Auditory Motion Perception: Investigation of Benefit in Multi-Talker Environments

By

Timothy J. Davis

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Hearing and Speech Sciences

May 8, 2020

Nashville, Tennessee

Approved:

Daniel H. Ashmead, Ph.D.

D. Wesley Grantham, Ph.D.

Melissa C. Duff, Ph.D.

Amy E. Needham, Ph.D.

ACKNOWLEDGEMENTS

This completed dissertation has been many years in the making. I owe a tremendous gratitude to my committee for their support of my completion of this endeavor. In particular, I want to thank Dr. Wes Grantham for being so giving of his time and expertise, especially in retirement, from afar. He was the greatest source of continuity throughout this process and his contributions to this dissertation as well as earlier projects were invaluable. Indeed, he was the reason I became interested in research all those years ago. Drs. Chris Stecker, Ben Hornsby, and Todd Ricketts all lent their technical expertise of acoustics and the hardware in the anechoic chamber, for which I am very appreciative. Finally, it is to Dr. Dan Ashmead that I owe the highest level of gratitude. At a time when my sense of direction in this program was faltering the most, he lent his full support to my completion of this degree. Beyond that, he sparked a new sense of dedication and belief in myself that had been missing for some time. Dr. Ashmead has been so extremely supportive and giving of his time and expertise. His kindness and generosity with his time was inspiring and without which, I would not have completed this work.

Finally, I owe the greatest appreciation to my wife, Heather. Her unwavering support of my best interests are inspiring, daily. She is my strength, my support, and motivation to maintain balance in life. It is for her, and because of her, that I have completed this work.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

Introduction

The human auditory system is remarkably tuned for understanding speech, not only in quiet but also in settings with various kinds of competing sounds. Listeners frequently need to understand what one talker is saying while other talkers in the vicinity are also speaking. When this situation is created in a laboratory context, speech from the additional talker(s) is considered to "mask" what the target talker is saying. This can be characterized as a signal-in-noise scenario - the signal is what the target talker says, and the noise is what the other talkers say (the other talkers are often called distracters, or maskers).

When a sound is presented to a listener from a direction off the median plane, it arrives sooner and with more intensity at the ear nearest to the sound source. These differences between ears are known as interaural time difference (ITD), and interaural level difference (ILD). With the use of two hearing ears, one could potentially perceive differences in timing and intensity between ears to determine the location of a sound. Raleigh (1907) is commonly cited as the first to publish findings from experiments that explored the abilities of the auditory system to discriminate sounds based on differences between ears. In the following century, we have learned a tremendous amount about the ways in which the auditory system is finely tuned for perception of sounds in space. In a speech perception context, presenting talkers from disparate locations allows the binaural system to distinguish one talker from another, thereby creating a release from masking and improving speech understanding, compared to presenting talkers from the same location (Hirsh, 1950).

Masking Types

In multi-talker environments, two types of masking are generally recognized: energetic and informational (Arbogast, Mason, & Kidd, 2002; Douglas S. Brungart, 2001; Freyman, Helfer, McCall, & Clifton, 1999; G. Kidd, Jr., Mason, Rohtla, & Deliwala, 1998). Energetic masking occurs when components of the signal and noise fall within the same critical frequency band, rendering the signal less audible. The auditory system is organized into a set of critical bands (Fletcher, 1940) that originate from how the cochlea responds to the frequency components of sounds. Neural processing occurs in a tonotopic (frequency-specific) mode at ascending levels of the auditory system. If the acoustic energy from the "noise" source within a critical band is greater than that from the "signal" source, then it is difficult to perceive the signal. The more two talkers overlap in frequency, the more difficult it is to track what they are saying. Energetic masking is thought to be a fairly peripheral process wherein the desired signal is masked at either the level of the cochlea, a monaural effect, or brainstem where neural signals from the two ears converge, a binaural effect. Informational masking is sometimes characterized as occurring beyond that which is attributed to energetic masking. A hallmark of informational masking is that it occurs when both signals are potentially discernible but higher-level processes cannot separate them. That is, the "target" signal is transmitted intact to the cortex, but perceptual similarities between it and other distracting signals render the target less recognizable. For example, if a man and woman are talking at the same time but saying different things, there would be moments of spectral and temporal overlap, but most of the time the acoustic information would be energetically distinct. Despite this, the perceptual similarities between the two voices make attending to one voice more difficult than if the

competing signal was a modulated speech-shaped noise. A number of parameters have been shown to influence informational masking, including the gender differences of the talkers (Douglas S. Brungart, 2001), fundamental frequency of talkers (Arbogast et al., 2002), actual and perceived spatial location of talkers (D. S. Brungart & Simpson, 2007; Freyman, Balakrishnan, & Helfer, 2001; Freyman et al., 1999), and familiarity with talker locations (D. S. Brungart & Simpson, 2007; G. Kidd, Arbogast, Mason, & Gallun, 2005). The concept of informational masking is broadly related to models of selective attention as applied to speech perception and understanding. While some studies have attempted to, as completely as possible, isolate one form of masking from the other, such manipulations are understandably limited to a laboratory setting. In our everyday listening environments, both forms of masking are constantly intertwined. Often, manipulation of one form changes the relative impact of the other. When informational masking is thought of in terms of selective attention, it is easy to imagine how a manipulation of the spatial layout of a listening environment might change the signal-to-noise ratio (energetic), and thereby make allocating attention to one talker easier (informational).

Spatial Separation of Talkers

One consideration about speech understanding in multi-talker settings that has been well studied is spatial separation of talkers (Bronkhorst & Plomp, 1988; Douglas S. Brungart, 2001; Cherry, 1953; Freyman et al., 1999; G. Kidd, Jr. et al., 1998). This occurs, for example, when people are seated around a dining table. Spatial release from masking is typically defined as an improvement in detection or speech understanding when the target and masker(s) are spatially separated, compared to when they are co-located. Spatial separation can allow for

robust improvements in speech understanding compared to a situation where talkers are coming from the same location. Arbogast et al. (2002) used the Coordinate Response Measure (CRM) sentences corpus to measure speech recognition in two spatial configurations. In the co-located condition, a male "target" talker and a competing distracter talker (or in some conditions a noise) were presented simultaneously from directly in front of the listener (the voices were played through the same loudspeaker). In the separately located condition, the target talker was in front of the listener, but the distracter was 90 degrees azimuth to the listener's right. These spatial conditions were compared for both speech and noise distracters (the noises were created to maximize or minimize energetic/informational masking). The signal-to-noise ratio of the speech and noise maskers was adaptively varied on a trial-by-trial basis centering on the 50% correct performance level. There was a benefit of spatial separation of talkers as large as 18 dB when the masker was minimally energetic (different frequency composition of target and distracter), and a smaller but still significant 6.9 dB benefit when the masker had the same frequency composition as the target. The authors concluded that the benefit of spatial separation is due, primarily, to a release from informational masking. Although the maskers used in this study were rather artificial, similar benefits have been documented in several other studies with more natural speech maskers (Allen, Carlile, & Alais, 2008; Bronkhorst & Plomp, 1988; Hawley, Litovsky, & Culling, 2004; Noble & Perrett, 2002). To that end, listening in an environment in which sounds come from distinct locations affords significant advantages for speech understanding.

While these studies have been compelling and largely consistent in demonstrating the benefit of spatial separation of talkers, they all involved stationary sound sources. Less is known

about understanding speech from moving sound sources (or when the listener is moving), even though such motion is common in our daily lives. People walk by, birds fly overhead, and cars drive past us, all providing meaningful information about our environment. In addition, self-motion from our own locomotion or head turns allow us to scan our auditory environment to gain information about the locations of sound sources. Without a doubt, our acoustic environment is often rather spatially dynamic, and we rely on the sophistication of our auditory system to separate an attended signal from the background.

Motion Perception

The utility of motion perception is not limited to the auditory system. Indeed, much of the primate visual system is sensitive to various types of motion (see Andersen (1997) for review). Because of our ability to locomote, perception of motion in our visual field is perhaps one of the most fundamental necessities of the visual system. This applies not only to the ability to understand the environment as we move through it, but also to detect and follow motion of other objects around us. In the visual domain, the ability to perceive motion goes beyond the simple ability to track a moving object. A growing body of research has focused on the phenomenon of visual "pop-out" (Christ & Abrams, 2008; Smith & Abrams, 2018; Sunny & von Muhlenen, 2011). Visual pop-out exists when the detectability of an object increases due to a unique characteristic of it compared to other objects around it. Changes in luminosity, texture, color, and position have all been shown to allow one visual object to perceptually stand out from the rest (Borst, 2000; Duncan & Humphreys, 1989; Hillstrom & Yantis, 1994; Nothdurft, 1993; Smith & Abrams, 2018). Imagine a hungry cat on a perch, surveying a field of stationary grass. When a scurrying mouse moves across the cat's visual field, the contrast of the motion

against the stationary background immediately draws the cat's attention so it can give chase. There is no doubt that visual motion perception is vital for survival, and as such there is a substantial portion of the visual system dedicated to it.

Given the auditory system's robust sensitivity to differences in locations of sound sources (Mills, 1958; Srinivasan, Jakien, & Gallun, 2016) and established sensitivity to motion (Carlile & Leung, 2016; Freeman et al., 2014; Grantham, 1986; Kaczmarek, 2005; Locke, Leung, & Carlile, 2016; Saberi & Perrott, 1990), the search has been on for the last few decades for a similar auditory motion-encoding area in the temporal lobe. Indeed, this area of research is receiving increasing attention as modern imaging techniques allow for greater access to the underlying neural mechanisms of auditory motion processing (see Grothe, Pecka, and McAlpine (2010) for review). Using electroencephalography (EEG), functional magnetic resonance imaging (fMRI), magnetoencephalography (MEG), positron emission tomography (PET), and other neuroimaging tools, much is being learned about how the auditory system processes spatial cues, both static and dynamic (Alink, Euler, Kriegeskorte, Singer, & Kohler, 2012; Battal, Rezk, Mattioni, Vadlamudi, & Collignon, 2019; Baumgart, Gaschler-Markefski, Woldorff, Heinze, & Scheich, 1999; Bednar & Lalor, 2020; T. D. Griffiths & Green, 1999; T. D. Griffiths et al., 1996; Zimmer & Macaluso, 2009).

Given that humans do perceive motion of auditory stimuli, it should not be surprising that this perception can be mapped to some area(s) of the brainstem and cortex. However, none of these studies to date have, on their own, corroborated the theory of visual "pop-out" in the auditory system. The visual system is finely honed to focus one's attention on a moving object. This attention-grabbing ability is the lifeblood of many predators. But also consider the

cat-and-mouse scenario in the dark. An animal without acute vision in the darkness of the night must rely on other senses to detect a looming predator. It is surprising that the pop-out phenomenon has not been observed to date in the auditory system of humans, given its potential biological significance. This study attempted to address some shortcomings of the previous literature on this issue. But even with this, much more research is needed in the area of auditory attention as it relates to motion perception. It is worth mentioning that it was common in early studies of cortical representation of sensitivity to auditory motion that motion was actually "simulated" rather than using an actual moving sound source (T. Griffiths, Bench, & Frackowiak, 1994; Stumpf, Toronchuk, & Cynader, 1992). There are some examples where actual motion was used as the stimulus (via a rotating speaker) to study neural encoding of this acoustic feature (Ahissar, Ahissar, Bergman, & Vaadia, 1992). More common in recent studies has been the use of KEMAR-recorded motion, or to use each participant's actual head-related transfer functions which are then used to re-create the perception of auditory motion under headphones while the listener's brain is imaged in some way (MRI, PET, etc.). While using simulated motion was a reasonable approach at the time, it certainly seems most valid to study motion perception using actual moving stimuli.

Perhaps the simplest measure of our auditory perception of motion is the minimum audible movement angle (MAMA). The MAMA is a measure of the smallest angular distance a sound source needs to traverse (at a given velocity) in order for a listener to either distinguish whether the sound was stationary or moving, or determine the direction of motion (J. D. Harris, 1972; J Donald Harris & Sergeant, 1971). The auditory system is less sensitive to changes in location of a continuously moving stimulus (thresholds can be as low as 2-4 degrees: (Chandler

& Grantham, 1992; Grantham, 1986; J Donald Harris & Sergeant, 1971; Perrott & Musicant, 1977; Perrott & Tucker, 1988) than to an abrupt shift between two otherwise stationary locations (thresholds around 1 degree: Mills (1958). It is clear that speech understanding is improved when the target and distracter(s) are presented from different but constant spatial locations. But since our ability to differentiate spatial positions is poorer when a sound source is in motion than when it is stationary, it is reasonable to wonder to what extent are we able to understand speech when spatial separations between talkers are manipulated.

Motion Perception: speech understanding

Allen et al. (2008) utilized the Coordinate Response Measure (CRM, Bolia et al. 2001) as the speech stimulus to examine the effect of changing the spatial configuration of talkers in the middle of a sentence. The CRM consists of sentences that follow the convention "Ready [callsign], go to [color] [number] now." An example sentence could be "Ready Ringo, go to blue three now". In the Allen et al (2008) study, the participant was asked to listen for a specific callsign (that of the target talker) and respond with the color and number spoken by that talker. They tested conditions in which the target talker and two distracter talkers: 1) were spatially separated (±30^{o}) throughout the sentence; 2) were co-located throughout the sentence; 3) switched positions in mid-sentence from co-located to separated (±30^{o}) (i.e. start-co-located); and 4) switched positions in mid-sentence from separated (±30^{o}) to co-located (i.e. start-separated). In conditions 3 and 4, talkers switched locations after the callsign but before the color-number occurred. The construction of these sentences was critical for that study in that presenting the target callsign towards the beginning of the utterance gave the listener an opportunity to identify the target talker's voice amongst the distracters prior to any changes in

talker location. Listeners were thus tasked with identifying the target talker and maintaining focus on that voice even after changes in the overall spatial configuration of talkers occurred. Listeners demonstrated a significant release from masking in the start-separated (3.6 dB), separated (12 dB), and start-co-located (11 dB) conditions, as compared to the co-located control conditions. Perhaps the most interesting finding was the 3.6 dB benefit in the start-separated condition because it demonstrated that the listener was able to identify the target when it was spatially separated from the distracters, and then maintain auditory stream segregation (see Shinn-Cunningham (2008) for review) when all the talkers become co-located.

Instead of using an instantaneous change in position, Davis, Grantham, and Gifford (2016) utilized continuous motion of either the target or distracters to evaluate the effect of dynamic spatial configurations on speech understanding. Like Allen et al. (2008), Davis et al. (2016) evaluated the extent to which an initial separation of talkers could allow a listener to identify a target talker and maintain that attention even when the spatial separation went away. In one condition most analogous to the "start-separated" condition in Allen et al. (2008), two distracters were presented from 0^0 azimuth, and the target started at 60^0 to the listener's left and moved at constant velocity (51^0/s) clockwise toward the distracters. In this case, the three talkers became co-located for a moment around the time that some key words were presented. The same type of motion paradigm was also applied to a stationary target presented from 0^0 azimuth and the two distracters started at the listener's left and moved clockwise. These motion conditions were repeated in the other direction (talker started in front of the listener and moved counter-clockwise). These motion conditions were compared to two control conditions in which all three talkers were presented from the same loudspeaker (either in front

of the listener or to the listener's left). These locations corresponded to the point where the talkers became co-located during the moving conditions. Finally, in order to make a more direct comparison to the results from Allen et al. (2008), a "Switch" condition was created where the target started separated from the distracters by 60 degrees, and instantly switched positions to become co-located with the distracters about mid-way through the utterance.

