Speech Communication 133 (2021) 1–8

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Perceptual learning of phonetic convergence

James W. Dias [a], Theresa C. Vazquez [b], Lawrence D. Rosenblum [b,*]

[a] Medical University of South Carolina, United States
[b] University of California, Riverside, United States

ARTICLE INFO

Keywords:
Phonetic convergence
Talker familiarity
Talker idiosyncracies
Perceptual learning

ABSTRACT

Phonetic convergence involves a talker subtly sharing their speaking style with another talker. It is known that speaking style is something perceivers can learn in order to better recognize a talker and a talker's speech. The question arises, can perceivers also learn to better recognize when a talker's speaking style is being shared with another speaker during convegence? To test this question, two convergence experiments were conducted to determine if perceivers (raters) could improve their ability to recognize when talkers were shadowing a specific model. The results showed that perceivers could improve this ability and this improvement generalized to new words and new shadowers of the same model. However, a follow-up experiment showed that this improvement did *not* generalize to new models and shadowers. This final result suggests that improvement is dependent on learning the shared speaking style of a specific model.

1. Introduction

Perceivers are sensitive to the talker-specific phonetic details of speech (for a review, see Smith, 2015). Research has shown that using the phonetic realizations specific to talkers (e.g., coarticulatory style; voice-onset time; vowel space; nasality) supports the learning required for talker recognition and for the facilitation of speech identification from a familiar talker (Nygaard, 2005; Smith, 2015). Arguably, phonetic convergence – the subtle and inadvertent imitation of an interlocuter's speaking style – also evidences perceptual sensitivity to many of these same phonetic realizations. In this sense, convergence can be considered the *sharing* of talker-specific phonetic details, with *sharing* defined as the partial adoption of some common phonetic detail produced by interlocuters (or models and shadowers). The sharing of such idiolectic information between talkers and the perceptibility of this sharing by others could facilitate social interactions and provide information regarding the social relationships between interlocutors (e.g., Babel et al., 2014; Giles et al., 1991; Pardo, 2006; Shepard et al., 2001). The question arises, just as perceivers can learn to better recognize talkers and their speech, can they also learn to better recognize when talker dimensions *are shared* in the context of convergence? Below, we describe two phonetic convergence experiments that provide an initial examination of this question.

1.1. Phonetic convergence of talker-specific phonetic details

Decades of research have shown that people unconsciously and subtly imitate the verbal and non-verbal behaviors of conversational partners (for reviews, see Heyes, 2011; Lakin et al., 2003; Pardo et al., 2017). Nonverbally, interlocutors are known to imitate each other's posture, gestures, and mannerisms. Verbally, interlocutors imitate each other's words, syntax, prosody, and pronunciation of speech segments. This last dimension—known as *phonetic convergence*—can occur at the segment and featural levels of speech based on either auditory or visual (lipread) information (Goldinger, 1998; Miller et al., 2010; Pardo, 2006; Nielsen, 2011; Sanchez et al., 2010).

Like other forms of unconscious imitation, research shows that phonetic convergence can be both spontaneous and ubiquitous (for reviews, see Babel et al., 2014; Dias and Rosenblum, 2016). For example, Pardo (2006) found that when performing an interactive map task, interlocutors often imitate aspects of each other's utterances. This phonetic convergence can also occur when an isolated participant simply listens to a recording of a talker. In a seminal study, Goldinger (1998) asked participants to listen to a model say words and then say those words themselves after hearing each, a task known as *shadowing*. Participants were never instructed to imitate, or even repeat the model, but rather to simply say aloud each word they heard. Still, participants' shadowing response words were rated as more similar to the model's

Research supported by National Science Foundation Grant #1632530 awarded to the third author.
* Corresponding author at: Department of Psychology, University of California, 900 University Ave., Riverside, CA 92521, United States.
E-mail address: lawrence.rosenblum@ucr.edu (L.D. Rosenblum).

https://doi.org/10.1016/j.specom.2021.07.004
Received 1 July 2020; Received in revised form 21 June 2021; Accepted 6 July 2021
Available online 15 July 2021

spoken words, relative to baseline words uttered by the participants before the shadowing task.

Both interactive- and shadowing-based phonetic convergence has been replicated multiple times and has been shown to work with both auditory and visual speech (Dias and Rosenblum, 2011, 2016; Miller et al., 2010; Sanchez et al., 2010). Like nonverbal imitation, phonetic convergence is always partial, with the degree of observable convergence dependent on many factors, including the social relationship between the interlocutors, cultural affiliation, facial attractiveness, and gender (e.g., Pardo et al., 2012; and for a review, see Babel et al., 2014).

Evaluation of phonetic convergence is typically conducted in two ways (Pardo et al., 2013). First, acoustic similarities between talker utterances can be evaluated. Research has shown that utterance duration (e.g., Brouwer et al., 2010), fundamental frequency (F0) (e.g., Babel and Bulatov, 2012), vowel spectra (e.g., Babel, 2012; Honorof et al., 2011), and voice onset time (VOT) (e.g., Nielsen, 2011; Sanchez et al., 2010) can become more similar as talkers converge.

More often however, phonetic convergence is evaluated using a set of naïve raters (e.g., Babel et al., 2014; Dias and Rosenblum, 2016; Goldinger, 1998; Goldinger and Azuma, 2004; Miller et al., 2010; Namy et al., 2002; Pardo, 2006; Pardo et al., 2012). Typically, these raters evaluate whether a shadowed utterance or another utterance of the same word is more similar to the shadowed model's utterance. Raters typically choose the shadowed utterance as a better match at greater than chance levels.

Using a rater method of evaluating convergence can provide useful information not available with acoustic measures (e.g., Pardo et al., 2013). Human raters can provide a more holistic and flexible evaluation of the different dimensions that can converge (e.g., Pardo et al., 2013). There are myriad acoustic dimensions on which talkers can converge, and it is likely that different talkers converge based on different subsets of these dimensions. This renders identifying the correct acoustic measures of convergence prohibitive and favors the flexibility with which human listeners may hear different types of interlocutor similarities. Further, if phonetic convergence does have some social relevance, as some have argued, then its existence should be *perceptible* (e.g., Babel and Bulatov, 2012; Goldinger, 1998; Pardo, 2006). As naïve raters can perceive phonetic convergence with some consistency, it would seem it is perceptible, at least to some degree. This suggests that raters do have some ability to determine when talker-specific dimensions are shared across individuals. What is not yet known, is whether raters can improve their ability to perceive shared talker-specific dimensions resulting from convergence and, if so, whether improvement can generalize to new words, shadowers, and models. The current experiments test these questions.

To which dimensions do talkers converge? As intimated, acoustic measures of convergence have shown that talkers can partially imitate each-other's utterance duration, F0, intensity, and vowel space, as well as more sub-phonemic, featural attributes such as VOT (Pardo et al., 2013). However, convergence can also be induced by non-acoustic realizations of speech. Research shows that visible (lipread) speech can modulate convergence: a) on its own (Miller et al., 2010); b) when integrating with auditory speech (Sanchez et al., 2010); and c) as a means to increase the degree of convergence over auditory speech alone (Dias and Rosenblum, 2011, 2016). These visual speech findings suggest that convergence may be more correctly conceived as not being toward acoustic dimensions as such, but toward a talker's articulatory (speaking) style – which can be conveyed either acoustically or visually (see also Honorof et al., 2011).

