A Mechanism for Sensorimotor Translation in Singing: The Multi-Modal Imagery Association (MMIA) Model

Peter Q. Pfordresher
University at Buffalo, State University of New York

Andrea R. Halpern
Bucknell University

Emma B. Greenspon
University at Buffalo State University of New York

We propose a new framework to understand singing accuracy, based on multi-modal imagery associations: the MMIA model. This model is based on recent data suggesting a link between auditory imagery and singing accuracy, evidence for a link between imagery and the functioning of internal models for sensorimotor associations, and the use of imagery in singing pedagogy. By this account, imagery involves automatic associations between different modalities, which in the present context comprise associations between pitch height and the regulation of vocal fold tension. Importantly, these associations are based on probabilistic relationships that may vary with respect to their precision and accuracy. We further describe how this framework may be extended to multi-modal associations at the sequential level, and how these associations develop. The model we propose here constitutes one part of a larger architecture responsible for singing, but at the same time is cast at a general level that can extend to multi-modal associations outside the domain of singing.

Received: June 7, 2014, accepted October 1, 2014.

Key words: poor-pitch singing, vocal imitation of pitch, auditory imagery, internal models, singing development

Singing is a complex, dynamic, multi-dimensional behavior that requires the integration of multiple motor, perceptual, and cognitive functions (Berkowska & Dalla Bella, 2009; Dalla Bella, Berkowska, & Sowinski, 2011; Levitin, 1999; Sundberg, 1989; Welch, 1979; Zarate, 2013). An overarching framework to understand this integration was presented at the beginning of this volume (Pfordresher et al.). We here focus on a specific component: *sensorimotor translation*. By sensorimotor translation, we refer to the mapping of a given sensory continuum onto a related continuum based on motor control. In a complex behavior like singing there are of course many such continua that one could address. As a starting point we consider two continua that are most closely linked to pitch accuracy: associations between phonatory motor control (muscle movements used to regulate vocal fold tension) and perceived pitch height.[1] In order to imitate pitches of a melody accurately while singing, one must be able to map pitch height onto the control of laryngeal muscles in a way that leads to accurate reproduction of the sequence from either long-term memory, or in imitating a model. We use the term "translation" because the continua under consideration differ in non-trivial respects.

Recent research suggests that individual differences in singing may in large part stem from problems of sensorimotor translation (Berkowska & Dalla Bella, 2009; Hutchins & Peretz, 2012; Pfordresher & Brown, 2007; Pfordresher & Mantell, 2014) that may reflect a problem of vocal pitch imitation that extends also to the imitation of speech prosody (Mantell & Pfordresher, 2013; Wisniewski, Mantell, & Pfordresher, 2013). As such, the model in this paper is not so much a model of singing as it is a model of sensorimotor associations that guide vocal imitation of pitch. In keeping with this more general theme, we refer to individuals who are unable to accurately replicate pitch vocally as experiencing a *vocal pitch-imitation deficit*, or VPID. Importantly, the model leads to a characterization of vocal imitation abilities as a continuum, such that VPID may be considered a deficit that varies in degree rather than dichotomously. Thus the model itself is agnostic with respect to the difficult issue of what kind of measures and cutoffs may constitute a good dividing line between "accurate" and "VPID" imitators (see Dalla Bella, this volume). Moreover, as we discuss later, the model presented here need not be limited simply to those associations involved in the vocal imitation of pitch, but can be extended to

[1] Other aspects of singing, such as control of respiration (motor) and vocal loudness (perceptual), are also critical to singing and are addressable in principle by the present framework.

Music Perception, volume 32, issue 3, pp. 242–253, issn 0730-7829, electronic issn 1533-8312. DOI: 10.1525/mp.2015.32.3.242

other forms of imitation, and possibly even cross-modal associations of mental images that are not imitative in nature (e.g., synaesthesia).

Despite the importance of sensorimotor translation, at present we lack a specific framework for understanding how this component works. Recent studies (Pfordresher, 2011; Pfordresher & Mantell, 2014) have invoked the construct of internal models, which include two components: forward and inverse models (e.g., Kawato, 1999; Wolpert, 1997). An internal model represents the way in which the brain internalizes causal associations between perception and action. Within such a framework, VPID may be said to come from a deficient inverse model: a model responsible for converting a perceptual event into the kind of motor plan necessary to reproduce it. (Forward models, which may play a less prominent role in vocal imitation, account for how motor planning generates an anticipated perceptual representation of an action's outcome.) However, saying that singing involves the use of an internal model is in a sense simply rephrasing the assertion that singing involves sensorimotor translation. The framework we propose here, we hope, specifies a framework with respect to functional sensorimotor relationships and the basis of the multi-modal associations that are involved, with a focus on sensorimotor translation in the vocal imitation of pitch.

We refer to our approach as the *multi-modal imagery association* (MMIA) model. In this framework, sensorimotor translation in singing is considered to represent one example of a broader class of mental associations that can play a role in motor planning and perception. A core assumption of this model is that multi-modal associations are typically not based on a 1:1 mapping across the associated continua. Mapping is typically noisy (imprecise) and may be biased toward specific values ("inaccurate," in the statistical sense of the term). We propose that the nature of this mapping results from interplay between associations learned through statistical environmental contingencies derived from past experiences, and generalizations based on these contingencies. Importantly, people showing a deficiency in making such associations may do poorly in tasks that rely on this mechanism. For those with VPID, the deficiency results from ineffective mapping of auditory images onto motor images for phonation.

Why Imagery?

The current literature on singing accuracy does not typically invoke mental imagery as an important component, at least not explicitly. Our proposal here is that mental imagery forms the core of sensorimotor associations that drive vocal pitch imitation and other abilities that rely on cross-modal associations. We have several reasons for proposing such a link.

First, many studies to date suggest that perceptually based mental imagery elicits motor associations. In neuroimaging studies, auditory imagery gives rise to activations in what are known as motor planning areas (Leaver, Van Lare, Zielinski, Halpern, & Rauschecker, 2009). Such results suggest that engaging in auditory mental imagery primes motor planning areas, much like an inverse internal model hypothesizes that perceptual input guides motor planning. Behavioral studies take such associations a step further by showing that motor tasks (e.g., subvocal articulation) can interfere with mental imagery (e.g., repeating speech sounds, Smith, Reisberg, & Wilson, 1992).

