Atten Percept Psychophys (2017) 79:637–659
DOI 10.3758/s13414-016-1226-0

Phonetic convergence across multiple measures and model talkers

Jennifer S. Pardo[1] · Adelya Urmanche[1] · Sherilyn Wilman[1] · Jaclyn Wiener[1]

Published online: 8 November 2016

Abstract This study consolidates findings on phonetic convergence in a large-scale examination of the impacts of talker sex, word frequency, and model talkers on multiple measures of convergence. A survey of nearly three dozen published reports revealed that most shadowing studies used very few model talkers and did not assess whether phonetic convergence varied across same- and mixed-sex pairings. Furthermore, some studies have reported effects of talker sex or word frequency on phonetic convergence, but others have failed to replicate these effects or have reported opposing patterns. In the present study, a set of 92 talkers (47 female) shadowed either same-sex or opposite-sex models (12 talkers, six female). Phonetic convergence was assessed in a holistic AXB perceptual-similarity task and in acoustic measures of duration, F0, F1, F2, and the F1 × F2 vowel space. Across these measures, convergence was subtle, variable, and inconsistent. There were no reliable main effects of talker sex or word frequency on any measures. However, female shadowers were more susceptible to lexical properties than were males, and model talkers elicited varying degrees of phonetic convergence. Mixed-effects regression models confirmed the complex relationships between acoustic and holistic perceptual measures of phonetic convergence. In order to draw broad conclusions about phonetic convergence, studies should employ multiple models and shadowers (both male and female), balanced multisyllabic items, and holistic measures. As a potential mechanism for sound change, phonetic convergence reflects complexities in speech perception and production that warrant elaboration of the underspecified components of current accounts.

Keywords Phonetic convergence · Speech imitation · Speech production · Speech perception

Phonetic convergence is emerging as a prominent phenomenon in many accounts of spoken communication. This tendency for individuals' speech to increase in similarity is also referred to as *speech imitation*, *accommodation*, *entrainment*, or *alignment*, and occurs across multiple settings of language use. From its beginnings in the literature on speech accommodation (see Giles, Coupland, & Coupland, 1991; Shepard, Giles, & Le Poire, 2001), to its adoption by the psycholinguistic literature (Goldinger, 1998; Namy, Nygaard, & Sauerteig, 2002), the phenomenon has informed a broad array of theories of social interaction and cognitive systems. Within psycholinguistics, studies of phonetic convergence have addressed questions involving speech perception (e.g., Fowler, Brown, Sabadini, & Weihing, 2003), phonological representation (e.g., Mitterer & Ernestus, 2008), memory systems (e.g., Goldinger, 1998), individual differences (Yu, Abrego-Collier, & Sonderegger, 2013), talker sex (e.g., Namy et al., 2002), conversational interaction (e.g., Pardo, 2006), sound change (e.g., Delvaux & Soquet, 2007), and neurolinguistics (e.g., Garnier, Lamalle, & Sato, 2013). This study aims to consolidate findings in the literature on phonetic convergence, to evaluate methodological practices, and to inform accounts of language comprehension and production.

Howard Giles's *communication accommodation theory* (CAT; Giles et al., 1991; Shepard et al., 2001) centers on external factors driving patterns of variation in speech produced in social interaction. Early investigations of convergence examined multiple attributes, such as perceived

✉ Jennifer S. Pardo
pardoj@montclair.edu

[1] Psychology Department, Montclair State University, 1 Normal Ave., Montclair, NJ 07043, USA

accentedness, phonological variants, speaking rate, and various acoustic measures (e.g., Coupland, 1984; Giles, 1973; Gregory & Webster, 1996; Natale, 1975; Putman & Street, 1984). Convergence in such parameters appears to be influenced by social factors that are local to communication exchanges, such as interlocutors' relative dominance or perceived prestige (Gregory, Dagan, & Webster, 1997; Gregory & Webster, 1996). Giles and colleagues also acknowledged the opposite pattern, accent divergence, under some circumstances (Bourhis & Giles, 1977; Giles, Bourhis, & Taylor, 1977; Giles et al., 1991; Shepard et al., 2001).

One explanation offered for accommodation is the similarity attraction hypothesis, which claims that individuals try to be more similar to those to whom they are attracted (Byrne, 1971). Accordingly, convergence arises from a need to gain approval from an interacting partner (Street, 1982) and/or from a desire to ensure smooth conversational interaction (Gallois, Giles, Jones, Cargiles, & Ota, 1995). Divergence is often interpreted as a means to accentuate individual/cultural differences or to display disdain (Bourhis & Giles, 1977; Shepard et al., 2001). Interlocutors also converge or diverge along different speech dimensions as a function of their relative status or dominance in an interaction (Giles et al., 1991; Jones, Gallois, Callan, & Barker, 1999; Shepard et al., 2001), which is compatible with the similarity attraction hypothesis. Typically, a talker in a less dominant role will converge toward a more dominant partner's speaking style (Giles, 1973). Finally, interacting talkers have been found to converge on some parameters while simultaneously diverging on others (Bilous & Krauss, 1988; Pardo, Cajori Jay, & Krauss, 2010; Pardo, Cajori Jay, et al., 2013).

It is clear that aspects of a social/cultural setting and relationships between interlocutors influence the form and direction of communication accommodation. However, research within the accommodation framework is mute regarding cognitive mechanisms that support convergence and divergence during speech production. To support phonetic accommodation in production, speech perception must resolve phonetic form in sufficient detail, and detailed phonetic form must persist in memory. Fowler's direct realist theory of speech perception asserts that individuals directly perceive linguistically significant vocal tract actions, or phonetic gestures (e.g., Fowler, 1986, 2014; Fowler et al., 2003; Fowler, Shankweiler, & Studdert-Kennedy, 2016; Goldstein & Fowler, 2003; Shockley et al., 2004). The motor theory of speech perception and Pickering and Garrod's interactive-alignment account both claim that speech perception processes recruit speech production processes to yield resolution of motor commands (e.g., Liberman, 1996; Pickering & Garrod, 2004, 2013). Pickering and Garrod (2013) further claimed explicitly that motor commands derived during language comprehension can lead to imitation in production during dialogue. Despite important differences across these approaches, they all propose that speech perception involves the resolution of phonetic form from vocal tract activity. Once phonetic form has been perceived, processes entailed in episodic memory systems maintain the persistence of phonetic details, lending further support for phonetic convergence in production (e.g., Goldinger, 1998; Hintzman, 1984; Johnson, 2007; Pierrehumbert, 2006, 2012). Therefore, accounts of both perception and memory systems have centered on the processes that support and/or predict phonetic convergence in speech production.

Speech perception and production

In their mechanistic approach, Pickering and Garrod (2004) proposed a model of language use in dialogue based on a simple idea. That is, automatic priming of shared representations leads to alignment at all levels of language—semantic, syntactic, and phonological. Moreover, alignment at one level promotes alignment at other levels. On those occasions when the default automatic priming mechanism fails to yield schematic alignment (e.g., during a misunderstanding), a second, more deliberate mechanism brings interlocutors into alignment. Pickering and Garrod supported their proposal of an automatic priming mechanism by citing evidence for between-talker alignment (i.e., convergence) in semantic components (e.g., Brennan & Clark, 1996; Wilkes-Gibbs & Clark, 1992), syntax (e.g., Branigan, Pickering, & Cleland, 2000; Branigan, Pickering, McLean, & Cleland, 2007), and phonetic form (e.g., Pardo, 2006).

In a more recent article, Pickering and Garrod (2013) drew out a critical component of their interactive-alignment account—that language comprehension and production processes are integrated within talkers due to a covert imitation process that generates inverse forward modeling of speech production commands during language comprehension. Accordingly, when a listener hears an utterance, comprehension relies on the same processes as production, leading to convergent production in a very straightforward manner (see also Gambi & Pickering, 2013). In their account, this so-called *simulation route* in action perception contrasts with an *association route*, which relies on past experiences in perceiving rather than producing utterances. Importantly, both versions of this account do not elaborate on the so-called secondary processes (deliberate alignment or association route) that also regulate linguistic form in speech production.

Whereas the interactive-alignment account explicitly predicts phonetic convergence, the *motor theory of speech perception* and Fowler's *direct-realist approach* provide indirect support for phonetic convergence in production. According to the motor theory, speech perception recruits the motor system to perform an analysis by synthesis that recovers a talker's intended gestures from coarticulated acoustic output (Galantucci, Fowler, & Turvey, 2006; Liberman & Mattingly, 1985). Fowler's direct-realist perspective departs from these accounts by not invoking use of the motor system (Fowler, 1986; Fowler & Galantucci, 2005; Goldstein & Fowler, 2003). Rather, speech perception and production are viewed as using a common currency— linguistically significant vocal tract actions. Although direct realism is not an account of speech production, Fowler has repeatedly proposed that the perception of speech directly and rapidly yields the same vocal tract gestures that are used when producing speech, effectively goading imitation (Fowler, 1986; Fowler et al., 2003; Sancier & Fowler, 1997). In contrast with Pickering and Garrod's integrated account of communication in dialogue, the motor theory and direct realism were developed mainly to account for speech perception, and Fowler has pointed out that phonetic forms serve multiple roles of specifying linguistic tokens, contributing to interpersonal coordination, and expressing social identity (Fowler, 1986, 2010, 2014). Taken together, accounts of language comprehension provide support for and predict phonetic convergence, within limits set by other factors that are outside their scope. Some of those factors are likely due to processes related to the persistence of phonetic detail in memory.

Phonetic detail in episodic memory

Models of memory systems often differ with respect to the level of abstraction of representations in memory stores (e.g., Hintzman, 1986; Posner, 1964). Abstraction in memory systems entails processes that normalize variable phonetic tokens to match phonological types, presumably facilitating lexical identification (Goldinger, 1996). For example, between-talker variation in pronunciation of the vowel in the word *pin* should be removed when determining word identity, according to an abstractionist account. However, many studies of memory have shown that changing the voice of a word or increasing the talker set size affects both implicit and explicit word memory (Goldinger, Pisoni, & Logan, 1991; Martin, Mullennix, Pisoni, & Summers, 1989; Nygaard, Sommers, & Pisoni, 1995; Palmeri, Goldinger, & Pisoni, 1993; Sommers, Nygaard, & Pisoni, 1994), as well as speech perception and production (Goldinger, 1998). Therefore, detailed talker information is not normalized away during memory encoding. It is likely that such effects are driven by the integration of perceptual processes that identify both linguistic and indexical properties of talkers (Mullennix & Pisoni, 1990). It appears that talker-related details affect speech perception, persist in memory, and could support convergence in production.

In a seminal study, Goldinger (1998) examined whether talker-specific phonetic details persist in memory to support convergent production in listeners who shadow speech, and whether a prominent episodic memory model could predict the observed patterns of phonetic convergence. In Goldinger's use of speech shadowing, a talker first produced baseline utterances prompted by text and then produced shadowed utterances prompted by audio recordings (also known as an auditory-naming task). In order to examine specific predictions from an exemplar-based episodic memory model (Hintzman, 1986), the study design manipulated the frequency of items presented to shadowers (using estimates of real-world exposure to words and direct manipulation of nonword frequency) and local task repetition (presenting items one or more times in an exposure phase prior to eliciting a shadowed utterance).

According to this episodic memory model, each encounter with a word leaves a trace, and words that are encountered more frequently result in more traces. When a listener hears a new version of a word, all similar traces are activated and averaged along with the recently heard version of the word, to generate an echo that forms the basis for recognition and (presumably) subsequent production. Echoes of high-frequency words reflect fewer idiosyncratic details of a recently heard version, thereby reducing their availability relative to lower-frequency words. Thus, exemplars of high-frequency words effectively drown out idiosyncratic details of each new exemplar. A series of experiments and modeling simulations demonstrated that shadowers converged to model talker utterances in production and verified the episodic memory model's predictions of convergence patterns. That is, talkers converged more to low-frequency items and to items that were repeated more times in the task. A follow-up study replicated the word frequency effects and demonstrated that idiosyncratic details supported convergence up to a week after exposure (Goldinger & Azuma, 2004). Therefore, speech perception and production support resolution of phonetic detail, which is encoded into exemplar-based memory systems, leading to phonetic convergence under some circumstances.

