bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Spectro-temporal acoustical markers differentiate speech from song across cultures

Philippe Albouy[1,2,3]*, Samuel A. Mehr[2,4,5], Roxane S. Hoyer[1], Jérémie Ginzburg[6] and Robert J. Zatorre[2,3,7]*

[1] CERVO Brain Research Centre, School of Psychology, Laval University, Québec, Canada

[2] International Laboratory for Brain, Music and Sound Research (BRAMS), Montréal, Canada

[3] Centre for Research in Brain, Language and Music and Centre for Interdisciplinary Research in Music, Media, and Technology; Montréal, Canada

[4] Haskins Laboratories, Yale University, New Haven, CT 06511, USA

[5] School of Psychology, University of Auckland, Auckland, New Zealand

[6] Lyon Neuroscience Research Center, CNRS, UMR5292, INSERM, U1028 - Université Claude Bernard Lyon 1, Lyon, F-69000, France

[7] Cognitive Neuroscience Unit, Montreal Neurological Institute, McGill University, Canada

*Corresponding authors. E-mail: robert.zatorre@mcgill.ca Montreal Neurological Institute, 3801 University St., Montreal, QC Canada H3A 2B4; philippe.albouy@psy.ulaval.ca CERVO brain research center, 2601 Chemin de la Canardière, Laval University, QC, G1J 2G3

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Humans produce two primary forms of vocal communication: speaking and singing. What is the basis for these two categories? Is the distinction between them based primarily on culturally specific, learned features, or do consistent acoustical cues exist that reliably distinguish speech and song worldwide? Some studies have suggested that important aspects of music can be distinguished from speech based on spectro-temporal modulation patterns, but this conclusion is based on Western music, leaving open the question of whether such a principle may apply more globally. Here, we studied the spectro-temporal modulation patterns of vocalizations produced by 369 people living in 21 urban, rural, and small-scale societies distributed across six continents. We show that specific ranges of spectral and temporal modulations differentiate speech from song in a consistent fashion, and that those ranges overlap within categories and across societies. Machine-learning analyses confirmed that this effect was cross-culturally robust, with vocalizations reliably classified solely from their spectro-temporal modulation patterns across all 21 societies. Listeners unfamiliar with most of the cultures could also classify the vocalizations, with similar accuracy patterns as the machine learning algorithm, indicating that the spectro-temporal cues used by the classifier are similar to those used by human listeners. Thus, the two most basic forms of human vocalization appear to exploit opposite extremes of the spectro-temporal continuum in a consistent fashion across societies. The findings support the idea that the human nervous system is specialized to produce and perceive two distinct ranges of spectro-temporal modulation in the service of the two distinct modes of human vocal communication.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Human vocal communication differs from that of other species in that humans vocalize in two distinct modes: speech and song [1-4]. A great deal of work has documented the remarkable variability in both the structural features of speech and song, as well as their acoustical manifestations [5-9], but debate continues about whether the two categories may be distinguished across societies on the basis of acoustical features alone. Speech and song are produced by the same vocal tract, yet each makes distinct demands on musculature, breathing, and motor control mechanisms [10], raising the possibility that certain acoustical cues could serve as markers of each category [11]. However, cross-cultural variability in the forms of music is large, and distinctions between speech and song within cultures are far from clear [12-14], so that such a claim is difficult to address. Indeed, even if speech and singing reliably exist as separate, recognizable entities, their cognitive representation could depend mostly on learned regularities that are particular to each cultural group.

One source of difficulty in comparing speech and song is that they each form part of broader communication systems of language and music, respectively. These systems share certain cognitive operations (e.g., recursive syntactical operations), but also differ in important ways (e.g., the hierarchical organization of metrical patterns) [15,16]. Pitch variation in music tends to be more discrete than in speech [17], leading to the formation of hierarchical tonal organization [18], which may be a fundamental property of music worldwide [7]. But it remains unclear whether such descriptive differences represent acoustic phenomena invariant enough to form a sufficient basis for categorization across different musical and linguistic systems, or, instead, are merely associated with the two

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

domains, perhaps mainly in those cultures that happen to be well-studied in the cognitive sciences [19].

Recent developments in neurophysiology and cognitive neuroscience offer a rigorous framework for testing how speech and song could differ. Complex sounds can be characterized according to the distribution of their spectro-temporal modulation power [20]. Neurons in auditory regions across various species can be described in terms of their spectro-temporal receptive fields, which have been shown to constitute an efficient coding scheme for complex acoustical patterns [21-23]. Moreover, spectro-temporal tuning functions correspond well to the most relevant acoustical features that characterize different animals' communicative signals, including birdsong [23], cat meows [24], and monkey calls [25], indicating a match between the acoustics of important sounds in the environment and the neural hardware needed to process them.

Might spectro-temporal modulation content constitute a fundamental, and sufficient difference to account for how speech and song differ from one another? Acoustical analysis shows that speech tends to contain faster temporal modulations as compared to music from Western genres [11,26], and temporal modulation cues are well-known to be sufficient for speech perception, even when spectral modulations are degraded [27]. Conversely, degradation of spectral modulations abolishes the perception of melodic content in song, while leaving speech comprehension intact, whereas degradation of temporal modulations renders the speech content of songs incomprehensible but has little effect on the melody [28]. These findings dovetail well with the idea that spectral and temporal features are processed in partially distinct neural populations within [29,30] and across the two hemispheres [28,31,32].

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Taken together, these results suggest that speech and song exploit different ends of the spectro-temporal continuum. But such a conclusion suffers from a major limitation, because although the high temporal rate of speech has been confirmed for many distinct languages [33], the spectrotemporal features of music have only been characterized in a limited Western musical repertoire, which is not necessarily representative of all human musical systems. Whether the role of spectro-temporal modulations in distinguishing speech from song is an idiosyncracy of some cultures, or whether it represents a more fundamental aspect of the biology of human communication — as one would expect, given the fundamentally different functional roles of speech [34,35] and music [2] in human evolution, and their partly distinct neural representations — is the question we address in this paper.

Specifically, we tested whether distributions of spectro-temporal modulation power in speech and song are sufficient to distinguish the two vocalization types within and across 21 societies sampled from all inhabited continents and comprising small-scale, rural, and urban societies. The recordings, produced in 18 languages from 12 language families, were gathered from native speakers of each language who each lived in the society where the recording was gathered (see [5] for full details and supplementary Table 1 and Supplementary Figure 1). Three hundred sixty-nine people from these societies were asked (i) to speak in a casual, ordinary fashion, on a mundane topic directed to the experimenter (e.g., describing their daily routine); and (ii) to sing a song of their choice, with the only requirement being that the song was not intended to be infant-directed.[1] Whether the vocalization was considered to be an example of speech or song was

[1] Additional recordings in the corpus were infant-directed, as the corpus has previously been used to study the acoustic features of infant-directed vocalizations, as in [5]

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

therefore determined by the person producing the vocalization, and not imposed by the researcher.