Importantly, there is also evidence that converged featural dimensions can generalize over non-heard segments (e.g., Nielsen, 2011, 2014; Zellou et al., 2017). Nielsen (2011) exposed participants to a talker whose words contained initial /p/ segments with artificially-lengthened VOTs (see also Shockley et al., 2004). After exposure, participants were asked to produce words containing initial /p/ segments, as well as words with initial /k/ segments. Results showed that participants extended their own VOTs in both /p/ and /k/ segment productions. This finding suggests that participants converged to the talker's initial consonant VOTs in general, and not just to those in the segments (/p/) that were actually heard. Similar findings have been observed for coarticulatory vowel nasalization (Zellou et al., 2017, 2016). The fact that participants converge to *generalized* sub-phonemic features suggests that they are incorporating talkers' talker-specific phonetic dimensions into their production responses. If a talker is, for example, heard articulating with longer VOTs for some initial voiceless stop consonants, a perceiver will tacitly assume that lengthened initial VOTs is a property of that talker's speaking style and will show a tendency to converge by subtly lengthening VOTs for all initial voiceless stop consonants. This type of featural generalization is also known to play a role in talker recognition and talker-facilitated speech perception (see below).

There is also evidence that perceivers, acting as convergence raters, can detect when talkers *share* VOTs as a result of convergence. Shockley et al. (2004) artificially lengthened the VOTs of a model's words and found that shadowers of these words lengthened their own VOTs (based on acoustic measures). In addition, the researchers found that raters could correctly judge which of these utterances were from shadowing vs. baseline (read aloud) versions, and did so with more accuracy than for stimuli that did not include extended VOTs. Potentially, raters made their judgments based, at least in part, on the lengthened VOTs produced by the shadowers.

These findings suggest that perceivers, in the form of convergence raters, can use VOT to detect shared talker-specific phonetic detail. This fact bodes well for the possibility that perceivers can *improve* this skill, as they are known to use VOT for improving talker recognition and talker-facilitated speech perception, as discussed below.

1.2. Learning of talker-specific phonetic information

While no research has examined if perception of (convergence-based) *shared* talker-specific phonetic dimensions can be improved, there is substantial evidence that perceivers can improve their ability to detect and use talker-specific information (for a review, see Smith, 2015). Potentially, improving perception of this information would allow listeners to better recognize when that talker-specific information is shared *between* talkers as a result of phonetic convergence.

There is a large literature showing that experiment participants can be taught to identify talkers (for reviews, see Schweinberger et al., 2014; Smith, 2015). The ease with which participants can learn to identify talkers depends on numerous factors, including the number of voices in a stimulus set, and whether the voices are presented along with their associated (talking) faces during training (e.g., Sheffert and Olson, 2004; von Kriegstein et al., 2006). There is also evidence that perceivers can implicitly learn talker information – the phonetic (articulatory) characteristics unique to individual talkers' speech – in the service of some other task (for reviews, see Creel and Bregman, 2011; Nygaard, 2005). Becoming familiar with a talker's voice allows perceivers to better recognize the speech of that talker (e.g., Nygaard and Pisoni, 1998; Nygaard et al., 1994). Also, words can be remembered more easily if the voice used for training and test are the same (e.g., Church and Schacter, 1994; Palmeri et al., 1993). Such talker familiarity effects suggest that the implicit learning of a talker's phonetic characteristics can facilitate processing of speech spoken by the familiar talker.

Regarding the form of the information used for these skills, it is known that voice learning can make use of myriad levels of information including acoustic, phonetic, lexical, semantic, and syntactic (e.g., Smith, 2015; Zarate et al., 2015). Voice quality and fundamental frequency are often cited as important acoustic dimensions for recognizing and identifying talkers. However, it is known that such talker learning can be performed when these dimensions are not available. For example, Sheffert et al. (2002) used sine wave speech re-synthesis to remove a majority of the acoustic dimensions typically thought to facilitate talker

learning (including voice quality and fundamental frequency). The technique involves reducing recorded talkers' signals to three pitch-varying sine waves that track speech formants (e.g., Remez et al., 1981). Previous research showed that while these types of stimuli do not contain normal voice or pitch characteristics, they can be understood as speech and be used to recognize familiar talkers (e.g., Remez et al., 2007, 1997; see also van Heugten et al., 2014). Sheffert et al. (2002) showed that perceivers could successfully learn to identify 10 novel sine wave talkers through a feedback-training task. Other research has shown that gaining familiarity with a talker's sine wave speech signals can be used to facilitate understanding of that talker's natural speech (e.g., Remez et al., 2011, 2018). While these stimuli are highly reduced, it has been speculated that they retain talker-specific phonetic information (Remez et al., 1997). If so, then this research suggest that perceivers can improve their ability to recognize and identify talkers and their speech using talker-specific phonetic information (sometimes called idiolect: Remez et al., 2007; Rosenblum et al., 2007b). This fact may bode well for learning to detect when talkers *share* this information in convergence contexts.

There is also research suggesting overlap in the talker-specific phonetic dimensions important for talker learning (and talker-facilitation of speech) and that show convergence. Recall that talker's VOT has been shown to shift toward that of the talker they hear (Nielsen, 2011; Sanchez et al., 2010; Shockley et al., 2004). Further, this shift occurs for produced VOTs of segments not previously heard from a talker (Nielsen, 2011, 2014). It turns out that characteristic VOT can similarly be used by perceivers for purposes of learning to recognize a talker and that talker's speech (e.g., Allen and Miller, 2004; Theodore and Miller, 2010; and for a review, see Smith, 2015). In addition, as in the context of convergence, learning a talker-specific VOT can generalize to recognition of a talker when they produce other voiceless stop consonants not previously heard (e.g., Theodore and Miller, 2010). Similar results have been observed for talker familiarity effects based on idiosyncratic vowel productions and phonetic detail at word boundaries (Dahan et al., 2008; Smith and Hawkins, 2012). Thus, for both phonetic convergence and learning of talker-specific phonetic details, idiosyncratic featural dimensions can generalize from heard to unheard segments. These facts also bode well for perceivers learning to better detect the sharing of talker-specific dimensions during phonetic convergence.

Finally, like for phonetic convergence, talker-specific phonetic details conveyed *visually* can be salient for talker and talker-facilitated speech perception. Our own laboratory has used a point-light methodology to isolate visual speech information (Rosenblum and Saldaña, 1996). This method produces videos comprised of white points moving against a black background, with the point motions following the articulations of the talker. Despite the reduced nature of these films, talkers can be recognized by their friends, and by novice perceivers (in a 2AFC matching task: Rosenblum et al., 2007b, 2002). Other research shows that it is easier to lipread from familiar faces, even for observers who have no formal lipreading experience (e.g., Lander and Davies, 2008; Schweinberger and Soukup, 1998; Yakel et al., 2000). Finally, one's experience with a talker can cross modalities to facilitate speech perception: Experience lipreading a talker for a period of time facilitates later comprehension of that talker's auditory speech and vice versa (Rosenblum et al., 2007a; Sanchez et al., 2013). Taken together, this research suggests that as is true for phonetic convergence, the relevant informational dimensions for talker learning and facilitation are best considered as talker-specific articulatory properties that can be conveyed auditorily or visually.

