Mechanisms for Communicating in a Marmoset 'Cocktail Party'

Vladimir Jovanovic & Cory T Miller

Cortical Systems and Behavior Laboratory
Neurosciences Graduate Program
University of California, San Diego

Abstract.

A key challenge for audition is parsing the voice of a single speaker amid a cacophony of other voices known as the Cocktail Party Problem (CPP). Despite its prevalence, relatively little remains known about how our simian cousins resolve the CPP for active, natural communication. Here we employed an innovative, multi-speaker paradigm comprising five computer-generated Virtual Monkeys (VM) whose respective vocal behavior could be systematically varied to construct marmoset cocktail parties and tested the impact of specific acoustic scene manipulations on subjects' natural conversations. Results indicate that marmosets not only employ auditory mechanisms – including attention – for speaker stream segregation, but also selectively change their own vocal behavior in response to the dynamics of the acoustic scene. These findings suggest notable parallels with human audition to solve the CPP and highlight the active role that individuals play to optimize communicative efficacy in complex real-world acoustic scenes.

Introduction.

Our ability to effectively communicate with others is often complicated by the co-occurrence of other speakers in an acoustic scene, classically illustrated by the Cocktail Party Problem [CPP] (Cherry, 1953; Bee and Micheyl, 2008; Mcdermott, 2009; Bronkhorst, 2015). The seeming effortlessness with which audition solves the myriad of challenges inherent to the CPP belies the suite of sophisticated mechanisms throughout multiple levels of the auditory system – including both peripheral and central neural processes - that must work in parallel to make it possible to engage in conversations with others in such acoustic environments (Pressnitzer et al., 2008). Although human and nonhuman primates share the core architecture of the cortical auditory system that distinguish our Order from other taxa (Kaas and Hackett, 1998, 2000; Hackett, 2009; Kaas, 2010) and face similar challenges of communicating in environments comprising multiple conspecifics vocalizing in concurrence (Waser and Brown, 1986), only a handful of neurophysiological and behavioral experiments have explored whether more general auditory scene analysis mechanisms are evident in our simian cousins (Miller et al., 2001; Micheyl et al., 2003; Petkov et al., 2003; Micheyl et al., 2005). Though suggestive, precisely how the myriad of perceptual and cognitive mechanisms needed to effectively communicate in dynamic cocktail party scenes are employed by nonhuman primates is not yet clear. Here we sought to bridge this notable gap by implementing an innovative, multi-speaker, interactive playback paradigm for common marmosets (*Callithrix jacchus*) that simulates a natural cocktail party while providing experimental control to systematically manipulate multiple dimensions of the acoustic scene. Our aim was not to determine psychoacoustic thresholds of the perceptual processes that reflect auditory scene analysis mechanisms applied generally in audition (Bregman, 1994), but rather to explicate how these mechanisms - and potentially others - are leveraged under real-world conditions to overcome the CPP for active communication in a nonhuman primate.

Studies suggest that humans are able to resolve the challenges of listening in multi-talker scenes using a handful of perceptual cues, including the spatial separation of the speakers and the acoustic idiosyncrasies of individual voices (Darwin, 1997; Bronkhorst, 2015). Under natural conditions, listeners are typically given important cues that facilitate parsing an acoustic scene, including each persons' voice and the fact that each person is speaking from a distinct location in space. Even relatively small distances between talkers can increase intelligibility significantly while differences in each speaker's voice pitch provides a reliable cue (Bregman, 1994; Brungart and Simpson, 2002; Brungart and Simpson, 2007). In more dynamic scenes involving numerous talkers, these cues may become less clear, requiring listeners to employ top-down perceptual mechanisms to selectively attend to a particular individual's voice (Darwin and Hukin, 2000; Hill and Miller, 2009). During speech, one could learn a talker's voice and segregate it into a single stream, potentially as a learned schema (Bregman, 1994; Bey and McAdams, 2002; Darwin, 2008; Johnsrude et al., 2013; Bronkhorst, 2015; Woods and McDermott, 2018), facilitating its segregation from other sounds in the acoustic landscape. Certainly observations indicate that primates are able to communicate in noisy environments, but whether this is accomplished because nonhuman primates are talented acoustic scene analyzers that rely primarily on bottom-up auditory mechanism, as is the case in some other nonhuman animals (Bee and Micheyl, 2008), or are able to use more top-down attentional processes characteristic of humans (Wood and Cowan, 1995; Bronkhorst, 2000; Shinn-Cunningham, 2008; Hill and Miller, 2009) is not yet known. Furthermore, because human experiments have typically relied on intelligibility, forced choice or similar methods, rather than active communicative exchanges between speakers, it is not yet known whether these auditory mechanisms alone are sufficient to solve the CPP or if they are complemented by processes that have not yet been described.

Common marmosets are highly voluble New World monkey who naturally engage in long-distance conversational exchanges within noisy, dynamic communication networks that reflect

the CPP (Eliades and Miller, 2017). Like human conversations, the temporal dynamics of these marmoset conversations are governed by learned social rules (Miller and Wang, 2006; Chow et al., 2015; Toarmino et al., 2017). Building on our previous interactive playback paradigm (Miller et al., 2009b; Miller and Thomas, 2012), we used a multi-speaker design to construct cocktail parties in which a single live monkey heard phee calls – the species-typical long-distance contact call (Miller and Wang, 2006; Miller et al., 2010) – produced by five Virtual Monkeys (VMs) whose respective vocal behavior differed relative to the subjects. In this innovative design, the behavior of one VM – the Target – was designed to directly interact with the live marmoset, emitting vocalizations in response to the subject in order to engage them in conversational exchanges, while the timing of the other VMs – the Distractors – were independent of the subject. Phee calls from pairs of VM Distractors were constructed in temporal sequences that simulated natural conversational exchanges, such that concurrent conversations between VM Distractor pairs were broadcast in each cocktail party. This innovative paradigm afforded a powerful opportunity to systematically manipulate features of the acoustic scene (e.g. spatial separability and predictability of VM location, distractor density, and the acoustic structure of the vocalizations themselves) in order to explicitly test their effect on subjects' propensity to engage in conversational exchanges with the Target VM; thus providing key insights into the mechanisms that this nonhuman primate employs to overcome the challenges of communicating in a cocktail party. Because marmoset phee calls are individually distinctive and recognizable in normal conversations (Miller et al., 2010; Miller and Thomas, 2012; Toarmino et al., 2017), we predicted that subjects here would be able to distinguish the individual VMs based on their respective vocal behavior unless specific manipulations of the scenes affected the mechanisms that support these key processes.

Results.

We tested 6 adult common marmoset monkeys (3 Male and 3 Female) in a series of experiments designed to examine the mechanisms that support communicating in a cocktail party. We completed two experiments comprising a total of fourteen different test conditions. Notably, we observed no statistically significant difference in the number of vocalizations produced by subjects across these condition (3-way ANOVA, total $df = 251$, $p = 0.693$). The mean amount of phee calls produced by each subject within each 30min recording session per condition was 60.4 calls with a standard deviation of 34.3 calls. This suggests that none of the constructed cocktail parties for these experiments caused an overt suppression of marmoset vocal behavior.

One dimension of their vocal behavior that did vary across the test conditions was subjects' propensity to engage in conversational exchanges with the Target VM. Conversations are a particularly insightful behavioral epoch to explicate how specific manipulations of the cocktail party because they are a coordinated, reciprocal vocal exchange that abides social rules (Stivers et al., 2009; Chow et al., 2015; Levinson, 2016). During conversations involving the marmoset long-distance contact calls used as stimuli in these experiments (phee calls), individuals perceive conspecific calls produced within 10s of their own calls as an 'antiphonal' response (Miller et al., 2009b; Toarmino et al., 2017). Thus, consecutive 'antiphonal' responses between a subject and VM would indicate a conversational exchange and be an indicator of active engagement (i.e. successful communication).

Here we defined a conversation as any instance in which two or more phee calls were produced in an alternating temporal sequence between the subject and a VM in which the inter-call interval was <10s. While short conversations could emerge inadvertently based on standard call rate of the marmosets, the probability of long conversations occurring by chance was lower due to the coordinated nature of the behavior. At the 95th percentile and above of conversation

lengths in a test session, we found a significant difference in these maximum maintained conversations. Across all spatial configuration conditions in Experiment 1 and 2, VM Distractors had a mean maximum of 4.97 calls in a conversation with subjects while Target VM had 13.6 calls in each conversation (1-Way ANOVA (VM type), total df = 514, VM df = 1, $p < 0.000$) suggesting that the deliberate interactive vocal behavior of the Target VMs facilitated long conversational exchanges.

To ascertain that 'conversations' represented active communication exchanges between the Target VM and subject, rather than a byproduct of periodic calling by the subject, we further examined the distributions of the conversation lengths. For each Experiment and Distractor Density, we tested subjects on a Baseline condition that was identical to the Fixed-Location condition, with one key difference. During Baseline sessions, we did not broadcast an audible vocalization stimulus from the Target VM as occurred in all other test conditions. Rather, a 'silent' stimulus occurred at the same response latency (1-5s) and probability as the Target VM in other conditions. The rationale was to distinguish between facets of vocal behavior that would emerge independent of the Target VM and those that were dependent on hearing Target VM vocalization stimuli and their related behavioral characteristics.

Figure 2A plots the probability density estimate of the conversation lengths for both the Baseline and Fixed-Location for all test conditions (normal kernel function via ksdensity in MATLAB R2020a). While overlap was evident between these distributions, particularly at the low end (i.e. shorter conversations), this occurred because the Target VM latency to respond was short (1-5s) and identical between the conditions. By focusing analyses on the 95th percentile of conversation lengths within each session, however, significant differences were evident. Whereas the mean max conversation length for Baseline was 11.4 calls, the Fixed condition yielded 14.8 calls (1-Way ANOVA (Condition), total df = 165, Condition df = 1, $p = 0.0232$). The two distributions plotted in Figure 2A are further color coded by the number of standard deviations from the median of the overall distribution. The inset in the upper right shows the amount of actual data from each distribution that fit into each bin in stacked bars. In both plots shown in 2A, the Fixed condition has a longer tail and is shifted to the right (i.e. longer conversations) in comparison to Baseline conversation lengths, explaining the significant difference in distributions. The occurrence of longer conversations accounting for a greater proportion of conversations in the Fixed-Location, indicates that subjects are able to maintain longer conversations on average when the Target VM is audible, while shorter conversations can emerge based on spontaneous, periodic calling from marmosets in the Baseline conditions.