Phonetic convergence in speech shadowing tasks

A review of the literature on phonetic convergence reveals that many potential sources influence phonetic form in speech production. Table 1 presents an analysis of methods employed in nearly three dozen published studies that have used shadowing or exposure tasks to assess phonetic convergence. Due to dramatic differences in purposes and methodologies that warrant a separate analysis, the table does not include studies that have examined convergence during conversational interaction (e.g., Abney, Paxton, Dale, & Kello, 2014; Aguilar et al., 2016; Dias & Rosenblum, 2011; Fusaroli & Tylén, 2016; Heldner, Edlund, & Hirschberg, 2010; Kim, Horton, & Bradlow, 2011; Levitan, Benus, Gravano, & Hirschberg, 2015;

Table 1 Summary of noninteractive shadowing/exposure studies of phonetic convergence

Year	Authors	Models		Shadowers		Items		Measures
		Female	Male	Female	Male	Mono	Multi	
1998	Goldinger	5	5	12	12	80	80	AXB
2002	Namy et al.	2	2	8	8		20 LF	AXB
2003	Nye & Fowler, Exp. 2		2	?/2	?/2		12	AXB
2004	Goldinger & Azuma (expo)	2	2	6	6		160	AXB
2004	Shockley et al.	1	1	12	12		80 LF	AXB, VOT ptk
2010	Miller et al.	1	1	8	8		74 LF	AXB
2012	Babel & Bulatov		1	12	7	15	24	AXB, F0
2013	Babel et al.		1	33	8	18		AXB, vowel
2013	Miller et al.	2	2	8	8		74 LF	AXB
2013	Pardo, Jordan, et al.	10	10	10	10	80		AXB, vowel F1F2, Dur, F0
2014	Babel et al.	4	4	10	10	15 LF		AXB
2015	Walker & Campbell-Kibler	4		36		70		AXB, vowel F1F2, rF3
2016	Dias & Rosenblum	1		32 & 24			120	AXB
2004	Vallabha & Tuller		3		3	Vs		vowel F1F2
2007	Delvaux & Soquet (expo)	5		12		3	1	vowel F1–F3
2007	Gentilucci & Bernardis, Exp. 2	2	2	10			/aba/	vowel F1F2, lip aperture, F0, dur, intensity
2009	Tilsen		1	6	6	Vs		vowel F1F2
2010	Babel		1	34	8	25		vowel F1F2 DID
2012	Babel		2	60	51	50 LF		vowel F1F2 DID
2012	Nguyen et al.		1	33	9	40		vowel F1 o aw
2013	Dufour & Nguyen	1		16	4		66	vowel F1 eE French
2003	Fowler et al., Exp. 4	1		?/24	?/24		48 VCVs	VOT ptk
2010	Sanchez et al.	1		35		6 CVs		VOT p
2011	Abrego-Collier et al. (expo)		1	?/48	?/48	17	55	VOT ptk
2011	Nielsen, Exp. 1 (expo)		1	?/27	?/27	?/120	?/120	VOT pk
2013	Olmstead et al.	1		20	12	11 CVs		VOT b-p
2013	Yu et al. (expo)		1	?/84	?/84	17	55	VOT ptk
2013	Garnier et al.	1	1	4	11	Vs		F0
2013	Mantell & Pfordresher	1	1	69	86		12	F0
2013	Postma-Nilsenová & Postma	2	2	67	21		16	F0
2013	Sato et al.	3	3	30	30	Vs		F0, F1 ieE
2013	Wisniewski et al.	1	1	8	8		12	F0
2008	Mitterer & Ernestus	1		?/18	?/18		28 CVVC	phonemic/r/allophones Dutch
2011	Honorof et al.		1	?/37	?/37		4 VCVs	phonemic/l/allophones English
2013	Mitterer & Müsseler	1		9	3		100	phonemic allophone pairs German

Studies in the table are grouped according to the measures used to assess phonetic convergence (last column)

Levitan & Hirschberg, 2011; Louwerse, Dale, Bard, & Jeuniaux, 2012; Pardo, 2006; Pardo, Cajori Jay, et al., 2013; Pardo, Cajori Jay, & Krauss, 2010; Paxton & Dale, 2013) and under conditions related to longer-term exposure to other talkers, to second language training, or to different linguistic environments (e.g., Chang, 2012; Evans & Iverson, 2007; Harrington, 2006; Harrington, Palethorpe, & Watson, 2000; Pardo, Gibbons, Suppes, & Krauss, 2012; Sancier & Fowler, 1997). Arguably, laboratory speech-shadowing tasks provide a favorable context to elicit phonetic convergence and assess its basic properties (i.e., without interference from conversational goals).

In the shadowing/exposure studies cataloged in Table 1, model talkers provided utterances that were presented to shadowers in either immediate shadowing tasks or in an exposure session with post-listening utterance production (marked "expo"). Individual studies appear in separate rows referenced by year of publication and authors. The next columns display the numbers of model talkers and shadowers employed in each study (each split by sex), and the penultimate column indicates the kinds of items used in each study (mono- vs. multisyllabic). Studies are grouped in the table according to the measures used to assess phonetic convergence (indicated in the last column)—AXB perceptual similarity tests, vowel spectra (F1, F2), VOT, F0, and particular phonemic variants. Although some have employed a holistic AXB perceptual-similarity task to assess phonetic convergence, the majority of studies have focused on specific acoustic–phonetic attributes (22 of 35 studies). Some studies have examined multiple measures, but most have focused on a single measure (20 of 35 studies).

Goldinger (1998) introduced an important adaptation of a classic AXB perceptual-similarity task to assess phonetic convergence. If a talker exhibits phonetic convergence, then utterances produced after hearing a model talker's utterances (either immediately shadowed or postexposure) should sound more similar in pronunciation to model utterances than those produced prior to hearing them (pre-exposure baseline). According to this logic, an AXB similarity task for assessing phonetic convergence involves comparing similarity of baseline utterances and shadowed/post-exposure utterances of shadowers (A/B) to model talker (X) utterances. On each trial, a listener hears three versions of the same item and decides whether the first or the last item (A/B) sounds more similar to the middle item (X) in pronunciation. Although Goldinger originally instructed listeners to judge imitation, most studies have asked listeners to judge similarity or similarity in pronunciation (Pardo et al., 2010, found no differences whether listeners judged imitation or similarity in pronunciation). Responses are then scored as proportion or percentage of shadowed/post-exposure items selected as more similar to model items than baseline items. Because this measure relies on perceptual similarity, it constitutes a holistic assessment of phonetic convergence that is sensitive to multiple acoustic attributes in parallel (Pardo & Remez, 2006). Holistic AXB assessment is useful for drawing broad conclusions regarding phonetic convergence, because it is not restricted to idiosyncratic patterns of convergence on individual acoustic attributes (see Pardo, Jordan, Mallari, Scanlon, & Lewandowski, 2013).

Examination of the table reveals that most of these studies employed very few model talkers—in 16 out of 35 studies, only a single female or male model talker's utterances were used to elicit shadowed utterances, 23 studies used two or fewer model talkers, and only five studies used more than four model talkers. Moreover, the number and balance of shadowing talkers used across studies have varied enormously (in some cases, the sex of the talkers was not reported, and these are marked with ?s in the table). Apart from limited generalizability, a potential issue with this practice is that differences across studies could be driven by differences in the degrees to which individual model talkers evoke phonetic convergence. As described below, there is some controversy over whether males or females are more likely to converge, as well as a potential for idiosyncratic effects related to model talkers. The current study examines these possibilities by employing a relatively large set of model talkers (12: six female), who were each shadowed by multiple talkers in same- and mixed-sex pairings.

Effects of word frequency and talker sex

Two factors found to influence phonetic convergence in initial reports have become lore in the field by virtue of repeated citation (albeit inconsistent replication): (1) that low-frequency words evoke greater convergence than high-frequency words, and (2) that female talkers converge more than males. Recall that Goldinger (1998) found that low-frequency words elicited greater phonetic convergence than high-frequency words, which was replicated in Goldinger and Azuma (2004). Largely as a result of the original finding, at least six studies have restricted their items to low-frequency words (Babel, 2012; Babel, McGuire, Walters, & Nicholls, 2014; Miller, Sanchez, & Rosenblum, 2010, 2013; Namy et al., 2002; Shockley, Sabadini, & Fowler, 2004). Three other studies have reported frequency effects on convergence (in voice onset time [VOT]: Nielsen, 2011; in vowel formants: Babel, 2010; in AXB: Dias & Rosenblum, 2016), but each of these studies used just one model talker who was shadowed by all or predominantly female listeners. Another study that used a much larger set of model talkers (20: ten female) and equal numbers of male and female listeners (ten each) in same-sex pairings failed to replicate frequency effects (in AXB: Pardo, Jordan, et al., 2013). The current study attempts another replication in an even more powerful design using Goldinger's bisyllabic word set.

Talker sex effects have an analogous treatment in the literature on phonetic convergence. In social settings, females might converge more due to a greater affiliative strategy (Giles, Coupland, & Coupland, 1991), and previous research has shown that women were more sensitive than men to indexical information in a nonsocial voice identification learning paradigm (Nygaard & Queen, 2000). A study by Namy et al. (2002) is frequently cited in support of the assertion that female talkers converge more than males. However, an examination of the method and findings of this study reveals that the reported effect cannot bear the weight of such a decisive conclusion. Indeed, Namy et al. acknowledged the limitations of their study, pointing out that the effect was completely driven by convergence of female shadowers to a single male model talker. Often overlooked is the fact that shadowers of both sexes converged at equivalent levels to the other three models in the study. Because the study used just 16 shadowers and four models, it should be replicated in a larger set of talkers. Moreover, similarly limited studies by Pardo (Pardo, 2006, using only 12 talkers; and Pardo et al., 2010, using 24 talkers) showed that males converged more than females. However, these studies were not exactly comparable, because Namy et al. used a shadowing task and Pardo's studies examined conversational interaction.

More recently, one study reported a marginally significant tendency for female shadowers to converge more than males (with 16 talkers shadowing two models in same-sex pairs), and the pattern was not replicated in a second experiment (Miller et al., 2010). Another study showed that females converged more than males and were more susceptible to differences in the vocal attractiveness and gender typicality of individual model talkers (Babel et al., 2014). It is noteworthy that all three studies reporting greater convergence of female talkers used only low-frequency words, limiting the generalizability of the finding. Instead of positing that females converge more than males only on low-frequency words, it is more likely the case that these weak and inconsistent effects of sex reflect limitations of the study designs, in terms of the item sets, numbers of model talkers, and numbers of listener/shadowers. In a recent shadowing study with a balanced word set, Pardo, Jordan, et al. (2013) failed to find sex effects on phonetic convergence. Although the original finding has been largely untested across the literature on phonetic convergence, a few studies have limited their talker sets to females as a result (Delvaux & Soquet, 2007; Dias & Rosenblum, 2016; Gentilucci & Bernardis, 2007; Sanchez, Miller, & Rosenblum, 2010; Walker & Campbell-Kibler, 2015).

In a recent study on phonetic convergence in shadowed speech, Pardo, Jordan, et al. (2013) examined talker sex and word frequency effects in a set of 20 talkers who shadowed 20 models (in same-sex pairings with one shadower/model). The measures of phonetic convergence included holistic AXB perceptual similarity, vowel spectra, F0, and vowel duration. Monosyllabic items differed in both word frequency and neighbor frequency-weighted density. Shadowers converged to their models overall (AXB $M = .58$), and convergence was not modulated by talker sex or lexical properties. Moreover, convergence was only reliable in the holistic AXB measure—no acoustic measure reached significance on its own. Despite their failure on average, mixed-effects regression modeling confirmed that variability in the convergence of multiple acoustic attributes predicted patterns of convergence in holistic AXB convergence. That is, listeners' judgments of greater similarity in pronunciation of shadowed items to model items were predicted by variation in the degrees of convergence across multiple acoustic attributes. The strongest predictor was duration, followed by F0 and vowel spectra.

Pardo, Jordan, et al.'s (2013) study was the first of its kind to directly relate convergence in multiple acoustic measures to a holistic assessment of convergence, developing a novel paradigm for examining phonetic convergence. As summarized earlier, many explanations of phonetic convergence focus on its role in promoting social interaction by reducing social distance or increasing liking of a conversational partner. Although it is often useful to examine convergence in an individual acoustic parameter when assessing questions related to specific accents or attributes of sound change (e.g., Babel, 2010; Babel, McAuliffe, & Haber, 2013; Delvaux & Soquet, 2007; Dufour & Nguyen, 2013; Mitterer & Ernestus, 2008; Mitterer & Müsseler, 2013; Olmstead, Viswanathan, Aivar, & Manuel, 2013; Nguyen, Dufour, & Brunellière, 2012; Walker & Campbell-Kibler, 2015), assessments of a single acoustic attribute are limited with respect to broader interpretations of the phenomenon. For example, studies of convergence in VOT have often reported small changes toward a model's extended VOT values (usually around 10 ms or less; Fowler et al., 2003; Nielsen, 2011; Sanchez, Miller, & Rosenblum, 2010; Shockley, Sabadini, & Fowler, 2004; Yu, Abrego-Collier, & Sonderegger, 2013). Although these effects were statistically reliable, it is unknown whether these small changes would be perceptible by listeners, and so could play a role in social interaction (note that Sancier & Fowler, 1997, reported that their talker's changes were detected as greater accentedness in sentence-length utterances, and these judgments were likely based on more than VOT alone). Moreover, it is increasingly apparent that talkers vary which attributes and how much to converge on an item-by-item basis (Pardo, Jordan, et al., 2013). A more comprehensive assessment emerges by relating patterns of convergence in acoustic measures to holistic perceived phonetic convergence. This paradigm can harness the inevitable variability across multiple attributes in parallel by evaluating the relative weight of each acoustic attribute's contribution toward holistically perceived convergence.

