Singing with yourself: Evidence for an inverse modeling account of poor-pitch singing

Peter Q. Pfordresher [a,*], James T. Mantell [b]

[a] Department of Psychology, University at Buffalo, State University of New York, United States
[b] Department of Psychology, St. Mary's College of Maryland, United States

ARTICLE INFO

ABSTRACT

Singing is a ubiquitous and culturally significant activity that humans engage in from an early age. Nevertheless, some individuals – termed poor-pitch singers – are unable to match target pitches within a musical semitone while singing. In the experiments reported here, we tested whether poor-pitch singing deficits would be reduced when individuals imitate recordings of themselves as opposed to recordings of other individuals. This prediction was based on the hypothesis that poor-pitch singers have not developed an abstract "inverse model" of the auditory–vocal system and instead must rely on sensorimotor associations that they have experienced directly, which is true for sequences an individual has already produced. In three experiments, participants, both accurate and poor-pitch singers, were better able to imitate sung recordings of themselves than sung recordings of other singers. However, this self-advantage was enhanced for poor-pitch singers. These effects were not a byproduct of self-recognition (Experiment 1), vocal timbre (Experiment 2), or the absolute pitch of target recordings (i.e., the advantage remains when recordings are transposed, Experiment 3). Results support the conceptualization of poor-pitch singing as an imitative deficit resulting from a deficient inverse model of the auditory–vocal system with respect to pitch.

© 2014 Elsevier Inc. All rights reserved.

---

* Corresponding author. Address: Department of Psychology, University at Buffalo, 355 Park Hall, Buffalo, NY 14260, United States. Fax: +1 716 645 3801.
  E-mail address: pqp@buffalo.edu (P.Q. Pfordresher).

1. Introduction

Individuals vary in their ability to imitate pitch accurately by singing. Although most individuals are on average accurate when imitating vocal pitch through singing (Dalla Bella, Giguere, & Peretz, 2007; Pfordresher & Brown, 2007; Pfordresher & Mantell, 2009), a minority of individuals (approximately 10–20%, Dalla Bella et al., 2007; Pfordresher & Brown, 2007) sing out-of-tune by more than ±100 cents (1 musical semitone on either side of the target pitch) on average and have been termed 'poor-pitch singers' (Welch, 1979a). Given the ubiquity of singing across cultures and the importance of vocal imitation for language acquisition (Kuhl, 2000; Kuhl & Meltzoff, 1996), such deficits are surprising and intriguing. Moreover, the cognitive bases of poor-pitch singing are still not well understood.

We propose that the problem of poor-pitch singing lies in the fact that vocal imitation, unlike the imitation of manual gestures, cannot rely on the direct observation (visual, kinesthetic or otherwise) of the motor gestures generated by the target. Thus, though singing may be considered a form of vocal imitation, the inability to perceive motor gestures directly causes the task of singing to better resemble what has been termed "emulation" (i.e., reproducing outcomes rather than imitating gestures, cf. Tomasello, 1990). It has been argued that vocal imitation may proceed entirely by error correction in the output (Heyes, 2005). However, recent evidence suggests that most individuals have only a limited capacity for using auditory feedback to improve pitch matching (Hutchins & Peretz, 2012). Thus, we propose an account of poor-pitch singing based on the problem of relating auditory events to kinesthetic motor states.

In the current paper we test the hypothesis that poor-pitch singing arises because these individuals have not formed an abstract *inverse model* of the auditory–vocal system, which allows singers to plan motor gestures based on target perceptual outcome. This framework, described later, leads to the prediction that poor-pitch singing should be reduced when the target exhibits a pitch pattern similar to one's own vocal gestures (for which better-formed associations should exist) as opposed to the vocalizations of others: a *self-imitation benefit*.

1.1. Poor-pitch singing: Background

Recent research has adopted acoustic measurements of production in an attempt to better document what poor pitch singers do while they are singing. As mentioned above, poor-pitch singers are primarily characterized by an overall tendency to distort absolute pitch consistently while singing ("mistuning"). Poor-pitch singers also compress the size of pitch intervals (normal singers exhibit compression to a lesser degree, Dalla Bella, Giguere, & Peretz, 2009; Pfordresher & Brown, 2007). Finally, poor-pitch singers are typically more variable in their production than are accurate singers, suggesting that inaccuracy and imprecision may jointly characterize poor-pitch singers (Pfordresher, Brown, Meier, Belyk, & Liotti, 2010). In contrast to these deficits, poor-pitch singers are often accurate in reproducing melodic contour (the pattern of upwards and downwards pitch changes in a melody, Pfordresher & Brown, 2007).

A variety of core deficits may lead to poor-pitch singing (for reviews see, Berkowska & Dalla Bella, 2009; Pfordresher & Brown, 2007; Pfordresher & Mantell, 2009; Tsang, Friendly, & Trainor, 2012). To date, evidence suggests that poor-pitch singing cannot be accounted for simply by some underlying perceptual deficit (as suggested by the colloquial term "tone deafness", Bradshaw & McHenry, 2005; Dalla Bella et al., 2007; Lévêque, Giovanni, & Schön, 2012; Pfordresher & Brown, 2007), or by problems with motor control (Pfordresher & Brown, 2007; Pfordresher & Mantell, 2009). Nor does an account based on an underlying memory deficit seem likely, particularly given that poor-pitch singing can emerge during matching of single pitches (Hutchins & Peretz, 2012) and poor-pitch singers are capable of performing complex, multi-note melodies even more accurately than simple melodies (Pfordresher & Brown, 2007). Thus, Pfordresher and Brown (2007) suggested that poor-pitch singing is a deficit of sensorimotor-translation in the vocal system—that is, a deficit specific to imitation. Of course, building a theory on negative evidence is not optimal. Thus, the current study represents an

attempt to provide positive evidence for a theoretical framework that has emerged from the implications of past research.

1.2. Inverse models and the self-imitation advantage

Critical to the present research is the *internal models* framework (Kawato, 1999; Wolpert, 1997), and in particular the *inverse model* component within this framework. Fig. 1 presents a hypothetical internal model for the control of vocal F0, stemming from recent approaches to speech phenomena such as stuttering (Max, Guenther, Gracco, Ghosh, & Wallace, 2004) and sensorimotor integration in speech (Hickok, Houde, & Rong, 2011). The "internal model" processes are those that connect the perceptually based target pitch on the left with the motor control of vocal folds in the larynx to the right. The two ovals in Fig. 1 represent different manifestations of such an internal model. Both constitute a cognitive representation of the perception/action interface, based on past experience of causal relationships between actions and their outcomes. Internal models are formed when we generalize from these past experiences to construct abstract algorithms that allow us to coordinate perception with action in novel contexts. With respect to vocal pitch production, an internal model would encode contingencies between a particular tension of the vocal folds and the pitch percept that would result.

An internal model can negotiate the perception/action interface in two ways. First, the internal model can generate an *inverse model* of a motor gesture based on the anticipated (goal-based) perceptual outcome of that action. With respect to pitch production, the inverse model allows the communicator to adjust vocal fold tension based on the desired fundamental frequency (F0) of the output. This process, for example, allows a singer to produce the correct pitch in advance of any auditory feedback. In addition, the inverse model generates an efference copy of the action, which is the expected outcome, and sends that copy as input to the *forward model*, which is the second way in which an internal model guides action. The forward model generates an expected outcome for the action that is being planned, against which feedback from the actual outcome can be compared. The forward model allows singers to judge whether their motor output differed from the output they intended to produce. In such cases, the controller (i.e., singer) may adjust the target F0 while continuing to generate pitch. If the efference copy is poorly formed (as may be the case in poor-pitch singing), this process will not work properly. Somewhat counter-intuitively, the inverse model is thought to be part of feed-forward control, whereas the forward model is thought to be part of feedback control.

In the present research, we focus on the inverse model component of this framework, because this component is most critical for the vocal imitation tasks used here. Specifically, in the task of vocal imitation an individual must reproduce auditory pitch–time contents using his or her own voice. In this case, the individual can perceive the pitch–time contents of the to-be-imitated target-sequence, but cannot directly observe the movements to produce the target-sequence. Thus, the kinesthetic motor states of the target-singer are hidden from the observer. In order to imitate the target sequence accurately, the observer must reconstruct the hidden movement states from the just perceived pitch–time
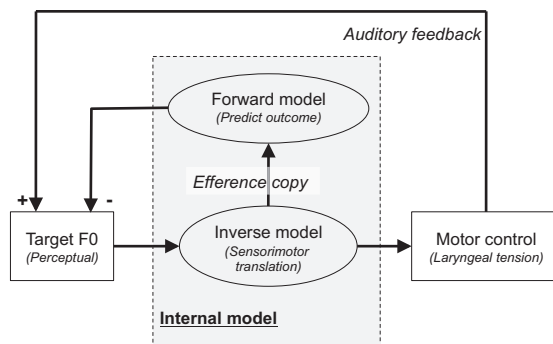
Fig. 1. Schematic diagram of the internal models framework, applied to the task of reproducing the pitch content of an auditory event (F0) through phonation.

information available in memory. In this way, the observer's process of sensorimotor translation can be described as inverse modeling. That is, in order to prepare the movement states necessary for imitative production, the observer must attempt to 'recreate' them from the perceptual contents. Henceforth, we refer explicitly to the inverse model component when discussing the basis of poor-pitch singing, though it is possible that a general internal model deficit (including inverse and forward components) exists for these individuals.

We propose that individuals vary in their ability to perform this inverse modeling process and that poor-pitch singers are especially deficient. Absent a properly functioning inverse model, how does a poor-pitch singer sing? A central assumption of the inverse model is that the controller (singer) has developed an abstract framework that can be used to associate perception and action in a broad range of contexts (similar to the motor schema construct, cf. Schmidt, 1975). For singing, this means that one can inversely model the motor gestures necessary to produce a pitch, or a sequence of pitches, even if one has not produced those pitches before. In the absence of such an abstract framework, the singer must rely on associations that have been acquired via singing experience. Such a singer would face great difficulty when being presented with pitches – or pitch sequences – that he or she has not produced in the past, whereas reproducing a pattern that the individual has sung before should not prove as difficult.

Based on this reasoning, we propose that the hypothesized inverse model deficit among poor-pitch singers should lead to a *self-imitation advantage* that is enhanced for poor-pitch singers relative to accurate singers. Past research has shown that singers in general can more effectively match pitch of their own voice than the pitch of another person's voice (Hutchins & Peretz, 2012) or to a synthetic target (Lévêque et al., 2012; Moore, Estis, Gordon-Hickey, & Watts, 2008; Watts & Hall, 2008). However, this advantage should increase for poor-pitch singers because, during self-imitation, the poor-pitch singer can successfully draw on learned associations, whereas the imitation of another singer's productions requires the use of an inverse model. Results suggesting such an interaction were reported by Moore and colleagues (2008), although a full statistical interaction between the accuracy of the imitator and the target being imitated was not reported in that study (the authors only identified four poor-pitch singers, leading to low power for such an effect). We predict that individuals that are poor imitators of pitch–time information will experience a greater self-imitation advantage than individuals who are accurate imitators.

In addition to evaluating the degree to which participants show an advantage in imitating self versus other-targets, we also assessed imitation accuracy for other-targets who represent accurate or poor-pitch singing. We had two reasons for doing this. First, the degree of advantage in imitating accurate versus inaccurate other-targets may be considered an advantage for imitating sequences that match prototypical musical patterns. The self-advantage thus can be assessed relative to any such prototype advantage. Second, as it is possible that poor-pitch singers in our sample may produce imitations that resemble other-targets who are also inaccurate, such possible coincidences (which did occasionally occur) allow us to evaluate whether an advantage similar to the self-advantage occurs when inaccurate singers imitate someone else who generates similarly flawed vocal performances.