Based on these analyses, we employed a single behavioral metric that encapsulated the number and length of conversations that occurred in each test session called the 'Conversation Index' (Figure 2B). Each call in a conversation was given a number that increased linearly at each consecutive call in the conversation. The first call received a 0, the second a 1, the third a 2, and so on. Any breaks in a conversation, defined as an inter-call interval greater than 10s, reset the value to zero. These conversation numbers were summed and divided by the total number of the subject's calls to yield the Conversation Index for each test session. Applying this metric to compare subjects' vocal behavior for all Baseline and Fixed-Location conditions we found that the VM Distractor Conversation Indices were significantly lower than the Target VM in both Conditions (2-Way ANOVA (Condition, VM), total df = 42894, VM df = 4, $p < 0.0001$; Condition*VM df = 4, $p < 0.0001$). As shown in Figure 2C, the Target VM Conversation Index was significantly higher than Distractor VMs in both the Baseline and Fixed-Location conditions (Tukey-Kramer multiple comparisons corrected $p < 0.0001$ for each, respectively). Moreover, the Target VM Conversation Index was significantly higher in the Fixed-Location than the Baseline condition (Tukey-Kramer multiple comparisons corrected $p < 0.0001$). These analyses indicate that the Conversation Index reliably represents marmosets' conversations and can be used to compare subjects' behavior across the different cocktail parties.

Experiment 1.

Here we sought to test how two dimensions of the cocktail party – Distractor Density and Spatial Configuration (Figure 3A) – affected marmosets' propensity to engage in conversational exchanges. Each of these facets of the acoustic scene are known to influence how humans resolve the CPP (Bronkhorst, 2015). These experiments broadcast 2-Pulse phee call stimuli from each VM at two Distractor Density levels – High and Low – in three spatial configurations – Fixed-Location, Random-Location, and Single-Location (Figure 3A) – as well as the Baseline condition. Results from Experiment 1 are shown in Figure 3.

Distractor Density was calculated as the ratio of the Target VM calls that temporally overlapped with Distractor VM calls. 'Low' Distractor Density had a mean interference ratio of 0.724 across all the sessions. In other words, on average,72% of the duration of the Target VM calls broadcast acoustically overlapped with one or more of the Distractor VMs calls. By contrast, the mean interference ratio for 'High' Distractor Density in this experiment was 0.903 (i.e. 90%). The standard error was 0.00358 for both Distractor Density levels. At both Distractor Density levels, the Conversation Index of subject calls with respect to Target VM was significantly higher than the Distractor VMs (1-way ANOVA, total df = 43114, VM df = 4, $p = 0$). As a result, we did not further include Distractor VM Conversation Index values in any additional analysis.

Figure 3B shows subjects' Conversation Index for Baseline and the three spatial configurations of Fixed-, Random-, and Single-Location at the Low Distractor Density level. Although subjects exhibited a significant broad increase in Conversation Index across all conditions relative to Baseline (1-Way ANOVA (Spatial), total df =4586, Spatial df = 3, $p < 0.0002$), this was largely driven by two of the three spatial conditions. For both the Fixed- and Single-Location conditions, subjects' Conversation Index was significantly higher relative to Baseline (Tukey-Kramer multiple comparison corrected $p = 0.0016$ and $p < 0.0004$, respectively), but not for the Random-Location condition ($p = 0.221$). These data suggest that at this Distractor Density, marmosets were able to correctly identify the Target VM and selectively engage that individual in conversational exchanges. The pattern of subjects' behavior across these conditions suggests that spatial configuration of the VMs does play a role in resolving the CPP in these settings, but not necessarily the separability of the VMs in space under these conditions. Rather, here the predictability of the Target VMs location was the key cue.

Figure 3C plots subjects' Conversation Index for the same conditions as above, but at the High Distractor Density level. Notably, marmosets' behavior was similar despite increased interference from the Distractor VMs. As in the low Distractor Density scenes, marmosets exhibited overall higher calling to the Target VM relative to Baseline (1-way ANOVA (Spatial), total df = 4041, Spatial df = 3, $p < 0.0001$). However, comparison of individual conditions revealed that subjects' Conversation Index was significantly higher than Baseline only for the Fixed and Single-Location conditions (Tukey-Kramer corrected $p<0.0001$ for each). The Conversation Index for the Random-Location condition was significantly lower than the Fixed and Single conditions (Tukey-Kramer corrected $p = 0.0125$ and $p = 0.0069$). The broad similarity between the High and Low Distractor densities in this experiment suggest that the increase in acoustic interference by conspecifics did not significantly impair marmosets' ability to resolve the challenges of communicating in these scenes, as long as the Target VM vocalizations were broadcast from a predictable location. This suggests that focusing attention to a spatial position may help to offset the challenge of parsing the Target VM from the Distractors when acoustic interference was nearly constant. Given the notably slow periodicity of marmoset conversational exchanges, with inter-call intervals up to 10s in duration, leveraging this mechanism may be crucial to resolving the challenges of communicating in a cocktail party

The ability of marmosets to correctly identify the Target VM and selectively engage with them in conversational exchanges irrespective of Distractor Density was somewhat surprising. However, marmosets effectively engaging in conversational exchanges does not reveal the more detailed nuances of how they accomplished this feat. We next performed a series of analyses to

determine whether these facets of their vocal behavior differed between the test conditions. As shown in Figure 3D, subjects produced a lower ratio of 1 pulse calls at the High Distractor Density level 55.2% to 47.6%, while 2 and 3 pulse calls modestly increased (2 pulse +6.70%, 3+ pulse +0.980%); a pattern found to be statistically significant (Kruskal-Wallis test, df = 6453, $p < 0.0001$). Notably, the median phee call variant produced in these experiments changed from 1-pulse phee at the Low Distractor Density to 2-pulse phee calls at the High Distractor Density. Figure 3E further shows that there was a significant change in both the average duration of phee calls (+9.12%), and the 1 pulse phee calls (+9.98%), but not 2 or 3+ pulse phee calls, from Low to High Distractor Density (3-Way ANOVA (Spatial, Acoustic, Pulse count), total df = 6453, acoustic $p < 0.0001$ and acoustic*pulse count $p < 0.0001$). Finally, we next compared the latency that subjects responded to the Target VM within conversations at the two Distractor Density levels (Figure 3F). Analyses indicated that the distribution of this latency was significantly shorter at the higher acoustic interference level (307ms; Kruskal-Wallis test, df = 2714, $p = 0.0207$). Together these results indicate that subjects increased the median duration of their phee calls and decreased the latency to respond to Target VMs when communicating at higher Distractor Density. This change in vocal behavior strategy may have been necessary to maintain conversational exchanges when acoustic interference from conspecifics vocalization is nearly constant.

Experiment 2.

The acoustic structure of long-distance contact calls – including the marmoset phee call (Morrill et al., 2013) – has been selected over evolution to maximize signaling efficacy in noisy environments (Waser and Brown, 1986; Mitani and Stuht, 1998). To this end, a common characteristic of this class of vocalizations is the repetition of an acoustically similar pulse. Such acoustic redundancy is speculated to have evolved because it functions to offset the inherent decline in the acoustic content of the vocalizations as they travel over long distances (Waser and Waser, 1977). Marmoset phee calls are consistent with this trend, comprising a series of acoustically similar repeated pulses (Miller et al., 2010). In Experiment 1, all VMs emitted 2-pulse phee calls because this is the most common variant of the call, accounting for nearly 70% of phee calls produced by marmosets (Miller et al., 2010). Here we tested whether this signal design characteristic was beneficial to marmosets communicating in a Cocktail Party. We hypothesized that if redundancy in call structure was beneficial to marmosets, eliminating this characteristic of the call would result in increased difficulty maintaining conversational exchanges. We tested subjects in the same Environments as in Experiment 1 but used 1-pulse phee calls as the stimulus produced by each VM rather than the 2-pulse phee calls used in the previous experiment. Given that subjects already struggled to communicate in the Random-Location condition under less challenging conditions, we did not repeat this test condition here. Instead we tested subjects only in the Fixed-Location and Single-Location spatial configurations.

Results for the Low Distractor Density in Experiment 2 are shown in Figure 4A. Consistent with Experiment 1, the Conversation Index for the Distractor VMs was significantly lower than the Target VMs in Baseline, Fixed, and Single conditions (2-Way ANOVA VM, total df = 16579, VM df = 4, $p < 0.0000$, Tukey-Kramer corrected, $p < 0.0000$), while all the distractors amongst each other had no difference (Tukey-Kramer corrected, $p > 0.917$). A comparison of Conversation Index with the Target VM between the test conditions showed that subjects engaged in significantly more conversations in both the Single and Fixed conditions relative to Baseline (2-Way ANOVA Spatial*VM, total df =16579, Spatial*VM df = 8, $p < 0.0001$), with Fixed and Single source conditions having a mean Conversation Index significantly above Baseline (Tukey-Kramer corrected, $p <0.0001$ for both), respectively. In contrast to parallel results in Experiment 1, the Fixed and Single source conditions were statistically different from each other (Tukey-Kramer corrected, $p = 0.0034$). These results suggest that, although marmosets could identify the Target VM and maintain conversational exchanges in both conditions, the spatial separation between the various VMs in the Fixed-Location condition may have afforded some perceptual advantages

despite the spatial predictability when only hearing 1-pulse phees emitted by the VMs even at the Low Distractor Density level.

A comparison of Conversation Index across the test conditions at the High Distractor Density is shown in Figure 4B. Similar to previous conditions, there was no difference in Conversation Index between the Distractor VMs (Tukey-Kramer correction $p = 1.00$) but a significant difference in the Target VM conversation indexes for Baseline, Fixed-Location, and Single-Location to distractors (Tukey-Kramer correction $p < 0.0001$). We observed a significant interactive effect of spatial configuration (2-Way ANOVA Spatial x VM, total df = 16434, $p < 0.0001$), but here only the Fixed-Location was statistically significant from Baseline. The Fixed condition had significantly higher mean Conversation Index than Baseline (Tukey-Kramer corrected, $p < 0.0001$), while the Single source condition was below Baseline (Tukey-Kramer corrected, $p = 0.668$). These results suggest that when the cocktail party consists of near constant acoustic interference from conspecifics producing only 1-pulse phee calls, the advantages afforded by selectively attending to a predictable location in space alone was insufficient for marmosets to consistently identify and engage in conversational exchanges with the Target VM. Under these conditions, spatial separability of the callers was critical for effective communication.

