Musical intervals in speech

Deborah Ross, Jonathan Choi, and Dale Purves*

Center for Cognitive Neuroscience and Department of Neurobiology, Duke University, Durham, NC 27708

Contributed by Dale Purves, April 5, 2007 (sent for review January 29, 2007)

Throughout history and across cultures, humans have created music using pitch intervals that divide octaves into the 12 tones of the chromatic scale. Why these specific intervals in music are preferred, however, is not known. In the present study, we analyzed a database of individually spoken English vowel phones to examine the hypothesis that musical intervals arise from the relationships of the formants in speech spectra that determine the perceptions of distinct vowels. Expressed as ratios, the frequency relationships of the first two formants in vowel phones represent all 12 intervals of the chromatic scale. Were the formants to fall outside the ranges found in the human voice, their relationships would generate either a less complete or a more dilute representation of these specific intervals. These results imply that human preference for the intervals of the chromatic scale arises from experience with the way speech formants modulate laryngeal harmonics to create different phonemes.

language | music | formants | scales | perception

Although periodic sound stimuli arise from a variety of natural sources, conspecific vocalizations are the principal source of periodic sound energy that humans have experienced over both evolutionary and individual time (1–3). It thus seems likely that the human sense of tonality and preferences for the specific tonal intervals are predicated on some aspect of speech. Indeed, several anomalies in the perception of pitch can be explained in terms of the human voice (2). Additional support for this idea has already been provided by the statistical presence of musical ratios in segments of voiced speech spectra that accord with many of the chromatic scale intervals, as well as evidence that consonance ranking is likely to be based on the distribution of energy in voiced speech (3). Despite pointing to the origin of chromatic intervals and relative consonance in the normalized distribution of energy in voiced speech, a more specific basis for these intervals in human vocalizations has remained unclear.

Intuitively, the most obvious place to look for musical intervals in human vocalizations would be in vocal prosody, i.e., the rising and falling pitches that characterize normal speech. When we examined recorded speech from this perspective, however, we failed to find any definitive evidence of musical intervals [see supporting information (SI) *Text*]. We thus turned to the possibility that the intervals of the chromatic scale are embedded in the spectral relationships within speech sound stimuli (called phones) that differentiate the phonemes perceived (4).

The periodicity in speech sound stimuli is generated primarily by the repeating peaks of energy in the vocal air stream produced by oscillations of the vocal folds in the larynx. The intensity carried by the harmonic series produced in this way is altered, however, by the resonance frequencies of the rest of the vocal tract, which change dynamically in response to neurally controlled movements of the soft palate, tongue, lips and other articulators (Fig. 1*A*). These variable vocal tract resonances, called formants, modulate the harmonic series generated by the laryngeal oscillations by suppressing some harmonics more than others (4, 5, 7, 8).† When coupled with unvoiced speech sounds (consonants), this modulation by the formants creates the different voiced speech sounds that give rise to the semantic content in all human languages. With respect to vowel phones, only the first two formants have a major influence on the vowel perceived: artificially removing them from vowel phones makes vowel phonemes largely indistinguishable, whereas removing the higher formants has little effect on the perception of speech sounds† (see *SI Text*). Indeed, the first and second formants of vowel sounds of all languages fall within well defined frequency ranges (4, 7–12). The resonances of the first two formants are typically between ≈200–1,000 Hz and ≈800–3,000 Hz, respectively, their central values approximating the odd harmonics of the resonances of a tube ≈17 cm in length open at one end, the usual physical model of the adult vocal tract in a relaxed state (4, 5, 7, 8).†

To test the hypothesis that chromatic scale intervals are specifically embedded in the frequency relationships in voiced speech sounds (i.e., phones whose acoustical structure is characterized by periodic repetition), we analyzed the spectra of different vowel nuclei in neutral speech uttered by adult native speakers of American English, as well as a smaller database of Mandarin.

Results

We first explored the ranges of the harmonics with the greatest intensity in the first and second formants in our database. Fig. 1*B* shows that, for English-speaking males uttering single words in a neutral emotional state, only harmonics 2–10 are possible intensity maxima in the first formant (F1) of vowels, and only harmonics 8–26 are possible maxima for the second formant (F2); for English-speaking females, these numbers are somewhat lower (harmonics 2–6 and 6–19, respectively) because the higher fundamental frequency of female vocalizations causes fewer harmonics to fall within the range of the first two formants in neutral speech (Fig. 1*C*).