1.3. The current experiments

We report three new experiments that test the self-imitation advantage among untrained singers representative of "accurate" and "poor-pitch" categories. Fig. 2 illustrates the critical features of each experiment's design. In each experiment, participants were initially recorded imitating four simple 4-note melodies (phase 1). These *self-target* recordings were then intermingled with *other-target* recordings of singers selected from a database of singers imitating the same melodies in previous experiments (Pfordresher & Brown, 2007, 2009). Half of the other targets were representative of accurate singing (average mistuning < |50| cents) whereas the other half were poor-pitch singers (average mistuning > |100| cents). Participants then imitated each of the four melodies as performed by themselves (self-targets) and each of the database singers (other-targets), and after each trial gave the experimenter a recognition rating indicating whether the recording they had just imitated was their own, or someone else's.
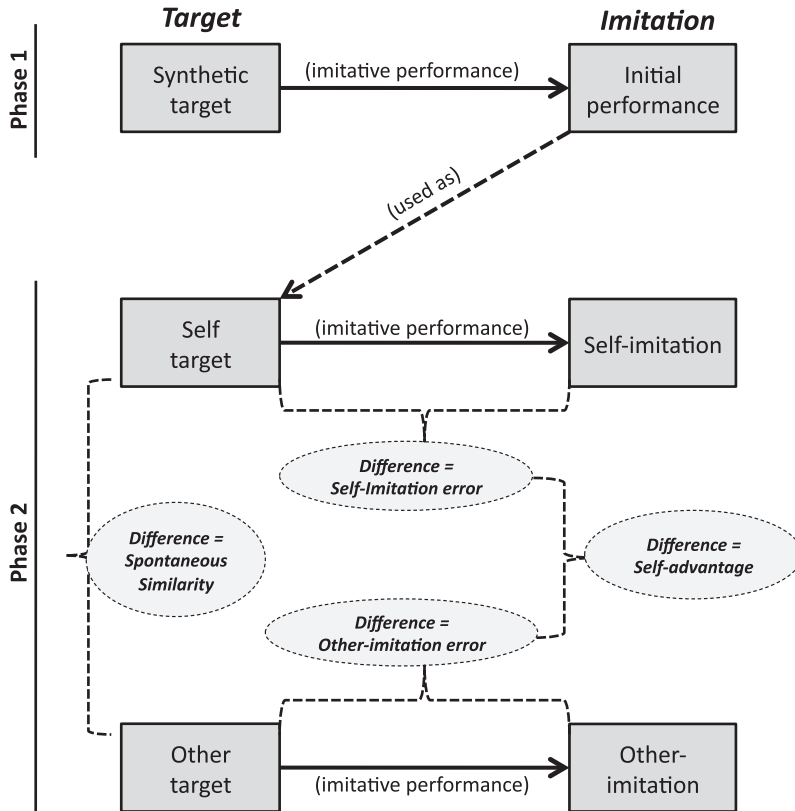
Fig. 2. Flow chart showing experimental design and performance measures. Dark gray boxes indicate recordings of targets or imitations of targets; dark arrows denote the sequence of events within a trial. Light gray ovals indicate measures based on difference scores based on connections formed by dashed lines.

Different experiments were designed to focus on the contribution of timbre, pitch height, and the pitch–time trajectory to the self-advantage. Experiment 1 included full recordings of self and other targets and thus featured a variety of potentially advantageous self-features. In Experiment 2, we eliminated the contribution of timbral self-similarity (cf. Hutchins & Peretz, 2012) by synthesizing all recordings into a neutralized vocal timbre. This procedure served the additional goal of reducing the contribution of self-recognition in the self-imitation advantage (cf. Knoblich & Prinz, 2001; Knoblich, Seigerschmidt, Flach, & Prinz, 2002; Repp & Knoblich, 2004). Finally, in Experiment 3 we focused on the degree to which the absolute pitch of one's voice leads to the self-advantage, by presenting re-synthesized recordings in three different transpositions: the original pitch, up three semitones, and down three semitones.

2. Experiment 1

In Experiment 1, participants first were recorded imitating synthesized recordings of simple 4-note melodies. These initial performances were then used as self-targets in a second phase of the experiment in which participants imitated both self- and other-targets. The other-targets were recordings of participants imitating the same synthesized melodies and were drawn from samples used in previous research. Importantly, other-targets were selected to reflect a range of accuracy from accurate to poor-pitch singing. Participants imitated other-targets of their own gender. Our primary hypothesis was that all participants would exhibit an advantage when imitating self-targets but that this advantage would be larger for poor-pitch than for accurate singers.

2.1. Methods

2.1.1. Participants

Forty-four participants were recruited from the University at Buffalo subject pool for course credit in Introductory Psychology.[1] Exclusion criteria for participants were the presence of a diagnosed hearing disorder or disorder of vocal motor control. The mean age of participants was 19.16 years (range: 18–23). Thirteen participants were female and 31 were male. Thirty-four participants reported at least one year of musical experience performing one or more musical instruments and/or via singing (M = 7.63 summed years of experience including voice and instrument, range = 1–28). However, only two participants reported formal vocal training outside of a school choir (19 participants reported singing in a choir at some point). One reported 2 years of training; the other reported 18 years of training, though this was likely a misunderstanding of the questionnaire, as the participant was only 18 years old. Ultimately, it is unlikely that performance differences among participants are simply due to formal training in singing although carry-over effects from other forms of musical training may contribute to individual differences (a possibility we address).

2.1.2. Stimuli

In phase 1 of the experiment, participants imitated idealized recordings of four 4-note melodies, generated by the voice synthesis packages Leon (for male participants) or Lola (for female; Vocaloid, Zero-G Limited, Okehampton, UK). All stimuli were based on selecting 4 different pitches from the first 5 scale degrees of the C major scale and varied with respect to starting pitch and melodic contour (the melodies were [C D E G], [G D C E], [C G F E], and [G E F C]). Pitches for male stimuli were one octave lower (starting at C3 = 131 Hz) than female stimuli (starting at C4 = 262 Hz). Each pitch was sung on the syllable "dah." All melodies began on either C or G, and maintained a rate of 1 second per syllable.

In phase 2 of the experiment, participants imitated recordings of themselves (from phase 1) and recordings of other participants (referred to as *other-targets*) who imitated the same set of melodies. Other-targets were drawn from a database of previously recorded performances that have been reported elsewhere (Pfordresher & Brown, 2007, 2009). We selected recordings from eight other-target individuals for use with both male and female participants (16 other-targets in all). Within each set of 8 (male or female) other-targets, four were selected to represent poor-pitch singing in that the mean signed difference between their sung pitch and the target pitch fell outside a window of ±100 cents. Their performances thus deviated from target pitches by more than 1 semitone in either direction on average, a criterion that has been used in the past to distinguish between accurate and poor-pitch singers (Pfordresher & Brown, 2007). These targets were therefore considered as *inaccurate-other targets*. By contrast, four more *accurate-other-target* individuals had average errors within 50 cents of the target pitch. Thus, each participant in the experiment imitated targets recorded from nine singers: the self (self-target recordings), four accurate-others (accurate-other target recordings), and four inaccurate others (inaccurate-other target recordings). Furthermore, there were four recordings from each singer, leading to 36 imitation trials in all.[2]

Table 1 summarizes important descriptive statistics for target participants, including their mean fundamental frequency (in Hz), the range of pitches produced (in cents), and two measures of pitch

---

[1] Experiment 1 was originally conducted as two experiments. The experiments differed with respect to some of the targets that were imitated and the recruitment procedure used. However, neither of these variables significantly influenced performance and so in order to simplify the exposition and to boost power we group both experiments together here. With respect to the recruitment procedure, the first original experiment (n = 32) included participants sampled at random, whereas in the second original experiment (n = 12) we adopted a prescreening procedure (described in Experiment 2 of Pfordresher & Brown, 2007, see also the description to Experiment 2 in the present article) to ensure a broad range of imitation abilities among participants even in a smaller sample. With respect to other-targets, in the first original experiment we included three levels of other-target: accurate, inaccurate (poor-pitch), and fair. After analyzing the data, we found that differences between "fair" and "accurate" other-targets were negligible, and so in the second original experiment we altered the set of other-targets to eliminate this intermediate level. All trials that involved the imitation of "fair" targets from the original first experiment are discarded here.

[2] The count of participants reported here held for the first 32 participants who participated in the original Experiment 1. After removing the "fair singer" participant category for the next 12 participants (see Footnote 1), we re-sampled the accurate and inaccurate other-targets and selected 6 singers from each category. This new set of other-targets was used in Experiments 2 and 3. Note that the descriptive statistics in Table 1 are averaged across all the other-targets that were used.

performance (as summarized in *Data Analysis*, below, both in cents). As can be seen, accurate and inaccurate targets differ with respect to their imitation ability but not with respect to the absolute pitch or range of pitches in the recordings.

2.1.3. Apparatus

Participants were recorded while sitting inside a Whisper Room SE 2000 recording booth. All instructions and stimuli were delivered to participants via a pair of Sennheiser HD 280 Pro headphones, which were used to present target sequences, instructions from the experimenter, and auditory feedback (which would have been attenuated by the headphones otherwise). The headphone intensity was kept at a constant comfortable level for all participants. All participant recordings were made using a Shure PG58 microphone. All sound levels were controlled using a Lexicon Omega I/O box. The experiment was conducted via Matlab (the Mathworks, Natick, MA) on a PC running Windows XP. Digital audio recordings of vocal imitations were acquired by Matlab at a sampling rate of 22,050 Hz.

2.1.4. Procedure

The experiment began with a series of vocal warm-up exercises. Participants were seated on a stool in the recording booth, fitted with headphones, and were given instructions on appropriate posture and breath support during singing. A microphone was positioned approximately two inches from their mouth. All stimuli were presented via headphones. First, participants completed warm-up exercises consisting of extemporaneous speech (e.g., "list what you had for dinner last night"), reading a page of text, and singing a familiar tune ("Happy Birthday"). Participants were then asked to produce vocal sweeps: a continuous change in pitch from the lowest note an individual can comfortably sing up to the highest note he or she can comfortably sing and then back down. Finally, participants produced a single, easily produced note representative of their typical vocal pitch (their "comfort pitch").

Following the warm-up exercises participants began phase 1 of the experiment, in which initial performance recordings were collected. Participants listened to and then imitated 4-note melodies produced by a stylized synthetic voice. Participants were instructed not to vocalize during presentation of the target. After the target was played, a noise burst sounded; it functioned as a cue to the participant to start vocally imitating the melody. After the recordings had been obtained, participants were asked to step out of the booth and complete a series of questionnaires concerning musical background and hearing sensitivity. While the participants were filling out the questionnaires, the experimenter edited the recordings that the participant had just produced by removing any silences or incidental sounds that might surround the recorded sequences.

Participants then reentered the recording booth to begin phase 2 of the experiment. Trials were identical to phase 1, except that participants imitated recordings of themselves (*self-targets*) from phase 1 and others (*other-targets*), as described above. All trials were pseudo-randomly intermingled with the constraint that no two successive trials could originate from the same target singer or be the same melody. After each trial, participants provided a numerical rating that indicated whether they thought the recording was their own voice or was of someone else. A rating of 1 indicated certainty that the recording was one's own, while a rating of 7 indicated certainty that the recording was someone else's. Participants spoke each rating aloud and the experimenter entered it into a data file. The

Table 1
Characteristics of targets used in all experiments.

	Mean F0 (Hz)	Pitch range	Signed deviation	Absolute deviation
Accurate	251 (20)	681 (15)	17 (12)	75 (10)
Poor-pitch	216 (17)	643 (40)	−201 (102)	361 (74)
All	234 (14)	663 (21)	−94 (53)	314 (44)
	p = .097	*p* = 0.183	***p* = .021**	***p* = .001**

Note: All figures averaged across male and female targets, who differed significantly with respect to mean Hz of pitch. Measures in cents (pitch range, signed pitch deviation, absolute pitch deviation) are computed relative to 131 Hz for male targets and 262 Hz for female targets. Statistical test yielding *p*-values are two-sample *t*-tests. Bold *p*-values are significant at the .05 level.

addition of the rating task in phase 2 nominally increased the duration of each recording trial. It is conceivable that this extra task may have had the effect of reducing subjects' overall accuracy (e.g., by splitting their attention between identifying the target and preparing an accurate imitation). However, acquiring rating data was necessary so that we could determine the efficacy of our stimulus manipulation techniques in Experiments 2 and 3.