We next analyzed how the challenges of communicating in these cocktail parties affected subjects' vocal behavior. Analyses revealed that like Experiment 1, marmosets systematically modified their communication behaviors. The pattern of changes, however, were notably different from what we observed in the previous experiment. Figure 4C shows that there was a significant change in the distribution of the number of pulses per call made by the subject (Kruskall-Wallis, df = 4424, $p < 0.0001$). Whereas here we observed a higher ratio of 1 pulse calls produced by the subjects in the High Distractor Density conditions, Experiment 1 had the opposite effect. The ratio of 1 pulse calls produced by subjects increased from 59.6% to 70.7%, the 2 pulse and 3+ pulse calls dropped (-11.1% and -0.06%). Again, in contrast to results seen in Experiment 1, Figure 4D shows this did not result in a significant overall change in the duration of calls produced by subjects; rather, the significant changes in duration was apparent when subject calls were broken down by the number of pulses (3-Way ANOVA (Spatial, Acoustic, Pulse count), total df = 4424, acoustic $p = 0.385$ and acoustic*pulse count $p < 0.0001$). The 1 pulse calls increased in duration from lower to higher by 10.6% (Tukey-Kramer corrected $p < 0.0001$). The 2 pulse and 3+ pulse calls did not change significantly from lower to higher at -1.17% and 7.04%, respectively (Tukey-Kramer corrected $p = 0.427$ and $p = 0.788$). We also observed a significant decrease in latency to respond to the Target VM at the High Distractor Density level relative to the lower level (Kruskall-Wallis test, df = 2119, $p = 0.0019$), similarly to Experiment 1 (Figure 4E). The median latency within the conversation decreased from 5.09 sec to 4.71 sec in Low to High Distractor Density conditions (-7.54% change or 384 msec faster response latency). One potential reason for the notable change to producing more 1-pulse phee calls in this experiment may be because of an unintended change in the acoustic scene statistics that emerged as a result of matching the amount of acoustic interference between the two experiments.

Emergent Acoustic Scene Dynamics Reveal Adaptive Changes in Vocal Behavior.

An unintended byproduct of our effort to control for acoustic interference across the two experiments was systematic changes to other dimensions of the acoustic scene statistics. Most notably was a systematic change in the timing and variability of the interval between the Distractor VM calls to achieve the desired Distractor Density when constructing these conversations. To explore their respective impact on marmoset vocal behavior, we limit our analyses only on the Fixed-Location and Single-Location conditions because the Random-Location was not performed in Experiment 2.

Figure 5A shows the distribution of the mean inter-call interval (ICI) against the calculated Distractor Density for each session within Fixed-Location and Single-Location (Low and High Distractor Density for both Experiment 1 and 2). Notably, significant negative correlations exist

between the two values for both Experiment 1 and 2 (rho = -0.796 & $p < 0.0000$, rho = -0.935 & $p < 0.0000$, respectively). The broad pattern revealed by these quantifications emerged because the shorter duration 1-pulse phee calls necessitated a shorter ICI between VM distractor pairs to ensure similar levels of Distractor Density across the experiments. This characterization formed the foundation for the subsequent statistical analyses aimed at explicating the relationship between the emergent scene structure and marmoset vocal behavior in these experiments.

We next applied a linear model to test how facets of marmoset vocal behavior covaried with dimensions of the acoustic scene. The following were input into the Linear Model – VM Pulse # (2-pulse:Expt 1, 1-pulse:Expt2), Low and High Distractor Density, and Fixed and Single conditions – for a total of 144 sessions. We also chose to include the calculated Distractor Density for each session along with the Distractor ICI. Given a strong positive correlation between Distractor ICI and standard deviation (rho = 0.931 and $p < 0.0001$), we took the coefficient of variance (COV, standard deviation divided by mean) as a way to encapsulate these two correlated factors while avoiding rank deficiency in any linear model (COV v Mean ICI, rho = -0.0956, $p = 0.254$. Figure 5B). This also gave an added benefit of enumerating the relative dispersion of the Distractor ICI. This analysis yielded six total predictor variables. The following 8 vocal behavior response variables were also inputted into the model: the mean duration of all calls, the duration of the 1-Pulse calls, Index of relative 1 and 2 pulse calls produced by subjects (Pulse Number Index), the Conversation Index, subjects mean latency to respond in a conversation, the number of subject calls produced, the number of conversations, and the mean length of those conversations.

We tested eight interactive linear models which included 22 terms (1 intercept, 6 linear predictor terms, and 15 pairs of distinct predictor terms). The statistical threshold for significant terms and models was corrected for multiple comparisons with Bonferroni correction based on 22*8=176 comparisons with a corrected P value threshold at 0.05/176 = 0.000284. Of these eight models, four models reached significance: duration of the subject 1-Pulse calls, Pulse Number Index, number of subject calls, and number of conversations (R^2 = 0.332, 0.396, 0.451, 0.489, adjusted R^2 = 0.216, 0.291, 0.357, 0.401). Of these four models, the duration of 1-Pulse subject calls did not have a significant term below the corrected threshold. Two significant terms were shared across the remaining three significant models. The Distractor ICI x COV Distractor ICI (which results in standard deviation Distractor ICI) for subject calls produced and number of conversations, and the mean distractor ICI x 1/2 Pulse VM Calls Condition for all three models. Figure 5C-G plots the five significant terms against the respective response variables in interaction effects plots. Each image plots the adjusted response function of the given response variables on the Y-axis against the values of the first predictor in the interactive term with the second predictor at fixed values (for categorical: all levels, and numeric: minimum, maximum, and average of minimum and maximum). Given that all five interactive terms have significant coefficients within their respective models, and that the slopes of the lines in all five plots are not parallel, there is significant interactive effect between the predictors for predicting the Pulse Number Index, the number of subject calls produced, and number of conversations.

Presenting subjects with VM calls comprising either 2 or 1 pulse phee calls – Experiments 1 and 2, respectively – resulted in opposite effects on the adjusted response variables. For Pulse Number Index (Figure 5C), Calls Produced (Figure 5D) and Conversation Count (Figure 5E), these behavioral metrics revealed a positive correlation with Distractor ICI in Experiment 1, but a negative relationship in Experiment 2. In other words, when hearing 2-pulse VM calls in Experiment 1, subjects were more likely to produce 2-pulse phees, produce more calls and engage in more conversations as the Distractor ICI increased in duration. By contrast, the opposite was true when hearing only 1-pulse phee calls in Experiment 2. This suggests that the strategy to optimize communication exchanges when the Distractor streams are heard at a certain interval is not static but is highly affected by the types of calls marmoset subjects heard in the cocktail party. In other words, vocal behavior did not simply change linearly as a function of the

Distractor VM inter-call interval, but that the optimal strategy was based on a combination of this feature and the phee variant they heard in the scene. It should be noted that the consistency in the change in the number of calls produced and conversations by subjects is notable because these need not necessarily be parallel. It suggests that the increased number of calls are specifically being committed to conversations rather than calls produced independent of these active communicative exchanges.

A further significant factor affecting marmoset vocal behavior in the linear model was COV Distractor ICI. Both the number calls produced (Figure 5F) and conversations (Figure 5G) showed a similar pattern relative to Distractor ICI. As the Distractor ICI increased, at low COV, the relationship of number of subject calls and conversations produced decreased. At the highest level of COV, the opposite relationship emerged with increasing calls produced and conversations (with a smaller relative change). This suggests that as the predictability of the Distractor ICI increased (high to low COV), shorter Distractor ICI were optimal for the subject to produce calls and engage in more conversations with the Target VM. Similarly, pertaining to the importance of spatial predictability for marmosets in Experiment 1, temporal predictability was advantageous for marmosets to navigate the complex acoustic scene and selectively engage with the Target VM.

Discussion.

Here we employed an innovative multi-speaker paradigm to construct a real-world cocktail party and explicate how a New World primate – common marmosets - resolve the acoustic and social challenges of these landscapes to engage in their natural conversational exchanges. We report that marmoset monkeys solved the CPP by complementing a suite of auditory mechanisms with active modification of their own behavior to optimize conversational efficacy. Evidence indicated that marmosets relied on facets of audition – including individual differences in the voices of the multiple marmoset callers and spatial separation of the callers (Darwin, 1997; Mcdermott, 2009; Bronkhorst, 2015) – but also employed attentional mechanisms for speaker stream segregation to enable their natural conversational exchanges. Marmosets were not, however, simply passive observers of the acoustic scene that grouped all non-conversational VM partners as a general masker. Rather, subjects actively adjusted their own vocal behavior in response to the dynamics of the immediate cocktail scene suggesting that these monkeys were not only attending to the Target VM but to the conversational exchanges occurring between the Distractor VMs as well. By leveraging the marmosets' natural vocal behavior to examine the CPP in an innovative experimental paradigm, our novel results highlight that resolving the challenges of these dynamic naturalistic acoustic scenes is not only the domain of audition in primates but is facilitated by adaptive modifications of individuals' own behavior to facilitate communication. These findings suggest that explicating the neural mechanisms supporting the CPP in primates – both human and nonhuman – may also need to consider how listeners actively adapt their own vocal behavior in response to the changing features of the acoustic scenes to optimize key auditory mechanisms.

The pattern of behavior in marmosets from Experiment 1 suggest that marmosets employed auditory attention to resolve the CPP. Experiments in humans involving multiple speakers found that when the spatial position of each talker randomly changed across locations, subjects' intelligibility scores decreased (Brungart and Simpson, 2007). Likewise, human subjects performed significantly better when the spatial location of the target was cued prior to hearing the sound (Kidd et al., 2005). In both cases, it was concluded that the predictability of a talker's position in space allowed subjects to focus attention to that position in space. When that predictability was eliminated, attention could not be focused, and it accordingly had a negative impact on subjects' capacity to understand what was spoken. Experiment 1 sought to test whether a nonhuman primate would exhibit a similar pattern of behavior under these conditions. In the

Random-Location condition, we randomized the spatial location across seven speaker locations each time one of the VM calls were broadcast. Importantly, the vocal behavior of the VMs (i.e. the acoustic scene) was identical across all three test conditions, and the only difference in the Random-Location condition was a lack of predictability for the location from which each phee call was broadcast from amongst the seven speakers. As a result, marmosets could not predict where in the scene the Target VM call would be broadcast at each stimulus broadcast. As shown in Figures 3B&C, subjects performed significantly worse under these test conditions than either of the other two spatial configurations. In fact, their conversational behavior was statistically indistinguishable from Baseline at both Distractor Density levels suggesting that a lower level of acoustic interference did not allow them to overcome the lack of predictability in the Target VMs spatial location. By contrast, marmosets readily engaged in conversational exchanges when VM calls were broadcast from a separate, but consistent locations like in Fixed-Location or where all VM calls broadcast from a single speaker like in Single-Location. While these results suggest that the spatial predictability, rather than the spatial separability, of the VM callers was key under these conditions, results from Experiment 2 (Figures 4A,B) that separation of speakers in space does play a role in parsing the Target VM under some conditions, as will be discussed later. The systematic decline in performance between marmosets and humans (Kidd et al., 2005; Brungart and Simpson, 2007) when a talker's spatial position could and could not be predicted was nearly identical and suggestive that similar attentional mechanisms were leveraged to resolve the CPP in both primate species.