Fig. 2 shows representative examples from the database for the three "point vowels" in English, i.e., the vowels whose formants are furthest apart in the F1 × F2 plot (vowel space) typically used in psycholinguistic studies (7); the most intense harmonic in the first and second formants of each utterance is indicated. The inset keyboards show that when the harmonic peak of the first formant of any vowel utterance in the database is set to a note represented on a piano tuned in just intonation, the peaks of intensity in the second formant often, but not always, fall on another note on the keyboard. Thus the ratio of the second to the first formant often represents one of the ratios that define chromatic scale intervals.

Fig. 3 shows the distribution of all F2/F1 ratios derived from the spectra of the 8 different vowels uttered by the 10 English-speaking participants (i.e., the relationships in 1,000

Author contributions: D.R., J.C., and D.P. designed research; D.R. and J.C. performed research; D.R. and J.C. analyzed data; and D.R., J.C., and D.P. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

*To whom correspondence should be addressed. E-mail: purves@neuro.duke.edu.

†Schouten, J. F., Fourth International Congress on Acoustics, August 21–28, 1962, Copenhagen, Denmark, 196:201–203.

This article contains supporting information online at www.pnas.org/cgi/content/full/0703140104/DC1.

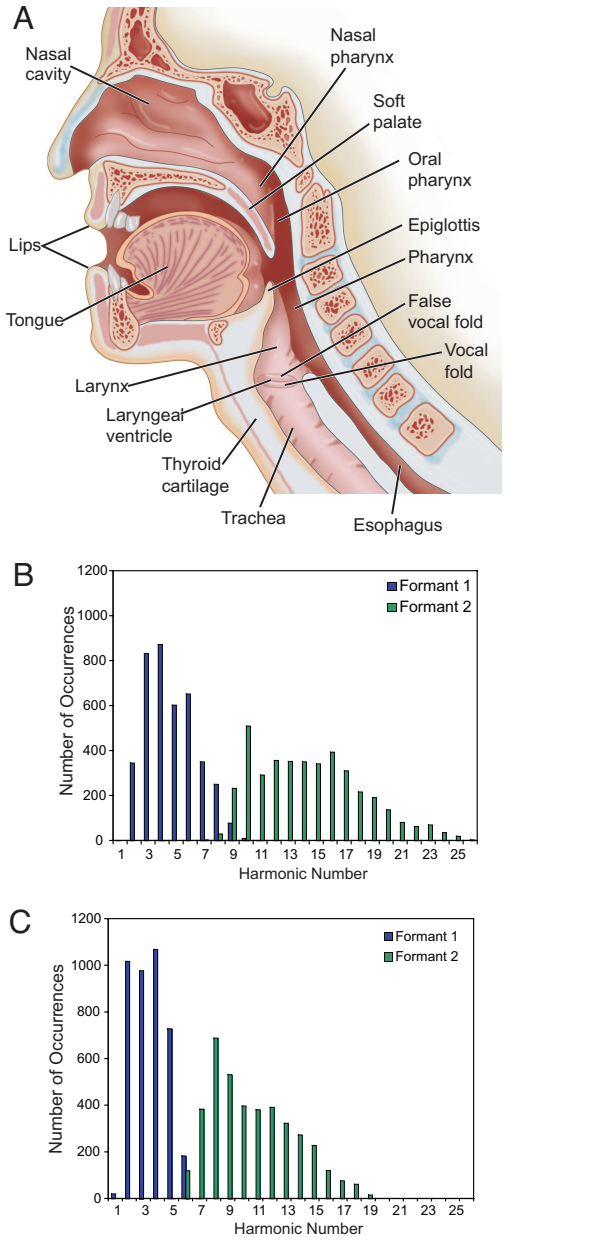

Fig. 1. Ranges of the peak harmonic in the first two formants (F1 and F2) for eight American English vowels uttered as single words in an emotionally neutral manner. (*A*) Diagram of the human larynx and vocal tract; see Introduction for explanation. (*B*) Distribution of the peak harmonics selected as the index for the first and second formant for the five male participants. (*C*) Distribution for the five female participants. The somewhat smaller harmonic ranges for females are due to the higher average fundamental frequency of female speech. The mean fundamental frequency for male speakers was 109 Hz (SD = 10) and for female speakers 171 Hz (SD = 20) (the diagram in *A* is adapted from ref. 6).

