AUTOMATIC COMPARISON OF HUMAN MUSIC, SPEECH, AND BIRD SONG SUGGESTS UNIQUENESS OF HUMAN SCALES

Jiei Kuroyanagi*[1], Shoichiro Sato*[1], Meng-Jou Ho[1], Gakuto Chiba[1], Joren Six[2], Peter Pfordresher[3], Adam Tierney[4], Shinya Fujii[1], Patrick E. Savage[1]**

[1]Keio University, Japan, [2]Ghent University, Belgium, [3]University at Buffalo, NY, USA, [4]Birbeck, University of London, UK

*Equal contribution, **Correspondence to: psavage@sfc.keio.ac.jp

ABSTRACT

The uniqueness of human music relative to speech and animal song has been extensively debated, but rarely directly measured. We applied an automated scale analysis algorithm to a sample of 86 recordings of human music, human speech, and bird songs from around the world. We found that human music throughout the world uniquely emphasized scales with small-integer frequency ratios, particularly a perfect 5th (3:2 ratio), while human speech and bird song showed no clear evidence of consistent scale-like tunings. We speculate that the uniquely human tendency toward scales with small-integer ratios may relate to the evolution of synchronized group performance among humans.

1.BACKGROUND

The origins of music and language have been debated for centuries. Both music and language are human universals found in all known societies, but language appears to be unique to humans while music has many parallels in non-human species such as songbirds and whales [6]. Why might this be, and what - if anything - is unique about human music?

Comparative analyses of music have produced conflicting results. Some studies argue that certain aspects of music like simple scales and rhythms are unique to human music and may have evolved to bond people together [19, 20]. Others argue against such ideas on the grounds that such features are not cross-culturally universal [5, 15] or that they are not specific to human music, but instead a byproduct of more general constraints on acoustic perception and production that are shared with speech and/or animal song [23].

The degree to which music and language share similar features has seen vigorous debate in the recent literature [9, 17]. An improved understanding of such boundaries requires empirical research to better understand their similarities and differences. However, while many have compared speech vs. song, human song vs. bird song, and human language vs. bird song [9, 10, 17, 21, 23], we know of only one previous empirical study that has simultaneously compared musical aspects of human music, human speech, and non-human vocalizations [24]. We decided to use automated scale analysis software to analyse and compared global samples of human music, human speech, and bird song. Because definitions of music, language, and animal song are controversial, we did not define and collect samples ourselves but used pre-existing databases (see Methods for details).

Previously, we used automated scale analysis software to demonstrate a strong cross-cultural tendency for human music to use scales containing pitches separated by intervals that approximate simple integers, particularly a perfect 5th (e.g., 700 cents, ~ 3:2 frequency ratio) [8]. If perfect fifths also predominate in bird song or human speech as well as human music, then they are likely a consequence of perceptual/motor constraints, whereas if they are specific to human music, this suggests that they could be an adaptation specifically for human music [23].

2.METHOD

2.1. Audio Samples

For this preliminary analysis, we aimed to assemble globally distributed samples of approximately 30 recordings each of 1) human music, 2) human speech, and 3) bird song. We were only able to identify 26 recordings of human speech, giving a total sample of 86 recordings (Fig. 1, Table 1). For all samples, only monophonic recordings were used to enable accurate automatic transcription.

(1) 30 music recordings from nine regions were obtained from the *Garland Encyclopedia of World Music* [16]. The audio files were recorded in diverse regions, covering a diverse mix of traditional genres (e.g., healing, love, religious). From the 124 monophonic Garland recordings assessed as usable, we randomly selected 3-4 recordings for this preliminary analysis from each of the following 9 regions designated by *Garland's* editors: Africa, South America, North America, Southeast Asia, South Asia, Middle East, East Asia, Europe, and Oceania. Our sample included both instrumental and vocal music.

