Cognition 127 (2013) 177–202

Contents lists available at SciVerse ScienceDirect

Cognition

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

Vocal imitation of song and speech

James T. Mantell *, Peter Q. Pfordresher

Department of Psychology, University at Buffalo, The State University of New York, United States

ARTICLE INFO

Article history:
Received 5 November 2010
Revised 7 December 2012
Accepted 21 December 2012
Available online 1 March 2013

Keywords:
Vocal imitation
Song
Speech
Modularity
Encapsulation
Domain specificity

ABSTRACT

We report four experiments that explored the cognitive bases of vocal imitation. Specifically, we investigated the accuracy with which normal individuals vocally imitated the pitch-time trajectories of spoken sentences and sung melodies, presented in their original form and with phonetic information removed. Overall, participants imitated melodies more accurately than sentences with respect to absolute pitch but not with respect to relative pitch or timing (overall duration). Notably, the presence of phonetic information facilitated imitation of both melodies and speech. Analyses of individual differences across studies suggested that the accuracy of imitating song predicts accuracy of imitating speech. Overall, these results do not accord with accounts of modular pitch processing that emphasize information encapsulation.

1. Introduction

Speech and song are forms of vocal communication. Each of these behaviors requires the coordinated use of the respiratory system, the larynx, and the vocal tract to provide variation in vocal intensity, pitch, and phonetic variation (Sundberg, 1999; Welch, 2005). In this context, it is not surprising that the distinction between speech and song is often blurred in practice, as in German sprechgesang and sprechstimme (sung speech and rhythmically heightened speech, respectively, which are utilized in certain operatic performances), and in the Japanese narrative forms of Nohgaki and Shinnai (Feld & Fox, 1994; List, 1963; Welch, 2005). Further, there is evidence to suggest that the perceptual identification of a vocal sequence as speech or song is plastic. Deutsch, Henthorn, and Lapidis (2011; see also Deutsch, Lapidis, & Henthorn, 2008; Falk & Rathcke, 2010) recently found that repeatedly presenting a spoken phrase causes that phrase to sound more like song; this suggests that context can influence the identification of a vocal sequence as speech or song. Yet, there are ways in which speech and song differ. For example, speech is a form of linguistic communication but song can serve as linguistic and/or musical communication. In everyday life, situational context underscores the distinction between speech and song. Individuals use speech when conversing but song is reserved for special occasions including celebration events, religious activities, and some social interactions (e.g., with young children). Some researchers have argued for shared processing of music and language (Koelsch, 2011; Patel, 2008; Sammler et al., 2009), some have emphasized that these modalities simultaneously present shared and distinct characteristics (Jackendoff, 2009; Jackendoff & Lerdahl, 2006), and some have suggested that music and language processing occur in separate cognitive modules (Peretz & Coltheart, 2003).

1.1. Modularity and music

The concept of modularity has been vigorously debated by scientists and philosophers since Fodor's (1983) landmark publication. In his essay, Fodor argued that mental input systems could be described as modules based on their possession of most or all of nine properties. For Fodor (1983, 2000), the single most important of these

* Corresponding author. Address: 207A Park Hall, North Campus, University at Buffalo, The State University of New York, Buffalo, NY 14226, United States. Tel.: +1 716 645 0225.
E-mail address: jtm29@buffalo.edu (J.T. Mantell).

0010-0277/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cognition.2012.12.008

characteristics is information encapsulation, the notion that not all information available to an organism informs operation of a modular system. Information encapsulation can be clearly imagined via a flowchart: when a researcher draws boxes to distinguish components of a processing system, it becomes clear that "only the inputs and outputs of functionally individuated systems can mediate their information exchanges" (Fodor, 1983, p. 87). Fodorian modularity gained early support by researchers. For example, Peretz and Morais (1989) argued that tonal encoding of pitch is accomplished by a cognitive processor that meets several of Fodor's modularity properties, including domain specificity (processing applies only to music), automaticity (operation is mandatory, given the input), and information encapsulation. However, several researchers (e.g., Pinker, 1997; Tooby & Cosmides, 1992, p. 113) explored the possibility that most or all of our mental faculties are evolutionarily adapted, domain specific, information processing modules; in so doing these researchers sought to expand the notion of modularity in ways that Fodor (1983) suggested were untenable. This approach, still under development today, is known as 'massive modularity' (Carruthers, 2006b).

Other researchers have eschewed Fodor's primary criterion, information encapsulation, in favor of another of Fodor's modularity characteristics, domain specificity. Coltheart (1999) proposed that a processing system is modular if it responds only to a particular class of stimuli (i.e., it is domain specific). However, Fodor (2000) rejected Coltheart's (1999) definition of modularity based on domain specificity (p. 113). For Fodor (1983), information encapsulation is "perhaps the most important aspect" (p. 37), "the essence" (p. 71), and "the key" (p. 98) to modularity. Other massive modularity theorists have dismissed the primacy of information encapsulation (Barrett & Kurzban, 2006, pp. 631–633; Carruthers, 2006a, pp. 12, 57–59). Barrett and Kurzban (2006) proposed a broad modularity based on functional specialization; their approach blends formal computationalism and evolutionary psychology. The authors assert that "Only information of certain types or formats will be processable by a specialized system... domain specificity is a necessary consequence of functional specialization" (p. 630).

But there is a problem with a modularity based only on domain specificity, and several researchers have recognized it (Besson & Schön, 2011; Fodor, 1983, 2000; Gibbs & Van Orden, 2010; Prinz, 2006). The problem is that declaring domain specificity as the essential quality of modularity trivializes the concept. In other words, a modularity based on specificity of input does not say anything useful about what modules do (see Fodor, 2000, p. 113; Prinz, 2006, p. 34). Instead, it posits a single characteristic as the definition of modularity and then points as "evidence" to the abundant cognitive systems that conform to this property. In line with Prinz's (2006) critique, Barrett and Kurzban appear to tacitly accept that most or all of the systems in the brain are modular (p. 630), writing "...whether an information-processing system "is or is not" modular is not useful. There is little doubt that different kinds of information are handled by different systems in the brain." This is probably what Fodor (1983, 2000) had in mind when he rejected domain specificity as the primary characteristic of a module. Today, modularity as a concept and a term continues to be debated (e.g., see the discussion between Carruthers, 2008 and Cowie, 2008; Machery, 2008; and Wilson, 2008), and it is clearly the case that neither massive modularity nor Fodorian modularity has been accepted by all researchers (Robbins, 2010).

The concept of cognitive modularity has not been decisively defined but there is considerable agreement that the specific information processing components that characterize modular processes must be information encapsulated, domain specific, or both. Thus, we have framed the empirical discussion within this paper around these two information processes. It is our hope that expanding knowledge of these two information processing characteristics will contribute to the debate on modularity in the cognitive processing of language and music. One modular model is particularly relevant to the current research because it makes empirical predictions about the performance and processing overlap between language and music. Peretz and Coltheart (2003) proposed a modular model of music processing based primarily on case studies of individuals with brain damage who together represent doubly dissociated music and language deficits. In their model, information from an initial acoustic analysis module is sent to specialized pitch, time, and speech modules. Separate modules facilitate the analysis of pitch, and of these distinct processors, one in particular—tonal encoding—is domain specific because it only accepts musical pitch information and likewise encapsulated to speech because phonological information cannot enter the module to influence pitch processing.[1] If a tonal encoding module exists as depicted in the model, it should handle tonality processing without access to phonological or linguistic information. Tonality is an informational property of music and not language; it is what determines why a single tone may sound good in one musical context and terrible in another (Krumhansl & Kessler, 1982). According to Patel (2008, p. 201), "At present there is no evidence of anything resembling scales or pitch hierarchies in speech melodies."

Evidence on the domain specificity and encapsulation of speech and song processing is mixed. Recent imaging research revealing substantial overlap in brain activations associated with speaking and singing (Callan et al., 2006; Saito, Ishii, Yagi, Tatsumi, & Mizusawa, 2006; Schön et al., 2010; Özdemir, Norton, & Schlaug, 2006) suggests that vocal processing may not be domain specific. However, these studies have also revealed non-overlapping areas with some exclusively right hemispheric activation for song tasks, indicating that there is something special about song. Moreover, the link between neural activations and modules is not entirely clear in part due to the fact that current imaging technology may not be capable of revealing the fine detail of adjacent neural networks (Peretz,

[1] It is conceivable that phonetic information could influence pitch processing, or that pitch information could influence phonetic processing, but neither of these possibilities is represented in the model in its current form. This is likely because no neuropsychology data has been collected to support such claims.

2009). Peretz and Coltheart (2003) and others (for reviews see Marin & Perry, 1999; Peretz & Zatorre, 2005) have suggested that the observed dissociations between music and language processing support music modularity, based in large part on evidence from perception tasks.

1.2. Imitation

Our focus shifts the emphasis from perception to imitative production. We address the accuracy with which people can vocally imitate pitch patterns in sung melodies and spoken sentences. The ability to vocally imitate musical pitch is of critical importance to musical communication. Most individuals in Western cultures believe they are deficient in imitating musical pitch patterns by singing (Pfordresher & Brown, 2007) but in reality, only a minority of individuals are incapable of imitating a sung melody within a semitone (typically 10–20% of a given sample, Pfordresher & Brown, 2007; see also Dalla Bella, Giguère, & Peretz, 2007; Pfordresher & Brown, 2009; Pfordresher, Brown, Meier, Belyk, & Liotti, 2010). The ability to vocally imitate musical pitch by singing may thus be considered a typical human trait.

Vocal imitation plays an important role in speech. For example, speech imitation is crucial for language acquisition (Kuhl, 2000; Kuhl & Meltzoff, 1996) and mature speakers utilize overt speech imitation for comedic or sarcastic effect (such as when telling a joke or a story). Previous research on vocal imitation of speech has focused on covert imitation tasks, such as the imitation of global syntactic properties (as in interactive alignment, Pickering & Garrod, 2004) and fine-grained phonetic imitation (Goldinger, 1998; Nielsen, 2005; Nielsen, 2007; Pardo, 2006; Shockley, Sabadini, & Fowler, 2004) during conversations or in shadowing tasks. Our perspective is on a different aspect of vocalization that is of direct importance both for song and speech: intentional imitation of pitch.

We consider how the ability to imitate pitch-time information is related to domain specificity and encapsulation. With respect to domain specificity, we compare the accuracy with which normal individuals (who are usually not trained singers) imitate pitch in a musical context with their ability to imitate the pitch of a spoken utterance. According to Peretz and Coltheart (2003), the music module contains processors (such as the tonal encoding processor) that are specialized for pitch information in music such as song. Because these processors are specialized for song, they should process song input but not speech input. The effect of specialized pitch processing for song targets should be enhanced imitation accuracy for song pitch that may lead to dissociated individual differences in the accuracy of imitation across the domains of speech and song, as found for perceptual deficits characteristic of presumed modular processing (e.g., Ayotte, Peretz, & Hyde, 2002). With respect to encapsulation, we predict that phonetic information, clearly in the linguistic domain, should not benefit pitch processing in song. If phonetic information facilitates song pitch processing, then the pitch processors may not be encapsulated to speech information. To test this hypothesis, we varied whether or not pitch trajectories from song and speech were combined with phonetic information. Specifically, participants imitated song and speech both in their original, worded forms as well as synthesized versions that included only pitch and time information.

It is possible that domain type can mediate the degree to which phonetic information influences imitation. The link between pitch-time trajectory and segmental information is arguably less flexible in speech than in song. After all, song can be produced without lyrics, but segmental phonetic information is the definitive characteristic of speech. Additionally, songs routinely vary the match between text and pitch, for instance by setting many different verses of text to the same melodic line. Given these generalizations, one might expect that the imitation of pitch-time trajectories from speech may be more dependent on phonetic information than the imitation of pitch-time trajectories from song. Overall, some research has indicated that melody and words are integrated in memory (Serafine, Crowder, & Repp, 1984; Serafine, Davidson, Crowder, & Repp, 1986; Wallace, 1994) but others have suggested that the relationship between lyrics and melody is not integrative but associative (i.e., speech and song are represented independently but can be readily associated via learning. See Ginsborg & Sloboda, 2007; Racette & Peretz, 2007). One study of singing showed that fine-grained timing of production reflects independent contributions of prosody and meter (Palmer & Kelly, 1992). Another recent study showed that production of pitch in folk songs was produced less accurately when notes were sung with words than on the syllable /la/ (Berkowska & Dalla Bella, 2009; however, for an opposite finding, see Racette, Bard, & Peretz, 2006, Experiment 1). Berkowska and Dalla Bella's finding accords with the claim that words and melody are represented separately and that combining them during production reduces performance accuracy (see also Racette & Peretz, 2007).

Research reported here addressed the performance of speech and song in the context of an intentional imitation paradigm: participants listen to a stimulus (the *target*) and then attempt to reproduce it as accurately as possible (the *imitation*). Targets were based on spoken sentences that were then transcribed into melodies with the same word content and global pitch contour (the overall pattern of upwards versus downwards pitch change over time). From these targets, we created "wordless" versions that lacked phonetic information by synthesizing the pitch-time trajectories from the worded versions as complex waveforms with resonances similar to that of the human voice. Although the synthesized pitch-time trajectories extracted from speech are not technically speech, we refer to them as wordless speech for brevity; the key point is that pitch-time information was the same between worded and wordless targets.

In addition, we introduce new measures of pitch imitation based on the accuracy of imitation across the entire trajectory. These measures are sensitive to imitation of pitch fluctuations within canonical rhythmic units, such as notes (for song) or syllables (for speech), and across the sequence. By contrast more traditional measures of pitch imitation (e.g., Dalla Bella, Giguère, & Peretz, 2009; Dalla Bella et al., 2007; Pfordresher & Brown, 2007, 2009; Pfordresher et al., 2010) extract a single point estimate

from each rhythmic unit, thereby treating pitch information within the unit as homogenous. Researchers have occasionally applied such simplifications for the speech signal by using the Prosogram (Mertens, 2004), which reduces pitch variability in speech and transforms F0 within syllables to either steady states or glides. Such simplifications are predicated on the autosegmental theory of prosody perception (Pierrehumbert, 1980/87) and are thus useful in studies that aim to understand the perception of music and language, such as the perception of tonal analogues for speech (e.g., Patel, Peretz, Tramo, & Labreque, 1998) or the use of pitch to convey emotion to the listener (e.g., Curtis & Bharucha, 2010). However, we suggest that such procedures oversimplify the signal for the purpose of assessing vocal imitation of pitch trajectories. Successful imitation involves tracking F0 within and also across rhythmic units. This is particularly important for speech, for which fluctuations in F0 can occur within a syllable, but can also be true of music for which a singer may "scoop" or use vibrato when sustaining a 'single' pitch. As such we focus on imitation of F0 across the entire trajectory for speech and song, and compare results from this analysis with other analyses that adopt more traditional techniques.

We report the results of four experiments that were designed to address the relative contributions of sequence type (song/speech) and phonetic information (worded/wordless) on vocal imitation of pitch and timing. Experiment 1 serves as a baseline for the other experiments; participants simply imitated the sequences as they heard them. Other experiments were designed to further explore two critical results of Experiment 1. Experiment 2 was designed to address why phonetic information facilitates imitation of pitch (as found in Experiment 1). In it, participants imitated all sequences using the neutral vowel *"ah"* [ɑ]. Experiments 3 and 4 were designed to address why the imitation of absolute pitch may be facilitated for songs as opposed to speech, focusing on temporal properties of speech versus music. Following our report of these experiments, we report individual differences analyses that result from pooling the data across all experiments, each of which included an independent sample of participants ($N = 148$).

2. General methods

2.1. Apparatus

For each of the reported experiments, vocal recordings were obtained in a sound attenuated booth (Whisper Room Inc., SE 2000 Series, Morristown, TN). Participants were seated on a stool and were instructed to maintain an upright posture during the recording session. Participants heard target stimuli and auditory feedback over Sennheiser HD 280 Pro headphones at a comfortable listening volume. Recordings were collected at a sampling frequency of 22050 Hz via a Shure PG58 microphone connected to a Lexicon Omega preamp and digitally stored as .wav files for future analysis.

2.2. Stimuli

The initial set of target stimuli for Experiments 1 and 2 were created by crossing the critical factors domain (speech versus song) and phonetic information (worded versus wordless) with the additional factors contour shape (statement versus question, intended to provide contour variation in the sequences), gender (male or female), and text (12 seed sentences, see Appendix A), yielding 192 targets (96 per gender). In Experiments 3 and 4 we modified the timing of these stimuli, as will be described later.