2.1.5. Data analysis

We analyzed the accuracy and precision with which people imitated target-pitch trajectories. The analyses we focus on here are based on the entire vector of sampled F0 values from a trial, after adjusting the timing of the imitation to match the total duration of the target (also used in Mantell & Pfordresher, 2013). This approach to pitch analysis differs from the standard way in which singing is analyzed, which is to extract a single point estimate to represent the F0 of an entire sung note and then to analyze accuracy and precision based on these point estimates. Although we do report such analyses here, we consider the use of the continuous measure to be advantageous because vocalists rarely produce consistently flat tones, and thus in an imitation task like this one it is important to measure participants' ability to imitate fluctuations of F0 within 'steady' sung tones and also transitions across tones.

The following specific procedure was used to compute error in pitch imitation (also described in Mantell & Pfordresher, 2013). We used the autocorrelation algorithm in Praat (Boersma & Weenink, 2013) to extract the F0 values from the vocal recordings. The F0 sampling frequency varied depending on the F0 extraction parameters that we chose to work best for each vocal recording. Furthermore, longer duration vocalizations produced more F0 samples than shorter duration productions. Thus, an imitation sequence and a target sequence did not have the same number of F0 samples. In order to make a pair (imitation and target) of F0 vectors directly comparable, the F0 values in both vectors (imitation and target) were transformed to cents (relative to the lowest F0 in the idealized target, which was 131 Hz for males and 262 Hz for females, corresponding to C3 and C4, respectively). Next, vectors were equated for duration by aligning the onsets of each vector and then re-sampling timepoints from the imitated F0 vector so that the number of samples was equal to the samples present in the target F0 vector. In practice, necessary adjustments to timing were modest given that the mean ratio of imitation duration to target duration was approximately 1:1 (M ratio = 1.03, SE = .01). After normalizing vectors with respect to time, difference scores were computed for each normalized sample (imitation – target). From the resulting vector of F0 difference scores, outliers were removed (defined as any data point in the imitation vector that was more than one half octave away from the target, after removing any tendency to sing sharp or flat overall). After removing outliers the *mean absolute error* for the trial was calculated as the mean of the absolute values from the vector of difference scores. This measure, which functions as our primary measure of imitative success, is influenced both by the accuracy and precision of imitation. Whereas accuracy refers to the proximity of the mean sung pitch, precision refers to variability around the participant's mean (whether accurate or not). The mean across a series of signed error scores reflects accuracy, but the mean across absolute values from the same set of scores can be influenced by accuracy and precision (for further discussion see Pfordresher et al., 2010).[3]

These scores were computed for each trial and were submitted to two kinds of analyses. First, we analyzed mean performance across participants as a function of the type of target as well as the overall performance of the participant (accurate or inaccurate). This analysis incorporated a two-way mixed model ANOVA with the within-subjects factor target type (self, accurate-other, inaccurate-other) and the between subjects factor participant type (accurate, inaccurate). This analysis speaks to the presence of a self-advantage and also indicates whether performance differed across accurate versus

---

[3] We also measured the accuracy and precision with which relative pitch was imitated by using the *pitch correlation* between the F0 values in the target and the corresponding F0 values during imitation (cf. Mantell & Pfordresher, 2013). The correlation coefficient reflects the imitation of relative rather than absolute pitch because correlations disregard the magnitude of the *Y*-intercept, which in the context of pitch imitation reflects mistuning of absolute pitch. However, these results were very similar to those reported for mean absolute pitch error, so in the interest of brevity we report only results from mean absolute error analyses.

inaccurate-other-targets. All ANOVA results reported as significant remain so after correcting for violations of sphericity through the Greenhouse–Geisser correction.

This omnibus ANOVA was followed by two further planned simple contrasts designed to focus on theoretically important comparisons within the design. The first of these focused on the self-advantage by computing the following self-advantage score for each participant (see also Fig. 2):

$$Self\text{-}advantage = .5[Error(accurate\text{-}other) + Error(inaccurate\text{-}other)] - Error(self)$$

where *Error* is the mean across all trials for that target type. Thus the self-advantage is high whenever error for self-targets is low relative to the average error for both other-target types. The magnitude of self-advantage scores was compared across groups using a one-way *t*-test (given that the inverse model deficit hypothesis generates a 1-tailed prediction). The second planned contrast was based on difference scores across accurate and inaccurate-other targets for each participant, discarding all trials involving self-imitation. These scores measure a *prototype advantage* (i.e., advantage for imitating accurate as opposed to inaccurate other-targets) and were likewise contrasted across groups.

In addition to analyses that treat poor-pitch and accurate singers as dichotomous groups, we also report results that represent imitation ability on a continuum, using a regression framework rather than analysis of variance. Our justification for this approach stems from both conceptual and methodological rationales. Conceptually, we are sensitive to recent arguments that individual differences in complex abilities such as music perception and performance may not be best understood as dichotomous variables (cf. Henry & McAuley, 2010). Methodologically, the use of a regression framework helps avoid the problem of floor effects (from low error) for accurate participants.

Regressions used mean absolute error across all trials as the predictor (*X*) variable. First, we regressed mean performance across all self-target trials (*Y*) on the mean across all trials (*X*), and compared the slope of that regression to a companion regression of mean performance across all other-target trials (*Y*) on the mean of all trials (*X*). The inverse model deficit hypothesis predicts a steeper positive slope for the latter regression than the former, under the assumption that poor-pitch singing deteriorates imitations of other-targets more so than self-targets. Second, we computed difference scores designed to measure the magnitude of the self-advantage (described above) and regressed those difference scores on the *X* variable.

2.2. Results

After initial inspection of the data, one participant's data were removed based on extreme pitch error scores relative to other participants in the sample (*M* for this participant = 723 cents, range across other participants = 64–243 cents). Such extreme cases, of course, are of interest for understanding poor-pitch singing. Thus, we discuss this participant's data separately from those of other participants.

The remaining participants were defined as accurate or poor-pitch based on the mean of signed difference scores (not absolute) across all trials. We used signed error scores to categorize participants because the specific relationship between signed error scores and accuracy (as opposed to precision), discussed earlier, make the application of cutoffs for categorizing participants more transparent. Participants whose signed error scores were outside a window of ±50 cents (1 semitone in all) were categorized as poor-pitch (*n* = 12) and the remaining participants were considered accurate (*n* = 32). In the past our lab has adopted a more conservative criterion for categorizing singers as poor-pitch (±100 cents, e.g., Pfordresher & Brown, 2007), but we utilize the more liberal criterion in the current work because of recent research that has questioned the validity of the previous criterion (Hutchins & Peretz, 2012; cf. Pfordresher et al., 2010). Participants categorized as accurate or poor-pitch did not differ significantly with respect to reported years of music lessons (*M* for accurate = 3.98 years, *M* for poor-pitch = 4.15 years, *p* = .48), or reported years of singing lessons (*M* for accurate <1 year, for inaccurate = 1.39 years, *p* = .19).

2.2.1. Means across participants

Fig. 3 shows mean absolute pitch error values across all participants and target recordings for the three different target types and two participant groups. Here, lower scores reflect better imitation. There was a significant main effect of target type, $F(2, 82) = 26.29, p < .01, \eta_p^2 = .39$, a main effect of
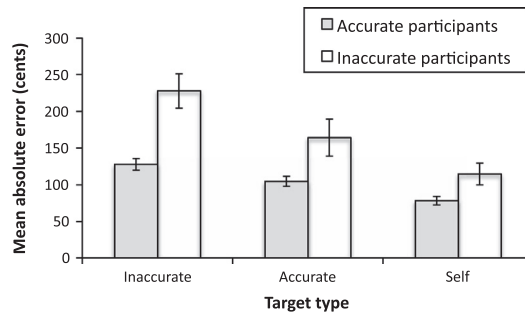
Fig. 3. Effect of target type on mean absolute pitch error in cents for imitative performances in phase 2 of Experiment 1. Error bars denote one between-subjects standard error of the mean.

participant type, $F(1, 41) = 26.40, p < .01, \eta_p^2 = .39$, and a target type × participant type interaction, $F(2, 82) = 4.81, p = .01, \eta_p^2 = .11$. As can be seen in Fig. 3, error decreases in both groups from left to right, being lower for self-imitation than for other imitation (with inaccurate-other targets generating higher errors than accurate-other targets). The mean data thus show both a self-advantage as well as an advantage for imitating accurate-other-targets over inaccurate-other targets.[4] However, the degree of improvement across conditions is considerably greater for poor-pitch than accurate participants. The planned contrast of self-advantage scores across participant groups likewise was significant, $t(41) = 2.66$, $p < .01$, although the contrast between groups for the prototype advantage score was non-significant, $t(41) = 0.83, p = .21$.

In addition to primary analyses of imitation, which address accuracy and precision across the entire F0 vector, we also performed more traditional analyses in which a single point-estimate for F0 is extracted for each target and imitated note. First tone onset times were identified through a custom Matlab algorithm that tracks fluctuations in pitch and intensity, followed by visual inspection for accuracy. Then a second Matlab algorithm measured the median F0 for the central portion of each sung syllable (middle 50% of sampled F0 values). This procedure removes the influence of glides that can occur at the beginning and ends of notes, and avoids possible biases from outlying F0 values by using the median. Four such values were extracted for every target and imitation for each trial. Absolute and relative pitch error scores were computed as described by Pfordresher and Brown (2007). The mean absolute difference between every imitated F0 and the corresponding target note was used to assess absolute pitch in imitation. The resulting ANOVA yielded the critical target x participant type interaction, $F(1, 41) = 5.45, p < .05, \eta_p^2 = 0.12$, reflecting a larger self-advantage among poor-pitch singers (M difference between self and other-targets = 112 cents) than among accurate singers (M difference = 39 cents).

2.2.2. Individual differences

Our primary interest here is to determine whether the magnitude of the self-advantage varies with the overall singing ability of a given participant (as measured by mean absolute pitch error across all trials).[5] Here we represent overall singing ability on a continuum rather than by dichotomizing participants into discrete "accurate" versus "poor-pitch" groups in order to better address the full spectrum

---

[4] The advantage for imitating accurate over inaccurate other-targets resembles other findings from music performance demonstrating an imitation advantage for prototypical performances (Clarke & Baker-Short, 1987). Repp (2000) demonstrated an important qualification to this "prototype advantage." Specifically, he found that if imitation success is computed as degree of change from a "spontaneous" performance (termed *relative imitation success*), the prototype advantage can reverse. We tested this prediction, using participants' initial performances as a proxy for a "spontaneous" performance. Results replicated those found by Repp: Mean relative imitation scores (computed as in Repp, 2000) were on average higher for the imitation of inaccurate other targets than for accurate other targets (inaccurate M = 0.63, SE = 0.04, accurate M = 0.37, SE = 0.05, p < .01 according to a paired *t*-test, two-tailed).

[5] We also ran correlation analyses that tested whether the advantage for accurate over inaccurate-other targets varied with overall singing ability. None of these correlations were significant, however.

of individual differences. This type of analysis can better address the possibility that floor effects among accurate singers may confound difference scores representing the self-advantage by showing the performance of all individuals on a continuum. Because we are principally interested in the self-advantage here, for these analyses we averaged across all other-target trials, including accurate and inaccurate-other-targets. Analyses like those reported below that measured degree of advantage for imitating accurate versus inaccurate-other-targets did not show this advantage varying with overall imitation ability.

