Music Training and Vocal Production of Speech and Song

Elizabeth L. Stegemöller, Erika Skoe, Trent Nicol, Catherine M. Warrier, and Nina Kraus
Northwestern University

Studying similarities and differences between speech and song provides an opportunity to examine music's role in human culture. Forty participants divided into groups of musicians and nonmusicians spoke and sang lyrics to two familiar songs. The spectral structures of speech and song were analyzed using a statistical analysis of frequency ratios. Results showed that speech and song have similar spectral structures, with song having more energy present at frequency ratios corresponding to those ratios associated with the 12-tone scale. This difference may be attributed to greater fundamental frequency variability in speech, and was not affected by musical experience. Higher levels of musical experience were associated with decreased energy at frequency ratios not corresponding to the 12-tone scale in both speech and song. Thus, musicians may invoke multisensory (auditory/vocal-motor) mechanisms to fine-tune their vocal production to more closely align their speaking and singing voices according to their vast music listening experience.

Received May 18, 2007, accepted January 22, 2008.

Key words: speech, song, music, training, noise

Music has recently become a popular topic in neuroscience research. Areas of focus include music perception (see Koelsch & Siebel, 2005 for a review), the effects of musical experience on biological processes (see Münte, Altenmüller, & Jänke, 2002 for a review), and the acoustics of speech and music. While research in each of these areas has improved the understanding of brain mechanisms involved in music processing, little is known about how musical experience may extend to vocal production of music and possibly even to speech.

A fundamental feature of music perception and the origin of the 12-tone scale is the concept of sensory consonance and dissonance. Much research has focused on how the perception of consonance and dissonance contributes to the perception of music, including how neural encoding principles contribute to their perception. Intervals perceived to be consonant stimulate the cochlea in a highly ordered pattern versus dissonant intervals stimulate the cochlea in a diffuse pattern (Greenwood, 1991). Synchronous, phase-locked activity of neurons in the primary auditory cortex may also contribute to the perception of consonance and dissonance (Fishman et al., 2001). This perceptual ability may be unique to higher-order primates (Fishman et al., 2001; McDermott & Hauser, 2004), and it may also be innate. Infants perceive consonant intervals better than dissonant intervals, suggesting that biological factors, such as neural encoding, modulate this ability (Trehub, 2003).

Recent work by Schwartz and colleagues (2003) suggests that a statistical link between periodic stimuli and their physical source may partially account for the neural encoding of consonance and dissonance in music. Schwartz and colleagues found strong similarities between the spectral structure of speech and the organization of the 12-tone scale. When comparing the normalized spectrum of speech sounds to the intervals of the 12-tone scale, Schwartz and colleagues found that the majority of the intervals correspond to the peaks in the spectrum. Moreover, consonance rankings predicted from the amplitude of the peaks in the normalized spectrum were similar to perceptual consonance rankings, suggesting a statistical link between the auditory perception of consonance and dissonance in music (i.e., musical universals) and speech (Schwartz, Howe, & Purves, 2003). However, this technique has not been used to examine if the spectral structure of other vocal signals, such as song, also correspond to the organization of the 12-tone scale.

Previous research on the acoustics of speech and song uses numerous methods to examine differences between vocal signals, as well as the effect of vocal training on these signals. Research that focuses on examining only the fundamental frequency (vocal fold vibration) has shown that vocal training does not influence the speaking voice of professional singers (Brown, Hunt, & Williams, 1988; Brown, Rothman, & Sapienza,

Music Perception volume 25, issue 5, pp. 419–428, issn 0730-7829, electronic issn 1533-8312 DOI:10.1525/mp.2008.25.5.419

2000; Mendes, Brown, Rothman, & Sapienza, 2004; Rothman, Brown, Sapienza, & Morris, 2001; Watson & Hixon, 1985). However, vocal tract resonance does influence vocal production. Long-term average spectra (LTAS), a fast Fourier transform (FFT) generated power spectrum, has been widely used to examine vocal tract resonance (Sundberg, 1974). This technique has revealed differences in speech and song (Barrichelo, Heuer, Dean, & Sataloff, 2001; Cleveland, Sundberg, & Stone, 2001; Stone, Cleveland, & Sundberg, 1999) and differences in singing genres (Borch & Sundberg, 2002; Cleveland et al., 2001; Stone, Cleveland, Sundberg, & Prokop, 2003; Sundberg, Gramming, & Lovetri, 1993; Thalén & Sundberg, 2001). In particular, the "singer's formant," an increase in power around 3 kHz, (Sundberg, 1974) has been linked to the classical or operatic voice (Barnes, Davis, Oates, & Chapman, 2004; Bloothooft & Plomp, 1986; Burns, 1986; Sundberg, 1974). The singer's formant is the result of clustering of the third, fourth, and fifth formants, suggesting that vocal training may affect vocal tract resonance.

To date, little research has examined the effect of different levels (novice or professional) and types (vocal or instrumental) of musical experience on vocal fold vibration and vocal tract resonance of both speech and song. The technique described by Schwartz and colleagues (2003) examines the relationship between the largest harmonic of the fundamental and successive harmonics above it, accounting for the effects of both vocal fold vibration and vocal tract resonance. Thus, the purpose of this study was to use the same approach as Schwartz and colleagues to examine the spectral differences between speech and song, and the effect of type and extent of music training on vocal production. We proposed that the spectral structures of an individual's speaking and singing voice are modulated by experience. We hypothesized that: (1) spectral structure would be similar for speech and song, but more energy would be focused in frequency ratios corresponding to the 12-tone scale in song samples; and (2) musical experience would affect the spectral structure of both speech and song.

