Supplemental Material S1.

Part A: Error Trial AEP Data From Task Condition

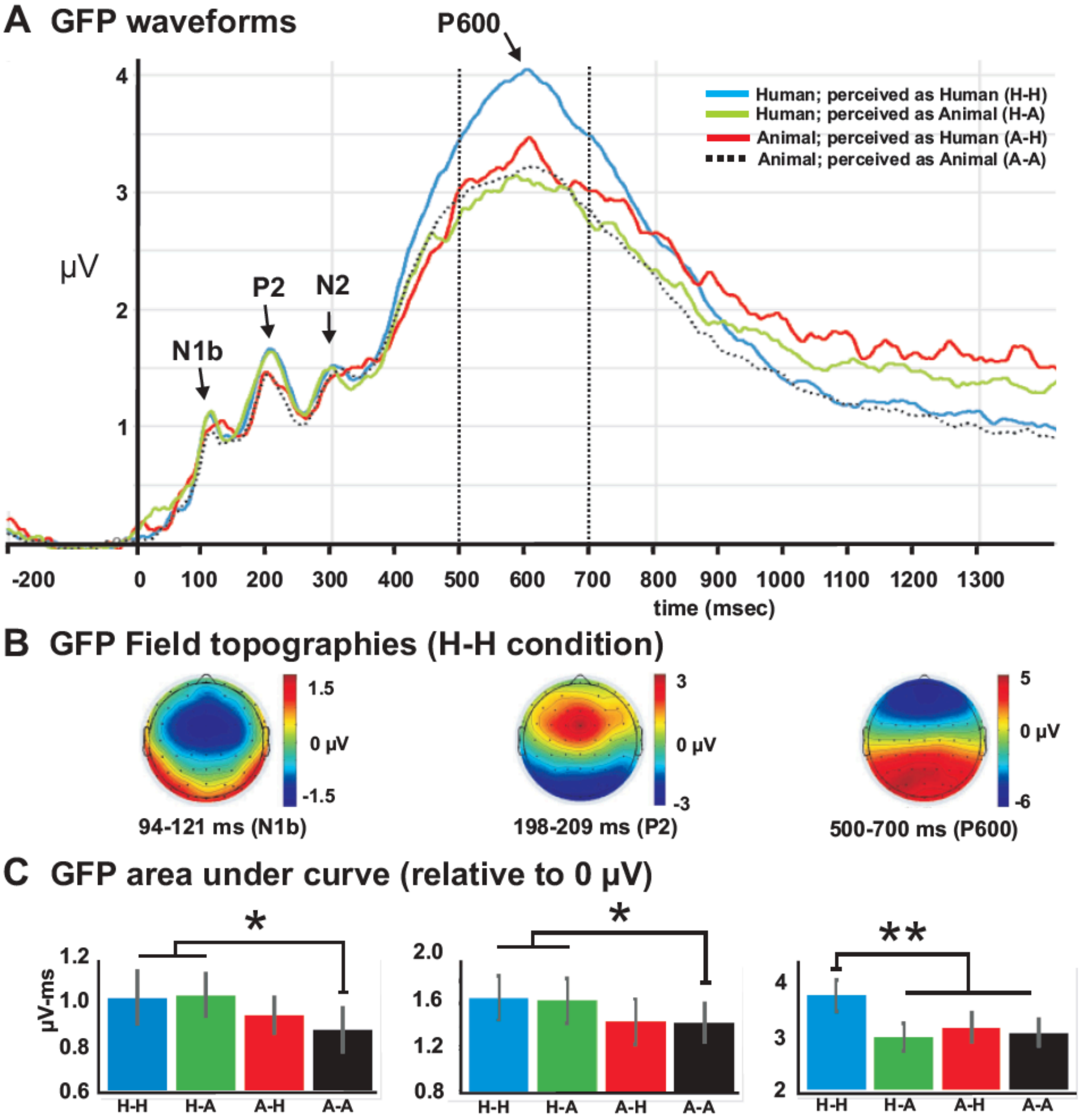
Methods for analyzing error trials from the task condition were identical to those describe in the main text. Participants with no retained error trials in one of the conditions, after artifact rejections, were further excluded from subsequent error trial analyses. A previous study of topographic consistency tests (TCT) for significance indicates that with 64 EEG sensors 50 or 100 observations (trials) yields significantly more reliable GFP/p-value results relative to, for example, 10 observations (Koenig & Melie-Garcia, 2010): The animal sounds perceived as human-mimics (A-H) during the two-alternative forced choice task condition of the present study resulted in a total of 24 trials, and thus represents a potential limitation to this aspect of the error trial analyses of this study. However, the other three conditions (H-H, A-A, and human-mimics perceived as animal vocalizations, H-A) fell within a range for greater reliability. Thus, all four waveforms were charted and analyzed. Figure S1 illustrates the group-averaged mean global field potential (GFP) responses to each of the four conditions. An N1b component in the range of 96–120 ms persisted for all four conditions: The H-A (green) error trial conditions was comparable in peak magnitude to the H-H condition (H-A vs A-A; two-tailed paired *t-test*$_{(53)}$ = 2.49, $p_{\text{uncorr}} < .02$), while the A-H (red) vs A-A (dotted black curve) condition only trended toward significantly greater N1b activation.