Fig. 4 shows the relationship between mean absolute error across all trials (X axis) and subsets of the data reflecting the target-type (Y-axis). Fig. 4A shows that the regression of mean absolute error for all other-target trials on mean absolute error in all trials yields a steeper slope ($b = 1.13$) than the regression of self-imitation trials on all trials ($b = 0.72$). The 95% confidence interval around each slope estimate did not overlap (for other-target trials upper CI = 1.20, lower CI = 1.05, for self-imitation trials upper CI = 0.87, lower CI = 0.56). The difference in slopes suggests that the magnitude of the self-advantage increases as overall error increases. This possibility was tested by the regression shown in Fig. 4B, which regresses difference scores measuring the self-advantage (see *Data Analysis*) on mean absolute error across all trials (for illustrative purposes, data points in Fig. 4B are distinguished according to categorization of the participant). The correlation was positive and significant, $r(41) = .50$, $p < .01$, supporting our prediction that the self advantage would increase with the degree of poor-pitch singing. The outlier participant who was removed from this analysis bore out this trend, having a mean absolute pitch error score of 681 cents and a difference score of 522 cents. It is worth pointing out that the self-advantage among poor-pitch singers does not suggest that differences across accurate and poor-pitch singers vanish in self-imitation trials. Indeed, the data in Fig. 4A show that poor-pitch singers remain poor when imitating themselves, relative to accurate singers. This observation is significant in that it suggests that the self-advantage is not simply due to poor-pitch singers repeating the same vocalization on every trial, which would of course lead to identical error for self-imitation trials in both phases of the experiment (Experiment 3 explores this general issue further).

2.2.3. Self-similarity in other-targets

Although poor-pitch singers on average were better able to imitate accurate-other-targets than inaccurate-other-targets (Fig. 3), it is possible that on certain trials a given inaccurate target ended up being a better "match" for the imitator. If so, one would predict that other-targets are better imitated if the target was similar to the initial performance of the participant, recorded during phase 1 of the experiment.

In order to address self-similarity on a trial-by-trial basis, we undertook the following data analysis procedure, which uses regression in order to measure an individual's sensitivity to self-similarity. This
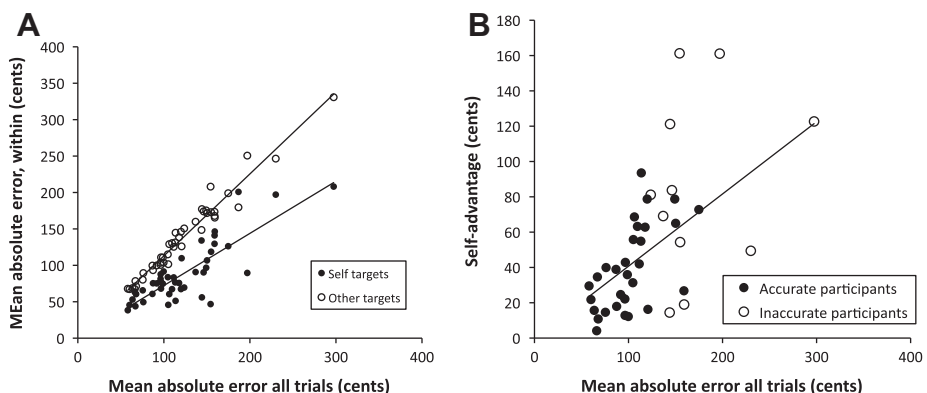
Fig. 4. Scatterplots displaying the relationship between measures of imitation success (X axis), averaged across all trials in phase 2 of Experiment 1, with different target trials (A), and with difference scores representing the magnitude of the self-advantage (B).

analysis comprised three steps. First, as illustrated in Fig. 2, we computed the degree to which these initial performances differed from each target performance of the same melody as an estimate of *spontaneous similarity* between the participant and the other-target. The logic of this measure is that participants who produce vocalizations that are similar to an other-target while not explicitly attempting to match that target are "self-similar" to that target (for similar approaches, see Loehr, Large, & Palmer, 2011; Repp, 2000).

Second, we regressed mean absolute error of vocal imitation (the measure reported in Figs. 3 and 4) on this measure of spontaneous similarity across all trials, separately for each participant. Slopes from these regressions indicate the degree to which imitation error depends on spontaneous similarity. We computed regressions separately for accurate-other and inaccurate-other targets, based on the notion that poor-pitch singers may be more influenced by self-similarity when imitating targets who are also poor-pitch singers than when imitating accurate targets. An example of two such regressions, for a single participant (a poor-pitch singer), are shown in Fig. 5A. Note that the slope for inaccurate-other-targets is much steeper than for accurate-other-targets, suggesting a stronger dependency of mean absolute error of imitation for these trials (*Y* axis) on spontaneous similarity of this participant to the target (*X* axis).

Finally, we treated slope values described in the previous paragraph as measures of *self-similarity dependence*, and regressed these slope values on mean absolute pitch error for absolute imitation. This final step was designed to measure the degree to which poor-pitch singers (who have high values on the *X*-axis, overall imitation error) are influenced by self-similarity.

The lower panels of Fig. 5 plot the slopes from this regression analysis separately for the imitation of inaccurate-other-targets (5B) and accurate-other-targets (5C). Slopes from the regressions shown in Fig. 5A are indicated with arrows. The relationship between the value of the regression slope and mean absolute pitch error for inaccurate-other targets was significant and positive, $r(41) = .66$, $p < .01$. The outlier removed from this analysis followed the general trend, with a mean absolute error value of 771 cents, and a slope of .71. However, the regression analysis performed for accurate-other targets was not significant, $r(41) = .02$. Thus, in addition to showing a self-advantage, poor-pitch singers also benefit when imitating an inaccurate-other-target whose vocalizations resemble their own.

2.2.4. Checking the validity of inferences

We now turn to two analyses designed to test in a more specific way whether the data support the inverse model deficit hypothesis. Data analyses reported above are consistent with this hypothesis but are vulnerable to alternative interpretations.

First, we consider an alternative hypothesis, mentioned earlier, that the self-advantage does not reflect true self-imitation but instead reflects a tendency simply to repeat the same pitch trajectory on every trial. Earlier results do not fully rule out such an account. For example, repetitions of the same errors could still lead to the self-similarity advantage if similar other-targets exhibit similar tendencies on every trial.

A more convincing test (suggested by an anonymous reviewer) is to compare performance on every trial not to the intended target but instead to the ideal performance (imitated in phase 1). If poor-pitch singers simply repeat themselves on each trial, then pitch error scores based on these comparisons should be uniform across all conditions and should not exhibit a self-advantage (or any other difference across conditions). By contrast, accurate imitators, who presumably vary their production according to the accuracy of the target, should exhibit high "error" scores when their imitations of inaccurate other-targets are compared to the ideal targets, and low error scores in other conditions (for which targets better resemble the ideal).

We computed such error scores by using the same data analysis technique described in the *Methods* section but using the idealized pitch trajectory as the "target" for every trial. Resulting mean absolute error scores were submitted to a 2×3 ANOVA with the same design as described for previous analyses. The ANOVA yielded only a significant main effect of target type, $F(2, 82) = 27.03, p < .01, \eta_p^2 = .40$, but no main effect of group ($p = .35, \eta_p^2 = .02$) and no interaction ($p = .12, \eta_p^2 = .05$). Both groups exhibited lowest error scores (indicating greatest similarity to ideal targets) for imitation of accurate other-targets ($M = 179$, $SE = 30$), highest error scores for imitation of inaccurate other-targets ($M = 292$, $SE = 13$), and intermediate for imitation of self-targets
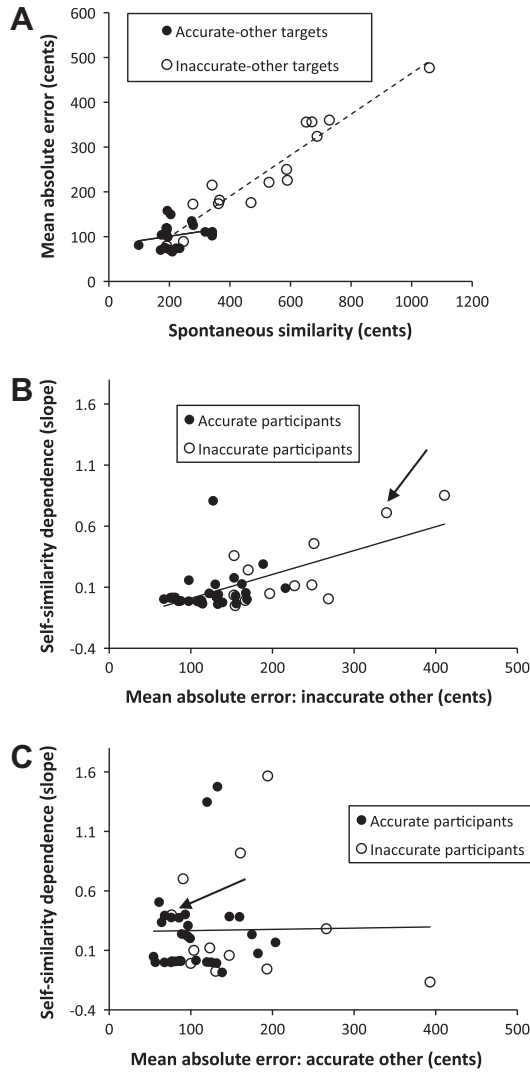
Fig. 5. The relationship between imitation accuracy and sensitivity to self-similarity. Panel (A) shows an illustrative example of how regression coefficients were computed to measure self-similarity dependence. The dotted line represents the regression of mean absolute error on spontaneous similarity for inaccurate-other target trials, and the solid line shows the comparable regression for trials with accurate-other targets. Lower panels show scatterplots displaying the relationship between mean absolute pitch error (X axis), averaged across all trials in phase 2, and regression coefficients based on the relationship between spontaneous similarity and production accuracy (Y axis), computed as in panel (A). This relationship is shown separately for trials with inaccurate other-targets (B) versus accurate other-targets (C). Arrows in lower panels highlight the data points from the example shown in panel (A).

($M = 179$, $SE = 30$). A follow-up ANOVA indicated that the effect of target type was significant within participants who were poor-pitch singers, $F(2, 22) = 7.11, p < .01, \eta_p^2 = .39$.

The second analysis in this section was designed to verify that the self-advantage observed here was based on the imitation of pitch patterns as opposed to variations in timing. Although our data analysis procedure was designed to adjust for differences in the overall duration of imitations (which were small), it did not adjust possible differences in the relative timing of note onsets within

sequences. This is an important concern because a performance that deviates from the target with respect to relative timing (e.g., the first tone is sustained longer and the second tone shorter than in the target) could elevate pitch error scores, but the elevation might be due to the timing rather than the pitch content of imitation. We addressed this concern, as done by Mantell and Pfordresher (2013), by identifying tone onsets and re-sampling each sung tone so that it matched the duration of tones in the target. We then re-computed error scores between these time-adjusted imitations and target trials. Error scores based on these time-adjusted imitations were highly correlated with error scores from the original method, $r = .93$, and means were within a few cents of each other (M for original method = 118, SE = 19, M for time-adjusted imitations = 115, SE = 17). Further analyses verified that the previously obtained interaction of the self-advantage with accuracy of the participant remained in the time-adjusted data.

2.2.5. Tests of third variables

Finally, we consider the role of variables that may provide alternative accounts to the inverse model deficit hypothesis proposed here. A particularly critical factor is episodic memory, which has been a concern of other studies that have demonstrated enhanced abilities to recognize and reproduce one's own actions (cf. Repp & Knoblich, 2004). It is possible that the ability to imitate one's own vocalization may result from the ability to recognize one's own voice (consciously or not) and use that recognition to retrieve a memory trace for the motor movements used to produce the sequence. If so, then the subjective recognition score provided by participants (7 = recognized as other, 1 = recognized as self) should predict accuracy of imitation.

Recognition ratings were significantly different between trials with other-targets (M = 5.83, SE = 0.16) versus self-targets (M = 2.45, SE = 0.20) in the expected direction, $t(43) = 11.36$, $p < .01$. Participants were thus able to discriminate other-targets from their own recordings. Linear detrending was used to remove the effect of self-recognition from imitation using the following procedure. First, mean absolute pitch error scores were regressed on recognition ratings for each participant and trial. Residuals from these regressions were then added to the original mean scores to generate detrended variables that were uncorrelated with recognition ratings. Following detrending, the effect of target type on mean absolute pitch error was still significant though smaller in size, $F(1, 42) = 6.88$, $p < .01, \eta_p^2 = .14$ (compared to $\eta_p^2 = .39$). The pattern of means matched the pattern found in the original data (M for inaccurate-other = 174 cents, SE = 20; M for accurate-other = 137, SE = 20; M for self = 124, SE = 8). Thus, although self-recognition may have played some role in differences in performance across target type, it appears not to be a necessary factor.