We predicted (i) that if speech and song are characterized by distinct spectro-temrporal modulation signatures, we should be able to observe distinct distributions of these patterns with appropriate acoustical analysis; (ii) that if such differences are truly common across societies, we should observe substantial overlap in the distribution of spectro-temporal modulation power for each category across all societies studied; (iii) that if these acoustical markers are sufficient to categorize the two classes of vocalizations, then a machine-learning classifier should be able to determine which sample corresponds to speech or song with adequate accuracy, based solely on their spectro-temporal modulation profile; (iv) that the information most used by the classifier should correspond to the spectro-temporal signatures derived from the initial acoustical analysis; and (v) that listeners unfamiliar with the language or music of the different societies should nevertheless be able to correctly classify speech and song, with a similar ordering of accuracy across samples as the machine learning classifier, if human judgments are based on spectro-temporal cues.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Results

We decomposed the acoustical signal of the vocalization samples using the Spectro-Temporal Modulation framework (Figure 1). Spectro-temporal modulations patterns for singing and speaking samples were extracted (ModFilter algorithm [36]) for each vocalization (see Methods and [28] for similar procedure), and then used for univariate and multivariate analyses. The identical pipeline was used for both speech and song samples, thus avoiding any kind of bias in the procedure.

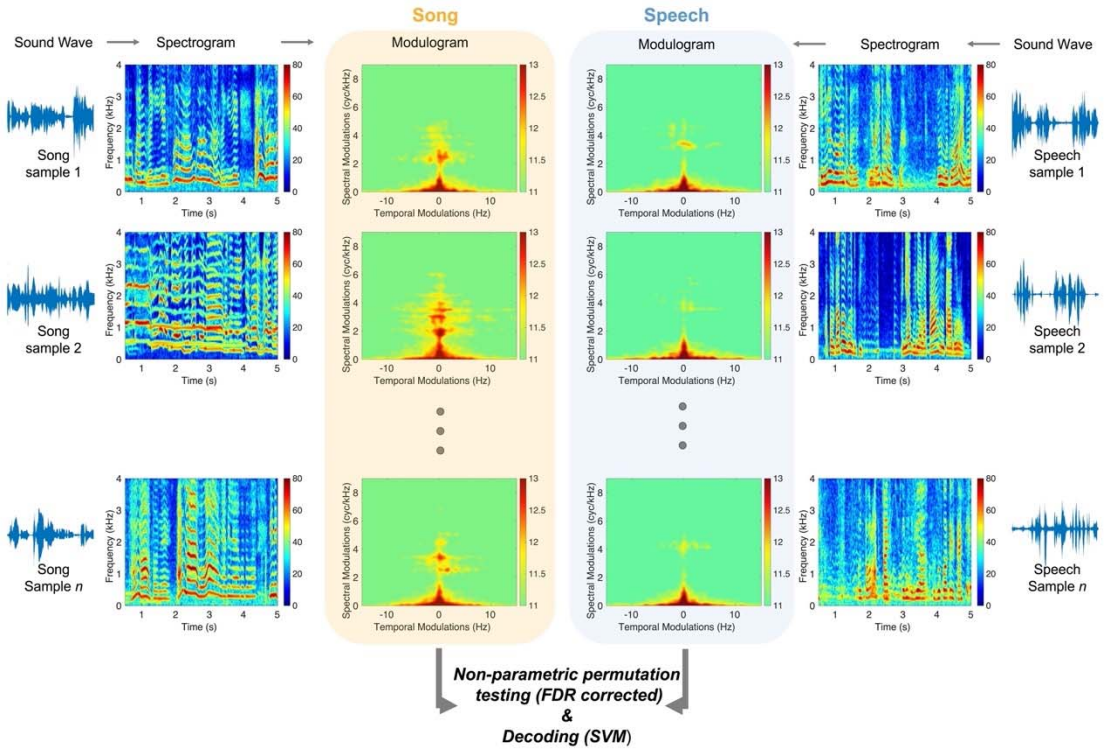

Figure 1. Extraction of spectro-temporal modulations patterns for singing and speaking vocalization samples. *Sound waves, spectrograms, and modulograms of representative vocalizations (here, in recordings from the Nyangatom of Ethiopia; left panel: song, right panel: speech) revealing the acoustic complexity of the song and*

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

speech samples. For each sample we extracted the spectro-temporal modulation patterns. We then contrasted the song and speech modulation patterns using non-parametric permutation statistics (FDR-corrected) and used the modulation patterns data as features to perform a 2-class SVM decoding of music and speech samples (see Figure 2).

We contrasted the spectro-temporal modulation patterns of song and speech vocalizations using non-parametric permutation statistics with FDR correction in the spectral and temporal domains (as implemented in FieldTrip[37] - see Methods). This analysis revealed two hotspots of increased spectral modulations in song as compared to speech samples (10^5 permutations, FDR corrected, $p < .0001$): hotspot 1: peak at 3.71 cyc/kHz in the spectral domain and 0.66Hz in the temporal domain; hotspot 2: peak at 6.86 cyc/kHz in the spectral domain and -0.66Hz in the temporal domain (note that human speech is symmetric between positive and negative temporal modulation frequencies,[36] which correspond to increasing and decreasing frequency trajectories, respectively). We also detected three hotspots of increased temporal modulation in speech as compared to song: hotspot 1: peak at 7.99 Hz in the temporal domain and 0 cyc/kHz in the spectral domain; hotspot 2: peak at -7.99 Hz in the temporal domain and 0 cyc/kHz in the spectral domain; hotspot 3: peak at 4.49 Hz in the temporal domain and 5.11 cyc/kHz in the spectral domain.

To assess the consistency of this effect across societies we generated a heatmap illustrating the overlap in number of societies that display a significant effect in the hotspots identified in Figure 2A. This analysis (Figure 2 B.) revealed that 20/21 societies

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

showed a significant increase of spectral modulations in singing samples vs speech samples at 3.71 cyc/kHz (in the spectral domain) and 0.66Hz (in the temporal domain). Moreover, 21/21 societies showed a significant increase of temporal modulations in speech samples vs singing samples at 7.83 Hz (in the temporal domain) and 0.09 cyc/kHz (in the spectral domain). The robustness of this effect was also confirmed with a k-means clustering analysis performed on the coordinates in the spectro-temporal domain of the statistical peaks of each society for the contrast song vs. speech. Note that for this analysis the absolute values of temporal modulations were used, as human vocalization are symmetric between positive and negative temporal modulation[36]. This analysis revealed 2 clear clusters with centroids at: cluster 1: 3.38 cyc/kHz (in the spectral domain) and -0.16 Hz (in the temporal domain, Figure 2 C.) and cluster 2: 0.04 cyc/kHz (in the spectral domain) and 6.99Hz in the temporal domain.