Engaging in conversational exchanges in these cocktail parties likely relied on a schema-based learning mechanism for speaker stream segregation (Bregman, 1994; Bey and McAdams, 2002; Woods and McDermott, 2018). First, the identity of the Target VM needed to be learned in each session. While the spectro-temporal structure of marmoset phee calls is relatively stereotyped, lending itself to potential specializations for parsing the signal from the myriad of potential acoustic interference in the natural environment (Mcdermott, 2009), each monkey's phee is individually distinctive (Miller et al., 2010; Miller and Thomas, 2012). As a result, segregating one caller's phee call from amongst the phee calls of many conspecifics presents a different challenge that relies on learning the identity of a willing conversational partner. Although all calls broadcast from VMs were produced by animals in the UCSD colony who were familiar to subjects, no phee calls from subjects' cage-mates were used. As a result, familiarity was likely consistent across the VM callers (Johnsrude et al., 2013). Second, learning the identity of the Target VM was based on its distinctive vocal behavior. While subjects heard high number of calls from Distractor VMs in all conditions, only the Target VM calls were designed to be broadcast interactively with subjects. The timing of Target VM calls conformed to the statistics of natural marmoset conversational exchanges and was designed to broadcast in response to a subject's call as an interactive vocal exchange (Miller and Wang, 2006; Miller et al., 2009b; Toarmino et al., 2017). Therefore, marmosets learned the identity of the Target based on the statistical occurrence of this critical temporal cue in the VM's behavior relative to their own rather than anything intrinsic to the vocalizations themselves. Third, once the Target VM identity was learned, marmosets must then continuously monitor conspecifics' behavior and temporally coordinate their own behavior for conversations to occur. A series of previous studies has shown that these temporal cues are governed by social rules that differ based on the age, sex and relatedness of the individuals and are crucial for the coordination at the core of these vocal interactions (Miller and Wang, 2006; Miller et al., 2009b; Chow et al., 2015; Toarmino et al., 2017). Amongst these temporal dynamics, however, is the relatively slow periodicity of these conversations. Marmosets abide turn-taking in these conversations but while the interval between calls is on average ~3s, it can range up to 10s (Miller and Wang, 2006; Miller et al., 2009b). Because of the cacophony of marmoset phee calls broadcast in these experiments, particularly at the high Distractor Density level, focusing attention to a predictable spatial location would have been notably advantageous considering a latency of several seconds between the offset of the subjects call and the Target

VM response, during which time the calls of multiple Distractor VMs could have been broadcast. Notably, however, the advantages of attention and a schema-based learning to stream the Target VM did have its limits, as evidenced by the difficulties of marmosets to communicate in the Single-Source condition for Experiment 2 when hearing only 1-pulse phee calls produced by VMs.

Results from Experiment 2 contrasted with Experiment 1 in several important ways that may reveal an evolutionary relationship between vocal signal design and audition in marmosets. While marmosets performed similarly in the Fixed- or Single-Location test conditions when hearing 2-pulse phee calls from all VMs in the first Experiment (Figure 3C), this pattern did not replicate in Experiment 2. When hearing only 1-pulse phee calls from VMs, marmosets did indeed continue to engage in conversational exchanges with the Target VM in the Fixed-Location condition irrespective of the level of acoustic interference, but exhibited a significant decline in Conversation Index when all VM calls were broadcast from a single speaker in the Single-Location condition (Figure 4B). In fact, at the higher Distractor Density level, marmosets' conversations were statistically indistinguishable from Baseline in the Single-Location condition. In other words, under these conditions spatial-release from masking was necessary to identify the Target VM and maintain conversational exchanges (Litovsky, 2005; Jones and Litovsky, 2011; Pastore and Yost, 2017). An explanation for this pattern likely pertains to the selective pressures on the phee calls themselves that occurred over evolution to maximize signaling efficacy (Morrill et al., 2013). Nonhuman primate long-distance contact calls – including the phee – often comprise the repetition of a single syllable, a signal design structure conjectured to limit degradation of the signals communicative content when transmitting long distances through noise acoustic environments (Waser and Waser, 1977; Mitani and Stuht, 1998). By reducing the number of pulses in each call, we effectively halved the amount of acoustic information available to both identify the Target VM and recognize it in the cocktail party for subsequent potential interactions. Indeed, experiments with tamarin monkeys – a close phylogenetic cousin with marmosets – reported that reducing the number of pulses in their contact call significantly impaired their ability to recognize the caller's identity (Miller et al., 2005). While marmosets were able to overcome the challenge of identifying the Target VM from a phee call with only a single pulse when each VM called from a different – but consistent – position in space, subjects had difficulty (i.e. Low Distractor Density) or were unable to converse with the Target VM (i.e. High Distractor Density) when all calls were broadcast from a Single-Location. This suggests that under conditions with the highest acoustic interference, the redundancy of a two-pulse phee call is crucial to maintaining active conversational changes. Selection for multi-pulsed phee calls in marmoset evolution, and potentially more broadly for other nonhuman primates (Mitani and Stuht, 1998), may have been driven by the limits of audition for parsing vocalizations and recognizing callers amid the myriad of biotic and abiotic noise common in the species forest habitat.

Our discussion to this point has focused on the role of auditory mechanisms for marmosets. Yet results from the GLM analysis indicated that these primates did not solely rely on audition to effectively communicate in cocktail parties. Rather subjects vocal behavior changed in response to specific characteristics of the acoustic scene that emerged from the Distractor VM conversational dynamics. To control for acoustic interference, it was necessary to decrease the inter-call interval (ICI) between phees in the Distractor VM conversations. The emergent effect on the acoustic scene was a systematic change in the periodicity of the Distractor VM conversation i.e. variance and inter-call interval. The effect of these cocktail party characteristics on marmoset vocal behavior was considerable. When Distractor VM Conversations comprised 1-pulse calls and occurred at a faster call rate (i.e. decrease in Distractor ICI), marmoset had a propensity to produce more calls in more conversations while also producing more 1 pulse calls (Figure 5D,E). By contrast, when Distractor VM conversations was comprised of 2-pulse phee calls, these same measures of volubility increased for slower Distractor VM conversations (i.e. increase in Distractor ICI). Furthermore, marmoset behavior was significantly influenced by the predictability of the Distractor ICI, as subjects exhibited significant biases to produce more calls (Figure 5F) and

conversations (Figure 5G) when Distractor VM conversations were the least variable. These patterns are notable for several reasons. First, across the Linear Model, changes in the number of calls produced and conversations occur in parallel with Distractor VM ICI. This suggests that marmosets are not simply calling more but are specifically producing more calls within the context of active conversations. Third, interactive effects revealed by the model suggests that optimizing communication in these challenging acoustic and social landscapes necessitates a dynamic strategy, one in which marmosets must exert control over both the call structure and the behavior (Pomberger et al.; Miller et al., 2009a; Pomberger et al., 2019; Zhao et al., 2019). Consistent with previous experiments (Roy et al., 2011) and results in Experiment 1, the predictability of the scene played a significant role in how marmosets solve the CPP for effective communication. Finally, these patterns suggest that marmosets are not simply treating the Distractor VM calls as a broad masker; instead, they are attending to the dynamics of those conversations to maximize conversational efficacy. In other words, marmoset attention appears to be divided between the Target VM and Distractor VM vocal behavior to resolve the challenges of the CPP.

The challenges of communicating in cocktail parties is a daily occurrence for human and nonhuman primates. While results presented here indicate that marmosets leveraged the advantages of attention and schema-based learning to segregate the Target VM stream from the other callers in the scene – similarly to humans – these auditory mechanisms did have their limits. Marmosets complemented audition by actively adapting their own vocal behavior to optimize their ability to effectively engage in long conversations. Together, these results suggest that solving the cocktail party in real-world, multi-talker environments may be a far more dynamic, active process than is typically considered (Brungart and Simpson, 2007). Notably, the unique insights reported here were possible because of the innovative multi-speaker paradigm developed for these experiments to construct cocktail parties and systematically manipulate key properties of these social and acoustic landscapes. A notable, broader implication of these findings is the opportunity to leverage this exciting paradigm to investigate the neural basis of these perceptual and cognitive mechanisms in the primate brain. Neural recordings in human auditory cortex have highlighted the role of attention in representing speakers in complex acoustic scenes comprising multiple talkers (Mesgarani and Chang, 2012; Khalighinejad et al., 2019), but relatively little remains known about how other neural substrates in the auditory system contributes to the myriad of related processes. Marmosets share the core functional architecture of the auditory system with all other human and nonhuman primates (Kaas and Hackett, 2000; de la Mothe et al., 2006; Bendor and Wang, 2008; Hackett, 2009), and has been a key primate model of sound processing, including vocalizations, for many years (Wang and Walker, 2012; Wang, 2013; Miller et al., 2016; Song et al., 2016 ; Eliades and Miller, 2017). By integrating existing technologies for recording neural activity in freely-moving marmosets with the current behavioral paradigm, the potential to explicate the circuit level mechanisms in the primate brain that underlie the CPP can be realized.

Materials & Methods.

Subjects.
Six adult marmosets (3 females and 3 males) participated as subjects in these experiments from September 2019 to May 2020. All subjects were social housed in pair-bonded family units that comprised of two adults, and up to two generations of offspring. The UCSD Institutional Animal Care and Use Committee approved all experimental procedures.

Experimental Design.
All experiments were performed in a ~4 X 3 m Radio-Frequency Shielded testing room (ETS-Lindgren). Individual subjects were transported from their home cage in clear acrylic transport boxes to the experimental chamber and tested individually. Subjects were placed in an acrylic and plastic mesh test cage (32 X 18 X 46 cm) designed to allow the animals to climb and jump freely along the front wall of the cage similarly to previous experiments (Miller and Thomas, 2012; Toarmino et al., 2017) . The cage was placed on a rectangular table against the shorter side of the room. Seven speakers (Polk Audio TSi100, frequency range 40-22,000 Hz) were placed on the opposite side of the room arranged to maximize distance relative to all other speakers in both the horizontal and vertical planes (Figure 1A). All vocal stimuli were broadcast at 80 dbSPL as measured 0.5 m in front of the speaker. A cloth occluder divided the room to prevent the subjects from seeing any of the speakers during testing. One directional microphone (Sennheiser, model ME-66) was placed approximately 0.3 m in front of the subject to record all vocalizations produced during a test session. Another directional microphone was placed in front of the central speaker as well. We tested subjects three times to each test condition across two experiments while randomized. The order of each condition within the individual Experiments was counterbalanced across subjects in a block design for the High and Low Distractor Density levels.