Their first experiment suggested that motion was a helpful cue as speech understanding was significantly improved in the motion conditions (mean performance = 64% correct) compared to the co-located control (mean performance = 31% correct). In addition to the benefit found when the target talker was in motion (mean performance = 69% correct), there was a smaller benefit when the target was stationary and the distracters were in motion (mean performance = 57% correct). Taken together, the observed benefit of motion of either target or distracters suggested that listeners were able to use motion as a perceptual cue to separate the target from distracters. Performance in the "Switch" condition was significantly poorer than the comparable motion condition.

One potentially problematic aspect of the study design of Experiment 1 in Davis et al. (2016) was that the motion path of passed over the stationary talker(s), such that co-location only occurred for an instant. As such, there was almost always some amount of spatial separation between talkers. It was difficult then to separate out the potential effects of motion vs even small spatial separations of talkers. In a follow-up experiment by Davis et al. (2016), three conditions were tested in an attempt to separate out the potential effects of small spatial separations from the motion itself. The new conditions included "Motion-Stop" where the target started at the listener's left and moved clockwise until it became co-located with the

distracters, at which point the motion stopped rather than passing by the distracters, and two conditions in which the target started at the listener's left and instantaneously changed position to either 10 degrees left of center ("Switch-Ten"), or co-located with the distracters. This experiment also included the co-located stationary control condition. There was essentially no benefit in the "Motion-Stop" condition compared to the co-located condition, and the most benefit in the "Switch-Ten" condition, where the target was never co-located with the distracters. The authors concluded that "small momentary spatial separations during key words, whether presented from stationary or moving targets led to consistently better performance than static, co-located key words". The current study was designed to address this shortcoming of Davis et al. (2016) by creating motion conditions in which the target remained spatially separated from the distracters. This was accomplished by having target and distracter stimuli be spatially separated by a large amount even in the stationary condition. In this way, we investigated the extent to which motion of a target may provide benefit above and beyond spatial separation when sufficiently separated from distracters.

Pastore and Yost (2017) studied of the effect of motion on speech understanding in multi-talker settings by specifically testing for the presence of the auditory pop-out effect. Unlike Allen et al (2008) and Davis et al. (2016), who used CRM stimuli lasting about 2 s, Pastore and Yost (2017) used a brief single word to assess whether short-duration motion could aid in speech understanding. It was hypothesized that using short-duration stimuli may help avoid utilizing "conditioning" or streaming of the target at one point in the utterance to better understand it at another point. Listeners were tasked with repeating the target word which was spoken by a female in the presence of 2 or 4 masker words spoken by males. Conditions tested

included a co-located condition, one with the target and maskers stationary but spatially separated, and one where the target word was moved 40 degrees in between two maskers. The two maskers were either presented co-located with the target, or were located at ±45 degrees. In the single motion condition, the target was amplitude panned between ±20 degrees. They found that performance was no better in the motion condition than when the target was stationary between maskers. In addition, performance in the motion condition was no better than a condition called "static pan" in which the target was stationary and located near one of the maskers. The static pan condition was like the target-center condition except the target was offset from center by about 20 degrees (via amplitude panning). Based on these findings, the authors concluded that motion is not a helpful cue for speech understanding.

Pastore & Yost (2017) acknowledged that use of single-word stimuli may have precluded perception of motion. As a control experiment, they asked listeners to simply identify the direction of motion. They found that listeners were able to identify the direction of motion with high accuracy and thus concluded that the lack of motion benefit could not be explained by an inability to detect it. Indeed, the duration of stimuli in this study (750 ms) were at least twice as long as what Grantham (1986) determined to be the minimum integration time for his participants (150-300 ms). However, it is not clear whether the minimum integration time is different for speech stimuli, particularly in the presence of competing speech.

In addition to the short stimulus used by Pastore & Yost (2017), a few other factors may have contributed to a lack of observed benefit of motion. A female target and male maskers were used, which has been shown to provide a robust release from informational masking (Douglas S. Brungart, 2001). It is feasible that the release from informational masking afforded

by using different-gender talkers left little remaining informational masking to be released by motion. A study design feature that Davis et al. (2016) and Pastore & Yost (2017) had in common was the co-location, or close proximity, of target and maskers. In contrast, the present study used wider separation to minimize any effects of informational masking based on close proximity of multiple talkers.

When testing moving auditory stimuli, parameters such as duration and velocity must be considered, as both are well known to affect our thresholds of detection (Grantham (1986); Perrott & Tucker (1988)). Although the movement durations used by Davis et al. (2016) and Pastore & Yost (2017) were quite different (~2 s vs. 0.75 s), the velocities were nearly identical (51 and 53 deg/s, respectively). Saberi and Perrott (1990) used broadband clicks to test motion sensitivity for a large range of velocities and showed that MAMA thresholds were less than 6 degrees of arc for velocities below 100 deg/s, and fairly consistent at velocities below 20 deg/s. Carlile and Best (2002) measured thresholds for changes in velocity based on three reference velocities using virtual auditory space. Using a two-interval forced-choice procedure, listeners were asked to report which interval contained the higher velocity of a broadband noise. Listeners were able to detect velocity differences of 9.1 and 14.8 deg/s for reference velocities of 30 and 60 deg/s, respectively. Locke et al. (2016) partially replicated those data by testing both discontinuous and continuous stimuli. They found that when presenting listeners with a stimulus that instantaneously changed velocities, average just noticeable differences (JND) to decreases in velocity were about 20 deg/s for both 30 deg/s and 60 deg/s reference velocities. Interestingly, listeners had significantly larger average JNDs for increases in velocity for both reference velocities (about 60 deg/s). Although a variety of methods and stimuli were used for

these studies, it seems clear that the auditory system is reasonably equally sensitive to velocities in the range of 30-60 deg/s.

Aims and Hypotheses

In the present study, we investigated new motion conditions that better isolate the effects of motion from potential confounding factors such as spatial separation and short utterance length. We tested speech understanding using sentence-material in a three-talker environment with a variety of spatial configurations for the target and distracters in order to address questions about the extent to which understanding may be positively or negatively affected by talker motion, target position relative to distracters, target spacing from distracters, greater spatial separation at the beginning or end of an utterance, and distracter motion. Test conditions using either motion of a target talker or a distracter talker provided insight to whether a pop-out effect, if present, could be harmful by guiding a listener's attention to the wrong talker. Unlike previous work using sentence material, the current study limited the spatial approximation of talkers so that potential effects of motion could be observed and not contaminated by effects of spatial release from masking.

It was hypothesized that performance would be better in a motion condition than in a comparable stationary condition. It was suspected that listeners would be better able to allocate attention to a target talker directly in front of them than positioned on the side. It was also hypothesized that positioning the target closer to a distracter would result in a decrease in performance compared to the baseline amount of separation. Finally, it was not clear whether greater spatial separation of talkers at the beginning or end of the utterance would be more

beneficial, since key information is contained in both segments of the stimuli used in this study. However, Allen et al. (2008) did report better performance in the "start co-located" condition than the "start-separated" condition, suggesting that spatial separation at the end of the utterance may be more important, at least for CRM stimuli.

Following Experiment 1, sound level measurements were conducted to assess possible contributions of SNR for each condition. These measurements served to validate the interpretation of results from Experiment 1, and to motivate a new approach for Experiment 2. Instead of viewing motion in and of itself as a potential cue for perceptual pop-out, it was considered in terms of a means for improving the overall SNR, particularly when the target was positioned with both distracter talkers on one side. Much like a listener might rotate her head slightly to direct one ear towards the target talker, it was hypothesized that an overall change in SNR brought on by the simulated head turn might improve speech understanding compared to a stationary control. Additional motivation for Experiment 2 is discussed below in Chapters 3 and 4.

CHAPTER 2

Experiment 1

Based on the presented literature, it was hypothesized that motion would be a helpful cue for speech understanding in a cocktail party listening environment. In that case, we expected to see certain patterns of results. For the case of target positioned in front of the listener, we expected the worst performance with the target stationary in the center of two distracters, with equivalent or worse performance with the target stationary but offset from center, and the best performance with the target moving between two distracters. A finding that performance in a target-motion condition is better than a stationary condition in which the target is offset from its "normal" position would be compelling evidence that motion itself is a helpful cue for speech understanding. Such a finding would rule out the possible explanation that an improvement in speech understanding in the motion condition had simply been caused by a momentary change in the signal-to-noise ratio, brought on by a moving talker's increasingly lateral position. Interestingly, if motion provided any sort of "pop out" effect or otherwise drew one's attention, then we would also expect performance to be worse if a distracter was in motion than when all three talkers were stationary, or the target was in motion.

In contrast, if motion is in some way harmful to speech understanding, then the best performance would be expected when the target is stationary and centered between two distracters, followed perhaps by the target stationary but offset, and then the worst performance in the moving target condition. It was more difficult to predict what sort of effect

a negative impact of motion would have in the case of distracter motion. It was possible that if motion does not “pop out”, listeners could still be equally able to attend to a stationary target and would thus demonstrate equal performance in distracter motion and target stationary conditions. Perhaps more likely was that listeners would be perceptive of distracter motion and would inaccurately attend to those key words, resulting in poorer performance than in a comparable target-stationary or target-motion condition.

Methods

Participants

Eight young adults (range = 22-33 years, mean = 25.5; one male) were recruited for this experiment according to Vanderbilt IRB protocol #190756. All participants had their hearing screened at 25dB HL at octave frequencies from 250-8,000Hz to confirm hearing within normal limits in both ears. Screenings were completed with a Grayson-Stadler GSI-61 diagnostic clinical Audiometer and TDH-39 headphones, in a double-walled sound booth. Participants read and signed consent forms approved by the Vanderbilt Institutional Review Board at the very beginning of their session.

Test Environment

The experiment was conducted in an anechoic chamber (4.65 x 6.55 x 7.47 m tall, measured from tips of 0.71 m deep foam wedges) with a horizontal circular array of 64 loudspeakers (Meyer MM-4) spaced at 5.625 deg. Participants sat at the center of the 1.95 m radius circle, at ear level with the loudspeakers, and entered responses via a hand-held computer number pad. Testing was controlled by a MacPro3,1 computer using a custom Matlab

(version R2017B) program. Multi-channel digital audio was delivered over gigabit ethernet (Audinate Dante protocol) via PCIe card (Focusrite Rednet PCIe) and Cisco SG200-18 managed switches to a bank of eight Dante-equipped 8-channel Ashly ne8250pe amplifiers (one channel per loudspeaker). The CRM stimuli were imported at their native sampling rate of 40KHz and up-sampled to 48 kHz. The spatial position of a sound source was simulated by direction and sound level information. The participant's head was not restrained, but they were instructed to face a single loudspeaker during stimulus presentation. Participants were monitored on a close-circuit TV for compliance with this instruction. Participation consisted of one visit of about 3 hours. Participants were compensated for their time.

Stimuli

Speech understanding was assessed with the coordinate response measure (CRM) corpus (Bolia, Nelson, Ericson, & Simpson, 2000). See Figure 1. The corpus contains sentences spoken by eight different talkers (4 female, 4 male). All sentences follow the structure "Ready [Callsign], go to [Color] [Number] now. Combinations of eight callsigns ("Arrow", "Baron", "Charlie", "Eagle", "Hopper", "Laker", "Ringo" & "Tiger"), four colors (red, white, blue, & green), and eight numbers (1-8) create 256 unique sentences for each talker. On each trial, a gender was selected at random, and then three of the four talkers of that gender were selected. In doing so, trials were comprised of only female talkers or only male talkers. Each participant was randomly assigned a target callsign at the beginning of their participation, and they used that single callsign for the duration of their listening. On each trial three CRM stimuli were presented at the same time, one with the target call sign and two with other call signs. Each stimulus had a unique color and number. For example:

Target: “Ready Ringo go to blue seven now”

Distracter 1: “Ready Baron go to red three now”

Distracter 2: “Ready Hopper go to white five now”

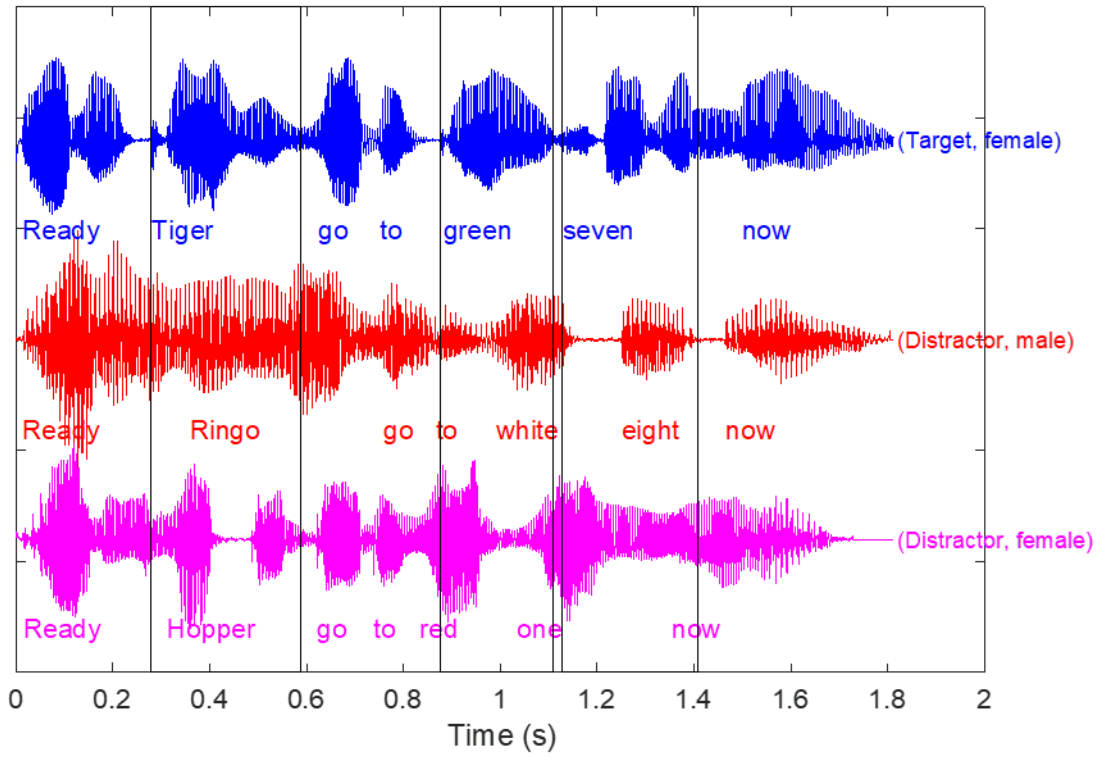

Figure 1: Illustration of three Coordinate Response Measure (CRM) stimuli presented concurrently, the target stimulus (blue, call sign is “Tiger”) and two distracter stimuli (red “Ringo”, magenta “Hopper”). The stimuli are offset vertically for clarity. Text words (“Ready”, “Tiger”, etc.) are positioned at the approximate beginning of each waveform segment where that word starts. (Technical note: In the CRM stimulus set, the signals shown in blue, red, and magenta are from talkers 5, 2, and 6, respectively.)