In sum, it is unclear whether phonetic convergence raters use the same strategies and information as perceivers in talker recognition and talker-facilitated speech experiments. However, as discussed above, there are hints that some of the same talker-specific phonetic dimensions are used for both purposes. To the degree that this is true, it would be predicted that perceivers (raters) should be able to improve their ability to detect the sharing of talker dimensions resulting from convergence. This skill may also provide a benefit for improving the aforementioned detection of the social alliances known to be related to phonetic convergence.

1.3. The current study

The current studies examine whether perceivers —performing as raters— can improve their ability to detect the sharing of talker-specific phonetic dimensions resulting from convergence. In addition, the experiments explore the degree to which this improvement can then generalize to new words, new shadowers, and new models. The experiments use performance feedback to train perceivers to better recognize that an utterance was a shadow of a heard model in the context of an AXB task. Following this training, perceivers then rated the similarity of new shadowed utterances spoken by new shadowers of trained models (Experiment 1) or new shadowers of new models (Experiment 2). Response feedback is a common learning tool used in many talker identification and talker-facilitated speech experiments (e.g., Sheffert et al., 2002; Theodore and Miller, 2010).

Four hypotheses are tested. First, if perceivers (raters) are able to use talker-specific phonetic information for convergence detection in the same manner used for talker identification, talker recognition, and talker-facilitated speech perception, then feedback training should improve their ability to detect convergence. Secondly, if this training works at the level of talker-specific *phonetic detail* (as it does for talker identification, talker recognition, and talker-facilitated speech perception), then improved convergence detection performance should be sustained when new words are heard from the same talkers. Thirdly, if perceiver strategies for improving detection of convergence involve learning talker-specific phonetic characteristics of some type, then improvement should be sustained if a trained talker remains in the stimulus set (and the others change). Finally, if performance improvement is dependent on learning talker-specific phonetic characteristics, rather than simple, general experience with an AXB matching task, then improved performance should *not* be sustained if all new talkers' (models and shadowers) utterances are judged.

2. Experiment 1

The purpose of Experiment 1 was to examine whether perceivers could learn to: a) better detect phonetic convergence between a model and set of shadowers; b) generalize this improvement to a new set of words; and c) generalize this performance with a new set of shadowers of the same model.

Participants were asked to perform an AXB matching test of model and shadower words. In much phonetic convergence research, participants who perform AXB matches are referred to as 'raters' and are typically treated as a methodological tool for evaluating the convergence of shadowers to models (or interlocutors to one another), with the shadowers (or interlocutors) treated as the study participants. However, in the current research, the raters themselves are the critical participants, and we will refer to them as raters or perceivers. For the present experiment the AXB matching task involved judging which of two shadowers' words (A or B) sounded most like the same word uttered by a model (X) (e.g., Miller et al., 2013). Thus, on each trial, participants heard three utterances of the same word. The middle (X) utterance was produced by a model, and the first and third (A or B) utterances were produced by 1) a shadower who had shadowed the utterance produced by the model heard in the X position; and 2) a shadower who had shadowed an utterance produced by a *different model* (whose word was *not* heard in the X position; see details below) (e.g., Miller et al., 2013). If perceivers correctly chose the utterance of the shadower who had shadowed the model heard in the X position, then they would be considered correct in their convergence judgment.

2.1. Method

Participants. Thirty ethnically/racially diverse undergraduate students (age range 18-22 years; 15 female, 15 male) of the University of California, Riverside, acted as the critical participants/raters in the experiment for course credit or $10. All participants had normal or corrected-to-normal vision and normal hearing. All participants spoke American English as a first language, were largely California native, and were naïve to the purposes of the study.

Stimuli. Creation of stimuli involved a two-part process. First, we used audio recordings of two female models saying 74 low-frequency, two-syllable words (e.g. kitten, pencil, tulip). The two female models were in their mid-twenties, spoke American English as a first language, were native to California, and were naïve to the purpose of the study. The words were adapted from the Shockley et al. (2004) word list, and the stimuli have been previously used to successfully elicit phonetic convergence detection (e.g. Miller et al., 2010).

Next, eight female undergraduates acted as shadowers. These shadowers spoke American English as a first language, were native to California, and were naïve to the purpose of the study. Female shadowers and models were initially chosen because of evidence that they show more consistent phonetic convergence behavior (e.g., Babel et al., 2014; Nye and Fowler, 2003). However, Experiment 2 also tested male shadowers and models to determine if the observed effects would generalize. Shadowers were Introductory Psychology students of the University of California, Riverside, who received course credit for their participation. Half of the shadowers listened to the recordings from one model, and half listened to the recordings from the other model. All shadowers were randomly assigned to the models. The shadowers listened to each word of the model and then said the word aloud. Shadowers were never instructed to imitate the model, or even to "repeat" what they heard. This procedure has previously been shown to produce phonetic convergence (e.g., Goldinger, 1998).

Models and shadowers spoke into a Shure SM57 microphone and were audio recorded at 44 kHz (16 bits) with Amadeus II software in a sound-attenuated chamber. Recordings were digitized and edited using FinalCut Pro software. Stimuli were presented through Sony MDR-V6 headphones using PsyScope software.

Procedure. In a single AXB trial, participants heard one shadower saying a word aloud, the same word said aloud by the model whom the shadower shadowed, and the same word said aloud by a different foil shadower (who had shadowed a different model). The order of the shadower and foil shadower (as either the first or third utterance) was counterbalanced across trials. Thus, on each trial, only one of the two shadowers (presented first or third) had actually shadowed the model in the middle (e.g., Miller et al., 2013). Testing a comparison utterance from a foil shadower (as they shadowed another model) has been used successfully in previous studies and avoids some of the concerns over using a comparison utterance that has been read aloud from a list of written words (for a discussion, see Miller et al., 2013).

Participants were told that for each trial, they would be hearing three talkers utter the same word. They were asked to determine which of the first or third utterances sounded most like the second utterance. They were asked to pay particular attention to the pronunciation of words when judging similarity. Participants then listened to stimuli via headphones and responded by pressing one keyboard button if the first utterance sounded most like the second, and a different button if the third utterance sounded most like the second. Stimulus order was randomized without replacement for word and for order of correct choice (A or B) in the AXB task. A pilot experiment (without response feedback) established that our stimuli and methods could induce phonetic convergence and that participants could correctly match models with their shadowers with 64% average accuracy.

