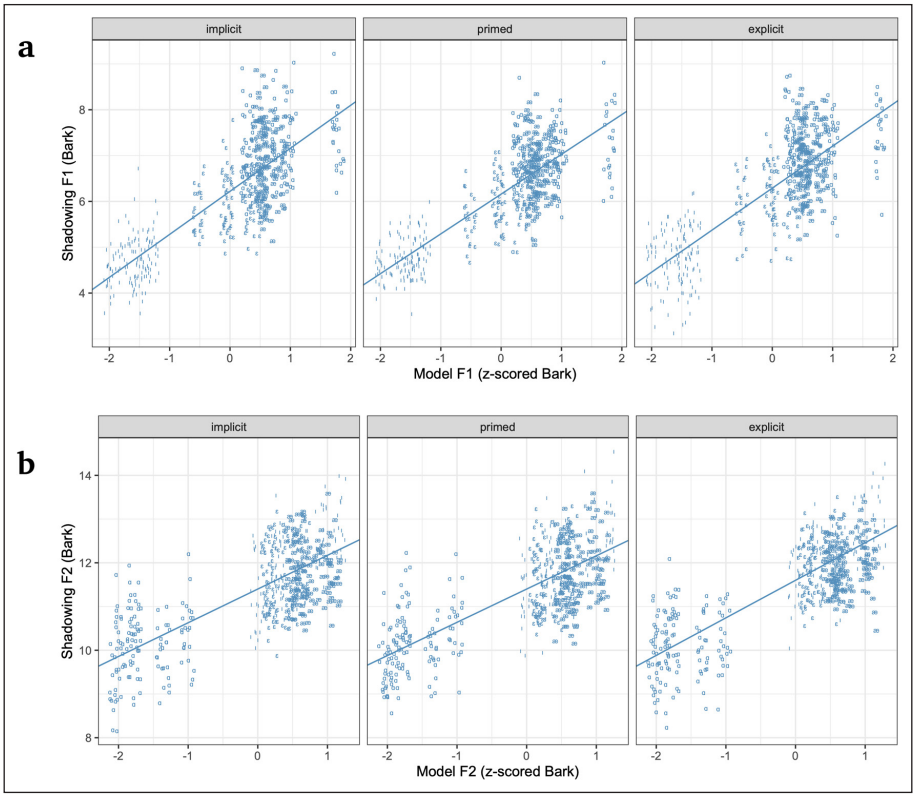

Figure 4: Summary of the relationship between the model talker utterances and the shadowers' shadowed utterances for F1 (top) and F2 (bottom) in each of the three instruction conditions (implicit, primed, explicit). Lines show linear-mixed effects model estimates. Symbols show individual tokens.

The linear mixed-effects model predicting F2 in the shadowing block revealed significant main effects of the baseline utterance ($F(1, 1162.9) = 528.04$, $p < .001$), the model talker utterance ($F(1, 16.3) = 9.55$, $p = .007$), and shadower gender ($F(1, 516.0) = 41.14$, $p < .001$), as well as significant interactions between model talker utterance and instruction condition ($F(2, 1443.1) = 3.39$, $p = .034$), and between vowel category and instruction condition ($F(2, 1442.8) = 2.34$, $p < .030$). The interaction between the vowel duration covariate and vowel category ($F(2, 1134.4) = 2.79$, $p = .039$) was also significant. None of the other main effects or interactions were significant. The maximal random effect structure that converged included by-shadower and by-word intercepts. As in the analysis of F1, the main effects confirm shadower consistency from the baseline to the shadowing block, significant token-by-token phonetic convergence to the model talker, and higher overall formant estimates for female shadowers than male shadowers. Post-hoc pairwise comparisons of estimated marginal means by vowel

category revealed no significant differences in any of the three instruction conditions, after controlling for the shadower baseline utterances. The significant interaction between vowel category and instruction condition therefore reflects modest variation across participant groups in overall vowel acoustics. The interaction of the vowel duration covariate with vowel category reflects fronting of longer vowels for /ɪ/ (β = 2.98, t(783) = 2.28, p = .023). Post-hoc pairwise comparisons of estimated marginal trends by instruction condition revealed significantly more token-by-token convergence in the primed condition than in the explicit condition (β = .202, t(1460) = 2.45, p = .039). None of the other pairwise differences across conditions were significant.