Second, our invocation of imagery coheres with recent theoretical proposals concerning sensorimotor associations in music. Keller (2012), for instance, explicitly linked "musical imagery" (by which he means multi-modal imagery) with the functioning of internal models for motor planning in music. Furthermore, in proposing a link between perception and action with respect to imagery, we propose that effective sensorimotor translation occurs at an abstract level of representation, more so than other models that suggest associations based on overt motor and perceptual representations (e.g., Berkowska & Dalla Bella, 2009, Hutchins & Moreno, 2013). Our more abstract level coheres with a major theoretical framework for sensorimotor integration, the Theory of Event Coding (Hommel, 2009; Hommel, Müsseler, Aschersleben, & Prinz, 2001; Prinz, Aschersleben, & Koch, 2009). A strong claim from this framework is that a common representation for perception and action exists at an abstract level in which actions are coded with respect to goals, and perception is coded with respect to the distal event. In keeping with this logic, brain areas subserving imagery almost always include higher-order association areas but less often are found to include primary sensory areas (e.g., Zatorre & Halpern, 2005).

Third, self-report indices suggest that individuals with motor imitation deficiencies also suffer from deficient imagery. A recent study by two of us (Pfordresher & Halpern, 2013) reported a positive correlation between self-reported vividness of auditory imagery and accuracy of pitch imitation. In the domain of manual control, individuals with deficient motor planning ability (ataxia) also exhibit deficiencies of motor-related imagery (Buxbaum, Johnson-Frey, & Bartlett-Williams, 2005).

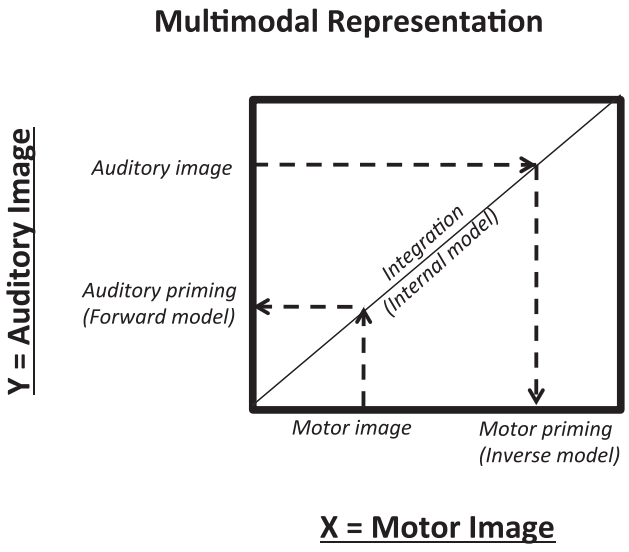

FIGURE 1. The multimodal representation framework for imagery and vocal production. Shown here is the interaction of two out of many dimensions, pitch height (Y), and laryngeal tension (X) which for simplicity are assumed to share a linear association. The formation of mental images within each modality generates a value on one axis. The multimodal representation associates this image with some other modality, through an internal model of how the modalities are associated. This plot illustrates the ideal hypothetical internal model. Forward and inverse modeling in this framework related to the directional relationship between modality of the image and priming of the associated modality.

Based on these points, we propose that the basis of sensorimotor translation in singing, and possibly in other motor imitation tasks, involves associations between dimensions of mental images based on perception, with corresponding dimensions associated with actions that could reproduce the percept. Based on this logic, individual differences in vocal pitch imitation (e.g., singing) may in large part be based on the nature of the associations between modalities involved; most critically, associations between pitch height and laryngeal muscle movements that adjust vocal fold tension.

MMIA Model: Basic Assumptions

The MMIA model is a general framework designed to account for the intersection of multiple modalities during mental imagery. As a starting point (based on its relevance to the present topic) we focus on the intersection between the dimension of pitch in the auditory modality, and laryngeal tension in the motor modality. These are represented in Figure 1 as two dimensions in a Cartesian coordinate system. We consider this configuration to be most plausible given that the modalities under discussion are in principle independent. We frame the interactions between modalities as resulting from the kind of function that relates the two dimensions (i.e., their mapping relationship) as opposed to the kind of coordinate system that defines the mapping relationships that can exist in any case.

Each axis represents a dimension along which different mental images within a single modality may be defined. In the *y*-axis of Figure 1, values represent differences in the pitch height of an auditory mental image, whereas *x*-axis values represent motor imagery associated with a certain level of vocal fold tension. Each of these dimensions is of course a simplification (pitch emerges from integration of spectral information, and vocal fold tension emerges from planning of synergies across a collection of muscles), but the simplifications here relate to a level of abstraction at which integration is likely to occur (a point we revisit later, cf. Hommel et al., 2001).

It is of course possible that distortions exist with respect to how distal perceptual events map onto points on these axes, and we will take up such possibilities in the next section. As a starting point, however, we focus specifically on the function relating points on each axis. This function is critical for us, as it constitutes the multi-modal association being modeled. Each point on the function represents the association between mental images relating to pitch height and laryngeal tension (as controlled by laryngeal muscles). Furthermore, the function effectively operates as an internal model of the auditory-motor system, though the implementation here differs in important ways from standard internal model architectures as we will discuss later.

The kind of association shown in Figure 1 is ideal in that auditory and motor modalities here are associated according to a 1:1 relationship. This mapping is both accurate (unbiased, given the slope and intercept), and precise (i.e., consistent: a given pitch height will always lead to the same associated motor gesture). Given such an internal representation, mapping relationships between auditory and motor modalities would be perfect. Note that we are here assuming that a linear relationship is appropriate. Though some data suggest a nearly linear relationship between F0 and Cricothyroid activity (Roubeau, Chevrie-Muller, & St. Guily, 1997) at present we must consider the linear mapping relationship shown here to be hypothetical.