(2) We obtained 26 recordings of human speech from the Linguistic Data Consortium [12] spoken language sampler. Because we were unable to locate 30 or more samples, we used all available samples without sampling equally by region as we did for the human music samples. The samples mainly consist of recorded telephone conversations. Each sample was edited to

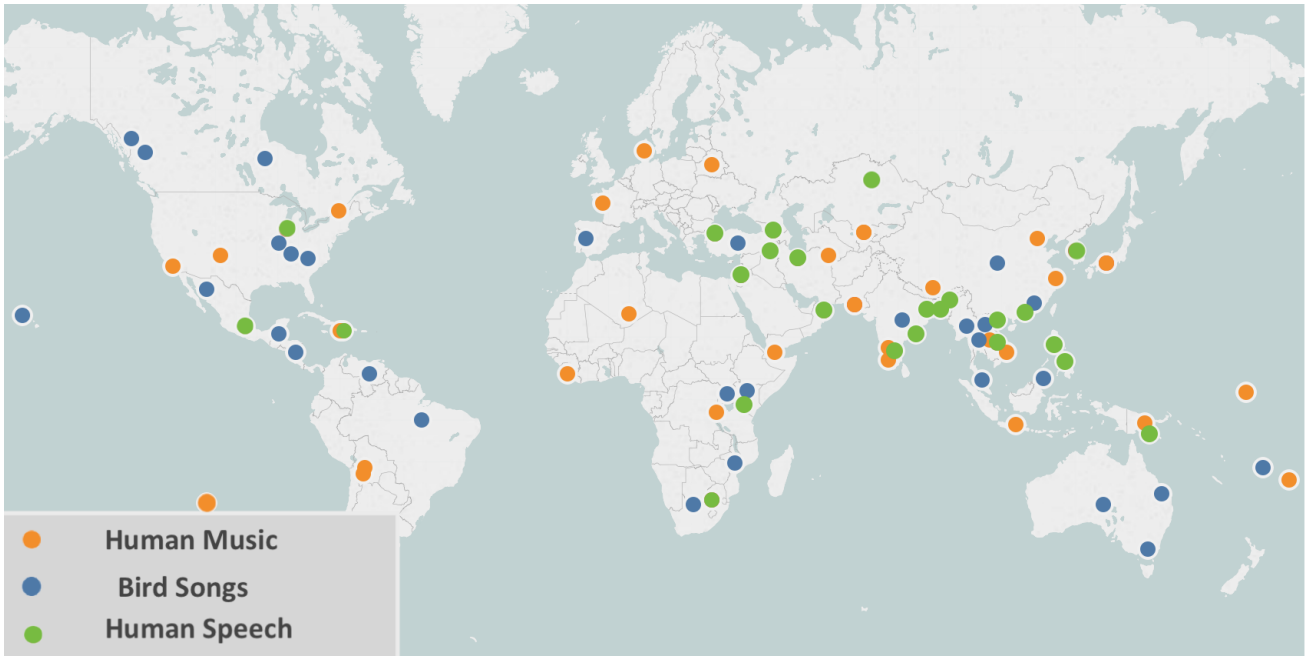

Figure 1. A map of the 86 recordings analysed. Human music (n=30), bird songs (n=30) and human speech (n=26) are respectively coloured orange, blue and green.

Table 1. Metadata about musical genre, bird species, or language name for the 86 recordings analysed. See original publications for additional metadata [8, 11, 18].

