Effects of virtual acoustics on target-word identification performance in multi-talker environments

Atul Rungta
University of North Carolina, Chapel Hill
rungta@cs.unc.edu

Nicholas Rewkowski
University of North Carolina, Chapel Hill
nr@unc.edu

Carl Schissler
Oculus & Facebook
carl.schissler@oculus.com

Philip Robinson
Oculus & Facebook
philip.robinson@oculus.com

Ravish Mehra
Oculus & Facebook
ravish.mehra@oculus.com

Dinesh Manocha
University of Maryland, College Park
dm@cs.umd.com

http://gamma.cs.unc.edu/ckp/

Figure 1: We present a user-study evaluating the effect of reverberation and spatialization on cocktail-party effect in multi-talker virtual environments. We vary the levels of reverberance (direct, direct+early, direct+early+late) and spatialization (mono, stereo, binaural) and evaluate how different combinations affect the target-word identification performance. Left: a subject taking our study solver using HMD. Middle: the subject identifying the target-words in the presence of background chatter. Right: a representational view of our virtual scene showing the sources acting as distractors (red) and target source (green)

ABSTRACT

Many virtual reality applications let multiple users communicate in a multi-talker environment, recreating the classic cocktail-party effect. While there is a vast body of research focusing on the perception and intelligibility of human speech in real-world scenarios with cocktail party effects, there is little work in accurately modeling and evaluating the effect in virtual environments. Given the goal of evaluating the impact of virtual acoustic simulation on the cocktail party effect, we conducted experiments to establish the signal-to-noise ratio (SNR) thresholds for target-word identification performance. Our evaluation was performed for sentences from the coordinate response measure corpus in presence of multi-talker babble. The thresholds were established under varying sound propagation and spatialization conditions. We used a state-of-the-art geometric acoustic system integrated into the Unity game engine to simulate varying conditions of reverberance (direct sound, direct sound & early reflections, direct sound and early reflections and late reverberation) and spatialization (mono, stereo, and binaural). Our results show that spatialization has the biggest effect on the ability of listeners to discern the target words in multi-talker virtual environments. Reverberance, on the other hand, slightly affects the target word discerning ability negatively.

SAP '18, August 10–11, 2018, Vancouver, BC, Canada

ACM ISBN 978-1-4503-5894-1/18/08…$15.00
https://doi.org/10.1145/3225153.3225166

CCS CONCEPTS

• **Computing methodologies** → **Simulation evaluation**;

KEYWORDS

Auditory Perception, Virtual Environments/Reality, Virtual Acoustics

ACM Reference Format:
Atul Rungta, Nicholas Rewkowski, Carl Schissler, Philip Robinson, Ravish Mehra, and Dinesh Manocha. 2018. Effects of virtual acoustics on target-word identification performance in multi-talker environments. In *ACM Symposium on Applied Perception 2018, August 10–11, 2018, Vancouver, BC, Canada.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3225153.3225166

1 INTRODUCTION

Multi-talker environments commonly occur in the physical world around us. Examples of these can be seen in air-traffic control [Hilburn 2004], teleconferencing [Botros et al. 1986], or social gatherings where multiple people talk at a party or meeting, commonly known as the 'Cocktail-Party effect' [Cherry 1953]. This effect describes our ability to selectively attend to a particular target speech in the presence of competing interfering voices. This phenomenon was first described by [Cherry 1953]. Over the decades, its importance and ubiquity have led to a vast amount of research related to the acoustical, physiological, and psychological aspects of it. Prior work in this area has identified the important cues, such as binaural/monaural hearing, temporal properties of the signal, spectral properties, environment acoustics, etc. that help us segregate a single voice from a clutter of competing background noise. While a comprehensive analysis of all factors involved would be well beyond the scope of this paper, we focus our attention on the salient acoustical cues, i.e., reverberation and binaural/monaural hearing or spatialization, that have important implications for designing virtual multi-talker environments.

From an acoustical point of view, an important factor to consider is the spatial separation between the target and the distractors, which is also referred to as 'spatial release from masking' [Culling and Summerfield 1995]. Another important factor contributing to the cocktail-party problem is the effect of the environment, which contains the sound sources and the listener. Reverberation, in particular, is known to degrade the ability to identify the target speech in the presence of background chatter [Moncur and Dirks 1967].

While the effects of monaural and binaural hearing (spatialization henceforth), as well as of reverberation have been explored vis-à-vis the cocktail-party effect, these studies were either conducted in the real world or in virtual environments with very simple acoustic simulation capabilities [Crispien and Ehrenberg 1995; Koehnke and Besing 1996] and only a few distractors. Prior virtual acoustics systems 'simulate' the sound in the environment through the use of precomputed filters, which may not model the complex interaction of sound waves with the environment. With the advent of more accurate and faster virtual acoustics methods [Mehra et al. 2015; Raghuvanshi et al. 2009; Schissler et al. 2014; Siltanen et al. 2007; Southern et al. 2013; Vorländer 1989], it has now become possible to simulate complex environments in an accurate and interactive manner, thereby making it easier to conduct high-fidelity psychoacoustical experiments. Sound propagation effects in an environment are composed of three distinct components: direct sound, early reflections, and late reverberation. The early reflections help in localization and conveys spatial information about an environment to a listener. On the other hand, late reverberation enhances immersion and gives an impression of the size of the environment and the absorptivity. One of our goals is to evaluate the impact of these components in a multi-talker virtual environments.