Method

Participants

Data were collected from 40 participants. All participants were native speakers of American English above the age of 18, had normal mental ability, normal speech, and no history of neurological illness. To examine the effect of *length* of musical experience, participants were divided into three groups, nonmusicians (n = 13), novice musicians (n = 13), and professional musicians (n = 14). Nonmusicians were defined as having fewer than five years of training on any musical instrument or voice (mean years of music training = 1.77, SD = 1.42). All professional musicians had over 10 years of training in music (mean years of music training = 13.07, SD = 2.55), and novice musicians had between five and nine years of music training (mean years of music training = 6.61. SD = 1.66). All groups differed significantly in years of training ($p < .001$ for all comparisons). To examine the effect of *type* of musical experience, a second grouping divided all novice and professional musicians into vocalists (n = 13; mean years of music training = 10.69, SD = 4.29) and instrumentalists (n = 14; mean years of music training = 9.28, SD = 3.56). Of the participants in the vocalist group, five participants were trained on voice only and eight were trained on voice and one to two additional instruments. All participants in the vocalist group had at least five years of vocal training. Participants in the instrumentalist group were trained on one to two instruments only, and all classes of instruments, percussion, woodwind, brass, and string were represented. Both the vocalist and instrumentalist groups significantly differed from the nonmusician group in years of training (vocalist vs. nonmusicians: $p < .001$; instrumentalists vs. nonmusicians: $p < .001$), but did not differ from each other (vocalists vs. instrumentalists: $p = .52$). All participants gave their written informed consent prior to inclusion into the study, and the procedures were approved by the Institutional Review Board of Northwestern University.

Data Collection

All recordings were completed in a professional sound booth designed with acoustic foam on all walls and ceiling, and carpet on the floor to minimize excess reverberations. The only equipment in the sound booth was a microphone stand, cord, and a Neumann KM 184 condenser microphone. The microphone was connected to a mixer with an 848 Motu interface box and then to the computer. Logic Pro 7 software was used for all recordings. The stand was adjusted to the proper height for each participant, and the participants were asked to stand approximately three inches in front of the microphone. The input volume of the microphone was tested, and the gain was adjusted on the mixer to the optimal level. Participants were asked to speak and sing the words to two well-known children's songs (Mary Had a Little Lamb; Row, Row, Row Your Boat).

To minimize the influence of melody and rhythm on the speaking voice, speaking samples were collected first followed by singing samples, and participants were specifically instructed not to speak rhythmically or melodically, maintaining their normal speaking voice throughout. Additionally, the melodies and rhythms of the two songs were not provided and no reference pitch was given. This was done so as to not influence the participants' natural singing or speaking voices. Participants were asked to review the words to both songs before recording. Two trials were collected for each condition. All trials were recorded at 24 bit, 48 kHz sampling rate.

Data Analysis

Data analysis followed the same methods as Schwartz et al. (2003). After obtaining sound recordings, all samples were downsampled to 16 kHz. Using Matlab, samples were divided into 100 ms segments and rated according to energy. Segments that had an amplitude <5% of the maximum amplitude of the sample were omitted to avoid silent intervals. The mean number of segments analyzed per sample was 74 (range = 48-143). The segments were analyzed using a fast Fourier transform. Each resulting amplitude (A) and frequency (F) value was normalized relative to the amplitude of the frequency exhibiting the maximal energy across the sample (A_{max}) and the frequency (F_{max}) at A_{max}, respectively.

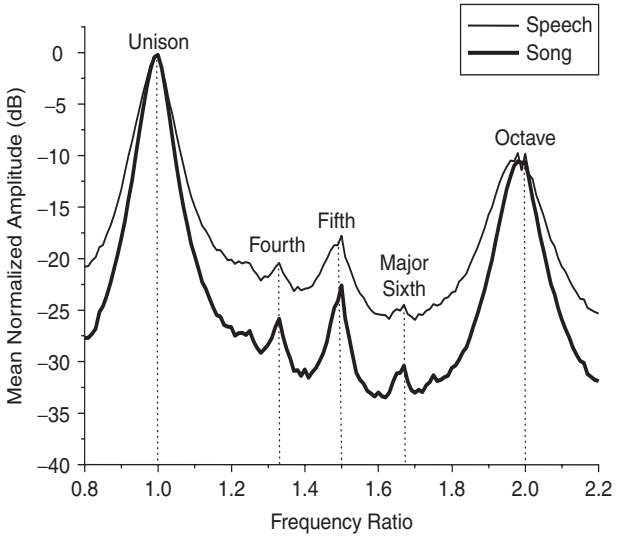

FIGURE 1. Grand average ratio spectra of the speech and song samples. Normalized speech (thin line) and song (thick line) spectra. Dashed lines demonstrate that the prominent peaks correspond to five intervals of the 12-tone scale.