To confirm the cross-cultural robustness of these effects, we then used a Support Vector Machine (SVM) classifier with fieldsite-wise k-fold cross-validation to classify song and speech vocalization samples, using only the spectro-temporal modulation patterns as input features (see Methods). This approach provides a strong evaluation of cross-cultural regularity: the model is trained only on data from 20 of the 21 societies to predict whether each vocalization in the 21st society is song or speech. The procedure is repeated 21 further times, with data from each society being successively held out, to estimate the classification performance across the full set of societies. The summary of the SVM's performance (average of all models) reflects, corpus-wide, the degree to which song and speech spectro-temporal modulation patterns are stereotyped, because high classification performance can only result from high cross-cultural regularities.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The models significantly classified song and speech above chance ($t(20) = 11.7$, $p < .001$; Cohen's $d = 2.55$; - Figure 2D; accuracy = 80.5% ± 11.9 (SD); sensitivity = 77.28% ± 20.9, specificity = 83.72% ± 9.3; ROC curves for each society are presented in Figure 2E). Evaluating classification performance within the recordings in each fieldsite showed a high degree of cross-cultural regularity, with the performance in all 21 fieldsites significantly above chance level (Figure 2.D.E), even though accuracy varied across different sites.

We then investigated what STM features the model relied upon to discriminate song and speech spectro-temporal modulation patterns. For each classifier, we extracted the feature weights to estimate their relative importance (z-scored, averaged across societies). We identified two spectro-temporal patterns (*i.* 7.83 Hz in the temporal domain and 0.09 cyc/kHz in the spectral domain and; *ii.* 0.83 Hz in the temporal domain and 3.16 cyc/kHz in the spectral domain) showing substantial di erences in the features the model relied upon to reliably classify speech and song across societies (Figure 2.F.). Furthermore, the two regions of the spectro-temporal modulation space most critical to the classifier's performance correspond well to the acoustical differences identified in the initial analyses (Figure 2 A. and B) as shown by the *a posteriori* overlap observable in Figure 2 C. and F.).

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

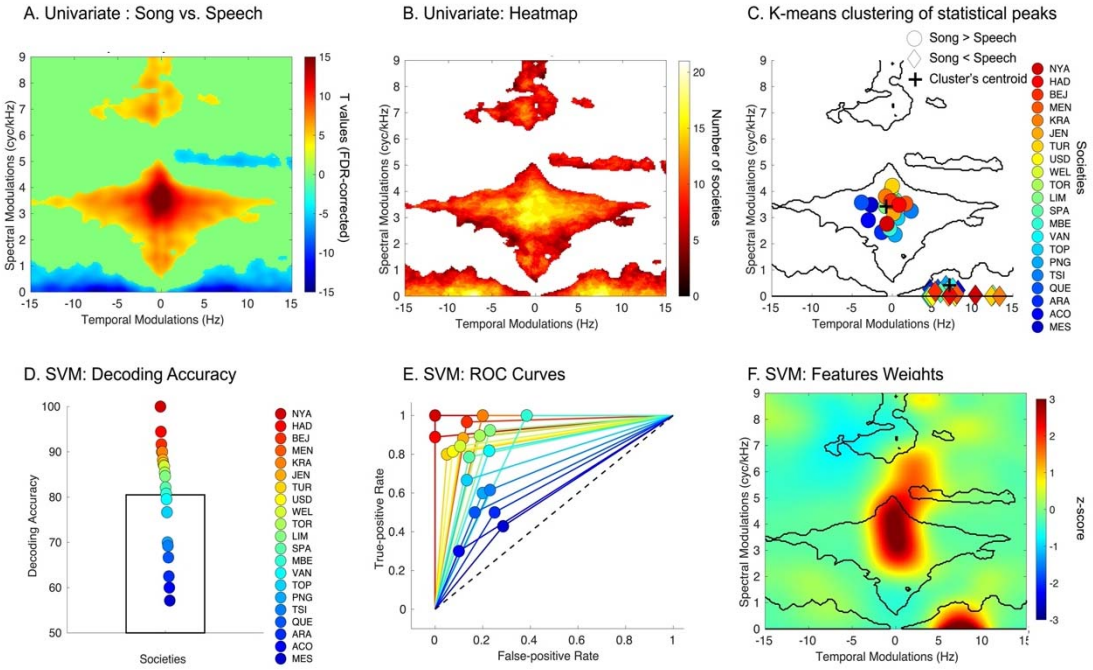

Figure 2. Cross-Cultural spectro-temporal markers of song vs. speech identified with univariate analyses and machine learning. ***A****. Song vs. Speech contrast in the modulation power spectrum domain across all societies (p< .001, FDR-corrected).* ***B****. Heatmap (smoothed) depicting the number of societies showing a significant effect in the clusters identified in (A.). Each value reports a numeric count, with larger counts associated with yellow/white coloring.* ***C****. K-means clustering of statistical peaks; dots represent each society. Dark lines illustrate the boundaries of the significant effects presented in (A.).* ***D****. Fieldsite-wise cross-validated (21 societies) support vector machine decoding accuracy (chance level: 50%). The colored dots represent the accuracy for each society (sorted as a function of accuracy with a jet colormap).* ***E****. Receiver operating characteristic curve (ROC) for each society (same color code as in A.). Black dashed line represents the chance level.* ***F****. Normalized feature weights in the modulation power spectrum domain showing features with the largest influence (z-score, average of*

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the 21 classifiers) for the classifier. Dark lines illustrate the boundaries of the significant effects presented in (A.).

To confirm the reliability of these findings, and to verify that the accuracy rates were not inflated by any incidental similarities between the samples used for cross-validation, we repeated the same analysis with four alternative cross-validation strategies, using the same cross-validation procedure but doing so across countries, language families, world subregions, and world regions instead of fieldsites (societies). The results robustly replicated in all cases with large effect sizes:

i) Countries ($t(17) = 11.7$, $p < .001$; Cohen's $d = 2.75$; - Figure 3A; accuracy = 79.7% ± 10.8 (SD); sensitivity = 76.4% ± 20.2 , specificity = 83.1% ± 10.9

ii) Language Families ($t(14) = 12.0$, $p < .001$; Cohen's $d = 3.10$; - Figure 3B; accuracy = 82.2% ± 10.4 (SD); sensitivity = 81.7% ± 17.9, specificity = 82.6% ± 11.5

iii) World subregions ($t(13) = 10.1$, $p < .001$; Cohen's $d = 2.69$; - Figure 3C; accuracy = 79.4% ± 10.9 (SD); sensitivity = 76.4% ± 22.4, specificity = 82.4% ± 11.4

iv) World regions ($t(5) = 12.4$, $p < .001$; Cohen's $d = 5.05$; - Figure 3D; accuracy = 82.5% ± 6.4 (SD); sensitivity = 80.3% ± 14.1, specificity = 84.7% ± 5.2