Other alternative accounts stem from individual differences in musical training and motivation during the experiment, both of which can influence performance overall and may contribute to the critical self-advantage. Reported years of musical experience was negatively correlated with mean absolute pitch error scores overall ($r = -.30$), as well as the magnitude of the self-advantage ($r = -.29$), as would be expected if musical training led to these findings. However, both correlations fell short of significance. With respect to motivation, we calculated correlations between participants' ratings of effort with the magnitude of the self-advantage. This correlation was not significant, and opposite in sign to what one would expect if poor-pitch singers (who exhibit a larger self-advantage) exhibit lower effort ($r = .33$). Thus, it seems unlikely that the present results can be attributed simply to musical training or to motivation.

2.3. Discussion

The results of Experiment 1 demonstrated an advantage for vocal imitation of pitch when the target is a recording of oneself as opposed to the recording of another singer. Whereas a similar advantage has been shown in the past for matching individual pitches (Hutchins & Peretz, 2012; Moore et al., 2008), this is the first time to our knowledge that the advantage has been shown for the imitation of sung sequences. More important, Experiment 1 demonstrated that the magnitude of this self-advantage varies with overall singing ability such that poor-pitch singers exhibit a larger self-advantage than accurate singers (Moore et al., 2008). This finding follows from the hypothesis described in the introduction: poor-pitch singers are deficient in the inverse modeling of motor control

based on the anticipated outcomes of their actions. Specifically, poor-pitch singers may only have access to vocal-auditory associations based on specific experience and may not be able to generalize vocal motor planning to novel vocal-auditory associations.

Results from Experiment 1 also revealed an overall advantage for imitating prototypical (accurate) other-targets over less prototypical (inaccurate) other targets. This finding is reminiscent of earlier literature on the imitation of timing patterns in piano performance. Clarke and Baker-Short (1987) found that imitation of timing patterns that were linked to phrase structure of music were performed more accurately than the imitation of disrupted timing patterns. The authors accounted for these findings by way of memory schemas. That is, a prototypical timing pattern resonates with memory schemas based on the phrase structure of music, whereas non-prototypical timing patterns can only be stored as episodic traces. However, unlike the self-advantage, the advantage for imitating accurate versus inaccurate-other-targets did not vary as a function of overall singing accuracy.

Finally, poor-pitch singers demonstrated heightened sensitivity to the similarity of other-targets to their own initial performance than did more accurate singers. This was particularly the case for inaccurate-other-targets. This result suggests that an inaccurate-other-target may yield better imitation than an accurate-other-target, as long as the vocalizations produced by the inaccurate-other-targets resemble what one would produce outside the context of vocal imitation.

3. Experiment 2

In Experiment 1, targets comprised the full original recordings of the participant or of others. As such, recordings provided information beyond F0, in particular vocal timbre. Timbre is an important variable in that it provides a cue to the source of the recording (self-recognition) that can influence the salience of pitch through harmonicity (cf. Cariani & Delgutte, 1996), and may interfere with the accuracy of vocal pitch matching if one tries to match the pitch of a timbre that is dissimilar to one's own (Hutchins & Peretz, 2012; Watts & Hall, 2008). However, previous studies on the role of timbre in vocal imitation used trials that involved the matching of single pitches; it is possible that the importance of timbre is reduced when participants imitate more complex pitch sequences. Thus, in Experiment 2 we removed timbral cues by synthesizing the pitch–time trajectories of all recordings using the "hum" command in Praat (Boersma & Weenink, 2013), which generates a 5-formant, voice-like tone. Importantly, this conversion removes articulation information and produces the same vocal timbre for all recordings. Thus, the pattern of pitch change over time becomes the only cue for self-identity, reducing the role of self-recognition. The influence of timbre on pitch matching is thus equated across trials. If the self-advantage observed in Experiment 1 is based on pitch patterns, as we predict it was, then we should find results in Experiment 2 similar to those in Experiment 1. However, if the effects in Experiment 1 were based on similarity of the target's vocal timbre to one's own voice, then the self-advantage (and, critically, the relationship between the self-advantage and overall accuracy) should disappear in Experiment 2.

3.1. Method

3.1.1. Participants

Fourteen participants who had not participated in Experiment 1 participated in exchange for course credit in Introductory Psychology, using the same exclusion criteria as before. Their mean age was 18.79 years (range: 18–23). Five participants were female and nine were male. Ten participants reported at least one year of musical experience at performing an instrument or singing (M = 12.5 summed years of experience across instruments and voice, range = 1–28). Two participants reported formal vocal training outside of a school choir (five participants reported singing in a choir at some point); one participant reported eight years of formal vocal training and the other reported three years of training.

The sample used in Experiment 2 was considerably smaller than the sample from Experiment 1 based on the fact that we incorporated a prescreening test in order to ensure that participants in Experiment 2 represented a wide range of vocal imitation skills, thus maximizing our potential

to address individual differences (this procedure was adopted for a small portion of participants from Experiment 1, see Footnote 1). In the prescreening test, participants ($N = 62$) vocally imitated six sequences in which a single pitch (different on subsequent trials) was repeated 4 times. Pitches on different trials spanned an octave range that surrounded the participant's "comfort pitch" (a single self-selected pitch used to represent a comfortably produced pitch from participant's vocal range). Participants used in Experiment 2 were those who either imitated at least 5 out of 6 trials within 100 cents of all target notes ($n = 8$ accurate singers) or imitated at least 5 of the 6 trials consistently outside a window of 100 cents surrounding each target note ($n = 6$ poor-pitch singers). The proportion of trials categorized in tune from this prescreening task correlated significantly with mean absolute imitation error scores in the main experiment, $r(12) = -.77$, $p < .01$. Although the prescreening data provide an alternate means for classifying participants, we used the same technique as in Experiment 1 (using signed imitation error scores from the main experiment, with a mean score of ±50 cents as a criterion for classification as poor-pitch) for consistency. Based on this criterion, seven participants in Experiment 2 were classified as accurate and seven were classified as poor-pitch singers. Participants categorized as accurate reported more years of music lessons ($M = 9.14$ years, range = 0–20) than participants categorized as poor-pitch singers ($M = 2.71$, range = 0–8), though this difference fell short of significance ($p = .08$). More important, groups did not differ with respect to reported years of vocal training (M for accurate = 1.57 years, for inaccurate = 0 years, $p = .22$).

3.1.2. Materials

The set of database singers matched the set from Experiment 1. As stated earlier, all recordings were synthesized to match a single synthetic vocal timbre using the "hum" command in Praat. Recordings using this timbre resemble the vowel "ah". All participant recordings from the first phase of the experiment were synthesized while participants completed questionnaires before phase 2.
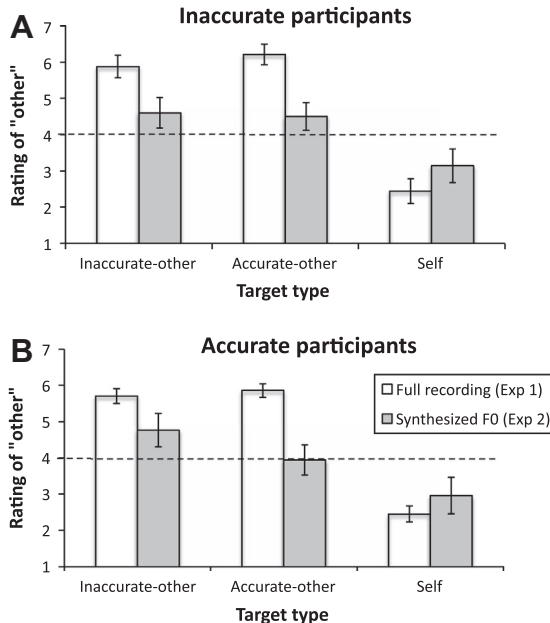
Fig. 6. The effect of target type (self versus other, averaging across accurate and inaccurate other-targets) on recognition scores (7 = other, 1 = self) for Experiments 1 and 2. Error bars display one between-subjects standard error of the mean.

3.1.3. Procedure

The procedure was identical to that used in Experiment 1 and data were analyzed in the same way. Participants imitated the pitch of the targets by producing the vowel "ah", which better matches the syllabification implied by synthesized targets than "dah" as was used in the previous experiment.

3.2. Results

3.2.1. Self-recognition

The neutralization of timbre in Experiment 2 was intended in large part to reduce cues for self-recognition. It is thus critical to determine whether this manipulation did in fact reduce self-recognition, so we report the results of that analysis first. Fig. 6 shows mean self-recognition ratings for all three target types (self, accurate-other, inaccurate-other) across both experiments, which range from 1 to 7. The recognition data from one participant in Experiment 2 were lost due to experimenter error. Ratings were analyzed with a 2 (experiment) $\times$ 3 (target type) $\times$ 2 (participant type) mixed-model ANOVA. Fig. 6 shows recognition ratings broken down by all three factors; note that ideal performance leads to a rating of 1 for all self-target trials and a rating of 7 for all other-target trials. There was a significant main effect of experiment, $F(1, 53) = 13.41, p < .01, \eta_p^2 = .20$, a significant main effect of target, $F(2, 106) = 59.45, p < .05, \eta_p^2 = .53$, and a target x experiment interaction, $F(2, 106) = 11.61, p < .01, \eta_p^2 = .18$. No other effects were significant; the imitation accuracy of participants did not influence recognition ability. The critical target by experiment interaction reflects the fact that recognition ratings in Experiment 2 were all closer to neutrality (a rating of 4), suggesting poorer recognition than in Experiment 1 where ratings were closer to ideal values. This difference across experiments was true for both poor-pitch participants (Fig. 6A) and accurate participants (Fig. 6B). Thus, synthesizing the pitch–time sequences served to reduce, but not completely diminish, recognition performance.

3.2.2. Measures across participants

Fig. 7 shows mean absolute imitation error scores as a function of target type and participant category. The ANOVA (conducted as in Experiment 1) revealed a significant main effect of target type, $F(2, 24) = 10.77, p < .01, \eta_p^2 = 0.47$, a main effect of group, $F(1, 12) = 9.15, p < .05, \eta_p^2 = 0.43$, and a significant interaction, $F(2, 24) = 5.48, p < .05, \eta_p^2 = 0.31$. Both groups exhibited lower errors for self than other targets and for accurate than inaccurate targets. However, this effect was stronger among poor-pitch participants than among accurate participants, leading to a reduction of group differences for self-targets than for other targets, as shown in Fig. 7. As in Experiment 1, the planned contrast of self-advantage scores across participant groups likewise was significant, $t(12) = 2.41, p < .05$, although the contrast between groups for the prototype advantage score was non-significant, $t(12) = 0.48, p = .32$.

In order to compare directly the magnitude of the self-advantage across experiments, we ran a between-subjects ANOVA with the single factor experiment on the mean self-advantage score for each
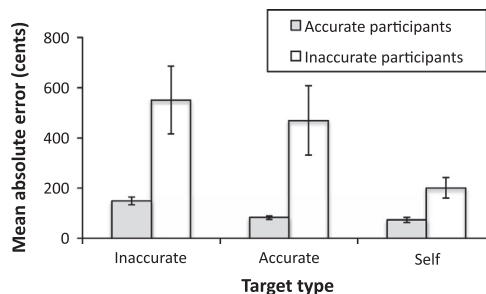

Fig. 7. Effect of target type on mean absolute pitch error in cents for imitative performances in phase 2 of Experiment 2. Error bars denote one between-subjects standard error of the mean.

participant. The main effect of experiment was significant, $F(1, 55) = 11.15, p < .01, \eta_p^2 = .17$. However, the magnitude of the self-advantage was in fact larger in Experiment 2 (M = 175.67, SE = 64.53) than Experiment 1 (M = 49.99, SE = 6.01), and thus does not support the view that reduced self-recognition (as found in Experiment 2) led to a reduced self-advantage. Similar results were found if the data from Experiment 1 reflected only that subset of participants from Experiment 1 who participated in the prescreening task also used in Experiment 2 (see Footnote 1).