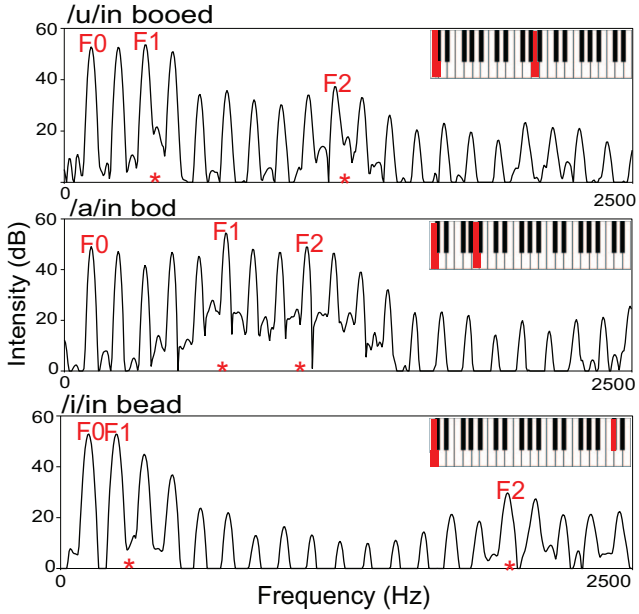

Fig. 2. Spectra of three different vowels uttered by a representative male speaker (the vowels are indicated in International Phonetic Alphabet nomenclature and phonetically). The repeating intensity peaks are the harmonics created by the varying energy in the air stream resulting from vibrations of the vocal folds (see Fig. 1*A*); the first peak indicates the fundamental frequency (F0). As in an ideal harmonic series, the intensity of successively higher harmonics tends to fall off exponentially; however, the resonances of the vocal tract above the larynx suppress some laryngeal harmonics more than others, thus creating the formant peaks. This differential suppression of the intensity in the air stream as a function of the configuration of the vocal tract generates the different vowel phones shown. The harmonic peaks of the first two formants are indicated by F1 and F2; asterisks are the formant values given by the linear predictive coding algorithm in Praat. (*Insets*) Keyboards showing that the intensity peaks in the first two formants often define musical intervals. Red keys indicate F1 and F2 values.

utterances of each of the vowels). Sixty-eight percent of these ratios fall on intervals of the chromatic scale (red bars), and all 12 chromatic intervals are represented over a span of 4 octaves. The black bars in Fig. 3 result from indexical pairs of harmonics for the first two formants that produce ratios not present in the chromatic scale, for example the first and second formants indexed by the 7th and 15th harmonics, respectively, which give a nonchromatic value of 1.10 [i.e., $\log_2(2.14)$; see *Methods* and *Discussion*].

Although the overall distribution of intervals on the chromatic scale is comparable in males and females, 75% of the intervals of female utterances were chromatic compared with 60% in male utterances (Table 1). The same prevalence of chromatic intervals was obtained if the harmonics adjacent to the maxima in the spectra were used as indices (see SI Fig. 6). Because up to a third of the intervals are not chromatic ratios (black bars in Fig. 3), the relationship between the first two formants in vowel phones only biases the distribution of interval ratios toward a representation of the chromatic scale.

To ensure that the biases favoring musical intervals are not peculiar to English, we analyzed the spectra of the six standard Mandarin vowels (13) uttered by native speakers of that language in the same way (see *Methods*). Mandarin has a very different history than English (14, 15) and represents a large group of languages that, unlike English, use tonality to convey semantic meaning (16, 17). Despite the Mandarin words in our data being spoken in the four different tones appropriate to their semantic meaning, the distribution of formant ratios derived from single words uttered by native Mandarin speakers was generally similar to the single English word data (Table 1).

We next assessed whether the bias toward a representation of chromatic scale ratios is specific to the modulation of the speech signal by the first two formants, or whether a similar bias would arise from an analysis of any harmonic series. This possibility arises because the ratios of the overtones of a harmonic series can, because they are integers, generate the 12 musical ratios that define the notes of the chromatic scale. SI Fig. 7 shows, however, that, when all possible comparisons of harmonic pairs up to the 26th harmonic are considered (the highest harmonic in the second formant; see Fig. 1), the