The current study examined the impacts of talker sex and lexical properties on phonetic convergence in a comprehensive set of model talkers, shadowers, and items. To assess the effects of talker sex, this study recruited a relatively large set of male and female model talkers (12: six female, six male), who were shadowed by multiple talkers in balanced same- and mixed-sex pairings (32 same-sex female, 30 same-sex male, and 30 mixed-sex shadowers). Previous studies have mostly employed same-sex pairings, when possible, but none have explicitly examined whether convergence differs in same- versus mixed-sex pairings in a study of this scope. Furthermore, by using multiple model talkers in a balanced design, it was possible to examine whether individual models evoked distinct patterns of phonetic convergence.

To be more directly comparable to the previous studies that reported word frequency effects, this study included Goldinger's (1998) bisyllabic item set along with the monosyllabic items used in Pardo, Jordan, et al. (2013). Although Pardo, Jordan, et al. replicated Munson and Solomon's (2004) finding that lexical properties influenced speech production, such that low-frequency words were produced with more dispersed vowels than were high-frequency words, the varied productions elicited equivalent degrees of phonetic convergence. However, it is possible that frequency effects in phonetic convergence would be more apparent in bisyllabic words, because they are longer in duration and comprise more opportunities for convergence. Therefore, in addition to word frequency, this study also explored a possible influence of word type (mono- vs. bisyllabic) on phonetic convergence. Finally, convergence in acoustic attributes of monosyllabic items was assessed and compared to convergence in holistic AXB perceptual convergence using mixed-effects regression modeling.

Method

Participants

Talkers A total of 108 talkers (54 female) were recruited from the Montclair State University student population to provide speech recordings. All talkers were native English speakers reporting normal hearing and speech, and were paid $10 for their participation. The full set of talkers was split into two groups—one set of 12 (six female) who provided model utterances, and a second set of 96 (48 female) who provided baseline and shadowed utterances in random same- and mixed-sex pairings with model talkers (32 female, 32 male, and 32 mixed). Three of the recruited shadowers failed to keep their recording appointments, and one shadower's recording was unusable due to extremely rushed and atypical utterances. Thus, the study employed a total of 92 (47 female) shadowers in 32 same-sex female, 30 same-sex male, and 30 mixed-sex pairings with their models. Most of the models (eight) were shadowed by eight talkers, and other models were shadowed by nine (one model), seven (one model), or six (two models) talkers. All of the model talkers and most of the shadowers were from New Jersey (*N* = 89), with others from Montana, Puerto Rico, and Jamaica. All talkers had resided in New Jersey for at least 3 years prior to completing the study.

Listeners A total of 736 listeners were recruited from the Montclair State University student population to participate in AXB perceptual similarity tests. All of the listeners were native English speakers reporting normal hearing and speech and were either paid $10 or received course credit for their participation.

Materials

To assess the impact of lexical properties on phonetic convergence, the word set comprised both mono- and bisyllabic words, which were each evenly split into high- and low-frequency sets. Monosyllabic words were taken from the consonant–vowel–consonant (CVC) word set developed by Munson and Solomon (2004, Exp. 2). This set was chosen because it sampled evenly across the vowel space (with frequency manipulated within vowels), permitting measures of vowel spectra and other acoustic attributes. Bisyllabic words were taken from the set developed by Goldinger (1998), which was the first study to report word frequency effects on phonetic convergence. Both sets comprised 40 words each in the high- and low-frequency groups, for a total of 160 words. In the Munson and Solomon word set, the high-frequency words averaged 148 (*SD* = 157; 20–750) and the low-frequency words averaged 6.8 (*SD* = 5.2; 1–17) uses/million. In Goldinger's bisyllabic words, the high-frequency words averaged 329 (*SD* = 200; 155–1,016) and the low-frequency words averaged 34 (*SD* = 34; 1–90) uses/million (Kučera & Francis, 1967). Thus, the Munson and Solomon set comprised lower-frequency items overall, and their distribution of high-frequency items partially overlapped in frequency with Goldinger's low-frequency items. Moreover, the frequency manipulation was stronger in the Goldinger bisyllabic word set. Comparisons across the two word sets will take these differences into consideration. The full set of words appears in Appendix A.

Procedures

For all recordings, each talker sat in an Acoustic Systems sound booth in front of a Macintosh computer presenting prompts via SuperLab 4.5 (Cedrus). Talkers wore Sennheiser HMD280 headsets, and recordings were digitized at a rate of 44.1 kHz at 16 bits on a separate iMac computer running outside the booth. Words were spliced into individual files and normalized to 80% of maximum peak intensity prior to all analyses using the Normalize function in SoundStudio (Felt Tip, Inc.) to equate for differences in amplitude across items that arise due to differences in recording conditions, list position, microphone distance, etc. All listening tests were presented over Sennheiser Pro headphones in quiet testing rooms, via SuperLab 4.5 (Cedrus) running on either Dell or iMac computers.

Model utterances A set of 12 talkers (six female) provided model recordings of all 160 words in three randomized blocks. Instructions directed talkers to say each word as quickly and as clearly as possible. Words appeared individually in print on the computer monitor and remained until the software detected speech. Items from the second iteration of the list were used to compose a set of auditory prompts for the shadowing session. This selection criterion ensured that items were not subject to potential lengthening effects of first mentions (Bard et al., 2000; Fowler & Housum, 1987). Very few errors were produced, and items from the third iteration of the list were sampled to fill in missing items.

Shadower utterances To assess the impact of talker sex on phonetic convergence, a total of 92 talkers (47 female) provided baseline and shadowed recordings of the word set in 32 same-sex female, 30 same-sex male, and 30 mixed-sex pairings. In two baseline blocks, words appeared individually in print on the computer monitor and remained until the software detected speech. In two subsequent shadowing blocks, utterances of words from a single model talker were randomly presented over headphones (nothing appeared on the screen). The instructions directed talkers to say each word as quickly and as clearly as possible, and they produced the word list four times in randomized blocks: twice for baseline recordings, followed by twice in the shadowing condition. Shadowers were given the same instructions for both the baseline and shadowed recordings—they were told that the words in the last two blocks would be presented through headphones instead of on the computer screen. The set of baseline items sampled words from the second iteration of the list, and the shadowed items sampled from the fourth iteration of the list (i.e., the second shadowing set). There were very few errors, and missing items were left out of further analyses.

AXB perceptual similarity A total of 736 listeners provided holistic pronunciation judgments in AXB perceptual similarity tests. This use of the AXB paradigm assessed whether shadowed items were more similar to model items than baseline items. On each trial, three repetitions of the same lexical items were presented, with a model's item as X and a shadower's baseline and shadowed versions of the same item as A and B, counterbalanced for order of presentation. Listeners were instructed to decide whether the first or the last item (A or B) sounded more like the middle item (X) in its pronunciation, and they pressed the 1 (*first*) or the 0 (*last*) key on the keyboard to indicate their response on each trial. If shadowers converged detectably to model talkers, then their shadowed utterances should sound more similar in pronunciation to model talker utterances (X) than their baseline items (which were collected prior to hearing the model talker). To keep the task to a manageable length for listeners, separate AXB tests were constructed for each model–shadower pair's monosyllabic and bisyllabic words, resulting in 184 separate tests of 80 words each, which were each presented to different sets of four listeners. Within each test, each word triad was presented four times, once in each order (shadowed first, baseline first) in two randomized blocks.

The decision to use four listeners per shadower test (monosyllabic and bisyllabic) was guided by a pilot study that assessed reliability in data collected using ten listeners/shadowers versus smaller groups of listeners (the first five, four, three, two, or one) for 24 of the current study's shadowers. Thus, separate groupings of AXB data were created as if the AXB task had been conducted with all ten listeners per shadower, or with the first five listeners per shadower, and so on, to using just one listener per shadower. Previous studies have used as few as two listeners per shadower (Miller et al., 2010), and as many as 64 listeners (Namy et al., 2002). Given the scope of the current study, which comprised 184 separate AXB tests, it was necessary to determine a minimal number of listeners that could provide reliable data in these tests. Reliability was assessed in split-halves of an AXB test (comparing measures obtained in Block 1 vs. Block 2 of an AXB test), and for overall levels of convergence (comparing the patterns obtained for an entire AXB test across subsets of listeners). Furthermore, the data were collapsed across listeners by shadowers ($N = 24$) and by words ($N = 80$), because this study assessed effects of sex that varied by shadower and effects of frequency and type that varied by word.

In the pilot study, ten listeners provided AXB perceptual-similarity data for each shadower's monosyllabic test (for 24 of the shadowers), and the data were collapsed across all ten listeners by shadowers and words. Then, five additional groupings of listeners were created by using only the data from the first one to five listeners who participated in the pilot study, and collapsing their data by shadowers and words. Overall, the averages and standard deviations for the AXB tests did not differ for datasets that used all ten listeners versus those that used subsets generated from the ten listeners. Thus, using fewer listeners would have resulted in equivalent average levels of convergence. Figure 1 plots the correlation coefficients for split-half reliability (solid lines compare across AXB Blocks 1 and 2 within each test) and for overall convergence (dashed/dotted lines compare the overall AXB data using five or fewer listeners/shadower with those using ten listeners/shadower). It is clear that reliability remains very high when reducing the number of listeners from ten to three across both shadowers and words, except for the split-half block-to-block comparison in data collapsed by word. In that case, the within-test reliability starts lower and declines more rapidly. This analysis indicates that the data patterns are more robust for variability across shadowers than for words, and that using sets of four listeners/shadower would be roughly equivalent to using ten with respect to within-test consistency and the overall reliability of the shadower and word effects (note that the averages and standard deviations were also equivalent).

Acoustic measures Measures of phonetic convergence in individual acoustic attributes focused on monosyllabic words (*N*

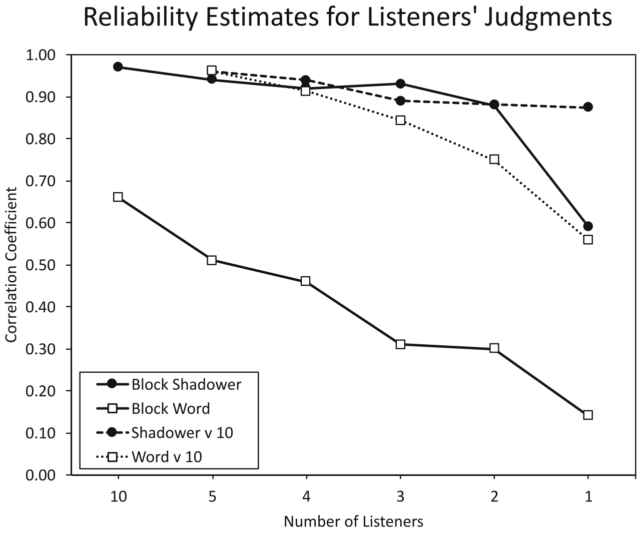

Fig. 1 Estimates of reliability (correlation coefficients) for phonetic convergence assessments when using ten versus one to five listeners per shadower. The data were collapsed by word (squares) and by shadower (circles). Two kinds of analyses are presented in the figure, split-half and shadower-set based. The solid lines starting at ten listeners report estimates for within-test split-half reliability that compare AXB Block 1 with AXB Block 2, and are labeled “Block Shadower” and “Block Word.” The dashed and dotted lines that start at only five listeners compare the average estimates across an entire AXB test using one to five listeners per shadower with the averages using ten listeners per shadower, and are labeled “Shadower v 10” and “Word v 10.”

= 80) because this word set balanced vowel identity and other segmental characteristics across word frequency categories. For all three sets of recordings (model, baseline, and shadowed items), trained research assistants measured vocalic duration as well as the fundamental frequency (F0) and vowel formants (F1 and F2) at the midpoint of each vowel. These measures were derived through visual inspection of the spectrograms and spectral plots using the default analysis settings in Praat (www.praat.org). Initial measures for the vowel spectra were cross-checked in F1 × F2 space for anomalous tokens by the second author, and anomalous measures were replaced with corrected measures. Anomalous measures were defined as those that resulted in vowel tokens that appeared in locations well outside of the cluster of points for an individual vowel, and/or that were more than two standard deviations from the mean. The final vowel formant measures were then normalized using the Labov technique in the *vowels* package (version 1.2-1; Kendall & Thomas, 2014) for R (version 3.1.3; R Development Core Team, 2015), yielding measures of F1’ and F2’. This technique scales the raw frequency measures for each talker’s vowels against a grand mean, permitting cross-talker comparisons that preserve idiolectal differences in vowel production (Labov, Ash, & Boberg, 2006).