After the stimuli were presented the participant attempted to identify the color and number associated with the target call sign, using a computer keypad. A response was considered correct only if both the correct color and number of the target were provided. We estimated chance level to be as low as $\frac{1}{4} \cdot \frac{1}{8} = .03125$ for pure guessing of color and number, but on some trials it could be as high as $\frac{1}{3}$ if the participant clearly heard one of the color/number pairs but could not associate it with a call sign. Participants were instructed to respond with the color and number spoken by the target via a small keypad. Four buttons on the keypad were modified with a small colored square corresponding to the four possible color choices. Feedback was not provided. The distracter talkers were presented at 59dB SPL. The signal to noise ratio (SNR) of the target talker relative to the two distracters was adaptively varied within a block of trials.

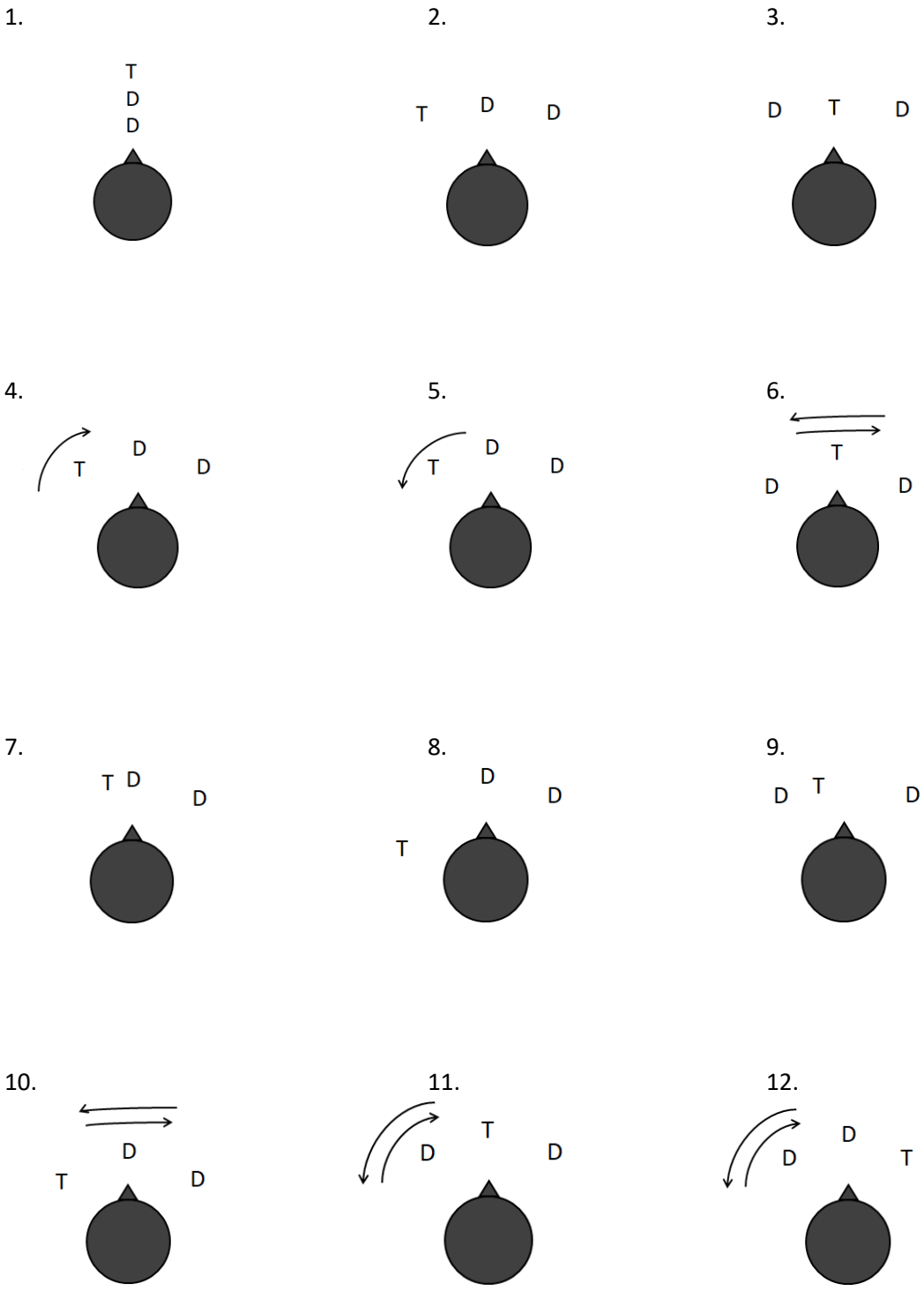

Figure 2: Stimulus conditions for Experiment 1. For all asymmetric conditions, a mirror opposite condition was also run.

Procedure

Twelve stimulus conditions (see Figure 2) were tested by concurrent, randomly interleaved threshold seeking algorithms. Each stimulus condition had a fixed number of 50 trials (so 600 trials in all), and the entire set of trials across all conditions occurred in a random order. Pilot work suggested that mixing conditions in this way prevented listeners from settling on listening strategies for particular conditions. Each of the 12 conditions in Experiment 1 were presented 50 times per run, and each participant did two such runs, each with independent threshold estimates for all the conditions. Participants were consistently able to reach asymptotic performance with 50 trials per condition (see Figure 3 for pilot data example). Threshold estimates from each participant's two runs were averaged for the purposes of analyses. Rest periods were offered every 100 trials to help minimize fatigue.

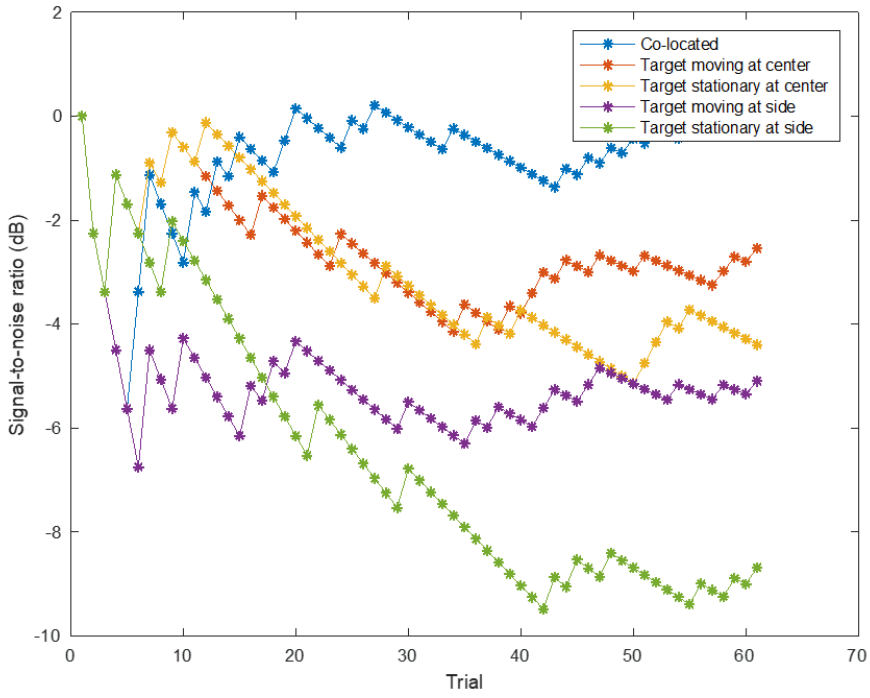

Figure 3: Illustration of testing multiple conditions concurrently in a single participant. Each of the five conditions was tested for 60 trials (the final stimulus levels are what would be used on a 61st trial). The total set of 5 x 60 = 300 trials was done in random order. This session took approximately 20 minutes. This illustration is from pilot work, so the conditions do not map exactly or fully to the set of conditions described in Figure 2.

The stimulus conditions of Experiment 1 were designed to assess how motion of the target or a distracter affects speech understanding in a multi-talker setting. In Condition 1 all three stimuli came from the same location, randomly chosen as 45 deg to the left, 0 deg (center), or 45 deg to the right. In this and all other conditions, these nominal directions were collectively jittered by up to 20 degrees in either direction to reduce the predictability of locations of talkers. The co-located condition was expected to be very difficult – it provides a baseline and a standard for comparison with previous findings. In Conditions 2 and 3 all three stimuli remained stationary, each at one of the positions 45 degrees left, center, or 45 degrees right. The target was at 45 degrees left or right (Condition2) or at center (Condition 3). The set of locations was jittered by up to 20 degrees in either direction, but the separation of 45 degrees between adjacent stimuli was maintained. These conditions provide the basic stationary reference against which motion conditions were evaluated. Conditions 2 and 3 were originally combined into a single condition, but they were split out because pilot work showed better performance with the target at center than at a side.

Conditions 4, 5, and 6 involved motion of the target, either at the side (4, 5) or at center (6). In order to avoid spatial overlap of the moving target with a stationary distracter, the target started its motion 20 degrees to one direction of its nominal position and moved at a constant velocity of approximately 20 degrees/second across its nominal position, ending 20 degrees to

the opposite side of its nominal position. Although the motion velocity used here was lower than previous studies, the separation was intended to be consistent with the study design of Pastore and Yost (2017). If motion of the target was beneficial, then performance in these conditions should be better (lower SNR) than in the corresponding stationary stimulus conditions.

Conditions 7, 8, and 9 were stationary conditions in which the target stimulus was shifted from its "home" position by about 20 deg, corresponding to the extent of motion away from the home position in the target motion conditions (4 through 6). If the target motion conditions showed a benefit (or loss) relative to the regular stationary conditions (2, 3), then conditions 7, 8, and 9 could be brought into the analysis to assess whether the benefit (or loss) from motion is due to motion per se or to target-distracter proximity effects.

Conditions 10, 11, and 12 involve motion of a distracter stimulus, either at the center (10), or at the side (11,12). To the extent that motion draws attention to a speech stimulus, distracter motion should lead to worse performance than in the corresponding stationary conditions (2, 3). Conditions 11 and 12 are differentiated based on whether the target is adjacent to, or farther away from the moving distracter. It was hypothesized that a moving distracter would be more distracting if the target was immediately adjacent, and less so if the target was separated from the moving distracter by the other, stationary distracter.

Another reason for including distracter motion is that we did not want motion to be associated exclusively with the target stimuli. In pilot work where there was that association, we often failed to hear the target call sign in the early part of the stimulus but noted that one

talker was moving and listened for the color and number from that talker. Although participants who were not familiar with the experimental design might not have been able to use this strategy, it seemed best to avoid having all moving stimuli be targets.

A consequence of the way these stimulus conditions were set is that there was a slight overall bias of 10% toward the target stimulus being in the center (that is, in front of the listener), rather than at either side. This is summarized in Figure 4. A discerning participant might pick up on this bias, knowingly or not, and allocate more attention frontward. And this could play on top of a general attentional bias toward the center or "in front of me" region. This complication was a cost of our interest in having separate stimulus conditions with the target at center vs. side. However, there was little reason to be concerned about this. First, it was doubtful that this potential bias would occur, given the random mixing of conditions from trial to trial and the fairly rapid pace participants were anticipated to maintain. Second, if there was a bias effect, performance should be better when the target is at the center than at the side, but if anything, our pilot work suggested the opposite (e.g., see Figure 3). If that pattern held, then it could not be explained by a bias that favors targets at the center.

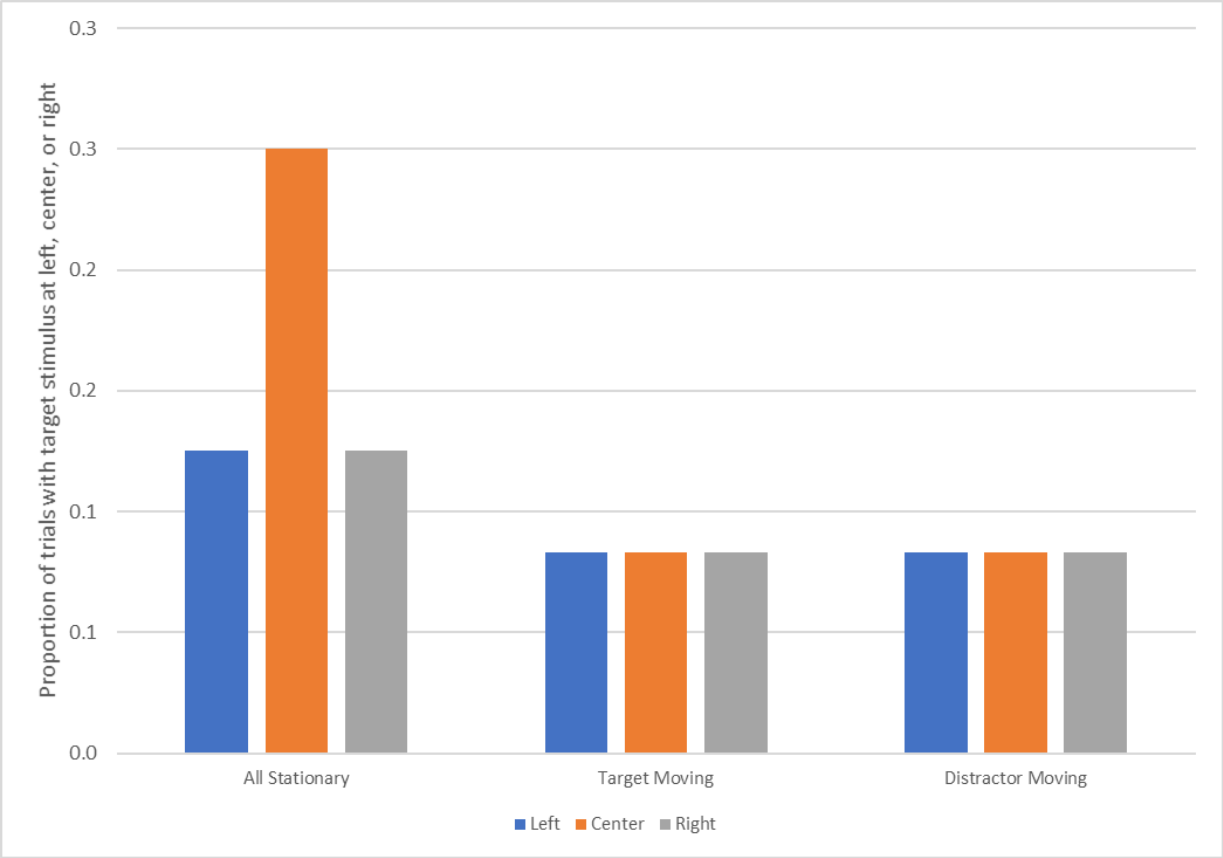

Figure 4: Distribution of trials on which the target stimulus is located at left, center, or right positions, by type of trial (all stimuli stationary, target moving, distracter moving) for Experiment 1. Overall, the target is at center on 41.7% of trials and at each of the two sides on 29.1% of trials. Thus, the target stimulus was about 12% more likely to be at center than at either side.

Prior to data collection, each participant ran through a practice session lasting about 5 minutes. After being familiarized with the task, they were presented with a shortened version of the full program that included four presentations of all 12 conditions, in random order, with a starting SNR for each condition of +4 dB. Following this practice, participants were asked if they were comfortable with the task and were ready to proceed, or if they would like to hear the practice again (none did so).