Note that the associations accounted for in this model are bidirectional. Auditory imagery (*y*-axis) may prime motor imagery (*x*-axis) or vice versa. As such the model predicts that deficits will be bidirectional in nature; although VPID manifests in associations that go in one

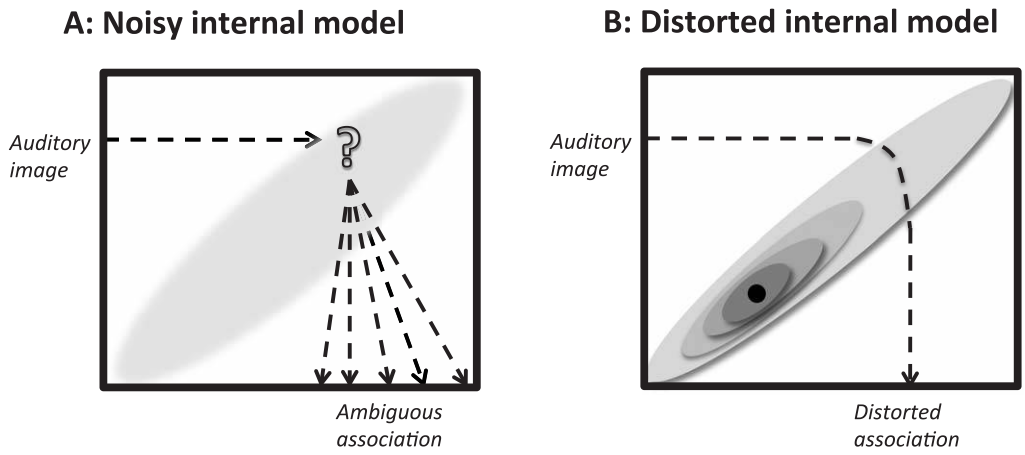

FIGURE 2. Two hypothetical multimodal representations that can lead to deficits such as the vocal pitch imitation deficit (VPID). The deficient internal model in panel A constitutes a broad equiprobable space, leading to noisy and thus unreliable cross-modal associations. The deficient internal model in panel B is also noisy compared to Figure 1, but includes a bivariate probability distribution (darker shades = higher probability) that introduces systematic distortions toward the mode of the distribution, hypothesized to be the singer's comfort pitch.

direction (auditory to motor), disrupted associations of imagery should also flow in the opposite direction. In this way, the MMIA model provides a simplified account for the different functions of an internal model. Within the model, the function of an "inverse model" amounts to the priming of motor imagery via auditory imagery, whereas the function of a "forward model" comes from associations in the reverse direction.

Distortions of Imagery and VPID

This framework supplies an account of VPID based on distortions in the mapping between perception and action. The distortions can be of several types. One possibility is that no mapping exists. This most extreme account would represent the multimodal imagery space as empty, such that input on one axis could be related to any point on the other axis with equal probability. Such an account is unlikely to be successful, given that pitch production even in the poorest singers that have been measured is not fully random. Although there is more random-like behavior in singers exhibiting VPID than among more accurate singers (Berkowska & Dalla Bella, 2013, Pfordresher, Brown, Meier, Beylk, & Liotti, 2010), a model of VPID should encompass other systematic behaviors. One is inflexibility. The direction of VPID errors has been shown to drift in the direction of the individual's "comfort pitch" (Sergeant, 1994; cf. Hutchins, Zarate, Zatorre, & Peretz, 2010; Pfordresher & Brown, 2007), the size of imitated pitch intervals is compressed (Dalla-Bella, Giguere, & Peretz, 2009, Liu et al., 2013, Pfordresher & Brown, 2007), and VPID singers exhibit a greater tendency to perseverate on past pitch patterns when transferring from one sequence to another (Wisniewski et al., 2013).

Figure 2 shows two candidate multi-modal representations for VPID individuals, both are consistent with aspects of existing data. In Figure 2A, the associations are diffuse, though the dimensions are significantly correlated. In place of the straight line of Figure 1, multimodal associations are determined via a grey oval, with each point on the oval representing a bivariate probability distribution. Within this framework, an image on one dimension does not intersect the other dimension at a perfectly predictable angle, as in Figure 1, but may "fan out" to any of a number of points on the alternate axis based on the curvature in the space. As a result, mapping of any point on one axis to the other axis forms a probability distribution rather than a 1:1 association. On average, this would lead to performance being accurate, but imprecise. The average of all sung pitches would be the correct pitch (thus accurate, in statistical terms), though there could be considerable variability in each individual sung pitch (imprecise, in statistical terms), based on the variability of the mapping relationship.

Computationally, the mapping in Figure 2A may be implemented by assuming that the likelihood that a level of laryngeal tension (x) is mapped to a level of pitch height (y) is based on a probability distribution, as in:

$$Zmap_{i,j} = \exp\left[-\frac{(y_j - x_i)^2}{2\sigma^2_{map}}\right] \quad (1)$$

Zmap is the probability of mapping *y* to *x*. Equation 1 predicts a mapping relationship that is unbiased, as in Figure 1, but that is also noisy (as determined by the parameter σ^2_{map}). Although the mapping in Figure 2A is plausible, it is not consistent with many important results from the present literature on VPID, suggesting that vocal production is not simply less consistent (imprecise) but also exhibits bias (i.e., is inaccurate) as in the aforementioned tendency for produced pitches to drift toward one's "comfort pitch." Thus, we incorporate bias into the system, as in Figure 2B.

Figure 2B illustrates a hypothetical representation that is consistent with recent findings from the literature on VPID. A good starting assumption is that the mapping between perception and action for poor-pitch singers is (a) more variable than for accurate singers, as in Figure 2A and (b) biased in the direction of their comfort pitch. This can be represented in the form of a bivariate distribution whose mode represents the distorting effect of the comfort pitch, with the width of the oval representing imprecision in the system. Figure 2B shows such a distribution as a contour plot, with the darkness of different ovals within the representation reflecting increased probability of accurate mapping between axes. The black dot, showing the mode of the distribution, reflects a hypothetical value for the participant's "comfort pitch," leading to best mapping between X and Y at that point, for other associations between X and Y, the mapping will be "drawn in" toward the comfort pitch, which would function like an attractor state for the multi-modal representation (Kelso, 1995). Likewise, auditory imagery for pitch content may also be associated less reliably with motor targets as well as being distorted. In both cases, motor activity will be both less reliably associated with production and compressed in range. Behaviorally, the biasing effect of the comfort pitch would lead to a compression of pitch range toward one's comfort pitch, as found in the data. In addition, the general oval shape in Figure 2B, as in Figure 2A, would lead to a reduction of precision in behavior.