Mixed-effects regression analyses Assessments of phonetic convergence employed mixed-effects binomial/logistic regression models to examine the impacts of talker sex, lexical factors, and acoustic convergence on AXB perceptual similarity. There are three main reasons to employ mixed-effects modeling over traditional analysis of variance with this dataset: (1) Mixed-effects regression handles multiple sources of variation simultaneously, which is not possible with traditional analysis of variance (Baayen, 2008; Baayen, Davidson, & Bates, 2008; Barr, Levy, Scheepers, & Tily, 2013); (2) binomial/logistic mixed-effects regression permits a more appropriate handling of binary data than does percent correct (see Barr et al., 2013; Dixon, 2008; Jaeger, 2008); and (3) like ordinary regression, mixed-effects regression permits analysis of continuous as well as categorical predictors. Analyses were conducted in R using the languageR (version 1.4.1; Baayen, 2015) and lme4 (version 1.1-7; Bates et al., 2014) packages. The modeling routines closely followed those prescribed by Baayen (2008), Jaeger (2008), and Barr et al. (2013).

In regression models, the AXB dependent measure was coded as the baseline versus shadowed item chosen on each trial, and the data from all listening trials were entered into the models. Thus, our regression analyses assessed the relative impact of each factor on the likelihood that a shadowed item sounded more similar to a model item than did a baseline item across all trials. Chi-square tests on the model parameters confirmed that inclusion of each significant factor improved the fit relative to a model without the factor. All categorical predictors were contrast-coded (–.5, .5) in the orders presented below, and all continuous predictors were *z*-scale normalized (and thereby centered). Thus, order was contrast-coded as first versus last in a trial, shadower sex was contrast-coded as female versus male, word frequency was contrast-coded as high versus low, and item type was contrast-coded as bi- versus monosyllabic. As was recommended by Barr et al. (2013), all models employed the maximal random-effects structure by including intercepts for all random sources of variance (shadowers, words, listeners, and models), and random slopes for all fixed effects, where appropriate. Detailed model parameters for the regression models reported below appear in Appendix B.

Results

AXB perceptual similarity

Descriptive statistics for the AXB perceptual similarity task reflect the proportion of trials in which a shadowed item was selected as more similar to a model utterance than a baseline item. The overall AXB phonetic convergence proportion averaged .56, which was significantly greater than chance responding of .50, as confirmed by a significant model intercept [Intercept = .245 (.035), Z = 7.043, p < .0001]. Thus, shadowers converged to model talkers, but the observed effect was characteristically subtle and comparable to those observed in other shadowing studies described above. Next, a model was constructed that included effects of model and shadower sex (female vs. male), pair type (same vs. mixed-sex pairings), word frequency (high vs. low), and word type (bisyllabic vs. monosyllabic) as predictors of phonetic convergence. Because model sex and pair type were nonsignificant factors that did not improve model fit or participate in significant interactions, they were eliminated from the final model (see Appendix B for the full details of the final model).

Overall, convergence was equivalent across female and male shadowers and model talkers (all *M*s = .56). With respect to pair type (same-sex vs. mixed-sex pairings), a numerical difference between same-sex and mixed-sex pairings was not significant (.55 < .56, p = .38), and there was no significant interaction between model sex and shadower sex (p = .39). The lack of reliable effects of talker sex in this study (among others) challenges a prevalent assertion that female talkers converge more than males. As discussed below, this assertion is not supported without qualification, both here and across the literature on phonetic convergence.

Additional predictors examined the impacts of lexical factors, including both word frequency (high vs. low) and word type (bisyllabic vs. monosyllabic). Again, a numerical difference in word frequency was not significant (high = .55, low = .56; p = .42; frequency was also not significant when treated as a continuous parameter). The lack of a difference due to word frequency held within both mono- and bisyllabic words (monosyllabic *M*s = .55; bisyllabic high = .562, low = .569) and when examining a subset of the data for the 31 shadowers with the highest AXB convergences (the averages for word frequency were equivalent at .61). Thus, word frequency findings were not due to overall performance levels or to using differently constructed word sets. However, phonetic convergence was influenced by word type—bisyllabic words evoked greater convergence than monosyllabic words (.57 > .55), and word type was a significant parameter in the model [β = –.090 (.032), Z = –2.849, p = .004; $\chi^2(3)$ = 80, p < .0001; the model also included random slopes for word type over shadowers].

Although a three-way interaction between shadower sex, word frequency, and item type was not significant (p = .79), there was a significant interaction between shadower sex and word frequency [β = –.050 (.023), Z = –2.199, p = .028; $\chi^2(6)$ = 27, p = .0001] and a marginal interaction between shadower sex and item type [β = .091 (.053), Z = 1.729, p = .084; $\chi^2(6)$ = 26, p = .0003]. As shown in Fig. 2, female shadowers were more susceptible to lexical effects. In each panel, the bars on the left correspond with convergence of female shadowers, with male shadower convergence on the right. The top panel shows that female shadowers converged more to low-frequency words, and that male talkers were not affected by word frequency. The bottom panel shows a similar pattern, in which female shadowers showed a marginally stronger difference in convergence to mono- versus bisyllabic words. There were no interactions between model sex and lexical factors in AXB convergence.

These interaction effects help explain some of the inconsistencies observed across the literature with respect to talker sex and word frequency. Recall that all three studies that reported greater convergence of female shadowers had used only low-frequency words (Babel et al., 2014; Namy et al., 2002; Miller et al., 2010). The present dataset replicates this pattern in the subset of bisyllabic low-frequency words—the mean convergence of female shadowers' bisyllabic low-frequency words was .58, whereas the convergence of male shadowers to the same items was .56. Therefore, if the present study had used only low-frequency words, a sex effect would have emerged. With respect to findings of effects of word frequency, one of the studies reporting an effect used only female shadowers (Dias & Rosenblum, 2016), one study used many more female than male shadowers (34 vs. 8; Babel, 2010), and one study did not provide information on shadower sex (Nielsen, 2011, Exp. 1). Thus, it appears that female shadowers tend to converge more than male shadowers on low-frequency words, and it is possible that some of the word frequency effects reported in the literature were driven by differences in the convergence of female talkers (but not Goldinger, 1998; Goldinger & Azuma, 2004). Although this interaction effect was reliable in the present study, it should be interpreted with caution, because it could be related to the context of collecting recordings of individual words in laboratory settings (see Byrd, 1994).

The AXB perceptual-similarity task revealed subtle holistic convergence of shadowers to model talkers. Word type influenced phonetic convergence, such that shadowers converged more to bisyllabic than to monosyllabic words. Recall that the bisyllabic word set was of higher frequency overall than the monosyllabic word set. Thus, the effect of word type goes against a prediction that low-frequency words should elicit greater convergence than high-frequency words. Talker sex and word frequency had no main effects on convergence, but shadower sex interacted with lexical properties such that female shadowers converged more to low-frequency words. The next set of analyses examined convergence in the individual acoustic attributes of monosyllabic words.

Convergence on acoustic attributes

To assess convergence in acoustic attributes, the duration, F0, and vowel spectra measures from the monosyllabic words were first converted into difference-in-distance (DID) scores. These scores compared baseline differences between each shadower and model with shadowed differences between each shadower and model. Thus, acoustic measures of phonetic

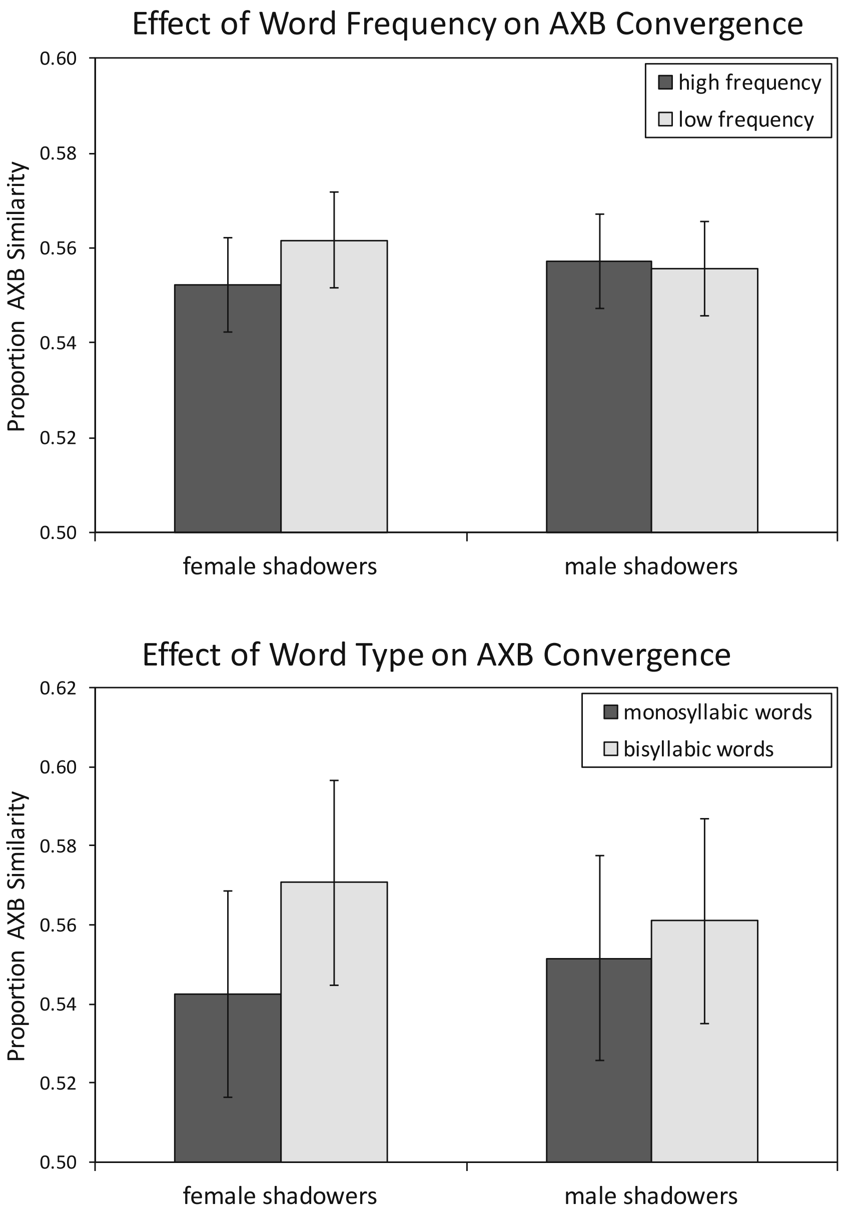

Fig. 2 Interactions between shadower sex and lexical properties in AXB perceptual convergence. Error bars span 95% confidence intervals. In the top panel, female shadowers converge to low-frequency more than to high-frequency words, whereas male shadowers show no impact of word frequency. In the bottom panel, female shadowers converge more to bisyllabic than to monosyllabic words, whereas male shadowers show a weaker effect.

convergence first derived differences in each parameter (duration, F0, and vowel spectra) between the baseline and model tokens (baseline – model) and between the shadowed and model tokens (shadowed – model). Then, absolute values of the differences for shadowed items were subtracted from absolute values of the differences for baseline items, yielding the DID estimates (DID = baseline distance – shadowed distance). Thus, values greater than zero indicate acoustic convergence, due to smaller differences during shadowing than during baseline. Because vowels are often described as points in two-dimensional space, an additional measure examined convergence in combined F1’ × F2’ vowel space by comparing interitem Euclidean distances (baseline to model minus shadowed to model).

In all measures, positive values indicate smaller differences for shadowed items to model items than for baseline items to model items, which should be interpreted as convergence during shadowing. To determine whether convergence in acoustic DIDs was influenced by talker sex or word frequency, all DID measures were submitted to linear mixed-effects modeling in R, analogous to treatment of the AXB perceptual-similarity data, with shadowers, words, and models entered as random sources and all fixed-effects factors contrast-coded (–.5, .5). The lmerTest package (version 2.0-25; Kuznetsova, Brockhoff, & Christensen, 2015) was used to obtain *p* values for these models, employing Satterthwaite's approximation for degrees of freedom.

Table 2 displays summary statistics for all acoustic models. The first column lists average DIDs for each acoustic attribute, with parameter estimates from mixed-effects regression modeling listed in adjacent columns. On average, acoustic DIDs converged for duration, there was marginal convergence in F1' × F2' vowel spectra and in F2' alone, and no significant convergence in F1' or F0. Thus, it appears that results are more robust (and arguably more valid) when treating vowel formant spectra as two-dimensional points rather than as separate parameters. Analogous to the pattern observed in AXB convergence, there were significant interactions between shadower sex and word frequency for every acoustic DID measure, and no main effects of shadower sex or word frequency (these nonsignificant main effect parameter estimates are omitted here for clarity). There were also no interactions between model sex and word frequency.