The experiment was run in two phases. During the training phase, participants listened to the two models and four shadowers (two per model). Across trials, shadowers were always paired with the same foil shadower (e.g., Miller et al., 2013). Participants heard 37 different words in 148 trials during the training phase. Each word was spoken by each model-shadower pair in both presentation orders (with the target shadower of the model [X] positioned as either the first [A] or last [B] utterance in the AXB trial). During training, participants received feedback about whether or not each of their responses was correct: The word "correct" appeared in green, or the word "incorrect" appeared in red after each judgment was made. Training with feedback was the same across participants: All participants received feedback on all 148 trials of the training phase of the experiment. No threshold in performance was required to move on to the testing phase.

During the testing phase of the experiment, participants again heard the two models but now with four *new* shadowers. During this phase, participants listened to 37 *new* words in 148 trials. Orders of stimuli were counterbalanced for words presented at training and testing and for shadowers presented at training and testing. Counterbalancing was implemented for these parameters, as presentation of all possible iterations of stimuli would be unduly cumbersome. No performance feedback was presented during the testing phase.

For data scoring, a correct response was coded when a participant chose the utterance produced by the shadower that had actually shadowed the model (presented as the X token).

2.2. Results

Performance in the training phase was highly related to performance in the testing phase of the experiment, $r(28)=0.817$, $p<0.001$. We fit logistic/binomial mixed effects models to the data across the training and testing phases of the experiment. First, we fit a control model that included participant as a random parameter. This control model yielded a significant intercept, indicating that the rate at which raters correctly matched the models with their shadower ($M = 0.637$, $SE=0.005$) was above chance, $B = 0.580$, $SE = 0.068$, $Z = 8.481$, $p < 0.001$. Adding experimental phase (training vs. testing phase) to the control model improved model fit, $\chi^2(1) = 5.106$, $p = 0.024$. The rate at which raters correctly matched the models with their shadower improved from the training ($M = 0.626$, $SE=0.007$) to testing ($M = 0.648$, $SE=0.007$) phases of the experiment, $B = 0.101$, $SE = 0.045$, $Z = 2.260$, $p = 0.024$. Adding task trial (trial 1 to 148) to the model with experimental phase also improved model fit, $\chi^2(2) = 10.895$, $p = 0.004$. The rate at which raters correctly matched the models with the shadowers improved over the course of the shadowing task, $B = 2.394\times10^{-3}$, $SE = 7.375\times10^{-4}$, $Z = 3.246$, $p = 0.001$. The interaction between experimental phase and task trial was significant, $B = -2.828\times10^{-3}$, $SE = 1.049\times10^{-3}$, $Z = -2.696$, $p = 0.007$.

To explore this interaction, separate logistic/binomial mixed effects models were fit to the training and experimental phases including participant as a random parameter and task trial as a fixed parameter. Over the course of the training phase, raters improved in their ability to match models with their shadowers, $B = 2.407\times10^{-3}$, $SE = 7.394\times10^{-4}$, $Z = 3.255$, $p = 0.001$. However, over the course of the testing phase, raters failed to demonstrate any improvement in their ability to match models with their shadowers, $B = -4.326\times10^{-4}$, $SE = 7.447\times10^{-4}$, $Z = -0.581$, $p = 0.561$. The results suggest that raters only improved in their ability to match models with their shadowers when trial feedback was provided. However, the improvement resulting from training with feedback lead to sustained improved performance when later matching the same models with different shadowers and using different words (see Fig. 1).

Results from Experiment 1 indicate that: a) perceivers could indeed learn to better detect phonetic convergence in the presence of feedback; b) this learning transferred to new words; and c) this learning transferred to new shadowers. Potentially, the fact that learning generalized to convergence judgments based on new words and shadowers could suggest that our participants were learning to better perceive the shared talker-specific phonetic dimensions of the model. However, there is an

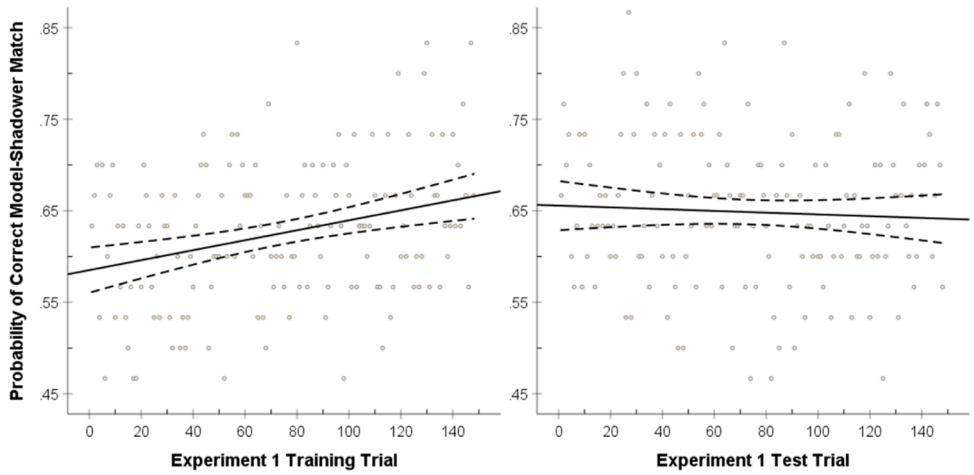

Fig. 1. Average trial-by-trial performance for the training and testing phases of Experiment 1. Performance improved over the course of training with feedback. This training resulted in overall improved performance in the testing phase. Solid lines represent the linear relationship between trial and the probability of correctly matching a model with their shadower. Dashed lines represent the 95% confidence interval.

alternative explanation for the current findings. It could be that the feedback provided during training simply allowed participants to better learn the *general task* of detecting convergence, regardless of the talkers. These alternatives are examined in the next experiment.

3. Experiment 2

The purpose of Experiment 2 was to examine the generalizability of the Experiment 1 results. Do improvements in similarity-matching of models with shadowers transfer to new shadowers *and* models? The experiment was designed to address multiple questions. First, we wanted to determine if the feedback-based learning during the training phase would replicate with some additional models and shadowers (see below). Second, we wanted to explore whether this learning could then transfer to a *test* phase composed of completely novel stimuli. Thus in Experiment 2, the test phase was composed of new words, spoken by new models and shadowers who were not heard during the training phase.

Based on the results of Experiment 1, participants would be expected to improve their performance across the training phase. However, regarding test phase performance, if training simply serves to improve participants' *general* ability to detect convergence, then an improvement during training should be maintained through the test phase, despite the change in models and shadowers. Alternatively, if the training serves to improve participants' ability to better detect shared talker-specific information, then test performance with new models and shadowers should *not* benefit from the training and performance should revert during the test phase.

3.1. Methods

Participants. Thirty-six ethnically/racially diverse undergraduate students (age range 17-23 years; 16 female, 20 male) of the University of California, Riverside, acted as participants/raters in the experiment for course credit or $10. All participants had normal or corrected-to-normal vision and normal hearing. All participants spoke American English as a first language, were largely California native, and were naïve to the purposes of the study. None of these individuals had participated in Experiment 1.