This produced the normalized ratio values A_n and F_n, where $A_n = A/A_{max}$ and $F_n = F/F_{max}$. Then F_n's in all 100 ms segments were aligned, and the normalized amplitude ratios were averaged and converted to a dB scale. Due to the normalizing procedure, A_n (0 dB) is always maximal at $F_n = 1$. The A_n's at all other frequency ratios are expressed in dB below A_{max}. Plotting A_n against F_n results in a dB-to-frequency ratio function where amplitude values peak at whole number frequency ratios and a number of predictable intermediate ratios (see Figure 1). Schwartz et al., (2003) found that the frequency ratios associated with peaks in the first octave range of this function correspond to the frequency ratios between intervals in the Western musical 12-tone scale.[1]

Because the minor and major seconds typically fall on the down slope of the unison and the major seventh falls on the up slope of the octave, peaks and troughs in the normalized spectrum for the 13 predicted scale intervals were manually and independently picked by two authors using the frequency ratios obtained by Schwartz and colleagues as guides. Discrepancies were rare and were settled by a third author. For all samples,

[1]Modeling the human vocal tract as a source and filter defines the laryngeal folds as a source of sound and the vocal tract as the filter of sound (Lieberman & Blumstein, 1988; Stevens, 1999). When the lungs expel air, it passes through the laryngeal folds and a sound pressure wave is generated which is periodic over short intervals of time (Ladefoged, 1962). This waveform has a maximum power at the rate of laryngeal fold vibration, the fundamental frequency (F0), and includes an integer harmonic series. This sound wave is then further modified as it travels through the vocal tract. The natural resonances of the vocal tract, as determined by its length and shape, produce formants where the power of the harmonic series is least attenuated (Fant, 1960). For example, a person whose vocal tract resonates at 500 Hz and has a fundamental frequency of 100 Hz would have the greatest amount of concentrated power at the fifth harmonic, which corresponds to 500 Hz. Using the methods of Schwartz and colleagues (2003), the fifth harmonic and would be assigned $F_n = 1$ and $A_n = 0$ and all frequency information would be normalized with respect to the fifth harmonic. Peaks at frequency ratios of 1, 1.2, 1.4, 1.6, 1.8, and 2 (i.e., 1 plus n/5, where n is an integer value between 1 and 5) would be expected to be the most prominent. Likewise, if the F0 of a given sample is 250 Hz, then the second harmonic would likely be the least attenuated. In this case, peaks at 1, 1.5, and 2 (i.e., 1 plus n/2. where n is an integer value between 1 and 2) would be the most prominent. Most human utterances have an F0 between 100 and 250 Hz, with a vocal tract resonance at approximately 500 Hz. Thus, the frequency ratios of the large corpus analyzed by Schwartz and colleagues are a direct consequence of the acoustics of human vocalization, which interestingly corresponds to the intervals of the 12-tone scale of Western music (Schwartz et al., 2003).

the most prominent peaks, corresponding to unison, fourth, fifth, major sixth, and the octave, were identified. To assess differences in the normalized spectra of speech and song and the effect of musical experience, peak-to-trough slope, a measure of power concentration, was calculated for these five prominent peaks. The upward and downward peak-to-trough slopes (|change in dB/change in frequency ratio|) were calculated for each participant. Using sign changes in the differential waveform, a second measure tallied the total number of peaks between unison and octave. The number of peaks present at the predicted 12-tone scale ratios for each participant was then subtracted from the total number of peaks in the waveform. This measure of extra, nonpredicted peaks served as a measure of how much energy was present at frequency ratios not corresponding to those of the 12-tone scale.

In order to compare our frequency ratio results to those employing more traditional methods of investigating speech, three additional measures were calculated in Praat (Boersma, 2001) and used to compare speech and song vocalizations among groups. The first two are measures of fundamental frequency (F0) variability. A measure of perturbation (in percent) between adjacent cycles of the F0 across each utterance was measured using Praat's local jitter technique. A related, slightly longer-term measure of F0 variability, more specifically investigated our 100 ms segments by means of F0 pitch extraction. The F0 contour of each speech and song sample was extracted using a time window of 33.3 ms (3 samples per segment) with the pitch floor and ceiling being set to their defaults (75 Hz, 600 Hz). Within-segment F0 variability was assessed by calculating the standard deviation of the three samples comprising the segment. Voiceless points were represented as zeros, and the standard deviation was calculated only if two of three or all three points in a segment were voiced; other segments were omitted from this analysis. A third measure, harmonic-to-noise ratio (HNR; Yumoto & Gould, 1982) was calculated on the 100 ms segments, excluding those that did not meet the minimum amplitude criterion applied in the ratio analyses. The result, expressed in dB, is a common measure of voice quality: the lower the HNR the breathier or hoarser the utterance. For these three measures, the data were exported from Praat for subsequent analysis in Matlab. For each measure, two composite scores were calculated per participant; an average score across the speech samples, and across the song samples.

A 2 × 3 repeated measures analysis of variance with posthoc analysis was completed for all statistical comparisons. Two separate analyses were completed for each measure. For the first analysis, the within-group factor was vocalization (speech vs. song) and the between-group factor was musical experience (professional musicians vs. novice musicians vs. nonmusicians). For the second analysis, the within-group factor again was vocalization (speech vs. song) and the between-group factor was training type (vocalists vs. instrumentalists vs. nonmusicians). Five measures were analyzed in each comparison; slope values (unison, major fourth, fifth, major sixth, and octave) and number of nonpredicted peaks from the frequency ratio data, and jitter, F0 variability and HNR from the raw waveforms.