Figure 3 displays interactions between shadower sex and word frequency for all acoustic DID measures. Each panel shows convergence of female shadowers to high- and low-frequency words on the left, with corresponding data for male shadowers on the right. Most acoustic measures of phonetic convergence aligned with AXB perceptual similarity with respect to effects of word frequency and talker sex. For female shadowers, all acoustic measures except duration showed at least a trend toward greater convergence to low- than to high-frequency words, and the effect was strongest in F0 and F1'. Male shadowers showed more complex trends, but most (except F1') were in the opposite direction from those of female shadowers.

Most acoustic DID attributes did not converge, on average, but examinations of interactions between talker sex and word frequency revealed complex patterns of convergence across these measures. These patterns are difficult to interpret without a clear rationale for choosing one measure over another. A potential solution to this problem would be to relate these measures to a more holistic assessment of phonetic convergence. Therefore, the next set of analyses examined the relationship between acoustic convergence and holistic convergence by using acoustic DID measures as predictors of variation in AXB perceptual similarity.

Convergence in multiple acoustic attributes predicts holistic phonetic convergence

A final set of logistic/binomial mixed-effects models assessed whether variability in AXB perceptual convergence could be predicted by convergence in acoustic attributes (see also Pardo, Jordan, et al., 2013). To conduct these analyses, each acoustic DID factor was first converted to *z* scores, which both centers them and permits comparisons of the relative contribution of each factor to predicting variability in AXB perceptual convergence. Because two-dimensional vowel DID and individual F1' and F2' DIDs were correlated across shadowers [F1' DID × vowel DID, *r*(90) = .28, *p* < .008; F2' DID × vowel DID, *r*(90) = .96, *p* < .0001], effects of vowel DID were assessed in a separate model from one that examined F1' and F2' (these measures were not correlated). In all cases, models that included multiple acoustic attributes were a better fit to AXB perceptual convergence than were models with fewer acoustic attributes.

Vowel DID First, a full model including duration DID, F0 DID, and vowel DID indicated that each parameter was a significant predictor of variation in AXB perceptual similarity.

Table 2 Acoustic difference-in-distance (DID) measures and significance tests

	Mean	Estimate	*SE*	*df*	*t*	*p*
Duration DID	8.46 ms	8.57 ms	2.319	13.1	3.694	.0030*
ShSex × Freq		1.68	0.446	104200	3.774	.0002*
Vowel DID	5.77 Hz'	5.30 Hz'	2.617	20	2.024	.0563*
ShSex × Freq		−2.85	0.996	104100	−2.862	.0042*
F2 DID	5.15 Hz'	4.81 Hz'	2.781	19	1.731	.1001
ShSex × Freq		−2.48	1.068	104200	−2.32	.0203*
F1 DID	1.66 Hz'	1.52 Hz'	1.200	160	1.269	.2062
ShSex × Freq		−2.07	0.571	104200	−3.625	.0003*
F0 DID	0.26 Hz	0.22 Hz	1.019	13	0.211	.8364
ShSex × Freq		−0.51	0.179	104200	−2.844	.0045*

The DID measures correspond to baseline minus shadowed differences between shadowers and their models. Interactions between shadower sex (ShSex) and word frequency (Freq) for each measure are included. The main effects of talker sex and word frequency were not significant and are omitted. Significance tests used Satterthwaite's approximation for the *df*s

* Significant results

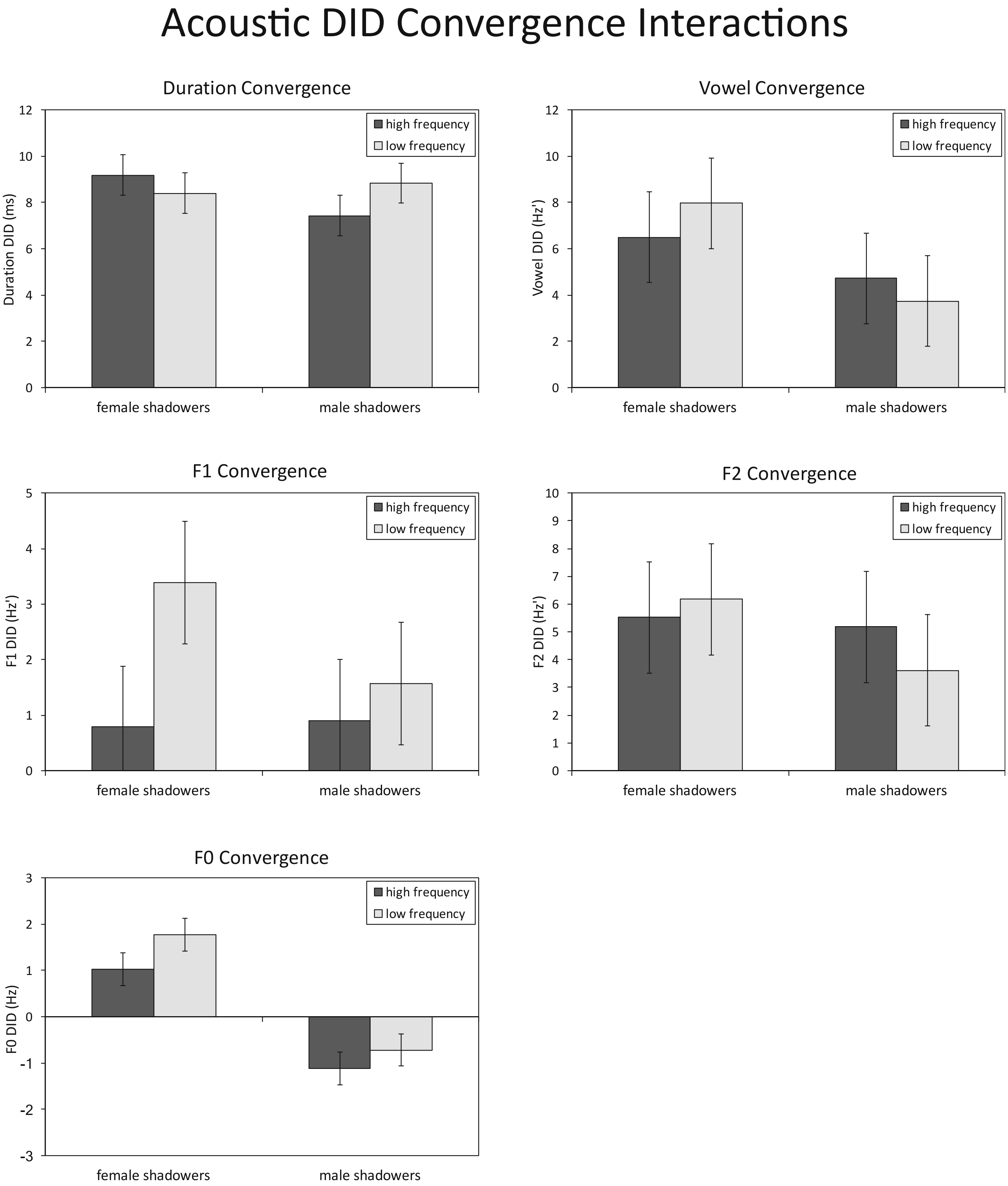

Fig. 3 Interactions between shadower sex and word frequency in multiple acoustic measures of phonetic convergence. Error bars span 95% confidence intervals. Female shadowers show trends toward greater convergence to low-frequency words on all acoustic attributes except duration. Male shadowers show more varied results across acoustic attributes.

Inclusion of each parameter improved model fit relative to a model without the parameter (see Appendix B for the full model details). Inspection of the beta weights indicated that duration DID was the strongest predictor [$\beta = .080$ (.013), $Z = 6.385$, $p < .0001$; $\chi^2(5) = 192$, $p < .0001$], followed by F0 DID [$\beta = .073$ (.011), $Z = 6.710$, $p < .0001$; $\chi^2(5) = 79$, $p < .0001$], and vowel DID [$\beta = .057$ (.010), $Z = 5.407$, $p < .0001$; $\chi^2(5) = 107$, $p < .0001$]. Measures of Somer's Dxy (.284) and concordance (.642) for the full model indicated a modest fit to the data. These data replicate the pattern of acoustic attribute predictions reported by Pardo, Jordan, et al. (2013), in a new and more extensive set of shadowers.

Formant DID When treated separately, both F1' and F2' DIDs were significant predictors of AXB perceptual convergence, along with duration DID and F0 DID. Compared to the prior vowel model, using F1' and F2' DIDs as separate parameters had a negligible impact on beta weights for duration DID and F0 DID, and the full model revealed the same relative influences, with F1' DID having a stronger impact than F2' DID [duration DID: $\beta = .079$ (.013), $Z = 6.308$, $p < .0001$, $\chi^2(6) = 186$, $p < .0001$; F0 DID: $\beta = .072$ (.010), $Z = 6.882$, $p < .0001$ $\chi^2(6) = 76$, $p < .0001$; F1' DID: $\beta = .057$ (.010), $Z = 5.498$, $p < .0001$, $\chi^2(6) = 126$, $p < .0001$; F2' DID: $\beta = .033$ (.010), $Z = 3.301$, $p < .0001$, $\chi^2(6) = 59$, $p < .0001$]. Measures of Somer's Dxy (.287) and concordance (.643) for the full model indicated a modest fit to the data that was slightly higher than that of the prior model with two-point vowel DID.

Overall, patterns of convergence in acoustic attributes predicted AXB perceptual similarity, and including multiple attributes together yielded better fits to the data than those of models with fewer parameters. Additional analyses indicated that these patterns were not modulated by model or shadower sex. These analyses confirmed that AXB perceptual convergence reflected holistic patterns of convergence in multiple acoustic dimensions simultaneously. It is notable that F0 and F1' DIDs were relatively strong predictors, despite having nonsignificant average convergence themselves, which probably contributed to relatively weak detection of holistic convergence. These data indicate that phonetic convergence reflects a complex interaction among multiple acoustic–phonetic dimensions, and that reliance on any individual acoustic attribute yields a portrait that is incomplete at best, and potentially misleading. For example, a study that only reported data from measures of vowel spectra would arrive at a very different conclusion than a study that examined duration or F0. Furthermore, the relatively modest overall fits of the models to the perceptual data indicate that additional and/or different kinds of attributes might also contribute to perceived convergence.

Model talker variability

A final consideration involves whether characteristics of the individual model talkers were more or less likely to evoke convergence from shadowers. Although interactions between model sex and shadower sex were not significant, examining phonetic convergence across individual model talkers revealed interesting patterns, shown in Fig. 4. Each pair of bars depicts convergence to a model talker by female shadowers (dark bars) and male shadowers (light bars). Female models appear in the left half of the figure, and models are ordered from left to right within sex by average AXB convergence levels. Most female models (four of six) evoked greater convergence from male than from female shadowers, and more convergence from their male shadowers than most male models. Most male models evoked high levels of convergence from female shadowers (four of six), more so than most female models (with the exception of F04ao).

Given the methodological choices across the literature, these patterns are important because they indicate that individual model talkers have consequences for overall convergence levels and for drawing conclusions about talker sex (see also Babel et al., 2014). For example, a study that used F04ao and M11bk as models and only examined same-sex pairings would lead to a conclusion that females converged and males did not. A different conclusion could be drawn using F07jt and M18rz. Although average differences by sex were small and not significant in this dataset, these trends merit further investigation with a larger set of model talkers. Finally, it is clear that avoidance of mixed-sex pairings in many designs is neither well-founded nor productive, because some of the highest levels of convergence occurred in mixed-sex pairs.

Discussion

This large-scale examination of phonetic convergence has shown that shadowers converged to multiple model talkers in multiple measures to varying degrees. By using 92 shadowers split into 32 same-sex female, 30 same-sex male, and 30 mixed-sex pairings with 12 model talkers, this study constitutes a rigorous assessment of the impacts of talker sex and word frequency on phonetic convergence. Thus, any failures to replicate previous findings are not simply due to a lack of power in the present study. Convergence occurred on average in holistic AXB perceptual assessment and duration measures, there was marginal convergence in measures of two-dimensional vowel space and F2 alone, and there was no significant average convergence in F1 and F0 measures.