Stimuli. The stimuli used in Experiment 1 were also used for Experiment 2. In addition, new stimuli were recorded. Stimuli from two new models, and eight new shadowers, all male, were added to the stimulus set used in Experiment 1. Like the female models used in Experiment 1, the two new models were also in their mid-twenties, spoken American English as a first language, and were naïve to the purpose of the study. Though the male models were not California native, they were native to North Atlantic states and exhibited no conspicuous accent, based on the experimenters' impressions. The eight new shadowers were Introductory Psychology students of the University of California, Riverside, who received course credit for their participation. All shadowers spoke American English as a first language, were native to California, and were naïve to the purposes of the study. Like for Experiment 1, all shadowers were randomly assigned to the models and half of the shadowers listened to the recordings from one male model, and half listened to the recordings from the other male model. The shadowers listened to each word of the model and then said the word aloud. Shadowers were never instructed to imitate the model, or even to "repeat" what they heard. These new recordings allowed us to create a second set of stimuli to add to our first. This new set of stimuli also allowed us to test completely different stimuli at training and test.

Procedure. The procedure used in Experiment 2 was identical to that used in Experiment 1 (including training feedback). Again, orders of stimuli were randomized within each phase (training and testing) for word and for order of correct choice in the AXB task. As in Experiment 1, all participants listened to different words during the training phase than during the testing phase. Orders of stimuli were counterbalanced (this time for words, shadowers, *and* models presented at training and testing). Male models were always paired with male shadowers, and female models were always paired with female shadowers.

During training, participants performed matches using utterances of shadowers and non-shadowers of one male model and one female model. During testing, participants rated the similarity of *new* shadowers and non-shadowers of one *new* male model and one *new* female model. Participants *never* heard the same person speak in both the training and testing phases of the experiment.

3.2. Results

Performance in the training phase was related to performance in the testing phase of the experiment, $r(34)=0.543$, $p<0.001$. We fit logistic/binomial mixed effects models to the data across the training and testing phases of the experiment. As before, we fit a control model that included participant as a random parameter. This control model yielded a significant intercept, indicating that the rate at which raters correctly matched models with their shadowers ($M = 0.639$, $SE=0.005$) was above chance, $B = 0.605$, $SE = .084$, $Z = 7.243$, $p < 0.001$. Adding experimental phase (training, testing) to the control model improved model fit, $\chi^2(1) = 16.014$, $p < 0.001$. The rate at which raters correctly

matched the models with their shadower declined from the training (M = 0.657, SE=0.007) to testing (M = 0.620, SE=0.007) phases of the experiment, B = -0.166, SE = 0.041, Z = -4.001, $p < 0.001$. Adding task trial (trial 1 to 148) to the model with experimental phase also improved model fit, $\chi^2(2) = 19.197$, $p < 0.001$. The rate at which raters correctly matched the models with their shadower improved over the course of the shadowing task, $B = 3.066\times10^{-3}$, $SE = 7.008\times10^{-4}$, $Z = 4.374$, $p < 0.001$. The interaction between experimental phase and task trial was significant, $B = -3.004\times10^{-3}$, $SE = 9.771\times10^{-4}$, $Z = -3.074$, $p = 0.002$.

To explore this interaction, separate logistic/binomial mixed effects models were fit to the training and experimental phases including participant as a random parameter and task trial as a fixed parameter. Over the course of the training phase, raters improved in their ability to match models with their shadowers, $B = 3.139\times10^{-3}$, $SE = 7.067\times10^{-4}$, $Z = 4.442$, $p < 0.001$. However, over the course of the testing phase, raters failed to demonstrate any improvement in their ability to match the models with their shadowers, $B = 7.927\times10^{-5}$, $SE = 6.822\times10^{-4}$, $Z = 0.116$, $p = 0.907$. As with Experiment 1, the results suggest that raters only improved in their ability to match models with their shadowers when actively provided feedback. The results further suggest that the improvement resulting from training with feedback does *not* lead to improved performance when later matching different models with their shadowers (see Fig. 2). The improvements resulting from feedback training do not seem to improve general model-shadower matching. Instead, feedback training seems to improve familiarity with the shared talker-specific phonetic dimensions of the models on which the raters are trained, resulting in improved matching based on those idiosyncratic characteristics.

4. Discussion

In two studies, we found that perceivers could learn to better detect phonetic convergence. Experiments 1 and 2 both showed that performance improved across training when participants received feedback. We also found in Experiment 1 that learning generalized to new words and to new shadowers of the same models. In Experiment 2, we found that learning did not transfer when new models—along with new shadowers and words—were involved. In Experiment 2, the gains achieved during training did not lead to improved convergence detection of new models and their shadowers. This suggests that the maintained learning observed in Experiment 1 was not simply based on participants becoming familiar with the AXB matching task.

Instead, the results suggest that participant improvements in detecting phonetic convergence were related to the models heard during training. Also, the fact that this improvement generalized to new words suggests that participants were learning to detect talker properties not related to specific utterances or words. Taken together, these results are consistent with the notion that participants learned to better detect talker-specific phonetic dimensions of the models' speech during training, as well as when these properties were shared by shadowers in the context of convergence.

The fact that learning generalized to new shadowers of the models suggests that whatever these properties may be, perceivers may possess some flexibility in learning to detect those properties across shadowers with different vocal tracts and natural idiolects. While many previous convergence studies have shown that raters can detect convergence between a model and multiple shadowers (e.g., Goldinger, 1998; Miller et al., 2013), the present study is unique in showing that *improvements* in detecting convergence can also generalize to multiple shadowers of a model. Again, this suggests that participants are learning to better perceive shared talker-specific phonetic dimensions.

4.1. What might perceivers be learning about shared talker-specific phonetic information?

The question naturally arises of the type of talker-specific dimensions of convergence perceivers learned during training. While these studies were designed as a first pass to simply test whether perception of shared talker-dimensions *could* be improved, some speculation is warranted. In a scenario ideal for a participant, a model could, in principle, produce speech with a few unique dimensions that are then imitated by *all* shadowers. For example, a given model may have uttered words with, say, long VOTs and hypernasalized vowels, and then all shadowers of that model imitated these same dimensions. When then faced with an AXB trial, a perceiver could then choose the utterance of the two (A or B) that was spoken with the longest VOTs and hypernasalized vowels. Learning, in this case, could involve a perceiver better attending to these dimensions of a particular model that is then replicated by all shadowers of that model.