The initial step in stimulus construction was to record utterances of the 12 seed sentences spoken as statements or questions. These sentences were three to five syllables each and constructed to be relatively short and easily comprehensible. The male speaker (used for male participants) produced American English with a midland dialect and the female speaker (used for female participants) utilized an inland North dialect (see Labov et al., 1997, as cited in Wolfram & Schilling-Estes, 1998, p. 122). Statements were spoken with a falling contour and questions were spoken with a rising contour. Fig. 1 (top) shows the F0 trace from

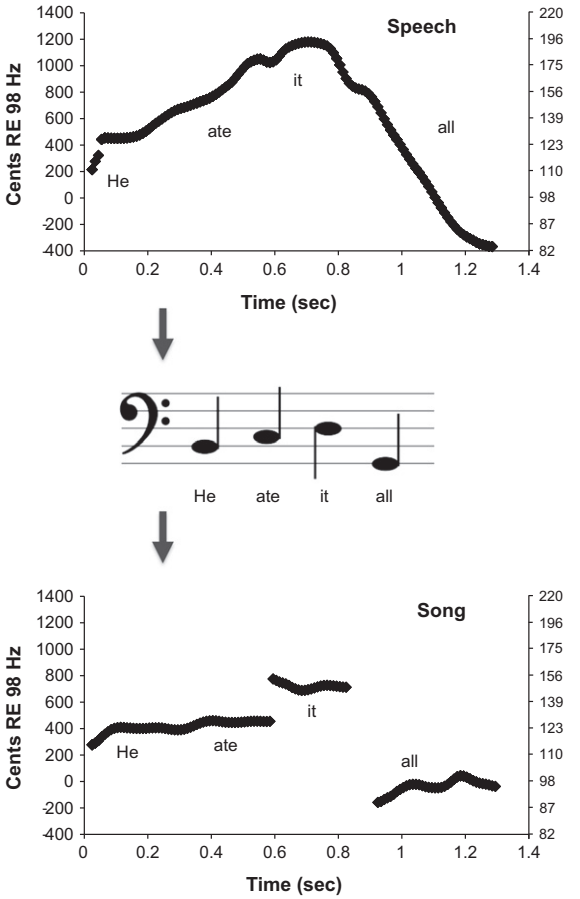

Fig. 1. F0 trajectory of the male spoken sentence, "He ate it all" (upper panel), the melodic transcription based on the spoken F0 contour (middle) and F0 trajectory for the recorded performance of the melody (lower). Primary y-axes show F0 in cents relative to a 98 Hz standard (G2).

a spoken statement by the male model speaker. Matched melodies were composed based on the pitch contour of syllables in the spoken sentences. First, the mean pitch values for each syllable in the sentences were used to assign approximated diatonic pitches, all from the key of G major. In order to elicit a sense of tonality, pitch classes were adjusted so that every melody featured the tonic and either the mediant (B) or dominant (D). The pitch contour (patterns of rising and falling across notes) of melodies matched the contour formed by successive syllables in the sentence. Fig. 1 (middle) shows notation from a matched melody as well as the F0 trace for the recorded performance of that melody (lower). The same two individuals who produced speech stimuli also recorded the melodic stimuli; each speaker had a moderate amount of vocal training and each were considered to be representative of accurate but not professional quality singers. In sum, the worded target stimuli are divided into two domains: speech and song. We operationally define these domains based on the intonation and timing properties of the stimuli. Speech targets were spoken naturally by the models and the pitch-time trajectory did not correspond to any diatonic scale. Song targets, on the other hand, were intoned, rhythmic stimuli. Each sung syllable received an approximately equivalent duration such that they invoked a metrical beat, and each note was intoned such that the sequence corresponded to a melody in the G major scale (see Fig. 1).

Finally, the speech and song stimuli were synthesized to create wordless versions. We used Praat (Boersma & Weenink, 2009) to extract the pitch-time trajectories and then transform them into "hums" that maintained the pitch-time information of the original sequences but did not contain any phonetic information. The hum sound includes five formants and is designed to mimic the timbre of a human voice. For descriptive statistics on the timing and pitch properties of target stimuli in Experiment 1, see Appendix B.

2.3. Procedure

Participants were seated in a sound-attenuated booth and instructed in good posture for vocalization. Next, participants performed several warm-up tasks, including reading a short passage of prose, singing 'happy birthday,' producing vocal sweeps, and vocalizing steady tone comfort pitches. The warm-ups helped to acclimate each participant to the recording environment.

Following the warm-ups, participants performed 96 vocal imitation trials. Each participant was assigned to one of two pseudorandom orders in which all experimental factors were intermingled. They were instructed to imitate the pitch of the target sequence to the best of their abilities. Male participants imitated the male-produced targets and female participants imitated the female-produced targets. Each trial began with the presentation of a target followed immediately by a short noise burst that served to cue the participant to begin his or her imitation of the target. After completing the imitation trials, participants were asked to complete questionnaires related to their musical background, cognitive abilities, language, and hearing sensitivity. Each experiment session lasted about 50 min.

2.4. Data analysis

Initial processing of recordings involved extracting vectors of F0 values and eliminating creaky voice data (see individual experiments for number of participants removed). Vocal creak is caused by loose contact of the vocal folds and results in decreased amplitude of F0 (Johnson, 2003, p. 136). All pitch data were converted from Hz to cents (100 cents = 1 semitone).

Accuracy was assessed by comparing the F0 vectors of the imitations with the F0 vectors of the matching targets; these comparisons were performed with in-house Matlab scripts (The MathWorks, Inc., Natick, MA). First, matched pairs of targets and imitations were equated for duration by resampling and aligning the imitation vector to match the rate of the stimulus vector. This temporal transformation adjusted the total duration of the imitation to match the stimulus target and afforded a quantitatively efficient means of calculating temporal accuracy which we refer to as the *duration ratio*; the ratio of imitation duration to target duration (well-timed imitation = 1; slower imitation < 1; faster imitation > 1).

After the alignment phase, outliers from pitch extraction were adjusted. Outliers were defined as any data point in the imitation that occurred at least 600 cents (one half octave) above or below the corresponding time-matched data point in the stimulus. As a first step, we equated the mean pitch of both target and imitation sequences by subtracting the average target-imitation pitch vector difference from the imitation pitch vector. This first step was used to avoid having outlier identification biased by mistuning of the imitation. Next, we transposed the octave of these data points up or down to match the appropriate target octave. Finally, we undid the initial transformation by adding back the initial pitch vector differences. On average, less than 4% of the data samples within each trial (across all experiments) needed to be adjusted.

Pitch accuracy was based on the entire vector rather than by segmenting the imitation into notes or syllables in order to account for imitation of fine-grained temporal fluctuations in F0 (particularly for speech). Moreover, we decided that segmentation into syllables was not appropriate for the imitation of wordless speech targets because those imitations were not based on phonetically delineated syllables in the first place (they consisted entirely of pitch-time varying hums). We report two measures of pitch accuracy: *mean absolute pitch error* was the average absolute difference between the target and imitation vectors. Both flat and sharp errors contributed to the total error; this is a measure that technically is influenced by both accuracy and precision (Pfordresher et al., 2010). Most importantly, mean absolute pitch error indexes how well participants match the absolute pitch values of song and speech, whether they sing or speak "in tune." Fig. 2A shows an example of a target melody and one participant's imitation; the mean absolute pitch error would be computed based on the mean absolute difference between all co-occurring samples. The singer illustrated in Fig. 2A imitates notes 2–4 flat, leading to a mean absolute pitch error of 176 cents.

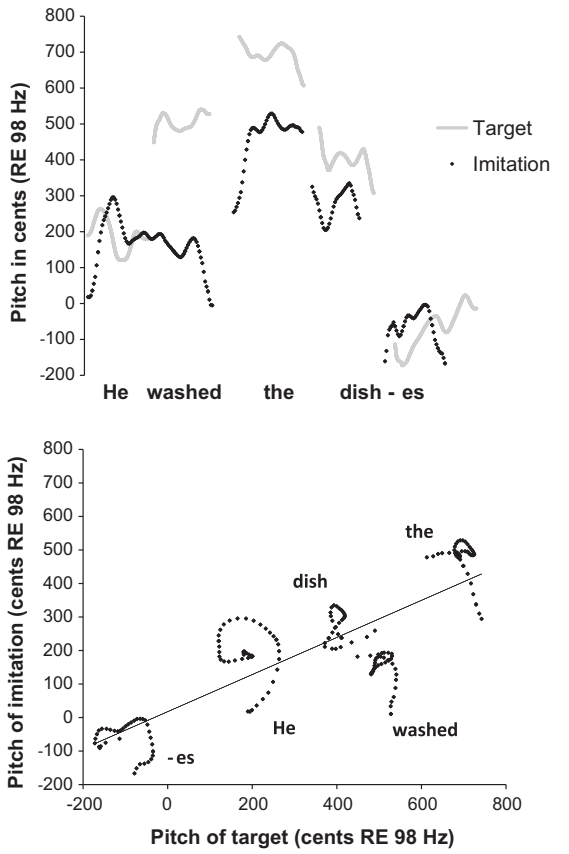

Fig. 2. Examples of performance measures. Plot showing the male target stimulus for the sung melody "He washed the dishes" (2A, upper panel), along with scatterplot relating the sung F0 pattern to the target pattern, which leads to the computation of the pitch correlation measure (2B, lower panel). Text indicates the sung content associated with each segment of the imitation.

Our second measure, *pitch correlation*, was used to measure the accuracy and precision of relative pitch in imitation. It was simply the Pearson correlation between produced and target pitch vectors (for a similar measure see d'Alessandro, Rilliard, & Le Beux, 2011). Theoretically, a perfectly accurate imitation would produce an r of 1; similarly, an imitation that was either *consistently* sharp or flat (i.e., matching contour but mismatching pitch) would also produce an r of 1, as such errors would simply shift the y-intercept of the regression line. Fig. 2B shows a scatterplot, based on the imitation shown in Fig. 2A, along with the correlation line describing the best-fitting linear relationship between target and imitated F0 (note that perfect imitation would lead to a 1:1 relationship as described by a regression line with a slope of 1). Although there is a general tendency for the singer to match the relative pitch height of F0 while imitating, there are also deviations from linearity, which reflect the fact that this singer does not imitate F0 fluctuations within each note (see e.g., the imitation of the pitch information corresponding to "He"). These fluctuations, along with the general tendency of this singer to compress pitch range while imitating (leading to a slope of the regression line of .56), causes the Pearson correlation to fall short of the ideal value (r = .82). We focus on correlation coefficients, rather than slope, because the correlation captures both the accuracy and the precision of imitated relative pitch.

Analyses were performed with a 2 (sequence type: sentences versus melodies) × 2 (phonetic information: worded versus wordless) repeated measures ANOVA. Significant interactions were examined using Tukey's honestly significant difference (HSD) test. Between-experiment comparisons were performed by calculating 95% confidence intervals for means in the baseline Experiment 1 and then determining which means from other experiments fell within those bounds. All statistical decisions were made with α = .05.

3. Experiment 1

The first experiment was intended to determine the effects of sequence type and phonetic information on vocal imitation performance. Participants imitated worded and wordless variants of song and speech targets to the best of their ability. If pitch processing during vocal imitation is domain specific, then we should observe differences across song and speech trials, likely resulting in an advantage for song. If pitch processing is also encapsulated, then there should be no effect of phonetic information on pitch accuracy.

3.1. Method

3.1.1. Participants

A total of 27 (female n = 12, male n = 15) University at Buffalo undergraduate students ranging in age from 18 to 29 (M = 19.82 years, SD = 2.25 years) participated. Five participants reported vocal training ranging from 2 to 12 years (M = 6.80, SD = 3.76). This training included individualized lessons and chorus participation. Two participants reported instrumental training of less than 6 years each. Overall, 20 of 27 participants reported 1 year or less of musical training either as part of a school curriculum or as private lessons; thus the sample was composed mainly of musical novices. Seven participants reported native languages other than English (French, Japanese, Indian, Korean, Asanti Twi, and Russian); all but one reported a high comfort rating for English use. Experimental results did not change when we removed these participants from data analyses, so their data were retained. Five other participants reported secondary fluency in languages including Cantonese, Hindi, Spanish, and Russian. All but two participants were right handed. Recordings from two participants (both musically untrained females with English as their only language) were not utilized because of excessively creaky voice, resulting in a total of 25 participants. We report analyses that average across all participants irrespective of musical training. For all experiments in this paper, we conducted additional analyses on data from the musically untrained participants (less than 1 year of music lessons) alone. These additional analyses did not yield qualitatively different

patterns than the reported results from the samples containing both trained and untrained participants, so they were not further explored.

3.1.2. Procedure

Participants listened to and imitated all 96 stimuli (see Section 2: General Methods) one time each. Participants were randomly assigned to one of two orders of trials. Before the experiment began, participants were reminded to imitate to the best of their ability. During worded trials, participants imitated by producing the words they heard. When the trial did not have words, participants imitated using [ɑ] (which was compatible with the sound of the wordless targets).

3.2. Results

3.2.1. Absolute pitch error

Mean absolute error values across the four sequence type × phonetic information conditions are shown in Fig. 3A. The ANOVA yielded a significant main effect of sequence type, $F(1,24) = 52.85$, $p < .01$, $\eta_p^2 = .69$, and of phonetic information, $F(1,24) = 18.78$, $p < .01$, $\eta_p^2 = .44$. There was no sequence × phonetics interaction. The main effect of sequence indicated better performance (lower error) for melodies ($M = 148.1$ cents, $SD = 102.63$ cents) than for sentences ($M = 214.74$, $SD = 73.39$). The significant main effect of phonetics indicated that worded trials were imitated better than wordless trials, (worded $M = 171.23$ cents, $SD = 99.83$ cents; wordless $M = 191.62$, $SD = 89.48$).

3.2.2. Relative pitch accuracy

Mean pitch correlations are shown in Fig. 3B. There was a significant main effect of phonetic information, $F(1,24) = 14.51$, $p < .01$, $\eta_p^2 = .38$, and a significant sequence × phonetics interaction, $F(1,24) = 5.92$, $p = .02$, $\eta_p^2 = .20$, but no main effect of sequence type. Worded trials ($M = .84$, $SD = .06$) were imitated significantly more accurately than wordless trials ($M = .80$, $SD = .08$), but melodies and sentences overall were imitated similarly well (both $M = .82$). The sequence × phonetics interaction suggested a greater effect of phonetic information on sentence than on song imitation. However, this implication was not fully verified in post hoc analyses, which simply confirmed that the main effect of phonetic information pertained to both sentence and melody conditions. Both pairwise differences between worded and wordless sequences were significant; no other differences reached significance although the difference between wordless sentences and wordless melodies approached significance (difference between these conditions = 0.0216, criterion for HSD = 0.0221).

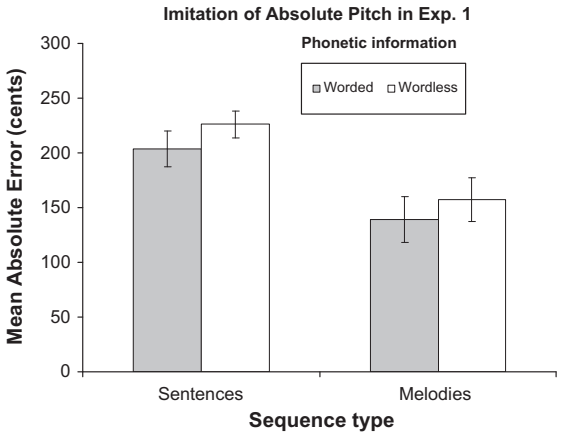

Fig. 3A. Mean absolute error in the sequence × phonetics interaction in Exp. 1; lower values indicate greater accuracy. Error bars represent one standard error of the mean.

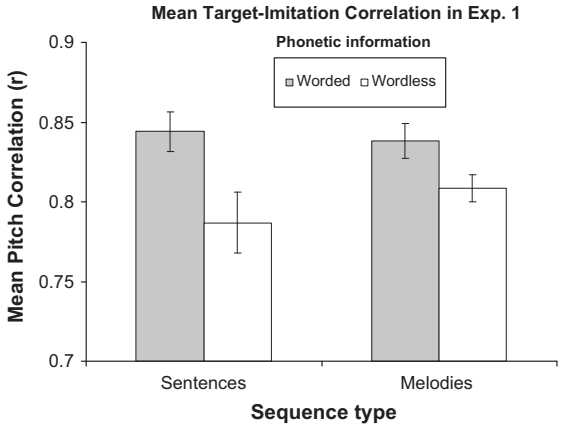

Fig. 3B. Mean target-imitation correlation in the sequence × phonetics interaction in Exp. 1; higher values indicate greater accuracy. Error bars represent one standard error of the mean.

3.2.3. Analyses with syllable-aligned trajectories

As noted before, the sequence length analyses we conducted only align the beginning of the imitation and target sequences in order to compare them. However, it is conceivable that minute timing errors, such as vowel elongation in the imitative production, could shift the entire pitch-time trajectory and negatively influence the analysis of an otherwise accurate production. In order to test this possibility, we used the syllable boundaries from the worded sequences to create syllable-aligned imitation and target pitch-time trajectories. For example, in a sequence with four syllables, the target and imitation would be aligned at the beginning of each of the four syllables. Within each imitation syllable, the trajectory was resampled and interpolated so that the number of samples matched the target trajectory. This analysis is not possible with the wordless sequences because they do not contain phonetically-defined syllable boundaries.