The Kesten (1958) stochastic procedure is a general algorithm for finding the value of a control variable that has a designated probability of a binary outcome. For example, in industrial production the control variable might be something that affects how a piece of metal resists deformation, and the outcome could be whether the piece deforms or not when a certain force is applied to it. In psychophysics applications, the control variable is the intensity of a stimulus and the outcome variable is whether a participant makes a correct response (e.g., in a detection, discrimination, or identification task). Similar to the commonly used adaptive staircase procedure (Levitt, 1971), the stochastic approximation method adaptively varies the SNR on a trial-by-trial basis. However, it differs from the staircase procedure in several ways.

As applied to the present experiment, we can specify the following variables:

S stimulus level (signal-to-noise ratio, higher is "easier")

R response on a given trial, coded as 0 if incorrect or 1 if correct

C a constant that affects how rapidly the stimulus level changes, we used 9

M a count of changes in the direction of stimulus level change (this is often called mshift)

p the probability of a correct response that the algorithm tracks to, we used 0.75

t trial number

The stimulus level was set to *S* = +2 dB for the first trial. After each trial the following three rules were applied:

If $t \leq 2$ then $S_{t+1} = S_t - \frac{C}{t}(R_t - p)$ [1]

If $t > 2$ then $S_{t+1} = S_t - \frac{C}{2+M}(R_t - p)$ [2]

If $t > 1$ and $R_t \neq R_{t-1}$ then $M = M + 1$ [3]

Kesten (1958) proved that this algorithm converges on the value of *S* for which *p* has a specified value, such as 0.75. In extensive pilot work, as well as computer simulations, we found that the algorithm converges well within about 50 trials. Various ways of estimating a threshold from such a run of trials were considered, and the simple approach of using the final stimulus level was adopted. Strictly speaking, threshold was defined as the signal-to-noise level that would have been used on the 51st trial (even though only 50 trials were run). The mathematical logic underlying this choice is Kesten's demonstration that, in the long run, the algorithm converges on the "true" stimulus level.

As was noted earlier, separate stochastic runs of 50 trials each were conducted for the twelve stimulus conditions, for a total of 600 trials. Each participant had a different random order of the 600 trials. Thus, the stimulus conditions were essentially random from trial to trial, rather than being presented in blocks. This was to discourage participants from attempting to work out condition-specific listening strategies.

Results

Individual threshold data and group means are provided in Table 1.

Condition	Run	Participant 1	2	3	4	5	6	7	8	Mean	Mean of Runs 1,2
1	1	1.86	4.29	1.03	1.38	-0.97	1.75	1.63	1.89	1.61	1.41
	2	1.21	1.66	1.51	0.47	0.41	2.23	1.12	1.06	1.21	
2	1	-3.03	-1.25	-8.10	-4.97	-3.66	-10.57	-4.47	0.20	-4.48	-4.72
	2	-4.88	-6.05	-12.37	-2.43	-6.19	-1.11	-4.53	-2.13	-4.96	
3	1	2.02	2.02	-1.29	-0.96	-0.91	1.40	-1.91	-4.11	-0.47	-0.82
	2	-0.86	2.21	-1.21	-1.48	0.12	-1.89	-1.60	-4.73	-1.18	
4	1	-3.50	0.62	-11.24	-2.36	-4.07	-3.71	-2.24	-1.01	-3.44	-3.99
	2	-4.67	-5.30	-8.95	-5.17	-4.22	-1.72	-3.15	-3.14	-4.54	
5	1	-4.68	-1.71	-11.85	-2.99	-4.28	-5.93	-1.18	-4.02	-4.58	-4.40
	2	-3.10	-3.48	-8.24	-3.71	-4.25	-3.29	-4.04	-3.58	-4.21	
6	1	-0.64	0.88	-2.35	-0.79	-0.56	-0.83	-1.80	-5.46	-1.44	-1.29
	2	-0.43	1.28	-1.67	-0.04	-0.16	-2.56	-0.85	-4.67	-1.14	
7	1	-2.29	-0.92	-5.76	-3.16	-2.21	-2.28	-1.83	-3.18	-2.70	-3.20
	2	-1.68	-3.60	-6.47	-3.83	-4.73	-2.25	-4.88	-2.15	-3.70	
8	1	-4.17	1.05	-6.19	-3.40	-5.25	-2.94	-2.02	-4.42	-3.42	-3.43
	2	-2.20	-0.71	-8.61	-3.56	-5.14	0.13	-4.86	-2.66	-3.45	
9	1	-0.45	2.72	0.27	1.20	0.23	0.17	-1.74	-5.56	-0.40	-0.79
	2	-0.78	1.18	-1.31	-0.47	-0.80	-0.87	-2.13	-4.29	-1.18	
10	1	-1.28	1.47	-7.18	-3.72	-4.07	-0.78	-2.41	-1.12	-2.39	-3.18
	2	-1.52	-4.35	-9.43	-2.45	-4.66	-3.49	-5.40	-0.45	-3.97	
11	1	-0.05	1.45	-0.03	-2.41	-1.47	1.99	-2.38	-2.82	-0.72	-0.38
	2	-0.36	1.84	0.06	-0.33	1.25	-0.34	-0.69	-1.72	-0.04	
12	1	-1.76	-0.25	-7.40	-2.67	-5.98	-5.79	-1.02	-1.30	-3.27	-3.85
	2	-2.93	-5.72	-8.50	-4.68	-6.32	-2.18	-3.72	-1.34	-4.42	

Table 1: Experiment 1, Individual participant threshold estimates and group means by listening condition and run. Threshold estimate units are signal to noise ratio in decibels (dB). See Figure 2 for description of listening conditions.

In a repeated measures analysis of variance with Condition (12 conditions) and Run (2 runs) as factors, there was a significant effect of Condition, $F(1,2.27) = 11.632$, $p < .001$, $\eta^2_{partial} = 0.624$, degrees of freedom per Huynh-Feldt correction for sphericity. Comparisons between listening conditions are described later in this results section. Neither the Run effect

nor the interaction were statistically significant. However, there was a small tendency toward better performance on the second run than the first. Averaged across listening conditions, the means and (standard errors) of the threshold signal to noise ratios were: Run 1, -2.141 dB (0.573), Run 2, -2.632 dB (0.441). This average improvement of about 0.5 dB across runs presumably indicates a modest practice effect.

Test-retest reliability was assessed using the intraclass correlation coefficient, which draws on findings from the analysis of variance. Figure 5 shows the entire data set in scatterplot form comparing the two runs. This figure illustrates that, despite a small improvement on the second run, the effects of listening condition and individual differences between participants were quite consistent across runs. The intraclass correlation coefficient was ICC = 0.846, which is considered very good by conventional ratings for ICC. This coefficient was computed using the ICC(A,k) model as described in McGraw and Wong (1996). Specifically:

$$ICC = \frac{MS_{Participant} - MS_{error}}{MS_{Participant} + \frac{MS_{Run} - MS_{error}}{n}} \quad [4]$$

where the MS_{error} term is the Participant x Run interaction. Based on these descriptive and statistical analyses, we considered the reliability across runs adequate. Accordingly, the contrasts between listening conditions reported below were based on each participant's average score across the two runs in each listening condition.

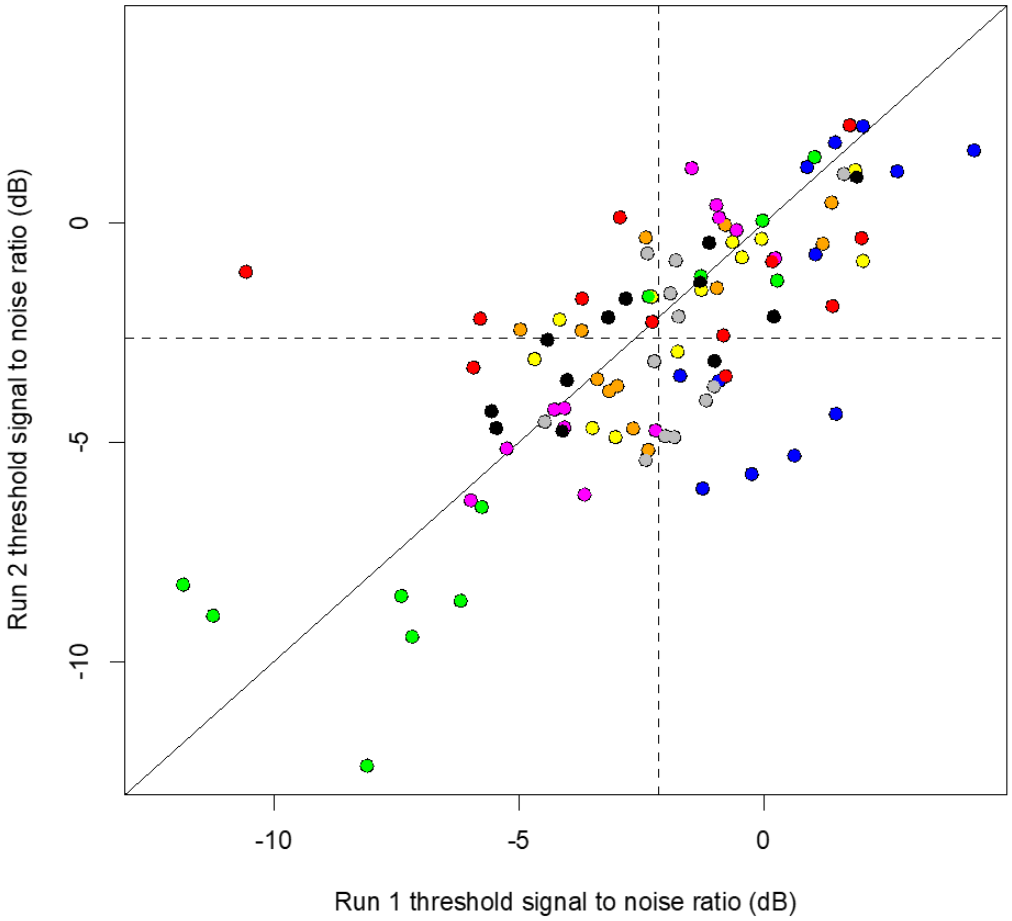

Figure 5: Experiment 1, individual participant thresholds on run 1 and run 2 within stimulus conditions. The solid diagonal line indicates where points would lie if the thresholds for a given participant and condition were identical on runs 1 and 2. The dashed lines show mean thresholds (across all participants and conditions) on runs 1 and 2. Individual participants are color coded: 1 - yellow, 2 - blue, 3 - green, 4 - orange, 5 - magenta, 6 - red, 7 - gray, and 8 - black. On average, thresholds were 0.491 dB lower (better) on run 2 than run 1.

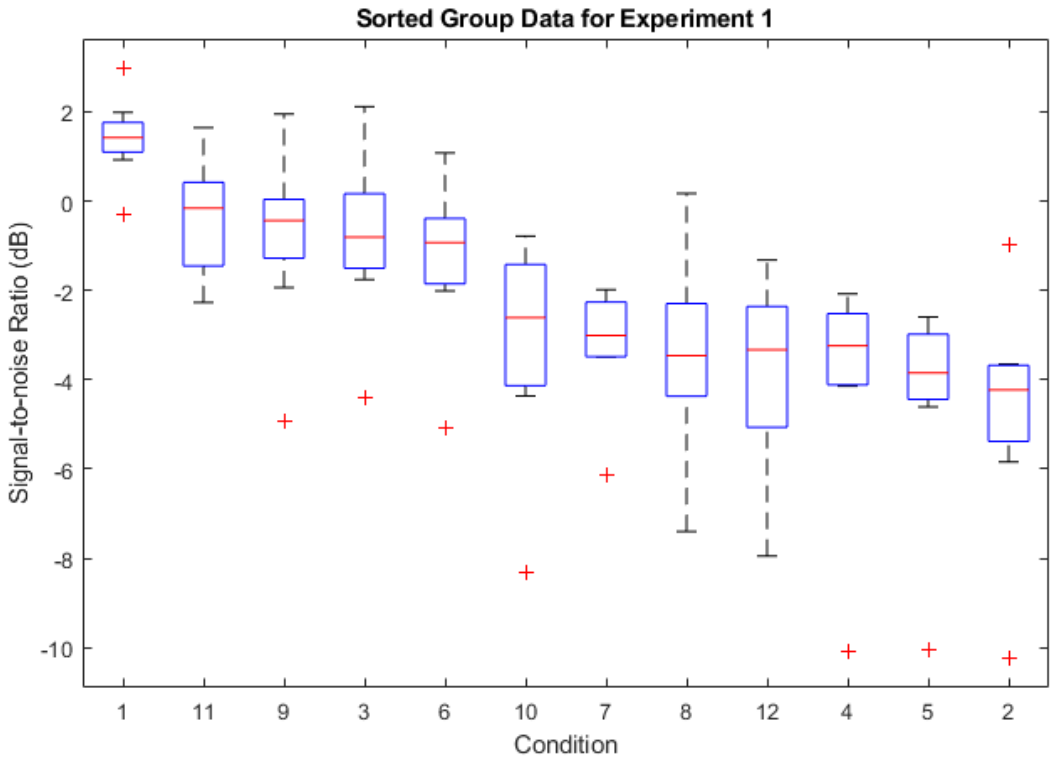

Figure 6: Group results for Experiment 1. Signal-to-noise ratio (dB) is plotted on the y-axis for each of the 12 conditions. Conditions are sorted on the x-axis from highest to lowest mean SNR. Red lines indicate the median for each condition, while boxes denote the 25th and 75th percentile and whiskers denote the 5th and 95th percentiles. Red plus signs indicate outliers for a given condition. From the left, Condition 1 involved co-location of all talkers. Conditions 11, 9, 3, and 6 involved a talker located between the two distracters, and the remainder involved a target located to one side or the other of the distracters.

Group mean performance by condition for Experiment 1 is shown in Figure 6. Group means were obtained by first averaging the results of each participant's two runs for a given condition, and then calculating the mean across all participants for each condition. The group mean results of Experiment 1 are sorted from worst to best performance in Figure 6. As shown in that Figure, mean signal to noise ratios sort into three groupings: (1) co-located target and distracters (condition 1); (2) target located between the distracters (conditions 11, 9, 3, and 6);

and target located to the side of the distracters (conditions 10, 7, 8, 12, 4, 5, and 2). These groupings were supported by a set of linear mixed effects models, using the *lmerTest* analysis package in *R* with the Satterthwaite method of estimating degrees of freedom and *p* values. (There is disagreement about the merits of using *p* values with linear mixed effects models, but we regard the *p* values as providing a useful overall indication of the effects of experimental factors. Where *p* values are given, those less than .001 are "truncated" to that value.) For these models, each stimulus condition was classified according to the following factors: (1) whether the target and distracters were co-located; (2) whether the target location was to one side of or between the distracters; (3) whether the target moved; and (4) whether a distracter moved.

The first model evaluated was the co-location effect, essentially condition 1 (co-located) vs. all other conditions. This effect was significant, $F(1,87) = 27.164$, $p < .001$. The average signal to noise ratios were $M = 1.41$ dB ($SE = 0.33$) in the co-located condition and $M = -2.73$ dB ($SE = 0.27$) in all other conditions combined. The difference of 4.14 dB between these means confirms the expected benefit derived from spatial separation between target and distracters.

