SIMILARITIES AND DIFFERENCES IN A CROSS-LINGUISTIC SAMPLE OF SONG AND SPEECH RECORDINGS

YUTO OZAKI[1], JIEI KUROYANAGI[2], JOHN MCBRIDE[3], POLINA PROUTSKOVA[4], ADAM TIERNEY[5], PETER PFORDRESHER[6], EMMANOUIL BENETOS[4], FANG LIU[7], PATRICK E. SAVAGE[2*]

[*]Corresponding Author: psavage@sfc.keio.ac.jp
[1]Graduate School of Media and Governance, Keio Univ., JP
[2]Faculty of Environment and Information Studies, Keio Univ., JP
[3]Center for Soft and Living Matter, Institute for Basic Science, KR
[4]School of EECS, Queen Mary Univ. of London, UK
[5]Department of Psychological Sciences, Birkbeck, Univ. of London, UK
[6]Department of Psychology, Univ. at Buffalo, NY, US
[7]School of Psychology and Clinical Language Sciences, Univ. of Reading, UK

1. Introduction: Music and language are prominent forms of acoustic communication across human societies. The relationship between these two modes of human sound communication has been investigated from the perspective of music and language evolution (e.g. Fitch, 2006; Patel, 2007). Song and speech are usually considered as distinct categories, but the boundary is not always clear (Brown, 2017; Cummins, 2020; Deutsch et al., 2011; Engelhardt & Bretèque, 2017; Feld & Fox, 1994). Many studies have documented acoustic differences between song and speech in specific languages, but few studies have identified consistent cross-cultural similarities or differences across a diverse set of languages (see the references cited in the Result section). Therefore, we aim to conduct a series of comparative analyses with diverse cross-cultural samples to explore more general acoustic similarities and differences. In particular, we analyzed the following acoustic features: fundamental frequency (F0) of voice, inter-onset interval (IOI) of vowel onsets, ratio of F0 between adjacent IOI segments (interval). Onset annotations were performed manually and intervals were calculated by taking the outer product of an F0 vector at an IOI segment and

the reciprocal of the F0 vector at the previous IOI segment which resulted in the distribution of pitch ratio of adjacent IOI segments[1].

2. Dataset: We analyzed 23 pairs of song and speech recordings (i.e. 46 audio files). Each pair was recorded by the same person. 36 recordings were sampled from Hilton et al. (in press) that include English, Mandarin and Spanish. The remaining 10 recordings were newly collected ones in Japanese, English, Mandarin and Korean. The latter data was created by singing a song first and then reciting the text of the sung lyrics.

3. Results: We observed that song has higher F0 and longer IOI than speech, as reported previously (Hansen et al., 2020; Merrill & Larrouy-Maestri, 2018; Sharma et al., 2021, but see also Ding et al., 2017 who used amplitude modulation). We also observed that song has a sharper concentration at an IOI ratio (Roeske et al., 2020) of 0.5 than speech though both data has a median around 0.5. Our nPVI analysis both failed to support Patel & Daniele (2003)'s song-speech relationship hypothesis and failed to sort languages into traditional syllable-/stress-/mora-timed categories. Speech and song showed similar ranges of melodic intervals (within ±700 cents). However, the interval distribution of song has distinct peaks at around ±200 cents in addition to 0 cent, while speech had only one peak at 0 cents. In addition, we measured the variability of F0 by entropy (like Ozaki et al., 2022), and English and Spanish showed that singing has lower entropy than speech indicating greater pitch stability of singing (Merrill & Larrouy-Maestri, 2018; Raposo de Medeiros et al., 2021; Sharma et al., 2021; Thompson, 2014), while Mandarin showed the opposite pattern[1].

4. Conclusion: In summary, we observed the above potential systematic differences and similarities between song and speech. Why and how these differences have emerged and how evolutionary theories of language and musicality can account for these are the key questions to be addressed in future research (Darwin, 1871; Brown, 2000; Tierney et al., 2011; Savage et al., 2021; Mehr et al., 2021). For example, slower and more regular vocal communication may facilitate synchronization (social bonding hypothesis, Savage et al., 2021), while similar melodic interval range and 1:1 IOI duration focus may be due to shared constraints on the vocalization mechanism (motor constraint hypothesis, Tierney et al., 2011). Future steps to explore more comprehensive and robust relationships between two universal human acoustic communication forms include increasing sample size and language diversity, timbre/formant analysis, and measuring inter-rater reliability for the syllable/note annotations.

---

[1] Further details and figures can be found in the supplementary materials.
https://drive.google.com/file/d/1MDnk9miUzRMupb8pMtqlpdcuQ-I7w-CY/view?usp=sharing

Acknowledgements

We thank Sangbuem Leonard Choo for providing Korean language data. This work was supported by Grant-in-Aid no. 19KK0064 from the Japan Society for the Promotion of Science and by startup grants from Keio University (Keio Global Research Institute, Keio Research Institute at SFC, and Keio Gijuku Academic Development Fund) to P.E.S, and by Grant Number JPMJSP2123 of Support for Pioneering Research Initiated by the Next Generation from the Japan Science and Technology Agency to Y.O.