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

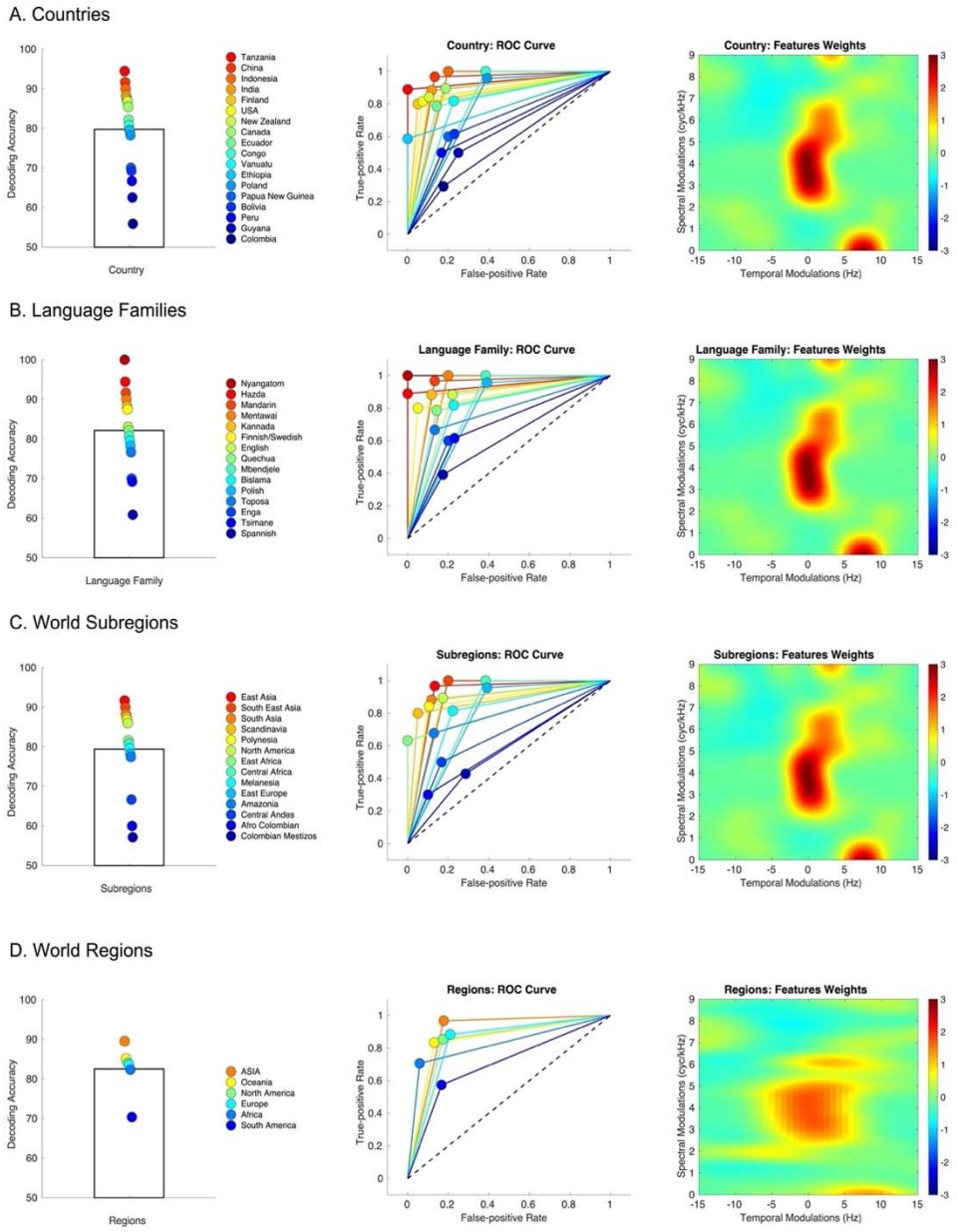

Figure 3. Cross-cultural regularities across countries, language families, world subregions, and world regions identified with machine learning. A. *Left Panel: Country-wise cross-validated decoding accuracy (chance level – 50%). The colored dots represent the performance accuracy for each country (sorted as a function of accuracy with a jet colormap). Middle Panel: Receiver operating characteristic curve (ROC) for*

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

each country (same color code as in the left panel). Black dashed line represents the chance level. Rigth Panel: Features weights in the MPS domain showing features with the largest influence (z-score, average of the 18 classifiers). B. C. D. same as (A.) for countries, world subregions, and world regions respectively

Behavioral Analysis

We then studied naïve listeners' sensitivity to these spectro-temporal features. We played the song and speech recordings to 80 individuals who were asked to rank, as rapidly as possible on a 5-point scale, whether each speaker was singing (code 1) or speaking (code -1) (see Figure 4A). These primarily French-speaking listeners from Quebec (Canada) and France were presumably unfamiliar with the languages or music of most of the societies from which the sounds were recorded. Their judgments were highly accurate, with large effect sizes for both Song ($t(79)= 93.8$, $p < .001$, Cohen's $d = 10.98$) and Speech ($t(79)= -25.1$, $p < .001$, Cohen's $d = -2.87$; Figure 4 B.)).

To test whether these listeners were using the spectro-temporal cues that distinguished song from speech in the prior analyses, we tested if the features identified on the modulation power spectrum (MPS; see Figures 2 and 3) could predict their behavioral ratings. To do so we computed the normalized difference between Song MPS and Speech MPS and between Song and Speech behavioral ratings (with a positive score representing large difference between song and speech ratings) for each of the 369 vocalizations/speakers (see Methods). We then computed the correlation (FDR corrected, $p< .05$, Figure 4C) between these difference scores and observed i) a positive relationship between increased spectral modulation for song relative to speech (-0.33 Hz in the temporal domain and 3.62 cyc/kHz in the spectral domain- – see Figure 4CD) and

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

positive behavioral difference scores (corresponding to large difference ratings between song and speech) and ii) a negative relationship between decreased temporal modulation for song relative to speech (4.83 Hz in the temporal domain and 0.04 cyc/kHz in the spectral domain) and positive behavioral difference scores (corresponding to large difference rating between song and speech– see Figure 4 C. D.).

To test the consistency of our listeners' inferences across cultures, we computed the fieldsite-level behavioral ratings. Within each of the 21 societies, listeners' judgments were accurate, again with large effect sizes, for both Song ($t(20)= 28.7$, $p < .001$, Cohen's $d = 6.27$) and Speech $t(20)= -45.6$, $p <.001$, Cohen's $d = -9.94$; - Figure 4 E., see supplementary Figure 2 for the same analysis for countries, language families, world sub regions and world regions).

Finally, to confirm that human judgments were based on similar spectro-temporal cues as those identified in the MPS, we investigated whether these listeners unfamiliar with the different societies were identifying speech and song samples with a similar ordering of accuracy across samples as the machine learning classifier (see Figure 2). To do so, we computed the correlation between SVM decoding accuracy (Figure 2 D.) and the normalized difference between Song and Speech behavioral ratings computed within each society. As expected, this analysis revealed that decoding accuracy of the classifier was positively correlated with the normalized behavioral scores ($r(20) = .73$, $p< .001$, Figure 4 F.).