Results

Statistical Structure of Speech and Song

Comparable to Schwartz's data, the normalized spectra of our speech samples produced peaks at frequency ratios corresponding to the majority of the 12-tone scale ratios. The same pattern of spectral structure was also found for song. There were no peaks evident for the minor and major seconds which fall on the down slope of the unison, and the major seventh which falls on the up slope of the octave (see Table 1 for frequency ratios). This is consistent with Schwartz's results. However, sampling only 40 participants compared to 630 participants in Schwartz's studied resulted in less prominent peaks, though still apparent, for the minor and major third. Figure 1 shows how the frequency ratios, corresponding to the scale intervals of unison, fourth, fifth,

TABLE 1. Ratio Values for the 12-Tone Scale.

Interval	Just Intonation
Unison	1.000
Minor second	1.067
Major second	1.125
Minor third	1.200
Major third	1.250
Fourth	1.333
Tritone	1.406
Fifth	1.500
Minor sixth	1.600
Major sixth	1.667
Minor seventh	1.750
Major seventh	1.875
Octave	2.000

major sixth, and octave, aligned with the most prominent peaks in both the speech and song grand average samples.

Differences in Slope

Although the harmonic structure of speech and song were similar, the frequency ratio peaks in the song sample were more prominent resulting in increased peak-to-trough amplitude and decreased peak-to-trough width (Figures 1 and 3). The slope calculation takes both amplitude and width into consideration. So to test for significant differences between speech and song, each slope value for the unison, fourth, fifth, major sixth, and the octave was calculated and analyzed. For each peak tested there was a main within-group effect of vocalization (speech vs. song, see Table 2). However, there were no between-group effects; both the level of musical experience (professional, novice, and nonmusician) and the type of musical experience (instrumental, vocal, or neither) were not significant. Figure 2 shows the mean slope values and standard errors for speech and song collapsed across participant groups of the five analyzed peaks. These data suggest that regardless of musical experience, resonances were more precise and more concentrated in power (i.e., steeper peak–to-trough slopes) at those five perceptually consonant intervals in song as compared to speech.

TABLE 2. Mean, Standard Deviation, and Within-Group Effect of Slope Value for Each Speech and Song Peak.

Interval	*M* (dB/ frequency ratio)	*SD*	*F* (1, 39)	*p*
Unison				
Speech	123.81	18.06	66.25	< .001
Song	152.39	21.99		
Fourth				
Speech	92.57	42.51	15.84	< .001
Song	137.23	67.21		
Fifth				
Speech	146.73	116.50	25.99	< .001
Song	208.21	136.72		
Major Sixth				
Speech	98.85	52.91	11.29	< .02
Song	129.14	65.13		
Octave				
Speech	85.96	14.42	124.43	< .001
Song	115.96	17.24		

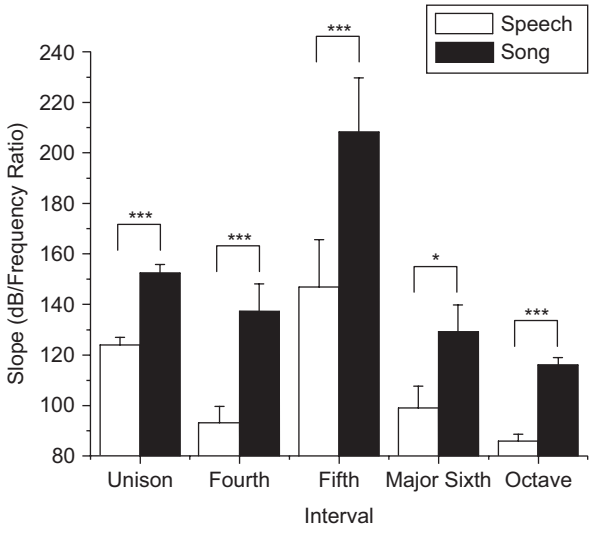

FIGURE 2. Mean slope values and standard error for speech and song. Each grouping (unison, fourth, fifth, major sixth, and octave) significantly differed between speech and song. One asterisk designates significance at $p < .05$ and triple asterisks designate significance at $p < .001$.

Differences in Nonpredicted Peaks

Further distinguishing the harmonic structure of speech and song was the amount of nonpredicted peaks in the samples, as defined by the number of peaks in excess of those ratios defining the 12-tone scale. To show an example of this difference, Figure 3 shows average functions for nonmusicians and vocalists, the most disparate groups, for both speech and song. The signals have been enlarged to show the difference in number of nonpredicted peaks between the frequency ratios of 1.2 and 1.8. The number of nonpredicted peaks decreased from speaking to singing and with musical experience. This effect can be seen clearly by the numerous extraneous peaks in the normalized spectrum of the nonmusicians' speech signal compared to that of vocalists (Figure 3).