The recent advent of social VR platforms, including Altspace VR, Facebook Spaces, Sansar, presents new challenges in dealing with multi-talker environments in a virtual world. A typical scenario could entail a situation where a group of people sharing a common virtual environment try to talk to each other, and thereby recreate the classic cocktail-party effect. As mentioned above, the acoustical factors can have a big impact on how accurately the effect is recreated. This makes it imperative to study the target speech identification performance in a virtual multi-talker environment while accurately simulating the acoustics of the environment. This served as our primary motivation to investigate the impact of sound propagation and spatialization in such a scenario.

Main Results: We present a novel experiment that is the first to examine the effect of real-time physically-based sound propagation and spatialization on the cocktail-party effect in complex virtual environments. We examine the effect on SNR (signal-to-noise ratio) threshold for target-word identification due to the combined effect of varying levels of reverberance and spatialization in an interactive virtual environment. Each of the two conditions had three levels associated with it - Reverberance(Direct sound only, Direct + Early reflections, Direct + Early reflections + Late reverberation) and Spatialization(Mono, Stereo, and HRTF-based binaural audio). The environment consisted of a rectangular room with ten sound sources in a circle around the subject, nine of which act as distractors and one of which acts as the target speech. The subjects experienced the virtual environment through an Oculus Rift CV1 head-mounted display with built-in headphones. We use a state-of-the-art interactive virtual acoustic simulation method [Schissler et al. 2014] to simulate the effect of sound propagating for all the ten sources through the environment. The resulting sound simulation system is capable of simulating the reverberance and spatialization levels for *all* the sound sources in the scene while maintaining interactivity. In particular, the novel components of our work include:

- First use of an interactive, physically-based sound propagation system with dynamic late reverberation to simulate and study the cocktail-party effect.
- Simulation of the full binaural room impulse response at the listener using generic HRTFs to study the effect of spatialization.
- First study to examine the effects of early reflections and late reverberation combined with various spatialization methods in a complex virtual environment with a high number of sources.

Our results show that spatialization and reverberance both have an effect on the target-word identification performance. Moving from monaural hearing to binaural hearing improves the identification performance, while going from anechoic to reverberant environment can negatively impact the target-word identification to a small extent.

2 BACKGROUND & RELATED WORK

In this section we give a brief overview of the cocktail-party effect and virtual acoustics and discuss the prior work done in these fields.

2.1 Cocktail-Party Effect

[Cherry 1953] was the first to identify the ability to selectively focus one's auditory attention on a target voice in presence of competing voices. Since then, a number of researchers have worked on this phenomenon seeking to identify the myriad of cues and processes that enable us to identify a target speech from competing background babble. In this section, we give a brief overview of some of that work.

Informational masking: This describes the masking caused by the linguistic content of the interfering speech and the target speech. Some of the work [Brungart 2001; Lutfi 1990; Pollack and Pickett 1958] shows that apart from energy masking, the interfering signal can mask the target by the content of the interfering speech. While most of the work in informational masking considers non-verbal stimuli, some work [Brungart 2001] explores verbal targets and maskers.

Fundamental frequency (F0): The differences in fundamental frequency (F0) of the interfering and target speeches have also been shown to positively affect the performance of target speech identification. Experiments have shown that understanding in a simultaneous vowel presentation task is dependent on the harmonic structure of the interfering sound. [Cheveigné 1997] and Summerfield et al. [Summerfield and Culling 1992], discovered that vowel identification for pairs containing inharmonic and harmonic vowels improved for the inharmonic vowels. However, the same effect was not observed for the harmonic vowel.

Temporal properties of masker: The ability to identify a target speech is affected by the temporal nature of the masker with a decrease in the temporal envelope of the distractor being beneficial for target identification. [Festen 1993] showed that a masking release of $6 - 8$ dB can be achieved with an interfering voice by adding a temporal shift to the varying width filter bands.

Binaural unmasking: Also known as 'spatial release from masking, this refers to spatial separation of the target and the interfering speech. This manifests as monaural [Bronkhorst and Plomp 1992; Shinn-Cunningham et al. 2001] and binaural advantage [Bronkhorst and Plomp 1988; Culling and Summerfield 1995]. In our experiment, we simulate spatial release from masking by simulating the target speech at random positions around the subject and rendering the audio either monaurally or binaurally.

Reverberation: Speech intelligibility suffers in reverberant environments. [Moncur and Dirks 1967] conducted experiments in reverberant environments and found that binaural hearing was superior to monaural hearing. Further, [Bronkhorst 2015] reported that reverberation and hearing impairment can negatively affect target speech identification by up to 10 dB.

Visual cues: [Gonzalez-Franco et al. 2017] investigated target identification in immersive multi-talker virtual environments and observed that visual-speech cues have a profound effect on our perception turning masker into target and vice-versa. Further, [MacDonald and McGurk 1978] showed that a target's lip movements strongly affects the auditory perception of natural speech. These findings caused us to consider a simplistic visual environment to perform an isolated study of the auditory cues involved in target speech performance in multi-talker environments.

