Behavioral/Cognitive

The Role of Harmonic Resolvability in Pitch Perception in a Vocal Nonhuman Primate, the Common Marmoset (Callithrix jacchus)

Michael S. Osmanski, Xindong Song (宋欣东), and Xiaoqin Wang
Laboratory of Auditory Neurophysiology, Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205

Pitch is one of the most fundamental percepts in the auditory system and can be extracted using either spectral or temporal information in an acoustic signal. Although pitch perception has been extensively studied in human subjects, it is far less clear how nonhuman primates perceive pitch. We have addressed this question in a series of behavioral studies in which marmosets, a vocal nonhuman primate species, were trained to discriminate complex harmonic tones differing in either spectral (fundamental frequency \(f_0\)) or temporal envelope (repetition rate) cues. We found that marmosets used temporal envelope information to discriminate pitch for acoustic stimuli with higher-order harmonics and lower \(f_0\) values and spectral information for acoustic stimuli with lower-order harmonics and higher \(f_0\) values. We further measured frequency resolution in marmosets using a psychophysical task in which pure tone thresholds were measured as a function of notched noise masker bandwidth. Results show that only the first four harmonics are resolved at low \(f_0\) values and up to 16 harmonics are resolved at higher \(f_0\) values. Resolvability in marmosets is different from that in humans, where the first five to nine harmonics are consistently resolved across most \(f_0\) values, and is likely the result of a smaller marmoset cochlea. In sum, these results show that marmosets use two mechanisms to extract pitch (harmonic templates [spectral] for resolved harmonics, and envelope extraction [temporal] for unresolved harmonics) and that species differences in stimulus resolvability need to be taken into account when investigating and comparing mechanisms of pitch perception across animals.

Introduction

Pitch is one of the most important percepts in audition and a principle attribute of spectrally and temporally periodic sounds, including music and speech (Plack et al., 2005; Moore, 2012). Many vertebrate species are sensitive to pitch, including songbirds (Cynx and Shapiro, 1986; Lohr and Dooling, 1998; Dooling et al., 2002), chinchillas (Shofner, 2011), cats (Heffner and Whitfield, 1976), ferrets (Kalluri et al., 2008; Walker et al., 2009), and monkeys (Tomlinson and Schwarz, 1988). Pitch is not explicitly represented in the cochlea; instead, it must be computed centrally. For example, pitch can be extracted by calculating the specific frequencies of, and harmonic spacing between, the vibrations produced by the interaction of those harmonics on the basilar membrane. A compelling case has been made for temporal coding mechanisms in humans (de Cheveigne, 1998; Meddis and Hewitt, 1991a, 1991b; Meddis and O’Mard, 1997; Yost et al., 1996; but see Oxenham et al., 2011) and animals (Cariani and Delgutte, 1996a, 1996b; Simmons and Buxbaum, 1996). However, combining spectral and temporal mechanisms provides a more parsimonious explanation of human psychophysical data (Flanagan and Gutman, 1960a, 1960b; Gutman and Flanagan, 1964; Carlyn and Shackleton, 1994; Shackleton and Carlyn, 1994), suggesting that pitch extraction is dependent on frequency resolution in the auditory periphery.

Recent work has described a putative area specialized for pitch processing in auditory cortex of marmosets (Bendor and Wang, 2005, 2010), similar to that in humans (Patterson et al., 2002; Penagos et al., 2004; Bendor and Wang, 2006; Wang and Walker, 2012). In marmosets, this area contains neurons that encode pitch using spectral or temporal envelope cues (Bendor et al., 1996; but see Oxenham et al., 2011) and animals (Cariani and Delgutte, 1996a, 1996b; Simmons and Buxbaum, 1996). However, combining spectral and temporal mechanisms provides a more parsimonious explanation of human psychophysical data (Flanagan and Gutman, 1960a, 1960b; Gutman and Flanagan, 1964; Carlyn and Shackleton, 1994; Shackleton and Carlyn, 1994), suggesting that pitch extraction is dependent on frequency resolution in the auditory periphery.