Cocktail parties were constructed using an innovative multi-speaker paradigm in which vocalizations were broadcast from five, software generated Virtual Marmosets (VMs) (Figure 1A). The unique individual identity of each VM was determined by (1) broadcasting prerecorded vocalizations from an individual marmoset in the UCSD colony and (2) its vocal behavior relative to the live subject and other VMs. With respect to this later characteristic, VM vocal behavior was determined by their designation as a Target or Distractor. Similar to our previous experiments (Miller and Thomas, 2012; Toarmino et al., 2017), the behavior of Target VM was specifically designed to directly engage subjects in the species-typical natural conversational exchanges by utilizing an interactive playback design. To this end, the Target VM would broadcast a phee call response within 1-5s with an 85% probability each time subjects produced a phee call. In successive vocal exchanges between the subject and target (e.g. a conversational exchange), the Target VM would broadcast a response with 100% probability to maintain the vocal interaction. If subjects did not produce a call within 15-30s, the Target VM would broadcast a spontaneous call. Custom-designed software recorded vocal signals produced by the test subject from the directional microphone positioned in front of the animal and identified when subjects produced a phee call. By contrast, the timing of Distractor VM phee calls were independent of subjects' behavior, occurring at a predetermined interval. In each test condition, we generated two pairs of Distractor VMs. Each pair was designed to directly engaged each other in conversational exchanges. The timing of phee calls within these conversations was determined by the parameters of the test condition.

VM Stimulus Sets. All phee calls used as stimuli in these experiments were recorded from animals in the UCSD colony using standardized methods in the laboratory described in previous work (Miller and Thomas, 2012; Toarmino et al., 2017). Briefly, two monkeys were placed in separate testing boxes positioned ~3m from each other with an opaque cloth occluder located

equidistant between the boxes to eliminate visual contact between the animals. Directional microphones (Sennheiser ME-66) were placed directly in front of each subject to record vocal output separately from each animal. Naturally produced calls were recorded direct to disk over a 30min session. At the conclusion of the session, custom-designed software was used to extract two-pulse phee calls produced during each session. Phee calls produced within 10s of a conspecific phee were classified as 'antiphonal' responses, while those produced after this threshold were classified as 'spontaneous' phee calls. These designations were based on previous research (Miller et al., 2009b). Each VM in a test session would only broadcast antiphonal and spontaneous phee calls from a single marmoset. The stimulus sets used as the basis for each Target and Distractor VM was randomized across test sessions. The VMs stimulus sets used to construct each cocktail party were never produced by animals in a subject's home cage because of confounds that might occur due to social relatedness (Miller and Wang, 2006).

Test Conditions.
We selectively manipulated two dimensions of the scenes - *Spatial Configuration* & *Distractor Density* - to directly test their respective impact on how marmosets resolved the challenges of communicating in a cocktail party in two separate experiments distinguished only by the phee call variant broadcast to subjects. Experiment 1 tested subjects using two-pulse phee calls as vocalization stimuli produced by VM, while Experiment 2 broadcast only 1-pulse phee calls from the VMs.

Spatial Configuration. The spatial location of the VMs was manipulated by broadcasting the phee stimuli in three different speaker configurations: *Fixed-Location*, *Random-Location* and *Single-Location* (Figure 3A). These configurations allowed us to contrast the effects of both the significance of spatial separation between the callers and the predictability of a caller's position in space on marmoset vocal behavior.

Fixed-Location: In this configuration, the calls of each VM were broadcast from among five distinct, spatially separated speakers. This scene afforded subjects spatial separability of each VM from a consistent spatial location for the duration of the experiment.

Random-Location: Like the Fixed-Source condition, VM calls were broadcast from distinct spatially separated speakers. Rather than each VM broadcast from their own speaker for the duration of the experiment, speaker location was randomized across all 7 potential speakers during each broadcast. No VM call would be broadcast from the same speaker twice in a row, nor was there any overlap in VM calls from the same speaker. As a result, subjects were afforded spatial segregation of the VMs, but with no predictability for where the VM would emit a call.

Single-Location: Here all VM stimuli were broadcast from a single speaker, thereby eliminating spatial separation of the different callers.

Distractor Density. Distractor density was manipulated to two levels – Low and High – by changing the relative inter-call interval between phees broadcast between VM Distractor pairs. In the 'Low' distractor density scene, Distractor VM conversations had an inter-VM call interval ranging 1 to 3.5 sec in Experiment 1 [2-pulse phee calls] and 1 to 2.5 sec in Experiment 2 [1-pulse phee calls]. In the 'High' distractor density scenes, Distractor VM conversations had an inter-VM call interval ranging from 0.5 to 1.0 sec in Experiment 1 [two-pulse phee calls] and 0.5-0.75 sec in Experiment 1 [one-pulse phee calls]. The shorter inter-VM call interval ranges for Experiment 2 were used to maintain the same level of Distractor Density when the shorter one-pulsed phee calls were used as stimuli.

Baseline Condition. To establish Baseline vocal behavior under these conditions, subjects were tested in the Fixed-Location using the same parameters as under with the only exception being the Target VM 'stimulus'. During Fixed-Location sessions, an audible vocalization was broadcast from the Target VM as described above, while during the Baseline sessions no audible stimulus was broadcast to subjects. The timing of the stimulus event was recorded, however, to determine the probability that Target VM and subject would, through the natural periodicity of marmoset vocal behavior that emerged in these scenes, occur in a temporal sequence consistent of conversational exchanges. By comparing subjects' behavior under between Fixed-Location and Baseline conditions, it would be possible to ascertain which properties of the conversation were indicative of active communication. Notably, because the Target VM response latency in both cases was relatively quick for these exchanges, this Baseline inherently biased the likelihood that putative 'conversations' would emerge because marmosets exhibited relatively high levels of volubility in these cocktail parties relative to previous experiments with considerably fewer VM call presented in both single and multi-speaker paradigms (Miller and Thomas, 2012; Toarmino et al., 2017). We pursued this course, however, because we sought to identify the characteristics of the marmoset conversations that were most reflective of active conversations. A Baseline condition was performed separately for High and Low Distractor Density levels in Experiments 1 and 2.

Data Analysis.

We calculated the following behavioral metrics to quantify changes in subject vocal behavior relative to the Target and Distractor VMs as well as standard acoustic parameters, such as call duration and response latency.

Conversation Index. This behavioral metric was developed to quantify the relationship phee calls produced in conversations weighted by its length relative to all phees produced in a session by the subject (Figure 2B).

Previous experiments in marmosets determined that phee calls produced within 10s following a conspecific phee call were perceived as a 'response' to the initial call by conspecifics and were significantly more likely to elicit a subsequent vocal response, while those produced after this threshold did not elicit vocal responses from conspecifics (Miller et al., 2009b). We defined a conversation as each behavioral epochs in which monkeys engage in a series of alternating, reciprocal phee exchanges during which the inter-call interval between phee calls is within the 10s threshold. Each conversation ended when the subject did not respond for more than ten seconds. We elected to use conversations as our key behavioral metric in these experiments because they are indicative of learned communication behavior that requires a coordinated, interactive effort between marmosets (Chow et al., 2015; Toarmino et al., 2017).

For each test session, the temporal relationship between subjects calls and each of the five VMs was measured to determine whether the subject and a VM engaged in a conversational exchange. To quantify the occurrence of conversations in each test session, we first identified all instances in which the timing of subjects' phee calls and each VM conformed to these parameters using custom software. Subjects calls in conversational exchanges were assigned a number based on their linear order in the vocal exchanges sequence. In other words, the first response was assigned 1, the second successive response was assigned 2, etc. Standalone subject calls and the initiation of a conversational exchanges by subjects were assigned 0. By taking the average of all these calls across an experimental condition for each Subject-VM pair we can calculate their respective Conversation Index (Figure 2B). This metric allowed us to compare the occurrence of conversations between the subject and each class of VM – Target and Distractors – as well as across test conditions.

Interference Ratio. We measured the temporal overlap between the Distractor VMs calls and the Target VM calls to determine the amount of acoustic interference that occurred. Each time a

Target VM call was broadcast, we measured the duration of time it temporally co-occurred with any Distractor VM call. The resultant ratio indicates the percentage of overlap in time between Target and Distractor VM calls.

Pulse-Number Index. Custom software extracted all phee calls produced by subjects in each test session and identified the number of pulses within these calls based on previously identified stereotyped spectro-temporal structure of these vocalizations (Miller et al., 2010). Once cataloged, we then compared the number phee calls produced that comprised 1, 2 or 3+ pulses. Previous studies have shown that the majority of marmoset phee calls consist of 2-pulses (~70%), while the other variants occur at lower frequency (Miller et al., 2010). Phee calls consisting of 3 or more pulse calls were rarely produced in the current experiments, accounting for <10% of calls, these were grouped together. Because the number of phee calls comprising 3+ pulses did not vary across the test conditions, these were excluded from this this metric. We generated the Pulse-Number Index by calculating the difference over the sum of the 1 and 2 pulsed phee calls produced in each session [$(1PulseRatio - 2PulseRatio)/(1PulseRatio + 2PulseRatio)$]. Positive values would indicated a bias towards 1-Pulse Phee calls, while a negative value reflected a bias towards 2-pulse Phee Calls.

Linear Model Analysis.

Response Variables. These metrics were used as response variables within our linear models as mentioned in the Results section. Each one was calculated for each recorded session within a given experimental condition (18 per condition).

- *Average Duration of Calls:* The mean duration of subject calls.
- *Duration of 1 Pulse Calls*: Mean duration of 1-pulse calls produced by the subject
- *Pulse-Number Index*: The difference over sum of the ratio of one pulse calls to two pulse calls produced by the subject.
- *Conversation Index:* The mean position of the subject calls as previously mentioned.
- *Response Latency in Conversation*: The mean latency of subjects to respond to Target VM within a conversational exchange.
- *Number of Calls*: Number of calls produced by the subject in a given session.
- *Number of Conversations*: The number of times the subject engaged in conversational exchanges.
- *Length of Conversations*: The mean number of subject calls produced within each conversation.

Design. MATLAB function 'fitlm' was used to fit six predictor variables to each of the 8 response variables thus creating 8 linear models of comparison on 144 observations per model. The six predictor variables were: the calculated Distractor Density (as seen in Figure 3B,C and Figure 4A,B), Distractor ICI, COV Distractor ICI, the categorical Distractor Density (Low or High), the categorical spatial configuration (Fixed or Single), and the categorical Experiment (2-Pulse or 1-Pulse). An interactive linear model was created that included an intercept term (1), linear term for each predictor (6), and products of pairs of distinct predictors excluding squared terms (15), for a total of 22 predictor terms. The 8 models created with 22 predictor terms were corrected for multiple comparisons using the Bonferroni correction. With a criterion at $\alpha = 0.05$, the new p-value threshold was calculated to be at 0.05/176 = 0.000284. Any model's F-test for a degenerate constant model that was below this threshold was included for further analysis of the terms. Four models reached this threshold as mentioned in the results. Of those four, only three had terms with coefficients that were significantly different from 0 below the corrected new threshold and were subsequently explored in Figure 5C-G.