Human Music	Bird Species	Human Speech
Inanga Chuchotee (Whispered Inanga)	*Cinclus mexicanus*	American English
Tuareg Tihadanaren	*Icterus galbula*	Georgian
Somali caayar "dhaanto," excerpt 1	*Poecile atricapillus*	Haitian
"Kulwa" (Kupla, cobla)	*Monarcha melanopsis*	Cebuano
Lichiwayu notch flutes	*Tchagra senegala*	Kazakh
Rara instrumental music	*Nectarinia kilimensis*	Levantine Arabic
"La Finada Pablita"	*Prunella fulvescens*	Malto
"Sabá Medley"	*Nilaus afer*	Mexican Spanish
"El Pájaro Verde" ("The Green Bird")	*Catherpes mexicanus*	Pashto
Chinese-Thai sizhu ensemble piece "Chung we meng" ("Moon Shining Brightly in the Spring")	*Bombycilla cedrorum*	Persian
Jarai gong ensemble with song "Yong Thoach" ("Brother Thoach, Please Come Back")	*Dendroica pensylvanica*	Tagalog
East Javanese "Srempeg, pelog patet wolu"	*Psophodes occidentalis*	Gulf Arabic
Rgvedic recitation by Nambudiri Brahmins	*Acridotheres tristis*	South Korean
South Indian devotional kriti, "Girirājasuta" 'Son of the mountain king's daughter, In rāga bangāla, desdi tāla	*Aegithina tiphia*	Japanese
Benjo 'keyed zither' perfoemance in Balochistan	*Galerida cristata*	Swahili
Persian narrative song, Sayyed Mohammad Khan	*Aethopyga siparaja*	Lao
Uzbek classical song, Sarabaxi Ōrōm-I Jōn 'Peace of the Soul' in maqām dugāh (or dugōh), saraxbōr (4/4) rythm	*Sturnella magna*	Bengali
Uzbek classical instrumental dance piece, Oynasin Dugah	*Spizella pusilla*	Assamese
Jewish-Yemenite liturgical、The song of the Sea	*Dicaeum ignipectus*	Cantonese
"Moonlight On The Ching Yang (Xunyang) River"	*Gymnomyza viridis*	Tamil
"The Revenge" (Lyrical area from Qi yaun bao)	*Regulus satrapa*	Telgu
"Song Foe Reparing Water Channels in Barley Fields"	*Pachycephala pectoralis*	Tok Pisin
Nozakimura	*Eminia lepida*	Turkish
Denmark: sailor's lovesong "Sode pige, du er sa laught fra mig" (Dear girl, I am so far away')	*Vireolanius pulchellus*	Vietnamese
France: male solo song, "Mon pére a fait faire un étang"	*Garrulax canorus*	Zulu
Belarus: polyphonic harvest song	*Oriolus kundoo*	Kurmanji Kurdish
Rapa Nui string-figure song (pāta' uta' u)	*Moho braccatus Mimus*	
Tongan formal dance-song (lakalaka)	*Chloropsis cyanopogon*	
"Anawa anawa" Kiribati women's hip-shaking dance (kabuti)	*icrurus remifer*	
Usarufa blood-song(naa-ímá)	*Ramphocaenus melanurus*	

approximately 5 to 10 seconds in order to capture only a single speaker, and noises that could affect the result of the analysis were removed (using Logic's Noise Gate function). The edited recordings were then slowed down by a factor of five to avoid under-sampling the rapidly changing pitch (in the future we aim to modify our software to allow for increasing the sampling rate to resolve this issue without having to manually slow down recordings).

(3) We obtained 30 bird song recordings by selecting a subset of recordings without considerable noise from 80 previously analysed recordings of taxonomically diverse songbirds [23 One of the limitations of our

automatic analysis software is that polyphonic melodies or audio with considerable noise cannot be properly extracted. Moreover, exceedingly high pitches or short sounds are unanalysable. Thus, human speech and bird songs had to be slowed down (by 5x) and the bird songs had to be transposed two octaves lower and have noise removed. In the future we aim to modify our software to allow for increasing the sampling rate, noise filtering, and transposing to resolve this issue without having to manually edit recordings.

2.2. Pitch Class Histogram

We used Tarsos [22] to extract and compare musical scales, since it was designed for automatic quantitative analysis of any music from around the world. Thus it allows us to analyse audio data in a way that allows comparison of frequency ratio relationships across cultures and even species. Most of the analysis is executed using pitch histograms and octave reduced pitch class histograms. Tarsos first extracts the pitch histogram, then combines this pitch histogram across octaves to a pitch class histogram which is expressed in cents [5] ranging from 0-1200. We used Tarsos's default YIN pitch estimation algorithm [3]. In the future, we plan to explore the effects of using pitch histograms without assuming octave equivalence, and of using newer algorithms such as pYIN [14] and CREPE [11]. However, we note that such algorithms contain additional assumptions that may not be appropriate for cross-cultural/cross-species analyses.