4. Discussion

The results of the by-block analysis provide evidence of /æ/ raising and /ɑ/ fronting in the shadowing block relative to the baseline block. These shifts in production are consistent with either raw acoustic or normalized phonetic convergence on /ɑ/, but, critically, are only consistent with normalized phonetic convergence on /æ/. These effects are also critically independent of vowel duration effects, confirming that the results do not simply reflect spectral reduction in the shadowing block relative to the baseline block, due to second mention reduction (Fowler & Housum, 1987). Thus, overall, the results of the by-block analysis suggest normalized phonetic convergence to the NCS, rather than raw acoustic convergence to the model talker's voice, consistent with previous work suggesting normalized phonetic convergence to spectral vowel variation (e.g., Babel, 2010, 2012; Clopper & Dossey, 2020; Nguyen et al., 2012; Pardo et al., 2017; Zellou et al., 2016).

Although the results for /æ ɑ/ suggest convergence to the NCS, by-block effects were only observed for one formant for each of these two vowels (F1 for /æ/, F2 for /ɑ/), and no evidence of convergence was observed for either formant for /ɪ ɛ/. This variability in convergence across measures is consistent with previous studies, where individual shadowers converge on some dimensions, but not others (e.g., Sanker, 2015; Pardo et al., 2017). The observed convergence on the low vowels /æ ɑ/, but not the non-low vowels /ɪ ɛ/, is also consistent with Babel's (2012) findings, which involved Northern American English shadowers and California English model talkers. Properties of the NCS in the speech of either the model talkers or the shadowers may therefore lend itself to convergence especially on low vowels, which may be more variable within and across American English dialects and therefore afford greater opportunities for convergence (Walker & Campbell-Kibler, 2015). Finally, it is also possible that the model talker in the current study produced more /æ/ raising and /ɑ/ fronting than other features of the NCS, leading to greater starting distances from the shadowers and therefore greater evidence of convergence (Babel, 2012; Ross et al., 2021). In the absence of a normalization transformation that can successfully capture the dialect variation in the model talker's and shadowers' vowel spaces, however, this possibility remains speculative.

The observed results of normalized phonetic convergence align with Zellou et al.'s (2016) findings for coarticulatory vowel nasalization, but differ from Hauser et al.'s (2023) findings, in which raw acoustic convergence, but not normalized phonetic convergence, to fricative spectra was observed. One possible explanation for this pattern of results, following Hauser et al., is that differences in perceptual normalization mechanisms underlie these differences in the target of convergence. In particular, whereas fricative spectra are characterized by a single peak frequency, which is normalized relative to surrounding vowels (Hauser et al., 2023; cf. Shadle, 2023), both vowel nasalization and vowel quality are characterized by multiple spectral peaks (i.e., A1 and P0 for vowel nasalization, F1 and F2 for vowel quality), which may be sufficient for vowel-intrinsic perceptual normalization (Johnson, 2020; Syrdal & Gopal, 1986). That is, raw acoustic convergence may be observed when perceptual normalization requires reference to other segments and normalized phonetic convergence may be observed when perceptual normalization is segment-internal. Although highly speculative, this proposal has clear implications for our understanding of perceptual normalization processes (see Barreda, 2020; Johnson & Sjerps, 2021, for recent reviews) and highlights the need for further research to explore the potential for word shadowing data to provide empirical support for theoretical claims about perceptual normalization.

One alternative explanation that Hauser et al. (2023) considered for their results is that normalized phonetic convergence on the lower /s/ spectral mean, where raw acoustic convergence was observed, would potentially impinge on the /s ʃ/ contrast. Contrast maintenance cannot explain our data, however, because /æ/ raising endangers the contrast with /ɛ/ and /ɑ/ fronting endangers the contrast with /æ/. Similar to our results, Hauser et al.'s results can also not be explained in terms of spectral reduction from baseline to post-exposure, because some shadowers increased their spectral mean of /s/ following exposure, or in terms of convergence to a hyperarticulate style in general, because other shadowers decreased their spectral mean of /s/ following exposure. Kwon (2015) observed between-shadower variability in her study of f0 convergence in a shadowing task, with three of the 12 female shadowers producing raw acoustic convergence to the raised f0 of the male model talker, but the other nine female shadowers (and all seven male shadowers) showing normalized phonetic convergence. However, individual shadower variability seems unlikely to explain the overall differences between Hauser et al.'s results and ours. Indeed, the raw acoustic convergence on /s/ spectral mean that Hauser et al. observed remains an outlier in the literature and warrants further investigation. One potentially fruitful direction for further study would be a more careful consideration of the methods. Whereas Hauser et al. used a listening exposure task without shadowing and then a post-exposure reading task to assess convergence, Kwon and Zellou et al. (2016) both used a shadowing task followed by a post-exposure reading task and observed variability in the magnitude of convergence across the shadowing and post-exposure blocks. Moreover, all three studies used artificially manipulated stimulus materials, whereas our materials were all naturally produced. Thus, the details of the