Figure 4 shows the average number of nonpredicted peaks and standard error in the averaged speech and song samples for the five groups. Each comparison for level and type of musical experience revealed both a within-group effect and between-group effects. Across all participants, the number of nonpredicted peaks was significantly smaller in song than speech, $F(1, 39) = 13.29$, $p = .001$. Post hoc comparisons for the level of musical experience revealed that professional musicians had a significantly smoother sample (fewer or no nonpredicted peaks) than nonmusicians for song. Post

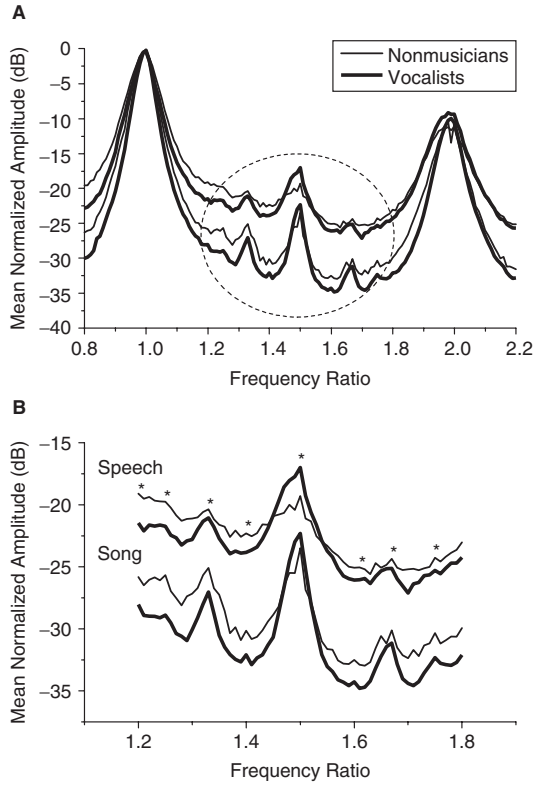

FIGURE 3. Averaged normalized ratio spectra of speech (top two traces) and song (bottom two traces) for nonmusicians and vocalists. The encircled portion of the signal in A is enlarged to show the decrease in the number of peaks from speech to song, and from no musical experience to trained vocal experience. Asterisks designated the location of 8 of the predicted 13 12-tone scale peaks that fall in between the frequency ratios of 1.2 and 1.8. Additional, nonpredicted, peaks serve as a measure of spectral noise.

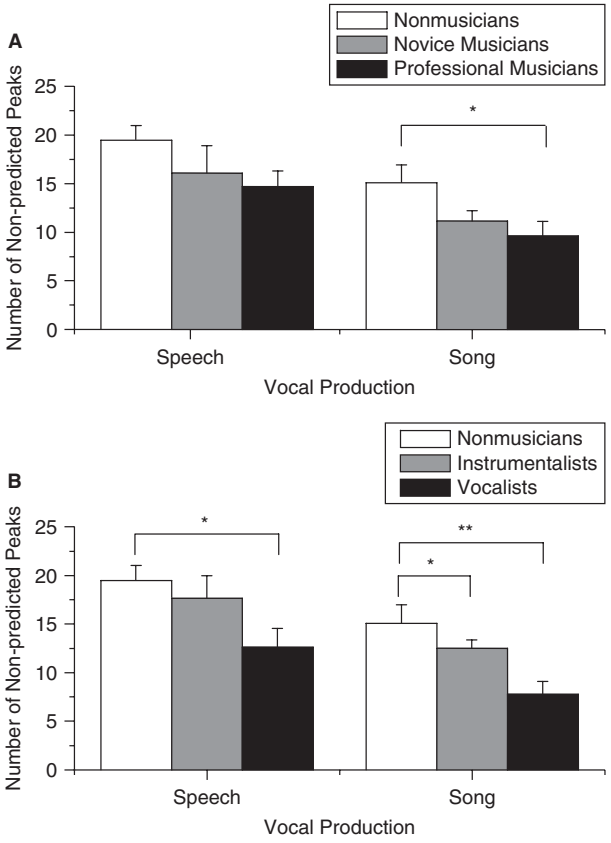

FIGURE 4. Mean value and standard error for the number of nonpredicted peaks, spectral noise, in speech and song. One asterisk designates significance at $p < .05$ and double asterisks designate significance at $p < .01$. **A.** Mean number of nonpredicted peaks for speech and song for nonmusicians, novice musicians, and professional musicians. **B.** Mean number of nonpredicted peaks for speech and song for nonmusicians, instrumentalists, and vocalists.

hoc comparisons for type of musical experience revealed that there were significantly fewer nonpredicted peaks for vocalists compared to nonmusicians for both speech and song, while instrumentalists had a significantly smoother sample than nonmusicians for song only (Table 3). Thus, musical experience, specifically vocal training, affected the number of nonpredicted peaks in the acoustic signal, increasing the musicality of the voice in both speech and song.

Jitter, F0 Variability, HNR

Jitter, short-term F0 variability, and Harmonic to Noise Ratio (HNR) all showed a main within-group effect of vocalization (speech vs. song), with song showing less F0 deviance and greater HNR. However, none of these measures corroborated the between-group effects that were seen in the nonpredicted peak analysis of the frequency-ratio spectra, nor were there any group by mode interactions (Table 4).

Discussion

Summary of Results

The above results showed that the overall spectral structure of speech and song is similar. However, the slope of the major peaks was greater for song. There was also a difference in the number of nonpredicted peaks, with a decrease in number of nonpredicted peaks in song samples. The extent of music training was related to the number of nonpredicted peaks in both speech and song. The number of nonpredicted peaks decreased as musical experience increased, with professional music training having the greatest effect. Additionally, F0 variability and HNR differed between speech and song, but were not affected by musical experience. Using the words to the same familiar songs

TABLE 3. Mean, Standard Deviation, and Post hoc Comparisons (Fisher's Least Significant Difference test) of the Number of Nonpredicted Speech and Song Peaks for Musical Experience Groups ($df = 37$).