2.2 Virtual Acoustics & Spatial Audio

In this section, we give a brief overview of prior work in sound propagation and spatial sound.

2.2.1 Virtual acoustics. Deals with the simulation and interaction of sound within the virtual environment. Since sound includes phenomena such as diffraction, focusing, and specular and diffuse reflection, virtual acoustics methods seek to simulate these phenomena in the virtual environment. The underlying mathematical principal governing the interaction and behavior of sound is governed by the acoustic wave equation given in the time-domain as:

$$\nabla^2 p - \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} = F(\mathbf{x}, t) \quad \mathbf{x} \in \Omega \tag{1}$$

where ∇^2 is the Laplacian, p is the pressure, $c = 343ms^{-1}$ is the speed of sound, $F(\mathbf{x}, t)$ is the forcing term corresponding to the source, and Ω is the domain on which the equation is defined. Solving Eq 1 above gives the sound pressure p at a point in the domain. Over the years, research in virtual acoustics has sought to solve the above equation. This has yielded two distinct ways in which sound is modeled: Geometric methods and Wave-based methods. We now take a brief look at these methods.

Geometric Sound Propagation: These methods assume that sound travels in straight lines and treats them like rays. This enables them to use methods such as ray-tracing [Krokstad et al. 1968], beam-tracing [Funkhouser et al. 1998], and frustum tracing [Chandak et al. 2008; Lauterbach et al. 2007]. These methods are fast and can be used for interactive computation of early reflections. Recent advances in geometric sound propagation use temporal and spatial coherence to generate higher order specular and diffuse reflections (i.e., late reverberation) at interactive rates [Schissler et al. 2014] and can also approximate diffraction around smooth objects [Rungta et al. 2018]. Further, a large number of sources can also be handled at interactive rates [Schissler and Manocha 2017]. We use these interactive methods to simulate early reflections and late reverberation from multiple sources.

Wave-based Sound Propagation: These methods directly solve the acoustic wave-equation (Eq. 1) using numeric methods and tend to be the most accurate. But the discretization conditions imposed by the equation makes these methods very compute intensive. As a result, they are limited to simulate the low-frequency effects and static scenes. Many precomputation-based methods have been proposed that store the pressure fields computed offline and then use them efficiently at runtime to compute the IRs from a moving source or a moving listener [Mehra et al. 2013, 2015; Raghuvanshi et al. 2009]. Current implementations of such wave-based sound propagation algorithms is typically limited to simulating low frequencies (i.e. up to $< 1kHz$) in static scenes and they can be combined with geometric methods for higher frequencies.

2.2.2 Spatial Audio. Our auditory system perceives the incoming sound's direction enabling us to localize its position in the

world. Spatial audio deals with simulating this phenomenon in virtual environments. The most commonly used spatialization method is called stereo and is frequently implemented using vector-based amplitude panning [Pulkki 1997]. More sophisticated methods include ambisonics [Gerzon 1973], which decomposes the sound-field in a first-order plane wave basis. Another commonly used spatial audio method is called binaural or Head-related transfer function (HRTF) based rendering. This method considers the geometry of the human head and ears. HRTF captures the scattering of sound due to the head and ears and uses that to render spatial audio [Begault and Trejo 2000]. Since HRTF is a function of the individual's head, it needs to be computed individually. Many works [Katz 2001; Meshram et al. 2014; Wightman and Kistler 1989] have investigated the measurement and simulation of individual HRTFs. Measuring an individual's HRTF accurately can be a slow and expensive process that requires specialized hardware setup or capture, hence limiting their applicability. [Algazi et al. 2001] provided an HRTF database for 45 individuals that can be used as an approximation of an individual's HRTF with good accuracy [Berger et al. 2018; Drullman and Bronkhorst 2000; Nelson et al. 1999]. Apart from that, HRTF measurements have also been provided for human head using mannequins such as KEMAR [Gardner and Martin 1995]. We use mono, stereo, and KEMAR-based HRTF (HRTF henceforth) based rendering methods in our experiments.

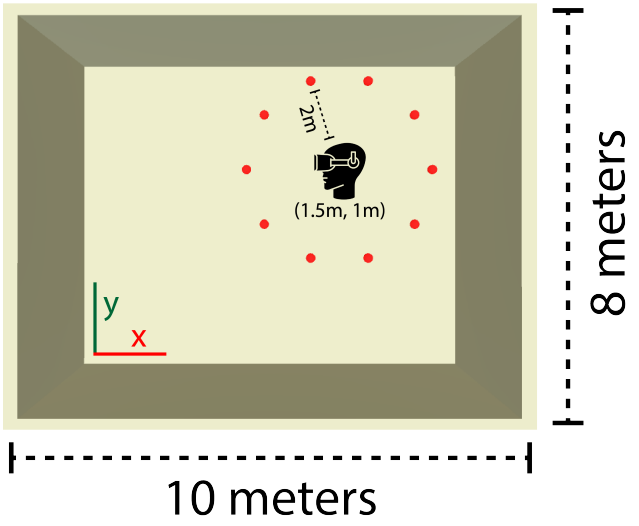

Figure 2: Our setup consisted of a virtual room of dimensions 10m x 8m x 3m centered at (0, 0, 0) m. The subjects were placed slightly off-center at (1.5, 1, 0.3) m. 10 sound sources (red dots) were placed around them in a circle of radius of 2m, varying the source positions only in the azimuth keeping the elevation fixed. One of these was selected at random to play the target speech while the rest simulated the distractors. The head position was fixed but the subjects were encouraged to rotate their head to glean additional cues from the environment. The walls had a reflection coefficient mimicking that of a physical room. The reverberation time (RT_{60}) was approximately 1s.