Further analyses omitted the data from condition 1, in order to assess factors that arise when the target and distracters are spatially separated. A model was run with main effects of target location, target motion, and distracter motion, as well as two-way interactions between target location and each of the motion factors. The only significant factor was target location, $F(1,75) = 36.304$, $p < .001$. As Figure 6 shows, the signal to noise ratio in the conditions (11, 9, 3, and 6) with the target between the distracters was higher (worse) than in the conditions with the target located to one side or the other of the distracters. The average signal to noise ratios were $M = -0.82$ ($SE = 0.30$) for conditions with the target between the distracters and $M = -3.82$

(*SE* = -3.82) for conditions with the target to one side of the distracters. The strong effect of target location, a "cost" of about 3 dB for the target being flanked by distracters on either side, was not an expected finding from this experiment. As will be explained later, this finding led us to consider how the target location affected the signal to noise ratios measured at each ear, and to design Experiment 2 to investigate this matter.

This experiment was designed primarily to test whether motion of a target speech token (or a distracter token) influences speech understanding. Table 2 summarizes a rather large set of planned comparisons to evaluate such effects. For example, Figure 2 shows that listening condition 4 had the target on the left (or right) and distracters at center and right (or left), with the target moving toward the center as the speech tokens played out. The mean threshold signal to noise ratio from this condition was compared to the mean from condition 2, which had the same relative locations of target and distracters, but no motion. As Table 2 (first row) shows, in a paired samples *t* test comparing these conditions, the difference was not significant. The mean difference between conditions was computed such that a negative value meant a benefit in speech understanding due to motion, while a positive value indicated a cost. For the comparison between conditions 4 and 2 the mean difference was 0.73 dB, so performance was a bit worse when the target moved.

	M_1	M_2	D	t	p	95% CI
Side target, motion toward center						
Conditions 4 vs. 2	-3.99	-4.72	0.73	1.599	0.163	-0.38 to 1.84
Conditions 4 vs. 7	-3.99	-3.20	-0.79	1.439	0.193	-2.08 to 0.51
Side target, motion away from center						
Conditions 5 vs. 2	-4.40	-4.72	0.33	0.652	0.535	-0.86 to 1.51
Conditions 5 vs. 8	-4.40	-3.43	-0.96	1.622	0.149	-2.36 to 0.44
Center target motion						
Conditions 6 vs. 3	-1.29	-0.82	-0.47	1.647	0.144	-1.14 to 0.20
Conditions 6 vs. 9	-1.29	-0.79	-0.50	1.927	0.095	-1.12 to 0.11
Side distracter motion, target at center						
Conditions 11 vs. 3	-0.38	-0.82	0.45	1.291	0.238	-0.37 to 1.27
Conditions 11 vs. 9	-0.38	-0.79	0.41	0.942	0.378	-0.62 to 1.45
Side distracter motion, target at other side						
Conditions 12 vs. 2	-3.85	-4.72	0.87	1.907	0.098	-0.21 to 1.96
Conditions 12 vs. 7	-3.85	-3.20	-0.65	1.310	0.232	-1.81 to 0.52
Conditions 12 vs. 8	-3.85	-3.43	-0.41	0.641	0.542	-1.94 to 1.11
Center distracter motion						
Conditions 10 vs. 2	-3.18	-4.72	1.54	3.503	0.010	0.50 to 2.59
Conditions 10 vs. 7	-3.18	-3.20	0.02	0.055	0.958	-1.00 to 1.05
Conditions 10 vs. 8	-3.18	-3.43	0.26	0.492	0.638	-0.98 to 1.49

Table 2: Experiment 1, Planned Contrasts for Motion Effects. M1 and M2 are the means for the conditions being compared, D is the difference between means, t and p are the t value and significance value for a paired samples t-test (df = 7, two-tailed), and the 95% confidence interval is on the difference between conditions. See Figure 2 for illustrations of the listening conditions.

The comparisons in Table 2 were done non-conservatively, with no protection for multiple comparisons and recognition that some listening conditions figured into more than one comparison, in order to maximize the prospect of finding effects of motion. Nevertheless, in keeping with the linear mixed effects testing reported above, there was almost no evidence that motion mattered. Just one of the fourteen comparisons showed a significant effect, for listening conditions 10 vs. 2. The difference of 1.54 dB suggested that motion of a distracter at the center location made speech understanding worse than if that distracter was stationary. This possible motion effect must be tempered, however, by the finding that the motion condition 10 was not significantly different from two other appropriate stationary control conditions, 7 and 8 (see details in Table 2). Across all fourteen comparisons the range of differences between motion and non-motion conditions was from -0.96 dB to 1.54 dB, with an average of 0.06 dB. A retrospective power analysis (taking into account the variability of the difference scores between conditions) suggested that there was adequate power (0.80) to find differences of about ±1.3 dB. We conclude that this experiment does not provide evidence that motion of target or distracter speech tokens makes a difference in the signal to noise ratio required for speech understanding. Motion does not seem to either enhance or detract from speech understanding.

Discussion

The primary aim of Experiment 1 was to determine if motion could be a useful cue for speech understanding. In two previous studies, motion was found to have little, if any, effect on speech understanding (Davis et al., 2016; Pastore & Yost, 2017). Consistent with those reports, there was essentially no evidence that the type of motion studied in Experiment 1 had any

effect on speech understanding. All three studies used constant velocity motion in a single direction, and listeners reported being able to hear the motion in all three studies. Despite that, there was no observed pop-out effect for moving auditory stimuli. This result is discussed in greater detail in the discussion section of Chapter 4.

The finding that performance was better when the target was positioned on the side of the distracters was puzzling at first, as it was expected that listeners would be able to focus on someone talking directly in front of them. After all, that seems to be a very common and ecologically valid listening scenario. Indeed, many research paradigms are set up with the target in front of the listener for this very reason (see Bronkhorst and Plomp (1990)). However, there is an explanation for these findings in terms of how the locations of the target and distracters affect the actual signal to noise ratios at the ears. When the target is presented from in front of the listener with a distracter on either side, the two distracters each have a fairly direct path into the ipsilateral ear. In contrast, the target's path to either ear, being presented from directly in front of the listener, is somewhat obstructed by the head. On the other hand, if the target's position is swapped with one of the distracters, it is now the target with a direct path to one ear.

Another interesting finding from Experiment 1 was a lack of significant difference between the target offset conditions (7, 8, and 9) and their equally spaced comparison conditions. Given the substantial literature on spatial release from masking, including the effects reported by Davis et al. (2016), it was suspected that positioning the target 20 degrees closer to a distracter would result in a significant decrease in performance compared to 45-degree the baseline amount of separation. Condition 7 included such a set up with the target

positioned 20 degrees closer to the distracters than in condition 2. The mean difference between conditions was 1.52 dB, but a paired t-test found this difference did not quite reach statistical significance ($t = 2.35$, $p = 0.051$,). Neither of the other two offset conditions were statistically significantly different from their equally spaced comparison conditions. Like the earlier examples, it seems the actual SNR at each ear was a greater driving factor of speech understanding performance than simple spatial separation, whether talkers were in motion or stationary.

During pilot testing, we considered combining conditions 4 and 5 into a single condition. They were ultimately separated into two separate conditions so that we could evaluate the potential benefit of increased spatial separation at the beginning of the utterance (presumably giving the listener a better opportunity to hear the target callsign), compared to the end of the utterance when the listener is tasked with identifying the color and number spoken by the target talker. It was later revealed through SNR calculations (discussed in the following Chapter), that there was relatively little difference in SNR between the two stationary control conditions (7 and 8). Presumably because of this, there was no significant difference observed in performance between conditions 4 and 5 in experiment 1 ($M = 0.41$ dB; $t = 1.29$, $p = 0.238$). The spatial separations used in this experiment were admittedly greater than those used by Allen et al. (2008) or Davis et al. (2016) who both observed significant effects of co-location or close approximation (10-30 degrees) of talkers at the end of the utterances. For example, Davis et al (2016) reported significantly higher performance in their “Switch-Ten” condition where the target remained 10 degrees separated from the distracters than in the “Switch-Zero” condition where all three talkers became co-located during the color-number complex. The

larger separation used in this study and the different distracter locations are likely reasons that such an effect was not observed in this experiment. By maintaining larger separations of talkers in this study, the target remained sufficiently isolated on one side of the head compared to the distracters, such that a beneficial SNR was maintained in the near ear.

Since it was apparent from Experiment 1 that SNR of the target in different positions likely played a bigger role than motion, it became clear that it would be of scientific value to make acoustic measurements at a series of sound source azimuths. With these measurements, we were then able to calculate effective SNR for any combination of talker locations. These results were then used to assess the results of Experiment 1 by evaluating the contribution of SNR to performance in each condition, and to serve as a foundation for the conditions tested in Experiment 2. Chapter 3 presents these analyses.

CHAPTER 3

Sound Measurements

Although there are several classic papers that have made measurements of the relationship between sound source azimuth and sound level (dB) (see Bronkhorst and Plomp (1988)), it was of value for this project for us to make measurements in our lab space, using stimuli related to our experiments, and from the same azimuths used in our experiments. These measurements were valuable not only for archival purposes, but also because they allowed us to validate the hypothesis that performance in Experiment 1 was better when the target was positioned on the side of the two distracters because of a more favorable SNR in one ear. Finally, these measurements served as motivation for an additional experiment, where we tested how motion of all three talkers could allow a listener to take advantage of momentary changes in SNR to better understand speech in background noise.

Methods

Acoustic measurements were made in order to assess, as directly as possible, how the directions of talkers with respect to the listener's head could affect SNR in a three-talker environment. Recordings were made from the right ear of a Knowles Electronics Mannequin for Acoustic Research (KEMAR) fitted with a G.R.A.S. artificial ear simulator, IEC 60318-4 (711) microphone, and a G.R.A.S. microphone amplifier (PowerModule Type 12AA, Holte, Denmark). The "stimulus" was a 5-second sample of gaussian noise, filtered to approximate the long-term spectral average of the male CRM speech tokens (thanks to Wes Grantham and Ben Hornsby for providing the filter coefficients, which were based on concatenating all of the male speech

tokens in the CRM corpus, and to Todd Ricketts for assistance with the recordings). The same filtered noise sample was played through a single loudspeaker that was positioned in sequence at each of the 64 loudspeaker locations (at 5.625 deg spacing). At each location a 5-second recording was made.

Analyses

With the loudspeaker and noise sample being identical across locations, the principal basis for differences across locations in the sound arriving at the right ear was the anatomy of KEMAR's head and ears. Each recording was spectrally analyzed using Adobe Audition software (version 13.0.3). This analysis resulted in sound levels in frequency bands of 46.88 Hz from 0 through 23953.13 Hz. Because of the sample rate of the original CRM tokens, there was meaningful information only below about 12 kHz, and as described below our focus was on the frequency range below about 4 kHz. Although our experiment used both the female and male CRM tokens, it seemed sufficient to conduct these acoustical analyses based on one gender because the long-term spectral composition of adult female and male voices is quite similar for frequencies above about 250 Hz (Byrne et al., 1994). For each of the 64 locations we averaged across frequencies to obtain two sound level estimates, one each for frequencies below and above about 1500 Hz. With these estimates in hand, we could approximate what the signal to noise ratio would be for any arbitrary set of locations of a target and two distracters. This approach focuses on what we think is the long-term average signal to noise ratio. For any specific set of three CRM speech tokens, there is additional variability in the signal to noise ratio based on which talkers are involved and on differences in the speech pacing. And since the

actual stimulus used in our experiments was recorded speech and not speech-shaped noise, the normal amplitude modulations of speech caused the moment-by-moment SNR to vary widely.

Although the purpose of these sound recordings was to support estimates of signal to noise ratio, some features of the analyses have intrinsic interest, so we present them here. Figure 7 shows sound levels by frequency for each of the 64 locations (covering 360 deg in steps of 5.625 deg). The sound levels decrease dramatically beyond 4 kHz, so frequencies higher than that are not shown. Each loudspeaker location is shown as a separate line, most in gray but three highlighted in color. The black line is for the location directly in front of KEMAR. It is noteworthy that the sound level from this direction is fairly high, but not the highest. Rather, sound level is higher for directions that are somewhat rightward because they have a more direct path to the right ear. The red line is for 56.25 deg to the right, which has the highest average sound level out of all the locations, and the blue line is for 112.5 deg to the left, which has the lowest average sound level. Sounds from that general direction are reduced in intensity because of the so-called head shadow. We recorded from only KEMAR's right ear, but we assume that the pattern would be a "mirror image" from the left ear.

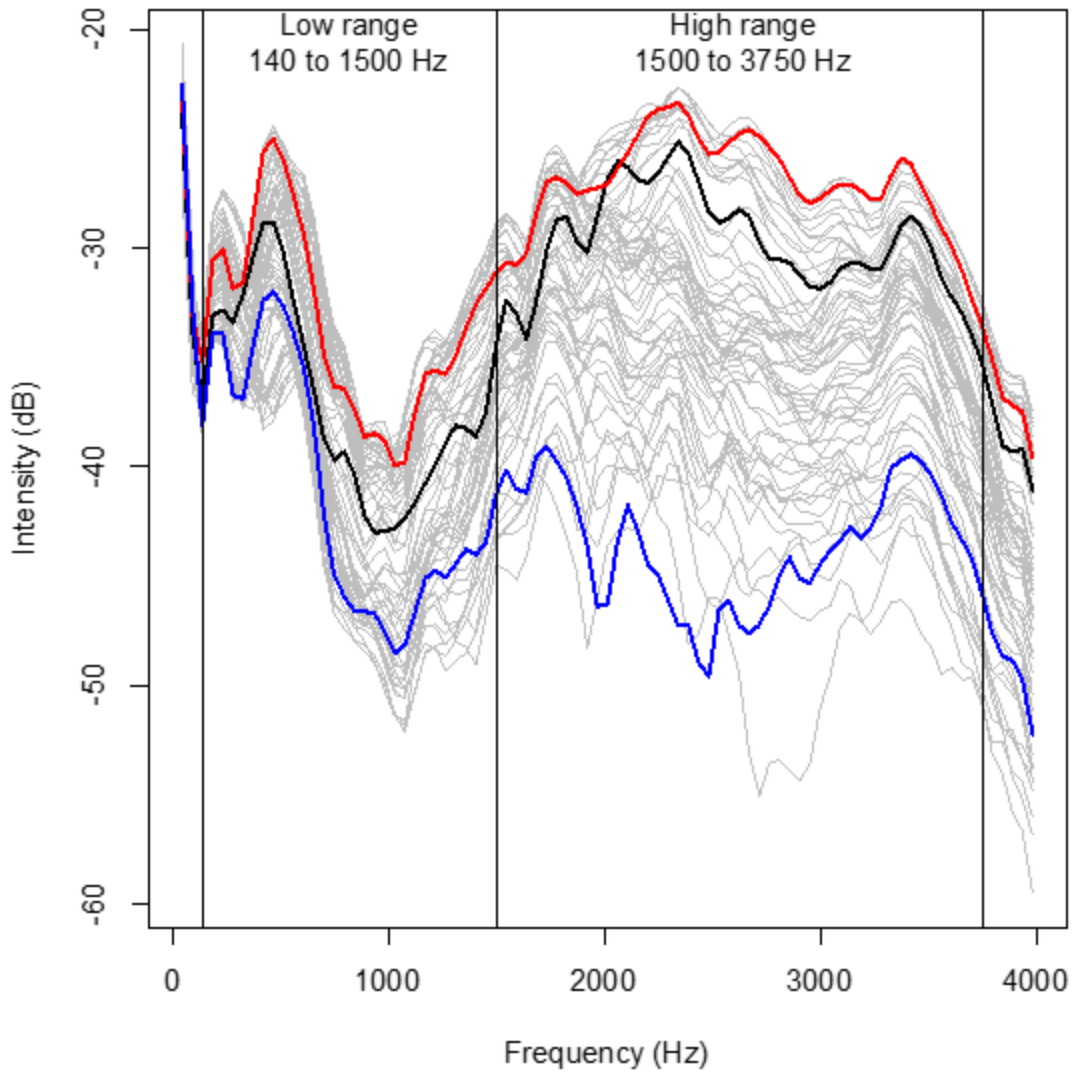

Figure 7: Sound level by frequency for each of the 64 loudspeaker locations covering a full circle, as recorded in the right ear of the KEMAR manikin. Three of the loudspeaker locations are highlighted: black is 0 deg (straight ahead), red is 56.25 deg to the right, and blue is 112.5 deg to the left. The intensity units are on an Adobe Audition scale in which 0 dB would be effectively the highest sound level that the computer sound card could support.