Interestingly, the GFP-derived P2 component appeared to be more sensitive to the human mimic sounds, regardless of whether they were correctly or incorrectly categorized (Fig. S1A, H-H blue curve, and H-A green curve), both relative to the animal vocalizations, whether they were correctly or incorrectly categorized (A-A black dotted curve, and A-H red curve). This effect was significant for both the H-H vs A-A contrast (two-tailed paired *t-test*$_{(28)}$ = 2.98, $p_{\text{uncorr}} < .006$) and H-A vs A-A contrast (*t*-test$_{(28)}$ = 2.48, $p_{\text{uncorr}} < .02$), but the A-H contrast (red curve) condition did not show statistical difference from any of the other three conditions in the P2 component time ranges.

Analyses of the averaged GFP-derived P600 response (in the 500–700 ms poststimulus range) was greatest for the H-H condition (Avg. ± Variance = 3.79 ± 2.52 µV; maximum peak at 602 ms), being significantly greater than each of the other three conditions (H-H vs. A-A; two-tailed paired *t-test*$_{(28)}$ = 6.74, $p_{\text{uncorr}} < 2.5 \times 10^{-7}$; H-H vs. A-H *t-test*$_{(28)}$ = 3.79, $p_{\text{uncorr}} < .0007$; and H-H vs. H-A, *t-test*$_{(28)}$ = 4.90, $p_{\text{uncorr}} < 3.7 \times 10^{-5}$), and these latter three conditions (H-A, A-H and A-A) did not significantly differ in averaged response amplitude from one another. A nonparametric Wilcoxon test indicated a significant difference between the H-H condition versus all other three at the 600 ms time point (at $p < .00001$ and $p_{\text{corr}} = .0001$ from nonparametric cluster permutation test adjusted for multiple comparisons).

The human mimic vocalizations, relative to animal vocalizations, showed a greater N1b response magnitude largely independent of task accuracy (e.g. blue and green versus red and dotted black curves). In contrast to the passive condition, however, auditory attention in the task condition altered the P2 component magnitude such that it became significantly greater for human vocalizations (H-H and H-A) regardless of participant perception accuracy, and thus appeared to have been under some level of top-down attentional modulation. Moreover, the addition of a discrimination task further evoked a P600 component that ostensibly was significantly greater only when human mimic vocalizations were *correctly* categorized as human-produced relative to all three of the other response conditions (H-A, A-H, and A-A). Together the differences in error trial (H-A and A-H) response profiles across the AEP components were suggestive of a temporal hierarchy of differential processing roles influenced by auditory attention. Thus, multiple stages (N1b, P2, possibly N2) of differential acoustic signal processing are presumably occurring along both spatial and temporal hierarchies to begin disentangling acoustic signal features characteristic of human mimic-voice, which occur prior to perceptual level differentiation related to the P600 component.

Figure S1. Group-averaged EEG responses to human mimic versus animal vocalizations during the task condition. (**A**) Averaged global field potential (GFP) waveforms (62 channels, $n = 29$ participants) corresponding to responses to the four possible outcomes from the task condition: human vocalizations correctly perceived as human (H-H); animal vocalizations incorrectly perceived as human (A-H); human vocalizations incorrectly perceived as animal (H-A); and animal vocalizations correctly perceived as animal-produced (A-A). (**B**) Scalp topography for the N1b, P2, and P600 components for the H-H condition (similar scalp topographies for the other three response conditions). (**C**) GFP area under the curve measures for the corresponding waveform intervals in panel B for N1b, P2, plus P600 (two-tailed paired t-tests: at least $^{*}p_{\text{uncorr}} < .02$, $^{**}p_{\text{uncorr}} < .007$). See text for other details.