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

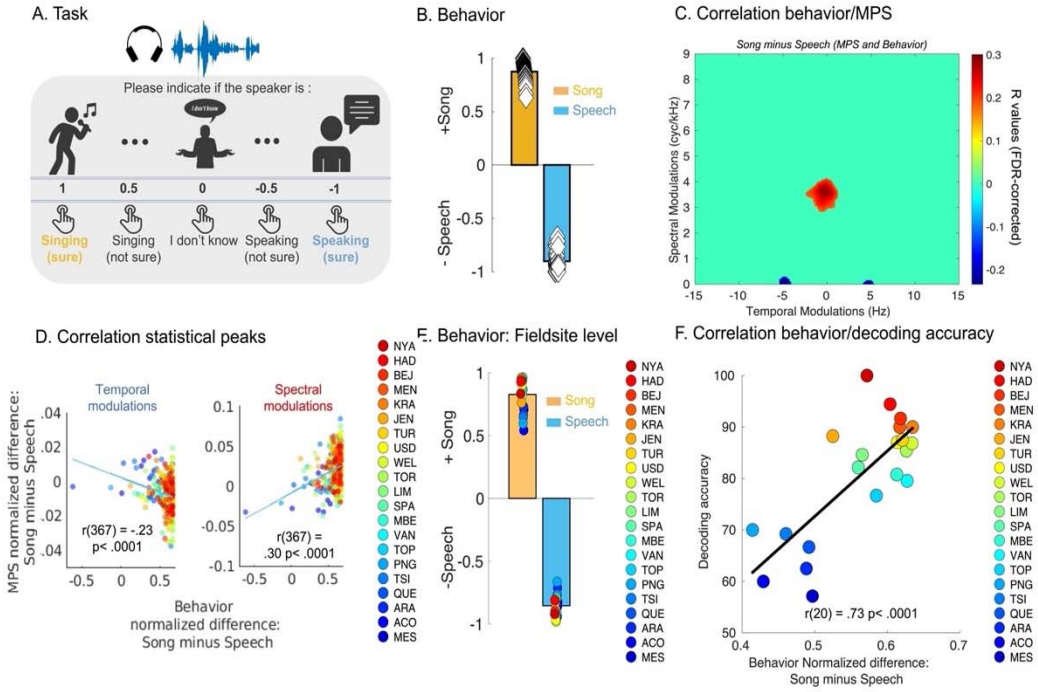

Figure 4: Naïve listeners distinguish song from speech vocalizations across cultures. A. *Behavioral task: 80 individuals were asked to rank, as rapidly as possible on a 5-point scale, whether each speaker was singing (code 1) or speaking code (-1).* ***B.*** *Behavioral ratings (chance level – 0) for song (orange) and speech (blue) samples. Diamonds represent the ratings for each listener.* ***C.*** *Correlation between normalized difference scores (Song MPS vs. Speech MPS and Song vs. Speech behavioral ratings) represented in the MPS domain. (FDR-corrected in the spectral and temporal modulation domains, p < .05).* ***D.*** *Scatter plot of MPS normalized difference (Song minus Speech) against Behavioral normalized difference (Song minus Speech) for the statistical peaks reported in C. Circles represents each speakers/vocalization (n = 369).* ***E.*** *Fieldsites-level behavioral ratings (chance level – 0) for song (orange) and speech (blue) samples. Colored circles represent each of the 21 societies/cultures (sorted as a function of the SVM decoding accuracy of Figure 2 D. - with a jet colormap).* ***F.*** *Scatter plot of SVM*

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

decoding accuracy (Figure 2 D.) against behavioral normalized difference (Song vs. Speech). Colored circles represent each of the 21 societies/cultures (sorted as a function of the SVM decoding accuracy of Figure 2 D. - with a jet colormap)

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Discussion

Using vocalizations drawn from a diverse set of languages and societies [5] we found that speech and song systematically differ in their typical acoustical signatures: songs contain greater energy than spoken utterances at higher spectral and lower temporal modulation rates, whereas speech shows the reverse effect (Figure 2 A.). This pattern was sufficiently consistent that, despite measurable variation in the distributions of spectro-temporal modulations in the vocalizations of each society tested (supplementary Figures 3, 4, 5, 6), we still observed overlap within each category (song and speech) in the two specific acoustical ranges across nearly all of the societies (Figure 2 B.); conversely, there was essentially no overlap between the two categories.

That these spectro-temporal cues suffice to classify the two categories was shown by the outcome of a machine-learning classifier, which was trained exclusively on the spectro-temporal features, and correctly identified both classes of vocalization above chance for all of the 21 societies (Figure 2 D.), albeit with differing degrees of accuracy. To verify that this outcome was not merely driven by similarities in the speech or song samples across societies that may have been geographically or linguistically related, we trained the classifier using only data from one country/region or language family, and tested on the others; the outcomes were essentially the same (Figure 3.). Furthermore, the information used by the classifier (Figure 2 F.) corresponded well to the ranges of modulation power that characterize the two classes, as identified in the initial aggregate acoustical analysis.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Last, human listeners who were unfamiliar with most of the speech and song systems sampled here performed close to ceiling when asked to indicate which vocalization corresponded to which category (Fig 4B). Their ratings were directly related to the distribution of energy in the modulation power spectrum (Figure 4 C and D), and the ranking of behavioral accuracy across societies was similar to that of the classification algorithm (Figure 4 E., see also supplementary Figure 2), suggesting that both the classifier and the humans relied on the same cues.

The findings support the existence of universals in the acoustical manifestations of the two principal modes of auditory-vocal communication found in our species. Because the differences in specific ranges of spectro-temporal modulation for speech and song are widely shared across unrelated groups of people, we may conclude that they represent a fundamental property of how sounds are generated by the human vocal tract, depending on the nature of the communication. To transmit denotative information using speech, a high level of temporal modulation is used, but spectral modulation is less prominent; whereas to communicate musical content and affective states using song, a high level of spectral modulation is used, but at lower temporal modulation rates. The fact that people unfamiliar with the most of the linguistic or musical systems in question were nevertheless easily able to identify which vocalization was which, and that they used essentially the same spectro-temporal cues as the machine-learning classifier had determined to be optimal, supports the conclusion that such cues are widely shared and readily available even in the absence of any culturally specific knowledge.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

One explanation for the distinct spectro-temporal signatures of speech and song is that they result from differences in how human vocal musculature is used for speaking or singing. The high rate of temporal modulation in speech reflects the syllable rate (opening and closing of the mouth), which tends to be faster when speaking than when singing [26]. The longer syllable duration in singing may allow for production of more stable pitch values, leading to better encoding of tonal relationships important for music [38]. Conversely, the high spectral modulation rate associated with song may be related to the complex physiology of phonation typical of singing that generates more energy in the upper harmonics [10,39].

Most songs, including those used here, incorporate both spoken and melodic content simultaneously. Thus, both types of modulation are typically present together. But what distinguishes the two is their different acoustical signature, as determined by comparing them against one another (Figure 2 A. and supplementary Figures 3 to 6). This is not to say that all cultures necessarily carve out the spectro-temporal space in exactly the same way. Although there was almost complete overlap of at least part of the distribution of spectro-temporal modulation for both speech and song across societies (Figure 2 B.), and the centroids of each distribution were clustered in close proximity (Figure 2 C.), a glance at the individual modulation difference plots for each culture (supplementary Figures 3 to 6) shows that there are important differences across them, especially in the songs, which exploit wide ranges of spectral and temporal modulation, even if they are always fairly far from the range of modulations used for speech. Further study of how and why these cues are deployed in different musical traditions could help to identify and explain such cross-cultural differences. Indeed, the spectro-temporal

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

framework may prove particularly valuable for examining questions of cross-cultural variability in language and music, since it does not require the selection of any particular acoustic or musical features, which are notoriously vulnerable to culture-specific assumptions [14]; see also Supplementary Information in [7].