3.2.3. Individual differences

We now turn to analyses of individual differences in the self-imitation advantage, carried out as in Experiment 1. Fig. 8A shows that the regression of mean absolute error for all other-target trials on mean absolute error in all trials yields a steeper slope (b = 1.32) than the regression of self-imitation trials on all trials (b = 0.36), as in Experiment 1. The 95% confidence intervals around each slope estimate did not overlap (for other-target trials upper CI = 1.38, lower CI = 1.26, for self-imitation trials upper CI = 0.49, lower CI = 0.23). The apparent increase in the self-advantage with overall error was further verified in a regression of difference scores reflecting the self-advantage on mean error across all trials (shown in Fig. 8B). The correlation was positive and significant, $r(12)$ = .95, p < .01, further suggesting that the reduction of self-recognition in Experiment 2 did not diminish the relationship between poor-pitch singing and the self-advantage.

3.3. Discussion

In Experiment 2, we synthesized all recordings to match a synthetic, voice-like tone. As such, self-recognition cues based on timbre were eliminated, and the influence of timbre matching on vocal imitation was equated across all self and other targets. As a result of this manipulation, self-recognition was reduced relative to Experiment 1. Even so, a significant self-advantage was still observed across all participants, and the self-advantage still varied significantly as a function of overall singing skill. Experiment 2 thus offers new support for the inverse model deficit hypothesis and further suggests that vocal timbre matching has a fairly minimal role for the imitation of sequences, though it may have a stronger influence for pitch matching of single tones (Hutchins & Peretz, 2012; Moore et al., 2008; Watts & Hall, 2008).

This being said, a question remains concerning whether the self-advantage is specific to absolute pitch. Although the standard inverse model framework (as in Fig. 1), suggests that internal models are based on absolute pitch (i.e., zero-order associations), associations based on relative pitch may be more salient in complex sequential behaviors such as music production. Beyond this, the results
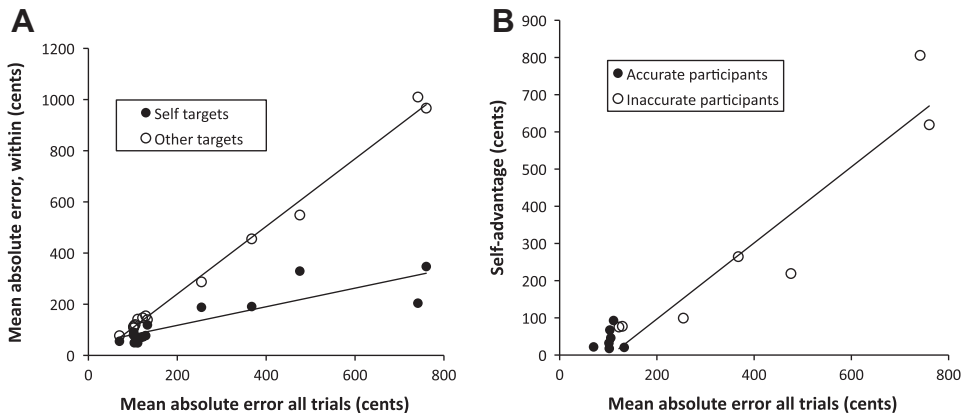

Fig. 8. Scatterplots displaying the relationship between measures of imitation success (*X*), averaged across all trials in phase 2 of Experiment 2, with different target trials (A), and with difference scores representing the magnitude of the self-advantage (B).

of Experiments 1 and 2 lead to an important question concerning the role of absolute pitch. Specifically, the self-advantage may be based on the tendency among poor-pitch singers to vocalize pitches close to a single "comfort pitch", and to experience difficulty straying from that pitch (cf. Pfordresher & Brown, 2007, Experiment 2). We addressed this issue in Experiment 3 by transposing timbre-neutralized recordings of self and other targets to be higher and lower in pitch than the original recordings.

4. Experiment 3

In Experiment 3, targets were re-synthesized (as in Experiment 2) and presented in three different transposition conditions: no transposition (a replication of Experiment 2), all pitches transposed up three semitones, and all pitches transposed down three semitones. If results from previous experiments are specific to absolute pitch information, then results from Experiment 3 should only replicate previous findings for the non-transposed condition. However, the encoding of musical sequences is dominated by relative pitch information. As such we predicted that previous results would be found across transposition conditions.
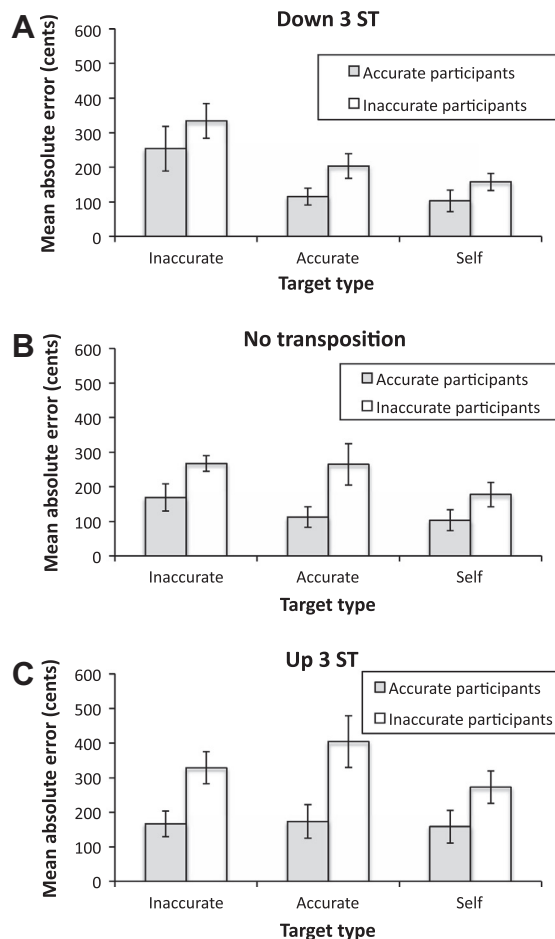
Fig. 9. Effect of target type and target transposition on mean absolute pitch error in cents for imitative performances in phase 2 of Experiment 3. Error bars denote one between-subjects standard error of the mean.

4.1. Method

4.1.1. Participants

Twenty participants who had not participated in Experiments 1 or 2 participated in exchange for course credit in Introductory Psychology, using the same exclusion criteria. As in Experiment 2, participants were drawn from a larger prescreened sample ($N = 57$), in order to ensure a broad range of imitation ability. Their mean age was 18.6 years (range: 18–21). Eleven participants were female and nine were male. Fifteen participants reported at least one year of musical experience at performing an instrument or singing ($M = 7.73$ summed years of experience summed across instruments and voice, range = 0–21). Four participants reported formal vocal training outside of a school choir (M years of training = 5, range 2–9). Six participants reported singing in a choir at some point. As in Experiment 2, the proportion of trials categorized in tune from the prescreening task correlated significantly with mean absolute error scores across all trials in the primary experiment, $r(18) = -.61$, $p < .01$. Based on the classification used in all experiments (signed error scores across all trials), eight participants were categorized as accurate and 12 were categorized as poor-pitch singers. Participants categorized as accurate or poor-pitch did not differ significantly with respect to reported years of music lessons (M for accurate = 5.25 years, M for poor-pitch = 4.13 years, $p = .34$), or reported years of singing lessons (M for accurate < 1 year, M for inaccurate = 1.33 years, $p = .40$).

4.1.2. Materials

In Experiment 3, all targets (including self-targets) were presented in three transposition conditions: normal, transposed up three semitones, and transposed down three semitones. A new set of targets was used so that we could keep the number of trials reasonable given the incorporation of a new factor. Thus we used six other-target singers from previous experiments (three accurate, three inaccurate) meant to represent a wide range of singing ability. Four recordings of each other-target singer were presented, comprising the four melodies used in other experiments.

4.1.3. Procedure

The procedure of Experiment 3 was identical to Experiment 2. Different target types, melodies, and transposition conditions were randomly intermingled across trials.

4.2. Results

4.2.1. Measures across participants

Means across participants were analyzed using ANOVAs based on those from previous experiments, including the additional factor transposition condition (normal, up 3 semitones, down 3 semitones). Corresponding means across conditions are shown in Fig. 9. The ANOVA, which included 3 levels of the factor target type, yielded significant main effects for all factors [target type, $F(2, 36) = 8.54$, $p < .01$, $\eta_p^2 = .32$, participant type, $F(1, 18) = 5.56$, $p < .05$, $\eta_p^2 = .24$, transposition, $F(2, 36) = 6.37$, $p < .01$, $\eta_p^2 = .26$]. Error magnitude varied with target type as in other experiments, were lower for accurate than for poor-pitch participants, and were lower for imitations of non-transposed targets than either transposition condition (with highest error in the upwards transposition condition). In addition there was a significant target type x transposition interaction, $F(4, 72) = 8.72$, $p < .01$, $\eta_p^2 = .33$. This reflects the fact that the effect of target (when averaging across participant type) was larger for downward transpositions (range of maximum to minimum means = 166 cents) than for the other conditions (range of means for no transpositions = 80 cents, for upward transpositions = 85 cents). No interactions with the factor participant category were significant in the ANOVA, including the critical target type x participant type interaction ($p = .28$, $\eta_p^2 = .07$). We ran planned contrasts, as in other experiments, averaging across the three transposition conditions. As in Experiments 1 and 2, the planned contrast of self-advantage scores across participant groups likewise was significant, $t(18) = 2.33$, $p < .05$, although the contrast between groups for the prototype advantage score was non-significant, $t(12) = 0.97$, $p = .17$.
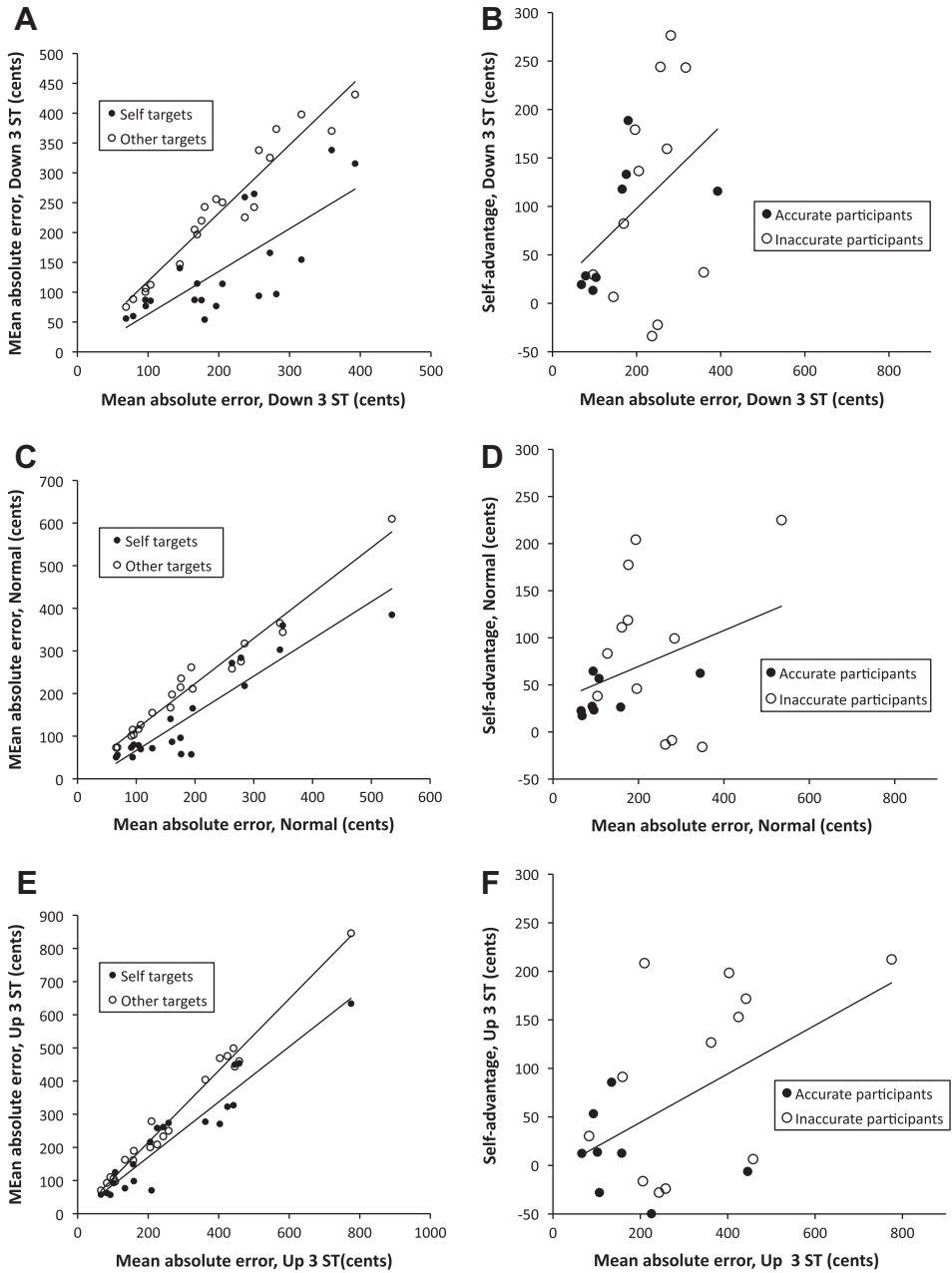
Fig. 10. Scatterplots displaying the relationship between measures of imitation success (X), averaged across all trials in phase 2 of Experiment 3 for a trials in which targets were transposed down 3 semitones (A and B), presented with no transposition (C and D), or transposed up 3 semitones (E and F). Ordinate values represent error scores for different target types (A, C, E), or difference scores measuring the self-advantage (B, D, F).

