

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. 1A). 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. 1B 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. 1C).

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 $F1 \times F2$ 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.

© 2007 by The National Academy of Sciences of the USA

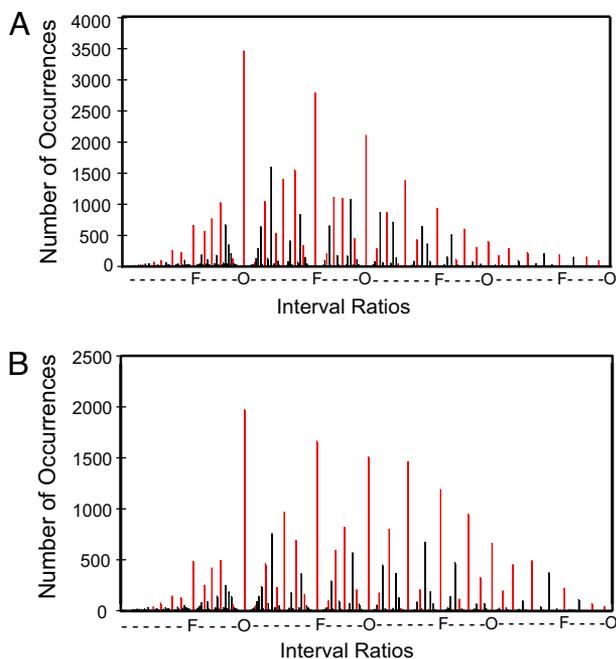


Fig. 5. Ratio relationships between the peak intensity of the first and second formants from the American English (A) and Mandarin (B) monologues, compiled from all of the participants. All 12 intervals (red bars) of the chromatic scale in just intonation are represented in both speech databases; black 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, bood, bud, and “bood,” the last pronounced like the word “good”). These vowels (/i, I, ε, 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 $\approx 6\%$ 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 $> \approx 5,000$ Hz for males, but up to $\approx 5,500$ 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: Soci t  Neophilologischer Verein 100 Jahre, M moires de la Soci t  Neophilologique de Helsinki XLV*, ed Kahlas-Tarkka L, pp 87–119.
10. Azami Z (1992) *Rapport d'Activit s de L'institut de Phon tique* (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.