For relative pitch accuracy, each imitation-target syllable pairing produced a correlation coefficient, which we averaged to derive a single sequence-length correlation (comparable to our original relative pitch accuracy measure). As was the case with the original analysis, the difference between worded speech and song was not statistically significant (worded song $M = .85$, $SD = .07$; worded speech $M = .84$, $SD = .09$; $p = .28$, $\eta_p^2 = .05$). We also

compared the results of the new syllable-aligned analysis for mean absolute error to our original absolute pitch accuracy results, and nothing changed; the difference between worded speech and worded song continued to be statistically significant: worded song $M = 132.04$, $SD = 102.18$; worded speech $M = 193.32$, $SD = 75.91$; $F(1,27) = 43.40$, $p < .01$, $\eta_p^2 = .63$. In total, the results of the new analysis with syllable-aligned boundaries does not provide any evidence that our original accuracy measures are overly sensitive to timing errors. Even so, we comment on the potential limitations of our accuracy analyses in the General Discussion (Section 8.4).

3.2.4. Analyses of pitch by note

As described earlier, we analyzed pitch accuracy by using continuous change in F0 across the entire recorded pitch vector because we wanted to measure the imitation of fine-grained temporal changes in F0. However, it is not clear whether the same effects found here would be identified by a more traditional note-by-note analysis. Thus, we ran a follow-up analysis based on central tendency measures of F0 within notes and syllables as we have done in previous research (Pfordresher & Brown, 2007, 2009; Pfordresher et al., 2010). To our knowledge, there is no valid way to segment the pitch-time trajectories of wordless sentences, due to the considerable fluctuations of F0 both within and across segments. Thus, we ran two separate analyses comparing the remaining three conditions. Comparisons between worded song and worded speech were conducted with segmentations based on phonetic information. Comparisons between worded songs with wordless songs were performed based on the pitch patterns of produced songs, using the abrupt changes between notes that characterize song production. For all analyses, we measured the median F0 for the central portion of each sung syllable (middle 50% of sampled F0 values), which removed the influence of glides that can occur at the beginning and ends of notes.

Pitch error measures reported above (Fig. 3A) are comparable to the measure referred to in Pfordresher and Brown (2007) as *mean absolute note error*: the mean absolute difference between sung and target notes across a sequence. We computed this measure for all song imitations (worded and wordless) for each participant. The advantage for worded above wordless songs was not statistically significant (M error for worded = 108.49 cents, $SD = 108.67$, M for wordless = 111.06, $SD = 112.25$), $F < 1$, $\eta_p^2 = .02$. Thus, the fact that a phonetic advantage has not been found in other studies that vary linguistic content (e.g., Racette & Peretz, 2007) may be related to the type of measurement that is used. However, the advantage for song over speech within worded trials remained when using the mean absolute note error measure (treating syllables as "notes" in speech), $F(1,24) = 72.01$, $p < .01$, $\eta_p^2 = .75$ (M error for song = 112.95 cents, $SD = 108.48$, M for speech = 205.09, $SD = 85.99$).

Next we compared pitch correlation measures used here to *mean absolute interval error* for each participant, which is the mean absolute difference between sung pitch intervals versus target pitch intervals across a sequence (see Pfordresher & Brown, 2007), and is equivalent to the "interval deviation" measure reported by Dalla Bella and colleagues (Berkowska & Dalla Bella, 2009; Dalla Bella et al., 2007, 2009). According to this measure, the phonetic advantage within song conditions was nominally present but did not reach significance (M error for worded = 88.76 cents, $SD = 37.76$, M for wordless = 98.73, $SD = 47.19$), $p = .12$, $\eta_p^2 = .10$. Interestingly, a robust song advantage emerged in the contrast between worded song and worded speech that was not present in the pitch correlation data (see Fig. 3B), $F(1,24) = 117.20$, $p < .01$, $\eta_p^2 = .83$ (M error for song = 97.91 cents, $SD = 50.89$, M for speech = 218.86, $SD = 59.39$).[2]

Analyses by notes and syllables thus differ in several ways from analyses based on the entire F0 vector, the most salient being the appearance of a song advantage within worded trials for the measure of relative pitch that was absent from the pitch correlation measure analyses. As discussed earlier, we believe that differences across measures are due to the reduction of information in traditional by-note analyses, which disregards the accuracy with which participants may imitate F0 patterns within rhythmic units. We suggest that this reduction of information places speech imitation at a particular disadvantage. As such, we ran additional analyses that address the accuracy with which participants imitated F0 within notes or syllables. Because we could only be confident about the precise location of our segmentations between rhythmic units for worded trials, we based this analysis on the comparison between worded speech and worded song.

We computed pitch correlations separately within each syllable or note of worded trials. This analysis disregards a participant's ability to track pitch variations across successive notes or syllables, and thus is only sensitive to pitch variations within these rhythmic units. Speech includes larger variability within syllables than song (Stegemöller, Skoe, Nicol, Warrier, & Kraus, 2008), and variations within each syllable for speech are more informative than song, where such variations are typically limited to vibrato and other incidental variations such as pitch "scoops." In keeping with these observations, correlations within segments were significantly higher for speech than for song in wordless trials (M correlation for song = .42, $SE = .01$, M for speech = .59, $SE = .02$), $F(1,24) = 87.76$, $p < .01$, $\eta_p^2 = .79$. Thus, the null effect of domain for pitch correlations across the entire F0 trajectory may reflect a tradeoff between the ability to imitate pitch information across segments (better for song) and the ability to imitate pitch information within segments (better for speech). It is plausible that the fast-moving pitch information within the speech syllables requires superior temporal resolution provided by left hemispheric speech processors (Zatorre, Belin, & Penhune, 2002).

[2] It is worth noting that the difference between this analyses and the pitch correlation measure is not due to the use of the correlation coefficient *per se*. We ran a follow-up analysis based on pitch correlations across the four point estimates used to derive mean absolute interval error measures ($N = 3$–5 samples per trial), and this measure also yielded a worded song over worded sentence imitation advantage.

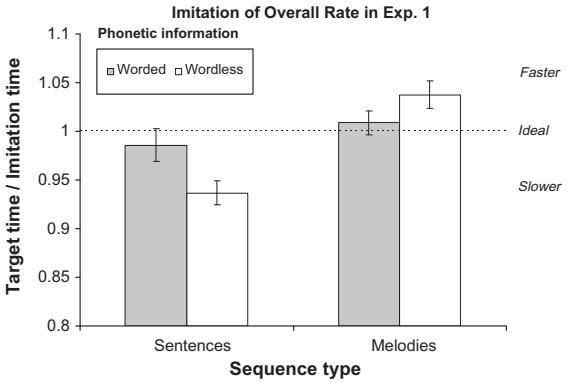

Fig. 3C. Overall rate accuracy in the sequence × phonetics interaction in Exp. 1. The dashed line indicates ideal imitation of rate while the areas above and below indicate faster and slower than ideal imitation timing, respectively. Error bars represent one standard error of the mean.

3.2.5. Imitation of production rate

Mean duration ratios are shown in Fig. 3C.[3] The ANOVA revealed a main effect of sequence type, $F(1,24) = 35.26$, $p < .01$, $\eta_p^2 = .60$ (melodies $M = 1.02$, $SD = .07$; sentences $M = .96$, $SD = .08$), and a sequence × phonetics interaction, $F(1,24) = 20.32$, $p < .01$, $\eta_p^2 = .46$. There was no main effect of phonetic information. Next, all means in the significant sequence × phonetics interaction were compared with post hoc analyses; the production rate during imitations of wordless sentences was significantly slower than worded sentences and both worded and wordless melodies. Worded and wordless melodies were not significantly different from each other.

We went on to measure duration error, independent of speeding versus slowing, by calculating the absolute difference between mean imitation and mean target duration (in seconds) divided by the mean target duration. This descriptive measure yields a value close to zero when the absolute difference between imitation and target duration is low; higher values indicate greater degree of error. The results indicated that timing in worded melodies was imitated most accurately (0.0016), followed by worded sentences (0.0248) and wordless melodies (0.0290), with wordless sentences least accurate overall (0.0695). These results further support the conclusion that phonetic information facilitates imitation of timing in melodies and sentences, and also that the imitation of sentence timing may rely on phonetic information more than the imitation of melody timing.

[3] It is conceivable that unusually slow or fast imitations could seriously alter the accuracy results. In order to assess this possibility, the duration ratios were scrutinized to determine how many of them fell outside the range of three standard deviations from the mean within each of the four current experiments. In total, less than 1.1% of the duration ratios (144 out of 13,438 trials) qualified as outliers in this sense. Furthermore, when these outliers were removed from the data and all ANOVAs were recomputed, the results were almost entirely the same as those currently reported. The upshot is that unusually slow or fast imitative productions do not strongly influence overall results.

3.2.6. Imitation of spectral information

An important consideration in interpreting the advantage for worded over wordless trials has to do with timbral information. Although the "hum" sounds used for nonworded trials were designed to mimic the formants of a human voice, the spectra of wordless recordings were not identical to the original recordings from which they were derived. Specifically, wordless spectra featured a dramatic spectral tilt up to 6 kHz, which was not present in the original recordings. Spectral differences could be related to the exclusion of phonetic information or to the relative artificiality of the "hum" sound. This distinction is potentially important because timbre and pitch have been shown to interact perceptually (Melara & Marks, 1990) and timbre interference may be worse for nonmusicians than musicians (Pitt, 1994). Pitch matching becomes difficult when participants compare tones of different timbres (Krumhansl & Iverson, 1992), and resulting confusions can lead to illusions of pitch height (Russo & Thompson, 2005) and may interfere with vocal pitch matching (Hutchins & Peretz, 2012).

In order to address these concerns, we computed long-term average spectra (LTAS) for all targets, and correlated these with corresponding LTAS measures for imitations. LTAS has been used in the past to measure spectral vocal performance of melodies (Sundberg, 1999). We assessed LTAS for frequencies from 0 to 10 kHz in steps of 10 Hz, which encompasses frequencies present in all targets (worded and wordless). Each target LTAS was correlated with every imitation of that target, and the resulting correlation coefficients (one for every participant and trial) were submitted to a 2 (phonetic information) × 2 (sequence type) ANOVA. Most importantly, there was a main effect of phonetic information, $F(1,25) = 106.78$, $p < .01$, $\eta_p^2 = .81$. Stronger correlations emerged for the imitation of wordless targets ($M = .87$, $SE = .01$), than for worded targets ($M = .79$, $SE = .01$). Thus, participants more effectively imitated timbre for targets that lacked phonetic information, even when their imitation of pitch was worse for these trials. The ANOVA also yielded a main effect of sequence type, $F(1,25) = 44.78$, $p < .01$, $\eta_p^2 = .64$, and a phonetic information × sequence type interaction, $F(1,25) = 19.10$, $p < .01$, $\eta_p^2 = .43$. The spectra of sentences were imitated more accurately ($M = .85$, $SE = .01$) than for melodies ($M = .81$, $SE = .01$). This is a notable finding given that Warrier and Zatorre (2002) showed that tonal context reduces the interfering influences of timbre (thus, we may have expected superior imitation of melodic spectra). The interaction in the current data reflected the fact that the difference between worded and wordless trials was larger for the imitation of melodies (difference = .10) than sentences (difference = .06). The critical outcome of this analysis is that difficulty with timbre matching cannot explain the observed phonetic advantage.

3.3. Discussion

In Experiment 1, participants imitated pitch-time trajectories typical of song and speech based on original versions with words and also wordless variants. Our primary interest was whether imitative performance would reflect

domain specific pitch processing and whether such processing is encapsulated from the effects of phonetic information. Results revealed partial support for the former prediction and no support for the latter prediction. Vocal imitation of pitch was more accurate for music than speech with respect to absolute pitch, but not relative pitch (for worded targets). Thus, any domain specific processing of pitch may be limited to absolute rather than relative pitch content. Furthermore, a phonetic advantage was found within each domain across all production measures. The accuracy of vocal imitation for pitch appears to be influenced by non-pitch information. However, one result that does accord with possible domain specific differences was found in the pitch correlation measure (accuracy of relative pitch), which hinted that the imitation of pitch in speech may rely on phonetic information more so than the imitation of melodies. The difference in pitch correlations between worded and wordless sentences (0.057) was nearly twice as large as the difference between worded and wordless melodies (0.029). However, post hoc analyses only supported the main effect of phonetics. As such, though this interaction is large in absolute terms, it may be of low reliability.

Comparisons across different types of analyses suggested a distinction between the imitation of speech and song: speech imitation, more so than song imitation, may be sensitive to fine-grained fluctuations in pitch that occur within syllables as well as across syllables. The song advantage was prominent in all analyses concerning the imitation of absolute pitch, but conclusions based on relative pitch were more variable. Whereas pitch correlations based on the entire trace suggested no differences across domains (for worded trials), pitch correlations within segments (syllables or notes) suggested a speech advantage, and analyses that disregarded fluctuations within a segment ("note-by-note" analyses) suggested a song advantage. These results suggest domain specific differences that are distinct from the modular model of Peretz and Coltheart (2003), which predicts a song advantage due to tonal encoding. Rather, we think the present results are best accounted for by signal-specific properties, and how these properties of a signal can be tracked by an integrated vocal imitation mechanism. Specifically, because pitch fluctuations in speech are more variable (Stegemöller et al., 2008) and linked to transitions between phonemes as well as suprasegmental information, pitch imitation is oriented toward smaller timescales. We think a similar property leads to the phonetic advantage, which was reduced (and was non-significant) when fine-grained variability was disregarded. Pitch fluctuations in speech, and to a reduced degree in song, are linked to variations in articulation. When these articulations are absent, as in wordless trials, these pitch fluctuations lose their meaning and are thus harder to imitate.

The phonetic advantage that participants experienced when imitating worded sequences is compelling, but it can be associated with two different explanations. One explanation is based on the hypothesis that segmental and suprasegmental levels are integrated in the representation of the to-be-imitated sequence; that is, phonetic information and pitch are processed interdependently. This interdependence may be particularly critical given the measures we used, in that participants needed to imitate fine-grained fluctuations in pitch to perform ideally. A second explanation focuses specifically on perception: Phonetic information may be associated with better imitation because segments partition the pitch-time contour into higher-order units that facilitate encoding. For instance, autosegmental theories of speech processing suggest that listeners categorize syllables discretely based on the accentual properties of phonetically defined segments (Pierrehumbert, 1980/87). Experiment 2 addressed these two interpretations by having participants imitate all sequences (worded and wordless) with a neutral vowel.

4. Experiment 2

If the speech and song phonetic advantage identified in Experiment 1 was a result of perceptual facilitation brought solely by perceiving phonetic information, then it might be replicated when participants imitate worded sequences but omit words in their produced imitations. That is, participants should be just as good at imitating the pitch-time trajectories of worded sequences, even when they do not reproduce the apprehended phonetic information, because the facilitative benefit of phonetic information has already been obtained during perceptual processing (recall that phonetic information is technically irrelevant for accuracy measures). By contrast, if the phonetic advantage were to diminish in Experiment 2, it would suggest that the cause of the phonetic advantage is based on the use of phonetic information during the process of imitative production and not just on perceptual segmentation. For the sake of brevity, data analyses for this and all remaining experiments focus on our primary measures of performance: pitch error, pitch correlation, and production duration evaluated across the entire F0 vector, with imitations and targets temporally aligned from the start of the sequences (see Section 3.2.3).

4.1. Methods

4.1.1. Participants

Thirty-one University at Buffalo students and one other adult participated in Experiment 2 (female $n = 13$, male $n = 19$). Participants' ages ranged from 18 to 27 years ($M = 19.9$, $SD = 2.18$). Six participants reported vocal training (lessons) of at least 4 years ($M = 5.33$, $SD = 1.97$) and sixteen participants reported instrumental training ranging from 2 to 9 years ($M = 3.94$, $SD = 2.3$). Overall, participants in Experiment 2 reported more instrumental music experience than those in Experiment 1, but this moderate level (4 years) is not unusual among college students, and the difference between experiments was not statistically significant. Six participants reported first languages other than English (including Mandarin, Burmese, Malayalam, Vietnamese, and Bengali) and another reported learning English and Spanish natively. All of these participants rated their English comfort level as high or moderately high. Four other participants reported fluency in other languages, and all but two participants were right handed.

4.1.2. Procedure

The general procedure and stimuli were the same as Experiment 1 except that participants were instructed *to imitate all sequences using* [ɑ]. Thus, even when the target contained words, participants had to imitate its pitch-time contour using the syllable "ah." Another difference from the first experiment was that participants in Experiment 2 did not imitate all of the target sequences once each. Instead, they imitated sequences in either of eight order conditions that contained 24 of the original 96 targets in a pseudorandom order within three blocks for a total of 72 targets. Stimuli in each condition were chosen so that participants never heard both worded and wordless versions of the same target. This constraint was designed to avoid carry-over effects that could cause phonetic information during one trial to facilitate production on a later nonword trial. For instance, if a participant heard the worded sentence "He ate it all" followed later by the wordless version of the same utterance, performance on the later trial might benefit from the participant's memory of the earlier trial.