Computationally, the framework shown in Figure 2B results from the product of the probabilities generated by Equation 1, with values from a second probability distribution based on differences between perceived pitch heights (*y*) and the biasing effect of a participant's comfort pitch.

$$Zbias_{i,j} = \exp\left[-\frac{(y_j - x_{bias})^2}{2\sigma^2_{bias}}\right] \quad (2)$$

In this equation, x_{bias} is a constant for all values of *i*. Our starting assumption is that x_{bias} represents the participant's comfort pitch although obviously it could reflect any source of bias that misdirects sensorimotor mapping (e.g., difficulty in separating one's own part from another part in a chorus). Taken on its own, Equation 2 leads to a single probability distribution that when convolved with the distribution from Equation 1 generates the kind of distribution shown in Figure 2B.

The joint product of *Zmap* and *Zbias*, defined for all *i* and *j*, determines the multi-modal mapping relationship. A particularly important factor in this relationship has to do with the relative balance of the two variance parameters: σ^2_{map} and σ^2_{bias}. Increasing variance associated with one factor in the mapping increases the impact of the other factor. As such, different relationships between these variance components yield predictions for different kinds of VPID behavior that may be assessed empirically. For example, if σ^2_{map} is very high relative to σ^2_{bias} then the probability of an appropriate mapping of X to Y becomes lower due to the greater variability, and performance as a result is increasingly dominated by the impact of the biasing parameter. At the extreme, this would lead to predicted monotone singing, as every heard pitch (Y) is mapped to an X value close to the individual's comfort pitch. Reversing this balance leads to less dominance of the biasing factor (e.g., comfort pitch). On average, mapping relationships between X and Y will be accurate, though there still can be considerable variability (imprecision) in this relationship. By way of example, consider the effects of increasing variance associated with bias. Assuming the bias reflects comfort pitch, a wider distribution leads to a range of pitches that are all "comfortable," which is of course what one tries to achieve as a singer. Thus, widening this range reduces the biasing effect of the comfort pitch, while also reducing possible negative influences from variance associated with mapping.

The Role of Time

A limitation of the representation described above is that it focuses on a single pitch at a single time. Most singing, by contrast involves the reproduction of sequences. Moreover, evidence suggests that sequential pitch structure has important and somewhat counter-intuitive effects on singing accuracy. Specifically, VPID singers improve when imitating sequences with pitch variation, in contrast to sequences involving a single repeated pitch. Accurate singers have shown the reverse pattern, and deteriorate in accuracy when sequences feature more diversity of pitch (Pfordresher & Brown, 2007). Furthermore, recent evidence suggests that vocal imitation may be sensitive to fluctuations within sung pitches, particularly when sung

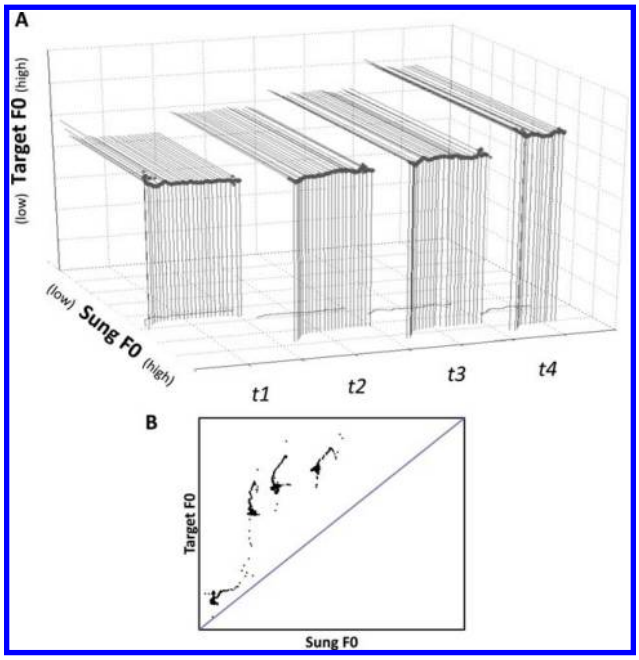

FIGURE 3. Multimodal representation applied to continuous variation of pitch across a sequence. Panel A: Mapping of pitch height (vertical) to laryngeal movement (depth) as a function of time (horizontal). The floor of the plot (representing motor activation) shows both sung pitch based on mapping (vertical lines going down) as well as lines representing correctly sung pitches. Labels t1-t4 higlight the timing of individual notes in the squence. Panel B: The same pitch trajectory now parameterized by time, showing multi-modal matching as a two-dimensional state space. The *x*-axis here comes from depth in panel A, and the *y*-axis is height in panel A.

notes are integrated with words (Mantell & Pfordresher, 2013; Pfordresher & Mantell, 2014).

A simple approach to addressing time would be to extend the framework shown in Figure 2 discretely to a series of sung tones. However, musical sequences in practice are not discrete, but include fine-grained variability in pitch and time. Moreover, the fact that poor singers perform better on melodies than in pitch-matching tasks suggests that multi-modal mapping is not performed independently at each time point and may be sensitive to relational information and not simply pitch height. Figure 3A illustrates how the multi-modal representation may guide mapping in real time (data come from one trial in Pfordresher & Mantell, 2014). This plot shows how the mapping between an auditory image of pitch (vertical axis) and motor planning for vocal fold tension (implied depth) evolve over time (horizontal). Lines connecting the back "wall" of the plot to the "floor" of the plot show sensorimotor mapping evolving over time, with asterisks showing specific connection points. The closeness of these points gives the appearance of a solid line. In addition, the ideal performance (which matches values on the vertical axis) is shown on the "floor" of the plot. As can be seen, this performance is generally "sharp," suggesting an upwards bias in sensorimotor mapping.

The evolution of a system like this over time is often represented as a "state space," which shows the relationships between two dimensions of a system with time as the parameter (i.e., collapsing across the horizontal axis). Figure 3B shows such a state space for the data shown in Figure 3A. A representation like this shows more clearly how two system dimensions (here auditory versus motor imagery) relate to each other over time. Specifically, Figure 3B shows where more bias in the mapping emerges, with bias occurring when intersections of X and Y diverge from the major diagonal.

An important point to consider is that the kind of trajectory shown in Figure 3A is not contingent on the presence of production, or even of perception. In particular, the motor-based associations shown here are presumed to result from mental images that are triggered by auditory imagery, and that may also exist during initial exposure to a target stimulus that the participant imitates later.