To simplify the signal to noise analysis, we collapsed the information shown in Figure 7 by obtaining sound level averages in two frequency ranges. As shown in the Figure, the low range was from about 140 to 1500 Hz and the high range was from 1500 to 3750 Hz. We

expected that sound level differences across locations would be larger in the high than the low frequency range. Figure 7 shows that although this was true, there were also substantial differences at lower frequencies as well. Therefore, it seemed appropriate to include a version of average sound level at two frequency ranges. Within each frequency range, the sound levels were averaged by converting the decibel units from the Adobe Audition into intensity units, computing the mean intensity, and converting the mean value back into decibel units. After this calculation was done for both frequency ranges and all 64 loudspeaker locations, the maximum sound level was found, and then all sound levels were expressed in dB units relative to that maximum. Thus, the highest sound level was 0 dB (by definition) and all others were negative (on this scale the lowest sound level was -16.52 dB).

Figure 8 shows the sound level findings in polar coordinates. It may helpful to think of this figure as an overhead view of the listening environment with the participant at center and facing upward. The directions of the plotted symbols relative to the center position correspond to the directions of the loudspeakers in the anechoic chamber. The farther away from the center a symbol is, the higher the sound level is at the right ear. Because the sound levels may have archival value, they are also provided in numerical form in Table 3.

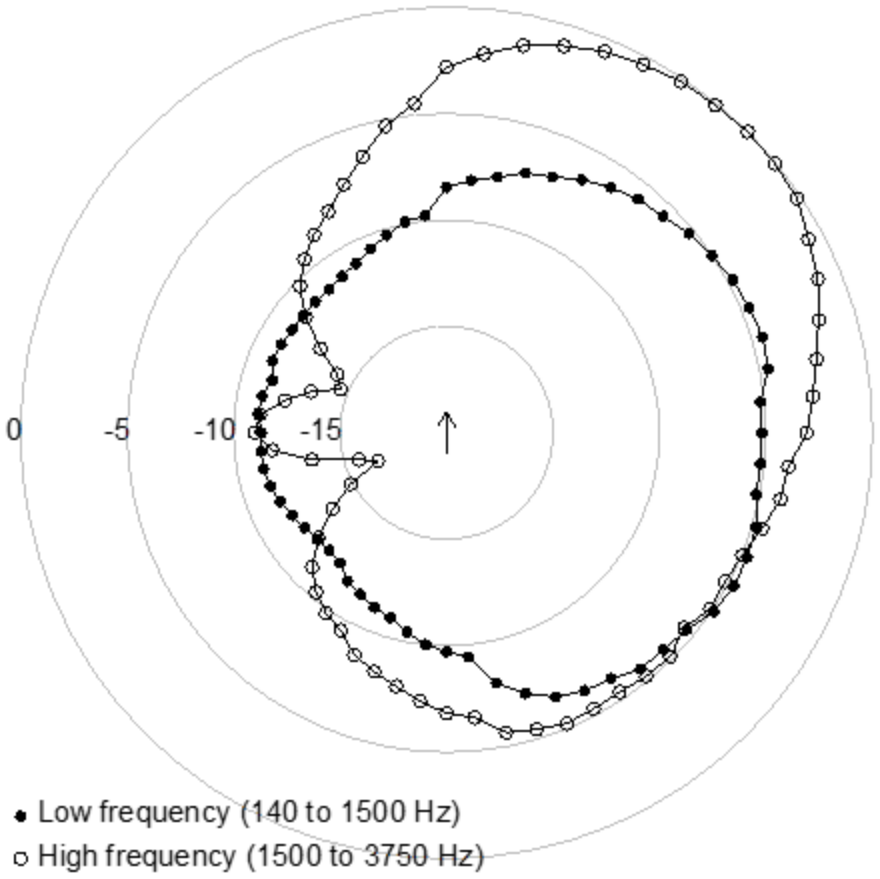

Figure 8: Average sound levels at the right ear for low and high frequency ranges, for sounds coming from each of the 64 loudspeaker locations (covering 360 deg in increments of 5.625 deg). The arrow at center indicates that the listener is facing upward or "north" in this representation. The concentric circles mark off sound levels in 5 dB increments. The farther a plotted symbol is from the center of the graph, the higher the sound level is (by definition, the highest level was set at 0 dB).

Loudspeaker number (clockwise starting at 0 deg)	Direction on 360 deg basis (increasing clockwise from 0 deg)	Direction on 180 deg basis (positive clockwise from 0 deg, negative leftward)	Low frequency average sound level (dB)	High frequency average sound level (dB)
1	0.000	0.000	-8.475	-2.823
2	5.625	5.625	-8.088	-2.117
3	11.250	11.250	-7.748	-1.457
4	16.875	16.875	-7.245	-1.012
5	22.500	22.500	-6.991	-0.611
6	28.125	28.125	-6.522	-0.396
7	33.750	33.750	-6.139	-0.165
8	39.375	39.375	-5.814	-0.096
9	45.000	45.000	-5.612	0.000
10	50.625	50.625	-5.256	-0.055
11	56.250	56.250	-5.010	-0.207
12	61.875	61.875	-4.766	-0.719
13	67.500	67.500	-4.619	-1.125
14	73.125	73.125	-4.489	-1.739
15	78.750	78.750	-4.603	-2.295
16	84.375	84.375	-5.193	-2.714
17	90.000	90.000	-5.194	-3.089
18	95.625	95.625	-5.197	-3.869
19	101.250	101.250	-5.174	-4.014
20	106.875	106.875	-4.812	-4.515
21	112.500	112.500	-4.709	-4.968
22	118.125	118.125	-4.710	-5.169
23	123.750	123.750	-4.886	-5.119
24	129.375	129.375	-5.415	-5.556
25	135.000	135.000	-5.587	-5.143
26	140.625	140.625	-5.661	-5.220
27	146.250	146.250	-6.104	-5.362
28	151.875	151.875	-6.268	-5.319
29	157.500	157.500	-6.583	-5.231
30	163.125	163.125	-7.209	-5.455
31	168.750	168.750	-8.016	-5.655
32	174.375	174.375	-9.430	-6.563
33	180.000	180.000	-9.715	-6.836
34	185.625	-174.375	-10.012	-7.336
35	191.250	-168.750	-10.463	-7.871
36	196.875	-163.125	-10.925	-8.287
37	202.500	-157.500	-11.134	-8.706
38	208.125	-151.875	-11.407	-9.469

39	213.750	-146.250	-11.619	-9.809
40	219.375	-140.625	-12.112	-10.336
41	225.000	-135.000	-12.210	-11.072
42	230.625	-129.375	-12.162	-12.212
43	236.250	-123.750	-11.981	-13.575
44	241.875	-118.125	-11.784	-14.895
45	247.500	-112.500	-11.547	-16.518
46	253.125	-106.875	-11.356	-15.691
47	258.750	-101.250	-11.247	-13.557
48	264.375	-95.625	-11.249	-11.788
49	270.000	-90.000	-11.298	-11.006
50	275.625	-84.375	-11.122	-11.185
51	281.250	-78.750	-11.167	-12.262
52	286.875	-73.125	-11.452	-13.342
53	292.500	-67.500	-11.169	-14.622
54	298.125	-61.875	-11.194	-14.182
55	303.750	-56.250	-11.286	-12.871
56	309.375	-50.625	-11.304	-11.446
57	315.000	-45.000	-11.286	-10.265
58	320.625	-39.375	-11.298	-9.470
59	326.250	-33.750	-11.166	-8.802
60	331.875	-28.125	-10.980	-8.220
61	337.500	-22.500	-10.661	-7.379
62	343.125	-16.875	-10.300	-6.440
63	348.750	-11.250	-9.904	-5.309
64	354.375	-5.625	-9.736	-4.468

Table 3: Average low and high rrequency sound levels by loudspeaker location as measured at right ear of KEMAR manikin.

For the frequency range below 1500 Hz (filled symbols) sound levels are higher on the right side than the left side by about 6 dB, and there is not much front/back difference. For the higher frequency range (open symbols) the right side sound levels are higher than the left side by about 6 to 9 dB, and within the right side, sounds from about 30 to 60 deg in front are about 5 dB higher in level than those from a similar range of directions to the rear. In summary, there are clear left-right and front-back asymmetries in sound level. Patterns like this have been reported previously, especially in the research literature on sound localization. We felt that it

was important to obtain measurements in our experimental setting and using stimuli that resemble the CRM speech tokens in spectral content. Although we did not do recordings from the left ear, it is reasonable to assume that they would have produced a mirror image of Figure 8. Indeed, in the signal to noise calculations reported below we explicitly assumed, for example, that the sound level at the left ear for a stimulus coming from 45 deg to the left is identical to what we recorded at the right ear for a sound coming from 45 deg to the right.

The sound level information summarized in Table 3 and Figure 8 can be used to explore how signal to noise ratios vary with the spatial locations of the target and distracter stimuli. For this analysis, we assume that the target and both distracter stimuli are all played at the same sound level, so that variations in the signal to noise ratio can be attributed to the locations of the stimuli. In general, signal to noise ratio is the average power of a signal relative to the average power (*P*) of noise that may be present:

$$SNR = \frac{P_{signal}}{P_{noise}} \text{ [5]}$$

This is commonly expressed in decibel units:

$$SNR_{dB} = 10log_{10}\left(\frac{P_{signal}}{P_{noise}}\right) \text{ [6]}$$

If the signal and noise have equal power, then $SNR = 1$ and $SNR_{dB} = 10log_{10}(1) = 10(0) = 0$. If the power of the signal is half that of the noise, $SNR = 0.5$ and $SNR_{dB} = 10log_{10}(0.5) = -3.01$. Sound is typically measured in amplitude (*A*) units, such as root mean square amplitude. The ratio of two power values is proportional to the ratio of their amplitudes squared, so:

$$SNR = \frac{P_{signal}}{P_{noise}} = \left(\frac{A_{signal}}{A_{noise}}\right)^2 \text{ [7]}$$

Expressing SNR in decibel units:

$$SNR_{dB} = 10log_{10}\left[\left(\frac{A_{signal}}{A_{noise}}\right)^2\right] = 20log_{10}\left(\frac{A_{signal}}{A_{noise}}\right) \text{ [8]}$$

In the context of our experiment, "A_{signal}" is the amplitude of the target CRM speech token and "A_{noise}" is the combined amplitude of the two distracter CRM speech tokens. The decibel values in Table 3 can be used to estimate these amplitudes. Suppose the target speech comes from directly ahead of the listener and that we want to focus on the high frequency range. The top row of Table 3 shows that the sound level at KEMAR's right ear is -2.823 dB. This sound level has the general form:

$$-2.823dB = 20log_{10}\left(\frac{A_{target}}{A_{reference}}\right) \text{ [9]}$$

For our calculations it doesn't matter exactly what the $A_{reference}$ value is, and it is convenient to just call it 1. We can then solve for:

$$A_{target} = 10^{(-2.823/20)} = 0.7225 \text{ [10]}$$

That establishes the amplitude of the target (also known as the "signal"). What about the distracters ("noise")? We need to consider the combined effect of both distracters. If the distracters also come from directly ahead, then each on its own will have the same amplitude as the target, 0.7225. Assuming that the two distracters are independent of one another, their combined amplitude is:

$$A_{distractors} = \sqrt{A_{disractor1}^2 + A_{distractor2}^2} = \sqrt{0.7225^2 + 0.7225^2} = 1.0218 \text{ [11]}$$

Finally, the signal to noise ratio is:

$$SNR_{dB} = 20log_{10}\left(\frac{A_{signal}}{A_{noise}}\right) = 20log_{10}\left(\frac{0.7225}{1.0218}\right) = -3.01dB \text{ [12]}$$

For any co-located set of a target and two distracters (assuming all three are presented at the same source sound level), the signal to noise ratio is about -3 dB. This is true at both ears and for any direction of the source loudspeaker. Of course, the overall sound level is higher at the ipsilateral ear when the source is off to one side, but each ear receives the same signal to noise ratio.

If the target and distracters come from different directions, signal to noise ratios vary in more complicated ways. Consider the case of a target from directly in front and distracters from 45 deg to the left and right. We continue to focus on the high frequency sound level measures. The target has the same sound level, -2.823 dB, and amplitude, 0.7225, as in the previous example. With reference to Table 3, the distracters are from loudspeakers 57 (45 deg to the left, sound level -10.265 dB) and 9 (45 deg right, sound level 0.000 dB). These have amplitudes:

$$A_{distractor1} = 10^{(-10.265/20)} = 0.3067 \text{ [13]}$$

$$A_{distractor2} = 10^{(0.000/20)} = 1.0000 \text{ [14]}$$

The combined amplitude of the distracters is:

$$A_{distractors} = \sqrt{0.3067^2 + 1.0000^2} = 1.0460 \text{ [15]}$$

and the signal to noise ratio is:

$$SNR_{dB} = 20log_{10}\left(\frac{0.7225}{1.0460}\right) = -3.21dB \text{ [16]}$$

At first glance it is surprising that with both distracters spatially separated from the target by 45 deg the signal to noise ratio is actually worse than the -3.01 dB for the co-located condition. However, as Figure 8 shows, the distracter coming from 45 deg to the right reaches the right ear with a very high sound level. The signal to noise ratio at the left ear would also be -3.21 dB because the distracters are from the symmetrical directions of 45 deg left and right.

For other scenarios the signal to noise ratios differ at the two ears. Suppose the target is 45 deg to the right and the distracters are 45 deg to the left and directly ahead. By calculations like those above, the signal to noise ratio at the right ear is 2.10 dB, about 5 dB higher than the examples above. This makes sense because the target now has a strong path to the right ear. However, the signal to noise ratio at the left ear is -12.09 dB. This unfavorable ratio at the left ear is because one of the distracters has a strong path to that ear, and the sound from target is strongly shadowed by the head.

Calculations like these were performed to estimate the signal to noise ratios that would occur at each ear for all twelve listening conditions of Experiment 1. As described above, these calculations assumed that the target and both distracters have identical source sound levels (this might be true, for example, at the beginning of a test session). For the target motion conditions (4, 5, 6) the signal to noise estimates were based on the end of the motion path, since the CRM color and number came toward the end of the sentence. For the distracter motion conditions (10, 11, 12) the estimates were made by averaging separate estimates for the two path directions, which were equally likely. The estimates were done separately for the low and high frequency ranges. For most of the listening conditions it was generally the case that when a signal to noise estimate was higher (beneficial) in one ear it was lower in the other

ear (the exceptions to this were conditions 1 and 3). For example, when the target is located to one side, the signal to noise ratio is more favorable at the ear on that side than the other ear.