2.3. Normalisation and comparison of averaged pitch class histograms

Normalisation is required to compare scales between different songs that have different keys or tuning systems. If we were confident that we could identify a tonal center for different recordings of human music, human speech, and bird song by, for example, selecting the final pitch of a recording, this might be a useful method of normalizing. However, because the idea of final notes as tonal centre is not necessarily applicable across cultures or species, we chose to normalise all recordings by setting the most frequent pitch class to 0. In a separate study, we have validated this approach by directly comparing results of normalizing human music using the final pitch vs. the most frequent pitch, finding that there is almost no difference between the two methods, leading us to use most frequent as it can be calculated more objectively (e.g., in cases where music fades out before the end) [4]. Figure 2 shows example analyses.

In addition, the raw count of pitch annotations is converted to a percentage so that longer recordings will not be weighted more than shorter ones. After normalizing, pitch class histograms were averaged across recordings separately for human music, human speech, and bird songs to determine whether there were any tuning intervals that were consistent across each sample.

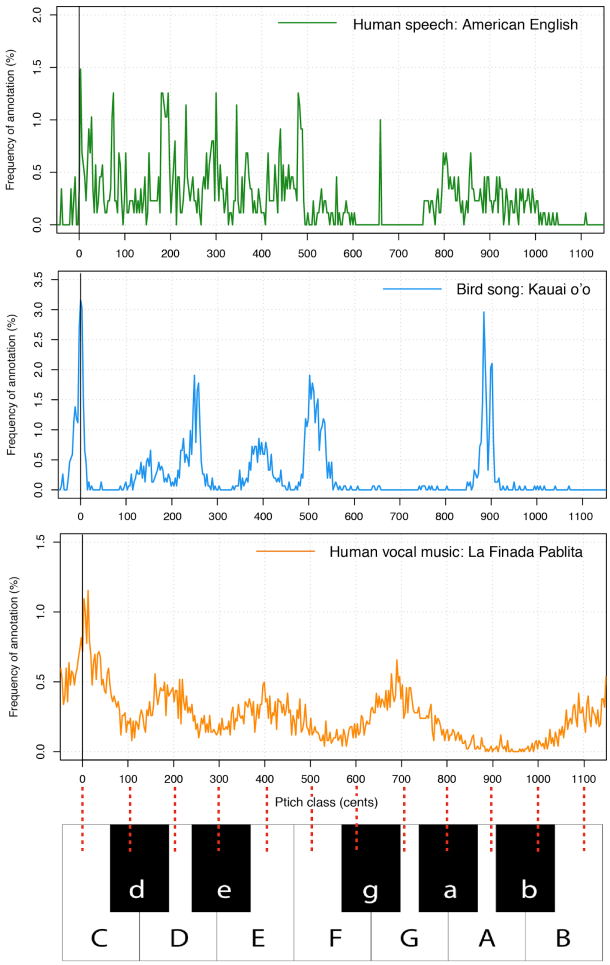

Figure 2. Bottom: A pitch class histogram of "La finada Pablita", a Mexican-American narrative song, aligned against a keyboard octave for comparison. The vertical axis represents how often a given pitch class occurs in the audio. The horizontal axis is plotted in cents over an octave range, from 0-1200, where the most common pitch class (~1.2% of annotations) is set to 0. In this figure the second most frequent scale degree appears a perfect fifth (~700 cents, equivalent to G in key of C) above the most frequent note. **Middle:** Pitch class histogram analysis of a bird song (Extinct bird in Kauai called "Kauaʻi ʻōʻō). **Top:** Pitch class histogram analysis of human speech (North American English).