3 EXPERIMENT

3.1 Participants

Sixteen participants took part in the study (14 Male). Their ages ranged from 20 to 32 (mean = 26.4, SD = 3.05). All reported normal hearing and were recruited from the students and staff at a university campus. All subjects were either native English speakers or had professional proficiency in the language.

3.2 Apparatus

The setup consisted of the Oculus Rift CV1 and an X-box One controller for input. The software consisted of an in-house realtime, geometric sound propagation engine integrated with the Unity game engine. The study framework was written in C#, but the analysis code was written in MATLAB. The setup was run on a Dell workstation with 4-core Intel Xeon E5620 processors and 24 GB memory. The operating system was Windows 8.

3.3 Stimuli

The coordinate-response-measure (CRM) corpus was used for target-word identification speech [Bolia et al. 2000]. The corpus selected had a male voice uttering a target speech in the following format: "Ready [Call sign] goto [Color] [Number] now". The call signs had eight different options [Charlie, Ringo, Laker, Hopper, Arrow, Tiger, Eagle, Baron] , color had four [Blue, Red, White, Green], and number had eight as well [One, Two, Three, Four, Five, Six, Seven, Eight] giving a total of 256 combinations of target CRM speech. (Fig 4). The background babble consisted of 10 different dry recorded sounds of human males talking.

3.4 Design & Procedure

The environment consisted of a simple rectangular room of dimensions 10m x 8m x 3m simulating a similar real-world room (Fig. 2) centered at (0, 0, 0) m. The walls had reflectivity similar to that of its real-world counterpart with an $RT_{60} \approx 1.0$s. The sound was created using a ray-tracing-based geometric sound propagation system [Schissler et al. 2014] that can simulate multiple sources in realtime. The three reverberance conditions were simulated by changing the number of rays traced in the scene (Fig. 3). Direct sound was simulated by considering a reflection order of 0, basically meaning that the sound would get absorbed as soon it hit the walls of the room. Early reflections were simulated by considering 3 orders of reflections, while late reverberation (or full reverberance) was simulated by considering very high order (2000) reflections. As for spatialization levels, stereo was simulated using vector-based amplitude panning. Binaural spatialization was simulated by convolving a generic, KEMAR HRIR (head-related impulse response) with the entire room impulse response giving us the binaural room impulse response (BRIR) at the listener. Monaural listening constituted both ears hearing the same sound.

The subject was placed at (1.5, 1, 0.3) with the sources placed at ear level. The ten virtual sources (red dots in Fig 2) were placed in a circle of radius 2m around the subject. In its default state, each of the ten sources looped different speech segments in human male voices. **This constituted our multi-talker environment**. The visuals of the environment were kept simple on purpose to prevent

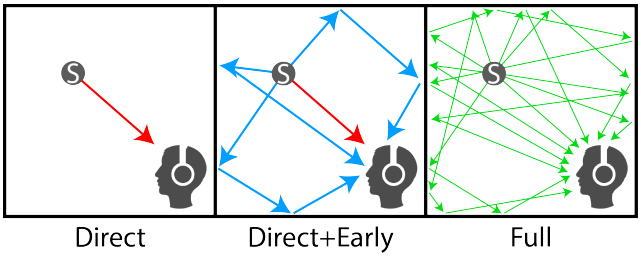

Figure 3: The figure shows a representative view of the various levels of reverberance simulated in our experiment. Direct sound corresponds to an anechoic environment, while direct + early reflection corresponds to an environment with a low reverberation time $RT_{60} \approx 0.2s$. Full (i.e. with late reverberation) corresponds to an environment where sound decays naturally after many reflections off the walls. ($RT_{60} \approx 1s$)

vision from becoming the dominant cue in deciphering the target speech [Gonzalez-Franco et al. 2017].

The subjects identified the target speech through input via an X-box One controller. Each of the categories of the CRM corpus (call sign, color, and number) was navigated using the left-stick and then selected using the A button. Corrections to the response could be made by pressing the B button. After making the selection, the response was submitted by pressing the menu button. A new trial also began with the pressing of the menu button. Before starting the experiment, the subjects underwent a short training phase. This consisted of the subjects taking a short version of the study for $15 - 20$ trials. We explained to them how the input worked and allowed them to get comfortable with the controller and answering the questions.