As implied earlier however, this simple scenario is very unlikely to occur. Nor is it likely that shadowers' convergence to our models was driven by dialect or accent, considering the overall homogeneity of our samples. Recall that previous research shows that the phonetic characteristics to which talkers converge differs across models and across shadowers of a given model (e.g., Pardo et al., 2013). Furthermore, the properties that are used to detect convergence will differ across raters. It is likely then, that the properties providing the basis of perceptual learning for the perceivers in these experiments also differed across models, shadowers, and perceivers, themselves. Given these multiple levels of variability, it is unlikely that single simple dimensions of a model were converged and detected across the experiment. In fact, it is this exact variability that has led many convergence researchers to

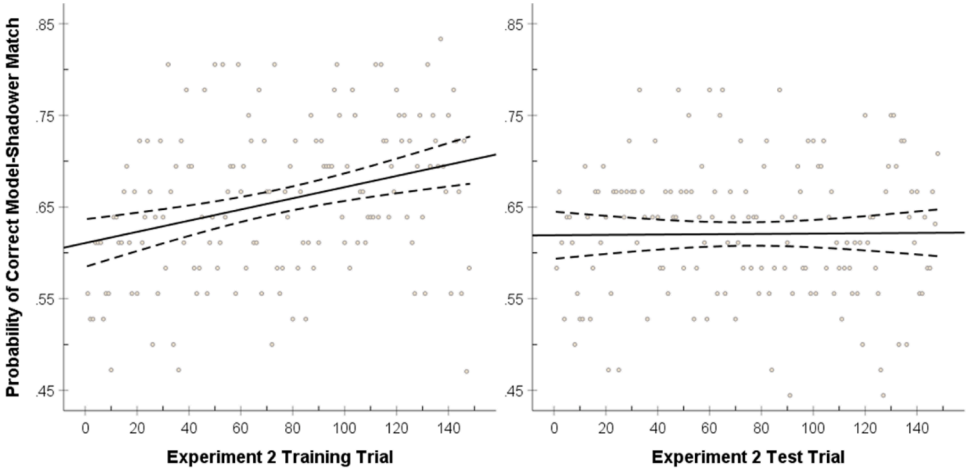

Fig. 2. Average trial-by-trial performance for the training and testing phases of Experiment 2. Performance improved over the course of training with feedback. However, the improved performance did not result in better model-shadower matching when matching new models to their shadowers in the testing phase. Solid lines represent the linear relationship between trial and the probability of correctly matching a model with their shadower. Dashed lines represent the 95% confidence interval.

depend on perceptual rating judgments instead of, or in addition to acoustical measures (for a review, see Pardo et al., 2013).

Regardless, this variability makes it difficult to infer which specific dimensions our participants were learning in our experiments. Based on the fact that they can generalize their learning to new shadowers, it is likely that perceivers were learning some constellation of talker-specific dimensions of the models. Learning a constellation of dimensions would then allow perceivers to generalize their improvements in detecting convergence for a model to a new set of shadowers of that model, who possess different natural talking styles and vocal tract dimensions. Again however, the learning of these constellations seems to be model-specific: improved performance was not maintained when new models (and shadowers) were involved (Experiment 2).

A related question is whether the convergence learning observed here would generalize to a situation in which talkers that have very different speaking characteristics. Recall that while in both experiments shadowers were randomly assigned to models (of the same gender), most all of our models, shadowers, and raters were from Southern California. The question arises of whether raters/perceivers would show similar convergence learning between talkers of different genders, dialects, or foreign accents. While it is known that phonetic convergence can occur in all of these contexts (for a review, see Pardo et al., 2017), the degree to which it occurs can differ. It could be that when learning to perceive convergence for talkers of very different voice characteristics, learning will not be as systematic as we observed in the current study and will not allow generalization to other shadowers. Future research can examine this question.

4.2. Social implications of improving phonetic convergence detection

As intimated, phonetic convergence may have relevance for social interactions. If so, then improving one's ability to detect convergence may provide some social psychological benefits. Recall that one reason phonetic convergence is often measured with rater matching is that the social relevance of the phenomenon should render it perceptible by outside perceivers (e.g., Babel and Bulatov, 2012; Goldinger, 1998; Pardo, 2006). As naïve raters can perceive phonetic convergence with some consistency and, as shown here, learn to improve this skill, the social ramifications of detecting convergence may be very real.

It may be a benefit to an outside observer if they can (either consciously or unconsciously) detect convergence between other people and then improve this skill with practice. Convergence may offer important information about social alliances – and the ability to detect it could help perceivers navigate the social world, even if unconsciously. If observers can detect imitation during interpersonal communication, whether verbal or nonverbal, it might suggest who is friends with whom and even which people are social leaders (Giles et al., 1991; Pardo et al., 2012; Shepard et al., 2001). If such skills could rapidly be improved and adapted in a new environment, better still.

In fact, there is evidence that outside observers can perceive rapport by viewing *nonverbal* behavior (e.g., Grahe and Bernieri, 1999; Kavanagh et al., 2011). Children can also use nonverbal cues to judge the rapport of interacting adults (e.g., Over and Carpenter, 2015). There is also evidence that adult observers can *improve* their judgments of interlocutor rapport through performance feedback (e.g., Gillis et al., 1995). The basis of this improvement seems at least partially related to an observer's ability to detect nonverbal mimicry and synchrony between interlocutors (e.g., Bernieri et al., 1996). Because phonetic convergence may also signal rapport between interlocutors (e.g., Pardo et al., 2012), learning to better detect this form of imitation might also benefit an outside observer.

Of course, improvements in convergence may be based on the same mechanisms used for talker recognition, identification, and the advantage in perceiving speech from a familiar talker. As stated, our participants may have learned to better detect the talker-specific phonetic dimensions of our models, which in turn, allowed them to more easily hear when another talker imitated those characteristics. In fact, this interpretation is consistent with the results of Experiment 2 for which learning did not generalize when new models were involved. This interpretation is also supported by the aforementioned research showing that talker-learning can occur through feedback and that this learning can be based on talker-specific phonetic properties (for a review, see Smith, 2015)—likely important for convergence.

Thus, it is quite possible that learning to better detect convergence exists as both an instance of talker familiarity while also serving a social purpose. In fact, the social function of convergence learning could be considered akin to Gould's *exaptation*, for which an attribute/skill arising originally for one purpose is co-opted for a new purpose not directly related to that original purpose (e.g., Buss et al., 1998; Gould, 1991, 1997). For the case of improving convergence detection, the function may have arisen as a by-product of talker familiarity, but could be co-opted for a social function. Future research can address this and other questions of how perceivers learn to better detect phonetic convergence.

Declaration of Competing Interest

The authors have no conflict of interest in submitting this report.

References

Allen, J., Miller, J.L., 2004. Listener sensitivity to individual talker differences in voice-onset-time. J. Acoust. Soc. Am. 115 (6), 3171–3183. https://doi.org/10.1121/1.1701898.

Babel, M., 2012. Evidence for phonetic and social selectivity in spontaneous phonetic imiation. J. Phonet. 40, 177–189. https://doi.org/10.1016/j.wocn.2011.09.001.

Babel, M., Bulatov, D., 2012. The role of fundamental frequency in phonetic accommodation. Lang. Speech 55 (2), 231–248. https://doi.org/10.1177/0023830911417695.

Babel, M., McGuire, G., Walters, S., Nicholls, A., 2014. Novelty and social preference in phonetic accomodation. Lab. Phonol. 5 (1), 123–150. https://doi.org/10.1515/lp-2014-0006.

Bernieri, F.J., Gillis, J.S., Davis, J.M., Grahe, J.E., 1996. Dyad rapport and the accuracy of its judgment across situations: a lens model analysis. J. Pers. Soc. Psychol. 71 (1), 110–129. https://doi.org/10.1037/0022-3514.71.1.110 https://doi.org/.