Condition	Best Either	Best Low	Best High	Worst Either	Worst Low	Worst High	Thresholds
C1	-3.01	-3.01	-3.01	-3.01	-3.01	-3.01	1.41
C2	2.10	1.03	2.10	-12.09	-7.48	-12.09	-4.72
C3	-3.21	-3.90	-3.21	-3.90	-3.90	-3.21	-0.82
C4	1.49	-0.34	1.49	-9.20	-6.86	-9.20	-3.20
C5	2.03	2.03	0.98	-16.45	-7.37	-16.45	-3.43
C6	-1.00	-2.42	-1.00	-7.77	-6.09	-7.77	-0.79
C7	1.49	-0.34	1.49	-9.20	-6.86	-9.20	-3.99
C8	2.03	2.03	0.98	-16.45	-7.37	-16.45	-4.40
C9	-1.00	-2.42	-1.00	-7.77	-6.09	-7.77	-1.29
C10	2.87	1.18	2.87	-11.99	-7.46	-11.99	-3.18
C11	-2.43	-3.40	-2.43	-4.37	-4.37	-3.26	-0.38
C12	2.03	0.90	2.03	-11.54	-7.40	-11.54	-3.85

Table 4: Estimates of signal to noise ratios in listening conditions of Experiment 1. Estimates of signal to noise ratios (dB) in columns 2 through 7 are based on the acoustic measurements reported in Chapter 3, assuming equal sound levels of target and distracters at their sources. "Best" and "worst" refer to whichever ear has the higher or lower signal to noise ratio. "Either," "Low," and "High" refer to the frequency ranges as presented in Chapter 3. The threshold values in column 8 are the mean threshold signal to noise ratios from Experiment 1 - these are the means shown in Figure 6.

Table 4 shows the estimated signal to noise ratios for the listening conditions of Experiment 1. Rather than showing values for the left and right ears, the "best" and "worst" values are shown regardless of ear. The observed mean threshold signal to noise ratios obtained from the participants are also shown in Table 4, in the far-right column. Figure 9 illustrates that there is a very strong association between the acoustically estimated signal to noise ratios and the threshold values from the participants. This figure focuses on the estimated best signal to noise ratios regardless of ear and frequency range, but similarly strong

associations with the observed thresholds are also seen with other information from Table 4. The clustering of listening conditions in Figure 9 is very similar to that in Figure 6. This suggests that what drives differences in performance across the listening conditions is, to a very substantial extent, the way that the spatial locations of target and distracters create signal to noise ratios at the two ears.

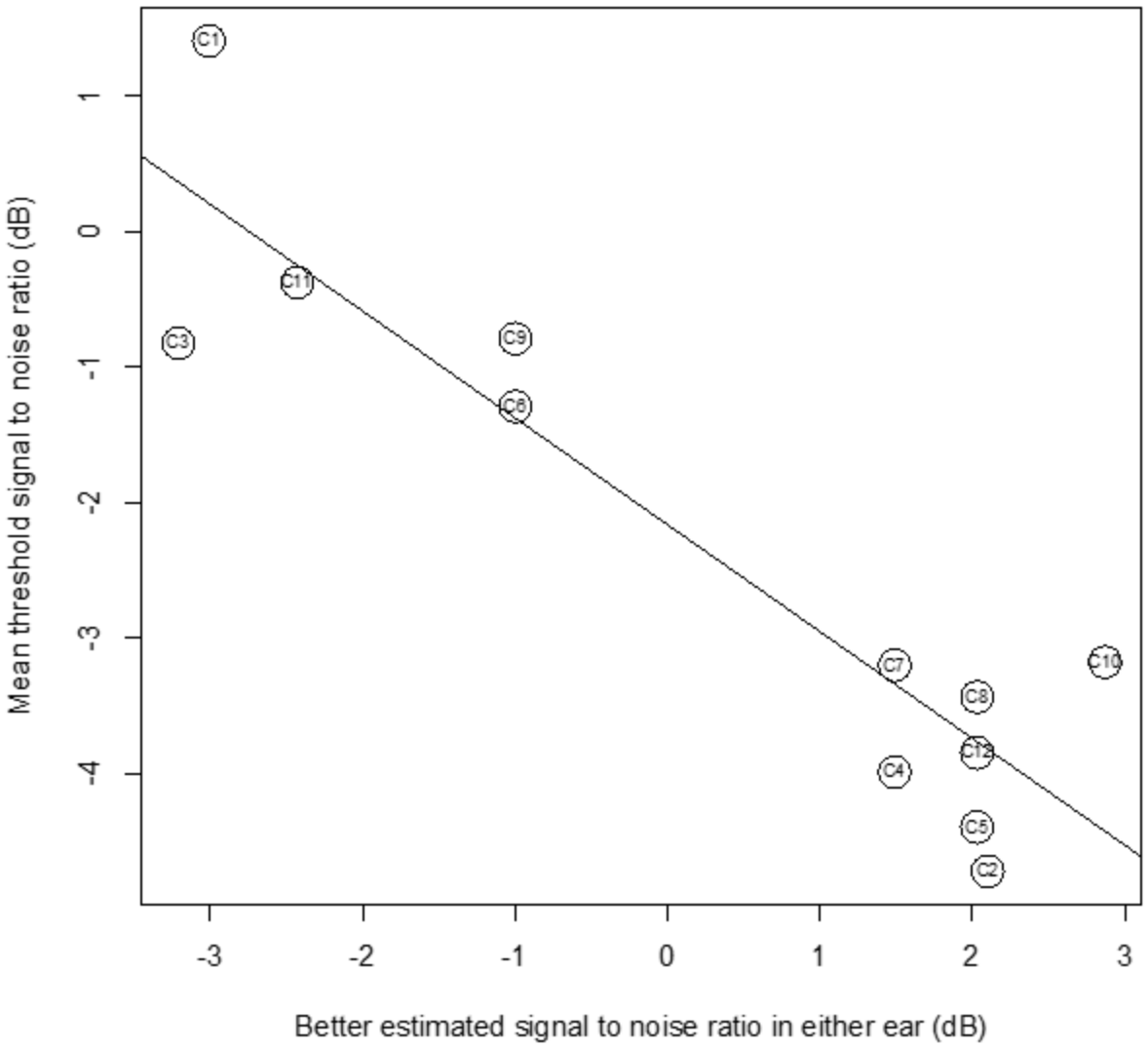

Figure 9: Experiment 1, association between observed threshold estimates (ordinate) and acoustically estimated signal to noise ratios for the ear with the more favorable ratio. Labels inside circles indicate listening conditions as summarized in Figure 2. Diagonal line is the best-fitting linear regression, y = -2.1638 - 0.7884x. The correlation is r = -0.917, p < 0.001. Note that the listening conditions cluster very similarly to Figure 6, suggesting that variations in the signal to noise ratios available under different listening conditions are a strong driver of speech understanding performance.

Discussion

In the existing literature on the cocktail party problem and spatial release from masking, much attention has been given to the amount of spatial separation between talkers. Comparisons are often between a co-located condition and talkers spaced apart by some amount. While these are completely valid measures, the patterns revealed in Figure 8 suggest that it is important to consider not only the separation between talkers but also the specific locations of talkers. Considering the differences in SNR between a scenario where the target is positioned with a distracter on either side and one where the target is positioned such that both distracters are off to one side, the relative positions of talkers can have a substantial impact on SNR, and thus on speech understanding. Indeed, this finding confirms that some of our findings in Experiment 1 were due to differences in SNR (Figure 9). Specifically, the finding that performance was better when the target was positioned on one side is fully supported by the approximate 5dB improvement in SNR at the ear closest to the target, relative to the condition where the target was positioned directly in front of the listener and between two distracters. From an ecological perspective, this finding suggests that in an environment with a particularly challenging SNR, it may not necessarily be best to point your head directly at the person you want to hear.

Given the substantial effect observed in our calculations of azimuth on SNR, we were motivated to take on a new perspective of the potential utility of motion. Instead of thinking about motion as a perceptual cue causing one talker to pop out from the background, perhaps motion of all talkers could be useful as a means for positioning the target such that a more favorable SNR is achieved. Imagine carrying on a conversation in a very noisy, crowded space.

The listener, struggling to understand what is being said, might lean in slightly while simultaneously turning the head slightly to direct one ear more to the sound source. This seemingly innate reaction to a difficult listening situation is likely based on the same observances made in our SNR calculations, since turning one ear slightly towards the sound source allows for a more direct path to that ear, theoretically optimizing the SNR. Indeed, laboratory studies have confirmed that both young adults with normal hearing (Grange & Culling, 2016a) and adults with hearing impairment (Grange & Culling, 2016b; Grange et al., 2018) can obtain significant head-orientation benefit by turning their head while listening. These data, and other studies such as Wightman and Kistler (1999) have demonstrated that the effects of simulated and real head turns are equivalent. There is, however, some evidence that auditory spatial acuity is slightly better during self-motion than source motion (Brimijoin & Akeroyd, 2012, 2014).

There have also been some mentions in the literature of the effect of the acoustic bright spot phenomenon and its potential effects on sound source localization (Macaulay, Hartmann, & Rakerd, 2010) and speech understanding (Grange & Culling, 2016a). We, too, observed the acoustic bright spot in our measurements. The bright spot is seen on the left side of Figure 8 by the hollow circles representing the high-frequency band. Since measurements were taken at the right ear, the narrow increase in amplitude in the high frequencies as the sound source approaches 90 degrees to the left is caused by the bright spot falling at or near the far (right) ear. Our observation of the bright spot effect was that it tended to have a minimal impact on the overall level at the far ear, at least given the way we analyzed by grouping frequencies into

two “bins”. The impact of the acoustic bright spot on speech understanding is discussed in more detail in the following chapter.

The observed effect of sound source azimuth on signal-to-noise ratio in complex listening environments served as a great backdrop for Experiment 2. It was clear from Experiment 1 that motion of a single talker was not enough to overcome any advantage or disadvantage of signal-to-noise ratio. Motivated by the recent work of Grange and Culling (2016a) on the potential advantages of turning one’s head to optimize SNR during a speech task, a new motion paradigm was used for a second experiment. The primary aim of Experiment 2 was to evaluate the extent to which a simulated head turn could allow a listener to take advantage of these momentary changes in SNR to better identify and understand speech of the target talker. By extension, we sought additional evidence for differential benefit of SNR improvement at the beginning vs end of the utterance. An additional motivation was to test the possible effect of the acoustic bright spot on speech understanding.

CHAPTER 4

Experiment 2

In order to assess whether a favorable SNR at the beginning or end of the trial was more helpful, conditions were designed in groups including two motion conditions and two stationary conditions. The two stationary conditions represent the start and end points, if you will, of the two motion conditions. Since the motion in this experiment was unidirectional, two motion conditions with the same start and end points as the two stationary conditions were included in each comparison. From these comparisons, we hoped to be able to determine whether it was more important for a listener to have a more favorable SNR at the beginning of the utterance (i.e. during callsign identification), or at the end (during the color-number complex). Whichever stationary condition had the best SNR was expected to outperform the other. With regard to the motion conditions, it was hypothesized that performance would be greater for the motion condition with the better SNR at the beginning of the utterance, as accurate target talker identification should allow the listener to maintain perceptual streaming during the color-number complex (see Shinn-Cunningham (2008)). These types of comparisons were made for several talker positions.

Based on the findings of Experiment 1, it was generally expected that performance would be greater when the target was on the end of the two distracters than when it was between them. Particularly, it was hypothesized that the best performance would be observed in conditions where the target moved in front of the listener with the two distracters before/after it. This spatial configuration, along with one of the stationary correlates, involved the presentation of the target from one side of the head and the distracters from in front and

the other side. Such a set up should produce the most robust head shadow effect and therefore the best SNR for the target.

Methods

The same three-talker paradigm and test environment from Experiment 1 were used in Experiment 2. Six new young adult participants and two repeat participants were recruited for Experiment 2 (age range = 22-33, mean = 26; two male). The stimulus conditions shown in Figure 10 were designed to assess the extent to which a simulated head turn would provide a listener with an adequate opportunity to change the SNR in a favorable way that would lead to an improvement in speech understanding. A head turn was simulated in the anechoic chamber by simultaneously rotating all three talkers by 90 degrees in the same direction. The velocity of motion was approximately 45 deg./sec. For the purposes of this study, simulating head turns was preferable to directing participants to move their heads because of the added complexity of having to monitor for head velocity, timing, and extent of motion on a trial-by-trial basis.

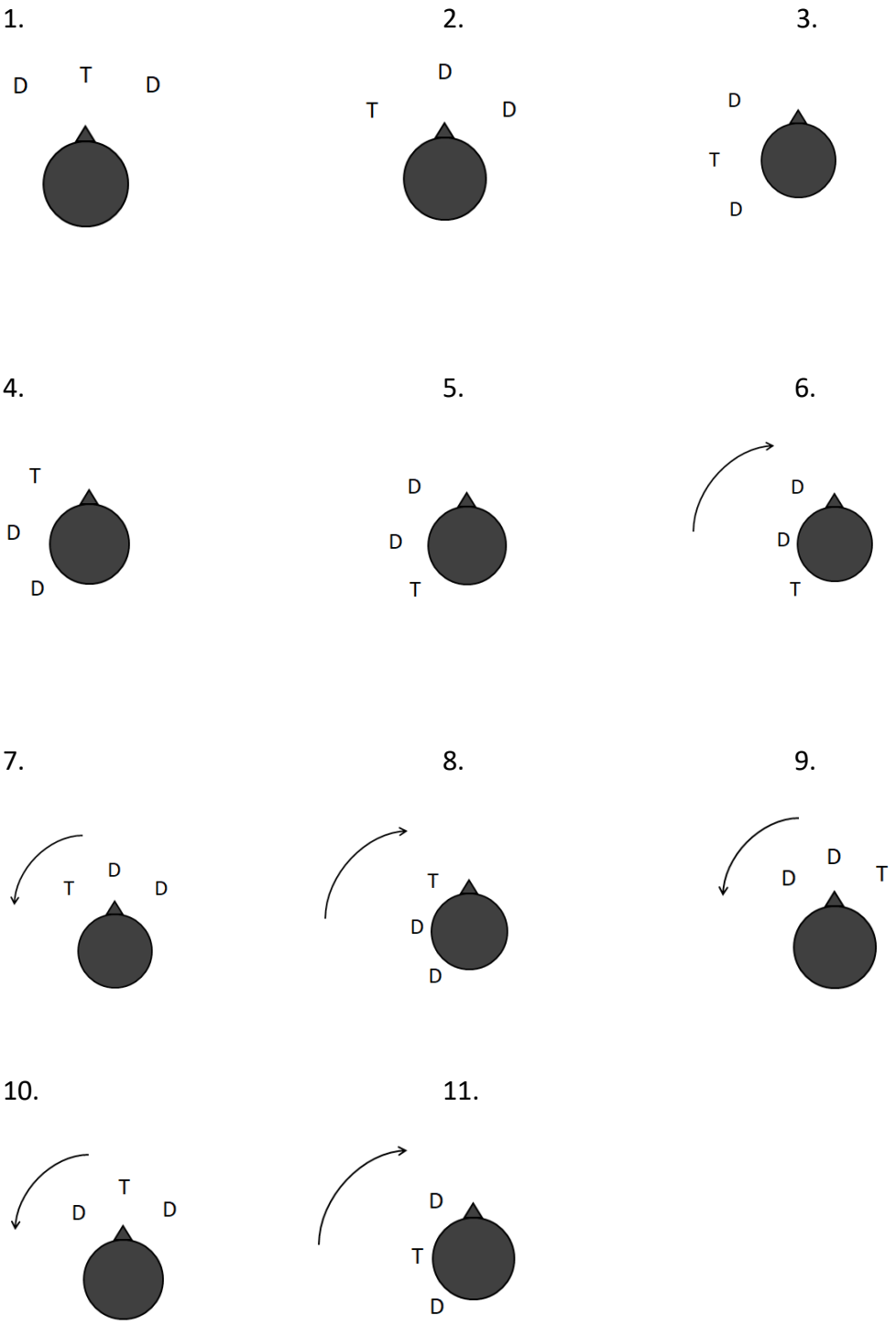

Figure 10: Stimulus conditions for Experiment 2. For the sake of simplicity, only left-sided positions are shown here. For each condition except for condition 1, a mirror opposite condition was also run.