3. RESULTS

Figure 3 shows the average pitch class histograms for human music, human speech, and bird song, plotted on the same axis for comparison. By definition, all samples show a peak at 0 cents, because 0 was defined as the most frequent note for each recording. Human speech and bird songs show no other clear peaks. Bird song does shows a stronger peak at the most frequent pitch

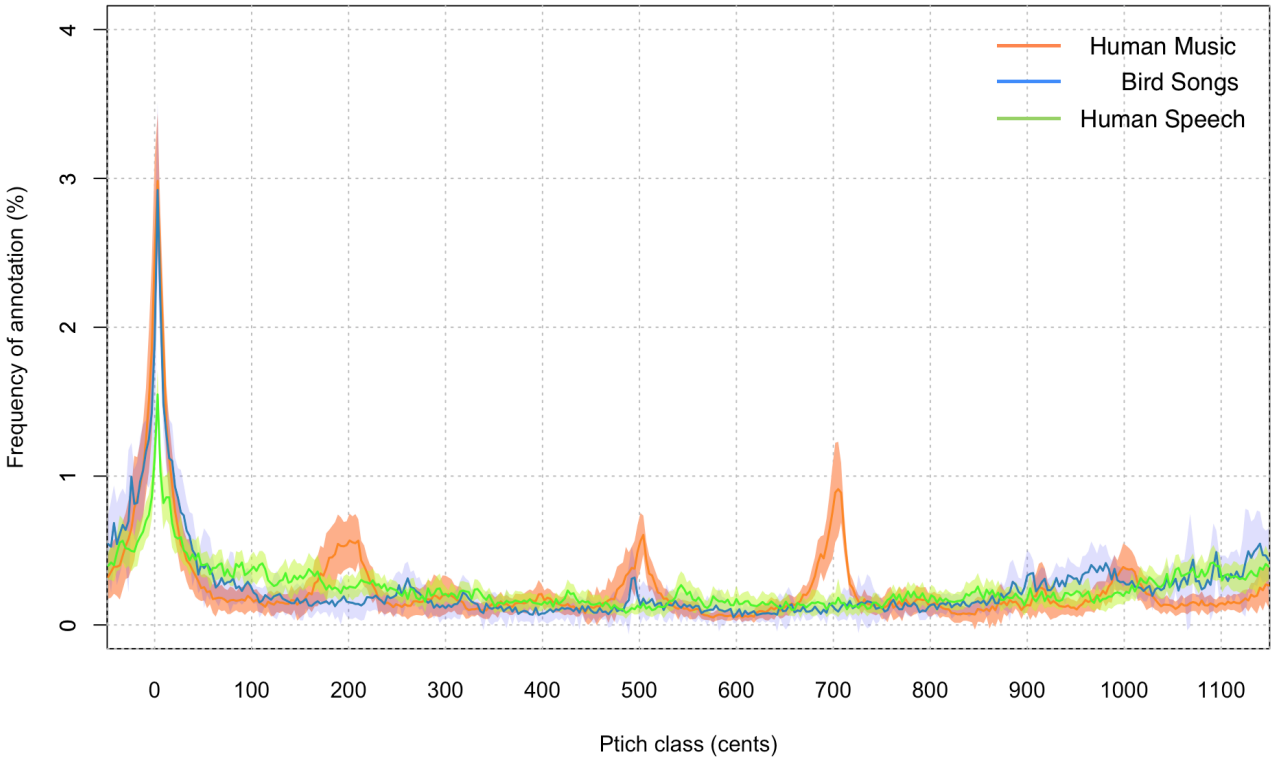

Figure 3. Averaged pitch class histograms for human music (n=30), bird song (n=30), and human speech (n=26). Shading indicates 95% confidence intervals.

(0 cents) and possible peaks at approximately 250, 500 and 1,000 cents, while human speech shows a possible peak at approximately 550 cents. These are small enough that it is not clear if they are true peaks or artefacts of the small sample size. However, human music shows much stronger peaks at intervals of approximately 700, 500, and 200 cents. These correspond approximately to small-integer ratios of 3:2 (perfect 5th), 4:3 (perfect 4th), and major 2nd (9:8), respectively.

4. DISCUSSION

Our results show that scale tunings in human music uniquely tend to emphasize intervals with small-integer ratios - particularly the perfect 5th (~ a 3:2 ratio) - while no consistent intervals emerge when the same analyses are applied to human speech or bird song (cf. Fig. 3).

Our previous analyses breaking up human music into sub-samples based on region and instrumentation [4, 7] showed that of these ratios, only the perfect 5th - the smallest possible integer-ratio within the octave - consistently predominated across all sub-samples. Taken together, these studies suggest that the perfect 5th uniquely predominates throughout the world's music but not in speech or bird song.