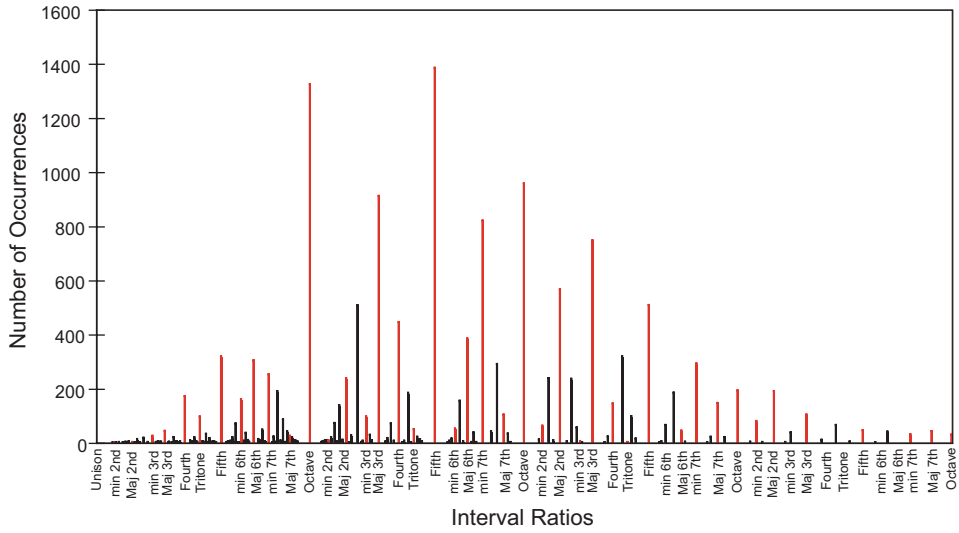

Fig. 3. Ratio relationships between the peak intensity of the first and second formants (see Fig. 2) for the eight vowels tested, compiled for the native English-speakers in the study. All 12 intervals of the chromatic scale in just intonation are represented (red bars); black bars show the frequency of occurrence of interval ratios that do not fall on chromatic scale tones. Sixty-eight percent of the occurrences are chromatic intervals (see *SI Text* for further discussion).

occurrence of musical intervals falls to 36% and the prevalence of nonmusical intervals increases to 64% (compare Fig. 3). Thus, the biases in Fig. 3 are specific to speech.

More significantly, Fig. 4*A* shows that, if the available range of harmonics were less than the range found in speech, the number of chromatic intervals represented would be diminished; in the example shown, 2 of the 12 chromatic intervals are missing altogether and 3 others are only weakly represented. Conversely, if the ranges of the harmonic peaks of the first two formants were greater than the range found in human speech (see Fig. 1), the intervals of the chromatic scale would be diluted by additional nonmusical intervals (black bars, Fig. 4*B*). Thus, the full range of chromatic intervals with minimal dilution by other intervals is specifically determined by the neural control of the vocal articulators in speech production.

Finally, to test whether these results with single word consonant-vowel-consonant utterances generalized to more natural forms of speech, we recorded both the native English and Mandarin participants speaking five ≈50-word monologues (see *Methods*). The results derived from an analysis of all voiced speech segments in the monologues were similar to the results for the single word utterances (Table 3 and Fig. 5).

Table 1. Comparison of the prevalence of chromatic formant ratios in English and Mandarin, based on single word analyses

	Percentage of chromatic intervals			
Interval	E-m	M-m	E-f	M-f
Unison	0.00	0.24	0.03	0.47
Octave	19.40	15.92	21.68	27.60
Fifth	13.24	19.66	15.41	20.48
Fourth	5.63	11.70	6.37	6.78
Maj third	14.44	9.75	6.40	15.65
Maj sixth	12.49	9.91	5.18	3.69
Min third	2.28	2.19	2.34	0.27
Min sixth	2.40	5.69	6.83	2.01
Tritone	2.07	2.92	3.86	3.56
Min seventh	8.27	8.04	14.62	10.48
Maj second	8.85	6.66	8.61	5.37
Maj seventh	4.80	3.74	6.57	2.22
Min second	6.12	3.57	2.08	1.41

The mean fundamental frequency for female Mandarin speakers was 206 Hz compared with 171 Hz for female English speakers; the average of the male speakers was 124 Hz and 109 Hz, respectively. These characteristics of the particular speakers in the samples resulted in a somewhat higher percentage of smaller F2/F1 ratios and a somewhat higher percentage of chromatic intervals in the Mandarin data. Interval rankings are in order of descending consonance (2, 21). E, English; M, Mandarin; m, male; f, female; Maj, major; Min, minor.

Discussion

Taken together, these results imply that the human preference for the specific intervals of the chromatic scale, subsets of which are used worldwide to create music (18–22), arises from the routine experience of these intervals during social communication by speech.