Recent work has described a putative area specialized for pitch processing in auditory cortex of marmosets (Bendor and Wang, 2005, 2010), similar to that in humans (Patterson et al., 2002; Penagos et al., 2004; Bendor and Wang, 2006; Wang and Walker, 2012). In marmosets, this area contains neurons that encode pitch using spectral or temporal envelope cues (Bendor et al., 1996; but see Oxenham et al., 2011) and animals (Cariani and Delgutte, 1996a, 1996b; Simmons and Buxbaum, 1996). However, combining spectral and temporal mechanisms provides a more parsimonious explanation of human psychophysical data (Flanagan and Gutman, 1960a, 1960b; Gutman and Flanagan, 1964; Carlyn and Shackleton, 1994; Shackleton and Carlyn, 1994), suggesting that pitch extraction is dependent on frequency resolution in the auditory periphery.
Materials and Methods

Subjects. Six common marmosets (five male, one female) were used in these experiments: four in Experiment 1 and five in Experiment 2 (three were subjects in both experiments). Animals were housed in individual cages in a large colony at the Johns Hopkins University School of Medicine and were maintained at ~90% of their free-feeding weight on a diet consisting of monkey chow, fruit, and yogurt and had ad libitum access to water. Subjects were tested five days per week between the hours of 0900 and 1800. All experimental procedures were approved by the Johns Hopkins University Animal Care and Use Committee and were in compliance with the guidelines of the National Institutes of Health.

Operant testing apparatus. Marmosets were tested while seated in a Plexiglas restraining chair mounted in the center of a single-walled sound isolation chamber (Industrial Acoustic, model 400A) lined with 3-inch acoustic absorption foam (Pinta Acoustics). A food delivery tube was positioned at the top of the restraint chair in front of the animal; and a custom bracket, which housed an infrared photobeam, was mounted to the end of the feeding tube. Responses were recorded whenever the animal broke the photobeam by licking at the tube. Reward food consisted of a mixture of Similac baby formula, Gerber single-grain rice cereal, and strawberry-flavored Nesquik, delivered via a syringe pump (New Era Pump Systems, model NE-500) mounted to the base of the chair.

All sound stimuli were generated using MATLAB software (MathWorks) and delivered at a nominal sampling rate of 100 kHz through a digital signal processor and programmable attenuator (Tucker-Davis Technologies, RX6 and PA5) followed by an audio amplifier (Crown Instruments, model D-75). Stimuli were played from a speaker (Tannoy, model 4191) positioned at the same location as the animal’s head. Test sessions were monitored via portable webcam (Logitech, model B905).

Operant methods. The basic operant task has been described in detail previously (Osmanski and Wang, 2011; Remington et al., 2012). Briefly, each sound’s harmonics (spectral) and temporal envelope cues (temporal) or using the spacing between the sound’s harmonics (spectral). An ALT stimulus, however, has the same harmonic spacing (and thus the same pitch derived from the sound’s envelope repetition rate (i.e., the pitch derived from temporal envelope cues is one octave higher) (Fig. 1). Stimuli were 500 ms (20 ms rise/fall time, 200 ms interstimulus interval) and presented at an average level of 70 ± 4 dB SPL. Intensity was varied for each stimulus presentation to remove potential amplitude cues between stimuli with differing phase relationships. Stimuli were presented in blocks of 10 trials. Each block contained five COS targets, two ALT targets, and three catch trials. ALT stimuli essentially functioned as low-probability probe sounds to determine whether they were perceived in the same way as the higher-probability COS stimuli. We quantified responses in terms of $d'$ ($d' = z(\text{hit rate}) - z(\text{false alarm rate})$) (higher $d'$ values indicate better discriminability), as well as a corrected hit rate [hit rate$_{corr}$ = (hit rate − false alarm rate)/(1 − false alarm rate)] (Geisheider, 1985). Final $d'$ and hit rate values are based on a minimum of 300 trials. Sessions with a false alarm rate >25% were discarded (<10% of all sessions). Significance tests were conducted with a Bonferroni correction.