Many scholars have proposed that there is something special about small-integer ratios in human music, with most explanations centering around the psychoacoustics of harmonic overtone structure. Whenever an object resonates to produce a fundamental pitch, it also can produce a series of "overtones" which appear at integer ratios above the fundamental pitch due to the physics of how objects vibrate. While birds and many other animals tend to produce pure tone vocalizations without complex harmonic structure, human vocalizations tend to have a rich harmonic structure emphasizing many overtones, and this has been proposed to explain preferences for small-integer ratios in music via statistical learning through exposure to the speech of other humans that contains such harmonic structure [1, 7]. However, this "vocal similarity" hypothesis does not explain why we find small-integer ratios in human music but not human speech.

We speculate that human music may be unique because it evolved to be performed in synchronized groups, possibly to bond group members [19, 20]. This may have selected for integer-ratio frequencies because the resulting harmonies are more likely to perceptually fuse and sound like one large auditory event [2, 20]. Neither speech nor bird song are regularly performed in synchronized groups, and thus harmonicity does not result in any perceptual fusion.

While our preliminary data suggest that human scales are unique to music, there are also intriguing suggestions of similarities between bird song and human music as distinct from human speech. In particular, both human music and bird song show a stronger tendency to prefer a single most frequent note that remains relatively stable

throughout a performance, and bird songs may suggest a possible peak at a perfect 4th (4:3 ratio, approximately 500 cents). We hope to investigate this further with larger samples and through perceptual experiments in humans and birds.

5. FUTURE WORK

The main challenge for future work is to expand the sample to include several hundred recordings and perform statistical testing to formally evaluate the degree to which the pitch class histograms depart from distributions that would be expected by chance. We also intend to perform sub-sample analyses to explore the degree to which any average trends are consistent across geographic regions and instrumentation (particularly vocal vs. instrumental music). Expanding the sample of human music to the full set of 124 monophonic recordings has confirmed that intervals of approximately 200, 500, and 700 cents tend to predominate throughout the world, and that 700 cents is particularly common across almost all world regions [4]. We aim to similarly expand our samples of human speech and animal song, as the current results must be considered preliminary due to their small sample sizes and somewhat uneven geographic distribution.

Our analysis methods are well-suited for analyzing scales with a stable tonal center throughout a recording, but in the future we would like to explore ways of analyzing pitch relationships at interval-by-interval and phrase-by-phrase levels, which will particularly help in cases such as unaccompanied singing in which tonal centers can drift over time [18]. We also plan to implement additional features in Tarsos to allow us to automate the analysis to make it applicable to larger samples and to make the pre-analysis process more objective.

While the universality or cultural specificity of scale intonation has been debated for centuries, there remains little cross-species and cross-domain data to identify musical features that may represent unique adaptations for human music. Our analysis, while limited in scale, provides suggestive evidence that scales with small-integer ratios may represent a candidate for such an adaptation, and allow us to speculate why they may have evolved. Future studies with larger samples should help to clarify the robustness of our present findings.

6. AUTHOR CONTRIBUTIONS

P.E.S., S.F., A.T., J.S., and P.Q.P. designed the study; J.K., S.S., M.-J.H., and G.C. analysed the data, supervised by P.E.S.; J.K. and P.E.S. drafted the manuscript.

7. ACKNOWLEDGMENTS

This work was supported by a Grant-in-Aid for Young Scientists from the Japan Society for the Promotion of Science, Keio Research Institute at SFC Startup Grant, and a Keio Gijuku Academic Development Fund grant to P.E.S.

8. REFERENCES

[1]. D. L. Bowling and D. Purves, “A biological rationale for musical consonance,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 36, pp. 11155–11160, 2015.

[2]. S. Brown, “Contagious heterophony: A new theory about the origins of music,” *Music. Sci.*, vol. 11, no. 1, pp. 3–26, 2007.

[3]. A. de Cheveigne and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917– 1930, 2002.