This conclusion is relevant to a number of unanswered questions in music, musicology, linguistics, and cognitive neuroscience. For example, if the source of musical intervals is indeed the formant ratios in speech, then the present results are pertinent to the longstanding argument in music about which of several tuning systems is "natural" (23). In so far as the observations here inform this argument, the observed ratios in speech spectra accord most closely with a just intonation tuning system. Ten of the 12 intervals generated by the analysis of either English or Mandarin vowel spectra are those used in just intonation tuning, whereas 4 of the 12 match the Pythagorean tuning and only 1 of the 12 intervals matches those used in equal temperament. The two anomalies in our data with respect to just intonation concern the minor second and the tritone. The interval ratio of the minor second defined by F2/F1 in speech is 1.0625 whereas, in just intonation (which is based on maintaining perfect fifths and major thirds in each octave), this interval is 1.0667. This difference occurs because 1.0667 is the ratio of 16:15, which does not occur in speech because the range of maximum intensity in the first formant peak extends only up to the 10th harmonic. Our value of 1.0625 for the minor second arises from formant ratios of 17:8, 17:4, and 17:2 (see Fig. 3 and *Methods*). Similarly, our value for the tritone is 1.400 whereas the just intonation value is 1.406. This difference arises because 1.406 is the ratio of 45:32, which again does not occur in speech, in this case because the range

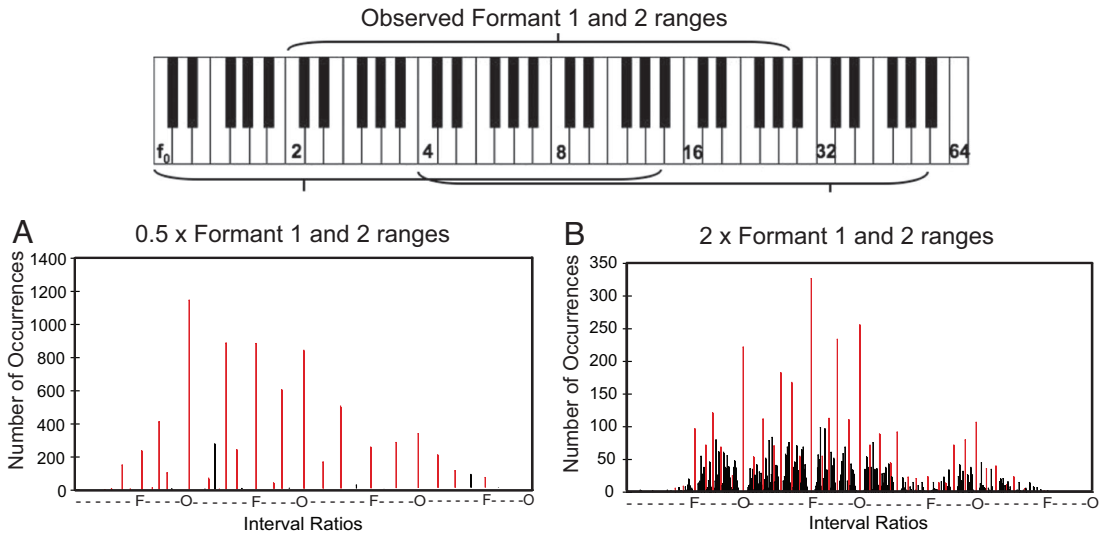

Fig. 4. Evidence that the ranges of the first two formants in speech specifically bias the distribution of formant ratios toward chromatic scale intervals. The diagram of the piano keyboard shows the ranges of the formants for the speakers in our data set (brackets); the numbers on the keyboard indicate harmonic overtones above the fundamental. If the formant ranges were lower than those found in speech (e.g., reduced by half as shown in *A*), then compared with emotionally neutral speech (Fig. 3), the intervals generated would represent only a subset of the chromatic scale (red bars; see *Results*). If the ranges were higher (e.g., doubled; as shown in *B*), all of the chromatic intervals would be represented (red bars), but their proportion would be diluted by additional nonchromatic intervals (black bars). The chromatic scale, however, is not optimized in the distribution; optimization would require formant peaks somewhat lower than those in our data (a reduction of ≈0.2 from the harmonic values generated by our subjects). An optimal representation of the chromatic scale would thus entail slightly higher fundamental frequencies of voiced phones, which presumably occur in the more energized natural speech that we routinely experience. ''F'' and ''O'' on the abscissa denote the position of fifths and octaves; the ticks are at chromatic intervals (see Fig. 3).

of maximum intensity in the second formant peak extends only up to the 26th harmonic. The values 1.400 in speech arise from the F2/F1 ratios in the database of 7:5, 14:5, 21:5, 14:10, and 21:10 (see again Fig. 3 and *Methods*). In summary, just intonation tuning closely fits the chromatic scale defined by speech data. The fact that instruments in just intonation tuning are widely agreed to sound ''brighter'' than in the equal temperament tuning used for the last three centuries (9) (a compromise that allows multiple instruments to play pieces that include notes in more than one key) is in keeping with our conclusion that the chromatic scale arises from formant ratios in speech.