Talker sex and word frequency had no effects on overall levels of convergence, but interactions between them revealed that female shadowers were more susceptible to lexical properties. That is, female shadowers converged more to low-frequency than to high-frequency words, and more than male shadowers to low-frequency words. Therefore, previously reported findings that female shadowers converge more than males could have been due to fact that those studies used only low-frequency items (Babel et al., 2014; Miller et al., 2010; Namy et al., 2002). Likewise, some previous studies reporting greater convergence to low-frequency words could have been due to the use of only female shadowers (Babel, 2010; Dias &

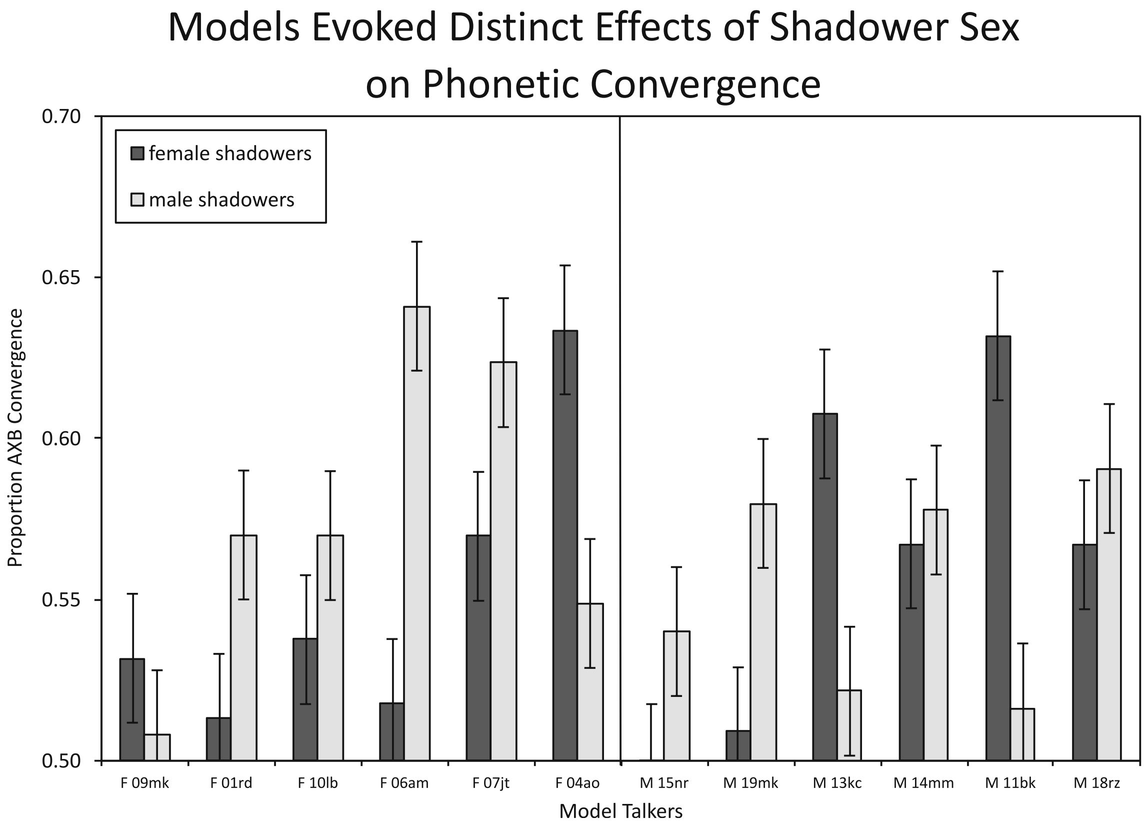

Fig. 4 Phonetic convergence collapsed by individual model talkers. Error bars indicate standard errors; note that the interaction between model and shadower sex was not significant. Female models are shown on the left side; dark bars depict convergence of female shadowers, and light bars depict convergence of male shadowers. Different models evoke different patterns of convergence across female and male shadowers. The average AXB for female shadowers of M15nr equals .50.

Rosenblum, 2016; and possibly Nielsen, 2011). It is clear from these results that the prevalent view that female talkers converge more than males must be qualified—the effect is weak and inconsistent, and only appears when studies use low-frequency words (see also Pardo, 2006; Pardo, Jordan, et al., 2013). It is not clear why this particular pattern occurs, but the inconsistency in the effects of talker sex preclude a simplistic interpretation that females converge more than males.

Reconciling effects of word frequency

Word frequency effects are more difficult to reconcile than talker sex effects. Pardo, Jordan, et al. (201b) also failed to find frequency effects in the same monosyllabic items used in the current study. To conduct a more comparable assessment, the current study included the same bisyllabic words that evoked the original finding reported by Goldinger (1998). However, word frequency effects were not robust in the present dataset, only emerging as a weak effect in female shadowers across all six measures of phonetic convergence. Goldinger (1998) also included a repetition manipulation, in which talkers heard prompts 0, 2, 6, or 12 times prior to shadowing. The most comparable data from that study to the current dataset would be those words with two repetitions (however Goldinger's talkers did not shadow during the first presentation block). In that cell, high-frequency words yielded approximately 63% correct detection of imitation, whereas low-frequency words yielded performance levels around 75% (estimates derived from inspection of Fig. 4 in Goldinger 1998). Goldinger and Azuma (2004) exposed talkers to the same words under the same repetition manipulation, but collected target utterances a week later. In that case, high-frequency words heard twice yielded approximately 50% correct detection of imitation, whereas low-frequency words yielded around 58% (estimates derived from inspection of Fig. 2 in Goldinger and Azuma 2004).

Overall, performance levels reported in the current study more closely resemble those of Goldinger and Azuma, who collected utterances a full week after exposure, but the frequency effect was stronger in their dataset. In Goldinger (1998), even the exposure condition with zero prior repetitions yielded 60% detection levels in high-frequency words. All of the AXB studies listed in Table 1 reported average convergence levels less than 62% (except Dias & Rosenblum, 2016), and four used low-frequency words, which should have elicited the highest levels of convergence (Babel et al., 2014; Miller et al., 2010, 2013; Shockley et al., 2004). It is worth noting that the higher performance levels reported in

Goldinger (1998) have only been observed in one other study using AXB convergence assessment—Dias and Rosenblum (2016) reported an overall AXB $M = .69$. Although overall performance levels in the current study do not align with those reported in Goldinger (1998), they are comparable to those of Goldinger and Azuma (2004) and to most other findings in the literature. Moreover, there were no frequency effects in the current dataset, even among the top-converging 31 shadowers ($M = .61$). Therefore, the current failure to replicate is unlikely to be due to floor effects or to poor power in the dataset.

Three other studies reported significant effects of word frequency on convergence. Two of these studies used acoustic measures of convergence and did not report the size of the effect on their measures (vowel spectra: Babel, 2010; and VOT: Nielsen, 2011). Moreover, the effect was not reliable in all conditions tested in these studies. A recent study by Dias and Rosenblum (2016) reported substantial effects of word frequency on AXB phonetic convergence (low .71 > high .67), but the study employed bisyllabic words produced by female talkers shadowing a single female model. In addition, their study included audiovisual presentation of prompts in some shadowing trials, which increased performance levels relative to audio-alone trials. Although they did not report examining interactions between presentation mode and lexical frequency, it is possible that frequency effects were enhanced by audiovisual presentation.

Examination of bisyllabic words in the present dataset revealed that some model talkers elicited greater convergence to low-frequency words from female shadowers (proportions differed by >.02 for six models), whereas others elicited equivalent degrees of convergence across low- and high-frequency words from female shadowers (proportions differed by <.02 for six models). Given the scope of the present study, as well as a previously reported failure to replicate frequency effects on phonetic convergence (Pardo, Jordan, et al., 2013), a conservative conclusion would be that effects of word frequency on phonetic convergence are inconsistent and possibly sensitive to talker sex.

Episodic memory models and word frequency effects

Frequency is a prominent attribute in episodic memory systems, often generating specific testable predictions, as exemplified in Goldinger (1998; see also Hintzman, 1984; Johnson, 2007; Pierrehumbert, 2006, 2012). As discussed earlier, frequency effects in episodic models of memory emerge from parallel activation of multiple stored traces during perception, which contribute to an echo that constitutes recognition, and as shown in Goldinger, influences speech production. An episodic echo incorporates elements from activated representations and the most recent item, in this case, a shadowing prompt. An echo of a more frequently encountered word comprises many more competitors to a prompt than that of a less frequently encountered word, thereby reducing the contribution of the prompt to the echo. Many examples of specificity effects in speech perception attest to the validity of episodic models of recognition memory.

Crucially, the set of exemplars that are activated depends on their similarity to a prompt (Hintzman, 1984). Because episodic echo generation depends on similarity of stored exemplars to a prompt, an account is needed of what attributes are encoded, how attributes are used to activate stored episodes, and of the scope of candidate traces that are activated. Goldinger (1998) achieved adequate fits to his nonword dataset by modeling vectors with both word elements and voice elements for all episodes. By incorporating voice elements, the model could also predict that greater exposure to a particular voice would lead to enhanced convergence to words produced by the same talker relative to those produced by a different dissimilar talker. Pierrehumbert's (2001, 2006) hybrid model adds important refinements to episodic models by imposing constraints on the number of activated traces; by proposing that exemplars are equivalence classes of perceptual experiences rather than the experiences themselves; and through preferential weighting of recent exemplars and preferred voices. Thus, whereas speech perception might yield episodic elements in an echo, the inconsistency of frequency effects indicates that these elements are unlikely to represent all previous encounters, do not comprise a fixed set of acoustic-phonetic attributes, and do not always evoke convergent speech production.

Integrated perception-production and phonetic convergence

Pickering and Garrod's (2013) simulation route for language comprehension centers on complete integration between perception and production processes, which supports and promotes phonetic convergence (among other kinds of alignment; see also Gambi & Pickering, 2013). This occurs because comprehension entails a process of forward modeling simulations involving covert imitation of perceived speech that can become overt imitation. Accordingly, these forward simulations are impoverished relative to actual production planning, and they are scaled to a talker's own production system. Based on these core features of the simulation route, the model predicts that talkers should converge at the phonological level, and be better able to imitate their own utterances and utterances of individuals more similar to them.

As was pointed out by Pickering and Garrod (2013), listeners should not repeat talkers' utterances verbatim during conversational interaction, rather, most of their contributions should be complementary. The purpose of simulation is to facilitate language comprehension and ultimately spoken communication. Despite this circumstance, the forward modeling component in this account predicts phonetic convergence at the phonological level. It has already been established that phonetic convergence does not require verbatim repetition—talkers converge on phonetic features that are apparent when comparing across different lexical items (Kim, Horton, & Bradlow, 2011) and that even span words from different languages (Sancier & Fowler, 1997). Furthermore, a noninteractive speech shadowing task that minimizes competing conversational demands should facilitate convergence. With respect to predictions regarding modulations of convergence based on similarity, same-sex pairs should converge more than mixed-sex pairs, but this pattern was not found in the present study. There is some evidence that convergence is stronger for within-language and within-dialect pairings (Kim et al., 2011; Olmstead et al., 2013), but others have found the opposite patterns (Babel, 2010, 2012; Walker & Campbell-Kibler, 2015). Furthermore, Pardo (2006) paired talkers from distinct dialect regions, and found comparably robust findings to studies using same-dialect talkers. Given the degree of support for convergence in a fully integrated perception-production model, it is surprising that observed levels of convergence are so weak and variable, even in circumstances that seem most favorable for eliciting convergent production.

It is arguable that weak levels of phonetic convergence are due to anatomical differences or to habitual speech production patterns, which limit a talker's ability to match another talker's acoustic-phonetic attributes, even in speech shadowing tasks (Fowler et al., 2003). If habitual speech patterns prevent a talker from matching another's speech, they should assist a talker in matching their own speech. However, a study examining directed imitation of individuals' own speech samples complicates an interpretation based on similarity, habits, or anatomical differences (Vallabha & Tuller, 2004). In that case, talkers were unable to match their own vowel formant acoustics. Crucially, self-imitations exhibited patterned biases that were not uniform across the vowel space, and were not explained by models of random noise either in production or perception. Thus, habitual patterns in speech production appear to drive systematic yet complex variation in production. Taken together, observed patterns of weak and variable phonetic convergence do not align well with predictions from this fully integrated model.

Attributes and measures of phonetic convergence

As in Pardo, Jordan, et al. (2013), the present study reveals that talkers do not imitate all acoustic–phonetic attributes in the same manner (see also Babel & Bulatov, 2012; Babel et al., 2013; Levitan & Hirschberg, 2011; Pardo et al., 2010; Pardo Gibbons, Suppes, & Krauss, 2012; Pardo, Cajori Jay, et al., 2013; Walker & Campbell-Kibler, 2015). No single attribute drives convergence, and talkers converge on some attributes at the same time that they diverge or fail to converge on others. Each talker exhibits a unique profile of convergence and divergence on multiple dimensions that is perceived holistically. For example, one talker might converge on duration, diverge on F0, and show little or no change in vowel formants, whereas another talker might converge on vowel formants, diverge on duration, and show no change in F0. Moreover, this variability can be observed across words within a single talker. Across the set of talkers examined here, all possible combinations were observed. Thus, individual acoustic attributes, considered alone, contribute little to an understanding of the phenomenon.

It is important to acknowledge the complexities and limitations involved in measuring phonetic convergence. The choice of attributes to measure in a study rests on often implicit assumptions about the nature of phonetic form variation and convergence. Measures of particular acoustic attributes have proven useful for examining sound changes in progress, but more comprehensive measures are necessary for addressing broader questions related to phonetic convergence. Because measurable acoustic attributes do not always align with vocal tract gestures, perceptual assessments are more likely to reflect actual patterns of phonetic variation and convergence. Future investigations would benefit from enhanced measures that explore articulatory parameters and/or acoustic parameters that better reflect articulatory dynamics.