The experiment was divided into three blocks one for each reverberation condition which was selected at random. For each block, trials were interleaved between the spatialization conditions chosen at random. On each trial, one of the ten sources was selected at random which played a randomly chosen 256 CRM clip once. After playing the CRM clip the menu (Fig 4) popped up prompting the subject to identify the target-words. The subjects made their selections and submitted their response by pressing the menu button. Each experiment lasted for an average of 300 trials with each trial lasting for around $6 - 7$ seconds. Subjects were asked to take a 5 minute break at the end of each block. The total experiment lasted for around 45 minutes. No fatigue was reported.

The starting power-level of the CRM clip was 85 dB while the background voices were fixed at 65 dB each. Before starting the experiments, the RMS level of the CRM clips and the background clips were matched. None of the subjects were familiar with the CRM corpus.

The thresholds were measured using a 2-down/1-up, interleaved adaptive-staircase method [Levitt 1971] with step-sizes of 6 dB for the first two reversals, 3 dB for the next three reversals, and 1 dB for the last four reversals (for a total of nine reversals). This was a within group study and all subjects experienced all the 9 (3 reverberance x 3 spatialization) conditions.

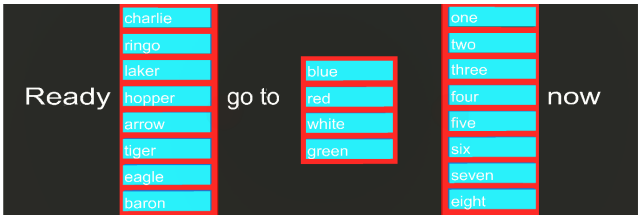

Figure 4: The menu presented to the subjects to select the target CRM words. The selection was made using an XBox One controller. We simulate the background chatter using virtual acoustic algorithms and evaluate the impact on selecting the target-words.

3.5 Results

We ran repeated measures ANOVA on the SNR thresholds for our 3×3 factorial design. Our analysis showed a significant main effect of reverberance $F(2, 30) = 9.68, p = 0.001$ indicating that subjects were able to utilize the propagation cues while determining their thresholds. Spatialization also showed a significant effect $F(2, 30) = 81.1, p < 0.001$, indicating that it subjects were able to use the spatial cues for target-word recognition. The interaction between spatialization and reverberance failed to show significance $F(4, 60) = 0.61, p = 0.656$.

A post-hoc analysis of reverberance with Bonferroni adjustment showed direct and full conditions varied significantly ($p = 0.002$) and so did early and full conditions ($p < 0.001$) but not in case of direct and early conditions. This indicates that the subjects' SNR thresholds were not affected when switching between direct sound and direct sound with added early reflections, but the thresholds were affected when switching from direct to full as well as when switching from early to full. Similar post-hoc analysis of spatializtion showed significant difference between mono & stereo ($p < 0.001$) and mono & HRTF ($p < 0.001$) but no difference between stereo and HRTF ($p = 0.17$). This indicates that subjects' SNR thresholds did not change significantly when going from stereo and HRTF conditions, but were affected significantly when switching from mono to stereo and mono to HRTF conditions.

4 ANALYSIS & DISCUSSION

Our results confirm existing results observed in other similar studies. Our experiment was meant to validate our use of an interactive sound propagation system for evaluating virtual, multi-talker environments. We now take a look at the individual factors we considered for our experiment.

Spatialization: Spatialization had the biggest effect on the target-word identification performance in our studies. Any monaural advantage that could've been obtained because of head shadowing was negated due to the spatial distribution of the sources in the scene. As is seen in Fig 5 and Fig 6, mono shows an SNR gain of 1.1 dB when combined with early reflections confirming the benefit of early reflections in monaural listening [Bradley et al. 2003]. Although, the standard error for monaural with early reflections is high indicating the variance of the responses which might have been the result of the spatial configuration and the high number

of distractors in the scene. Binaural listening fares much better than monaural condition and proved to be robust to spatial configuration of the distractors. SNR gain with binaural listening is between 3-5 dB compared with monaural listening in line with existing data [Hawley et al. 2004]. Overall, binaural listening is also robust to reverberance conditions presented in our experiment with a loss of less than 1 dB when going from early reflections to late reverberation.

An interesting observation to be made is the performance of stereo vs. HRTF. Although HRTF-based rendering takes advantage of the directional cues provided by early reflections, our experiment showed no statistically significant advantage of using HRTF over stereo ($p = 0.17$). This basically indicates that the shape of the function defining the effect of binaural listening is similar for stereo and HRTF but HRTF based listening essentially provides a modest gain of 0.4 dB on average over stereo. The gain should increase with changing elevation of the sources since stereo listening does not incorporate elevation cues while HRTF does.

Reverberance: Reverberance had a small but significant interaction with the SNR thresholds. In general, increase in reverberance is correlated with a decrease in speech intelligibility and our results clearly demonstrate that. Further, as can be seen in Fig 5, adding late reverberation causes a degradation in performance for all conditions. This can be easily explained by the decrease in robustness of both monaural and binaural cues in presence of reverberation. An interesting observation is the lack on significance on the thresholds when moving from the direct and direct + early condition. Although, monaural listening can clearly be seen benefiting from added early reflections, binaural listening does not show a similar effect. We believe that the degradation would become more pronounced as the reverberance or RT_{60} of the environment increases.