Part B: Machine Learning Classification of Animal Vocals and Their Human Mimics

For the classifier learning algorithm, mean spectral peak values and standard deviations (in Hz) for the animal sounds and human mimics were derived (Table S1). Using eight acoustic-signal attributes and features (4 signal attributes plus the 4 derived spectral peaks) an overall classification accuracy (CA) was 82.1%, with classification accuracy of animal sounds at 80.2% (Table 1; 0 = *incorrect*, 1 = *correct categorization*) and corresponding human mimics at 84%. These results were comparable to the performance of both sets of human listeners (cf. Table 1: n33, n50, and CA columns). At each step of the stepwise procedure, the predictor that was found to be significant ($p < .05$) and that minimized unexplained variance (Wilks’ Lambda) in the model was entered. Spectral peak four (~6–7 kHz range) was the single best predictor for classifying the sound as produced by an animal or mimicked by a human (65.4% classification accuracy). However, other signals contributed such that no single attribute tested could fully categorize human mimics versus animal vocalizations, but in combination produced classification accuracy similar to the overall behavioral results. Accuracy across the groups ($n = 33$ EEG participants, $n = 50$ non-EEG participants) and the classifier algorithm were also examined (from Table 1, last three columns for each category). The two groups of human listeners showed greater similarity to each other in overall accuracy for correctly classifying the human mimic sounds ($R^2 = .63$) than for the animal sounds ($R^2 = .46$). Interestingly, several of the incorrect classifications of the algorithm were classified well by humans, and conversely some of the poorly classified stimuli by humans were correctly classified by the algorithm. Quantitatively, the classifier algorithm showed no correlation for overall accuracy in classification of human mimic sounds with the $n = 33$ EEG group ($R^2 = .0007$) nor the $n = 50$ non-EEG group ($R^2 = .00005$), though did show a trend for correlation in accuracy for the animal sounds with the $n = 33$ group ($R^2 = .17$) and the $n = 50$ non-EEG group ($R^2 = .15$). Overall, the classifier algorithm’s use of the six acoustics signal attributes we selected (HNR, F_0, WE, SSV, and four spectral peak bands) did not reveal any clear result regarding what specific acoustic features that the *human auditory system* may be utilizing to distinguish human mimic-voice from corresponding animal sounds.

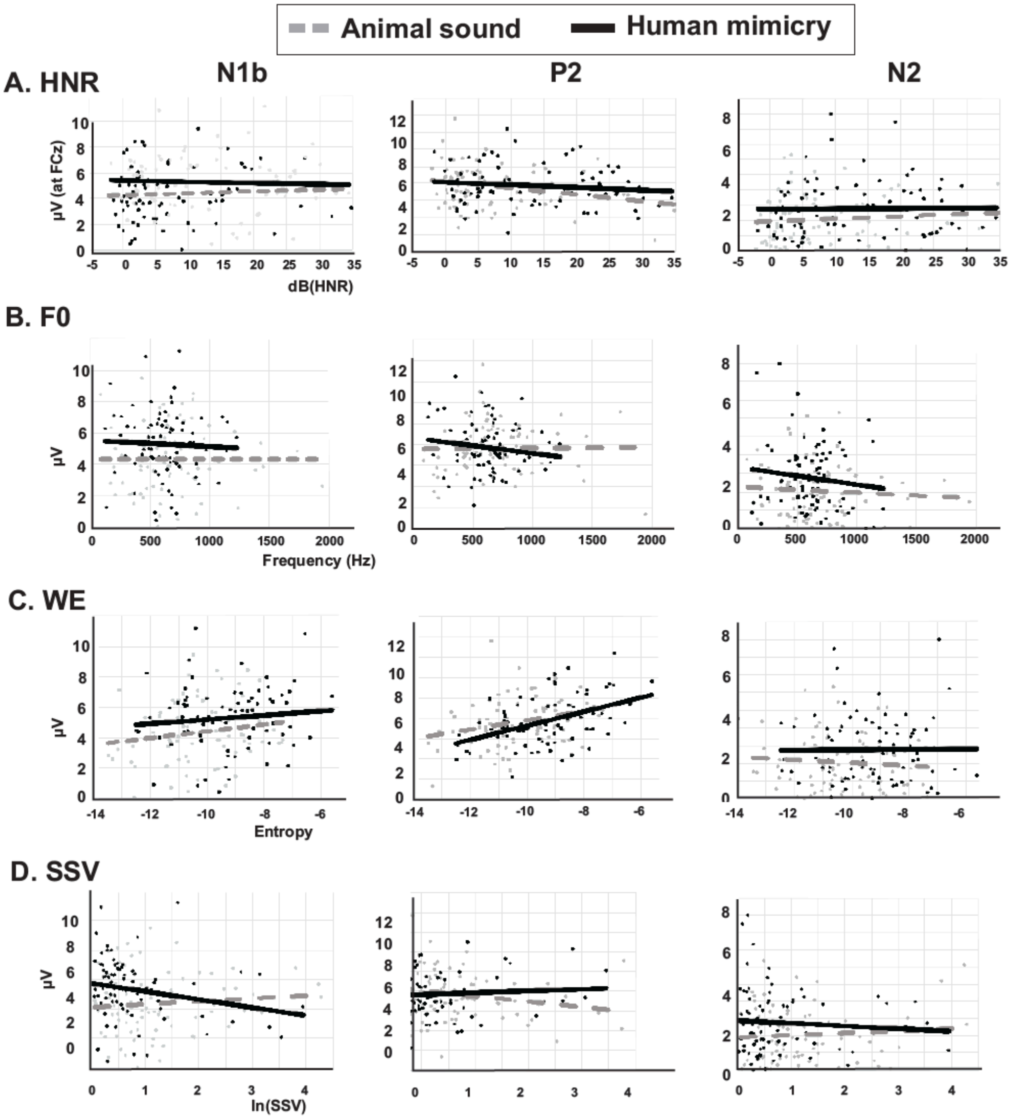
Table S1. Mean spectral peak values and standard deviation (in Hz) for animal sounds and human mimics.