For the purposes of drawing general conclusions, the AXB perceptual similarity task provides a ready means for calibrating phonetic convergence across multiple acoustic–phonetic dimensions, and avoids potentially misleading interpretations based on patterns found in a single attribute. The present examination of multiple acoustic–phonetic attributes in parallel reveals that the landscape of phonetic convergence is extremely complex. Analogous to episodic memory echoes, forward modeling simulations do not necessarily evoke phonetic convergence, as other factors intervene between perception and production on some occasions.

Talkers as targets of convergence

Thus far, investigations of phonetic convergence have focused on the converging talker. For example, studies have explored the impact of individual differences in talkers on their degrees of phonetic convergence (Aguilar et al., 2016; Mantell & Pfordresher, 2013;

Postma-Nilsenová & Postma, 2013; Yu et al., 2013) and have related talker attitudes toward models to phonetic convergence (Abrego-Collier et al., 2011; Babel et al., 2013; Yu et al., 2013). A related and equally important consideration involves aspects of talkers who are the targets of convergence. As demonstrated here, some models evoke greater degrees of convergence from shadowers, and distinct patterns of convergence from male and female shadowers. When relating patterns of immediate phonetic convergence to broader contexts of language use, it is important to consider both sides of the phenomenon.

A recent study by Babel et al. (2014) offers a promising perspective. They first collected ratings of vocal attractiveness and measures of gender typicality for 60 talkers and selected a set of eight talkers (four females) who yielded the lowest and highest scores for each attribute. These model talkers were then shadowed by others, and phonetic convergence was influenced by attractiveness and typicality of model talkers. Given the current state of research in the field, in which most studies use very few model talkers, additional investigations are warranted to evaluate the characteristics of model talkers that might evoke more or less convergence from multiple shadowers.

Conclusion

Research on phonetic convergence both promotes and challenges accounts of integrated speech perception and production, and exemplar-based episodic memory systems. On the one hand, a listener must perceive and retain phonetic attributes in sufficient detail to support convergent production; on the other, phonetic convergence is subtle and highly variable across individuals, both as talkers and as targets of convergence. Perceptual assessment harnesses variability across multiple acoustic–phonetic attributes, calibrating the relative contribution of each attribute to holistic phonetic convergence. To draw broad conclusions about phonetic convergence, studies should employ multiple models and shadowers with equal representation of male and female talkers, balanced multisyllabic items, and comprehensive measures. As a potential mechanism of language acquisition and sound change, phonetic convergence reflects complexities in spoken communication that warrant elaboration of the underspecified components of current accounts.

Author note Completion of this project was supported in part by NSF PAC program Grant No. BCS-1229033. The authors thank Alexa Decker, Keagan Francis, Hannah Gash, Nicholas Mason, Sara Parker, and Tainah Torres for assistance with data collection. The narrative benefitted greatly from comments provided by the editor, Dr. Lisa Sanders, and three anonymous reviewers. Portions of this project were presented at the 167th, 169th, and 171st meetings of the Acoustical Society of America; at the 55th, 56th, and 57th Annual Meetings of the Psychonomic Society; and at AMLaP 2015.

Appendix A: Word sets

Bisyllabic		Monosyllabic	
Low Frequency	High Frequency	Low Frequency	High Frequency
active	basis	babe	bad
balance	become	bathe	bag
beacon	before	beak	beach
bicep	better	bean	beam
captain	between	boot	beat
career	beyond	cage	bet
careful	city	cake	bone
cavern	common	cop	check
coffee	country	cot	death
cousin	father	dab	dock
deport	figure	dad	foot
dozen	final	dame	gain
fashion	later	deaf	game
favor	market	debt	gave
forage	matter	dome	get
forget	music	dot	got
garden	nature	fad	half
garter	never	gene	known
gusto	number	hoof	laugh
handle	order	hook	loan
hazel	party	hoot	lock
jelly	people	keen	mean
listen	person	knock	moon
master	picture	leach	note
mingle	police	mash	pot
nectar	power	moan	put
novel	program	moat	rock
nugget	public	mop	room
parcel	rather	nape	rose
patron	recent	pep	sad
permit	report	pet	sang
pigeon	river	rash	save
portal	second	roam	scene
rustic	single	robe	shape
staple	social	rope	suit
symbol	spirit	sag	tape
title	system	siege	team
venom	table	sock	top
vision	value	tune	wrote
wedlock	water	womb	youth

Appendix B: Mixed-effects regression modeling

AXB perceptual similarity model

The order parameter controls for listener biases in choosing first versus last items in AXB tests. A significant effect of word type indicates that bisyllabic words were associated with an increased likelihood that a shadowed item would sound more similar to a model item than would a baseline item in AXB tests. Chi-square statistics confirmed that the inclusion of significant fixed effects and interactions in the model yielded a significant improvement in model fit relative to a model that excluded that parameter. As was recommended by Barr et al. (2013), all models employed maximal random-effects structures by including both intercepts and random slopes where appropriate.

Somer's Dxy = .287, Concordance = .644

Fixed Effects	*β*	*SE*	*Z*	*p*(*Z*)	$\chi^2(df)$	$p(\chi^2)$
(Intercept)	.245	.033	7.450	9.3e–14		
order.effect: first	.021	.031	0.685	.493	5,962 (3)	2.2e–16
Shadower Sex: female	–.005	.058	–0.084	.933		
Frequency: high	.017	.021	0.814	.416		
Item Type: bi	–.090	.032	–2.849	.004	80 (3)	2.2e–16
Sex X Frequency: high	–.050	.023	–2.199	.028	27 (6)	.0001
Sex X Item Type: bi	.091	.053	1.729	.084	26 (6)	.0003

Random Effects

Group	Source	Variance	*SD*	Corr	
Listener	(Intercept)	.0312	.1766		
	order.effect: first	.6424	.8015	–.06	
Word	(Intercept)	.0123	.1109		
Shadower	(Intercept)	.0705	.2655		
	itemType.effect: bi	.0409	.2021	–.47	
	freq.effect: high	.0047	.0683	–.34	.04
Model	(Intercept)	.0021	.0461		
AIC	BIC	LogLik	Deviance	df Resid	
299,730	299,916	–149,847	299,694	226,027	

Acoustic models

Order parameters control for listener biases in choosing first versus last items in AXB tests. Significant parameters for acoustic difference-in-distance (DID) measures indicate that larger distances between baseline and model utterances compared to shadowed and model utterances were associated with an increased likelihood that a shadowed item would sound more similar to a model item than would a baseline item in AXB tests. Chi-square statistics confirmed that the inclusion of each parameter in the model yielded a significant improvement in model fit relative to a model that excluded that parameter. As was recommended by Barr et al. (2013), models employed maximal random-effects structures by including both intercepts and random slopes where appropriate.

Vowel model: Somer's Dxy = 0.284, Concordance = 0.642

Fixed Effects	*β*	*SE*	*Z*	*p*(*Z*)	$\chi^2(df)$	$p(\chi^2)$
(Intercept)	.190	.030	6.331	2.4e–10		
order.effect: first	–.066	.041	–1.630	.103	2,389 (3)	2.2e–16
zdurDID	.080	.013	6.385	1.7e–10	192 (5)	2.2e–16
zF0DID	.073	.011	6.710	2.0e–11	79 (5)	1.2e–15
zvowelDID	.057	.010	5.407	6.4e–08	107 (5)	2.2e–16

Random Effects

Group	Source	Variance	*SD*	Corr		
Listener	(Intercept)	.0204	.1428			
	order.effect: first	.5442	.7377	–.12		
Shadower	(Intercept)	.0398	.1994			
	zvowelDID	.0057	.0753	.01		
	zF0DID	.0033	.0578	–.17	.19	
	zdurDID	.0088	.0938	.04	.14	.10
Word	(Intercept)	.0145	.1205			
Model	(Intercept)	.0021	.0462			
AIC	BIC	LogLik	Deviance	df Resid		
139,438	139,629	–69,699	139,398	104,319		

Formants model: Somer's D = 0.287, Concordance = 0.643

Fixed Effects	*β*	*SE*	*Z*	*p(Z)*	$\chi^2(df)$	$p(\chi^2)$
(Intercept)	.192	.029	6.649	2.96E–11		
order.effect: first	–.066	.041	–1.635	.1021	2,391 (3)	2.2e–16
zdurDID	.079	.013	6.308	2.83E–10	186 (6)	2.2e–16
zF0DID	.072	.010	6.882	5.90E–12	76 (6)	2.3e–14
zF1DID	.057	.010	5.498	3.85E–08	126 (6)	2.2e–16
zF2DID	.033	.010	3.301	.0010	59 (6)	8.4e–11

Random Effects

Group	Source	Variance	*SD*	Corr			
Listener	(Intercept)	.0204	.1429				
	order.effect: first	.5454	.7385	–.12			
Shadower	(Intercept)	.0411	.2028				
	zF1DID	.0055	.0741	–.55			
	zF2DID	.0050	.0708	.12	.06		
	zF0DID	.0026	.0511	–.09	.30	.03	
	zdurDID	.0088	.0938	.05	.06	.11	.02
Word	(Intercept)	.0139	.1180				
Model	(Intercept)	.0012	.0342				
AIC	BIC	LogLik	Deviance	df Resid			
139,370	139,618	–69,659	139,318	104,313			

References

Abney, D. H., Paxton, A., Dale, R., & Kello, C. T. (2014). Complexity matching in dyadic conversation. *Journal of Experimental Psychology: General, 143,* 2304–2315. doi:10.1037/xge0000021

Abrego-Collier, C., Grove, J., Sonderegger, M., & Yu, A. C. L. (2011, August). *Effects of speaker evaluation on phonetic convergence.* Paper presented at the 17th International Congress of the Phonetic Sciences, Hong Kong.

Aguilar, L., Downey, G., Krauss, R., Pardo, J., Lane, S., & Bolger, N. (2016). A dyadic perspective on speech accommodation and social connection: Both partners' rejection sensitivity matter. *Journal of Personality, 84,* 165–177. doi:10.1111/jopy.12149

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* New York, NY: Cambridge University Press.

Baayen, R. H. (2015). *languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics."* R package version 1.4.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412. doi:10.1016/j.jml.2007.12.005

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

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics, 40,* 177–189. doi:10.1016/j.wocn.2011.09.001

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

Babel, M., McAuliffe, M., & Haber, G. (2013). Can mergers-in-progress be unmerged in speech accommodation. *Frontiers in Psychology, 4,* 653. doi:10.3389/fpsyg.2013.00653

Babel, M., McGuire, G., Walters, S., & Nicholls, A. (2014). Novelty and social preference in phonetic accommodation. *Laboratory Phonology, 5,* 123–150. doi:10.1515/lp-2014-0006

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language, 42,* 1–22. doi:10.1006/jmla.1999.2667

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255–278. doi:10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., & Dai, B. (2014). *lme4: Linear mixed-effects models using Eigen and S4 classes (R package version 1.1-7).* Retrieved from http://cran.r-project.org/package=lme4

Bilous, F. R., & Krauss, R. M. (1988). Dominance and accommodation in the conversational behaviours of same-and mixed-gender dyads. *Language and Communication, 8,* 183–194.

Bourhis, R. Y., & Giles, H. (1977). The language of intergroup distinctiveness. In H. Giles (Ed.), *Language, ethnicity, and intergroup relations* (pp. 119–135). London, UK: Academic Press.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic coordination in dialogue. *Cognition, 75,* B13–B25. doi:10.1016/S0010-0277(99)00081-5

Branigan, H. P., Pickering, M. J., McLean, J. F., & Cleland, A. A. (2007). Syntactic alignment and participant role in dialogue. *Cognition, 104,* 163–197. doi:10.1016/j.cognition.2006.05.006

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1482–1493. doi:10.1037/0278-7393.22.6.1482

Byrd, D. (1994). Relations of sex and dialect to reduction. *Speech Communication, 15,* 39–54.

Byrne, D. (1971). *The attraction paradigm*. New York, NY: Academic Press.

Chang, C. B. (2012). Rapid and multifaceted effects of second-language learning on first-language speech production. *Journal of Phonetics, 40,* 249–268. doi:10.1016/j.wocn.2011.10.007

Coupland, N. (1984). Accommodation at work: Some phonological data and their implications. *International Journal of the Sociology of Language, 46,* 49–70.

Delvaux, V., & Soquet, A. (2007). The influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica, 64,* 145–173.

Dias, J. W., & Rosenblum, L. D. (2011). Visual influences on interactive speech alignment. *Perception, 40,* 1457–1466. doi:10.1068/p7071

Dias, J. W., & Rosenblum, L. D. (2016). Visibility of speech articulation enhances auditory phonetic convergence. *Attention, Perception, & Psychophysics, 78,* 317–333. doi:10.3758/s13414-015-0982-6

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language, 59,* 447–456. doi:10.1016/j.jml.2007.11.004

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

Evans, B. G., & Iverson, P. (2007). Plasticity in vowel perception and production: A study of accent change in young adults. *Journal of the Acoustical Society of America, 121,* 3814–3826. doi:10.1121/1.2722209

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14,* 3–28.