[4]. G. Chiba, M.-J. Ho, S. Sato, J. Kuroyanagi, J. Six, P. Pfordresher, A. Tierney, S. Fujii, and P. E. Savage, “Small-integer ratio scales predominate throughout the world’s music.” 2019. *PsyArXiv.* Preprint doi: 10.31234/osf.io/5bghm

[5]. A. J. Ellis, “On the Musical Scales of Various Nations,” *Journal of the Society of Arts,* vol.23, no.1688, pp. 435-527, 1885.

[6.] W. T. Fitch, “The biology and evolution of music: A comparative perspective,” *Cognition*, vol. 100, no. 1, pp. 173–215, 2006.

[7]. K. Z. Gill and D. Purves, “A biological rationale for musical scales,” *PLOS ONE*, vol. 4, no. 12, p. e8144, 2009.

[8]. M. Ho, S. Sato, J. Kuroyanagi, J. Six, S. Brown. S Fujii, P. E. Savage. “Automatic analysis of global music recordings suggests scale tuning universals,” in *Extended abstracts for the Late-Breaking Demo Session of the 19th International Society for Music Information Retrieval Conference (ISMIR 2018)*, 2018.

[9]. H. Honing (Ed.). *The origins of musicality*. Cambridge, MA: MIT Press, 2018.

[10]. E. D. Jarvis, “Learned birdsong and the neurobiology of human language,” *Annals of the New York Academy of Science*, vol. 1016, no. 1, pp. 749-777, 2004.

[11]. Kim, J. W., Salamon, J., Li, P., & Bello, J. P. “CREPE: A convolutional representation for pitch estimation”. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 161-165), 2018.

[12]. M. Lieberman, Ed., “*Linguistic Data Consortium*,” 2019. [Online]. Available: https://www.ldc.upenn.edu. [Accessed: 20-Mar-2019].

[13]. A. Lomax (Ed.). *Folk song style and culture.* Washington, DC: American Association for the Advancement of Science, 1968.

[14]. M. Mauch and S. Dixon, “pYIN: A fundamental frequency estimator using probabilistic threshold distributions,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 659–663, 2014.

[15]. J. H. McDermott, A. F. Schultz, E. A. Undurraga, and R. A. Godoy, “Indifference to Dissonance in Native Amazonians Reveals Cultural Variation in Music Perception,” *Nature*, vol. 535, pp. 547–550, 2016.

[16]. B. Nettl, R. Stone, J. Porter, and T. Rice, Eds., *The Garland encyclopedia of world music* [10 volumes; 9 CDs]. New York: Garland Pub., 1998-2002.

[17]. A. D. Patel, *Music, language and the brain*. Oxford: Oxford University Press, 2008.

[18]. P. Q. Pfordresher and S. Brown, “Vocal mistuning reveals the origin of musical scales,” *J. Cogn. Psychol.*, vol. 29, no. 1, pp. 35–52, 2017.

[19]. P. E. Savage, S. Brown, E. Sakai, & T. E. Currie, ”Statistical universals reveal the structures and functions of human music,” *Proceedings of the National Academy of Sciences of the United States of America*, vol.112, no.29, pp. 8987–8992, 2015.

[20]. P. E. Savage, P. Loui, B. Tarr, A. Schachner, L. Glowacki, S. J. Mithen, and W. T. Fitch, “Music as a coevolved system for social bonding.” In preparation.

[21]. P. E. Savage, A. T. Tierney, A. D. Patel. “Global music recordings support the motor constraint hypothesis for human and avian song contour,” *Music Perception*. 34. 327-334, 2017.

[22]. J. Six, O. Cornelis, and M. Leman. “Tarsos, a Modular Platform for Precise Pitch Analysis of Western and Non-Western Music,” *Journal of New Music Research*, vol.42, no.2, pp. 113-129, 2013.

[23]. A. T. Tierney, F. A. Russo, and A. D. Patel, “The motor origins of human and avian song structure,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 37, pp. 15510–15515, 2011.

[24]. A. T. Tierney, F. A. Russo, and A. D. Patel, “Empirical comparisons of pitch patterns in music, speech, and birdsong,” in *Proceedings of the Acoustics’08 Paris conference*, 2008.