4.2. Results

4.2.1. Absolute pitch accuracy

Mean absolute error values from Experiment 2 are shown in Fig. 4A. The ANOVA revealed a main effect of sequence type, $F(1,31) = 46.07$, $p < .01$, $\eta_p^2 = .60$, but no main effect of phonetic information and no interaction. As in Experiment 1, mean absolute error was lower in melodies ($M = 150.61$ cents, $SD = 117.83$) than sentences ($M = 235.7$, $SD = 91.1$).

In order to examine the difference in accuracy of absolute pitch imitation between experiments, each of the four means in the sequence × phonetics interactions in Experiments 1 and 2 were compared by calculating 95% confidence intervals for each of the means in Experiment 1 and determining which of the respective means from Experiment 2 fell within that range. Importantly, the results indicated that the only difference was between worded sentence means (Experiment 1 $M = 169.13$ cents ±34.42; Experiment 2 $M = 242.31$). Worded melodies were not similarly disrupted (Experiment 1 $M = 138.91$ cents ±43.58; Experiment 2 $M = 153.49$).

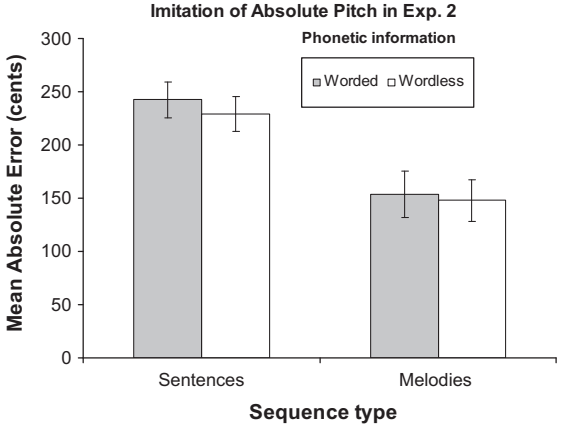

Fig. 4A. Mean absolute error in the sequence × phonetics interaction in Exp. 2. Lower values indicate greater accuracy. Error bars represent one standard error of the mean.

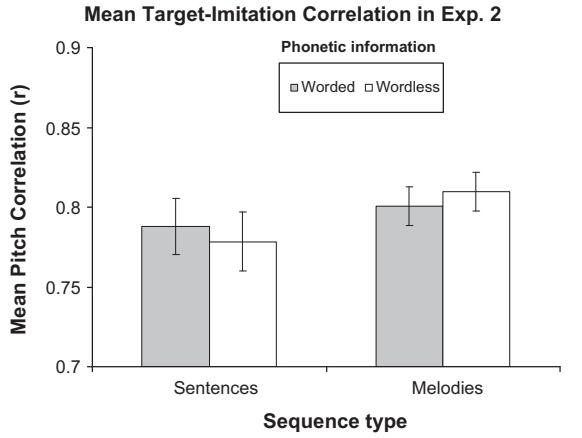

Fig. 4B. Mean target-imitation correlation in the sequence × phonetics interaction in Exp. 2; higher values indicate greater accuracy. Error bars represent one standard error of the mean.

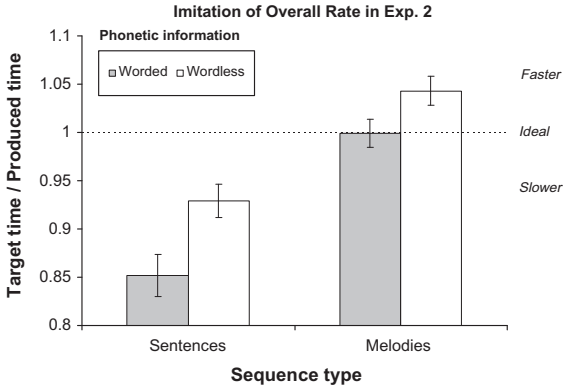

Fig. 4C. Overall rate accuracy in the sequence × phonetics interaction in Exp. 2. The dashed line indicates ideal imitation of rate while the areas above and below indicate faster and slower than ideal imitation timing, respectively. Error bars represent one standard error of the mean.

4.2.2. Relative pitch accuracy

Mean pitch correlation values from Experiment 2 are shown in Fig. 4B. The ANOVA revealed no significant effects. All four means from Experiment 2 were compared to Experiment 1 using confidence intervals. As in the analysis of absolute pitch accuracy, relative pitch in worded sentences was imitated worse in Experiment 2 (Experiment 1 $M = 0.8440 \pm 0.0258$; Experiment 2 $M = 0.7878$). However, imitation of relative pitch was also worse for worded melodies (Experiment 1 $M = 0.8382 \pm 0.0227$; Experiment 2 $M = 0.8004$).

4.2.3. Imitation of production rate

Mean duration ratios from Experiment 2 are shown in Fig. 4C. The ANOVA revealed a main effect of sequence

type, $F(1,31) = 51.15$, $p < .01$, $\eta_p^2 = .62$ (melodies $M = 1.02$, $SD = .09$; sentences $M = .89$, $SD = .12$), and a main effect of phonetic information, $F(1,31) = 27.22$, $p < .01$, $\eta_p^2 = .47$ (worded $M = .93$, $SD = .13$; wordless $M = .99$, $SD = .11$), but no interaction. Wordless sequences were always faster than their counterpart worded sequences. However, whereas worded melodies were imitated closer to ideal timing than wordless melodies, worded sentences here appear to be less accurate than wordless sentences. A descriptive analysis of absolute duration error further indicated that timing in worded melodies was imitated most accurately (0.0099), followed by wordless melodies (0.0330) and wordless sentences (0.0984). The timing of worded sentences was imitated worst overall (0.2308). Thus, in contrast to Experiment 1, these results indicate that imitation of timing for speech targets suffered compared to melodic targets. Further, the disruptive effect was especially pronounced for wordless sentences.

In order to further examine the difference in accuracy of timing imitation between experiments, we utilized the confidence interval procedure to compare duration ratios across Experiments 1 and 2. Similar to the absolute pitch accuracy results, the only between-experiment difference was for worded sentences (Experiment 1 $M = 0.986 \pm 0.338$; Experiment 2 $M = 0.852$), indicating that the imitation of timing in worded sentences slowed when participants heard words in the stimulus but did not use them in their imitation.

4.3. Discussion

Experiment 2 was designed to test whether the phonetic advantage was based on segmentation of the perceptual signal. Such an explanation would predict the results of Experiment 2 to match those of Experiment 1, given that the stimuli were identical and that the measures of production focus entirely on F0 during imitation. However, the results of Experiment 2 were unlike Experiment 1 in that the phonetic advantage disappeared when participants reproduced worded sequences with minimal articulation during production. This effect was found across both pitch accuracy measures for both song and speech. Thus, the phonetic advantage is sensitive to matches between perception and production with respect to phonetic information.

We also conducted comparisons across means from Experiments 1 and 2. Two of these analyses, absolute pitch accuracy and timing, suggested greater disruption of worded speech than worded song targets. According to the confidence interval analyses, mean absolute error for worded sentences, but not worded melodies, was worse in Experiment 2 than Experiment 1; similarly, imitation of production rate for worded sentences, but not worded melodies, was more inaccurate in Experiment 2. These results hint that the phonetic advantage may have a stronger perceptual basis for song than for speech imitation. Yet, if this is the case, the phonetic advantage for song must be small because we did not observe a phonetic advantage for song imitation within Experiment 2. In sum, minimizing the phonetic content of imitative productions had the effects of reducing accuracy of timing and absolute pitch matching for worded sentences but not for worded melodies.

The results from Experiment 2 do not support a strictly perceptual basis for the phonetic advantage found in Experiment 1, but the results leave open at least two production-based explanations. Specifically, the findings in Experiment 2 could be interpreted as resulting from the absence of a phonetic advantage, or from an interference effect, based on the fact that participants in Experiment 2 essentially had to "filter out" the apprehended phonetic information from worded targets when forming a vocal performance plan. It is difficult to tease apart these interpretations. However, the fact that imitation of worded conditions in Experiment 2, which could have led to interference, were never worse than wordless conditions, for which no interference was present, suggests that the results stem from an absent advantage rather than interference.

Whereas Experiment 2 focused on the basis for the phonetic advantage, the remaining experiments further probed differences across domains with respect to temporal properties of pitch patterns. We were curious about the extent to which the targets' syllable timing properties might affect the accuracy of speech and song imitation. In Experiments 3 and 4 we investigated the influence of target syllable timing by directly manipulating the temporal contents of the stimuli by equating overall duration (both experiments), or by manipulating relative timing of speech or song to match the other domain (Experiment 4).

5. Experiment 3

Although the speech and song sequences in the first two experiments were matched for pitch contour and word content, they were not equated for overall duration; the melodies were actually produced slower than the speech sequences (See Appendix B). Experiment 3 investigated the influence of overall sequence duration by equating the temporal length of the matched speech and song pairs. Duration was equated by altering the duration of component syllables while preserving their relative timing. The details of the procedure are described below.

5.1. Methods

5.1.1. Participants

Thirty-three University at Buffalo students participated (female $n = 18$, male $n = 15$). Their ages ranged from 18 to 33 years ($M = 19.81$, $SD = 3.21$). Six participants reported vocal training of at least 4 years ($M = 6.33$, $SD = 1.97$) while 21 participants reported instrumental training ranging from 1 to 15 years ($M = 5.62$, $SD = 3.8$). Three participants reported a first language other than English (all Chinese) but each rated their English comfort level as at least moderately high. Two participants were left handed. The data from one participant (a female) was lost due to computer malfunction; data analysis was performed on data from 32 participants.

5.1.2. Stimuli

We equated the overall duration of matched pairs of sentences and melodies (e.g., "He ate it all" spoken as a question by the male speaker, and sung as a question by the same individual) while preserving the relative timing of syllables. Specifically, the duration of each spoken or sung sequence (including all phonetic information) was transformed such that sentences were lengthened by 30% of the difference in total duration across matched sentence/melody pairs and melodies were shortened by 70% of the difference. For example, if in a matched pair the melody was 2000 ms in total duration and the sentence was 1500 ms, the sentence would be lengthened by 150 ms and the melody would be shortened by 350 ms to form a common duration of 1650 ms. We chose to alter melody timing more than sentence timing because larger changes to sentence timing led to degraded intelligibility and reduced naturalness; the duration transformations did not produce such noticeable effects in melodies. New wordless sequences were synthesized from the duration transformed stimuli.

In order to determine the influence of rate manipulations on the naturalness of targets from Experiment 3 versus Experiment 1, we conducted a follow-up study in which 29 participants (who had not participated in any of the imitation experiments) rated the naturalness of all targets from Experiments 1 and 3. Participants used a scale ranging from 1 ("from natural speech") to 7 ("from natural song"). The middle value in the scale (4) was labeled "neutral" so that participants could choose this response if they were unsure about the naturalness of a target. The mean rating of every speech target type differed from the mean rating for every song target type in the expected direction; thus alterations of rate (and phonetic information) did not influence the distinctiveness between song ($M = 5.77$; $SD = .83$) and speech ($M = 2.22$; $SD = .85$) targets. It is important to note that the continuum of response alternatives ranged from speech to song and not from "natural" to "unnatural." Thus, the response scale is better for comparing domain differences instead of naturalness differences within a domain. Based on the logic that ratings further from the middle neutral response imply higher naturalness (i.e., less domain ambiguity) we transformed the scores to represent naturalness by using absolute values of the rating scores centered around zero (leading to an ordinal scale from 0 to 4 representing low to high naturalness). According to Tukey's HSD, all of the worded targets were more natural sounding than the wordless targets. Worded speech targets from Experiment 1 were rated most natural overall ($M = 2.61$, $SD = .39$), followed by worded melodies ($M = 2.34$, $SD = .62$) from the same experiment (a significant difference). The worded speech targets from Experiment 1 were statistically more natural sounding than worded speech targets ($M = 2.25$, $SD = .54$) and worded melody targets from Experiment 3 ($M = 2.20$, $SD = .66$). Finally, the worded melodies from Experiment 1 were statistically significantly more natural sounding than the worded melodies from Experiment 3. In summary, the analysis of transformed ratings shows that the original stimuli were more natural sounding than the rate manipulated stimuli used in Experiment 3. However, the analyses also show that these differences were small in magnitude compared to the differences between domains. In other words, even the target stimuli rated as low in naturalness (compared to the original stimuli) were easily identified as speech or song.

5.1.3. Procedure

The procedure was the same as Experiment 1.

5.2. Results

5.2.1. Absolute pitch accuracy

Mean absolute error values are shown in Fig. 5A. The ANOVA revealed a significant main effect of sequence type, $F(1,31) = 155.76$, $p < .01$, $\eta_p^2 = .83$, (melodies $M = 110.85$ cents, $SD = 31.6$ cents; sentences $M = 208.86$, $SD = 62.52$), and a significant sequence × phonetics interaction, $F(1,31) = 8.24$, $p < .01$, $\eta_p^2 = .21$. There was no main effect of phonetic information. The interaction reflects the fact that phonetic information influenced melody imitation but not sentence imitation. This was confirmed with post hoc analyses; the only nonsignificant paired contrast difference was between worded and wordless sentences. The post hoc tests showed that worded melodies were imitated most accurately overall, while both worded and wordless sentences were imitated least accurately.

The large sequence effect ($\eta_p^2 = .83$) shows that melodies were imitated much more accurately than sentences. This melodic sequence advantage was greater than the significant effects identified in Experiment 1 ($\eta_p^2 = .69$) and Experiment 2 ($\eta_p^2 = .60$), possibly because of decreased variance in the current data (note the standard error bars across Figs. 3A, 4A and 5A).

5.2.2. Relative pitch accuracy

Target-imitation pitch correlation data are shown in Fig. 5B. There was a main effect of sequence type, $F(1,31) = 14.02$, $p < .01$, $\eta_p^2 = .31$, a main effect of phonetic information, $F(1,31) = 16.88$, $p < .01$, $\eta_p^2 = .35$, and a significant sequence × phonetics interaction, $F(1,31) = 49.39$,

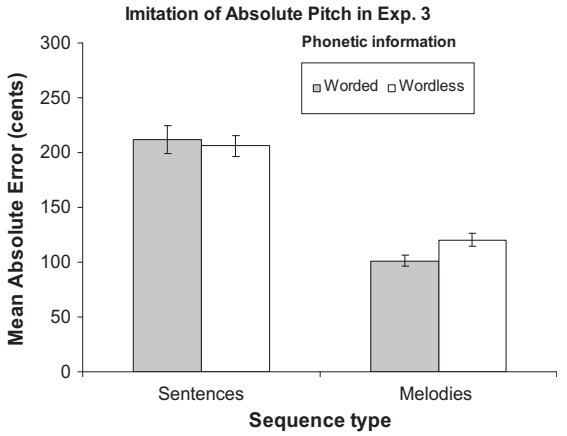

Fig. 5A. Mean absolute error in the sequence × phonetics interaction in Exp. 3 (duration equated targets). Lower values indicate greater accuracy. Error bars represent one standard error of the mean.

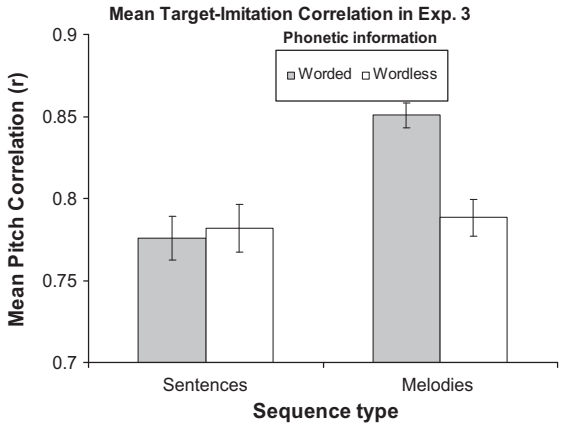

Fig. 5B. Mean target-imitation correlation in the sequence × phonetics interaction in Exp. 3 (duration equated targets); higher values indicate greater accuracy. Error bars represent one standard error of the mean.

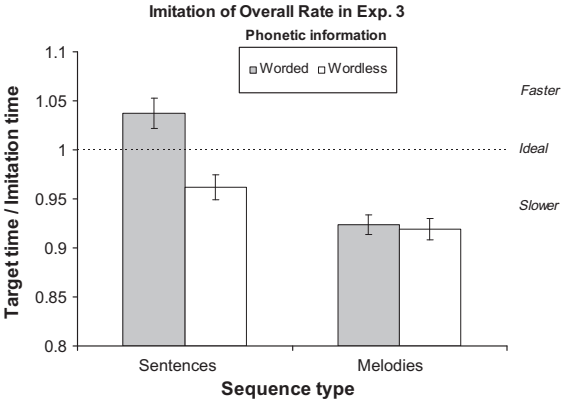

Fig. 5C. Overall rate accuracy in the sequence × phonetics interaction in Exp. 3 (duration equated targets). The dashed line indicates ideal imitation of rate while the areas above and below indicate faster and slower than ideal imitation timing, respectively. Error bars represent one standard error of the mean.