Experiment 2. Marmosets were trained to detect a 500 ms pure tone target (20 ms rise/fall time, 200 ms interstimulus interval) against a silent background. Targets were alternated with equal-duration periods of silence during the response window. We tested frequencies of 500, 1000, 4000, and 16,000 Hz. Seven different amplitudes (in 5 dB SPL steps) of a given tone frequency were presented per session and were chosen based on previously measured thresholds (Osmanski and Wang, 2011). Threshold was defined as the target amplitude corresponding to a hit rate of 50% correct detection, after a correction for false alarm rate. If a threshold could not be determined, tone amplitudes were adjusted in the subsequent session to bracket the presumed threshold. Animals were tested until threshold values stabilized (i.e., three successive sessions in which the obtained threshold was within ±1/3 step size). Final threshold estimates are based on a minimum of 300 trials. Sessions with a false alarm rate >25% were discarded (<12% of all sessions). Once thresholds were determined in quiet, the detection task was modified to include a continuous broadband noise played across the effective bandwidth of the speaker (0.8–54 kHz). Noise level was increased over several sessions until thresholds were elevated ~35 ± 5 dB relative to thresholds in quiet. The level of the noise was the same for all animals at a given frequency but varied across frequencies. The average noise spectrum level required to attain the criterion threshold shift across all frequencies tested was 34.8 ± 2.3 dB. Subsequent thresholds were measured by adding spectral notches to the noise background. Notched noise filter coefficients were generated in MATLAB and input to the
digital signal processor. Notches were logarithmically centered on the tone frequency with bandwidths of 0.125, 0.25, 0.5, 1, and 2 octaves. Only one bandwidth was presented per session, and the same bandwidth was used until threshold values stabilized.

Derivation of auditory filter shapes and ERB estimation. Auditory filter shapes were derived using a rounded exponential (roex) filter fitting function (using custom MATLAB software adapted from Glasberg and Moore, 1990). This procedure has been described in detail previously (Lina and Lauer, 2013). Briefly, the technique involves a minimization algorithm for calculating fitting parameters against the interpolated masked threshold data, which is determined by minimizing the root-mean-square error between the recorded threshold values and the filter-generated predicted values.

The power spectrum model of masking (Fletcher, 1940; Glasberg and Moore, 1990) is mathematically summarized as follows:

\[ P_s = K \int_{-\infty}^{\infty} N(f) W(f) df \]  

where the masked threshold of a signal \( P_s \) is defined by the power spectrum of the noise \( N(f) \) and the weighted function of the signal \( W(f) \) in the frequency domain. The constant \( K \) is a scalar adjustment factor equal to the signal-to-masker ratio at the output of the filter required to achieve threshold. Previous psychoacoustic studies (Patterson, 1976) using noise with spectral notches have found that the shape of the auditory filter can be approximated by a roex \( \text{roex}(p,r) \) fit as follows:

\[ W(g) = \text{roex}(p,r) = (1 - r) (1 + pg e^{-rg} + r) \]  

where \( p \) describes the filter skirt slopes and \( r \) flattens the filter at remote frequencies. The parameter \( g \) is the frequency deviation of the edge of the notch normalized by its center frequency \( f_c / f_0 \).

Equations 1 and 2 can be combined to yield Equation 3, which is analogous to Equation 1 converted into the power spectrum domain:

\[ P_s = K + 10 \log_{10}(N_f2 \int_{-\infty}^{\infty} \text{roex}(p,r)) \]

Equation 3 is then used to compare masked thresholds to the optimized roex \( \text{roex}(p,r) \) filter output values. Masked thresholds and filter magnitudes are plotted in terms of the deviation at the near (lower) edge. Finally, using the resultant roex \( \text{roex}(p,r) \) filter parameters, the ERB of the auditory filter shape can be calculated according to the following:

\[ \text{ERB} = 2f_0 \left[ \frac{1 - r}{p} \left( 2 - (2 + p)e^{-rg} \right) \right] + r \]  

Harmonic resolvability model. The lowest \( f_0 \) an auditory channel centered at \( f_c \) can resolve is assumed equal to 1 ERB (Moore and Ohgushi, 1993; Moore, 2012). Thus, the highest resolved harmonic for this \( f_0 \) can be estimated by the following equation:

\[ f_{\text{HRH}} = f_0 / \text{ERB} \]  

where \( f_0 \) is center frequency. Resolvability curves were cubic spline interpolated from the measured ERB \( f_0 \) points. Additionally, we used previously published neural frequency tuning data measured at each neuron’s best sound level (Bartlett et al., 2011) to estimate the resolvability of A1 neurons in marmoset. For these neural data, we substituted the ERB term in Equation 5 with the frequency response bandwidth (calculated by Bartlett et al., 2011) to derive the highest resolved harmonic of marmoset A1 neurons.