We note that song and speech being acoustically distinct does not imply that top-down factors have no influence on the perception of a vocalization as song or as speech. Indeed, the well-known "speech-to-song" illusion [40] demonstrates that speech may sometimes be perceived as song with repeated presentation, even if the acoustics are held constant. This phenomenon has been attributed both to particular acoustical features of sounds susceptible to the illusion, as well as to individual differences across listeners [41-43]. The spectro-temporal framework may provide a useful approach to investigate vocalizations that are intermediate between canonical speech and song, and which may share features of both, not only in the context of the speech-to-song illusion, but also more broadly to study artistic forms in which speech and song features are blended (e.g. rap), or in speech with more prominent song-like features (e.g. infant-directed speech).

The findings presented here fit well with previous empirical work examining the perceptual relevance of spectral vs temporal cues for speech and music. In a classic paper, Shannon and colleagues [27] distorted normal speech by removing spectral information and replacing it with amplitude-modulated noise passed through a limited number of filter banks centered at different frequencies, which preserves primarily the temporal cues. The results indicated very good speech perception with as few as 3-4 spectral channels, thereby showing that temporal modulation cues suffice to extract

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

relevant information from a speech. Other studies of speech have also suggested that temporal modulation is closely related to phonetic identity [36].

The temporal modulation rate of speech samples from many different languages shows a consistent peak in the temporal modulation distribution at about 4-6 Hz [11,33], but the equivalent rate for a variety of Western musical genres (classical, rock, jazz) is generally less than half the speed of speech [11]. These observations are close to ours, in which speech temporal modulation occupied a range of 5-8 Hz, while song temporal modulations were close to 1Hz. But whereas prior studies only examined Western music, we show that this slower rate is characteristic of many global musical systems. Furthermore, we also show that spectral modulation rates are higher for songs than for speech across many cultures, which, to our knowledge, had not previously been shown, and which indicates that the distinction between the two kinds of vocalization is not only based on temporal, but also on spectral information.

In a direct test of the importance of spectral and temporal cues for song and speech, we [28] found that perception of speech content remained largely intact with spectral degradation, but quickly deteriorated with temporal degradation, whereas melody perception was largely abolished with spectral degradation but was not much affected by temporal degradation. That study, however, used only English and French speech, and Western-style melodies. The current results replicate and extend those findings beyond Western linguistic and musical systems, to encompass a widely distributed set of cultures, including ones with little or no contact with Western societies.

The differences we observed in the present study for speech and song can be interpreted within the context of neuroscience findings that suggest partially dissociable

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

neural representations of the two types of signals. Recent functional MRI data [44] using a voxel decomposition approach suggest that speech and music have distinct cortical representations as cognitive domains, rather than on the basis of acoustical cues. Indeed, intracranial recordings suggests the existence a cortical region, located bilaterally within the anterior temporal lobes, that is specifically sensitive to song over all other sound categories [45]; interestingly, the same dataset also shows specific sensitivity to spectral and temporal modulation in different cortical regions.

A competing idea is that speech content vs song melody are processed in distinct auditory cortical regions as a function of hemispheric differences in sensitivity to spectral and temporal modulations [28]. Numerous studies have adduced evidence that the neuronal populations in left auditory cortex have higher temporal resolution but lower spectral resolution, whereas the right auditory cortex has the reverse specialization [31,46-48]. According to this view, speech and song are represented in distinct neural substrates not because of domain-specific aspects, but rather because of their tendency to utilize opposite ends of the spectro-temporal continuum. The data from the present study would be in line with this conclusion, insofar as the spectro-temporal acoustical signatures of speech and song are shown to be sufficient to distinguish the two categories across many different linguistic and musical systems, suggesting that they reflect a fundamental organizational specialization of the human brain to process the two acoustical dimensions.

Our findings point to a biological origin of speech and song, upon which cultural influences act to produce the rich, varied, and beautiful forms of language and music found throughout the world. This conclusion fits with two basic ideas about the design of

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

human auditory perception. First, it fits with the idea of efficient coding, according to which the nervous system optimizes its representation of the environment based on the most salient features necessary for success [49]. Thus, neural responses are well-matched to the statistical properties of the most important aspects of both the visual [50] and auditory worlds [51] of a given species. Second, it supports the conclusion that music and speech tend to have distinct functional roles [2,34,35] in human evolution. Humans talk and sing, thanks to the organization of a nervous system that allows us to generate and perceive those signals that occupy different portions of the spectro-temporal acoustical continuum.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Methods

Vocalization corpus

We used a corpus of 738 recordings of adult-directed song, and adult-directed speech (all audio is available at https://doi.org/10.5281/zenodo.5525161) from [5]. People (N= 369) living in 21 societies produced each of these vocalizations, respectively, with a median of 15 individuals per society (range 6-57). From those for whom information was available, 86% were female.

Recordings were collected by the investigators of [5] and/or staff at their field sites, all using the same data collection protocol. They translated instructions to the native language of the participants, following the standard research practices at each site. Fieldsites were selected partly by convenience (i.e., via recruiting principal investigators at fieldsites) and partly to maximize cultural, linguistic, and geographic diversity (see supplementary Table 1).

For speech recordings, participants spoke to the researcher about a topic of their choice (e.g., they described their daily routine). For song, participants sang a song that was not intended for infants (see [5], for details); they also stated what that song was intended for (e.g., “a celebration song”). Participants vocalized in the primary language of their fieldsite, with a few exceptions (e.g., when singing songs without words; or in locations that used multiple languages, such as Turku, which included both Finnish and Swedish speakers).

Participants were free to determine the content of their vocalizations. This was intentional: imposing a specific content category on their vocalizations would likely alter

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the acoustic features of their vocalizations, which are known to be influenced by experimental contexts [5].

All recordings were made with Zoom H2n digital audio recorders, using foam windscreens (where available). To ensure that participants were audible along with researchers, who stated information about the participant and environment before and after the vocalizations, recordings were made with a 360° dual x-y microphone pattern. This produced two uncompressed stereo audio files (WAV) per participant at 44.1 kHz; we only analyzed audio from the two-channel file on which the participant was loudest.

The investigator at each fieldsite provided standardized background data on the behavior and cultural practices of the society (e.g., whether there was access to mobile-phones/TV/radio, and how commonly people used ID speech or song in their daily lives). Most items were based on variables included in the D-PLACE cross-cultural corpus[5]. The 21 societies varied widely in their characteristics, from cities with millions of residents (Beijing) to small-scale hunter-gatherer groups of as few as 35 people (Hadza). All of the small-scale societies studied had limited access to TV, radio, and the internet, mitigating against the influence of exposure to the music of other societies. Four of the small-scale societies (Nyangatom, Toposa, Sápara/Achuar, and Mbendjele) were completely without access to these communication technologies.

Our strategy was to analyze the 5 first seconds of the raw recording of speech and song vocalization produced by the same individual; this ensured that the findings were not unbalanced, e.g., because some recordings were much longer than others.