$p < .01$, $\eta_p^2 = .61$. The interaction was driven by the large facilitating influence of phonetics within melodies but not sentences, as found in the absolute accuracy measure. Post hoc tests confirmed that worded melodies ($M = .85$, $SD = .04$) were imitated significantly more accurately than all other sequence types. There were no other significant contrasts.

5.2.3. Imitation of production rate

Mean duration ratios are shown in Fig. 5C. The ANOVA revealed a main effect of sequence type, $F(1,31) = 76.01$, $p < .01$, $\eta_p^2 = .71$, a main effect of phonetic information, $F(1,31) = 16.1$, $p < .01$, $\eta_p^2 = .34$, and a sequence × phonetics interaction, $F(1,31) = 23.27$, $p < .01$, $\eta_p^2 = .43$. The interaction suggested that both worded and wordless melodies were imitated slowly but that imitation of rate in speech depended on phonetic information. This interpretation was supported by post hoc analyses, which indicated that all paired contrasts were significant except for worded versus wordless melodies. We do not report a full analysis of absolute timing in Experiments 3–4 because durations of matched stimuli were equated. However it is worth noting that the tendency to speed up during the imitation of worded sentences in Experiment 3 (targets for which were slower than is typical of speech) nevertheless led to speech rates that were considerably slower ($M = 1.52$ s total sequence duration) than imitative speaking rates for the same condition from Experiment 1 ($M = 1.00$ s).

5.3. Discussion

In Experiments 1–2, participants were better able to imitate absolute pitch information in melodies than in sentences. A possible reason for this difference was that melodies were slower than sentences (it has been shown previously that singing accuracy improves considerably when people sing at slower as opposed to faster tempos, Dalla Bella et al., 2007). Yet, as can be seen from the results of Experiment 3, the melodic advantage for the imitation of absolute pitch remained when durations were equated across speech and music targets. We return to the relationship between speed and accuracy across all experiments in the General Discussion (Section 8.3).

The effect of equating duration had an unexpected effect on the imitation of speech; the phonetic advantage disappeared in measurements of pitch accuracy. Although the sentence durations were altered less than melodies, it is possible that even small changes in overall sequence timing upset sentence imitation. Experiment 4 further investigated the influence of relative timing of target syllables.

6. Experiment 4

Experiment 4 investigated the influence of the relative timing of syllables on pitch imitation, while equating total sequence duration as in Experiment 3. Given prior evidence that timing can affect pitch perception (Jones, Boltz, & Kidd, 1982; Jones, Moynihan, MacKenzie, & Puente, 2002; Prince, Schmuckler, & Thompson, 2009), it is possible that the complexity of temporal structure for speech may contribute to the disadvantage for imitating the absolute pitch of speech versus song. In the fourth experiment, relative timing was altered so that speech targets incorporated the relative timing of songs (Exp. 4a) or song targets incorporated the relative timing of speech (Exp. 4b).

6.1. Methods

6.1.1. Participants

Thirty-two University at Buffalo students participated in Experiment 4a (female $n = 10$, male $n = 22$) and 30 participated in Experiment 4b (female $n = 15$, male $n = 15$), leading to a total N in Experiment 4 of 62. Their ages ranged from 18 to 28 years (4a: $M = 19.41$, $SD = 1.04$; 4b: $M = 20.61$, $SD = 2.69$). Eleven participants reported vocal training of at least 4 years ($n = 5$ in 4a, years reported in this group $M = 5$, $SD = 1.73$; $n = 6$ in 4b, $M = 6.75$, $SD = 4.32$) and 36 participants reported instrumental

training ranging from 1 to 14 years (n = 16 in 4a, years reported in this group M = 5.5, SD = 3.37; n = 20 in 4b, M = 6.75, SD = 4.32). All participants reported being comfortable using English, although 7 participants reported a first language other than English (4 from 4a; 3 from 4b). The data from two participants from Experiment 4a were removed due to vocal creak, and the data from one participant in Experiment 4b were removed because the participant did not complete the procedure.

6.1.2. Stimuli

In Experiment 4a, melodic targets were identical to Experiment 3 and we adjusted the relative timing of sentence targets so that the duration of each syllable in a given target sentence was equal to the duration of the same syllable in the matching song target. In Experiment 4b we did the reverse; speech targets were identical to Experiment 3 and we adjusted the relative timing of song targets so that the duration of each syllable was equal to the same duration from the matched sentence target. After equating syllable duration for the worded targets, we synthesized new wordless targets for each experiment. Thus, sentence and song targets for worded and wordless conditions were equal with respect to relative and absolute time.

6.1.3. Procedure

The procedure was the same as Experiment 1.

6.2. Results

The effects of sequence type and phonetic information were highly stable across manipulations of relative timing. Thus, for each analysis type, we present the results of Experiment 4a and 4b adjacently.

6.2.1. Absolute pitch accuracy

Mean absolute error values are shown in Fig. 6A, showing results from Experiment 4a (left panel) and 4b (right panel). In each case, the ANOVA revealed a main effect of sequence type, Experiment 4a, $F(1,29) = 112.94$, $p < .01$, $\eta_p^2 = .80$, Experiment 4b, $F(1,28) = 98.45$, $p < .01$, $\eta_p^2 = .78$, and a main effect of phonetic information, Experiment 4a, $F(1,29) = 16.07$, $p < .01$, $\eta_p^2 = .36$, Experiment 4b, $F(1,28) = 34.0$, $p < .01$, $\eta_p^2 = .55$, but no interaction. As in Experiment 1, melodies were imitated more accurately than sentences, and worded sequences were imitated more accurately than wordless sequences. Thus, surprisingly, holding relative and absolute timing constant across sequence types "brings back" the beneficial influence of phonetic information that was not present for the sentences in Experiment 3.

6.2.2. Relative pitch accuracy

Mean pitch correlation values are shown in Fig. 6B. The ANOVA from Experiment 4a (left) revealed one significant finding: a main effect of phonetic information, $F(1,29) = 17.75$, $p < .01$, $\eta_p^2 = .38$. In Experiment 4b (right) there was a main effect of phonetic information, $F(1,28) = 68.18$, $p < .01$, $\eta_p^2 = .71$ and also a main effect of sequence type, $F(1,28) = 4.33$, $p < .05$, $\eta_p^2 = .13$. Neither experiment yielded a significant interaction. In both Experiments, worded sequences were imitated more accurately than wordless sequences. The main effect of sequence type in Experiment 4b suggests better imitation of sentences than melodies when melodies inherit the relative timing of sentences. However, it should be noted that this main effect (which did not appear in any other experiment) disappears when especially long or short imitative productions (outside of three standard deviations from the mean) are removed from analysis (see footnote 2 in Section 3.2.4). Just as in the analysis of mean absolute error, the phonetic benefit for sentences "returned" when relative and absolute timing were held constant across targets. Moreover, the disappearance of the phonetic benefit for sentences in Experiment 3 cannot simply be due to the reduction of naturalness in certain stimuli brought about by temporal transformations.

6.2.3. Imitation of production rate

Mean duration ratios are shown in Fig. 6C for Experiment 4a (left) and 4b (right). Both ANOVAs yielded a significant main effect of sequence type, Experiment 4a, $F(1,29) = 73.17$, $p < .01$, $\eta_p^2 = .72$, Experiment 4b, $F(1,28) = 59.92$, $p < .01$, $\eta_p^2 = .68$, and a significant

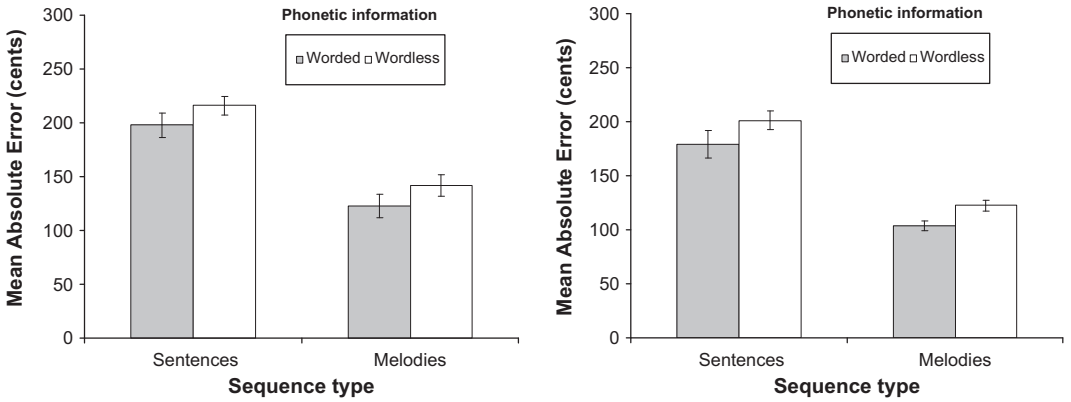

Fig. 6A. Mean absolute error in the sequence × phonetics interaction in Exp. 4a (melodic-timed targets, left) and 4b (speech-timed targets, right). Lower values indicate greater accuracy. Error bars represent one standard error of the mean.

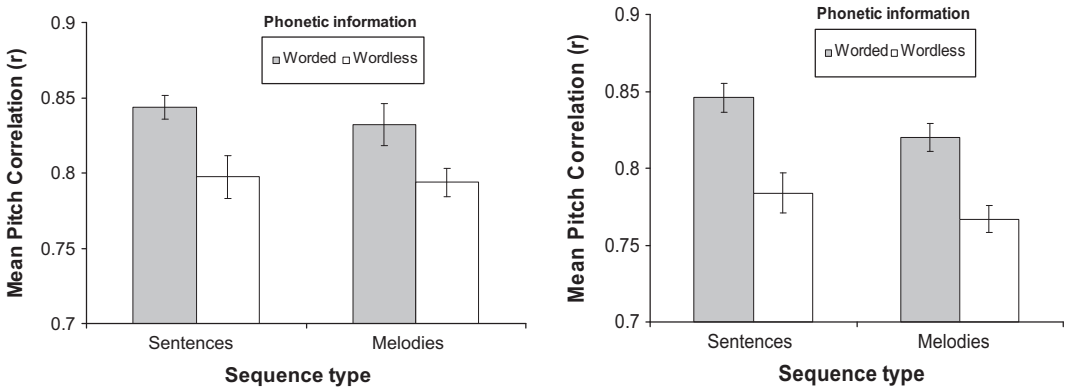

Fig. 6B. Mean target-imitation correlation in the sequence × phonetics interaction in Experiment 4a (melodic-timed targets, left) and Experiment 4b (speech-timed targets, right); higher values indicate greater accuracy. Error bars represent one standard error of the mean.

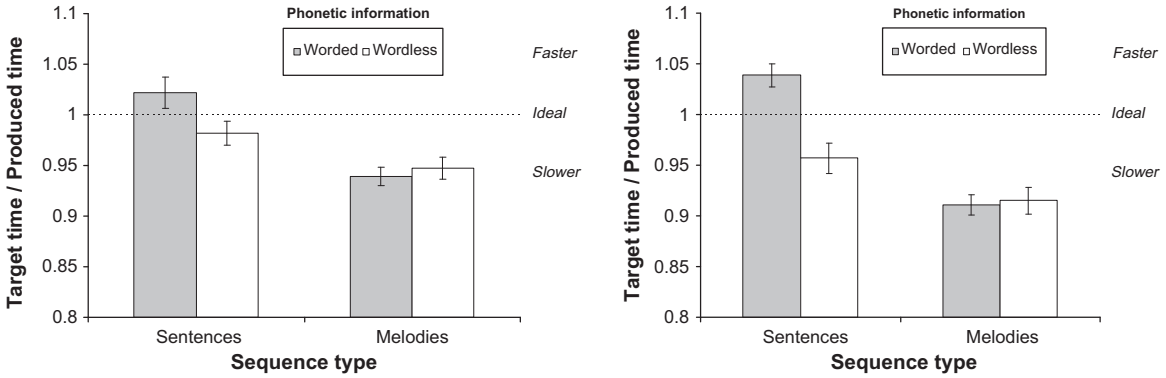

Fig. 6C. Overall rate accuracy in the sequence × phonetics interaction in Exp. 4a (melodic-timed targets, left) and 4b (speech-time targets, right). Error bars represent one standard error of the mean.

sequence × phonetics interaction, Experiment 4a, $F(1,29) = 20.48$, $p < .01$, $\eta_p^2 = .41$, Experiment 4b, $F(1,28) = 44.29$, $p < .01$, $\eta_p^2 = .61$. In addition, Experiment 4b, but not 4a, yielded a significant main effect of phonetic information, $F(1,28) = 13.35$, $p < .01$, $\eta_p^2 = .32$. Results in general resemble those of Experiment 3. Melodies were imitated more slowly than targets, and were not influenced by phonetic information. By contrast, the imitation of sentences was influenced by phonetic information, leading to imitations that were faster than targets for worded sentences, but slower than targets for wordless sentences.

6.3. Discussion

The imitation of pitch in melodies in Experiment 4 was similar to what we found in Experiments 1 and 3. Unexpectedly, the imitation of pitch in sentences for Experiment 4 was similar to Experiment 1 and unlike Experiment 3, in that sentences showed a phonetic advantage and were similar to melodies with respect to the accuracy of relative pitch, despite the fact that sentence stimuli in Experiment 4b were identical to Experiment 3. Of course, this difference could be a chance result. Another possible explanation is that participants in Experiment 3 did not notice differences in relative timing across sentences and melodies given the similar overall durations, and thus imitated sentences using timing more appropriate for melodies. By contrast, the same strategy, similar relative timing for speech and music, would not hinder performance in Experiment 4 when relative timing was constant across domains.

7. Pooled results across experiments

We now consider whether individual differences in vocal imitation within the domain of song correlate with individual differences in the domain of speech. Although certain results varied across experiments, a consistent theme was that there was a domain specific advantage for the imitation of song over speech with respect to absolute pitch matching. The research that has supported the notion of music or language specific modules has focused on deficits within individuals, including music-specific and language-specific deficits (Peretz & Coltheart, 2003). Such perceptually-based individual differences have been taken to support modular processing in general. However, it is not clear whether individual differences in imitative production would provide similar support for the modular-

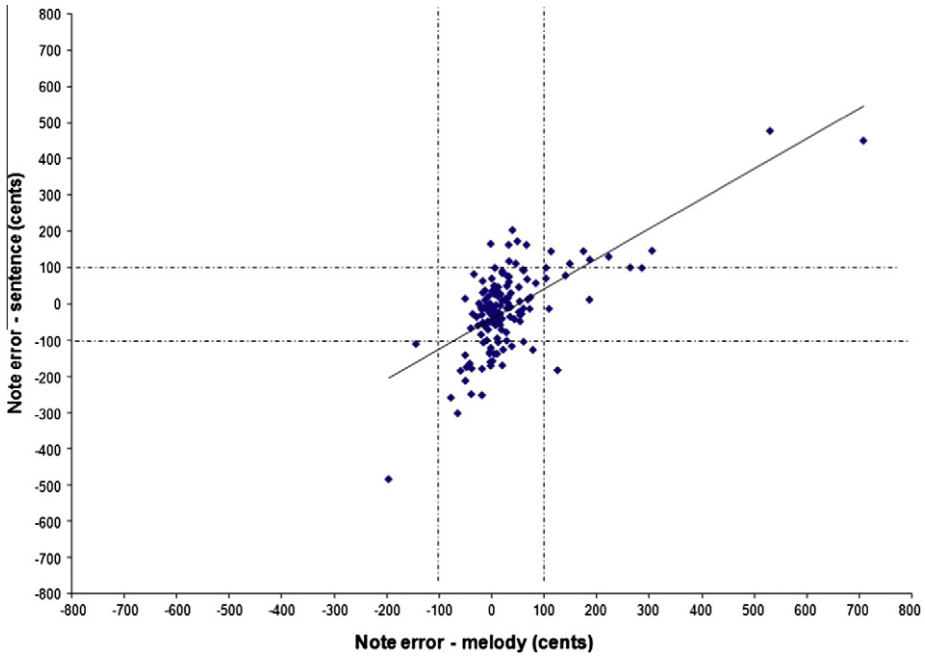

Fig. 7A. Mean pitch error scores (signed error) across all participants in all experiments. Dashed lines highlight boundaries for accurate performance within ±100 cents for each domain. The center square represents accurate performance within ±100 cents.