**Results**

**Experiment 1a: pitch discrimination.**

Animals were trained to discriminate one of two “targets” (2\( f_0 \) COS or \( f_0 \) ALT) from a repeating “background” (\( f_0 \) COS) (see Materials and Methods). We assumed that, if marmosets rely on spectral cues to extract pitch, they would respond to the 2\( f_0 \) COS target but not the \( f_0 \) ALT target. However, if they rely on temporal envelope cues, they would respond to both targets. Animals were tested using three stimulus sets: first nine harmonics present (H1–H9), a missing \( f_0 \) set containing the fourth through 12th
harmonics of a common f₀ (H4-H12), and first four harmonics present (H1-H4). Set H4-H12 was presented with a 3-octave-wide band of noise (10 dB signal-to-noise ratio) centered at f₀ to minimize potential harmonic distortion effects at the level of the cochlea. In all three stimulus sets, background sounds had f₀ values of 150, 300, 450, 600, and 900 Hz (2f₀ COS targets had f₀ values ranging from 300 to 1800 Hz).

Performance for ALT targets was highly dependent on f₀ in both H1-H9 and H4-H12 conditions, but not so for 2f₀ COS targets (Fig. 2A,B). We quantified performance in terms of d’ (see Materials and Methods) and found a significant decrease in discriminability for ALT stimuli as f₀ increased (H1-H9, F(4,12) = 12.59, p < 0.001; H4-H12, F(4,12) = 33.29, p < 0.001) (Fig. 2A, B, top). Performance for ALT and COS targets was indistinguishable at 150 Hz (H1-H9, t(3) = −1.09, p = 0.36, corrected; H4-H12, t(3) = −2.74, p = 0.071, corrected), but significantly different at 900 Hz (H1-H9, t(3) = −16.86, p < 0.001, corrected; H4-H12, 900 Hz: t(3) = −17.25, p < 0.001, corrected). Subjects required significantly more time to respond to ALT targets as f₀ increased (H1-H9, F(4,12) = 6.82, p < 0.01; H4-H12, F(4,12) = 10.42, p = 0.001) (Fig. 2A,B, bottom). However, discriminability was impaired across all five f₀ values in the H1-H4 condition, suggesting that subjects could not make discriminations based on temporal envelope cues in these stimuli (150 Hz, t(3) = −7.97, p < 0.01, corrected; 300 Hz, t(3) = −6.67, p < 0.01, corrected; 450 Hz, t(3) = −18.18, p < 0.01, corrected; 600 Hz, t(3) = −42.63, p < 0.01, corrected; 900 Hz, t(3) = −5.34, p < 0.05, corrected) (Fig. 2C). Overall, these results suggest that marmosets are sensitive to temporal envelope cues at f₀ values <450 Hz but rely more strongly on spectral cues for f₀ values >450 Hz, and that the first four harmonics of these complex tones are probably resolved.

**Experiment 1b: the role of harmonic order on temporal pitch processing**

We further examined the influence of stimulus resolvability on temporal envelope processing by testing discrimination of complex tones with harmonics confined to one of three spectral regions: low (H1-H4), medium (H5-H8), or high (H9-H12). Stimuli in the medium and high conditions were presented with a 3 octave noise band centered at f₀. Only f₀ values of 150 Hz and 900 Hz were used. If all harmonics are resolved, performance should be dependent on spectral cues alone and marmosets should respond only to COS (not ALT) stimuli. If all harmonics are unresolved, or if there is a mixture of resolved and unresolved harmonics (which produces an ambiguous pitch in human listeners) (Shackleton and Carlyon, 1994), temporal envelope cues should cause marmosets to respond to both targets.