Acknowledgements.

We thank Victoria Ngo and Madeline Gagne for assistance in data collection and Drs. Yi Zhou and Vatsun Sadagopan for comments and discussion on an earlier draft of this manuscript. This work supported by grants from NIH (R01 DC012087) and DARPA (SSC-5029) to CTM.

Literature Cited.

Bee MA, Micheyl C (2008) The cocktail party problem: What is it? How can it be solved? And why should animal behaviorists study it? J Comp Psych 122:235-251.

Bendor DA, Wang X (2008) Neural response properties of the primary, rostral and rostrotemporal core fields in the auditory cortex of marmoset monkeys. J Neurophys 100:888-906.

Bey C, McAdams S (2002) Schema-based processing in auditory scene analysis. Perception & Psychophysics 64:844-854.

Bregman AS (1994) Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, MA: MIT Press.

Bronkhorst AW (2000) The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. Acustica 86:117-128.

Bronkhorst AW (2015) The cocktail-party problem revisited: early processing and selection of multi-talker speech. Attention, Perception, & Psychophysics 77:1465-1487.

Brungart DS, Simpson BD (2002) The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal. J Acoust Soc Am 112:664-676.

Brungart DS, Simpson BD (2007) Cocktail party listening in a dynamic multitalker environment. Perception & Psychophysics 69:79-91.

Cherry EC (1953) Some experiments on the recognition of speech, with one and with two ears. J Acoust Soc Am 25:975-979.

Chow C, Mitchell J, Miller CT (2015) Vocal turn-taking in a nonhuman primate is learned during ontogeny. Proceedings of the Royal Society, B 282:210150069.

Darwin CJ (1997) Auditory grouping. Trends in Cognitive Sciences 1:327-333.

Darwin CJ (2008) Listening to speech in the presence of other sounds. Philosophical Transactions of the Royal Society B: Biological Sciences 363:1011-1021.

Darwin CJ, Hukin RW (2000) Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention. The Journal of the Acoustical Society of America 108:335-342.

de la Mothe LA, Blumell S, Kajikawa Y, Hackett TA (2006) Cortical connections of the auditory cortex in marmoset monkeys: core and medial belt regions. J Comp Neurol 496:27-71.

Eliades SJ, Miller CT (2017) Marmoset vocal communication: Neurobiology and behavior. Developmental Neurobiology 77:286-299.

Hackett TA (2009) The evolution of primate and human auditory system. In: Evolutionary Neuroscience (Kaas JH, ed), pp 893-903. San Diego, CA: Academic Press.

Hill KT, Miller LM (2009) Auditory Attentional Control and Selection during Cocktail Party Listening. Cerebral Cortex 20:583-590.

Johnsrude IS, Mackey A, Hakyemez H, Alexander E, Trang HP, Carlyon RP (2013) Swinging at a Cocktail Party:Voice Familiarity Aids Speech Perception in the Presence of a Competing Voice. Psych Sci 24:1995-2004.

Jones GL, Litovsky RY (2011) A cocktail party model of spatial release from masking by both noise and speech interferers. The Journal of the Acoustical Society of America 130:1463-1474.

Kaas JH (2010) Sensory and motor systems in primates. In: Primate Neuroethology (Platt M, Ghazanfar AA, eds), pp 177-200. New York, NY: Oxford University Press.

Kaas JH, Hackett TA (1998) Subdivisions of auditory cortex and levels of processing in primates. Audiology and Neuro-Otology 3:73-85.

Kaas JH, Hackett TA (2000) Subdivisions of auditory cortex and procesing streams in primates. PNAS 97:11793-11799.

Khalighinejad B, Herrero JL, Mehta AD, Mesgarani N (2019) Adaptation of the human auditory cortex to changing background noise. Nature Communications 10:2509.

Kidd G, Arbogast TL, Mason CR, Gallun FJ (2005) The advantage of knowing where to listen. The Journal of the Acoustical Society of America 118:3804-3815.
Levinson SC (2016) Turn-taking in human communication: Origins and impliations for language processing. . Trends in Cognitive Science 20:6-14.
Litovsky RY (2005) Speech intelligibility and spatial release from masking in young children. The Journal of the Acoustical Society of America 117:3091-3099.
Mcdermott JH (2009) The cocktail party problem. Current Biology 19:R1024-R1027.
Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485:233-236.
Micheyl C, Tian B, Carlyon RP, Rauschecker JP (2005) Perceptual organization of tone sequences in the auditory cortex of awake Macaques. Neuron 48:139-148.
Micheyl C, Carlyon RP, Shtyrov Y, Hauk O, Dodson T, Pullvermuller F (2003) The neurophysiological basis of the auditory continuity illuion: A mismatch negativity study. J Cog Neurosci 15:747-758.
Miller CT, Wang X (2006) Sensory-motor interactions modulate a primate vocal behavior: antiphonal calling in common marmosets. Journal of Comparative Physiology A 192:27-38.
Miller CT, Thomas AW (2012) Individual recognition during bouts of antiphonal calling in common marmosets. Journal of Comparative Physiology A 198:337-346.
Miller CT, Dibble E, Hauser MD (2001) Amodal completion of acoustic signals by a nonhuman primate. Nature Neuroscience 4:783-784.
Miller CT, Iguina C, Hauser MD (2005) Processing vocal signals for recognition during antiphonal calling. Anim Behav 69:1387-1398.
Miller CT, Eliades SJ, Wang X (2009a) Motor-planning for vocal production in common marmosets Anim Behav 78:1195-1203.
Miller CT, Mandel K, Wang X (2010) The communicative content of the common marmoset phee call during antiphonal calling. . Am J Primatol 72:974-980.
Miller CT, Beck K, Meade B, Wang X (2009b) Antiphonal call timing in marmosets is behaviorally significant: Interactive playback experiments. . Journal of Comparative Physiology A 195:783-789.
Miller CT, Freiwald W, Leopold DA, Mitchell JF, Silva AC, Wang X (2016) Marmosets: A Neuroscientific Model of Human Social Behavior. Neuron 90:219-233.
Mitani J, Stuht J (1998) The evolution of nonhuman primate loud calls: acoustic adaptation for long-distance transmission. Primates 39:171-182.
Morrill R, Thomas AW, Schiel N, Souto A, Miller CT (2013) The effect of habitat acousics on common marmoset vocal signal transmission. Am J Primatol:904-916.
Pastore MT, Yost WA (2017) Spatial Release from Masking with a Moving Target. Frontiers in Psychology 8.
Petkov CI, O'Connor KN, Sutter ML (2003) Illusory sound perception in macaque monkeys. J Neurosci 23:9155-9161.
Pomberger T, Löschner J, Hage SR Compensatory mechanisms affect sensorimotor integration during ongoing vocal motor acts in marmoset monkeys. European Journal of Neuroscience n/a.
Pomberger T, Risueno-Segovia C, Gultekin YB, Dohmen D, Hage SR (2019) Cognitive control of complex motor behavior in marmoset monkeys. Nature Communications 10:3796.
Pressnitzer D, Sayles M, Micheyl C, Winter IM (2008) Perceptual Organization of Sound Begins in the Auditory Periphery. Current Biology 18:1124-1128.
Roy S, Miller CT, Gottsch D, Wang X (2011) Vocal control by the common marmoset in the presence of interfering noise. J Exp Biol 214:3619-3629.
Shinn-Cunningham BG (2008) Object-based auditory and visual attention. Trends in Cognitive Sciences 12:182-186.

Song X, Osmanski MS, Guo Y, Wang X (2016) Complex pitch perception mechanisms are shared by humans and a New World monkey. PNAS 113:781-786.

Stivers T, Enfield N, Brown PF, Englert C, Hayashi M, Heinemann T, Hoymann G, Rossano F, de Ruiter J, Yoon K, Levinson S (2009) Universals and cultural variation in turn-taking in coversation. PNAS 106:10587-10592.

Toarmino C, Wong L, Miller CT (2017) Audience affects decision-making in a marmoset communication network. Biology Letters 13:20160934.

Wang X (2013) The harmonic organization of auditory cortex. Frontiers in Neuroscience 7:2013.00114.

Wang X, Walker KMM (2012) Neural mechanisms for the abstraction of pitch information in auditory cortex. J Neurosci 32:13339-13342.

Waser PM, Waser MS (1977) Experimental studies of primate vocalization - specializations for long-distance propagation. Zeit Tierpsychol 43:239-263.

Waser PM, Brown CH (1986) Habitat acoustics and primate communication. Am J Primatol 10:135-154.

Wood NL, Cowan N (1995) The cocktail party phenomenon revisited: Attention and memory in the classic selective listening procedure of Cherry (1953). Journal of Experimental Psychology: General 124:243-262.

Woods KJP, McDermott JH (2018) Schema learning for the cocktail party problem. PNAS 115:E3313.

Zhao L, Rad BB, Wang X (2019) Long-lasting vocal plasticity in adult marmoset monkeys. Proceedings of the Royal Society B: Biological Sciences 286:20190817.