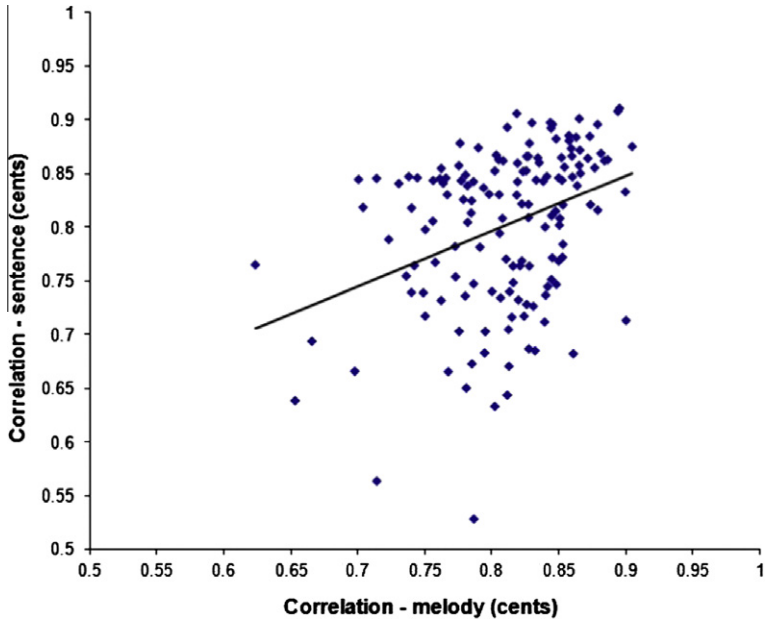

Fig. 7B. Mean pitch correlation for each individual and experiment, across the domains of speech and song.

ity thesis. To date, there is just one report of an individual who exhibited deficient imitation of song but not speech (Dalla Bella, Berkowska, & Sowiński, 2011, p. 8). Here we test whether this single observation generalizes to the large number of participants pooled across the experiments reported here. The group means we have reported so far support domain specificity for the imitation of absolute pitch, but do not support domain specificity with respect to the imitation of relative pitch. Thus, we focus on correlations across speech and song imitation tasks and across experiments on these measures.

Fig. 7A shows the correlation for mean pitch error scores across all participants in all experiments. We show the signed error scores here, which more clearly relate to

Table 1
Correlations (r) across song and speech imitation for each experiment.

Experiment	Pitch accuracy			Pitch correlation		
	All	Word	Wordless	All	Word	Wordless
1	0.86 **	0.82 **	0.85 **	0.16	0.02	0.52 **
2	0.75 **	0.68 **	0.79 **	0.46 **	0.35 *	0.55 **
3	0.59 **	0.59 **	0.50 **	0.58 **	0.24	0.73 **
4a	0.55 **	0.55 **	0.54 **	0.19	0.27	0.36 (*)
4b	0.67 **	0.73 **	0.55 **	0.40 *	0.28	0.49 **

** $p < .01$.
* $p < .05$.
(*) $p = .05$.

accuracy on an individual basis (Pfordresher et al., 2010) but can be problematic when used to summarize group performance given that positive and negative values may cancel each other. The correlation is positive and significant, $r(146) = .68$, $p < .01$; a similarly strong correlation emerged for the mean absolute pitch error scores across participants, $r(146) = .76$, $p < .01$. Outliers are retained here because they are theoretically significant, insofar as poor-pitch singers are typically outliers.[4] Nevertheless, we confirmed that the correlation is still significant when statistical outliers are removed (defined here as those who fall outside two standard deviations of the overall mean), $r(140) = .44$, $p < .01$. Correlations within each experiment, as shown in Table 1, were also positive and significant. With respect to deficits representative of "poor-pitch singing" it is instructive to note that 12% of the total sample would be considered poor-pitch singers given the 100-cent criterion that has been used elsewhere; this margin is similar to what has been found before (Pfordresher & Brown, 2007). Furthermore, if we apply the same criterion to vocal imitation of speech, 61% of poor-pitch singers would also be considered poor-pitch imitators of speech, a significant margin according to a binomial sign test ($p < .01$).

Fig. 7B shows the average pitch correlation measure for each individual and experiment, across the domains of speech and song. This relationship was weaker than the relationship for the absolute pitch accuracy measures, primarily due to the fact that individual differences are not as extreme for the imitation of relative pitch as for the imitation of absolute pitch (cf. Pfordresher & Brown, 2007). However, the relationship was still positive and significant $r(146) = .36$, $p < .01$. With just two exceptions, correlations within experiment (shown in Table 1) were positive and significant. Across speech and song, pitch correlations were generally stronger for wordless than worded trials, with all experiments yielding p of equal to or less than .05.

We further investigated whether phonetic information would modulate correlations between speech and song imitation by separating worded from wordless trials. Given the possibility that articulations are more closely associated with pitch for speech than music (as discussed in the Introduction, Section 1), one might expect that imitation of speech is more "music-like" for the wordless speech trials than for worded speech, leading to higher correlations for wordless than worded trials. Such a pattern was evident for measures of relative pitch accuracy but not for absolute pitch accuracy. When pooling across all experiments, the correlation between speech and song trials for the pitch correlation measure was higher for wordless trials, $r = .54$ than for worded trials, $r = .23$ and this difference was significant according to a z-test for independent r's (Cohen & Cohen, 1983, pp. 53–54), $z = 4.41$, $p < .01$. As can be seen in Table 1, the increased associations for wordless versus worded trials was evident in pitch correlation measures across all experiments, even in Experiment 2 where participants did not use phonetic information while imitating. A similar, but smaller and nonsignificant, trend was found for the pitch error measure (for wordless trials, $r = .67$, for worded trials $r = .65$).

A possible problem with correlational analyses like these is whether a third variable might account for individual differences. One possibility we consider here is whether individual differences reflect the degree of effort exerted by participants, rather than individual differences in imitation ability. In every experiment, participants rated their level of effort on a scale of 1–7 (7 = highest effort), and in Experiments 2–4 the experimenter recorded his or her own subjective rating of each participant's effort (this rating was recording immediately after the experiment ended). We examined correlations between effort ratings and measures of imitation performance for songs and sentences. No correlations with participant self-ratings were significant, nor were correlations of experimenter ratings with pitch error scores (Fig. 7A). However, there was a significant correlation between experimenter effort ratings and pitch correlations for the imitation of songs, $r(120) = .30$, $p < .01$, though not for the imitation of sentences (note that degrees of freedom reflect the absence of data from Experiment 1, for which no experimenter ratings were collected). In order to control for the role of effort in the relationship between pitch correlations for song imitation and pitch correlations for sentence imitation (Fig. 7B), we removed variance associated with experimenter ratings from each variable through linear detrending. The resulting correlation between detrended pitch correlation measures remained significant,

[4] If we take mean signed error of ±100 cents (one semitone) to be the criterion of poor-pitch performance, 18 of 148 subjects (12.16%) qualify as poor-pitch singers. This figure is similar to several previous estimates of poor-pitch singing (15% in Pfordresher & Brown, 2007; 17% in Pfordresher & Brown, 2009; 13% in Pfordresher et al., 2010), but substantially lower than some others (24% in Pfordresher & Mantell, 2009; 40% in Hutchins & Peretz, 2012). If we apply the same music-based criterion to our speech data, 46 subjects (31.08%) in our sample qualify as "poor-pitch speech imitators."

$r(120) = .30$, $p < .01$, suggesting that the relationship shown in Fig. 7B is not an artifact of participant motivation.

8. General discussion

The purpose of the current work was to investigate the accuracy with which individuals imitate the pitch-time trajectories of matched speech and song sequences. Furthermore, we addressed the degree to which the presence of segmental, phonetic information influences vocal imitation of suprasegmental, pitch information. Our primary concern in doing this was to determine whether pitch processing during vocal imitation is coordinated in a modular fashion that is domain specific and/or encapsulated from other information in the signal. In general, results do not support the notion that pitch processing is entirely modular in either respect, with the possible exception that the imitation of musical pitch probably benefits from specialized processing of absolute pitch information. Experiment 1 functioned as a baseline, Experiment 2 addressed how phonetic information contributes to imitation, and Experiments 3 and 4 addressed the way in which temporal characteristics of speech and song contribute to vocal imitation. Table 2 provides a qualitative summary of results across all experiments, for each performance measure. Two especially reliable findings are notable in the table. First, the melodic advantage for absolute pitch accuracy was identified in all four experiments. Second, a phonetic advantage for absolute and relative pitch was identified in three experiments, and in Experiment 3, it applied only for songs. In short, phonetic information improves pitch processing in song imitation.

In this general discussion, we first summarize the results pertaining to the two primary manipulations across experiments, focusing on their implications for the cognitive bases of vocal imitation. We then consider other issues that arise from the current research and their implications for future work in this area.

8.1. Partial support for domain specific pitch processing

As outlined in the introduction, if pitch processing during vocal imitation is domain specific, there should be differences between pitch accuracy for the imitation of song and speech, likely leading to an advantage for song because of the activation of specialized pitch processors reserved for tonal encoding (Peretz & Coltheart, 2003). Indeed, all four experiments produced a song advantage, but only for one of two pitch accuracy measures: mean absolute error (see Table 2). Thus, these results are supportive of claims about specialized processing of absolute pitch information in music.

An important question is why did we fail to observe a reliable advantage for relative pitch in the imitation of song targets? A song advantage for relative pitch was only apparent when imitation was assessed through measures that disregard variability within rhythmic units of the sequence (notes or syllables; Section 3.2.4). When performance measures account for imitation of within-syllable pitch fluctuations, performance differences across domains vanish due to the apparent fact that participants are better able to imitate such fluctuations in speech than in song. One possible theoretical implication of the current findings is that relative pitch processors may be activated by both song and speech input. In fact, this possibility was also suggested by Peretz and Coltheart (2003, p. 689). The present data add further to this claim in demonstrating that the timescale at which imitators are sensitive to pitch fluctuations may vary across domains, possibly due to target signal properties or task demands.

Another way to state this finding is that participants were similarly able to imitate the relative pitch of both speech and song, but they did not necessarily align their average spoken pitch with the average of the sentence targets. It is unlikely that this result emerged from confusions regarding the task because all participants were instructed to imitate the pitch of targets. We also do not believe that this result merely reflects the fact that the pitch structure of sentences is more variable, and thus more complex, than the structure of songs given that participants imitated relative pitch similarly well in both domains. Instead, we suggest this difference reflects the functional significance of relative versus absolute pitch in each domain. Although songs can be reproduced in any key, people often sing in unison which requires pitch matching and in practice people typically sing songs in the key in which songs are most commonly heard (Levitin, 1994). Thus there is reason to believe that absolute pitch is substantively important for music, even if it is not as important as relative pitch. By contrast, the ability to match absolute pitch may be unimportant for speaking (at least for imitating English sentences); only relative pitch conveys meaning for intonation languages like English, the primary language for most of our participants. Even when people speak in unison (e.g., when reciting a pledge or oath) there is no overt attempt to match pitch. Furthermore, even professional impersonators may match relative pitch more faith-

Table 2
Summary of significant results.

	Melodic advantage (YES = across phonetic types)			Phonetic advantage (YES = across sequence types)		
	Timing	Abs. pitch	Rel. pitch	Timing	Abs. pitch	Rel. pitch
Exp. 1		YES		YES	YES	YES
Exp. 2 (always on "ah")	worded only	YES		melodies only		
Exp. 3 (duration equated)		YES	worded only		melodies only	melodies only
Exp. 4 (relative time equated)		YES			YES	YES

fully than absolute pitch while imitating speech (cf. Zetterholm, 2002). However, our results may differ for the production of tone languages. Deutsch, Henthorn, and Dolson (2004) showed that tone language speakers reliably reproduce absolute pitch contents when reciting words on separate occasions. It is possible that this pitch consistency may influence performance in an imitation paradigm such as our own. Furthermore, Deutsch, Le, Shen, and Henthorn (2009) identified distinct spoken pitch levels for individuals in different Chinese villages; this finding suggests that pitch level, at least for tone language speakers, is influenced by developmental linguistic context (for an examination of the relationship between absolute pitch possession and tone language fluency, see Deutsch, Dooley, Henthorn, & Head, 2009).

It is also important to consider that deficits do not always reflect the functioning of an underlying representation. Deficits may be based on deficient representations or on a lack of resources on which these representations rely (cf. Patel, 2003, 2008, pp. 282–285). In this respect it may be the case that individual differences in vocal imitation that lead to "poor-pitch" singing (and speaking) are based on resources whereas the overall advantage for music is based on representational differences across domains that hold for all participants. Why, then, is the advantage for music specific to absolute pitch?

We suggest that sensorimotor associations are influenced by domain specific constraints. Whereas sensorimotor associations for song-like patterns are attuned to both relative and absolute pitch, in speech these associations are weaker for absolute pitch. This proposal emerges in part from the fact that the song advantage was found for both worded and wordless trials. Whereas one could argue that the imitation of worded speech draws on speech-specific processes, it is unlikely that F0 vectors extracted from speech on their own would similarly resonate with such presumed modules, in that the lack of formant transitions leads to the removal of acoustic–phonetic information. In conclusion, the song-specific advantage for absolute pitch appears to be consistent with the prediction afforded by a domain specific tonal pitch processor: song imitation ought to be more accurate than speech imitation. However, we feel that the current results are not consistent with the notion of a tonal encoding module that is both domain specific and encapsulated to information outside of its domain. Our conclusion is based in part on accuracy results from the wordless speech targets and in part on the effects of phonetic information, which we turn to next.

8.2. No support for encapsulated pitch processing

According to numerous cognitive scientists, (Besson & Schön, 2011; Fodor, 1983, 2000; Gibbs & Van Orden, 2010; Prinz, 2006; see also Cowie, 2008; Machery, 2008; Wilson, 2008, for criticisms of 'reduced' modularity), the most characteristic feature of a module is information encapsulation. We probed whether the imitation of pitch is encapsulated with respect to phonetic information by presenting participants with both worded and wordless targets that shared identical pitch-time trajectories. The current results strongly support a consistent advantage for imitation of worded song targets; in several cases, a similar advantage was found for the imitation of spoken pitch as well (see Table 2). Similar to effects related to domain specificity, this result was influenced by the way in which production was measured, and was enhanced for measures that take into account fine-grained fluctuations in pitch within rhythmic units. Thus, the results do not support the notion that pitch processing during vocal imitation of pitch is performed independently of available phonetic information. In other words, if pitch processing is performed by a module, then that module apparently modulates its processing output based on incoming phonetic information. Pitch processing does not appear to be encapsulated from phonetic information.

The results of Experiment 2 are critical in this respect; that was the only experiment in which a phonetic advantage was not found for any task. In Experiment 2, participants were instructed to imitate using a single vowel; they did not imitate the perceived acoustic–phonetic information. Importantly, the data from Experiment 2 did not suggest that the presence of to-be-ignored phonetic information interfered with production, in that pitch matching on worded trials was equivalent to, rather than below, performance on wordless trials. Rather, these results suggest that the phonetic advantage seen in the other experiments is related to congruence between the target (in terms of both pitch and phonetic information) and motor gestures in production (for a demonstration of perceptuomotor compatibility in speech, see Galantucci, Fowler, & Goldstein, 2009).

We found some support for the idea that vocal imitation of speech is more greatly integrated with phonetic information than vocal imitation of song, but that support was limited to one measure of accuracy and only in one experiment: the imitation of relative pitch in Experiment 1. There we found an interaction of sequence and phonetics in which performance across song and speech imitation was equivalent for word trials but the decrement for wordless trials was greater for the imitation of speech than for song. However, this result was not robust to manipulations of timing. Further, it is not entirely consistent with the modular architecture of Peretz and Coltheart (2003), which predicts no facilitative effect of phonetic information on song imitation. Of course, it is possible that Peretz and Coltheart's model could be revised to account for these results by adding an information exchange arrow between the phonetic and tonal processing modules. However, such an inclusion would prevent the pitch processors from being described as information encapsulated because the processor would accept and use information from outside of its defined domain. Further, that modification would also question the description of the processing modules as domain-specific. It seems clear that a module that processes both pitch and phonetic information is not specific to either domain. Another possibility suggested from the current work is that speech-specific processing during imitation is highly sensitive to the naturalness of the speech

signal. In Experiments 3–4 the overall durations of song and speech trials were equated. Although the timing of speech was not manipulated as strongly as was the timing of song, there is evidence suggesting that speech timing is less flexible than the timing of music performance (Jungers, Palmer, & Speer, 2002). This leads us to our next point: the role of timing in the vocal imitation of pitch.

8.3. Temporal influences on the imitation of pitch

Although our main focus in this article was on the imitation of pitch, timing figured into this research in two important respects. First, we measured imitation of timing as well as pitch and found that manipulations of sequence type and phonetic information influenced the timing of imitations. In general, patterns of accuracy in imitating duration were not significantly related to patterns of accuracy in imitating pitch across experiments and conditions. Second, we manipulated the temporal characteristics of speech and song in Experiments 3 and 4 to determine whether temporal characteristics associated with the absolute timing of speech and song targets influence pitch imitation. Although some effects of these temporal manipulations emerged, the principal effects (melodic advantage for absolute pitch matching and phonetic advantage) of interest remained unchanged. Taken together, these results suggest that the imitation of pitch functions independently from the imitation of timing (cf. Dalla Bella et al., 2007, 2009; Drake & Palmer, 2000). However, there were additional unanticipated effects of target timing that are more complex. Whereas the phonetic advantage seen in Experiment 1 was maintained across Experiments 4a and 4b, wherein both absolute and relative timing for speech and song were equated, the phonetic advantage for speech disappeared in Experiment 3 for both the imitation of relative and absolute pitch. Imitators apparently respond in subtle ways to both absolute and relative timing of sequences, and they are particularly sensitive to the appropriateness of both forms of timing in speech.