methods might matter, with respect to either the task in which convergence is assessed (i.e., shadowing vs. post-exposure reading) or the materials (i.e., natural vs. manipulated) or both.

The observed shifts in production in the by-block analysis did not vary across instruction condition, suggesting that participants interpreted the instructions in the explicit imitation condition as being about normalized phonetic imitation rather than raw acoustic imitation, at least for the NCS features we examined. Given that our instructions were to repeat the words the way that the talker from Chicago, Illinois, said them, this outcome is perhaps not surprising. That is, a reasonable interpretation of our instructions would be to produce normalized phonetic imitations of the model talker's productions (i.e., "the way she said the words") or even normalized phonetic imitations of a Chicago accent more generally (i.e., "the way someone from Chicago would say the words"), and not raw acoustic imitations of the model talker's voice (i.e., "repeat the words while imitating her voice"). Alternatively, the explicit imitation instructions may have encouraged raw acoustic imitation of other properties of the model talker's voice that we did not examine (e.g., f0, voice quality, consonant quality, etc.), but not greater normalized phonetic imitation of the NCS. As Schertz et al. (2023) noted, the target of explicit imitation may vary across shadowers, depending on whether they interpret the goal as being about imitating the model talker's accent or their voice. Notably, the by-block analysis of vowel duration did not reveal any effects of instruction condition, comparable to the formant analyses, suggesting that the explicit instructions also did not enhance convergence on vowel duration. Further research is needed using a variety of "explicit" imitation instructions and a wider range of acoustic-phonetic measures to determine under which conditions participants adopt a more acoustically-based strategy vs. a more phonetically-based strategy of imitation.

The results of the linear combination analysis provide evidence of significant token-by-token phonetic convergence in the F1 × F2 vowel space, with a similar magnitude of convergence to F1 across vowels, shadower gender, and instruction condition, but variability in the magnitude of convergence to F2 across instruction conditions. Social information may have contributed to the modest effect of instruction condition on F2, in which more token-by-token phonetic convergence on F2 was observed in the primed condition than in the explicit condition. One possible explanation for this effect is that priming the Northern dialect in the instructions increased the shadowers' attention to the features of the NCS in the model talker's speech. Since the NCS is not negatively stereotyped, this increased attention could have led to greater token-by-token phonetic convergence, rather than reduced phonetic convergence as in the case of stereotyped variants (Babel, 2010; Mitterer & Müsseler, 2013; Walker & Campbell-Kibler, 2015). This interpretation is consistent with Wade et al.'s (2023) "expectation-driven convergence" (see also Wade, 2022), in which non-Southern participants produced more monophthongal /aɪ/, consistent with Southern American English, when they were told that the model talker was

Southern vs. when they were not. That is, the Southern label primed non-Southern participants to produce Southern features, similar to the increased token-by-token convergence observed in the primed condition in the current study.

In contrast, the weaker evidence of token-by-token phonetic convergence in the explicit instruction condition is more surprising, but, as in the results of the by-block analysis discussed above, may reflect how the shadowers interpreted the instructions. Specifically, instructions to repeat the words in the way the model talker said them may have encouraged a more global (i.e., across-trial) imitation strategy, rather than a more local trial-by-trial imitation strategy. Zellou et al. (2017) observed this kind of global imitation strategy for nasal coarticulation in a priming task, in which participants heard a word produced by the model talker (the prime) and then read aloud a different word (the target). Zellou et al.'s (2017) task therefore involved implicit imitation, whereas we observed global imitation in our explicit condition, where local trial-by-trial imitation might be predicted. As noted above, further research is needed using a variety of "explicit" imitation instructions to determine under which conditions participants adopt different acoustic vs. normalized phonetic imitation strategies. Further research is also needed to explore global vs. local imitation strategies, using a combination of analyses as in the current study. Overall, our results do not provide strong evidence for social constraints on normalized phonetic convergence to the NCS. This lack of social constraints on phonetic convergence was predicted, given that the NCS is not socially stereotyped in the American Midwest (Campbell-Kibler, 2012; Niedzielski, 1999).