Discriminability of ALT targets was highly dependent on spectral region at 150 Hz (F(2,6) = 24.23, p < 0.01), but not at 900 Hz (F(2,6) = 1.01, p = 0.42) (Fig. 3A,B, top). At 150 Hz, d’ values for ALT targets were significantly lower than COS targets in both low (t(3) = −7.79, p < 0.01, corrected) and medium (t(3) = −5.65, p < 0.05, corrected), but not high (t(3) = −2.08, p = 0.129, corrected), conditions, and response latency decreased as spectral region increased (F(2,6) = 0.777, p < 0.05) (Fig. 3A, bottom). d’ values for ALT stimuli at 900 Hz were significantly different from COS stimuli at all f₀ values (low, t(3) = −6.24, p < 0.01, corrected; medium, t(3) = −8.56, p < 0.01, corrected; high, t(3) = −4.95, p < 0.05, corrected) (Fig. 3B, top). These data suggest that at 150 Hz, but not 900 Hz, harmonics are probably unresolved above ~fifth.

**Experiment 1c: background-target stimulus reversal**

Experiment 1a showed that discrimination performance for f₀ ALT and 2f₀ COS targets is the same at 150 Hz for both H1-H9 and H4-H12 conditions (Fig. 2A,B, top), but it is unclear whether these stimuli are perceptually equivalent. We changed the background-target relationships so that a correct response required discriminating f₀ ALT and f₀ COS targets from a 2f₀ COS background. Animals thus had to report on an octave decrease. If f₀ ALT and 2f₀ COS stimuli are perceptually indistinguishable, then subjects should not respond to f₀ ALT targets. Results show that, for both H1-H9 and H4-H12 stimuli, marmosets were able to easily discriminate both target types with equal latency (hit rate: H1-H9, t(3) = 0.54, p = 0.628; H4-H12, t(3) = 0.88, p = 0.442; d’: H1-H9, t(3) = 0.74, p = 0.516; H4-H12, t(3) = 0.88, p = 0.444; response latency: H1-H9, t(3) = 0.059, p = 0.957; H4-H12, t(3) = −0.721, p = 0.523) (Fig. 3C).

**Experiment 2a: psychophysical estimates of frequency selectivity**

Results of Experiment 1 suggest that frequency resolution in the auditory periphery primarily determines whether a spec-
A central or temporal mechanism is used to extract pitch. We estimated frequency resolution by measuring ERBs using notched noise procedures (see Materials and Methods). Previous work has shown that auditory filter shapes can be approximated by applying roex filter fits to notched noise data (Glasberg and Moore, 1990). The ERB transforms this filter into a rectangular shape with the same peak and total transmission. We tested four tone frequencies (500, 1000, 7000, and 16,000 Hz) drawn from across the marmoset hearing range (Osmanski and Wang, 2011).

Psychometric functions were measured in seven conditions (quiet, broadband noise, and five different spectral notch widths). Examples of individual psychometric functions are shown in Figure 4A. Thresholds compared favorably with previously measured audiograms for this species (Osmanski and Wang, 2011). Masking effects are indicated by a rightward shift in the psychometric functions measured in noise relative to quiet. The largest shift is seen for broadband noise, and adding spectral notches reduced the magnitude of this shift. Individual thresholds for each notch width were averaged to derive the input to the roex function (see Materials and Methods). These thresholds, along with the best fitting function to the masked threshold data, are shown in Figure 4B. Filter parameters obtained from the fitting algorithm were used to derive the filter shapes, which were then converted into ERBs (Fig. 4C).

Psychophysical measures of frequency selectivity using notched noise data have been described for only a few species, and these data are shown in Figure 5 (mouse, May et al., 2006; chinchilla, Niemiec et al., 1992; guinea pig, Evans et al., 1992). ERBs are typically between 10% and 20% of center frequency across species. Those measured in the current experiment vary from 6.6% (7 kHz) - 25.4% (0.5 kHz) (Fig. 5). Marmoset ERBs tend to be wider than those measured in

---

**Figure 4.** ERB estimation. A, Summary psychometric functions for each of the four tone frequencies in seven conditions (quiet, broadband noise, and five bandwidths of notched noise). Representative performance from example subjects is plotted as the proportion of correct responses as a function of tone amplitude. Threshold = 50% correct. B, Effects of notched noise bandwidth on detection thresholds for all five subjects (colored symbols) for each tone frequency. The best fitting function is shown (solid black line) fitted to the average of each deviation (black open circles). C, Estimated roex(p,r) filter functions plotted against the interpolated average notched-noise thresholds across all five subjects.
humans, except at 7000 Hz, where bandwidths were narrower than those of the other four species.