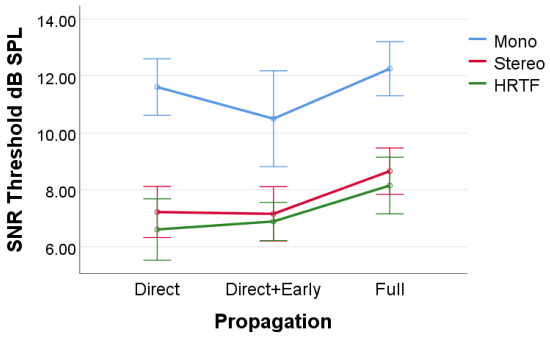

Figure 5: The plot shows the mean SNR Thresholds with the error plots for the various reverberance and spatialization conditions. Mono performs the worst among the three spatialization conditions. The effects of early reflections do not manifest prominently when stereo/binaural listening is available but are beneficial in case of monaural listening. For stereo and HRTF, the threshold increase is low when going from direct to direct+early, but increases more rapidly when late reverberation is introduced.

5 CONCLUSIONS, LIMITATIONS AND FUTURE WORK

We present a novel experiment to evaluate the cocktail-party effect using an interactive, physically-based sound propagation system along with different spatialization conditions. Our setup combined two important acoustic characteristics that are known to have an effect on this phenomenon, viz., reverberance and spatialization. Our setup involves a rectangular room with 10 sources, *each* of which acted as either distractor or the target speech. Our results show that spatialization has a big positive effect on the target-word identification performance with subjects showing a gain of up to 5 dB for stereo & binaural over monaural listening. Reverberence, on the other hand, shows a modest loss of 1.5 dB for HRTF in full reverberation vs. HRTF in anechoic (or direct only) condition. These results show the same trends as seen in previous isolated studies, although a direct comparison would be difficult since no other study has considered more than four interfering sound sources.

Our experiment has a few limitations. Our sources were placed at ear-level thereby reducing the advantages of HRTF-based binaural over simple stereo rendering. Placing the sources with some elevation would be an interesting venue for future work. The distance between the sources and subject was kept fixed and it would be interesting to examine the effect of staggering the sources around the subject. Apart from these, all our sources were in the direct line-of-sight of the subject and the effect of sound occlusion or diffraction effects could not be observed. In real world, not all interfering sources may be in the line-of-sight of the subject and it would be interesting to design a study considering occluded sources with significant diffraction effects In the future, we would like to evaluate the effect in more sophisticated virtual environments that couple high acoustic fidelity with high visual fidelity. Personalized HRTFs could also be evaluated and compared with generic HRTFs, especially when the interfering sources lie at an elevation with respect to the subject. Finally, we would like to integrate these techniques into a Social VR setting and perform more detailed evaluations.

ACKNOWLEDGMENTS

The authors would like to thank the subjects who took part in the user-study. The work was supported in part by ARO grant W911NF14-1-0437, NSF grant 1320644, and Oculus Research.

REFERENCES

V Ralph Algazi, Richard O Duda, Dennis M Thompson, and Carlos Avendano. 2001. The cipic hrtf database. In *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the.* IEEE, 99–102.

Durand R Begault and Leonard J Trejo. 2000. 3-D sound for virtual reality and multimedia. (2000).

Christopher C Berger, Mar Gonzalez-Franco, Ana Tajadura-Jiménez, Dinei Florencio, and Zhengyou Zhang. 2018. Generic HRTFs may be good enough in Virtual Reality. Improving source localization through cross-modal plasticity. *Frontiers in Neuroscience* 12 (2018), 21.

Robert S Bolia, W Todd Nelson, Mark A Ericson, and Brian D Simpson. 2000. A speech corpus for multitalker communications research. *The Journal of the Acoustical Society of America* 107, 2 (2000), 1065–1066.

Radamis Botros, Onsy Abdel-Alim, and Peter Damaske. 1986. Stereophonic speech teleconferencing. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86.*, Vol. 11. IEEE, 1321–1324.

JS Bradley, Hiroshi Sato, and M Picard. 2003. On the importance of early reflections for speech in rooms. *The Journal of the Acoustical Society of America* 113, 6 (2003), 3233–3244.