A second fascinating question is whether the tonal preferences in the music of a culture can be rationalized in terms of the formant relationships of the voiced speech sounds prevalent in the relevant language. If the chromatic scale derives from experience with the formant relationships used to elicit different phonemes, then the speech sounds of a particular language might be expected to influence the subsets of the chromatic scale used in the music of that culture (24–27). Analyses of cultural scale preferences in relation to the spectral characteristics of the language or languages of a given culture should be possible using the approach described here.

A third question of interest concerns the widespread preference across cultures for diatonic (seven-note) and pentatonic (five-note) subsets of the chromatic scale in creating music (18–22, 27). The pentatonic scale in particular is the basis for much ethnic (''folk'') music worldwide. It is noteworthy in this respect that, of the chromatic intervals in our data, ≈70% are components of the pentatonic scale and ≈80% of the diatonic

Table 2. Comparison of the prevalence of chromatic formant ratios in English and Mandarin based on the monologue analyses

	Percentage of chromatic intervals			
Interval	E-m	M-m	E-f	M-f
Unison	0.54	0.36	0.33	0.44
Octave	22.37	21.55	23.26	22.51
Fifth	16.78	16.93	17.76	20.66
Fourth	7.80	6.66	8.94	4.56
Maj third	11.95	13.20	11.53	18.36
Maj sixth	6.53	8.40	8.22	4.32
Min third	3.15	2.29	1.70	0.60
Min sixth	2.86	1.94	2.87	1.34
Tritone	2.43	2.09	1.78	0.89
Min seventh	10.34	9.26	11.18	14.39
Maj second	7.62	9.08	9.07	9.77
Maj seventh	4.84	4.86	2.54	1.78
Min second	3.18	3.63	1.07	0.78

E, English; M, Mandarin; m, male; f, female; Maj, major; Min, minor.

Table 3. Example of one of the monologues in English and Mandarin translation

The building I work in is very old. It was built in the fifties, is four stories and has high ceilings with electric fans. There are no elevators, however. I work on the fourth floor, so I have a lot of stairs to climb. That gives me a little bit more of the exercise I should be doing.

我在一个非常古老的楼里工作。 这个楼在50年代建成，它有四层高，高高的天花板配有吊扇。 由于没有电梯，每天我要爬很多的楼梯到四层的办公室。当然多一些锻炼，对我来说也不错。

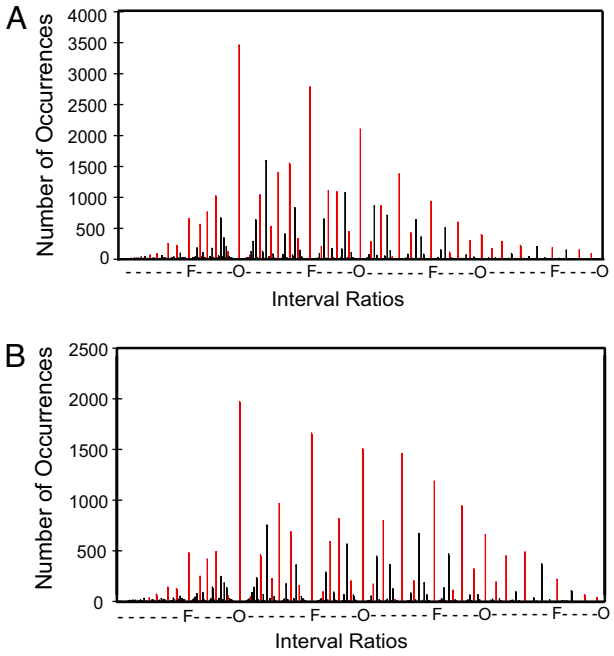

Fig. 5. Ratio relationships between the peak intensity of the first and second formants from the American English (*A*) and Mandarin (*B*) monologues, compiled from all of the participants. All 12 intervals (red bars) of the chromatic scale in just intonation are represented in both speech databases; black bars show the frequency of occurrence of interval ratios that do not fall on chromatic scale tones (see also Tables 1 and 2).

scale (see SI Table 4). This prevalence suggests that the general preference for diatonic and pentatonic scales arises from the greater familiarity with these formant ratios in the speech of any language.

Further questions that can be explored in these terms arise from other aspects of the phenomenology of musical scales and their impact on listeners. For example, could the different emotional impact of major and minor musical scales be based on variations in the predominant intervals among vowel formants uttered in different physiological states (e.g., excitement versus the subdued physiology that characterizes sadness)? And what, in these terms, is the significance of the tonic anchor in musical composition and performance?