Group	*M*	*SD*	Post hoc Comparisons (level of musical experience)	*t*	Post hoc Comparisons (type of musical experience)	*t*
Non.			*Non. vs. Nov.*		*Non. vs. Instr.*	
Speech	19.46	5.13	Speech	1.21	Speech	0.69
Song	15.08	6.42	Song	1.94	Song	1.39
Nov.			*Non. vs. Prof.*		*Non. vs. Voc.*	
Speech	17.31	9.10	Speech	1.83	Speech	2.55*
Song	10.46	6.29	Song	2.86*	Song	3.90**
Prof.			*Nov. vs. Prof.*		*Instr. vs. Voc.*	
Speech	16.36	7.71	Speech	0.60	Speech	1.90
Song	9.57	5.20	Song	0.90	Song	2.55*
Instr.						
Speech	17.64	8.33				
Song	12.50	2.90				
Voc.						
Speech	12.62	6.64				
Song	7.77	4.57				

Non. = nonmusician; Nov. = novice musician; Prof. = professional musician; Instr. = instrumentalists; Voc. = vocalists. *$p < .05$. **$p < .01$.

as our speech sample may have limited the effect of musical experience. It is possible that mental imagery of the songs may have influenced the speech samples, making the speech samples more musical (Zatorre & Halpern. 2005), especially for musicians. However, significant differences between our speech and song samples were still found and would suggest that the differences between speech and song may be even greater for "pure" speech.

Speech vs. Song

Each technique revealed differences between speech and song. However, only the technique described by Schwartz and colleagues revealed differences in the vocal spectrum of musicians and nonmusicians, while traditional techniques, F0 variability and HNR, did not. The F0 of speech is likely to change more frequently than the F0 of song. Thus, if the F0 of speech changed more frequently on either a cycle-by-cycle basis or slightly more slowly over the course of the 100 ms analysis window, then the energy in the spectrum would be distributed across a wider range of frequencies, resulting in broader peaks. Likewise, a greater HNR while singing may also contribute to the increased energy seen at the five prominent peaks of the normalized spectrum. Therefore, short-term F0 perturbations or HNR differences may sufficiently describe the differences in slope between speech and song. However, the differences in the number of nonpredicted peaks seen between groups is not due to these factors, but rather a different factor that may be observed when using the technique described by Schwartz and colleagues. The technique examines the relationship between the largest harmonic of the fundamental and successive harmonics above it. This type of analysis accounts for the effects of both vocal fold vibration (F0) and vocal tract resonance. Thus, differences seen here in the number of nonpredicted peaks associated with musical experience may be, at least in part, the result of changes in vocal tract resonance. Using the method described by Schwartz and colleagues may provide a more accurate and objective quantitative measure for singing and vocal quality, revealing changes in both vocal fold vibration and vocal tract resonance.

TABLE 4. Mean, Standard Deviation, and Within-Group Effect for Jitter, F0 Variability, and HNR for Speech and Song.

Measure	*M*	*SD*	$F(1, 39)$	*P*
Jitter (%)				
Speech	0.09	0.02	43.96	< .01
Song	0.07	0.02		
F0 Variability (Hz)				
Speech	6.60	2.47	16.53	< .01
Song	5.33	1.80		
HNR (dB)				
Speech	15.9	2.44	458.38	< .01
Song	21.3	2.50		

Effects of Musical Experience

Previous research has reported that vocal training does not influence various physiologic and acoustic parameters of the speaking voice in professional singers (Brown et al., 1988, 2000; Mendes et al., 2004; Rothman et al., 2001; Watson & Hixon, 1985). These researchers used more traditional methods that generally only account for vocal fold vibration (Yumoto & Gould, 1982) to investigate the differences between speech and song. Results showed no difference between speech vocalizations of trained and untrained singers. In contrast, previous research has suggested that vocal tract resonance while singing changes with vocal training.

In classically trained and operatic singers, there is an increase in energy around 3 kHz, which is the result of a clustering of the third, fourth, and fifth harmonics (Barnes et al., 2004; Bloothooft & Plomp, 1986; Burns, 1986; Sundberg, 1974). Our results are in keeping with this research. It is important to note that our results include instrumental training (wind, string, and percussive), while most of the previous research has focused on vocal training only. While not as significant as vocal training, our results did show that instrumental training affected the number of nonpredicted peaks. This would suggest that the changes in vocal tract resonance may also be, at least in part, related to auditory experience.

Effects of musical experience on the nervous system are well documented at cortical (Gaser & Schlaug, 2003; Ohnishi et al., 2001; Pantev, Oostenveld, Engelien, Ross, Roberts, & Hoke, 1998; Peretz & Zatorre, 2005; Schlaug, Jancke, Huang, & Steinmetz, 1995; Wong, Skoe, Russo, Dees, & Kraus, 2007; Zatorre, 1998) and subcortical levels (Musacchia, Sams, Skoe, & Kraus, 2007; Wong et al., 2007). Musical experience inherently engages coordination across sensory modalities, leading to the notion that the interplay between modalities is stronger in musicians. Enhanced audiovisual task performance has been reported in musicians (Saldana & Rosenblum, 1993) and in conductors and is related to enhanced activity in multisensory cortical areas (Hodges, Hairston, & Burdette, 2005). Audiovisual enhancement in musicians has been shown to extend to subcortical sensory structures (Musacchia et al., 2007). Moreover, enhanced brainstem function occurred for both music and for speech stimuli, consistent with common, dynamic subcortical mechanisms for speech and music. Finally, the rich neural systems that underlie auditory-motor interactions in music perception and production also differ in musicians (Zatorre, Chen, & Penhune, 2007). One interpretation of the findings reported here is that musicians may invoke multisensory (auditory/vocal-motor) mechanisms to fine-tune their vocal production to more closely align their speaking and singing voices according to their vast experience listening to music. These mechanisms may also drive music preferences.