Peak 1		**Peak 2**		**Peak 3**		**Peak 4**	
Animal	Mimic	Animal	Mimic	Animal	Mimic	Animal	Mimic
1330(480)	1195(427)	2769(543)	2846(744)	4526(624)	4788(196)	6238(652)	6665(576)

Part C: Parametric charting of acoustic signals versus AEP components

We next sought to identify acoustic signal attribute differences between animal and human mimic stimuli that might have been driving the early AEPs, using data from the passive condition (there were insufficient numbers of error trials for the task condition results to address this question). A peak and average response magnitude was derived in response to each of the 81 animal and 81 human mimic stimuli for the N1b, P2, and N2 components. These in turn were plotted parametrically (Fig. S2, using averages over corresponding component ranges) against various acoustic attribute values (from Table 1), including HNR, F_0, WE, and SSV attributes. While the linear slopes did not show significant R^2 fits (without binning), the trends were nonetheless informative. For instance, while HNR content (harmonicity) is reported to be a major driver of responses in auditory cortex (Kikuchi et al., 2014; Lewis et al., 2009, 2012), this attribute did not appear to influence differential responses of the N1b, P2, or N2 components across category since the slopes for the animal and human stimuli were nearly parallel (Fig. S2A). In contrast, parametric sensitivity to SSV (Fig. S2D) was different for animal and human sounds as suggested by slopes that intersected. While these results did not reveal specific bottom-up acoustic signal attributes that could distinguish conspecific human from non-conspecific voice, they do nonetheless reveal some novel avenues for future research on what signal attributes may ultimately drive the early AEP components associated with correct perception of "human voiceness."

Figure S2. Parametric correlations between various acoustic signal attributes of human-mimic (black lines) and animal vocalizations (dashed gray lines) with N1b, P2 and N2 component response amplitudes (microvolts) from the passive condition AEP data. Charts that show roughly parallel lines suggest that any parametric sensitivity of that component with that attribute may not be contributing to category distinction. Refer to text for other details. F_0 = fundamental frequency; HNR = harmonics-to-noise ratio; SSV = Spectral structure variation; WE = Wiener Entropy.



References

Kikuchi, Y., Horwitz, B., Mishkin, M., & Rauschecker, J. P. (2014). Processing of harmonics in the lateral belt of macaque auditory cortex. *Frontiers in Neuroscience, 8,* 204. https://doi.org/10.3389/fnins.2014.00204

Koenig, T., & Melie-Garcia, L. (2010). A method to determine the presence of averaged event-related fields using randomization tests. Brain Topography, 23(3), 233–242. https://doi.org/10.1007/s10548-010-0142-1

Lewis, J. W., Talkington, W. J., Tallaksen, K. C., & Frum, C. A. (2012). Auditory object salience: Human cortical processing of non-biological action sounds and their acoustic signal attributes. Frontiers in Systems Neuroscience, 6(27), 1–16. https://doi.org/10.3389/fnsys.2012.00027

Lewis, J. W., Talkington, W. J., Walker, N. A., Spirou, G. A., Jajosky,, A., Frum, C., & Brefczynski-Lewis, J. A. (2009). Human cortical organization for processing vocalizations indicates representation of harmonic structure as a signal attribute. Journal of Neuroscience, 29(7), 2283–2296. https://doi.org/10.1523/JNEUROSCI.4145-08.2009