We analyzed timing in performance by assessing the degree to which imitations were faster or slower than the original target, independently of how fast or slow the target was. Interestingly, whereas performance in wordless trials slowed down for wordless speech imitation versus worded speech, performance in wordless song trials sped up relative to worded song trials. This finding is important in two respects. First, it suggests that the phonetic advantage cannot be entirely due to a speed/accuracy tradeoff (a point we will return to shortly). Second, it suggests that timing in wordless trials may drift toward a common neutral pattern. Recall that speech targets were initially faster than song targets, reflective of these domains in real-world settings (see Appendix B). Thus, the opposite pattern of deterioration in relative timing suggests that performance in wordless trials was more similar in absolute terms.

An important issue to consider in any performance domain, including vocal imitation, is the speed/accuracy tradeoff. Indeed, one reason why we equated target rate in Experiments 3–4 was to determine if the melodic advantage for imitation of absolute pitch could be attributed to this tradeoff. If there were effects of timing that suggested faster imitation of speech than song, there is a possibility that the song advantage for imitation of absolute pitch reflects a speed/accuracy tradeoff. Thus, we assessed speed/accuracy relationships (for mean pitch error) across all experiments by using mean values for timing and pitch for each experiment, sequence type, and phonetic information condition ($n = 20$ because for each of the five separate experiments, there were four conditions: worded melodies, wordless melodies, worded sentences, wordless sentences). Timing values were transformed from duration ratios to mean overall durations using the values reported in Appendix B; for Experiments 3 and 4, duration ratio means were adjusted by the appropriate percentage for duration difference (see Section 5.1.2). In general, there was a significant negative correlation between mean durations and mean absolute pitch error, $r(18) = -.51$, $p < .05$, though not between durations and pitch correlations ($r = .13$, n.s.). Based on the speed/accuracy tradeoff for pitch error, we tested whether the song advantage for pitch exists independently of this relationship by removing the linear trend associated with this speed/accuracy relationship from mean pitch error scores. The mean difference between speech and song imitation conditions was still significant after this adjustment, $t(9) = 4.71$, $p < .01$. Thus, in general, it appears that the advantage for singing cannot simply be attributed to differences in the rate at which people produce imitations.

8.4. Prospects for future research

To our knowledge, this is one of the first studies to explore systematically the imitation of sentences and melodies that are designed to share critical features in common. Yet, the current project is not entirely unprecedented. Racette, Bard, and Peretz (2006) recorded brain-damaged aphasics singing and saying the words to familiar and novel songs. Although Racette et al. were primarily interested in the number of words recalled, they also measured the number of musical notes correctly recalled. When aphasic individuals performed familiar melodies, "sung notes were more accurate when sung with words than on /la/ in recall" (Racette et al., 2006, p. 2576). However, there was no effect on note accuracy when the subjects sang newly learned melodies. In related work, Racette and Peretz (2007) tested the recall abilities of university students in three different song learning conditions. The authors wanted to know if performing a song's lyrics and melody simultaneously would lead to superior word recall (it did not). However, unlike Racette et al. (2006), Racette and Peretz reported that "The percentages of notes correctly recalled in singing with and without words did not differ" (2007, p. 249). There are important differences between these related studies and our own. First, we found a reliable worded advantage for pitch accuracy in individuals singing novel melodies. Second, our quantitative mea-

sures of pitch accuracy are much more fine-grained than the subjective measures used by Racette et al. (2006) and Racette and Peretz (2007). Finally, we are measuring pitch accuracy, not the number of notes correctly recalled. Thus, the novelty of our work and the results suggest several new questions as well as conclusions.

One issue has to do with the degree to which the phonetic advantage we found merely results from articulations in production, as opposed to the generation of meaningful linguistic content. In this context, we highlight a difference between the present results and some recent studies that have compared singing melodies with lyrics to singing melodies on the syllable "la" (Berkowska & Dalla Bella, 2009; Racette & Peretz, 2007; Racette et al., 2006). Unlike the current results (with one exception in one experiment from Racette et al., 2006), those studies yielded no effect of phonetic information (Racette et al., 2006, Experiment 2; Racette & Peretz, 2007, Experiment 2) or even degraded production with words (Berkowska & Dalla Bella, 2009). Comparisons across different measures of performance in the current Experiment 1 suggest that one factor relating to these different results could be the measure of production. By this account, facilitative effects of phonetic information appear when measures take into account fine-grained pitch fluctuations because it is these fluctuations that are most closely related to articulatory variations. This explanation could explain the null effect of words reported by Racette and Peretz (2007), but not the interfering effect found by Berkowska and Dalla Bella (2009). One possible further difference has to do with the fact that melodies used in those studies were longer and more familiar than the sequences our participants sang. Future research should continue to address the influence of melody length and familiarity with respect to phonetic facilitation.

Another question has to do with whether our imitation task encouraged more "music-like" processing of speech sequences. Although domain specific differences were perceptually salient (in wordless as well as worded trials), it has been shown that the neural processing of speech can shift from left-lateralized (more speech-like processing) to right-lateralized activations depending on the context, such as linguistic function (Wong, Parsons, Martinez, & Diehl, 2004; Zatorre, Evans, Meyer, & Gjedde, 1992). One could argue that this claim runs counter to the current data, in that we found strong domain specific differences in the imitation of absolute pitch for wordless sequences whereas the context account may be taken to predict a null result in the absence of obvious phonetic context. However, it is possible that sequence complexity may have influenced performance beyond any possible influence expected to result from lateralized neural processing. Although we controlled for temporal aspects of complexity and overall pitch contour across speech and music, speech sequences in these experiments included more variation in pitch.

With few exceptions in previous literature (e.g., d'Alessandro et al., 2011), the sequence-length pitch accuracy measurements used in this paper are novel. Up to now, most research on pitch accuracy in song production has emphasized mean or median pitch within each note. Our decision to use sequence-length measurements was contingent on two major criteria. First, pitch-time information in speech syllables is typically more variable than in song notes (Stegemöller et al., 2008), and a central tendency, by-note analysis disregards this variability. An ideal accuracy measure for comparing speech and song would have to account for the increased variability in speech. Second, our experiments introduced wordless pitch-time trajectories that were synthesized from their worded counterparts. Wordless targets do not contain any phonetically-delineated syllable boundaries, and neither do their imitations (produced on the syllable "ah"). Thus, word syllables cannot be used to segment the wordless imitative productions, and central-tendency accuracy measures would be impossible to apply.

We believe that our sequence-length accuracy measures, which emphasize variability within syllable and note productions, are more informative than traditional note-based methods. However, this informational benefit does not come without cost. Because our measures incorporate information from the entire pitch-time trajectory, they are sensitive to more variables than traditional note-based methods. One such variable is produced timing—the duration of syllables and notes in the imitation. It is reasonable to assume that timing errors could upset the accuracy of our sequence-length measures because they could shift the imitation trajectory relative to the target. We addressed this issue by performing syllable-aligned sequence accuracy analyses for the worded productions in Experiment 1 (see Section 3.2.3). However, the newly aligned data did not produce different results, suggesting that our original measures are not biased or overly sensitive to timing errors (possibly because all of our sequences were limited to five syllables or less). Still, researchers should be cautious about the possibility of temporal contamination in sequence-length pitch accuracy analyses, especially if they use longer vocal sequences. Future work should attempt to expand and improve upon our approach for aligning and comparing target and imitation pitch-time trajectories.

9. Conclusions

The ability to vocally imitate the pattern of change in pitch across time within a sequence requires the translation of perceptual information (stored in working memory) into motor planning (Pfordresher & Mantell, 2009). We assessed how well people can do this for both melodies and sentences, matched for linguistic content and pitch contour, and whether the presence of phonological information (phonetic content) modulates this ability. Our primary interest in doing this was to determine whether vocal imitation of pitch incorporates domain specific mechanisms, and whether pitch processing is encapsulated

to phonetic information during vocal imitation. Results in general support the view that vocal imitation is integrative rather than modular, and that imitation abilities in one domain (e.g., song) predict imitation in another domain (e.g., speech). As highlighted in the introduction, many vocal forms blur the line between speech and song; the current work expands this notion from perception into production and contributes to the ongoing debate on the application of modularity concepts in the domains of speech and song.

Author's note

Some of the data from the studies described in this paper were previously presented as posters, presentations, and proceedings. Specifically, experiments 1–2 were presented in part at a poster session at 7th annual Auditory, Perception, Cognition, and Action (APCAM) meeting in 2008 in Chicago, IL. Some of the findings from experiments 1, 2, and 4 were delivered as a presentation at the Society for Music Perception and Cognition (SMPC) 2009 biennial conference in Indianapolis, Indiana and also briefly introduced in the proceedings of the seventh triennial conference of the European Society for the Cognitive Sciences of Music (ESCOM) 2009 in Jyväskylä, Finland. This work was presented in part at the August 2010 11th International Conference on Music Perception and Cognition (ICMPC) in Seattle, Washington.

Acknowledgments

The authors wish to express their grateful thanks to several research assistants. Jennifer Walsh assisted with stimuli preparation and data collection in Experiment 1. Ece Yildirim and David Ricotta assisted with data collection in Experiments 1 and 2. Marianna Sobczak ran a pilot of Experiment 4, and Rebecca O'Connor assisted in data collection and analysis of Experiments 3 and 4. The authors greatly appreciate the helpful critical remarks from four anonymous reviewers. This work was funded in part by NSF Grant BCS-0642592.

Appendix A

Text for target stimuli

1 She was here
2 They went home
3 He ate it all
4 He lost his boots
5 She bought apples
6 She parked the car
7 She wrote a book
8 He ran a mile
9 He washed the dishes
10 They finished the test
11 They forgot her name
12 They went to the store

Appendix B

Seq #	Mean duration (s)		Syllable duration nPVI		Coefficient of variation*		Mean pitch in Hz (cents)			Pitch range in Hz		Pitch SD in Hz	
	Sentences	Melodies	Sentences	Melodies	Sentences	Melodies	Sentences	Melodies	Difference	Sentences	Melodies	Sentences	Melodies
Exp. 1 stimulus parameter table, male values averaged across contour type													
1	0.97	2.25	38.54	12.82	0.49	0.15	153.70	108.18	(−608.1)	83.20	32.10	30.86	12.51
2	1.05	1.95	49.20	35.95	0.58	0.30	130.76	107.22	(−343.67)	80.19	28.24	32.49	12.00
3	0.94	2.25	66.01	20.58	0.65	0.23	131.64	116.99	(−206.93)	83.61	55.41	28.56	18.73
4	1.27	2.10	79.95	14.03	0.56	0.14	127.81	120.90	(−96.24)	92.31	52.54	27.61	20.12
5	1.06	2.58	25.56	17.39	0.30	0.18	126.46	119.37	(−100.02)	136.93	54.32	43.79	20.61
6	1.04	2.56	55.75	23.70	0.35	0.24	133.72	121.49	(−166.13)	69.22	55.47	24.19	20.62
7	1.06	2.24	32.99	4.35	0.30	0.05	142.58	117.08	(−341.14)	70.51	53.05	23.13	18.17
8	1.05	2.47	124.71	39.45	0.92	0.32	129.43	113.68	(−224.57)	81.49	39.84	25.29	14.68
9	1.16	2.85	58.53	19.97	0.51	0.18	133.26	120.73	(−170.93)	137.06	52.89	38.76	19.34
10	1.19	3.02	51.76	21.75	0.63	0.16	134.10	121.00	(−178)	108.81	56.03	36.37	19.23
11	1.34	2.94	69.45	16.36	0.77	0.21	134.22	120.37	(−186.56)	83.64	51.27	26.23	18.34
12	1.22	3.19	79.04	39.11	0.82	0.33	131.86	120.48	(−156.16)	111.78	57.99	36.08	19.57
Average	1.11	2.53	60.96	22.12	0.57	0.21	134.14	117.29	(−232.47)	95.31	49.10	31.11	17.83
t test	$p < 0.01$		$p < 0.01$		$p < 0.01$		$p < 0.01$			$p < 0.01$		$p < 0.01$	
Exp. 1 stimulus parameter table, female values averaged across contour type													
1	0.97	1.78	19.30	12.27	0.22	0.10	236.99	248.01	(78.67)	61.10	67.42	21.63	27.17
2	1.00	1.88	39.39	22.27	0.50	0.12	223.25	247.73	(180.2)	49.85	62.17	16.05	25.68

(continued on next page)

Appendix B (*continued*)

Seq #	Mean duration (s)		Syllable duration nPVI		Coefficient of variation*		Mean pitch in Hz (cents)			Pitch range in Hz		Pitch SD in Hz	
	Sentences	Melodies	Sentences	Melodies	Sentences	Melodies	Sentences	Melodies	Difference	Sentences	Melodies	Sentences	Melodies
3	1.08	2.43	49.33	17.83	0.41	0.17	221.64	257.78	(261.53)	55.81	89.31	19.26	30.24
4	1.29	2.12	80.66	20.25	0.57	0.16	228.41	248.22	(143.98)	68.49	68.66	21.73	25.93
5	1.36	3.01	34.73	29.02	0.29	0.31	226.59	252.61	(206.34)	127.32	88.69	40.04	30.58
6	1.21	2.54	53.64	28.75	0.42	0.20	226.16	254.79	(221.09)	41.32	114.87	14.58	37.26
7	1.13	2.61	33.80	35.39	0.55	0.23	219.93	249.89	(221.09)	38.19	65.55	14.19	24.62
8	1.08	2.45	116.29	30.58	0.85	0.20	228.97	252.51	(169.39)	51.96	71.46	16.61	26.51
9	1.29	3.10	77.76	21.21	0.55	0.21	222.33	268.57	(327.17)	152.61	109.75	41.68	38.14
10	1.34	2.72	62.46	35.61	0.70	0.26	231.23	258.88	(195.57)	65.28	88.10	18.82	30.24
11	1.35	2.96	40.83	25.73	0.44	0.30	233.05	258.25	(177.73)	56.84	71.01	16.49	27.99
12	1.20	2.67	66.99	28.44	0.79	0.28	218.68	259.26	(294.7)	51.01	76.86	15.26	28.32
Average	1.19	2.52	60.43	25.61	0.52	0.21	226.44	254.71	(203.7)	68.32	81.15	21.36	29.39
t test		$p < 0.01$		$p < 0.01$		$p < 0.01$		$p < 0.01$		$p = 0.28$		$p = 0.02$	

*CV for syllable duration.

References

Ayotte, J., Peretz, I., & Hyde, K. (2002). Congenital amusia: A group study of adults afflicted with a music-specific disorder. *Brain, 125*(2), 238–251. http://dx.doi.org/10.1093/brain/awf028.

Barrett, H. C., & Kurzban, R. (2006). Modularity in cognition: Framing the debate. *Psychological Review, 113*(3), 628–647. http://dx.doi.org/10.1037/0033-295X.113.3.628.

Berkowska, M., & Dalla Bella, S. (2009). Reducing linguistic information enhances singing proficiency in occasional singers. *The Neurosciences and Music III–Disorders and Plasticity: Annals of the New York Academy of Sciences, 1169*, 108–111. doi:10.1111/j.1749-6632.2009.04774.x.

Besson, M., & Schön, D. (2011). What remains of modularity? In P. Rebuschat, M. Rohmeier, J. Hawkins, & I. Cross (Eds.), *Language and music as cognitive systems*. New York: Oxford University Press.

Boersma, P., & Weenink, D. (2009). *Praat: Doing phonetics by computer* (Version 5.1) [Computer software]. <http://www.praat.org/>.

Carruthers, P. (2006a). *The architecture of the mind*. New York: Oxford University Press.

Carruthers, P. (2006b). The case for massively modular models of mind. In R. Stainton (Ed.), *Contemporary debates in cognitive science* (pp. 3–21). Malden, MA: Blackwell.

Carruthers, P. (2008). On Fodor-fixation, flexibility, and human uniqueness: A reply to Cowie, Machery, and Wilson. *Mind & Language, 23*(2), 293–303. http://dx.doi.org/10.1111/j.1468-0017.2008.00344.x.

Callan, D. E., Tsytsarev, V., Hanakawa, T., Callan, A. M., Katsuhara, M., Fukuyama, H., et al. (2006). Song and speech: Brain regions involved with perception and covert production. *NeuroImage, 31*, 1327–1342. http://dx.doi.org/10.1016/j.neuroimage.2006.01.036.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum.

Coltheart, M. (1999). Modularity and cognition. *Trends in Cognitive Sciences, 3*(3), 115–120. http://dx.doi.org/10.1016/S1364-6613(99)01289-9.