The somewhat puzzling result that poor singers do better when singing multiple pitches in a melody, as opposed to matching the same pitch repeatedly, may be based on the amount of information in the state space. Consider a typical task, in which the participant hears a short melody (4-6 notes) and then immediately repeats that melody. According to the MMIA model, the participant generates a multi-modal image associated with that melody while listening to it. That image is retained in memory, and then used to guide production during recall. When the sequence consists of a single pitch (or a single pitch repeated several times), the state space in memory will include a minimal amount of information, most likely a cluster of closely associated points. By contrast, the stored state space for a melody may be richer in content, such as the state space shown in Figure 3B. Possibly it is this richer store of information that leads to the advantage found for short melodies with variable pitch as opposed to single-pitch mapping for VPID singers. The reverse effect found for accurate singers (higher error for melodies with more pitch changes) may simply reflect adjustments in motor states as opposed to challenges in sensorimotor mapping.

Relationship to the Internal Models Construct

Some may wonder if the proposed imagery model marks a break from recent proposals that have accounted for VPID as an internal models deficit (Pfordresher, 2011;

Pfordresher & Halpern, 2013; Pfordresher & Mantell, 2014). This is not the case. Rather, we view the MMIA model as a proposal for how the abstract notion of an internal model may be implemented within the auditory-vocal system. We view the present model as primarily an elaboration of the internal models construct, driven by what we understand from recent results.

One deviation in our model from the internal models framework is that the present model synthesizes the inverse and forward model components. These components are typically represented as independent (e.g., Kawato, 1999). However, in the MMIA model both components simply represent two directions in which information may flow. Multi-modal associations representative of an inverse model occur when a pitch percept (*y*-axis) is mapped to some level of laryngeal tension (*x*-axis). Conversely, forward modeling associations occur when a motor plan relating to the laryngeal muscles is mapped to a pitch height value.

The separation of inverse and forward model components in other models is based in large part on evidence for neural separation of these functions (Kawato, 1999). Nevertheless, it is not known at present whether multi-modal associations are bidirectional in an integrated representation (as implied by the present model), or separated, as in previous internal model frameworks. Ultimately it will be up to future research to determine whether mental images and internal models are one and the same, or whether one leads to the other.

Development of Multi-Modal Associations

A major point of concern in the literature on singing has to do with development. The first, most groundbreaking research on VPID came from music education, deriving in large part from the formidable practical issue of how to handle such individuals in early music education (Welch, 1979). Thus, any model of vocal pitch imitation ought to offer some account for the development of pitch imitation ability, and how such abilities may thrive in some yet founder in others. According to the MMIA model, two factors are critical: (1) integration of modalities, (2) generalization of the sensorimotor map to novel overt vocalizations.

The first factor, integration of modalities, is what allows the internal mapping to represent joint relationships between coordinates, which is central to the representation shown in all Figures. Without such integration, the entire space is equiprobable, and no reliable mapping is possible. It is doubtful that such extreme cases occur in typically developing individuals. However, it is highly likely that this ability varies across individuals, with some individuals having an integration ability closer to the ideal map in Figure 1, and others with a noisier mapping function as shown in Figure 2A. We propose that modality integration is the primary factor in development, and that without it, the prospects for developing generalization ability in sensorimotor mapping is dubious.

The second factor, generalization, is what allows one to have a sensorimotor map that is not limited by past sensorimotor-associations. Consider a highly simplified scenario in which an individual is able associate modalities with some degree of (imperfect) accuracy, but has no ability to generalize. In such a case, the mapping relationship between modalities will be dominated by those associations that are most prominent in the individual's experience. This would lead to biasing effects such as regression to a poor-pitch singer's comfort pitch, mentioned before (and simulated in Figure 2B).

But where does the mapping come from in the first place? We suggest that the building blocks for multi-modal associations come from a simple source: learning by association. In the earliest stages of development, infants associate self-vocalizations with auditory feedback from these vocalizations through cooing and babbling (Figure 4A, cf. Guenther, 1995). Over time these associations form a constellation, taking the shape of a probability distribution. Given limitations in vocal range, this distribution will be limited in its range, as illustrated in Figure 4B. Note that the formation of these associations is in itself a critical, yet basic process. It is plausible that some individuals will have difficulty encoding and retaining these overt associations. Such individuals will fail to develop the first factor in the formation of multi-modal associations in mental imagery. Moreover, even those who successfully retain overt associations in memory may fail to generalize.

The process of generalization is illustrated in Figure 4C. Critical to this process is the ability to extrapolate the overt associations one experiences through self-vocalizations to other possible associations. This process is akin to the classic Piagetian framework of schema building in development. Ideally, over time, the abstract schematic mapping across modalities will come to dominate over memory for experienced mapping (Figures 4A-B). A possible mechanism for this kind of generalization is Bayesian inference, which has been introduced recently in models of motor learning (e.g., Wolpert, Diedrichsen, & Flanagan, 2011).

Simulations of a Preliminary Computational Framework

In order to assess whether the MMIA model could mimic characteristics of VPID in the literature we

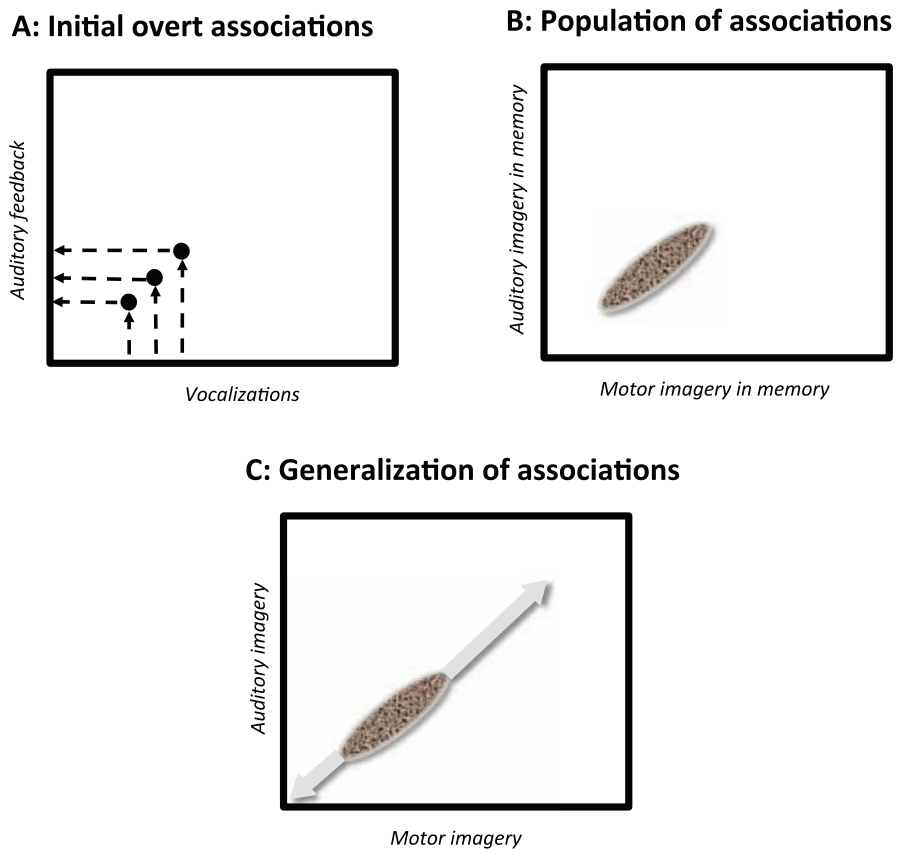

FIGURE 4. Developmental stages in the successful formation of multi-modal associations.