References

Brown, S. (2000). The "Musilanguage" model of music evolution. In N. L. Wallin, B. Merker, & S. Brown (Eds.), *The origins of music* (pp. 271–300). MIT Press.

Brown, S. (2017). A Joint Prosodic Origin of Language and Music. *Frontiers in Psychology*, *8*. https://www.frontiersin.org/article/10.3389/fpsyg.2017.01894

Cummins, F. (2020). The Territory Between Speech and SongA Joint Speech Perspective. *Music Perception*, *37*(4), 347–358. https://doi.org/10.1525/mp.2020.37.4.347

Darwin, C. (1871). *The descent of man*. Watts & Co.

Deutsch, D., Henthorn, T., & Lapidis, R. (2011). Illusory transformation from speech to song. The Journal of the Acoustical Society of America, 129(4), 2245–2252. https://doi.org/10.1121/1.3562174

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews*, *81*, 181–187. https://doi.org/10.1016/j.neubiorev.2017.02.011

Engelhardt, J., & Bretèque, E. A. de la. (2017). Guest Editors' Preface: Speech, Song, and In-Between. *Yearbook for Traditional Music*, *49*, xv–xix. https://doi.org/10.5921/yeartradmusi.49.2017.00xv

Feld, S., & Fox, A. A. (1994). Music and Language. *Annual Review of Anthropology*, *23*, 25–53.

Fitch, W. T. (2006). The biology and evolution of music: A comparative perspective. *Cognition*, *100*(1), 173–215. https://doi.org/10.1016/j.cognition.2005.11.009

Hansen, J. H. L., Bokshi, M., & Khorram, S. (2020). Speech variability: A cross-language study on acoustic variations of speaking versus untrained singing. *The Journal of the Acoustical Society of America*, *148*(2), 829. https://doi.org/10.1121/10.0001526

Hilton, C. B., Moser, C. J., Bertolo, M., Lee-Rubin, H., Amir, D., Bainbridge, C. M., Simson, J., Knox, D., Glowacki, L., Galbarczyk, A., Jasienska, G., Ross, C. T., Neff, M. B., Martin, A., Cirelli, L. K., Trehub, S. E., Song, J., Kim, M.,

Schachner, A., … Mehr, S. A. (in press). Acoustic regularities in infant-directed speech and song across cultures. *Nature Human Behaviour*. https://doi.org/10.1101/2020.04.09.032995

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

Merrill, J., & Larrouy-Maestri, P. (2017). Vocal Features of Song and Speech: Insights from Schoenberg's Pierrot Lunaire. *Frontiers in Psychology*, *8*, 1108. https://doi.org/10.3389/fpsyg.2017.01108

Ozaki, Y., Sato, S., Mcbride, J., Pfordresher, P. Q., Tierney, A. T., Six, J., Fujii, S., & Savage, P. E. (2022). Automatic acoustic analyses quantify pitch discreteness within and between human music, speech, and birdsong. *Proceedings of the 10th International Workshop on Folk Music Analysis*. The 10th International Workshop on Folk Music Analysis, Sheffield, United Kingdom.

Patel, A. D. (2007). *Music, Language, and the Brain*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195123753.001.0001

Patel, A. D., & Daniele, J. R. (2003). An empirical comparison of rhythm in language and music. Cognition, 87(1), B35–B45. https://doi.org/10.1016/S0010-0277(02)00187-7

Raposo de Medeiros, B., Cabral, J. P., Meireles, A. R., & Baceti, A. A. (2021). A comparative study of fundamental frequency stability between speech and singing. *Speech Communication*, *128*, 15–23. https://doi.org/10.1016/j.specom.2021.02.003

Roeske, T. C., Tchernichovski, O., Poeppel, D., & Jacoby, N. (2020). Categorical Rhythms Are Shared between Songbirds and Humans. *Current Biology*, *30*(18), 3544-3555.e6. https://doi.org/10.1016/j.cub.2020.06.072

Savage, P. E., Loui, P., Tarr, B., Schachner, A., Glowacki, L., Mithen, S., & Fitch, W. T. (2021). Music as a coevolved system for social bonding. *Behavioral and Brain Sciences*, *44*. https://doi.org/10.1017/S0140525X20000333

Sharma, B., Gao, X., Vijayan, K., Tian, X., & Li, H. (2021). NHSS: A speech and singing parallel database. *Speech Communication*, *133*, 9–22. https://doi.org/10.1016/j.specom.2021.07.002

Thompson, B. (2014). Discrimination between singing and speech in real-world audio. *2014 IEEE Spoken Language Technology Workshop (SLT)*, 407–412. https://doi.org/10.1109/SLT.2014.7078609

Tierney, A. T., Russo, F. A., & Patel, A. D. (2011). The motor origins of human and avian song structure. *Proceedings of the National Academy of Sciences*, *108*(37), 15510–15515. https://doi.org/10.1073/pnas.1103882108