Taken together, our results suggest a perception-production link that is phonetically-detailed in acoustic or articulatory space and mediated by normalized phonetic representations. With respect to phonetic detail, the linear combination models reveal a trial-by-trial link between variability in the model talker's productions and the shadowers' productions (see also Hauser et al., 2023; MacLeod, 2021; Tobin et al., 2018; Wade et al., 2020). This observation provides strong evidence for phonetic convergence to normalized phonetic detail in the F1 × F2 space. With respect to phonetic representations, the by-block models reveal normalized phonetic convergence to features of the NCS, rather than raw acoustic convergence to the model talker's high overall formant values. The convergence to the NCS is especially visible in the raising of /æ/ in the shadowing block in **Figure 3**. In addition, the small effect of instruction condition suggests that priming the model talker's regional background enhanced token-by-token phonetic convergence, although instruction condition had no effect on overall convergence patterns. This result is consistent with an effect of social knowledge or expectation on speech processing in general (Hay & Drager, 2010; Hay et al., 2006; McGowan, 2015; Niedzielski, 1999; Portes & German, 2019) and provides further, albeit modest, evidence for the role of social information in phonetic convergence (Babel, 2010; Clopper & Dossey, 2020; Ross et al., 2021; Walker &

Campbell-Kibler, 2015). Thus, as has long been noted (e.g., Babel & Bulotov, 2012; Shockley et al., 2004), phonetic convergence does not reflect direct acoustic imitation, but rather a perception-production link mediated by normalized phonetic representations. The perception-production link itself may involve gestures, as in motor theory (e.g., Fowler, 1996; Liberman & Mattingly, 1985), or a normalized mapping in an acoustic space, as in some exemplar theoretic proposals (e.g., Pierrehumbert, 2002). However, that link must also have access to phonological representations (Nielsen, 2011; Mitterer & Ernestus, 2008), as well as social information (Babel, 2010; Clopper & Dossey, 2020; Ross et al., 2021; Walker & Campbell-Kibler, 2015).

Data accessibility

The data and analysis code for this study are available on the Open Science Framework repository for this project: https://osf.io/ye7nw/.

Acknowledgments

We would like to thank Laura Beebe, Emily Behm, Sarah Chilson, Ben Elsbrock, Chance Goodwin, Conor Higgins, Rachel Monnin, and McKenna Dowdell for assistance with data collection and analysis.

Competing interests

The authors have no competing interests to declare.

Author contributions

Cynthia G. Clopper: Conceptualization, Experiment design, Data collection, analysis, visualization, and curation, Writing. Ellen Dossey: Conceptualization, Experiment design, Data collection, analysis, and visualization, Writing. Roberto Gonzalez: Data collection and analysis, Writing.

References

Adank, P., Smits, R., & van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America, 116*, 3099–3107. https://doi.org/10.1121/1.1795335

Babel, M. (2010). Dialect divergence and convergence in New Zealand English. *Language in Society, 39*, 437–456. https://doi.org/10.1017/S0047404510000400

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics, 40*, 117–189. https://doi.org/10.1016/j.wocn.2011.09.001

Babel, M., & Bulatov, D. (2012). The role of fundamental frequency in phonetic accommodation. *Language and Speech, 55*, 231–248. https://doi.org/10.1177/0023830911417695

Barreda, S. (2020). Vowel normalization as perceptual constancy. *Language, 96*, 224–254. https://doi.org/10.1353/lan.2020.0018

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. H. (2015). Parsimonious mixed models. arXiv:1506.04967.