Cowie, F. (2008). Us, them, and it: Modules, genes, environments and evolution. *Mind & Language, 23*(3), 284–292. http://dx.doi.org/10.1111/j.1468-0017.2008.00342.x.

Curtis, M. W., & Bharucha, J. J. (2010). The minor third communicates sadness in speech, mirroring its use in music. *Emotion, 10*(3), 335–348. http://dx.doi.org/10.1037/a0017928.

Dalla Bella, S., Berkowska, M., & Sowiński, J. (2011). Disorders of pitch production in tone deafness. *Frontiers in Psychology, 2*(164). http://dx.doi.org/10.3389/fpsyg.2011.00164.

Dalla Bella, S., Giguère, J. F., & Peretz, I. (2007). Singing proficiency in the general population. *Journal of the Acoustical Society of America, 121*(2), 1182–1189. http://dx.doi.org/10.1121/1.2427111.

Dalla Bella, S., Giguère, J. F., & Peretz, I. (2009). Singing in congenital amusia. *Journal of the Acoustical Society of America, 126*(1), 414–424. http://dx.doi.org/10.1121/1.3132504.

d'Alessandro, C., Rilliard, A., & Le Beux, S. (2011). Chironomic stylization of intonation. *The Journal of the Acoustical Society of America, 129*, 1594–1604. http://dx.doi.org/10.1121/1.3531802.

Deutsch, D., Dooley, K., Henthorn, T., & Head, B. (2009). Absolute pitch among students in an American music conservatory: Association with tone language fluency. *Journal of the Acoustical Society of America, 125*(4), 2398–2403. http://dx.doi.org/10.1121/1.3081389.

Deutsch, D., Henthorn, T., & Dolson, M. (2004). Absolute pitch, speech, and tone language: Some experiments and a proposed framework. *Music Perception, 21*(3), 339–356. http://dx.doi.org/10.1525/mp. 2004, 21(3), pp. 339.

Deutsch, D., Henthorn, T., & Lapidis, R. (2011). Illusory transformation from speech to song. *Journal of the Acoustical Society of America, 129*(4), 2245–2252. http://dx.doi.org/10.1121/1.3562174.

Deutsch, D., Lapidis, R., & Henthorn, T. (2008). The speech-to-song illusion. *Journal of the Acoustical Society of America, 124*, 2471. <http://link.aip.org/link/?JAS/124/2471/2>.

Deutsch, D., Le, J., Shen, J., & Henthorn, T. (2009). The pitch levels of female speech in two Chinese villages. *Journal of the Acoustical Society of America, 125*(5), EL208–EL213. doi:10.1121/1.3113892.

Drake, C., & Palmer, C. (2000). Skill acquisition in music performance. Relations between planning and temporal control. *Cognition, 74*(1), 1–32. http://dx.doi.org/10.1016/S0010-0277(99)00061-X.

Falk, S., & Rathcke, T. (2010). The speech-to-song-illusion: Empirical findings. In S. M. Demorest, S. J. Morrison, & P. S. Campbell (Eds.), *Proceedings of the 11th international conference on music perception and cognition* (pp. 338–339).

Feld, S., & Fox, A. A. (1994). Music and language. *Annual Review of Anthropology, 23*, 25–53. http://dx.doi.org/10.1146/annurev.an.23.100194.000325.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.

Fodor, J. (2000). *The mind doesn't work that way: The scope and limits of computational psychology*. Cambridge, MA: MIT Press.

Galantucci, B., Fowler, C. A., & Goldstein, L. (2009). Perceptuomotor compatibility effects in speech. *Attention, Perception, & Psychophysics, 71*(5), 1138–1149. http://dx.doi.org/10.3758/APP.71.5.1138.

Gibbs, R. W., Jr., & Van Orden, G. C. (2010). Adaptive cognition without massive modularity. *Language and Cognition, 2*(2), 149–176. http://dx.doi.org/10.1515/LANGCOG.2010.006.

Ginsborg, J., & Sloboda, J. A. (2007). Singers' recall for the words and melody of a new, unaccompanied song. *Psychology of Music, 35*(3), 421–440. http://dx.doi.org/10.1177/0305735607072654.

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

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*(1), 76–97. http://dx.doi.org/10.1037/a0025064.

Jackendoff, R. (2009). Parallels and nonparallels between language and music. *Music Perception, 26*(3), 195–204. http://dx.doi.org/10.1525/mp. 2009, 26(3), pp. 195.

Jackendoff, R., & Lerdahl, F. (2006). The capacity for music: What is it, and what's special about it? *Cognition, 100*, 33–72. http://dx.doi.org/10.1016/j.cognition.2005.11.005.

Johnson, K. (2003). *Acoustic and auditory phonetics* (2nd ed.). Cornwall, UK: Blackwell Publishing.

Jones, M. R., Boltz, M., & Kidd, G. (1982). Controlled attending as a function of melodic and temporal context. *Perception & Psychophysics, 32*(3), 211–218. http://dx.doi.org/10.3758/BF03206225.

Jones, M. R., Moynihan, H., MacKenzie, N., & Puente, J. (2002). Temporal aspects of stimulus-driven attending in dynamic arrays. *Psychological Science, 13*(4), 313–319. http://dx.doi.org/10.1111/1467-9280.00458.

Jungers, M. K., Palmer, C., & Speer, S. R. (2002). Time after time: The coordinating influence of tempo in music and speech. *Cognitive Processing, 1-2*, 21–35. <http://francais.mcgill.ca/files/spl/jungers02.pdf>.

Koelsch, S. (2011). Toward a neural basis of music perception—A review and updated model. *Frontiers in Psychology, 2*, 1–20. http://dx.doi.org/10.3389/fpsyg.2011.00110.

Krumhansl, C., & Iverson, P. (1992). Perceptual interactions between musical pitch and timbre. *Journal of Experimental Psychology: Human Perception and Performance, 18*(3), 739–751. http://dx.doi.org/10.1037/0096-1523.18.3.739.

Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review, 89*(4), 334–368. http://dx.doi.org/10.1037/0033-295X.89.4.334.

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences, 97*(22), 11850–11857. <http://www.pnas.org/content/97/22/11850.abstract?sid=7f296f91-f5b9-4bf1-8b3e-7d3054b67526>.

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*(4), 2425–2438. http://dx.doi.org/10.1121/1.417951.

Levitin, D. J. (1994). Absolute memory for musical pitch: Evidence from the production of learned melodies. *Perception & Psychophysics, 56*(4), 414–423. http://dx.doi.org/10.3758/BF03206733.

List, G. (1963). The boundaries of speech and song. *Ethnomusicology, 7*(1), 1–16. http://dx.doi.org/10.2307/924141.

Machery, E. (2008). Massive modularity and the flexibility of human cognition. *Mind & Language, 23*(3), 263–272. http://dx.doi.org/10.1111/j.1468-0017.2008.00341.x.

Marin, O. S. M., & Perry, D. W. (1999). Neurological aspects of music perception and performance. In D. Deutsch (Ed.), *The psychology of music* (pp. 643–724). Sand Diego, CA: Academic Press.

Melara, R. D., & Marks, L. E. (1990). Interaction among auditory dimensions: Timbre, pitch, and loudness. *Perception & Psychophysics, 48*(2), 169–178. http://dx.doi.org/10.3758/BF03207084.

Mertens, P. (2004). The prosogram: Semi-automatic transcription of prosody based on a tonal perception model. In B. Bel & I. Marlien (Eds.), *Proceedings of speech prosody* 2004, Nara, Japan, March 23–26. <http://bach.arts.kuleuven.be/pmertens/papers/sp2004.pdf>.

Nielsen, K. Y. (2005). Generalization of phonetic imitation across place of articulation. In *Proceedings from ICSA workshop on plasticity in speech perception* (PSP2005) (pp. 47–50). London, UK. <http://www.linguistics.ucla.edu/people/grads/kuniko/index_files/nielsenday1.pdf>.

Nielsen, K. Y. (2007). Implicit phonetic imitation is constrained by phonemic contrast. In *Proceedings from 16th International Congress of Phonetic Sciences* (ICPhS XVI) (pp. 1961–1964). Saarbrücken, Germany. <http://www.icphs2007.de/conference/Papers/1641/1641.pdf>.

Özdemir, E., Norton, A., & Schlaug, G. (2006). Shared and distinct neural correlates of singing and speaking. *NeuroImage, 33*, 628–635. http://dx.doi.org/10.1016/j.neuroimage.2006.07.013.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America, 119*(4), 2382–2393. http://dx.doi.org/10.1121/1.2178720.

Palmer, C., & Kelly, M. H. (1992). Linguistic prosody and musical meter in song. *Journal of Memory and Language, 31*(4), 525–542. http://dx.doi.org/10.1016/0749-596X(92)90027-U.

Patel, A. D. (2003). Language, music, syntax, and the brain. *Nature Neuroscience, 6*(7), 674–681. http://dx.doi.org/10.1038/nn1082.

Patel, A. D. (2008). *Music, language, and the brain*. New York: Oxford University Press.

Patel, A. D., Peretz, I., Tramo, M., & Labreque, R. (1998). Processing prosodic and musical patterns: A neuropsychological investigation. *Brain and Language, 61*(1), 123–144. http://dx.doi.org/10.1006/brln.1997.1862.

Peretz, I. (2006). The nature of music from a biological perspective. *Cognition, 100*, 1–32. http://dx.doi.org/10.1016/j.cognition.2005.11.004.

Peretz, I. (2009). Music, language, and modularity framed in action. *Psychologica Belgica, 49*, 157–175. <http://www.ingentaconnect.com/content/acad/psyb/2009/00000049/F0020002/art00007>.

Peretz, I., & Coltheart, M. (2003). Modularity of music processing. *Nature Neuroscience, 6*(7), 688–691. http://dx.doi.org/10.1038/nn1083.

Peretz, I., & Morais, J. (1989). Music and modularity. *Contemporary Music Review, 4*(1), 279–293. http://dx.doi.org/10.1080/07494468900640361.

Peretz, I., & Zatorre, R. J. . (2005). Brain organization for music processing. *Annual Review of Psychology, 56*, 89–114. http://dx.doi.org/10.1146/annurev.psych.56.091103.070225.

Pfordresher, P. Q., & Brown, S. (2007). Poor-pitch singing in the absence of "tone deafness". *Music Perception, 25*(2), 95–115. doi:10.1525/mp.2007.25.2.95.

Pfordresher, P. Q., & Brown, S. (2009). Linguistic background influences the production and perception of musical intervals. *Attention, Perception, & Psychophysics, 71*, 1385–1398. http://dx.doi.org/10.3758/APP.71.6.1385.

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. <http://www.acsu.buffalo.edu/~pqp/pdfs/Pfordresher_etal_2010_JASA.pdf>.

Pfordresher, P. Q., & Mantell, J. T. (2009). Singing as a form of vocal imitation: Mechanisms and deficits. In J. Louhivuori, T. Eerola, S. Saarikallio, T. Himberg, & P. -S. Eerola (Eds.), *Proceedings of the 11th international conference on music perception and cognition* (pp. 821–824). <http://urn.fi/URN:NBN:fi:jyu-2009411309>.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*(2), 169–190. http://dx.doi.org/10.1017/S0140525X04000056.

Pierrehumbert, J. B. (1980/87). *The phonology and phonetics of English intonation*. Ph.D thesis, Massachusetts Institute of Technology. Published by Indiana University Linguistics Club (1987).

Pinker, S. (1997). *How the mind works*. New York, NY: Norton.

Pitt, M. A. (1994). Perception of pitch and timbre by musically trained and untrained listeners. *Journal of Experimental Psychology: Human Perception and Performance, 20*(5), 976–986. http://dx.doi.org/10.1037//0096-1523.20.5.976.

Prince, J. B., Schmuckler, M. A., & Thompson, W. F. (2009). The effect of task and pitch structure on pitch-time interactions in music. *Memory and Cognition, 37*(3), 368–381. http://dx.doi.org/10.3758/MC.37.3.368.

Prinz, J. J. (2006). Is the mind really modular? In R. Stainton (Ed.), *Contemporary debates in cognitive science* (pp. 22–36). Malden, MA: Blackwell.

Racette, A., Bard, C., & Peretz, I. (2006). Making non-fluent aphasics speak: Sing along! *Brain, 129*, 2571–2584. http://dx.doi.org/10.1093/brain/awl250.

Racette, A., & Peretz, I. (2007). Learning lyrics: To sing or not to sing? *Memory & Cognition, 35*, 242–253. <http://mc.psychonomic-journals.org/content/35/2/242.abstract>.

Robbins, P. (2010). Modularity of mind. In E. Zalta (Ed.), *The stanford encyclopedia of philosophy (Summer 2010 Edition)*. <http://plato.stanford.edu/archives/sum2010/entries/modularity-mind/>.

Russo, F. A., & Thompson, W. F. (2005). An interval-size illusion: The influence of timbre on the perceived size of melodic intervals. *Perception & Psychophysics, 67*, 559–568. http://dx.doi.org/10.3758/BF03193514.

Saito, Y., Ishii, K., Yagi, K., Tatsumi, I. F., & Mizusawa, H. (2006). Cerebral networks for spontaneous and synchronized singing and speaking. *NeuroReport, 17*(18), 1893–1897. http://dx.doi.org/10.1097/WNR.0b013e328011519c.

Sammler, D., Koelsch, S., Ball, T., Brandt, A., Elger, C. E., Friederici, A. D., et al. (2009). Overlap of musical and linguistic syntax processing: Intracranial ERP evidence. *The Neurosciences and Music III—Disorders and Plasticity: Annals of the New York Academy of Sciences, 1169*, 494–498. doi:10.1111/j.1749-6632.2009.04792.x.

Schön, D., Gordon, R., Campagne, A., Magne, C., Astésano, C., Anton, J.-L., et al. (2010). Similar cerebral networks in language, music and song perception. *NeuroImage, 51*, 450–461. http://dx.doi.org/10.1016/j.neuroimage.2010.02.023.

Serafine, M. L., Crowder, R. G., & Repp, B. H. (1984). Integration of melody and text in memory for songs. *Cognition, 16*, 285–303. http://dx.doi.org/10.1016/0010-0277(84)90031-3.

Serafine, M. L., Davidson, J., Crowder, R. G., & Repp, B. H. (1986). On the nature of melody-text integration in memory for songs. *Journal of Memory and Language, 25*, 123–135. http://dx.doi.org/10.1016/0749-596X(86)90025-2.

Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception and Psychophysics, 66*(3), 422–429. <http://app.psychonomic-journals.org/content/66/3/422.abstract>.

Stegemöller, E. L., Skoe, E., Nicol, T., Warrier, C. M., & Kraus, N. (2008). Music training and vocal production of speech and song. *Music Perception, 25*(5), 419–428. http://dx.doi.org/10.1525/MP.2008.25.5.419.

Sundberg, J. (1999). The perception of singing. In D. Deutsch (Ed.), *The psychology of music* (2nd ed., pp. 171–214). San Diego: Academic Press. http://dx.doi.org/10.1016/B978-012213564-4/50007-X.

The MathWorks, Inc. (2006). *MATLAB* (Version R2006a) [Computer software].

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In J. Barkow, L. Codmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19–136). New York: Oxford University Press.

Wallace, W. T. (1994). Memory for music: Effect of melody on recall of text. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), 1471–1485. http://dx.doi.org/10.1037/0278-7393.20.6.1471.

Warrier, C. M., & Zatorre, R. J. (2002). Influence of tonal context and timbral variation on perception of pitch. *Perception & Psychophysics, 64*(2), 198–207. http://dx.doi.org/10.3758/BF03195786.

Welch, G. F. (2005). Singing as communication. In D. Miell, R. MacDonald, & D. J. Hargreaves (Eds.), *Musical communication* (pp. 239–260). Oxford: Oxford University Press.

Wilson, R. A. (2008). The drink to have when you're not having a drink. *Mind & Language, 23*(3), 273–283. http://dx.doi.org/10.1111/j.1468-0017.2008.00343.x.

Wolfram, W., & Schilling-Estes, N. (1998). *American English: Dialects and variation*. Malden, MA: Blackwell Publishers, Inc..

Wong, P. C. M., Parsons, L. M., Martinez, M., & Diehl, R. L. (2004). The role of the insular cortex in pitch pattern perception: The effect of linguistic contexts. *Journal of Neuroscience, 24*(41), 9153–9160. http://dx.doi.org/10.1523/JNEUROSCI.2225-04.2004.

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Sciences, 6*(1), 37–46. http://dx.doi.org/10.1016/S1364-6613(00)01816-7.

Zatorre, R. J., Evans, E. C., Meyer, E., & Gjedde, A. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science, 256*(5058), 846–849. <http://www.jstor.org/stable/2877045>.

Zetterholm, E. (2002). A case study of successful voice imitation. *Logopedics Phoniatrics Vocology, 27*, 80–83. http://dx.doi.org/10.1080/140154302760409301.