Finally, it will be of interest to examine in this framework how formant relationships in the vocalizations of nonhuman primates and other animals compare with those in humans, and what such evidence could indicate about the origins of both speech and music.

Methods

Recording. Speech was recorded from 10 native speakers of American English (five males and five females) who ranged in age from 18–68 years of age and had no known speech or hearing pathology. The participants gave informed consent, as required by the Duke University Health System. Each participant was asked to repeat eight words that had a different vowel embedded between the consonants "b" and "d" (i.e., bad, bod, bead, bed, bid, booed, bud, and "bood," the last pronounced like the word "good"). These vowels (/i, I, ɛ, æ, a, Λ, U, u/) and consonants (/b, d/) were chosen based on the rationale of Hillenbrand and Clark (28) (in particular, vowel phone intelligibility is maximized by this consonant framing). The words were spoken at a conversational level of intensity (≈70 dB) and speed (mean duration, 523 ms; SD = 159 ms) in an emotionally neutral manner. Each word was repeated seven times; by analyzing only the central five of these utterances, we could avoid onset and offset effects. Participants paused for 30 s between saying each of four differently ordered lists of the words. After a break of at least 30 min, this entire procedure was repeated four more times; thus, we obtained 100 samples of each of the eight words for each participant. In the Mandarin control, only six words representing the major vowels in this language (ba, ge, bo, bi, du, and jü) (29) were used; the words were spoken by three male and three female native speakers ranging in age from 22–31 years of age. The procedure was the same as for English except that each word was uttered in each of the four major tones used in Mandarin (the fifth neutral tone form was not included because it is rarely used, comprising only ≈6% of vowel utterances in Mandarin speech (30). Both the English and Mandarin speaking participants also read aloud five monologues[‡] that contained ≈50 words each (Table 3), recording each monologue twice in an emotionally neutral manner.

All utterances were recorded in a closed, sound-attenuating chamber by using an Audio-Technica AT4049a omnidirectional capacitor microphone fed into a Marantz (Martel Electronics, Yorba Linda, CA) PMD670 solid-state recorder. The participants followed a series of simple instructions presented graphically, and the quality of their performance was monitored remotely. Sound files were saved to a Scandisk 1 flash memory card in uncompressed digital .wav format at a sampling rate of 22.05 kHz, and transferred from the flash memory card to a Dell Dimension 9150 computer for analysis.

Analysis. The recorded samples were analyzed by using Praat software (v.4.5) (32). A Praat script was used to generate a text grid and to automatically mark pauses at the onset/offset of each word; vowel identifier and positional information were then inserted manually for each utterance. The text grid was stored with the associated .wav file, and a second script was implemented to extract values (in hertz) for the fundamental frequency, as well as for the first and second formants from a 50-ms segment at the midpoint of each vowel utterance (thus yielding one value for each word uttered; 50 ms is the standard integration window in Praat). The frequency range analyzed was individually adjusted for male and female speakers (5 formants >≈5,000 Hz for males, but up to ≈5,500 Hz for females). To extract the formant values, Praat uses a Gaussian-like window to compute the linear predictive coding coefficients using the algorithm in ref. 33.

For the monologue data, Praat's pitch- and formant-listing utilities were used to extract and time-stamp the F0 (if present), F1, and F2 values at 10-ms intervals. Tracking the formants in this way is necessary in natural speech because of the greater degree of coarticulation compared with the somewhat artificial utterance of single words. The frequencies that define the formants vary less over the mid-region of the vowel nucleus, where the effects of coarticulation are minimal (34). Standard pitch settings were used and the frequency range was set at 75–600 Hz. The formant settings were adjusted in the same manner as was used for the single word condition. Any 10-ms time interval that contained no F0 was removed from the data.

For both the word and monologue data, the nearest harmonic peak to the underlying formant maximum given by Praat was used as an index of the formants: the formant value assigned by linear predictive coding was divided by the fundamental frequency, and the result was rounded to the nearest integer. The

[‡]McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M., Stroeve, S., International Speech Communication Association Tutorial and Research Workshop (ITRW) on Speech and Emotion, September 5–7, 2000, Newcastle, Northern Ireland, U.K., pp. 207–212.

ratios of the indices of the first two formants were then calculated as B/A where B = the formant 2 harmonic index and A = formant 1 harmonic index [the data were plotted as $\log_2(B/A)$, as is conventional]. Ratios were counted as chromatic if they corresponded to just intonation values for the chromatic scale (see *Discussion*).