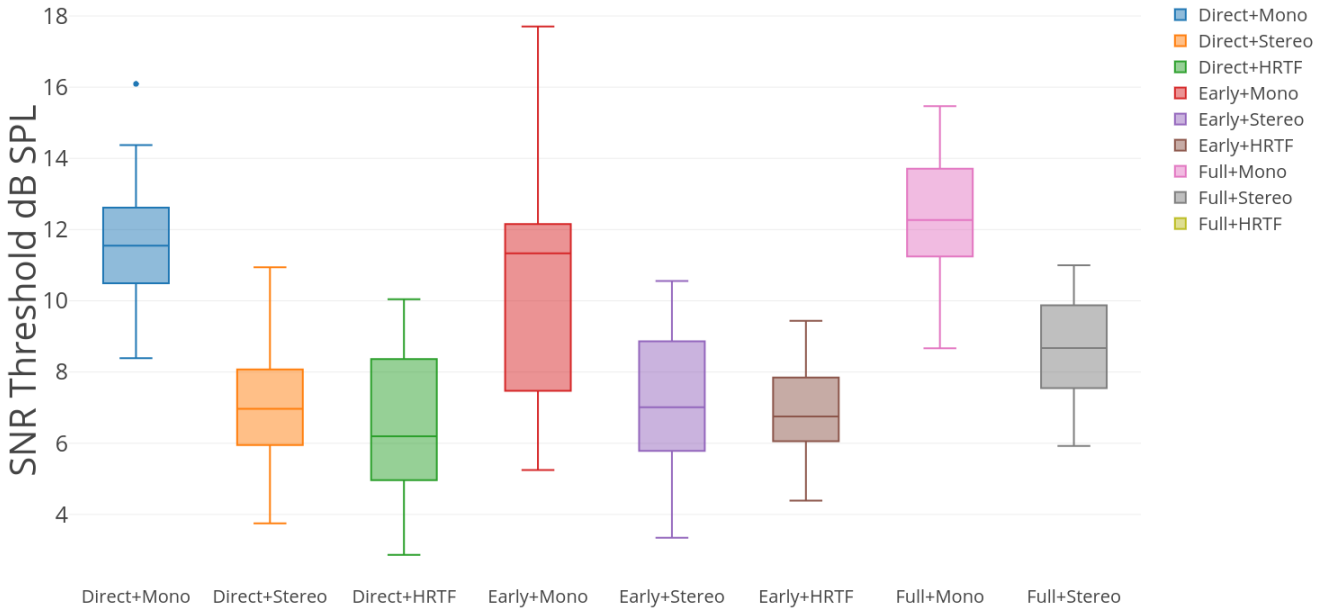

Figure 6: The box plots shows a flattened representation of the nine experimental conditions in our experiment. The horizontal line inside each of the boxes represents the median SNR threshold. As can be clearly seen, early reflections are beneficial for the monaural listening condition but show high variability. Binaural listening (Stereo and HRTF) perform similarly throughout.

AW Bronkhorst and R Plomp. 1988. The effect of head-induced interaural time and level differences on speech intelligibility in noise. *The Journal of the Acoustical Society of America* 83, 4 (1988), 1508–1516.

AW Bronkhorst and R Plomp. 1992. Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing. *The Journal of the Acoustical Society of America* 92, 6 (1992), 3132–3139.

Adelbert W Bronkhorst. 2015. The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics* 77, 5 (2015), 1465–1487.

Douglas S Brungart. 2001. Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America* 109, 3 (2001), 1101–1109.

Anish Chandak, Christian Lauterbach, Micah Taylor, Zhimin Ren, and Dinesh Manocha. 2008. Ad-frustum: Adaptive frustum tracing for interactive sound propagation. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (2008), 1707–1722.

E Colin Cherry. 1953. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America* 25, 5 (1953), 975–979.

Alain de Cheveigné. 1997. Concurrent vowel identification. III. A neural model of harmonic interference cancellation. *The Journal of the Acoustical Society of America* 101, 5 (1997), 2857–2865.

Kai Crispien and Tasso Ehrenberg. 1995. Evaluation of the-cocktail-party effect-for multiple speech stimuli within a spatial auditory display. *Journal of the Audio Engineering Society* 43, 11 (1995), 932–941.

John F Culling and Quentin Summerfield. 1995. Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay. *The Journal of the Acoustical Society of America* 98, 2 (1995), 785–797.

Rob Drullman and Adelbert W Bronkhorst. 2000. Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. *The Journal of the Acoustical Society of America* 107, 4 (2000), 2224–2235.

Joost M Festen. 1993. Contributions of comodulation masking release and temporal resolution to the speech-reception threshold masked by an interfering voice. *The Journal of the Acoustical Society of America* 94, 3 (1993), 1295–1300.

Thomas Funkhouser, Ingrid Carlbom, Gary Elko, Gopal Pingali, Mohan Sondhi, and Jim West. 1998. A beam tracing approach to acoustic modeling for interactive virtual environments. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*. ACM, 21–32.

William G Gardner and Keith D Martin. 1995. HRTF measurements of a KEMAR. *The Journal of the Acoustical Society of America* 97, 6 (1995), 3907–3908.

Michael A Gerzon. 1973. Periphony: With-height sound reproduction. *Journal of the Audio Engineering Society* 21, 1 (1973), 2–10.

Mar Gonzalez-Franco, Antonella Maselli, Dinei Florencio, Nikolai Smolyanskiy, and Zhengyou Zhang. 2017. Concurrent talking in immersive virtual reality: on the dominance of visual speech cues. *Scientific Reports* 7, 1 (2017), 3817.

Monica L Hawley, Ruth Y Litovsky, and John F Culling. 2004. The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *The Journal of the Acoustical Society of America* 115, 2 (2004), 833–843.

Brian Hilburn. 2004. Cognitive complexity in air traffic control: A literature review. *EEC note* 4, 04 (2004).

Brian FG Katz. 2001. Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation. *The Journal of the Acoustical Society of America* 110, 5 (2001), 2440–2448.