performed the following simulations, using equations 1 and 2. We used as input to the model a four-note melody that is one of the melodies often used in our lab [C D E G] coded in cents as [0 200 400 700]. Pitches were coded as cents relative to the low G. Values of σ^2_{map} and σ^2_{bias} were varied across ranges that according to preliminary runs led to simulations that approximated singers' behavior. Ten four-note "trials" were run for each variance combination. For each input note, an output note was selected at random from the probability distribution associated with the input. The pitch value (in cents) of that output was interpreted to represent a possible sung pitch. In addition, we varied the "comfort pitch" of the model x_{bias} in increments of 200 cents from -400 (A-flat below C) to +400 cents (E above C). Combining these variables (trials, notes within trials, two variance sources, comfort pitch) led to 14,000 simulated "notes."

First, we were interested in how the two variance sources influenced the mean of absolute differences between target and sung pitch (called "mean absolute note error"). This is a common way currently to evaluate singing accuracy for tasks that involve a fixed pitch referent. Figure 5 shows the mean absolute note error from simulations as a function of each variance source (represented as standard deviations to contain

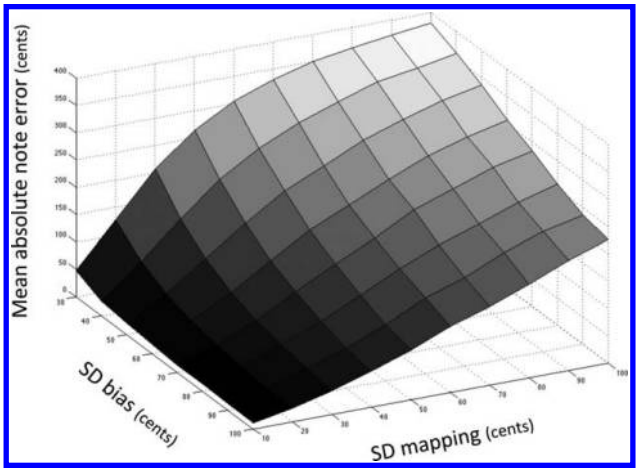

FIGURE 5. MMIA model simulations predicting mean absolute note error from two sources of variance. See text for further details.

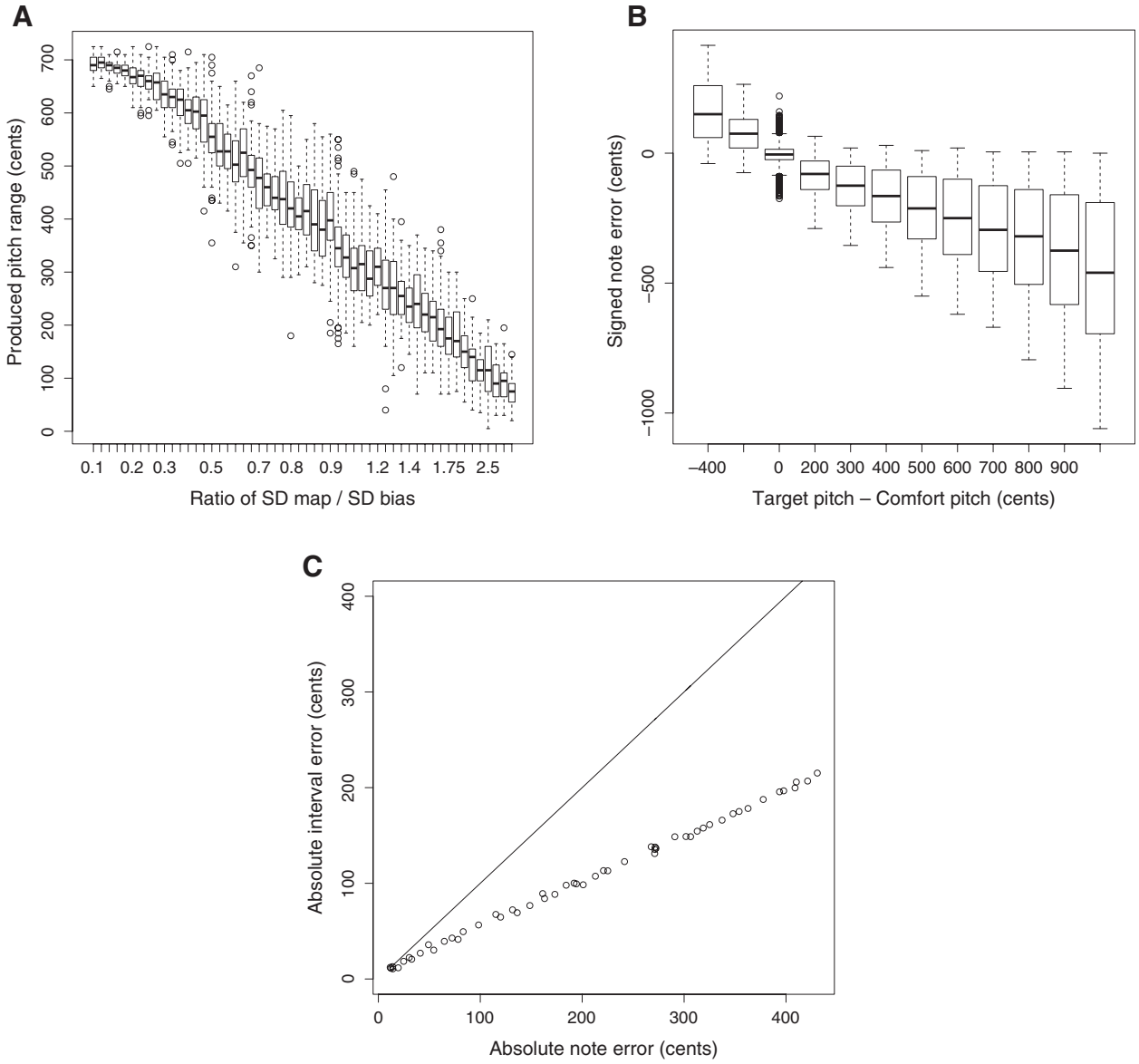

FIGURE 6. MMIA model simulations of three important behavioral effects: compressed pitch range for poor-pitch singers (A), drift toward comfort pitch (B), and differing sizes for note versus interval errors (C). See text for further details.