Brouwer, S., Mitterer, H., Huettig, F., 2010. Shadowing reduced speech and alignment. J. Acoust. Soc. Am. 128 (1), EL32–EL37. https://doi.org/10.1121/1.3448022.

Buss, D.M., Haselton, M.G., Shackelford, T.K., Bleske, A.L., Wakefield, J.C., 1998. Adaptations, exaptations, and spandrels. Am. Psychol. 53 (5), 533–548. https://doi.org/10.1037/0003-066X.53.5.533 https://doi.org/.

Church, B.A., Schacter, D.L., 1994. Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. J. Exp. Psychol. 20 (3), 521–533. https://doi.org/10.1037/0278-7393.20.3.521.

Creel, S.C., Bregman, M.R., 2011. How talker identity relates to language processing. Lang. Linguist. Compass 5 (5), 190–204. https://doi.org/10.1111/j.1749-818X.2011.00276.x.

Dahan, D., Drucker, S.J., Scarborough, R.A., 2008. Talker adaptation in speech perception: adjusting the signal or the representations? Cognition 108, 710–718. https://doi.org/10.1016/j.cognition.2008.06.003.

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

Dias, J.W., Rosenblum, L.D., 2016. Visibility of speech articulation enhances auditory phonetic convergence. Atten. Percept. Psychophys. 78, 317–333. http://doi.org/10.3758/s13414-015-0982-6.

Giles, H., Coupland, N., Coupland, J., 1991. Accommodation theory: communcation, context, and consequence. In: Giles, H., Coupland, J., Coupland, N. (Eds.), Contexts of Accommodation. Press Syndicate of the University of Cambridge, New York, pp. 1–162.

Gillis, J.S., Bernieri, F.J., Wooten, E., 1995. The effects of stimulus medium and feedback on the judgment of rapport. Organ. Behav. Hum. Decis. Process. 63 (1), 33–45. https://doi.org/10.1006/obhd.1995.1059 https://doi.org/.

Goldinger, S.D., 1998. Echoes of echoes? An episodic theory of lexical access. Psychol. Rev. 105 (2), 251–279. https://doi.org/10.1037/0033-295X.105.2.251.

Goldinger, S.D., Azuma, T., 2004. Episodic memory reflected in printed word naming. Psychon. Bull. Rev. 11 (4), 716–722. https://doi.org/10.3758/BF03196625.

Gould, S.J., 1991. Exaptation: A crucial tool for an evolutionary psychology. J. Soc. Issues 47 (3), 43–65. http://doi.org/10.1111/j.1540-4560.1991.tb01822.x.

Gould, S.J., 1997. The exaptive excellence of spandrels as a term and prototype. Proc. Natl. Acad. Sci. 94 (20), 10750–10755. http://doi.org/10.1073/pnas.94.20.10750.

Grahe, J.E., Bernieri, F.J., 1999. The importance of nonverbal cues in judging rapport. J. Nonverb. Behav. 23 (4), 253–269. http://doi.org/10.1023/a:1021698725361.

Heyes, C., 2011. Automatic imitation. Psychol. Bull. 137 (3), 463–483. http://doi.org/10.1037/a0022288.

Honorof, D.N., Weihing, J., Fowler, C.A., 2011. Articulatory events are imitated under rapid shadowing. J. Phonet. 39, 18–38. http://doi.org/10.1016/j.wocn.2010.10.007.

Kavanagh, L.C., Suhler, C.L., Churchland, P.S., Winkielman, P., 2011. When it's an error to mirror: the surprising reputational costs of mimicry. Psychol. Sci. 22 (10), 1274–1276. http://doi.org/10.1177/0956797611418678.

Lakin, J.L., Jefferis, V.E., Cheng, C.M., Chartrand, T.L., 2003. The chameleon effect as social glue: evidence for the evolutionary significance of nonconscious mimicry. J. Nonverb. Behav. 27 (3), 145–162. http://doi.org/10.1023/a:1025389814290.

Lander, K., Davies, R., 2008. Does face familiarity influence speechreadability? Q. J. Exp. Psychol. 61 (7), 961–967. http://doi.org/10.1080/17470210801908476.

Miller, R.M., Sanchez, K., Rosenblum, L.D., 2010. Alignment to visual speech information. Atten. Percept. Psychophys. 72 (6), 1614–1625. http://doi.org/10.3758/APP.72.6.1614.

Miller, R.M., Sanchez, K., Rosenblum, L.D., 2013. Is speech alignment to talkers or tasks? Atten. Percept. Psychophys. http://doi.org/10.3758/s13414-013-0517-y.

Namy, L.L., Nygaard, L.C., Sauerteig, D., 2002. Gender differences in vocal accommodation: the role of perception. J. Lang. Soc. Psychol. 21 (4), 422–432. http://doi.org/10.1177/026192702237958.

Nielsen, K., 2011. Specificity and abstractness of vot imitation. J. Phonet. 39, 132–142. http://doi.org/10.1016/j.wocn.2010.12.007.

Nielsen, K., 2014. Phonetic imitation by young children and its developmental changes. J. Speech Lang. Hear. Res. 57, 2065–2075. http://doi.org/10.1044/2014_JSLHR-S-13-0093.

Nye, P.W., Fowler, C.A., 2003. Shadowing latency and imitation: the effect of familiarity with the phonetic patterning of English. J. Phonet. 31, 63–79. http://doi.org/10.1016/S0095-4470(02)00072-4.

Nygaard, L.C., 2005. Perceptual integration of linguistic and non-linguistic properties of speech. In: Pisoni, D., Remez, R. (Eds.), Handbook of Speech Perception. Blackwell, Malden, MA, pp. 390–419.

Nygaard, L.C., Pisoni, D., 1998. Talker-specific learning in speech perception. Percept. Psychophys. 60 (3), 355–376. http://doi.org/10.3758/BF03206860.

Nygaard, L.C., Sommers, M.S., Pisoni, D.B., 1994. Speech perception as a talker-contingent process. Psychol. Sci. 5 (1), 42–46. http://doi.org/10.1111/j.1467-9280.1994.tb00612.x.

Over, H., Carpenter, M., 2015. Children infer affiliative and status relations from watching others imitate. Dev. Sci. 18 (6), 917–925. http://doi.org/10.1111/desc.12275.

Palmeri, T.J., Goldinger, S.D., Pisoni, D.B., 1993. Episodic encoding of voice attributes and recognition memory of spoken words. J. Exp. Psychol. 19 (2), 309–328. http://doi.org/10.1037/0278-7393.19.2.309.

Pardo, J.S., 2006. On phonetic convergence during conversational interaction. J. Acoust. Soc. Am. 119 (4), 2382–2393. http://doi.org/10.1121/1.2178720.

Pardo, J.S., Gibbons, R., Suppes, A., Krauss, R.M., 2012. Phonetic convergence in college roommates. J. Phonet. 40, 190–197. http://doi.org/10.1016/j.wocn.2011.10.001.