Experiment 2b: modeling harmonic resolvability

Our ERB estimates were used to model harmonic resolvability in marmosets. The highest resolved harmonic of a complex tone is approximately equal to the center frequency of the auditory filter divided by its ERB (see Materials and Methods). We compare resolvability for marmosets and humans in Figure 6A. The first five to nine harmonics of a complex tone are consistently resolved in humans (Plomp, 1964; Glasberg and Moore, 1990). In marmosets, however, the first four or five harmonics are resolved at an f₀ of 150 Hz, and this number increases to the first 14–16 harmonics between 600 and 800 Hz (Fig. 6A) before decreasing again at higher f₀ values. These values show a remarkable correspondence to our pitch discrimination data, where temporal cues could be reliably used to discriminate complex tones at 150 Hz f₀ (but only when stimuli contained harmonics > fifth), and this ability decreased as f₀ increased (up to 900 Hz, presumably as the stimuli became entirely composed of fully resolved harmonics).

Previous work has shown agreement between ERBs measured psychophysically and those derived from the tuning bandwidths of cortical neurons (Evans et al., 1992; Fishman and Steinschneider, 2006; Bitterman et al., 2008). We estimated the harmonic resolvability of marmoset A1 neurons based on previously measured cortical tuning bandwidths in this species (Bartlett et al., 2011) and show that these data show a qualitative agreement with our psychophysical estimates (Fig. 6B). Behavioral thresholds, which presumably reflect spectral resolvability in the auditory periphery, generally describe a lower resolvability boundary for these cortical neurons. This relationship is consistent with the idea that a sensory threshold is determined by the most sensitive neurons contributing to a given behavior (Parker and Newsome, 1998). The overall poorer spectral resolvability measured across neurons in cortex compared with psychophysical thresholds is likely the result of information being increasingly integrated at successive processing stages of the auditory system.

Discussion

We trained common marmosets to discriminate complex harmonic tones differing in their phase relationships to disentangle the relative use of temporal envelope and spectral cues in extracting pitch from complex sounds. Our results suggest that marmosets are sensitive to both cues and that the primary determinants of which mechanism is used for pitch extraction are f₀ and harmonic order. Specifically, marmosets were able to use changes in the envelope repetition rate of these sounds to extract pitch when stimuli contained low f₀ values and high harmonic orders, whereas spectral information was used for higher f₀ values with low harmonic orders. In a second experiment, we estimated auditory filter ERBs using notched noise techniques. Results show that marmoset ERBs are generally comparable to other mammalian species. Marmosets had wider ERBs compared with humans, except at 7000 Hz, which is approximately the frequency of best hearing in this species and also the frequency where most of the spectral energy in marmoset vocalizations is concentrated (Osmanski and Wang, 2011).
Threshold levels (average stimulus level was 70 dB SPL). Calculation, previously studied in humans, which operate well above pitch task was designed to examine mechanisms of pitch perception at relatively high levels (Moore and Glasberg, 1976). However, much sharper estimates of auditory filter tuning at low levels. Auditory filter bandwidths have been described in humans using a variant of the forward, compared with simultaneous stimuli, including changes in the position of the highest harmonic. Further work will be required to tease apart these possible alternatives to determine whether marmosets are truly sensitive to spectral mechanisms of pitch extraction.

In conclusion, these behavioral data support a dual pitch processing mechanism whereby both harmonic templates (spectral) and envelope extraction (temporal) are used to compute pitch depending on the resolvability of the stimulus spectrum. Although previous work has shown that primates are able to use pitch cues to make judgments about sound (May et al., 1989; Wright et al., 2000; Izumi, 2001; Brosch et al., 2004), we think that these data represent the first rigorous psychophysical tests of pitch perception in nonhuman primates that begin to address questions of underlying mechanism. Marmosets, an emerging model system in auditory neuroscience with a large vocal repertoire and an easily accessible auditory cortex (Wang, 2000, 2007), are an ideal species for questions of this kind. Treating apart the various mechanisms that give rise to pitch perception in marmosets will provide valuable comparative data, which will advance our understanding of how the auditory system works to create a unified pitch percept in different animal species, including humans.

References