the range of values), averaged across comfort pitch, trials, and notes within trials. As can be seen, both variance sources contribute to mean absolute note error according to a monotonic, but nonlinear, relationship. Importantly, each variance source yields an opposite relationship with mean absolute note error. Whereas increasing σ^2_{map} leads to an increase in error, increasing σ^2_{bias} has the opposite effect. This reflects what we mentioned before, that increasing each variance source diminishes the influence of that component.

Figure 6 shows three further analyses in which we tested whether the model predicts basic features of VPID from the literature. Figure 6A addresses compression of pitch range, mentioned earlier. Ordinate values represent difference scores of the highest versus the lowest "sung" pitch in each trial (the ideal value is 700 cents based on the stimulus), whereas the abscissa is the ratio formed by each variance source, with higher values reflecting a stronger influence of bias. The boxplots in this figure show the mean and interquartile range in the

center rectangle, with whiskers showing the range of values that extend up to 1.5 times the interquartile range (circles are outliers beyond the whiskers). As can be seen, pitch range decreases as the mapping relationship is increasingly dominated by the biasing effect of the comfort pitch, as has been documented elsewhere. Furthermore, it is worth noting that the dominant tendency of the system in general is toward compression, as is seen in human performance.

Simulations of a second hallmark of VPID are shown in Figure 6B, averaged across ratios of σ map to σ bias. Here, ordinate values are difference scores between sung and target pitches, with positive values indicating "sharp" performance and negative values indicating "flat." The abscissa shows difference scores between target pitches and the comfort pitch of the model fit (x_{bias}). The negative relationship between Y and X values replicates the attracting influence of comfort pitch on sung pitch errors reported by Pfordresher and Brown (2007, cf. Hutchins, Larrouy-Maestri & Peretz, 2014). Namely, when participants attempt to sing a pitch that is higher than their comfort pitch, errors drift in the flat direction, toward their comfort pitch, and vice versa when they attempt to sing a pitch that is lower than their comfort pitch. Note that the preponderance of positive signed values on the *x*-axis reflects differences between the range of pitches in the target melody (0 to 700 cents) and the range of comfort pitches simulated (-600 to 600 cents).

Figure 6C illustrates a test of whether the model simulates the fact that errors based on comparisons between single pitches tend to be larger than errors based on relative pitch (Berkowska & Dalla Bella, 2013; Vurma & Ross, 2006). For this analysis, we computed mean absolute errors of sung pitch intervals as described elsewhere (e.g., Berkowska & Dalla Bella, 2013; Pfordresher and Brown, 2007). For this scatterplot, the parameter is the ratio shown in figure 6A (*x*-axis), with each point reflecting the average across trials, notes within trials, and comfort pitches. As the size of absolute note error values increase, so does interval error. However, the slope is less than 1 (compare the relationship to the solid line). Thus, as seen in the human data, the model predicts smaller interval errors than note errors. This is particularly important in that the model at present does not directly code interval information and proceeds note by note.

Conclusions and Future Directions

In this article, we have argued for a possible role of mental imagery in singing, detailed a possible mechanism for multi-modal imagery associations, proposed a developmental trajectory for this mechanism, and presented simulations of a preliminary computational framework for this mechanism. We consider this direction fruitful from several respects. In linking vocal imitation of pitch to imagery, one can capitalize on the rich literature concerning mental imagery in testing sources of VPID. Moreover, although the processes modeled here serve the same function as an internal model, the framework proposed here is simpler in architecture (involving only one component) and can be tested using known experimental manipulations. Further research on imagery can give us insight into how the internal model may represent information and lastly, the multi-modal nature of imagery can be informative in understanding activities that integrate perception and action, such as singing.

Future directions for the model will involve further testing and refining of the model's existing constraints, and (as needed) exploration of additional factors. The model in its present form can be compared directly to individual data. Preliminary results are promising, though some important questions remain regarding the bias parameter. Specifically, the bias parameter can be based on explicit measurements of "comfort pitch," or can be adopted as a free parameter based on the difficulty of assessing comfort pitch objectively. Likewise, variance measures could be manipulated as free parameters or based on production data (cf. Pfordresher et al., 2010, for measures of precision). We will detail these efforts in a future paper.

The integrative and bidirectional nature of the multimodal mapping proposed in MMIA likewise has interesting practical implications. Whereas in this paper we focused on mapping from auditory to motor imagery, based on the kind of tasks we wished to simulate, the model predicts that associations in the opposite direction abide by the same mapping. Thus, the model predicts that effective use of auditory feedback may be based on the ability to map a planned motor image onto an auditory image (akin to the function of a forward model, as mentioned before). Second, our model suggests that imagery tasks may be useful as a way of improving vocal imitation skills. Although we tested this model using data from overt motor tasks, the fact that the model is ultimately based on internal images implies that imagery may facilitate mapping on its own, following the initial sensorimotor-based period of learning (as described in the section on development).

The preliminary assumption that the bias parameter is a fixed value for a participant is a simplification that we plan to test through further simulations and experimentation. It may be the case that bias is influenced by conditions, such as the dominant key presented in an experiment, for instance. Likewise, levels of anxiety brought about by an experiment may or may not

influence the source of bias exhibited by a participant. A particularly interesting and complex matter for modeling has to do with the aforementioned tendency for VPID participants to perseverate recently produced pitch patterns (Wisniewski et al., 2013). It is possible that such perseveratory behavior falls out of the pervasive influence of "comfort pitch" on all performances. However, it is also possible to reformulate the model so that bias is influenced by past sensorimotor associations that take a certain amount of time to decay and thus lead to perseverations (cf. Dell, Burger, & Svec, 1997).