4.2.2. Individual differences

We analyzed individual differences separately for each transposition condition, correlating overall imitation performance within that transposition condition (X) with scores relating to the self

advantage for that transposition condition (*Y*). Correlation plots showing *Y*-axis values that separate other-targets from self-targets are shown in the left panels of Fig. 10 (Fig. 10A, C and E), whereas regressions involving difference scores for the self-advantage are shown in the right-side panels (Fig. 10B, D and F). For each transposition condition, the magnitude of the self-advantage increased with overall mean absolute error scores, in keeping with relationships found in Experiments 1 and 2. Most important, there was no evidence that this relationship diminished in either of the transposition conditions. In fact, the weakest association was found in the condition that had no transposition (Fig. 10D), which fell slightly short of significance, $r(18) = .33$ (critical value for α of .05, one tailed = .38). Correlations for the other conditions were positive and significant, when transposed down three semitones (Fig. 10B), $r(18) = .51$, $p < .05$, and when transposed up three semitones (Fig. 10F), $r(18) = .42$, $p < .05$. In each transposition condition, the slope relating error scores for trials with other-targets ($b = 1.08, 1.06, 1.14$ for upwards transpositions, no transposition, and downward transpositions, respectively) were steeper than slopes from self-target trials from the corresponding transposition condition ($b = 0.83, 0.87, 0.72$). Unlike Experiments 1 and 2, 95% confidence intervals for the no transposition and downward transposition conditions did overlap. However, in all three conditions the slope estimate for other-target trials exceeded the upper CI for the self-target regres-
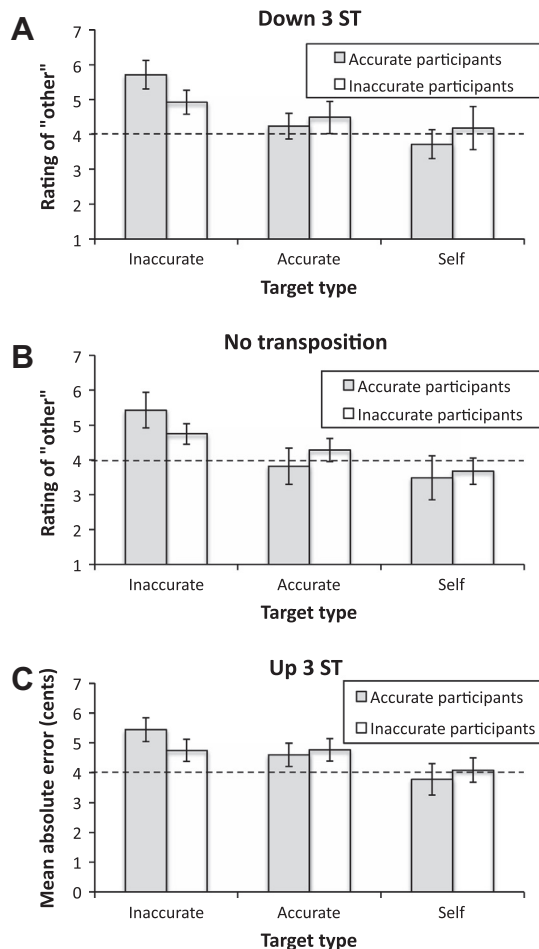
Fig. 11. The effect of target type (self versus other) and transposition condition on recognition scores (7 = other, 1 = self) for Experiment 3. Error bars display one between-subjects standard error of the mean.

sion (upper CI for self-target trials = 0.97, 1.05, 1.02 for upwards transpositions, no transposition, and downward transpositions, respectively), and the slope estimate for self-target trials fell below the lower CI for the other-target regression (lower CI for other-target trials = 1.01, 0.97, 0.99 for upwards transpositions, no transposition, and downward transpositions, respectively).

4.2.3. Recognition

Finally, we address the effects of transposition on self-recognition, using the same design as for mean performance data. Fig. 11 shows recognition ratings (on a scale of 1–7, with 7 indicating participants' certain recognition of the target as "other") as a function of target type, participant category, and transposition. An ANOVA revealed a main effect of target type, $F(2, 36) = 18.02$, $p < .01, \eta_p^2 = .50$, reflecting a tendency for participants to (appropriately) give higher ratings for other targets (inaccurate-other $M = 5.07$, $SE = 0.16$, accurate-other $M = 4.41$, $SE = 0.17$) than self-targets ($M = 3.87$, $SE = 0.19$). There was also a significant target type x participant category interaction, $F(2, 36) = 4.10, p < .05, \eta_p^2 = 0.18$. This interaction reflected the fact that accurate participants were better at identifying the inaccurate-other targets as "other" ($M = 5.53$, $SE = .24$), than were participants who were poor-pitch singers ($M = 4.72$, $SE = .19$). By contrast, performance across participant groups was similar for other target types (for accurate-other targets, accurate participant $M = 4.22$, $SE = .25$, poor-pitch participant $M = 4.41$, $SE = .20$; for self targets, accurate participant $M = 3.66$, $SE = .33$, poor-pitch participant $M = 3.92$, $SE = .19$). Other effects were non-significant, including any effects related to transposition.

4.2.4. Pooled results across experiments

Finally we turn to two analyses that test the specificity of the self-advantage across experiments. We first consider whether individual differences in the size of the self-imitation advantage are correlated with individual differences in the ability to distinguish self from other-targets in recognition. We computed difference scores for ratings of other-targets (averaged across accurate and inaccurate) minus self-targets to measure self-recognition, modeled after difference scores used to measure the self-imitation advantage. A strong correlation between self-recognition and the self-imitation advantage would suggest a relationship between perception and production with respect to recognition. Data from all experiments were pooled, and data from Experiment 3 were averaged across all three transposition conditions within an individual. The resulting correlation was not significant, $r(55) = -.14$, further supporting the notion that self-recognition was not a major determiner of imitation accuracy.

A related concern has to do with whether the self-imitation advantage may be related to individual differences in simple pitch discrimination, in which case the self-advantage may be linked to basic pitch perception mechanisms. Although a pitch discrimination task was not a part of the primary experiment, the prescreening procedure did incorporate a task in which participants discriminated which tone in a sequential pair yielded the higher pitch, and in Experiments 2 and 3 we were able to link data from prescreening with data from the primary experiment. Overall accuracy in the pitch discrimination task ($M = 69\%$ correct, $SD = 17\%$) was not significantly correlated with the magnitude of the self-advantage, $r(29) = .07$, arguing against an interpretation of the self-imitation advantage based on basic pitch perception ability.

4.3. Discussion

The primary goal of Experiment 3 was to determine whether the self-advantage shown in Experiments 1 and 2 is constrained by absolute pitch information in the target. Across participants, the self-advantage was preserved for imitations of transposed targets. In contrast to the self-advantage, the advantage of imitating accurate over inaccurate-other-targets was only present when targets were transposed down three semitones.

5. General discussion

The experiments reported have demonstrated a general advantage in the vocal imitation of one's own sung performances than of another individual's performance, with respect to the absolute pitch content of the target performance. In addition, this self-advantage increases for poor-pitch relative to accurate singers, even when the lack of timbre cues impedes self-recognition (Experiment 2) and when absolute pitch was transposed higher or lower than the original recording (Experiment 3). This is the first study to our knowledge that shows a statistically reliable increase in the self-advantage across a large sample of poor-pitch singers, relative to accurate singers, that scales with overall accuracy (cf. Hutchins & Peretz, 2012; Moore et al., 2008). In addition, the current research is distinctive in that the self-advantage is defined based on sequential pitch patterns, as opposed to individual pitch events. As such, the self-advantage in the present data was robust to alterations of timbre, whereas timbre appears to play a larger role in the self-advantage for matching single pitch events (Hutchins & Peretz, 2012).

Furthermore, all three experiments demonstrated an advantage for imitating other-targets who better approximate musical pitch prototypes (i.e., accurate-other-targets) as opposed to inaccurate-other-targets, although this advantage was not greater for poor-pitch than for accurate singers. Closer examination of performance across different other-targets revealed that poor-pitch singers are in fact better at imitating inaccurate-other-targets if the other-target happens to be more similar to the poor-pitch singer's own spontaneous performance. Thus, even the advantage in imitating accurate-other-targets was qualified by the primary finding of this research, namely the advantage conferred by self-similarity between an individual's spontaneous vocalization tendencies (presumably exhibited in the initial imitative performance of participants in phase 1) and the target.

These results support the hypothesis described in the introduction that poor-pitch singing in most cases reflects a deficit of inverse modeling during vocal imitation. That is, a poor-pitch singer's ability to approximate a vocal pitch–time pattern is limited to the kinds of patterns he or she has produced in the past (for the current experiments, earlier in the session), and the singer has difficulty generalizing imitative performance to novel pitch–time patterns. One way to conceptualize this hypothesis is that poor-pitch singers are inflexible with respect to their ability to mimic pitch patterns, and thus may have not developed the kind of vocal flexibility that generally occurs during childhood (Welch, 1979b, 2006; Welch et al., 2008). In the past this inflexibility has been documented in the form of compression of pitch intervals during imitation, and difficulty imitating pitches other than one's own "comfort pitch" (Pfordresher & Brown, 2007). Here we demonstrate this inflexibility in a more abstract context. Moreover, just as past research suggests that the inflexibility of poor-pitch singers does not represent a limitation in the range of producible pitches by the motor system but is instead specific to vocal imitation (Pfordresher & Brown, 2007; Pfordresher & Mantell, 2009), the results of Experiment 3 show that poor-pitch singers demonstrate a self-advantage, even when the absolute pitch values in a self-target are transposed. In other words, results of Experiment 3 suggest that the influence of inverse modeling is not specific to zero-order matches between the target F0 and motor gestures. Instead, results across all experiments suggest that the inverse model deficit is based on the overall pattern of change in F0 across time throughout the sequence. The inverse model used thus may be an inverse model of a sequence.