Extraction of spectro-temporal modulations

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

For the 738 selected samples (369 speech and 369 song) we decomposed the first five seconds of the acoustical signal using the framework of spectrotemporal modulation power [36]. The modulation domain results from the 2D fast Fourier transform of the autocorrelation matrix of the sound stimulus in its spectrographic representation and represents the energy modulation across the temporal and spectral axes (Figure 1). This results in 738 MPS data that were then used for univariate and multivariate analyses.

Univariate analyses

Fieldtrip [37] functions were used to perform non-parametric permutation statistics with FDR correction ($p < .001$) for the contrast between song and speech MPS.

Multivariate analyses

Multivariate analyses were performed using MATLAB and LibSVM's linear support vector machine (SVM) implementation (www.csie.ntu.edu.tw/~cjlin/libsvm/). A linear classifier was chosen as MPS data contains many more features than examples, and classification of such data is generally susceptible to over-fitting. One way of alleviating the danger of over-fitting is to choose a simple function (such as a linear function) for classification, where each feature affects the prediction solely via its weight and without interaction with other features (rather than more complex classifiers, such as nonlinear SVMs or artificial neural networks, which can let interactions between features and nonlinear functions thereof drive the prediction).

Our strategy was to use the Support Vector Machine (SVM) classifier with fieldsite-wise k-fold cross-validation to classify song and speech vocalization samples,

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

using the MPS as features. The model is trained only on data from 20 of the 21 societies to predict whether each vocalization in the 21st society is song or speech. The procedure is repeated 21 further times, with each society being held out, to estimate the classification performance across the full set of societies. Results were expressed as accuracy of category identification that was calculated using an average of the cross-validation folds. For each classifier, we extracted the features weights (zscore) to evaluate the relative contribution of each feature in the classification. This procedure was performed across societies (21), across countries (18), language families (16), world subregions (15) and regions (6).

Behavioral experiment

Participants: 80 adults participated in the behavioral experiment. The group was composed of 80 native French speakers from France and Canada (33 female, 4 non-binary, mean age = 32.4 years ± 10.86). Some of them (10 out of 80) were musically trained (more than 5 years of formal musical training). Participants reported no history of neurological or psychiatric disease. Ethical approval was obtained from the Ethics Review Board of the CIUSSS de la Capitale Nationale (2022-2476).

Procedure: The experiments took place in a sound-attenuated booth. Auditory stimuli were presented binaurally via Sennheiser HD 280 pro headphones at a comfortable sound level (~75 dB SPL). PsychoPy [52] was used to control the stimulus presentation and record responses.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We played the song and speech recordings to these individuals who were asked to rate, as rapidly as possible on a 5-point scale on their keyboard, whether each speaker was singing (code 1) or speaking (code - 1). Participants had 9 seconds to respond and received no feedback (i.e., we did not tell them whether or not their rating was accurate).

The experiment lasted approximately 15 minutes. We used 3 different blocks that were pseudo-randomly presented to the participant. Each bloc contained the same number of examples of speech and song for each society, with a total of 246 trials per block. This way a given listener was also rating the vocalization of the same speaker. Example of this task can be found online: : https://run.pavlovia.org/palbouy/spectrotemp_bloc1

Behavioral data analysis: Data were processed with MATLAB (The Mathworks), and statistical analyses were performed with Jamovi (https://www.jamovi.org). For each participant, the ratings were extracted and sorted as a function of society, language family, countries, world subregions and world regions. Ratings were analyzed with one sample t-tests and we performed Pearson's correlation corrected with FDR when necessary.

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1 Zatorre, R. J. & Baum, S. R. Musical melody and speech intonation: singing a different tune. *PLoS Biol* **10**, e1001372, doi:10.1371/journal.pbio.1001372 (2012).

2 Mehr, S. A., Krasnow, M. M., Bryant, G. A. & Hagen, E. H. Origins of music in credible signaling. *Behavioral and Brain Sciences* **44**, e60, doi:10.1017/S0140525X20000345 (2021).

3 Eibl-Eibesfeldt, I. Human ethology: concepts and implications for the sciences of man. *Behavioral and Brain Sciences* **2**, 1-26, doi:10.1017/S0140525X00060416 (1979).

4 List, G. The Boundaries of Speech and Song. *Ethnomusicology* **7**, 1-16, doi:10.2307/924141 (1963).

5 Hilton, C. B. *et al.* Acoustic regularities in infant-directed speech and song across cultures. *Nature Human Behaviour* **6**, 1545-1556 (2022).

6 Savage, P. E., Brown, S., Sakai, E. & Currie, T. E. Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences* **112**, 8987, doi:10.1073/pnas.1414495112 (2015).

7 Mehr, S. A. *et al.* Universality and diversity in human song. *Science* **366**, eaax0868, doi:10.1126/science.aax0868 (2019).

8 Singh, M. & Mehr, S. A. Universality, domain-specificity, and development of psychological responses to music.
. *Nature Reviews Psychology* (in press).

9 Jacoby, N. & McDermott, J. H. Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. *Current Biology* **27**, 359-370 (2017).

10 Sundberg, J. *The Science of the Singing Voice*. (Northern Illinois University Press, 1989).

11 Ding, N. *et al.* Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews* **81**, 181-187, doi:https://doi.org/10.1016/j.neubiorev.2017.02.011 (2017).

12 Wood, A. L. C. *et al.* The Global Jukebox: A public database of performing arts and culture. *PLOS ONE* **17**, e0275469, doi:10.1371/journal.pone.0275469 (2022).

13 Nettl, B. *The study of ethnomusicology: Twenty-nine issues and concepts*. (University of Illinois Press, 1983).

14 Jacoby, N. *et al.* Cross-Cultural Work in Music Cognition: Challenges, Insights, and Recommendations. *Music Perception* **37**, 185-195, doi:10.1525/mp.2020.37.3.185 (2020).

15 Jackendoff, R. Parallels and Nonparallels between Language and Music. *Music Perception* **26**, 195-204, doi:10.1525/mp.2009.26.3.195 (2009).

16 Patel, A. D. *Music, language, and the brain*. (Oxford university press, 2010).

17 Fitch, W. T. On the biology and evolution of music. *Music Perception* **24**, 85-88 (2006).

18 Krumhansl, C. L. *Cognitive foundations of musical pitch*. Vol. 17 (Oxford University Press, 2001).

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

19 Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D. & Majid, A. Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences* **26**, 1153-1170, doi:https://doi.org/10.1016/j.tics.2022.09.015 (2022).

20 Elhilali, M. in *Timbre: Acoustics, Perception, and Cognition* (eds Kai Siedenburg *et al.*) 335-359 (Springer International Publishing, 2019).

21 Shamma, S. On the role of space and time in auditory processing. *Trends in Cognitive Sciences* **5**, 340-348, doi:https://doi.org/10.1016/S1364-6613(00)01704-6 (2001).

22 Singh, N. C. & Theunissen, F. E. Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America* **114**, 3394-3411 (2003).

23 Woolley, S. M. N., Fremouw, T. E., Hsu, A. & Theunissen, F. E. Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nature Neuroscience* **8**, 1371-1379, doi:10.1038/nn1536 (2005).