Boersma, P., & Weenink, D. (2023). Praat: Doing phonetics by computer [Computer program]. Version 6.4.01. http://www.praat.org/

Campbell-Kibler, K. (2012). Contestation and enregisterment in Ohio's imagined dialects. *Journal of English Linguistics, 40*, 281–305. https://doi.org/10.1177/0075424211427911

Clopper, C. G., Burdin, R. S., & Turnbull, R. (2019). Variation in /u/ fronting in the American Midwest. *Journal of the Acoustical Society of America, 146*, 233–244. https://doi.org/10.1121/1.5116131

Clopper, C. G., Carter, A. K., Dillon, C. M., Hernandez, L. R., Pisoni, D. B., Clarke, C. M., Harnsberger, J. D., & Herman, R. (2002). The Indiana Speech Project: An overview of the development of a multi-talker multi-dialect speech corpus. *Research on Spoken Language Processing Progress Report No. 25* (pp. 367–380). Speech Research Laboratory, Indiana University.

Clopper, C. G., & Dossey, E. (2020). Phonetic convergence to Southern American English: Acoustics and perception. *Journal of the Acoustical Society of America, 147,* 671–683. https://doi.org/10.1121/10.0000555

Clopper, C. G., Pisoni, D. B., & de Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *Journal of the Acoustical Society of America, 118*, 1661–1676. https://doi.org/10.1121/1.2000774

Cohen Priva, U., & Sanker, C. (2019). Limitations of difference-in-difference for measuring convergence. *Laboratory Phonology, 10*(15), 1–29. https://doi.org/10.5334/labphon.200

Cole, J., & Shattuck-Hufnagel, S. (2011). The phonology and phonetics of perceived prosody: What do listeners imitate? *Proceedings of Interspeech 2011*. https://doi.org/10.21437/Interspeech.2011-395

Dailey, M. (2018). *Dialect classification and speech intelligibility in noise* [Unpublished bachelor's research thesis]. Ohio State University, Columbus.

Delaney, M., Savji, S., & Babel, M. (2010). An acoustic and auditory comparison of implicit and explicit phonetic imitation. *Canadian Acoustics, 38*, 132–133.

Dufour, S., & Nguyen, N. (2013). How much imitation is there in a shadowing task? *Frontiers in Psychology, 4*(346), 1–7. https://doi.org/10.3389/fpsyg.2013.00346

Edlund, J., Heldner, M., & Hirschberg, J. (2009). Pause and gap length in face-to-face interaction. *Proceedings of Interspeech 2009*, 2779–2782. https://doi.org/10.21437/Interspeech.2009-710

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America, 99*, 1730–1741. https://doi.org/10.1121/1.415237

Fowler, C. A., & Housum, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language, 26*, 489–504. https://doi.org/10.1016/0749-596X(87)90136-7

German, J. S., Carlson, K., & Pierrehumbert, J. B. (2013). Reassignment of consonant allophones in rapid dialect acquisition. *Journal of Phonetics, 41*, 228–248. https://doi.org/10.1016/j.wocn.2013.03.001

Gessinger, I., Raveh, E., Steiner, I., & Möbius, B. (2021). Phonetic accommodation to natural and synthetic voices: Behavior of groups and individuals in speech shadowing. *Speech Communication, 127*, 43–63. https://doi.org/10.1016/j.specom.2020.12.004

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251–279. https://doi.org/10.1037/0033-295X.105.2.251

Gonzalez, R., & Clopper, C. G. (2022). Phonetic convergence in temporal organization during shadowed speech. *Proceedings of Meetings on Acoustics, 45*(060012), 1–9. https://doi.org/10.1121/2.0001585

Hauser, I., Graham, E., & Zhang, X. (2023). Competing targets in English sibilant imitation. *JASA Express Letters, 3*(7), 075201. https://doi.org/10.1121/10.0019996

Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics, 48*, 865–892. https://doi.org/10.1515/ling.2010.027

Hay, J., Nolan, A., & Drager, K. (2006). From fush to feesh: Exemplar priming in speech perception. *Linguistic Review, 23*, 351–379. https://doi.org/10.1515/TLR.2006.014

Honorof, D. N., Weihing, J., & Fowler, C. A. (2011). Articulatory events are imitated under rapid shadowing. *Journal of Phonetics, 39*, 18–38. https://doi.org/10.1016/j.wocn.2010.10.007