Music preferences are thought to be determined by environmental factors, although some of the basic features of music, consonance and dissonance, may be determined by the hard wiring of the auditory system (Fishman et al., 2001; Greenwood, 1991). Noise could be considered a type of dissonance, and may be rated and encoded neurologically as dissonant. In comparison, we found the spectrum of song had less spectral noise (decreased number of nonpredicted peaks), suggesting a more precise harmonic series. Therefore, song may be rated and encoded neurologically as more consonant than speech. Individual preferences for musical styles and artists may be driven by the level of noise in the signal. In the above data, an interesting finding was that trained musicians, specifically vocalists, had less spectral noise in their speech and song signals. There was also a trend, though not significant, towards more precise resonance (more prominent and sharper peaks) in the vocal signals of professional musicians. Thus, one could speculate that people may prefer certain singers and performers based on their ability to minimize the amount of spectral noise and optimize the resonating precision in their vocal or instrumental sound (Watts, Barnes-Burroughs, Estis, & Blanton, 2006). Although, this concept may contribute to musical preference, other factors must also exist, as some popular musicians employ noisier signals, including purposeful signal distortions.

Music is a part of human culture, yet its purpose from evolutionary and biological perspectives is somewhat mysterious. Schwartz et al. (2003) suggested that the 12-tone scale was derived from the speaking voice due to a necessary statistical relationship between the stimuli (sound) and source (vocal tract) that allows humans to gain relevant information, such as speaker identification or emotion, from other sound sources. Here we have shown that song is a less variable acoustic signal than speech, having increased energy concentrated at frequency ratios corresponding to the 12-tone scale, and that musical experience reduced the number of nonpredicted peaks in both speech and song signals. Thus the 12-tone scale is more strongly conveyed through song, suggesting that song may have lead to

the development of the 12-tone scale. Moreover, these results suggest a direct link between musical experience and vocal production, most notably speech. It may be possible that music training increases the number of times an individual is exposed to a musical signal, and this exposure in turn influences the individual's speech. Perhaps then, one purpose of music and song in human culture and evolution is for the learning and refinement of speech and language (Saffran, Johnson, Aslin, & Newport, 1999; Schön, Boyer, Moreno, Besson, Peretz, & Kolinsky, 2007). While speech may predict musical universals (Schwartz et al., 2003), also plausible is that the statistical structure of human song predicts musical universals that influence human speech (Mithen, 2005).

Author Note

We are thankful to Professor Colum MacKinnon and Professor Ric Ashley for their useful comments.

Correspondence concerning this article should be addressed to Elizabeth L. Stegemöller, who is now at the Department of Physical Therapy and Human Movement Sciences, 645 North Michigan Avenue, Suite 1100, Chicago, Illinois 60611. E-MAIL: e-stegemoller@northwestern.edu

References

Barnes, J., Davis, P., Oates, J., & Chapman, J. (2004). The relationship between professional operatic soprano voice and high range spectral energy. *Journal of the Acoustical Society of America*, *116*, 530-538.

Barrichelo, V., Heuer, R., Dean, C., & Sataloff, R. (2001). Comparison of singer's formant, speaker's ring and LTA spectrum among classical singers and untrained normal speakers. *Journal of Voice*, *15*, 344-350.

Bloothooft, G., & Plomp, R. (1986). The sound level of the singer's formant in professional singing. *Journal of the Acoustical Society of America, 79*, 2028-2033.

Boersma, P. (2001). Praat: A system for doing phonetics by computer. *Glot International*, *5*, 341-345.

Borch, D., & Sundberg, J. (2002). Spectral distribution of solo voice and accompaniment in pop music. *Logopedics Phoniatrics Vocology*, *27*, 37-41.

Brown, W. S., Hunt, E., & Williams, W. (1988). Physiological differences between trained and untrained speaking and singing voice. *Journal of Voice*, *2*, 102-108.

Brown, W. S., Rothman, H., & Sapienza, C. (2000). Perceptual and acoustic study of professionally trained versus untrained voices. *Journal of Voice*, *14*, 301-309.

Burns, P. (1986). Acoustical analysis of the underlying voice differences between two groups of professional singers: Opera and country and western. *Laryngoscope, 96*, 549-554

Cleveland, T., Sundberg, J., & Stone, R. (2001). Long-Term-Average-Spectrum characteristics of country singers during speaking and singing. *Journal of Voice*, *15*, 53-60.

Fant, G. (1960). *Acoustic theory of speech production.* The Hague: Mouton.

Fishman, Y. I., Volkov, I. O., Noh, M. D., Garell, P. C., Bakken, H., Arezzo, J. C., et al. (2001). Consonance and dissonance of musical chords: neural correlates in auditory cortex of monkeys and humans. *Journal of Neurophysiology*, *86*, 2761-2788.

Gaser, C., & Schlaug, G. (2003). Brain structures differ between musicians and non-musicians. *Journal of Neuroscience, 23*, 9240-9245.

Greenwood, D. D. (1991). Critical bandwidth and consonance in relation to cochlear frequency-position coordinates. *Hearing Research*, *54*, 164-208.