The current findings complement recent research suggesting that poor-pitch singers are sensitive to the kind of model that they imitate. As mentioned before, other data suggest poor-pitch singers show an advantage for imitating human as opposed to synthetic vocal (Lévêque et al., 2012), or non-vocal (Moore et al., 2008; Watts & Hall, 2008) targets, whereas accurate singers are not as sensitive to such manipulations. Related to this, a recent study of singing in congenital amusia (a perceptual disorder that is usually associated with poor-pitch singing) suggests that congenital amusics improve significantly for imitative singing tasks as opposed to singing from memory, whereas normal individuals (who in this study were accurate singers) perform similarly well on both tasks (Tremblay-Champoux, Dalla Bella, Phillips-Silver, Lebrun, & Peretz, 2010). Tremblay-Champoux and colleagues also found that amusic singers perform better when the initial pitch of their production matches the target, a relationship that bears some resemblance to the finding reported here that imitation of

inaccurate-other-targets among poor-pitch singers is sensitive to spontaneous similarity (see Fig. 5). Both findings suggest a reliance on the physical form of the target, and in particular the relatedness of the target to vocalizations of the imitator.

What distinguishes the present results is the focus here on sequential information related to the pattern of changes in pitch over time, and individual differences in the production of such patterns. In contrast to studies showing a human-model advantage, which have focused on matching of single pitches, we did not find that alterations of timbre eliminated the self-advantage effect found here. This is likely due to the nature of our task, which emphasized sequential information. That being said, it is worth noting that the human-voice advantage found earlier was also present in our data, with mean absolute error scores in Experiment 2 (for example) being nearly double those in Experiment 1 (cf. Figs. 3 and 7). A follow-up t-test comparing the data from Experiments 1 and 2 with respect to mean absolute pitch error (averaging across target types) was significant, $t(56) = 2.71$, $p < .01$, with lower mean error in Experiment 1 ($M = 134.76$, $SE = 9.9$) than Experiment 2 ($M = 254.64$, $SE = 43.5$). A similar result was identified by Mantell and Pfordresher (2013), who found that individuals could more accurately imitate the pitch–time information of worded speech and song sequences than the same sequences presented with neutralized, wordless timbre.

Overall, an accumulating set of results suggests that poor-pitch singers are inflexible with respect to the conditions under which they can accurately reproduce a pitch pattern by singing. It is unlikely that this inflexibility is simply due to restrictions in the range of F0 values poor-pitch singers can produce. Poor-pitch singers do not typically show a restriction of vocal range in non-imitative tasks (Pfordresher & Brown, 2007), and in Experiment 3 of the present study poor-pitch singers exhibited a self-advantage when target recordings were transposed. As such, this inflexibility seems to be a property of sensorimotor translation that is necessary for vocal imitation. The internal models framework described in the introduction provides a plausible explanation of such phenomena.

We presented results using analyses that dichotomized participants into groups based on overall production, as well as analyses using a regression framework. The approach to analyzing individual differences by dichotomizing participants was used in order to draw connections between the present research and other studies that have dichotomized participants (e.g., signed pitch error, used by Pfordresher & Brown, 2007). However, we consider the regression framework, though less familiar for this kind of research, to have two advantages over the traditional approach. First, because our primary measures are difference scores, we wanted to demonstrate that differences across groups do not simply reflect floor effects in mean absolute error scores for accurate singers. Second, because performance was predicted to vary across target conditions, some uncertainty emerges with respect to what conditions might best be used to classify group differences. As can be seen in our results, individual differences do appear to fall along a continuum, although in every experiment there is a cluster of accurate scores. Ultimately, our view is that treating the categories "accurate" and "poor-pitch" as dichotomies can be useful for the purpose of complicated analyses but at the same time is inherently limited. A fuller understanding of the deficit(s) involved can be derived from addressing individual differences along a continuum as well as across groups. Similar procedures have been followed in understanding other music-related deficits, such as congenital amusia, which is characterized by striking qualitative differences but is at the same time associated with differences in test performance on the Montreal Battery for the Evaluation of Amusia (MBEA) that fall on a continuum (Henry & McAuley, 2010; Peretz, Champod, & Hyde, 2003). Moreover, studies addressing the relationship between singing and pitch discrimination among school children have revealed different kinds of effects when participants were separated into three rather than two groups (Demorest & Clements, 2007).

These results have important implications for the treatment of poor-pitch singing. In educational settings, it has been noted anecdotally that poor-pitch singers may be excluded from music making, sometimes being asked to "mouth" the words to a song rather than sing along (Welch, 2006). The current results suggest that such practices are inappropriate, based on the fact that poor-pitch singers can imitate pitch contours accurately under certain conditions (Pfordresher & Brown, 2007). Specifically, treatments of poor-pitch singing should begin with a foundation based on self-imitation – which is advantageous to the individual – then proceed gradually to sequences that increasingly diverge from the singer's spontaneous tendencies. Such training protocols might facilitate the development of internal models for the vocal imitation of pitch.

In conclusion, we have reported evidence consistent with the hypothesis that poor-pitch singing is primarily due to a sensorimotor translation deficit involving the ability to plan vocal-motor actions based on their anticipated outcomes (i.e., inverse modeling). The construct of inverse models has been used in the past to account for speech production deficits (Max et al., 2004), however, this is the first research to our knowledge that tests the self-advantage prediction in the context of melody imitation. This set of experiments thus adapts the internal models framework to a new performance context, based on a novel prediction from this framework. The internal models construct thus provides a robust framework for understanding a broad range of behaviors. An important issue for the future is to determine whether internal modeling of vocal production proceeds separately for different levels of organization (e.g., pitch contours versus articulation), such that imitative deficits at one level do not lead to deficits at a different level.

Acknowledgments

References

Berkowska, M., & Dalla Bella, S. (2009). Acquired and congenital disorders of sung performance: A review. *Advances in Cognitive Psychology, 5*, 69–83.

Boersma, P., & Weenink, D. (2013). *Praat: Doing phonetics by computer. Version 5.3.23* (Computer program).

Bradshaw, E., & McHenry, M. A. (2005). Pitch discrimination and pitch-matching abilities of adults who sing inaccurately. *Journal of Voice, 14*, 431–439.

Cariani, P. A., & Delgutte, B. (1996). Neural correlates of the pitch of complex tones. I. Pitch and pitch salience. *Journal of Neurophysiology, 76*, 1698–1716.

Clarke, E., & Baker-Short, C. (1987). The imitation of perceived rubato: A preliminary study. *Psychology of Music, 15*, 58–75.

Dalla Bella, S., Giguere, J. F., & Peretz, I. (2007). Singing proficiency in the general population. *Journal of the Acoustical Society of America, 121*, 1182–1189.

Dalla Bella, S., Giguere, J. F., & Peretz, I. (2009). Singing in congenital amusia. *Journal of the Acoustical Society of America, 126*, 414–424.

Demorest, S. M., & Clements, A. (2007). Factors influencing the pitch-matching of junior high boys. *Journal of Research in Music Education, 55*, 190–203.

Henry, M. J., & McAuley, J. D. (2010). On the prevalence of congenital amusia. *Music Perception, 27*, 413–424.

Heyes, C. (2005). Imitation by association. In S. H. N. Chater (Ed.), *Perspectives on imitation: From mirror neurons to memes* (pp. 157–176). Cambridge, MA: MIT Press.

Hickok, G., Houde, J. F., & Rong, F. (2011). Sensorimotor integration in speech processing: Computational basis and neural organization. *Neuron, 69*, 407–422.

Hutchins, S., & Peretz, I. (2012). A frog in your throat or in your ear? Searching for the causes of poor singing. *Journal of Experimental Psychology: General, 141*, 76–97.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current Opinion in Biology, 9*, 718–727.

Knoblich, G., & Prinz, W. (2001). Recognition of self-generated actions from kinematic displays of drawing. *Journal of Experimental Psychology: Human Perception and Performance, 27*, 456–465.

Knoblich, G., Seigerschmidt, E., Flach, R., & Prinz, W. (2002). Authorship effects in the prediction of handwriting strokes: Evidence for action simulation during action perception. *The Quarterly Journal of Experimental Psychology, 55A*, 1027–1045.

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences, 97*, 11850–11857.

Kuhl, P. K., & Meltzoff, A. N. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. *Journal of the Acoustical Society of America, 100*, 2425–2438.

Lévêque, Y., Giovanni, A., & Schön, D. (2012). Pitch-matching in poor singers: Human model advantage. *Journal of Voice, 26*, 293–298.

Loehr, J. D., Large, E. W., & Palmer, C. (2011). Temporal coordination and adaptation to rate change in music performance. *Journal of Experimental Psychology: Human Perception and Performance, 37*, 1292–1309.

Mantell, J. T., & Pfordresher, P. Q. (2013). Vocal imitation of song and speech. *Cognition, 127*, 177–202.

Max, L., Guenther, F. H., Gracco, V. L., Ghosh, S. S., & Wallace, M. E. (2004). Unstable or insufficiently activated internal models and feedback-biased motor control as sources of dysfluency: A theoretical model of stuttering. *Contemporary Issues in Communicative Sciences and Disorders, 31*, 105–122.

Moore, R., Estis, J., Gordon-Hickey, S., & Watts, C. (2008). Pitch discrimination and pitch matching abilities with vocal and nonvocal stimuli. *Journal of Voice, 22*, 399–407.

Peretz, I., Champod, A. S., & Hyde, K. (2003). Varieties of musical disorders: The Montreal battery of evaluation of amusia. *Annals of the New York Academy of Sciences, 999*, 58–75.

Pfordresher, P. Q., & Brown, S. (2007). Poor-pitch singing in the absence of tone deafness. *Music Perception, 25*, 95–115.

Pfordresher, P. Q., & Brown, S. (2009). Enhanced production and perception of musical pitch in tone language speakers. *Attention Perception & Psychophysics, 71*, 1385–1398.

Pfordresher, P. Q., Brown, S., Meier, K. M., Belyk, M., & Liotti, M. (2010). Imprecise singing is widespread. *Journal of the Acoustical Society of America, 128*, 2182–2190.

Pfordresher, P. Q., & Mantell, J. (2009). Singing as a form of vocal imitation: Mechanisms and deficits. In J. Louhivouri, T. Eerola, S. Saarikallio, T. Himberg, & P.-S. Eerola (Eds.), *Proceedings of the 7th triennial conference of European Society for the Cognitive Sciences of Music* (pp. 425–430). Jyväskylä, Finland: Jyväskylä University.

Repp, B. H. (2000). Pattern typicality and dimensional interactions in pianists' imitation of expressive timing and dynamics. *Music Perception, 18*, 173–211.

Repp, B. H., & Knoblich, G. (2004). Perceiving action identity: How pianists recognize their own performances. *Psychological Science, 15*, 604–609.

Schmidt, R. A. (1975). A schema theory of discrete motor skill learning. *Psychological Review, 82*, 225–260.

Tomasello, M. (1990). Cultural transmission in the tool use and communicatory signaling of chimpanzees? In S. Parker & K. Gibson (Eds.), *Language and intelligence in monkeys and apes: Comparative developmental perspectives* (pp. 274–311). Cambridge, UK: Cambridge University Press.

Tremblay-Champoux, A., Dalla Bella, S., Phillips-Silver, J., Lebrun, M.-A., & Peretz, I. (2010). Singing proficiency in congenital amusia: Imitation helps. *Cognitive Neuropsychology, 27*, 463–476.

Tsang, C. D., Friendly, R. H., & Trainor, L. J. (2012). Singing development as a sensorimotor interaction problem. *Psychomusicology: Music, Mind & Brain, 21*, 31–44.

Watts, C. R., & Hall, M. D. (2008). Timbral influences on vocal pitch-matching accuracy. *Logopedics Phoniatrics Vocology, 33*, 74–82.

Welch, G. F. (1979a). Poor pitch singing: A review of the literature. *Psychology of Music, 7*, 50–58.

Welch, G. F. (1979b). Vocal range and poor pitch singing. *Psychology of Music, 7*, 13–31.

Welch, G. F. (2006). Singing and vocal development. In G. McPherson (Ed.), *The child as musician: A handbook of musical development* (pp. 311–329). New York: Oxford University Press.

Welch, G. F., Himonides, E., Saunders, J., Papageorgi, I., Rinta, T. C. S., et al (2008). *The national singing programme for primary schools in England: An initial baseline study*. London.: Institute of Education.

Wolpert, D. M. (1997). Computational approaches to motor control. *Trends in Cognitive Sciences, 1*, 209–216.