Janet Koehnke and Joan M Besing. 1996. A procedure for testing speech intelligibility in a virtual listening environment. *Ear and Hearing* 17, 3 (1996), 211–217.

Asbjørn Krokstad, Staffan Strom, and Svein Sørsdal. 1968. Calculating the acoustical room response by the use of a ray tracing technique. *Journal of Sound and Vibration* 8, 1 (1968), 118–125.

Christian Lauterbach, Anish Chandak, and Dinesh Manocha. 2007. Interactive sound rendering in complex and dynamic scenes using frustum tracing. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (2007), 1672–1679.

HCCH Levitt. 1971. Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical society of America* 49, 2B (1971), 467–477.

Robert A Lutfi. 1990. How much masking is informational masking? *The Journal of the Acoustical Society of America* 88, 6 (1990), 2607–2610.

John MacDonald and Harry McGurk. 1978. Visual influences on speech perception processes. *Perception & Psychophysics* 24, 3 (1978), 253–257.

Ravish Mehra, Nikunj Raghuvanshi, Lakulish Antani, Anish Chandak, Sean Curtis, and Dinesh Manocha. 2013. Wave-based sound propagation in large open scenes using an equivalent source formulation. *ACM Transactions on Graphics (TOG)* 32, 2 (2013), 19.

Ravish Mehra, Atul Rungta, Abhinav Golas, Ming Lin, and Dinesh Manocha. 2015. Wave: Interactive wave-based sound propagation for virtual environments. *IEEE transactions on visualization and computer graphics* 21, 4 (2015), 434–442.

Alok Meshram, Ravish Mehra, Hongsheng Yang, Enrique Dunn, Jan-Michael Franm, and Dinesh Manocha. 2014. P-HRTF: Efficient personalized HRTF computation for high-fidelity spatial sound. In *Mixed and Augmented Reality (ISMAR), 2014 IEEE International Symposium on*. IEEE, 53–61.

John P Moncur and Donald Dirks. 1967. Binaural and monaural speech intelligibility in reverberation. *Journal of Speech, Language, and Hearing Research* 10, 2 (1967), 186–195.

W Todd Nelson, Robert S Bolia, Mark A Ericson, and Richard L McKinley. 1999. Spatial audio displays for speech communications: A comparison of free field and virtual acoustic environments. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 43. SAGE Publications Sage CA: Los Angeles, CA, 1202–1205.

Irwin Pollack and JM Pickett. 1958. Stereophonic listening and speech intelligibility against voice babble. *The Journal of the Acoustical Society of America* 30, 2 (1958), 131–133.

Ville Pulkki. 1997. Virtual sound source positioning using vector base amplitude panning. *Journal of the audio engineering society* 45, 6 (1997), 456–466.

Nikunj Raghuvanshi, Rahul Narain, and Ming C Lin. 2009. Efficient and accurate sound propagation using adaptive rectangular decomposition. *IEEE Transactions on Visualization and Computer Graphics* 15, 5 (2009), 789–801.

Atul Rungta, Carl Schissler, Nicholas Rewkowski, Ravish Mehra, and Dinesh Manocha. 2018. Diffraction Kernels for Interactive Sound Propagation in Dynamic Environments. *IEEE transactions on visualization and computer graphics* (2018).

Carl Schissler and Dinesh Manocha. 2017. Interactive sound propagation and rendering for large multi-source scenes. *ACM Transactions on Graphics (TOG)* 36, 1 (2017), 2.

Carl Schissler, Ravish Mehra, and Dinesh Manocha. 2014. High-order diffraction and diffuse reflections for interactive sound propagation in large environments. *ACM Transactions on Graphics (TOG)* 33, 4 (2014), 39.

Barbara G Shinn-Cunningham, Jason Schickler, Norbert Kopčo, and Ruth Litovsky. 2001. Spatial unmasking of nearby speech sources in a simulated anechoic environment. *The Journal of the Acoustical Society of America* 110, 2 (2001), 1118–1129.

Samuel Siltanen, Tapio Lokki, Sami Kiminki, and Lauri Savioja. 2007. The room acoustic rendering equation. *The Journal of the Acoustical Society of America* 122, 3 (September 2007), 1624–1635.

A. Southern, S. Siltanen, D. T. Murphy, and L. Savioja. 2013. Room Impulse Response Synthesis and Validation Using a Hybrid Acoustic Model. *IEEE Transactions on Audio, Speech, and Language Processing* 21, 9 (2013), 1940–1952.

Quentin Summerfield and John F Culling. 1992. Periodicity of maskers not targets determines ease of perceptual segregation using differences in fundamental frequency. *The Journal of the Acoustical Society of America* 92, 4 (1992), 2317–2317.

Michael Vorländer. 1989. Simulation of the transient and steady-state sound propagation in rooms using a new combined ray-tracing/image-source algorithm. *The Journal of the Acoustical Society of America* 86, 1 (1989), 172–178. DOI: http://dx.doi.org/10.1121/1.398336

Frederic L Wightman and Doris J Kistler. 1989. Headphone simulation of free-field listening. I: stimulus synthesis. *The Journal of the Acoustical Society of America* 85, 2 (1989), 858–867.