Hoole, P., & Honda, K. (2011). Automaticity vs. feature-enhancement in the control of f0. In G. N. Clements & R. Ridouane (Eds.), *Where Do Phonological Features Come From? Cognitive, Physical and Developmental Bases of Speech Categories* (pp. 131–171). John Benjamins. https://doi.org/10.1075/lfab.6.06hoo

Johnson, K. (2020). The ΔF method of vocal tract length normalization for vowels. *Laboratory Phonology 11*(10), 1–16. https://doi.org/10.5334/labphon.196

Johnson, K., & Sjerps, M. J. (2021). Speaker normalization in speech perception. In J. S. Pardo, L. C. Nygaard, R. E. Remez & D. B. Pisoni (Eds.), *The Handbook of Speech Perception*, second edition (pp. 145–176). John Wiley. https://doi.org/10.1002/9781119184096.ch6

Jungers, M. K., & Hupp, J. M. (2009). Speech priming: Evidence for rate persistence in unscripted speech. *Language and Cognitive Processes, 24*, 611–624. https://doi.org/10.1080/01690960802602241

Kim, D., & Clayards, M. (2019). Individual differences in the link between perception and production and the mechanisms of phonetic imitation. *Language, Cognition and Neuroscience, 34*, 769–786. https://doi.org/10.1080/23273798.2019.1582787

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. https://doi.org/10.18637/jss.v082.i13

Kwon, H. (2015). Spontaneous speech imitation and cue primacy. *Proceedings of the 18th International Congress of Phonetic Sciences*.

Labov, W., Ash, S., & Boberg, C. (2006). *Atlas of North American English: Phonetics, phonology and sound change.* Mouton de Gruyter. https://doi.org/10.1515/9783110167467

Lee-Kim, S.-I., & Chou, Y.-C. (2024). Unmerging the sibilant merger via phonetic imitation: Phonetic, phonological, and social factors. *Journal of Phonetics, 103*(101298), 1–19. https://doi.org/10.1016/j.wocn.2024.101298

Lenth, R. (2024). emmeans: Estimated marginal means, aka least-squares means. R package version 1.10.0. https://CRAN.R-project.org/package=emmeans.

Lewandowski, E. M., & Nygaard, L. C. (2018). Vocal alignment to native and non-native speakers of English. *Journal of the Acoustical Society of America, 144*, 620–633. https://doi.org/10.1121/1.5038567

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1–36. https://doi.org/10.1016/0010-0277(85)90021-6

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *Journal of the Acoustical Society of America, 49*, 606–608. https://doi.org/10.1121/1.1912396

MacLeod, B. (2021). Problems in the difference-in-distance measure of phonetic imitation. *Journal of Phonetics, 87* (101058), 1–21. https://doi.org/10.1016/j.wocn.2021.101058

Marekova, L., Kruyt, J., & Beňuš, Š. (2023). The effect of (non-)native language and task complexity on speech entrainment. *Proceedings of the 20th International Congress of Phonetic Sciences*, 1548–1552.

McGowan, K. B. (2015). Social expectation improves speech perception in noise. *Language and Speech, 58*, 502–521. https://doi.org/10.1177/0023830914565191

Michelas, A., & Nguyen, N. (2011). Uncovering the effect of imitation on tonal patterns of French Accentual Phrases. *Proceedings of Interspeech 2011*. https://doi.org/10.21437/Interspeech.2011-396

Mitterer, H., & Ernestus, M. (2008). The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition, 109*, 168–173. https://doi.org/10.1016/j.cognition.2008.08.002

Mitterer, H., & Müsseler, J. (2013). Regional accent variation in the shadowing task: Evidence for a loose perception-action coupling in speech. *Attention, Perception, & Psychophysics, 75*, 557–575. https://doi.org/10.3758/s13414-012-0407-8

Moon, S.-J., & Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *Journal of the Acoustical Society of America, 96*, 40–55. https://doi.org/10.1121/1.410492

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics, 47*, 379–390. https://doi.org/10.3758/BF03210878

Nearey, T. M. (1978). *Phonetic feature systems for vowels*. Indiana University Linguistics Club.