Hodges, D. A., Hairston, W. D., & Burdette, J. H. (2005). Aspects of multisensory perception: the integration of visual and auditory information in musical experiences. *Annals of the New York Academy of Sciences*, *1060*, 175-85

Koelsch, S., & Siebel, W. A. (2005). Towards a neural basis of music perception. *Trends in Cognitive Sciences, 9*, 578-584.

Ladefoged, P. (1962). *Elements of acoustic phonetics*. Chicago: University of Chicago.

Lieberman, P., & Blumstein, S. E. (1988). *Speech physiology, speech perception and acoustic phonetics.* Cambridge, UK: Cambridge University Press.

McDermott, J., & Hauser, M. (2004). Are consonant intervals music to their ears? Spontaneous acoustic preferences in a nonhuman primate. *Cognition*, *94*, B11-B21.

Mendes, A. P., Brown, W. S., Rothman, H. B., & Sapienza, C. (2004). Effects of singing training on the speaking voice of voice majors. *Journal of Voice*, *18*, 83-89.

Mithen, S. (2005). *The singing Neanderthals: The origins of music, language, mind, and body.* London: Weidenfeld and Nicolson.

Münte, T. F., Altenmüller, E., & Jänke, L. (2002). The musician's brain as a model of Neuroplasticity. *Nature Reviews Neuroscience, 3*, 473-478.

Musacchia, G., Sams, M., Skoe, E., & Kraus, N. (2007). Musicians have enhanced subcortical auditory and

audiovisual processing of speech and music. *Proceedings of the National Academy of Sciences*, *104*, 15894-158948.

Ohnishi, T., Matsuda, H., Asada, T., Aruga, M., Hirakata, M., Nishikawa, et al. (2001). Functional anatomy of musical perception in musicians. *Cerebral Cortex, 11*, 754-760.

Pantev, C., Oostenveld, R., Engelien, A., Ross, B., Roberts, L. E., & Hoke, M. (1998). Increased auditory cortical representation in musicians. *Nature, 392*, 811-814.

Peretz, I., & Zatorre, R. (2005). Brain organization for music processing. *Annual Review in Psychology, 56*, 89-114.

Rothman, H., Brown, W. S., Sapienza, C., & Morris, R. (2001). Acoustic analysis of trained singers perceptually identified from speaking samples. *Journal of Voice*, *15*, 25-35.

Saffran, J. R., Johnson, R. E. K., Aslin, N., & Newport, E. L. (1999) Abstract statistical learning of tone sequences by human infants and adults. *Cognition*, *70*, 27-52.

Saldana, H. M., & Rosenblum, L. D. (1993). Visual influences on auditory pluck and bow judgments. *Perception and Psychophysics, 54*, 406-16.

Schlaug, G., Jancke, L., Huang, Y., & Steinmetz, H. (1995). In vivo evidence of structural brain asymmetry in musicians. *Science, 267*, 699-701.

Schön, D., Boyer, M., Moreno, S., Besson, M., Peretz, I., & Kolinsky, R. (2007). Songs as an aid for language acquisition. *Cognition*, *106*, 975-983.

Schwartz, D. A., Howe, C. Q., & Purves, D. (2003). The statistical structure of human speech sounds predicts musical universals. *Journal of Neuroscience, 23*, 7160-7168.

Stevens, K. (1999). *Acoustic phonetics*. Cambridge, MA: MIT.

Stone, R., Cleveland, T., & Sundberg, J. (1999). Formant frequencies in country singers' speech and singing. *Journal of Voice*, *13*, 161-167.

Stone, R., Cleveland, T., Sundberg, J., & Prokop, J. (2003). Aerodynamic and acoustical measures of speech, operatic, and Broadway vocal styles in a professional female singer. *Journal of Voice*, *17*, 283-297.

Sundberg, J. (1974). Articulatory interpretation of the "singer's formant." *Journal of the Acoustical Society of America*, *55*, 838-844.

Sundberg, J., Gramming, P., & Lovetri, J. (1993). Comparisons of pharynx, source, formant, and pressure characteristics in operatic and musical theatre singing. *Journal of Voice*, *7*, 301-310.

Thalén, M., & Sundberg, J. (2001). Describing different styles of singing: a comparison of a female singer's voice source in "Classical", "Pop", "Jazz" and "Blues." *Logopedics Phoniatrics Vocology*, *26*, 82-93.

Trehub, S. (2003). The developmental origins of musicality. *Nature Neuroscience, 6*, 669-673.

Watson, P., & Hixon, T. (1985). Respiratory kinematics in classical (opera) singers. *Journal of Speech and Hearing Research, 28*, 104-122.

Watts C., Barnes-Burroughs K., Estis J., & Blanton D. (2006). The singing power ratio as an objective measure of singing voice quality in untrained talented and nontalented singers. *Journal of Voice*, *20*, 82-88.

Wong, P. C., Skoe, E., Russo, N. M., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience, 10*, 420-422.

Yumoto, E., & Gould, W. J. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *Journal of the Acoustical Society of America*, *71*, 1544-1550.

Zatorre, R. (1998). How do our brains analyze temporal structure in sound? *Brain, 121*, 1817-1818.

Zatorre, R., Chen, J., & Penhune, V. (2007). When the brain plays music: Auditory-motor interactions in music perception and production. *Nature Reviews Neuroscience*, *8*, 547-558.

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