Octave Collapse. The perceived similarity of tones an octave apart is so pronounced that it is termed octave equivalence (31). On this basis, we collapsed the results in Tables 1 and 2 into a single octave to allow a more direct comparison of the distribution of intervals found in speech in the two languages being compared.

We thank Sheena Baratono, Nigel Barrella, Catherine Howe, Reiko Mazuka, Rich Mooney, Elliott Moreton, and Jim Voyvodic for much helpful criticism and advice; Zhang Zheng for assistance in translating the English monologues into Mandarin; and Yale Cohen, Mark Tramo, and Robert Zatorre for thoughtful and constructive reviews.

1. Fletcher NH (1992) *Acoustic Systems in Biology* (Oxford Univ Press, New York).
2. Schwartz DA, Purves D (2004) *Hear Res* 194:31–46.
3. Schwartz DA, Howe CQ, Purves D (2003) *J Neurosci* 23:7160–7168.
4. Petersen GE, Barney HL (1952) *J Acoust Soc Am* 24:175–184.
5. Stevens KN, House AS (1961) *J Speech Hear Res* 4:303–320.
6. Purves D, Augustine GJ, Fitzpatrick D, Hall WC, LaMantia A-S, McNamara JO, Williams SM (2004) *Neuroscience* (Sinauer, Sunderland MA), 3rd Ed.
7. Ladefoged P (1962). *Elements of Acoustic Phonetics* (Univ of Chicago Press, Chicago).
8. Hillenbrand J, Getty LA, Clark MJ, Wheeler K (1995) *J Acoust Soc Am* 97:3099–3111.
9. Iivonen A (1987) in *Neophilologica Fennica: Société Neuphilologischer Verein 100 Jahre*, *Mémoires de la Société Néophilologique de Helsinki XLV*, ed Kahlas-Tarkka L, pp 87–119.
10. Azami Z (1992) *Rapport d'Activités de L'institut de Phonétique* (Universite Libre de Bruxelles, Brussels), Vol 28.
11. Reuter M (1971) *Festskrift till Olav Ahlbeck* 28:240–249.
12. Gu Z, Mori H, Kasuya H (2003) *Acoust Sci Tech* 24:192–193.
13. Howie J (1976) *Acoustical Studies of Mandarin Vowels and Tones* (Cambridge Univ Press, Cambridge, UK).
14. Maddieson I (1978) in *Universals of Human Language: Phonology*, ed Greenberg JH (Stanford Univ Press, Stanford, CA), Vol 2.
15. Hombert J, Ohala JJ, Ewan WG (1979) *Language* 55:37–58.
16. Fromkin VA, ed (1978) *Tone: A Linguistic Survey* (Academic, New York).
17. Xu Y (1997) *J Phonetics* 25:61–83.
18. Nettl B (1956) *Music in Primitive Culture* (Harvard Univ Press, Cambridge, MA).
19. Burns EM (1999) in *The Psychology of Music*, ed Deutsch D (Academic, San Diego), 2nd Ed, pp 215–264.
20. Kallman HJ, Massaro DW (1979) *Percept Psychophys* 26:32–36.
21. Krumhansl CL (1980) *Cognitive Foundations for Musical Pitch* (Oxford Univ Press, New York).
22. Justus T, Hutsler J (2005) *Music Percept* 23:1–27.
23. Isacoff S (2001) *Temperament: The Idea that Solved Music's Greatest Riddle* (Knopf, New York).
24. Patel AD, Iversen JR, Rosenberg JC (2006) *Empir Musicol Rev* 1:166–169.
25. Patel AD (2003) *Nat Neurosci* 6:674–681.
26. Wenk BJ (1987) *Linguistics* 25:969–981.
27. Krumhansl CL (2000) *Music Percept* 17:461–479.
28. Hillenbrand JM, Clark MJ (2000) *J Acoust Soc Am* 109:748–763.
29. Chao Y (1932) *Bull Inst Hist Philos* 1(Suppl):105–156.
30. Suen CY (1979) in *Linguistic Series: Coling*, ed Horecky J (Academia North–Holland, Amsterdam), Vol 82.
31. Burns EM, Ward WD (1978) *J Acoust Soc Am* 63:456–468.
32. Boersma P, Weenik D (2006) *Praat: Doing Phonetics by Computer*, Version 4.5, www.praat.org.
33. Burg JP (1978) *A New Technique for Time Series Data* (IEEE Press, New York), pp 252–255.
34. Turner GS, Hutchings DT, Sylvester B, Weimer G (2003) *J Acoust Soc Am* 113:1965–1974.