Nguyen, N., Dufour, S., & Brunellière, A. (2012). Does imitation facilitate word recognition in a non-native regional accent? *Frontiers in Psychology, 3*(480), 1–7. https://doi.org/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*, 62–85. https://doi.org/10.1177/0261927X99018001005

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics, 39*, 132–142. https://doi.org/10.1016/j.wocn.2010.12.007

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. https://doi.org/10.1016/j.jml.2013.06.002

Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics, 79*, 637–659. https://doi.org/10.3758/s13414-016-1226-0

Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *Journal of the Acoustical Society of America, 32*, 693–703. https://doi.org/10.1121/1.1908183

Pierrehumbert, J. B. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology 7* (pp. 101–139). Mouton de Gruyter. https://doi.org/10.1515/9783110197105.1.101

Podlipský, V. J., & Šimáčková, S. (2015). Phonetic imitation is not conditioned by preservation of phonological contrast but by perceptual salience. *Proceedings of the 18th International Congress of Phonetic Sciences.*

Portes, C., & German, J. S. (2019). Implicit effects of regional cues on the interpretation of intonation by Corsican French listeners. *Laboratory Phonology, 10*(22), 1–26. https://doi.org/10.5334/labphon.162

Ross, J. P., Lilley, K. D., Clopper, C. G., Pardo, J. S., & Levi, S. V. (2021). Effects of dialect-specific features and familiarity on cross-dialect phonetic convergence. *Journal of Phonetics, 86*(101041), 1–23. https://doi.org/10.1016/j.wocn.2021.101041

Sanker, C. (2015). Comparison of phonetic convergence in multiple measures. In *Cornell working papers in phonetics and phonology* (pp. 60–75). Department of Linguistics, Cornell University. https://doi.org/10.5281/zenodo.3726190

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), 1–14. https://doi.org/10.3389/fpsyg.2013.00422

Schertz, J., Adil, F., & Kravchuk, A. (2023). Underpinnings of explicit phonetic imitation: Perception, production, and variability. *Glossa Psycholinguistics, 2*(4), 1–51. https://doi.org/10.5070/G601123

Shadle, C. H. (2023). Alternatives to moments for characterizing fricatives: Reconsidering Forrest et al. (1988). *Journal of the Acoustical Society of America, 153*, 1412–1426. https://doi.org/10.1121/10.0017231

Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics, 66*, 422–429. https://doi.org/10.3758/BF03194890

Song, Y. J., & Clopper, C. G. (2023). Implicit imitation of intonation contours in word shadowing. *Proceedings of the 20th International Congress of Phonetic Sciences,* 1365–1369.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America, 79*, 1086–1100. https://doi.org/10.1121/1.393381

Tobin, S. (2022). Effects of native language and habituation in phonetic accommodation. *Journal of Phonetics, 93*(101148), 1–18. https://doi.org/10.1016/j.wocn.2022.101148

Tobin, S., Hullebus, M., & Gafos, A. (2018). Immediate phonetic convergence in a cue-distractor paradigm. *Journal of the Acoustical Society of America, 144*, EL528–EL534. https://doi.org/10.1121/1.5082984

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *Journal of the Acoustical Society of America, 88*, 97–100. https://doi.org/10.1121/1.399849

Wade, L. (2022). Experimental evidence for expectation-driven linguistic convergence. *Language, 98*, 63–97. https://doi.org/10.1353/lan.2021.0086

Wade, L., Embick, D., & Tamminga, M. (2023). Dialect experience modulates cue reliance in sociolinguistic convergence. *Glossa Psycholinguistics, 2*(19), 1–30. https://doi.org/10.5070/G6011187

Wade, L., Lai, W., & Tamminga, M. (2020). The reliability of individual differences in VOT imitation. *Language and Speech, 64*, 576–593. https://doi.org/10.1177/0023830920947769

Walker, A., & Campbell-Kibler, K. (2015). Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task. *Frontiers in Psychology, 6*(546), 1–18. https://doi.org/10.3389/fpsyg.2015.00546

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

Zellou, G., Dahan, D., & Embick, D. (2017). Imitation of coarticulatory vowel nasality across words and time. *Language, Cognition and Neuroscience, 32*, 776–791. https://doi.org/10.1080/23273798.2016.1275710

Zellou, G., Scarborough, R., & Nielsen, K. (2016). Phonetic imitation of coarticulatory vowel nasalization. *Journal of the Acoustical Society of America, 140*, 3560–3575. https://doi.org/10.1121/1.4966232