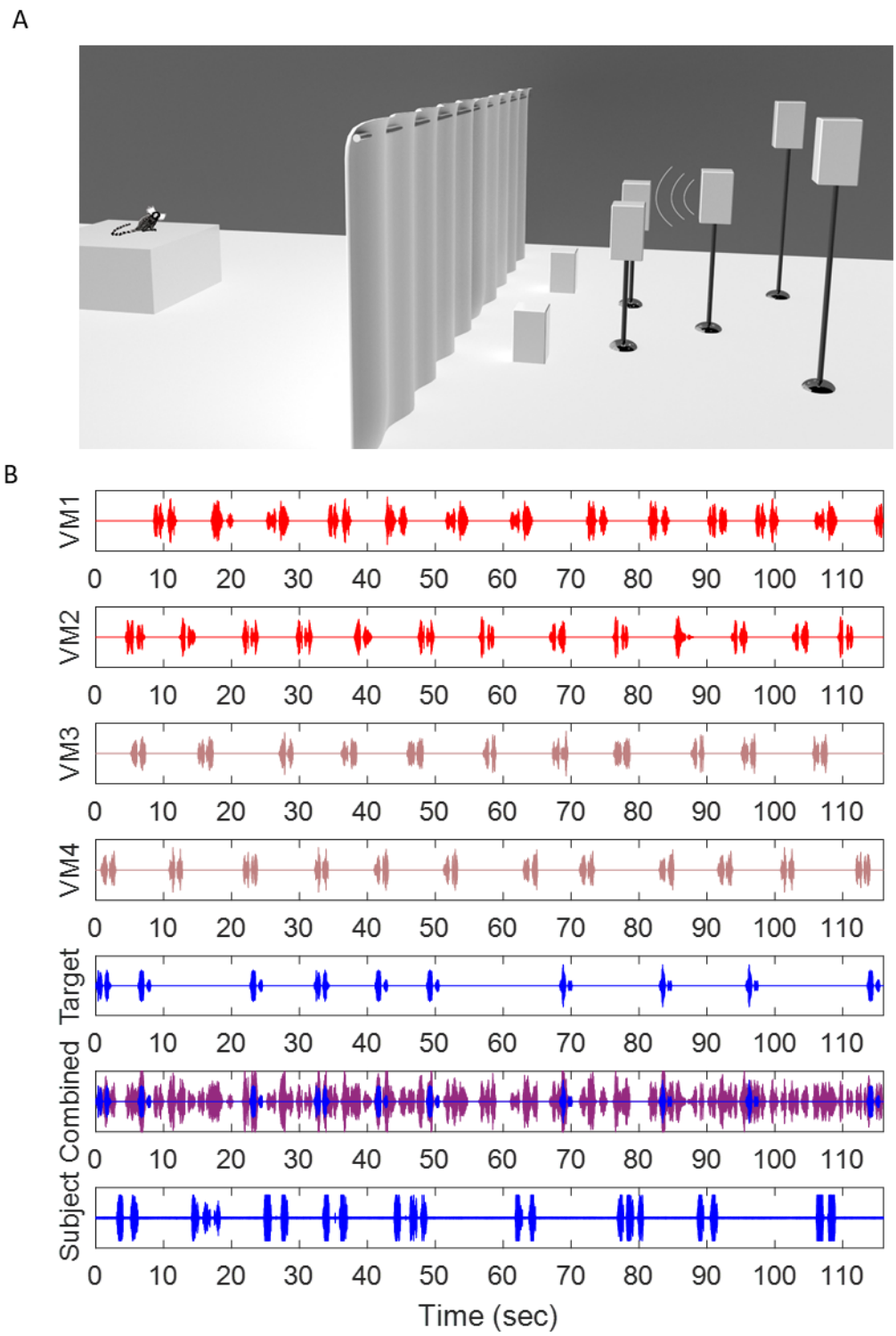

Figure 1. Design of the marmoset Cocktail Party experiments. (A) Schematic drawing of the spatial configuration of the testing room. Subjects were placed in a clear acrylic box with a mesh front (box around subject not pictured). Seven speakers were positioned to have spatial separation in height, distance and width. An opaque curtain was placed equidistant between the subject and speakers to occlude visual access. (B) An exemplar two-minute sample of the

vocalizations broadcast by the Virtual Monkeys (VM) and a live marmoset subject from a High Distractor Density, Fixed-Source session in Experiment 1. VM 1-4 are Distractors. VM1 and VM2 (shown in red) have been designed to broadcast 2-pulse phee calls that reflect a conversation with each other, while VM3 and VM4 (shown in brown) are likewise designed to engage in a reciprocal conversational exchange. The Target VM (blue) is engaged with the live marmoset Subject in an interactive reciprocal exchange based on subjects' vocal behavior. The combined view shows the summation of all VM phee calls – Distractors (purple) and Target (blue).

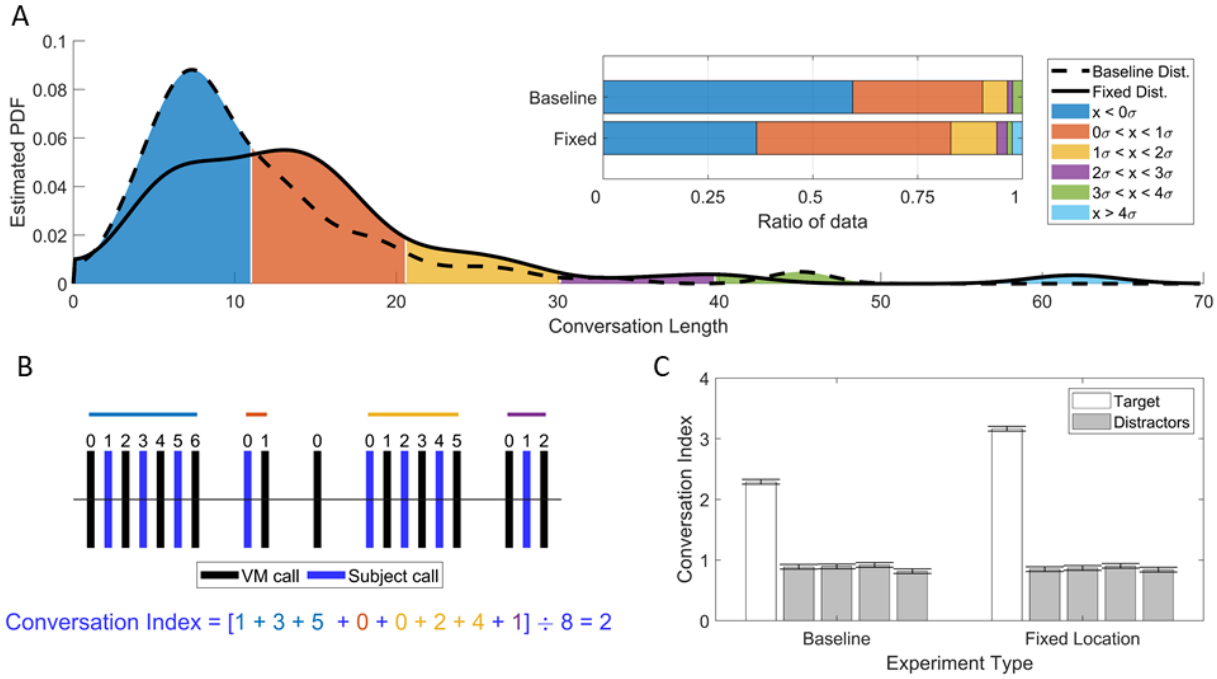

Figure 2. Conversation lengths to Conversation Index. (A) The estimated probability density of the Baseline conditions and Fixed conditions in the 95th percentile and above of conversation lengths. Shaded areas represent the boundaries of the combined distribution of Baseline and Fixed. The inset shows how much of the actual data within Baseline and Fixed fit within those described boundaries. 0σ represents the median of that distribution while 1-4σ represents the standard deviations above that median. (B) Schematic of an example recording session. Blue bars represent calls by the subject and black by a given VM. The numbers that would be assigned to each bar is written above. Bars next to each other without a gap are considered to be within the 10 second antiphonal response window and thus are counted consecutively. Horizontal colored lines above the bars indicate conversations. The average of all subjects' calls determines the Conversation Index. (C) Grouped bar graph showing the subject conversation index as it changes across the condition (Baseline versus Fixed) and VM (Target vs the 4 distractors). Error bars represent the 95% Confidence Interval.

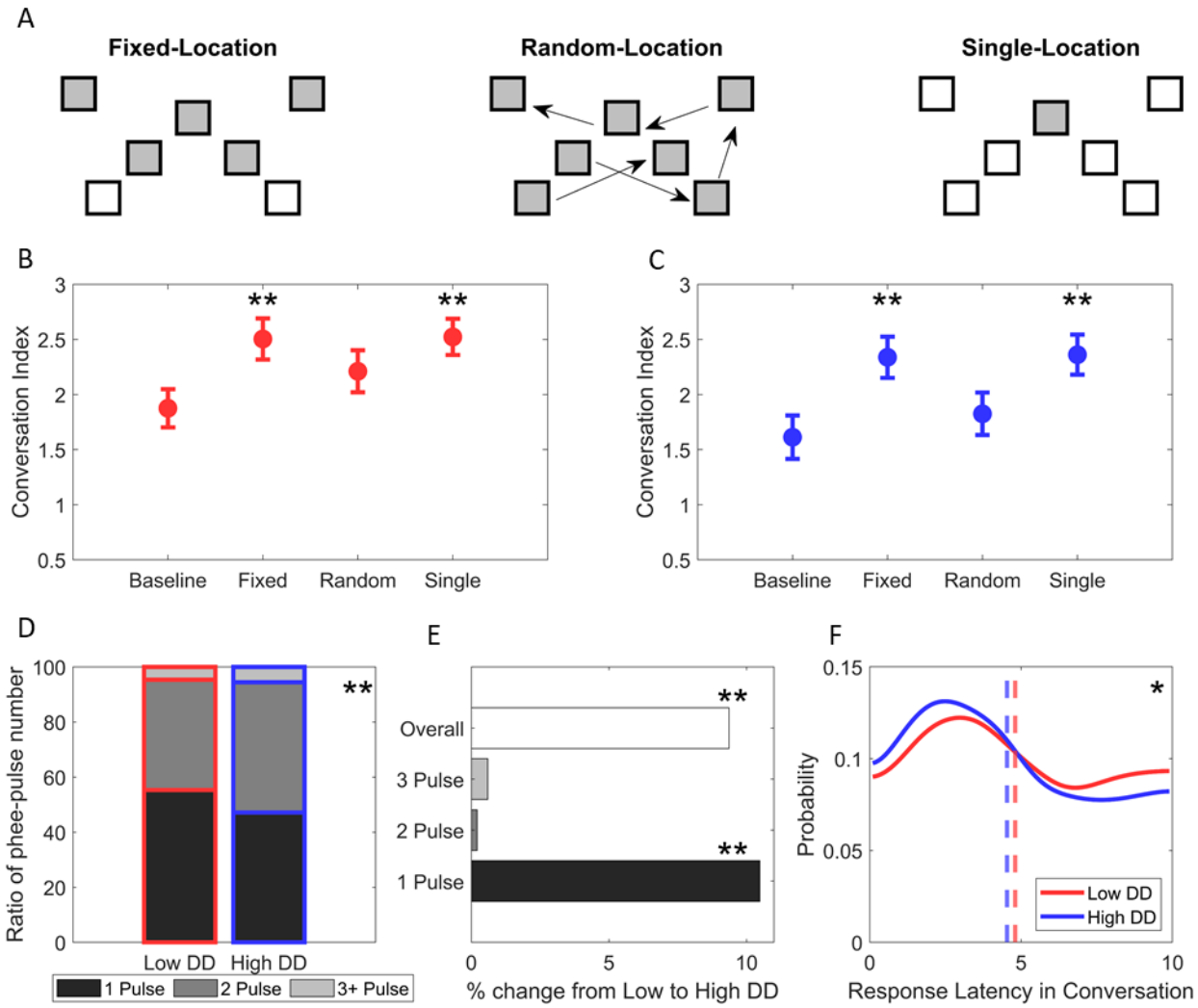

Figure 3. Experiment 1 Results. (A) A schematic drawing of the spatial configuration of the seven speakers used in these three test conditions from above: Fixed-Location, Random-Location, and Single-Location. Grey shading indicates which speakers broadcast phee calls for that condition. Arrows in the Random-Location condition indicate the fact that the speaker location from which each VM phee was broadcast was randomized for each stimulus presentation across the seven-speakers. (B, C) Plots the Mean Conversation index [95% CI] for Baseline, Fixed-Location, Random-Location, and Single-Location test conditions. ** Significant difference of condition from Baseline, $p < 0.0001$. (B) Plots Conversation Index for the Low Distractor Density condition, while (C) plots the High Distractor Density condition. (D) Stacked bar graph showing the distribution phee calls produced by subjects that comprised 1-Pulse (black), 2-Pulses (dark-grey) and 3 or more pulses (light-grey) in both the Low DD and High DD environments. ** Significant difference between distributions, $p < 0.0001$ (E) The change in duration of all calls, and sub-groups of phee-pulse calls from Low to High Distractor Density (DD) is shown as percent change. ** Significant difference for that category, $p < 0.0001$. (F) Probability density estimate plots of subjects' latency to respond to the Target VM in conversations in both Low DD (red) and High DD (blue) conditions. The median value is shown as a dashed vertical red bar – Low DD – and blue bar – High DD. * Significant difference between distributions, $p < 0.05$.