Fowler, C. A. (2010). Embodied, embedded language use. *Ecological Psychology, 22,* 286–303. doi:10.1080/10407413.2010.517115

Fowler, C. A. (2014). Talking as doing: Language forms and public language. *New Ideas in Psychology, 32,* 174–182. doi:10.1016/j.newideapsych.2013.03.007

Fowler, C. A., & Galantucci, B. (2005). The relation of speech perception and speech production. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 633–652). Malden, MA: Blackwell.

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, 49,* 396–413.

Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language, 49,* 396–413. doi:10.1016/S0749-596X(03)00072-X

Fowler, C. A., Shankweiler, D., & Studdert-Kennedy, M. (2016). "Perception of the speech code" revisited: Speech is alphabetic after all. *Psychological Review, 123,* 125–150. doi:10.1037/rev0000013

Fusaroli, R., & Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive Science, 40,* 145–171. doi:10.1111/cogs.12251

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review, 13,* 361–377. doi:10.3758/BF03193857

Gallois, C., Giles, H., Jones, E., Cargiles, A. C., & Ota, H. (1995). Accommodating intercultural encounters: Elaboration and extensions. In R. Wiseman (Ed.), *Intercultural communication theory* (pp. 115–147). Newbury Park, CA: Sage.

Gambi, C., & Pickering, M. J. (2013). Prediction and imitation in speech. *Frontiers in Psychology, 4,* 340. doi:10.3389/fpsyg.2013.00340

Garnier, M., Lamalle, L., & Sato, M. (2013). Neural correlates of phonetic convergence and speech imitation. *Frontiers in Psychology, 4,* 600. doi:10.3389/fpsyg.2013.00600

Gentilucci, M., & Bernardis, P. (2007). Imitation during phoneme production. *Neuropsychologia, 45,* 608–615. doi:10.1016/j.neuropsychologia.2006.04.004

Giles, H. (1973). Accent mobility: A model and some data. *Anthropological Linguistics, 15,* 87–109.

Giles, H., Bourhis, R. Y., & Taylor, D. M. (1977). Towards a theory of language in ethnic group relations. In H. Giles (Ed.), *Language, ethnicity, and intergroup relations* (pp. 307–348). London, UK: Academic Press.

Giles, H., Coupland, J., & Coupland, N. (1991). Accommodation theory: Communication, context, and consequence. In *Contexts of accommodation: Developments in applied sociolinguistics* (pp. 1–68). Cambridge, UK: Cambridge University Press.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1166–1183. doi:10.1037/0278-7393.22.5.1166

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

Goldinger, S. D., & Azuma, T. (2004). Episodic memory reflected in printed word naming. *Psychonomic Bulletin & Review, 11,* 716–722. doi:10.3758/BF03196625

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 152–162. doi:10.1037/0278-7393.17.1.152

Goldstein, L., & Fowler, C. A. (2003). Articulatory phonology: A phonology for public language use. In A. S. Meyer & N. O. Schiller (Eds.), *Phonetics & phonology in language comprehension & production: Differences & similarities* (pp. 1–53). Berlin, Germany: Mouton.

Gregory, S. W., Jr., & Webster, S. (1996). A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions. *Journal of Personality and Social Psychology, 70,* 1231–1240. doi:10.1037/0022-3514.70.6.1231

Gregory, S. W., Jr., Dagan, K., & Webster, S. (1997). Evaluating the relation of vocal accommodation in conversation partners' fundamental frequencies to perceptions of communication quality. *Journal of Nonverbal Behavior, 21,* 23–43.

Harrington, J. (2006). An acoustic analysis of "happy-tensing" in the Queen's Christmas broadcasts. *Journal of Phonetics, 34,* 439–457.

Harrington, J., Palethorpe, S., & Watson, C. (2000). Monophthongal vowel changes in Received Pronunciation: An acoustic analysis of the Queen's Christmas broadcasts. *Journal of the International Phonetic Association, 30,* 63–78.

Heldner, M., Edlund, J., & Hirschberg, J. (2010). Pitch similarity in the vicinity of backchannels. In *Proceedings of INTERSPEECH 2010* (pp. 3054–3057). Baixas, France: International Speech Communication Association.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers, 16,* 96–101. doi:10.3758/BF03202365

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93,* 411–428. doi:10.1037/0033-295X.93.4.411

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

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59,* 434–446. doi:10.1016/j.jml.2007.11.007

Johnson, K. (2007). Decisions and mechanisms in exemplar-based phonology. In M. J. Sole, P. Beddor, & M. Ohala (Eds.), *Experimental approaches to phonology: In honor of John Ohala* (pp. 25–40). Oxford, UK: Oxford University Press.

Jones, E., Gallois, C., Callan, V., & Barker, M. (1999). Strategies of accommodation: Development of a coding system for conversational interaction. *Journal of Language and Social Psychology, 18,* 123.

Kendall, T., & Thomas, E. R. (2014). *Vowel manipulation, normalization, and plotting*. R package version 1.2-1.

Kim, M., Horton, W. S., & Bradlow, A. R. (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Laboratory Phonology, 2,* 125–156.

Kučera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). *Tests in linear mixed effects models*. R package version 2.0-25.

Labov, W., Ash, S., & Boberg, C. (2006). *Atlas of North American English: Phonetics, phonology, and sound change*. Berlin, Germany: Mouton de Gruyter.

Levitan, R., & Hirschberg, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proceedings of INTERSPEECH 2011* (pp. 3081–3084). Baixas, France: International Speech Communication Association.

Levitan, R., Benus, S., Gravano, A., & Hirschberg, J. (2015). *Entrainment and turn-taking in human-human dialogue*. Paper presented at the AAAI Spring Symposium, Stanford, California.

Liberman, A. M. (1996). *Speech: A special code*. Cambridge, MA: MIT Press.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1–36.

Louwerse, M. M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive Science, 36,* 1404–1426. doi:10.1111/j.1551-6709.2012.01269.x

Mantell, J. T., & Pfordresher, P. Q. (2013). Vocal imitation of song and speech. *Cognition, 127,* 177–202. doi:10.1016/j.cognition.2012.12.008

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 676–684. doi:10.1037/0278-7393.15.4.676

Miller, R. M., Sanchez, K., & Rosenblum, L. D. (2010). Alignment to visual speech information. *Attention, Perception, & Psychophysics, 72,* 1614–1625. doi:10.3758/APP.72.6.1614

Miller, R. M., Sanchez, K., & Rosenblum, L. D. (2013). Is speech alignment to talkers or tasks? *Attention, Perception, & Psychophysics, 75,* 1817–1826. doi:10.3758/s13414-013-0517-y

Mitterer, H., & Ernestus, M. (2008). The link between perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition, 109,* 168–173.

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. doi:10.3758/s13414-012-0407-8

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

Munson, B., & Solomon, N. P. (2004). The effect of phonological neighborhood density on vowel articulation. *Journal of Speech, Language, and Hearing Research, 47,* 1048–1058.

Namy, L. L., Nygaard, L. C., & Sauerteig, D. (2002). Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology, 21,* 422–432. doi:10.1177/026192702237958

Natale, M. (1975). Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology, 32,* 790–804.

Nguyen, N., Dufour, S., & Brunellière, A. (2012). Does imitation facilitate word recognition in a non-native regional accent. *Frontiers in Psychology, 3,* 480. doi:10.3389/fpsyg.2012.00480

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

Nye, P. W., & Fowler, C. A. (2003). Shadowing latency and imitation: The effect of familiarity with the phonetic patterning of English. *Journal of Phonetics, 31,* 63–79.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1995). Effects of stimulus variability on perception and representation of spoken words in memory. *Perception & Psychophysics, 57,* 989–1001.

Nygaard, L. C., & Queen, J. S. (2000). *The role of sentential prosody in learning voices*. Paper presented at the meeting of the Acoustical Society of America. Atlanta, GA.

Olmstead, A. J., Viswanathan, N., Aivar, M. P., & Manuel, S. (2013). Comparison of native and non-native phone imitation by English and Spanish speakers. *Frontiers in Psychology, 4,* 475. doi:10.3389/fpsyg.2013.00475

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 309–328. doi:10.1037/0278-7393.19.2.309

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America, 119,* 2382–2393. doi:10.1121/1.2178720

Pardo, J. S., & Remez, R. E. (2006). The perception of speech. In M. Traxler & M. A. Gernsbacher (Eds.), *The handbook of psycholinguistics* (2nd ed., pp. 201–248). New York, NY: Academic Press.

Pardo, J. S., Cajori Jay, I., & Krauss, R. M. (2010). Conversational role influences speech imitation. *Attention, Perception, & Psychophysics, 72,* 2254–2264. doi:10.3758/APP.72.8.2254

Pardo, J. S., Gibbons, R., Suppes, A., & Krauss, R. M. (2012). Phonetic convergence in college roommates. *Journal of Phonetics, 40,* 190–197. doi:10.1016/j.wocn.2011.10.001

Pardo, J. S., Cajori Jay, I., Hoshino, R., Hasbun, S. M., Sowemimo-Coker, C., & Krauss, R. M. (2013). The influence of role-switching on phonetic convergence in conversation. *Discourse Processes, 50,* 276–300. doi:10.1080/0163853X.2013.778168

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. doi:10.1016/j.jml.2013.06.002

Paxton, A., & Dale, R. (2013). Argument disrupt interpersonal synchrony. *Quarterly Journal of Experimental Psychology, 66,* 2092–2102. doi:10.1080/17470218.2013.853289

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27,* 169–190.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36,* 49–64. doi:10.1017/S0140525X12003238

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency effects and the emergence of linguistic structure* (pp. 137–158). Philadelphia, PA: Benjamins.

Pierrehumbert, J. B. (2006). The next toolkit. *Journal of Phonetics, 34,* 516–530. doi:10.1016/j.wocn.2006.06.003

Pierrehumbert, J. B. (2012). The dynamic lexicon. In A. Cohn, M. Huffman, & C. Fougeron (Eds.), *Handbook of laboratory phonology* (pp. 173–183). Oxford, UK: Oxford University Press.

Posner, M. I. (1964). Information reduction in the analysis of a sequential task. *Psychological Review, 71,* 491–504. doi:10.1037/h0041120

Postma-Nilsenová, M., & Postma, E. (2013). Auditory perception bias in speech imitation. *Frontiers in Psychology, 4,* 826. doi:10.3389/fpsyg.2013.00826

Putman, W. B., & Street, R. L. (1984). The conception and perception of noncontent speech performance: Implications for speech-accommodation theory. *International Journal of the Sociology of Language, 46,* 97–114.

R Development Core Team. (2015). *R: A language and environment for statistical computing (Version 3.1.3). Vienna, Austria: R Foundation for Statistical Computing*. Retrieved from www.R-project.org

Sanchez, K., Miller, R. M., & Rosenblum, L. D. (2010). Visual influences on alignment to voice onset time. *Journal of Speech, Language, and Hearing Research, 53,* 262–272.

Sancier, M. L., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics, 25,* 421–436.

Sato, M., Grabski, K., Garnier, M., Granjon, L., Schwartz, J. L., & Nguyen, N. (2013). Converging toward a common speech code: Imitative and perceptuo-motor recalibration processes in speech production. *Frontiers in Psychology, 4,* 422. doi:10.3389/fpsyg.2013.00422

Shepard, C. A., Giles, H., & Le Poire, B. A. (2001). Communication accommodation theory. In W. P. Robinson & H. Giles (Eds.), *The new handbook of language and social psychology* (pp. 33–56). New York, NY: Wiley.

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

Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. (1994). Stimulus variability and spoken word recognition: Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America, 96,* 1314–1324.

Street, R. L. (1982). Evaluation of noncontent speech accommodation. *Language & Communication, 2,* 13–31.

Tilsen, S. (2009). Subphonemic and cross-phonemic priming in vowel shadowing: Evidence for the involvement of exemplars in production. *Journal of Phonetics, 37,* 276–296. doi:10.1016/j.wocn.2009.03.004

Vallabha, G. K., & Tuller, B. (2004). Perceptuomotor bias in the imitation of steady-state vowels. *Journal of the Acoustical Society of America, 116,* 1184–1197. doi:10.1121/1.1764832

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

Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language, 31,* 183–194.

Wisniewski, M. G., Mantell, J. T., & Pfordresher, P. Q. (2013). Transfer effects in the vocal imitation of speech and song. *Psychomusicology: Music, Mind, and Brain, 23,* 82–99. doi:10.1037/a0033299

Yu, A. C. L., Abrego-Collier, C., & Sonderegger, M. (2013). Phonetic imitation from an individual-difference perspective: Subjective attitude, personality and "autistic" traits. *PLoS ONE, 8*, e74746. doi:10.1371/journal.pone.0074746