What we propose here is preliminary, and involves just one part of a more complicated sensorimotor architecture used for singing (see introduction in the present volume). There are some behaviors that the present model is not designed to model, such as the ability to make conscious decisions about pitch changes. The fact that MMIA involves a mechanism that is distinct from this ability is consistent with dissociations between discrimination and imitation tasks reported elsewhere (e.g., Loui, Guenther, Mathys, & Schlaug, 2008). Similarly, this model does not account for storage of sequences in long-term memory, although it does propose an account for short-term storage of a multi-modal sequence. This model also does not account for the ability to imitate text, although a similar mapping mechanism could be developed toward that purpose. Ultimately we see the MMIA model approach introduced here as an approach with the potential to account for a rich set of vocal imitation behaviors within a simple framework, and thus an important step in our understanding of VPID.

Author Note

This research was supported in part by NSF grant BCS-1256964. We thank Steven Demorest, Psyche Loui and Graham Welch for helpful comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Peter Q. Pfordresher, Department of Psychology, 362 Park Hall, University at Buffalo, Buffalo, NY 14260. E-mail: pqp@buffalo.edu

References

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

Berkowska, M., & Dalla Bella, S. (2013). Uncovering phenotypes of poor-pitch singing: The Sung Performance Battery (SPB). *Frontiers in Psychology, 4*, 714.

Buxbaum, L. J., Johnson-Frey, S. H., & Bartlett-Williams, M. (2005). Deficient internal models for planning hand-object interactions in apraxia. *Neuropsychologia, 43*, 917-929.

Dalla Bella, S., Berkowska, M., & Sowinski, J. (2011). Disorders of pitch production in tone deafness. *Frontiers in Psychology, 2*, 164.

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

Dell, G. S., Burger, L. K., & Svec, W. R. (1997). Language production and serial order: A functional analysis and a model. *Psychological Review, 104*, 123-147.

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review, 102*, 594-621.

Hommel, B. (2009). Action control according to TEC (theory of event coding). *Psychological Research, 73*, 512-526.

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences, 24*, 849-937.

Hutchins, S., Larrouy-Maestri, P., & Peretz, I. (2014). Singing ability is rooted in vocal-motor control of pitch. *Attention, Perception and Psychophysics, 76,* 2522-2530.

Hutchins, S., & Moreno, S. (2013). The Linked Dual Representation model of vocal perception and production. *Frontiers in Psychology, 4*, 825.

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.

Hutchins, S., Zarate, J. M., Zatorre, R. J., & Peretz, I. (2010). An acoustical study of vocal pitch matching in congenital amusia. *Journal of the Acoustical Society of America, 127*, 504-512.

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

Keller, P. E. (2012). Mental imagery in music performance: Underlying mechanisms and potential benefits. *Annals of the New York Academy of Sciences, 1252*, 206-213.

Kelso, J. A. S. (1995). *Dynamical patterns: The self-organization of brain and behavior.* Cambridge, MA: MIT Press.

Leaver, A. M., Van Lare, J., Zielinski, B., Halpern, A. R., & Rauschecker, J. (2009). Brain activation during anticipation of sound sequences. *Journal of Neuroscience, 29*, 2477-2485.

Levitin, D. J. (1999). Tone deafness: Failures of musical anticipation and self-reference. *International Journal of Computing and Anticipatory Systems, 4*, 243-254.

Liu, F. Jiang, C., Pfordresher, P. Q., Mantell, J. T. Xu, X., Yang, Y., & Stewart, L. (2013). Individuals with congenital amusia imitate pitches more accurately in singing than in speaking: Implications for music and language processing. *Attention, Perception and Psychophysics, 75*, 1783-1798.

Loui, P., Guenther, F. H., Mathys, C., & Schlaug, G. (2008). Action-perception mismatch in tone-deafness. *Current Biology, 18*, R331-332.

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

Pfordresher, P. Q. (2011). Poor pitch singing as an inverse model deficit: Imitation and estimation. In A. Williamon, D. Edwards & L. Bartel (Eds.), *Proceedings of the International Symposium on Performance Science* (pp. 539-544). Utrecht, the Netherlands: Association Européenne des Conservatories.

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., Meier, K., Belyk, M., & Liotti, M. (2010). Imprecise singing is widespread. *Journal of the Acoustical Society of America, 128*, 2182-2190.

Pfordresher, P. Q., & Halpern, A. R. (2013). Auditory imagery and the poor-pitch singer. *Psychonomic Bulletin and Review, 20*, 747-753.

Pfordresher, P. Q., & Mantell, J. T. (2014). Singing with yourself: Evidence for an inverse modeling account of poor-pitch singing. *Cognitive Psychology, 70,* 31-57.

Prinz, W., Aschersleben, G., & Koch, I. (2009). Cognition and action. In E. Morsella, J. Bargh & P. M. Gollwitzer (Eds.), *Oxford handbook of human action* (pp. 35-71). Oxford, UK: Oxford University Press.

Roubeau, B., Chevrie-Muller, C., & St. Guily, J. L. (1997). Electromyographic activity of strap and cricothyroid muscles in pitch change. *Acta Otolaryngologica, 117*, 459-464.

Sergeant, D. (1994). Towards a specificaion for poor pitch singing. In G. F. Welch & T. Murao (Eds), *Onchi and singing development* (pp. 63-73). London, UK: David Fulton.

Smith, J. D., Reisberg, D., & Wilson, M. (1992). Subvocalization and auditory imagery: Interactions between the inner ear and inner voice. In D. Reisberg (Ed.), *Auditory imagery* (pp. 95-119). Hilsdale, NJ: Earlbaum.

Sundberg, J. (1989). *The science of the singing voice*. DeKalb, IL: Northern Illinois University Press.

Vurma, A., & Ross, J. (2006). Production and perception of musical intervals. *Music Perception, 23*, 331-344.

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

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

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

Wolpert, D. M., Diedrichsen, J. & Flanagan, J. R. (2011). Principles of sensorimotor learning. *Nature Reviews Neuroscience, 12*, 739-751.

Zarate, J. M. (2013). The neural control of singing. *Frontiers in Human Neuroscience, 7*, 237.

Zatorre, R. J., & Halpern, A. R. (2005). Mental concerts: Musical imagery and auditory cortex. *Neuron, 47*, 9-12.