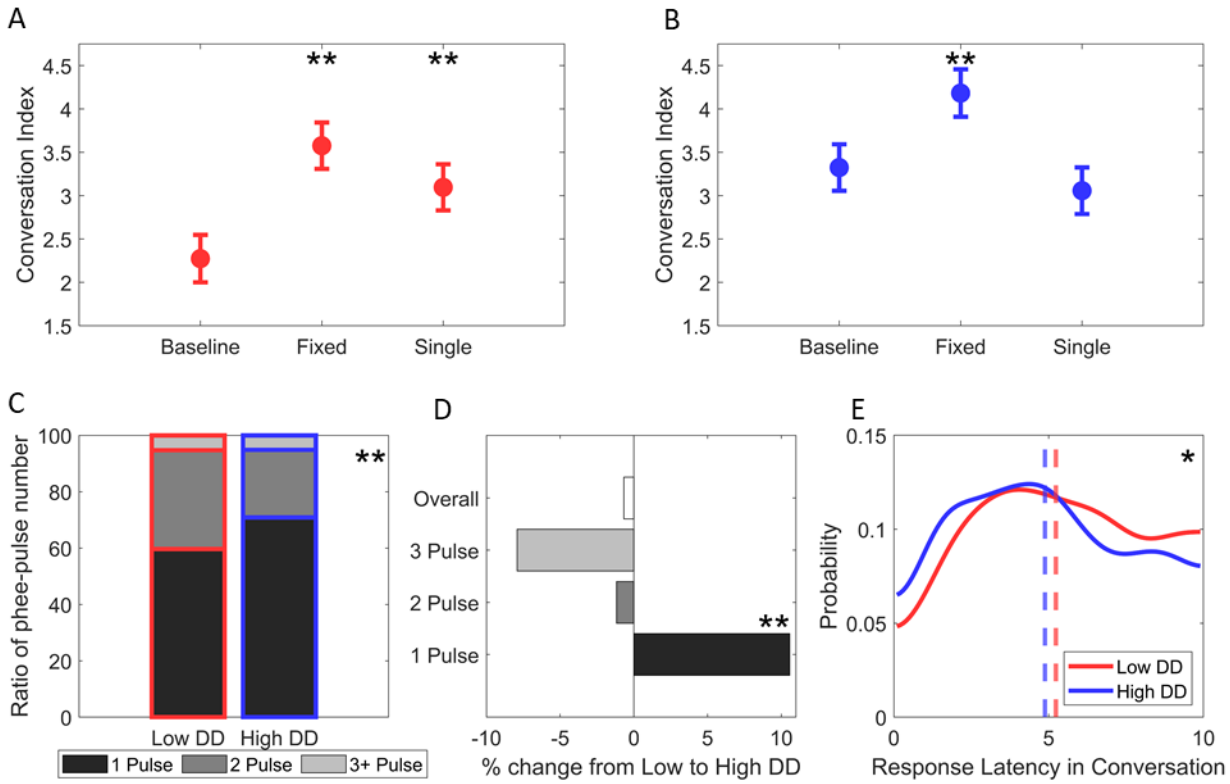

Figure 4. Experiment 2 Results. (A, B) Plots the Mean Conversation index [95% CI] for Baseline, Fixed-Location and Single-Location test conditions. ** Significant difference of condition from Baseline, $p < 0.0001$. (A) Plots Conversation Index for the Low Distractor Density condition in red, while (B) plots the High Distractor Density condition in blue. (C) Stacked bar graph showing the distribution phee calls produced by subjects that comprised 1-Pulse (black), 2-Pulses (dark-grey) and 3 or more pulses (light-grey) in both the Low and High Distractor Density (DD) environments. ** Significant difference between distributions, $p < 0.0001$ (D) The change in duration of the phee calls comprising 1, 2, 3 and Overall duration is shown as percent change from Low DD to High DD conditions. ** Significant difference for that category, $p < 0.0001$ (E) Probability density estimate plots subjects' latency to respond to the Target VM in conversations in both Low DD (red) and High DD (blue) conditions. The median value is shown as a dashed, vertical red bar – Low DD – and blue bar – High DD. * Significant difference for that category, $p < 0.01$.

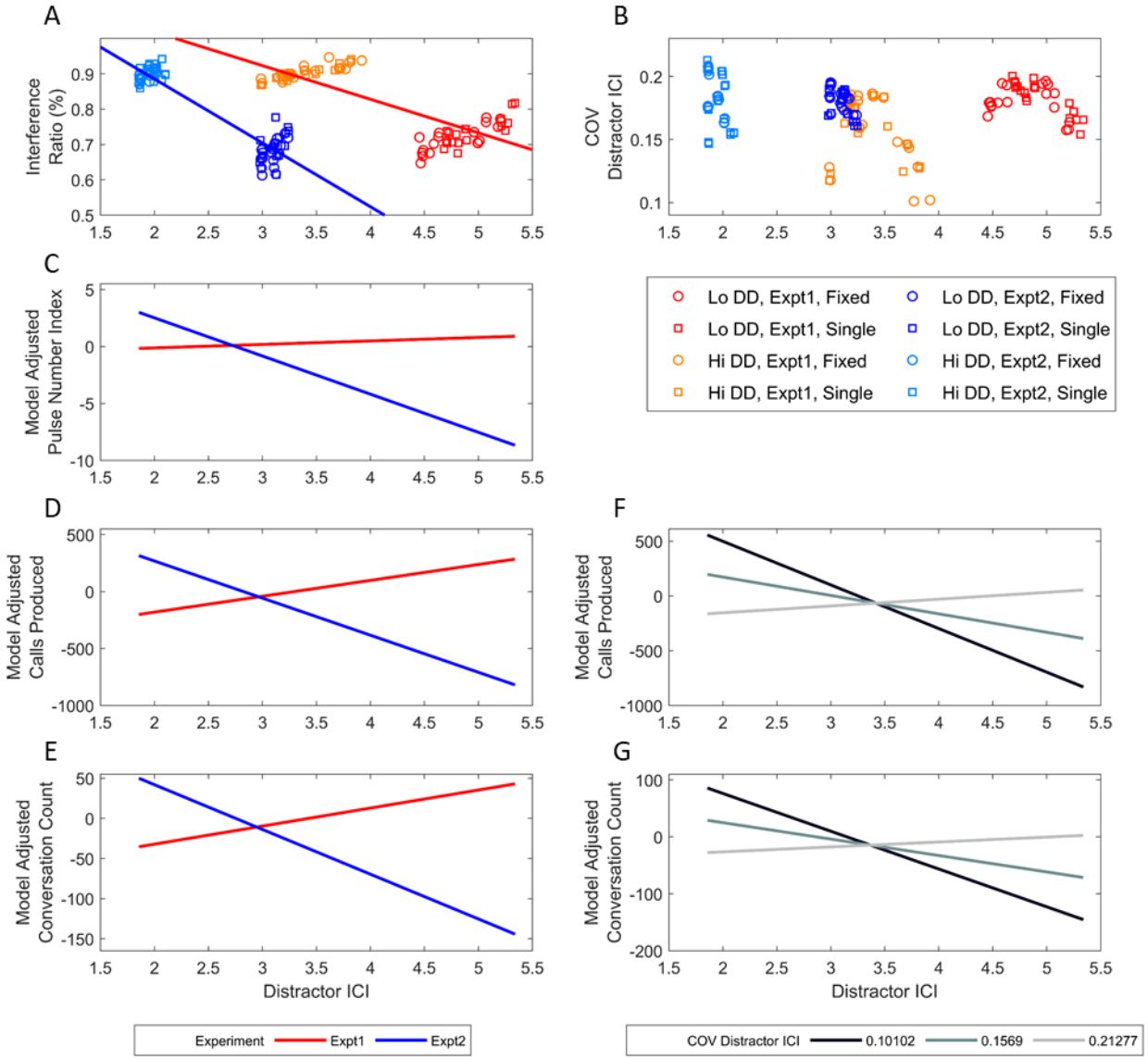

Figure 5. Linear Model Outcome (A) Scatter plot displaying Interference Ratio for the Distractor ICI measured during in each test session. Lines represent the least-squares fit for each Experiment. (B) Plots the COV Distractor ICI for the Distractor ICI measured during in each test session. Figure legend for (A & B) is shown below (B). (C-E) Significant interactive effects of Distractor ICI with different metrics of vocal behavior revealed by the linear model are shown. Results of the model from Experiment 1: 2-pulse VM phee calls (red line) and Experiment 2: 1-pulse VM phee calls are shown (blue line). The adjusted response value accounts for the average values of all other terms except Distractor ICI x Experiment within the linear model. (C) Plots Distractor ICI by the adjusted response variable of Pulse Number Index. Pulse Number Index refers to the difference over sum of the portion of 1-pulse subject calls to 2-pulse subject calls. The more positive a value the higher the portion of 1-pulse phee calls subjects produced, while more negative values indicate a bias towards subjects producing 2-pulse phee calls. (D) Plots the relationship between the model adjusted calls produced by subjects by the Distractor ICI. (E) Plots the relationship between model adjusted Conversation Count (i.e. the number of at least two calls produced by a subject in succession with the target) by the Distractor ICI. (F,G) Distractor ICI x

COV Distractor ICI term is plotted against its effect on the number Calls Subjects Produced (F) and Conversation Count (G). COV values plotted include minimum (light grey), maximum (dark grey), and the average of the two (mid-grey).