Pardo, J.S., Jordan, K., Mallari, R., Scanlon, C., Lewandowski, E., 2013. Phonetic convergence in shadowing speech: the relation between acoustic and perceptual measures. J. Memory Lang. 69, 183–195. https://doi.org/10.1016/j.jml.2013.06.002 https://doi.org/.

Pardo, J.S., Urmanche, A., Wilman, S., Wiener, J., 2017. Phonetic convergence across multiple measures and model talkers. Atten. Percept. Psychophys. 79, 637–659. http://doi.org/10.3758/s13414-016-1226-0.

Remez, R.E., Dubowski, K.R., Broder, R.S., Davids, M.L., Grossman, Y.S., Moskalenko, M., Hasbun, S.M., 2011. Auditory-phonetic projection and lexical structure in the recognition of sine-wave words. J. Exp. Psychol. Hum. Percept. Perform. 37 (3), 968–977. http://doi.org/10.1037/a0020734.

Remez, R.E., Fellowes, J.M., Nagel, D.S., 2007. On the perception of similarity among talkers. J. Acoust. Soc. Am. 122 (6), 3688–3696. http://doi.org/10.1121/1.2799903.

Remez, R.E., Fellowes, J.M., Rubin, P.E., 1997. Talker identification based on phonetic information. J. Exp. Psychol. Hum. Percept. Perform. 23 (3), 651–666. https://doi.org/10.1037/0096-1523.23.3.651.

Remez, R.E., Rubin, P.E., Pisoni, D.B., Carrell, T.D., 1981. Speech perception without traditional speech cues. Science 212, 947–950. http://doi.org/10.1126/science.7233191.

Remez, R.E., Thomas, E.F., Crank, A.T., Kostro, K.B., Cheimets, C.B., Pardo, J.S., 2018. Short-term perceptual tuning to talker characteristics. Lang. Cognit. Neurosci. 33 (9), 1083–1091. http://doi.org/10.1080/23273798.2018.1442580.

Rosenblum, L.D., Miller, R.M., Sanchez, K., 2007a. Lip-read me now, hear me better later: cross-modal transfer of talker-familiarity effects. Psychol. Sci. 18 (5), 392–396. https://doi.org/10.1121/1.4788294.

Rosenblum, L.D., Niehus, R.P., Smith, N.M., 2007b. Look who's talking: recognizing friends from visible articulation. Perception 36 (1), 157–159. http://doi.org/10.1068/p5613.

Rosenblum, L.D., Saldaña, H.M., 1996. An audiovisual test of kinematic primitives for visual speech perception. J. Exp. Psychol. Hum. Percept. Perform. 22 (2), 318–331. https://doi.org/10.1037/0096-1523.22.2.318.

Rosenblum, L.D., Yakel, D.A., Baseer, N., Panchal, A., Nodarse, B.C., Niehus, R.P., 2002. Visual speech information for face recognition. Percept. Psychophys. 64 (2), 220–229. https://doi.org/10.3758/BF03195788.

Sanchez, K., Dias, J.W., Rosenblum, L.D., 2013. Experience with a talker can transfer across modalities to facilitate lipreading. Attent. Percept. Psychophys. 75 (7), 1359–1365. http://doi.org/10.3758/s13414-013-0534-x.

Sanchez, K., Miller, R.M., Rosenblum, L.D., 2010. Visual influences on alignment to voice onset time. J. Speech Lang. Hear. Res. 53, 262–272. https://doi.org/10.1044/1092-4388(2009/08-0247.

Schweinberger, S.R., Kawahara, H., Simpson, A.P., Skuk, V.G., Zäske, R., 2014. Speaker perception. WIREs Cognit. Sci. 5 (1), 15–25. https://doi.org/10.1002/wcs.1261.

Schweinberger, S.R., Soukup, G.R., 1998. Asymmetric relationships among perceptions of facial identity, emotion, and facial speech. J. Exp. Psychol. Hum. Percept. Perform. 24 (6), 1748–1765. https://doi.org/10.1037/0096-1523.24.6.1748.

Sheffert, S.M., Olson, E., 2004. Audiovisual speech facilitates voice learning. Percept. Psychophys. 66 (2), 352–362. https://doi.org/10.3758/BF03194884.

Sheffert, S.M., Pisoni, D.B., Fellowes, J.M., Remez, R.E., 2002. Learning to recognize talkers from natural, sinewave, and reversed speech samples. J. Exp. Psychol. Hum. Percept. Perform. 28 (6), 1447–1469. https://doi.org/10.1037/0096-1523.28.6.1447.

Shepard, C.A., Giles, H., Le Poire, B.A., 2001. Communication accommodation theory. In: Robinson, W.P., Giles, H. (Eds.), The New Handbook of Language and Social Psychology. Wiley, New York, NY, pp. 33–56.

Shockley, K., Sabadini, L., Fowler, C.A., 2004. Imitation in shadowing words. Percept. Psychophys. 66 (3), 422–429. https://doi.org/10.3758/BF03194890.

Smith, R., 2015. Perception of speaker-specific phonetic detail. In: Fuchs, S., Pape, D., Petrone, C., Perrier, P. (Eds.), Individual Differences in Speech Production and Perception. (Vol. 3, pp. 11-38)Petere Lang, Oxford.

Smith, R., Hawkins, S., 2012. Production and perception of speaker-specific phonetic detail at word boundaries. J. Phonet. 40, 213–233. https://doi.org/10.1016/j.wocn.2011.11.003.

Theodore, R.M., Miller, J.L., 2010. Characteristics of listener sensitivity to talker-specific phonetic detail. J. Acoust. Soc. Am. 128 (4), 2090–2099. http://doi.org/10.1121/1.3467771.

van Heugten, M., Volkova, A., Trehub, S.E., Schellenberg, E.G., 2014. Children's recognition of spectrally degraded cartoon voices. Ear Hear. 35 (1). http://doi.org/10.1097/AUD.0b013e3182a468d0.

von Kriegstein, K., Kleinschmidt, A., Giraud, A.-L., 2006. Voice recognition and cross-modal responses to familiar speakers' voices in prosopagnosia. Cereb. Cortex 16 (9), 1314–1322. http://doi.org/10.1093/cercor/bhj073.

Yakel, D.A., Rosenblum, L.D., Fortier, M.A., 2000. Effects of talker variability on speechreading. Percept. Psychophys. 62 (7), 1405–1412. https://doi.org/10.3758/BF03212142.

Zarate, M.J., Tian, X., Woods, K.J.P., Poeppel, D., 2015. Multiple levels of linguistic and paralinguistic features contribute to voice recognition. Sci. Rep. 5 (1), 11475. http://doi.org/10.1038/srep11475.

Zellou, G., Dahan, D., Embick, D., 2017. Imitation of coarticulatory vowel nasality across words and time. Lang. Cognit. Neurosci. 32 (6), 776–791. http://doi.org/10.1080/23273798.2016.1275710.

Zellou, G., Scarborough, R., Nielsen, K., 2016. Phonetic imitation of coarticulatory vowel nasalization. J. Acoust. Soc. Am. 140 (5), 3560–3575. https://doi.org/10.1121/1.4966232 https://doi.org/.