24 Gehr, D. D., Komiya, H. & Eggermont, J. J. Neuronal responses in cat primary auditory cortex to natural and altered species-specific calls. *Hearing research* **150**, 27-42 (2000).

25 Wang, X., Merzenich, M. M., Beitel, R. & Schreiner, C. E. Representation of a species-specific vocalization in the primary auditory cortex of the common marmoset: temporal and spectral characteristics. *Journal of neurophysiology* **74**, 2685-2706 (1995).

26 Poeppel, D. & Assaneo, M. F. Speech rhythms and their neural foundations. *Nature Reviews Neuroscience* **21**, 322-334, doi:10.1038/s41583-020-0304-4 (2020).

27 Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J. & Ekelid, M. Speech recognition with primarily temporal cues. *Science* **270**, 303-304 (1995).

28 Albouy, P., Benjamin, L., Morillon, B. & Zatorre, R. J. Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science* **367**, 1043, doi:10.1126/science.aaz3468 (2020).

29 Santoro, R. *et al.* Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS Comput Biol* **10**, e1003412 (2014).

30 Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E. & Chang, E. F. Human Superior Temporal Gyrus Organization of Spectrotemporal Modulation Tuning Derived from Speech Stimuli. *The Journal of Neuroscience* **36**, 2014, doi:10.1523/JNEUROSCI.1779-15.2016 (2016).

31 Flinker, A., Doyle, W. K., Mehta, A. D., Devinsky, O. & Poeppel, D. Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries. *Nature human behaviour* **3**, 393-405 (2019).

32 Zatorre, R. J., Belin, P. & Penhune, V. B. Structure and function of auditory cortex: music and speech. *Trends Cogn Sci* **6**, 37-46, doi:10.1016/s1364-6613(00)01816-7 (2002).

33 Varnet, L., Ortiz-Barajas, M. C., Erra, R. G., Gervain, J. & Lorenzi, C. A cross-linguistic study of speech modulation spectra. *The Journal of the Acoustical Society of America* **142**, 1976-1989, doi:10.1121/1.5006179 (2017).

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

34 Pinker, S. & Bloom, P. Natural language and natural selection. *Behavioral and brain sciences* **13**, 707-727 (1990).

35 Fitch, W. T. *The evolution of language*. (Cambridge University Press, 2010).

36 Elliott, T. M. & Theunissen, F. E. The Modulation Transfer Function for Speech Intelligibility. *PLOS Computational Biology* **5**, e1000302, doi:10.1371/journal.pcbi.1000302 (2009).

37 Oostenveld, R., Fries, P., Maris, E. & Schoffelen, J. M. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput Intell Neurosci* **2011**, 156869, doi:10.1155/2011/156869 (2011).

38 Mantell, J. T. & Pfordresher, P. Q. Vocal imitation of song and speech. *Cognition* **127**, 177-202, doi:https://doi.org/10.1016/j.cognition.2012.12.008 (2013).

39 Kob, M. *et al.* Analysing and Understanding the Singing Voice: Recent Progress and Open Questions. *Current Bioinformatics* **6**, 362-374, doi:10.2174/157489311796904709 (2011).

40 Deutsch, D., Henthorn, T. & Lapidis, R. Illusory transformation from speech to song. *The Journal of the Acoustical Society of America* **129**, 2245-2252 (2011).

41 Tierney, A., Patel, A. D. & Breen, M. Acoustic foundations of the speech-to-song illusion. *Journal of Experimental Psychology: General* **147**, 888-904, doi:10.1037/xge0000455 (2018).

42 Rathcke, T., Falk, S. & Dalla Bella, S. Music to Your Ears: Sentence Sonority and Listener Background Modulate the "Speech-to-Song Illusion". *Music Perception* **38**, 499-508, doi:10.1525/mp.2021.38.5.499 (2021).

43 Jaisin, K., Suphanchaimat, R., Figueroa Candia, M. A. & Warren, J. D. The Speech-to-Song Illusion Is Reduced in Speakers of Tonal (vs. Non-Tonal) Languages. *Frontiers in Psychology* **7**, doi:10.3389/fpsyg.2016.00662 (2016).

44 Norman-Haignere, S., Kanwisher, Nancy G. & McDermott, Josh H. Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. *Neuron* **88**, 1281-1296, doi:https://doi.org/10.1016/j.neuron.2015.11.035 (2015).

45 Norman-Haignere, S. V. *et al.* A neural population selective for song in human auditory cortex. *Current Biology*, doi:https://doi.org/10.1016/j.cub.2022.01.069 (2022).

46 Zatorre, R. J. & Belin, P. Spectral and temporal processing in human auditory cortex. *Cereb Cortex* **11**, 946-953, doi:10.1093/cercor/11.10.946 (2001).

47 Jamison, H. L., Watkins, K. E., Bishop, D. V. & Matthews, P. M. Hemispheric specialization for processing auditory nonspeech stimuli. *Cerebral cortex* **16**, 1266-1275 (2006).

48 Schönwiesner, M., Rübsamen, R. & Von Cramon, D. Y. Hemispheric asymmetry for spectral and temporal processing in the human antero□lateral auditory belt cortex. *European Journal of Neuroscience* **22**, 1521-1528 (2005).

49 Gervain, J. & Geffen, M. N. Efficient Neural Coding in Auditory and Speech Perception. *Trends in Neurosciences* **42**, 56-65, doi:https://doi.org/10.1016/j.tins.2018.09.004 (2019).

50 Simoncelli, E. P. & Olshausen, B. A. Natural image statistics and neural representation. *Annual review of neuroscience* **24**, 1193-1216 (2001).

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.29.526133; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

51 Smith, E. C. & Lewicki, M. S. Efficient auditory coding. *Nature* **439**, 978-982 (2006).
52 Peirce, J. *et al.* PsychoPy2: Experiments in behavior made easy. *Behav Res Methods* **51**, 195-203, doi:10.3758/s13428-018-01193-y (2019).

Acknowledgments and Funding: This work was supported a CIHR Foundation grant to R.J.Z., and NSERC Discovery grants to P.A. and R.J.Z. R.J.Z. is a fellow of the Canadian Institute for Advanced Research and is funded via the Canada Research Chair program, and by the Fondation pour l'Audirion (Paris). P.A. is supported by FRQS Junior 1 grant.

Authors' contributions: Conceptualization, R.J.Z., P.A., S.A.M.; Methodology, P.A., S.A.M., R.J.Z.; Analysis, P.A.; Investigation, P.A., S.A.M., R.S.H., J.G.; Resources, P.A., R.J.Z.; Writing – Original Draft, R.J.Z., P.A.; Writing – Review & Editing: P.A., S.A.M., R.S.H, J.G., R.J.Z; Visualization, P.A.; Supervision and Project Administration, R.J.Z., P.A., S.A.M.

Conflict of interest: none declared.

Data and materials availability: Behavioral data and MATLAB code are freely available at the following URL: https://osf.io/XXXX (address delivered upon publication). Example of the judgment task can be found here: :

https://run.pavlovia.org/palbouy/spectrotemp_bloc1

Raw vocalizations are freely available at https://zenodo.org/record/5525161