Procedure

Conditions 1 and 2 were stationary conditions in which the three talkers were centered in front of the listener. In condition 1, the target was positioned directly in front of the listener with a stationary distracter at ±45 degrees on either side. In condition 2 the target was positioned at either 45 degrees to the left or right of center with one distracter directly in front of the listener and the 45-degree location not occupied by the target. Conditions 3-5 were also stationary conditions, but all three talkers were randomly selected to be either on the listener's left or right. If the talkers were on the listener's right, one talker would be positioned at 45 degrees azimuth, one at 90 degrees, and the third at 135 degrees. In condition 3, the target was positioned in between the two distracters at ±90 degrees. In condition 4, the target was positioned at ±45 degrees, and in condition 5 at ±135 degrees. Conditions 1-5 served as the baseline measures for Experiment 2. Among these conditions, it was hypothesized that the worst performance would be observed in condition 1. Although the distracters were not co-located with the target as in Experiment 1, the symmetrically positioned distracters did not create any better-ear advantage. In contrast, condition 2 created a better SNR in the ear closest to the target, and condition 3 caused the target to fall in the bright spot at the far ear. Therefore, the best performance of these stationary conditions was expected in condition 2, followed by condition 3 and condition 1. Conditions 1 and 2 in Experiment 2 were also tested in Experiment 1.

Conditions 6-11 were all motion conditions. Recall that a major difference in the motion paradigm from Experiment 1 is that all three talkers were in simultaneous motion in Experiment 2. Conditions 6, 8, and 11 started with all three talkers off to one side, moved

towards the listener's front, and ended centered directly in front of the listener. In condition 6, the target moved from ±135 degrees to 45 degrees on that same side. In Condition 8, the target started at ±45 degrees and crossed in front of the listener, ending at 45 degrees on the opposite side. In condition 11, the target was centered between the distracters, starting at ±90 degrees and ending directly in front of the listener. Conditions 7, 9, and 10 were essentially the opposite of the other three motion conditions. In these last three, the talkers began in front of the listener and moved at a constant velocity to one side or the other. It was unclear prior to the start of data collection whether having a more favorable SNR at the beginning or end of the utterance would be more helpful for a listener. A favorable SNR at the beginning of an utterance could give a listener an opportunity to capture the location and vocal characteristics of the target talker by way of hearing out the call sign but could also make understanding the key words at the end of the utterance more difficult. Since important information is spread throughout the sentence in the CRM stimuli, it was not clear how listeners would respond to the dynamic SNR in the motion conditions of Experiment 2.

Results

In a repeated measures analysis of variance with Condition (11 conditions) and Run (2 runs) as factors, there was a significant effect of Condition, $F(1,5.319) = 20.201$, $p < .001$, $\eta^2_{partial} = 0.743$, degrees of freedom per Huynh-Feldt correction for sphericity. As in Experiment 1, the main effect of Run and the interaction effect were not significant. The mean (and standard error) signal to noise ratios on the first and second runs were -0.533 dB (0.418) and -0.803 dB (0.561), respectively, so there was an average improvement of 0.27 dB across runs. The intraclass correlation coefficient was 0.966, indicating that there was excellent

reliability across the first and second runs. Accordingly, as in Experiment 1, further analyses were based on each participant's scores averaged across runs within each listening condition. Group mean performance by condition for Experiment 2 is shown in Figure 11. Group means for each condition were derived using the same methods as in Experiment 1.

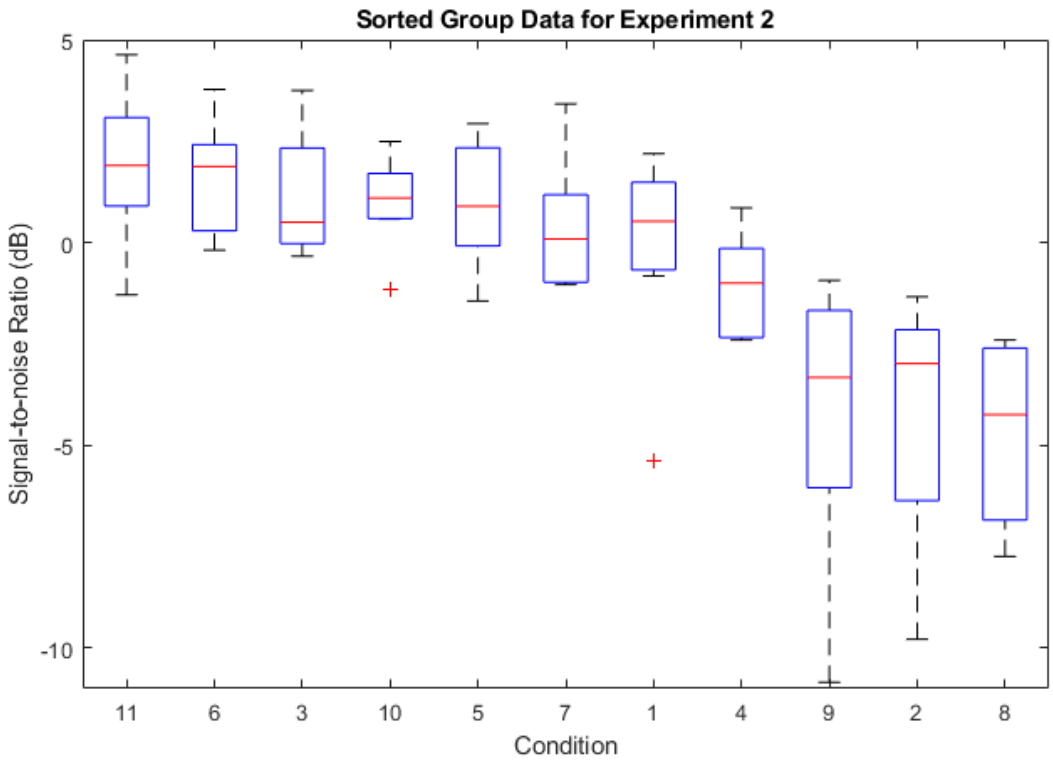

Figure 11: Group results for Experiment 2. Signal-to-noise ratio (dB) is plotted on the y-axis for each of the 11 conditions. Conditions are sorted on the x-axis from highest to lowest mean SNR. Red lines indicate the median for each condition, while boxes denote the 25th and 75th percentile, whiskers denote the 5th and 95th percentiles. Red plus signs indicate outliers for a given condition.

Linear mixed modeling was performed with the following factors: target location (target located between distracters or on one end of them); overall orientation (stimuli starting or remaining in front of or to one side of the listener); and motion (stimuli stationary or moving).

There were significant effects of target location ($F(1,76) = 26.917$, $p < .001$) and overall orientation ($F(1,76) = 10.902$, $p < .002$), as well as an interaction between overall orientation and motion ($F(1,76) = 5.187$, $p < .026$). Descriptively, signal to noise thresholds were lower (better) when the target was on one end of the distracters (M = -1.62 dB, SE = 1.16) than when the target was centered between the distracters (M = 0.99 dB, SE = 0.64). This difference of about 2.6 dB favoring the target-at-end configuration replicates a similar finding from Experiment 1, where the effect was about 3 dB in magnitude. With respect to the overall orientation of the stimuli, thresholds were better when the stimuli started in front of the listener (M = -1.44 dB, SE = 1.16) than when they started to one side (M = -.03 dB, SE = 0.99). The difference of about 1.4 dB might reflect an attentional advantage for events occurring generally in a listener's heading direction. The interaction between motion and general orientation was, to some extent, opposite to our expectation that motion simulating a head turn might be beneficial for speech understanding. When the stimuli were in front of the listener to start with, thresholds were better when the stimuli remained stationary at the front (M = -2.16 dB, SE = 1.20) than when the stimuli moved toward one side (M = -0.95 dB, SE = 1.13). Thus, there was a fairly substantial benefit of 2 dB when the stimuli remained in place. However, when the stimuli started at the side of the listener, the motion effect was in the opposite direction - thresholds were somewhat worse when the stimuli remained stationary (M = 0.35 dB, SE = 0.61) than when they moved (M = -0.40, SE = 1.27), an advantage of about 0.75 dB for motion.

The stationary listening conditions (1 through 5) of Experiment 2 were a constructive replication of one aspect of Experiment 1 - the finding that thresholds were better when the

target was to one side of the distracters than when it was centered between them. Indeed, conditions 1 and 2 of Experiment 2 were a direct replication of conditions 3 and 2 from Experiment 1. Once again, the average threshold was lower (better) when the target was to the side of the distracters (Experiment 2, condition 2, M = -4.26 dB, SE = 1.06 dB) than when the target was centered between the distracters (condition 1, M = -0.06 dB, SE = 0.84 dB), a significant difference, $t(7) = 4.698$, $p < .001$. In these conditions the stimuli were generally in front of the listener. In designing Experiment 2, we speculated that the effect of whether the target was between the distracters or to one side of them might differ (or even "flip") if the stimuli were off to one side of the listener. This reasoning was based on the possibility that the signal to noise ratios at the listeners' ears might differ when the stimuli were to one side of the listener rather than in front. There was evidence to support this speculation. Figure 12 shows threshold signal to noise ratios plotted against estimated signal to noise ratios at the ears for the five stationary conditions of Experiment 2. Conditions 1 and 2 had stimuli in front of the listener, while conditions 3, 4, and 5 had stimuli at the side. The pattern of findings suggests that the signal to noise ratio at the ears (specifically, at the ear receiving a more favorable ratio) has a very strong influence on speech understanding. This is perhaps best understood in terms of energetic masking. Although the spatial locations of the target and distracters relative to each other may also play a role in terms of attentional focus, the acoustical signal to noise ratios at the ears seem very important. Those acoustical factors are determined by a combination of where the target and distracter(s) are located, and how the listener's head is oriented.

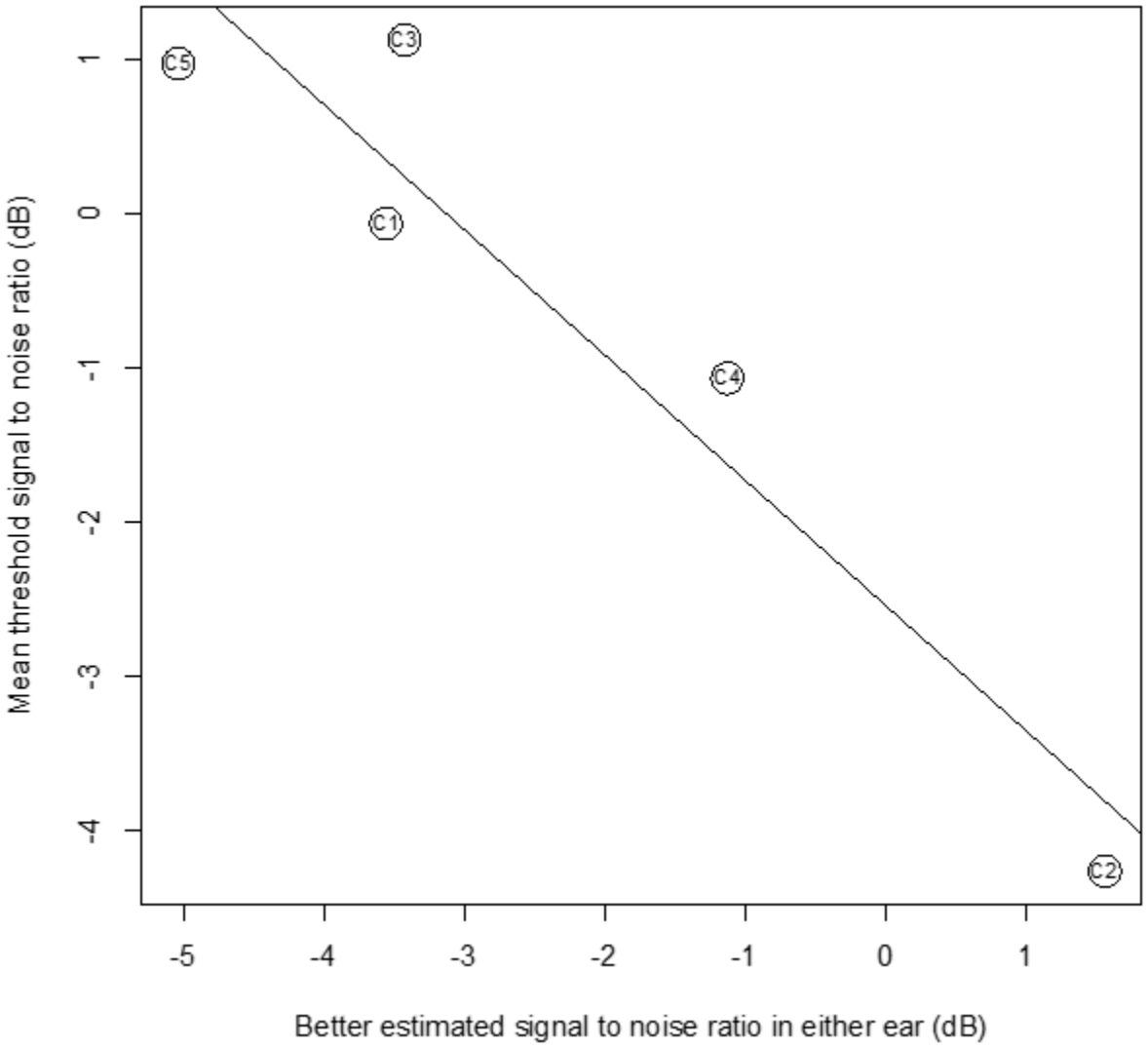

Figure 12: Experiment 2, association between observed threshold estimates (ordinate) and acoustically estimated signal to noise ratios for the five stationary listening conditions. The abscissa shows the mean estimated signal to noise ratio (averaged across the low and high frequency ranges) for the ear with the more favorable ratio. Labels inside circles indicate listening conditions as summarized in Figure 10. Diagonal line is the best-fitting linear regression, y = -2.5369 - 0.8114x. The correlation is r = -0.952, p < 0.001.

Group mean results in Experiment 2 tended to fall into two main categories of performance. Starting with the worst performance (most favorable SNR needed), conditions 1,

3, 4, 5, 6, 7, 10, and 11 had similar average thresholds (mean = 0.73 dB, range -1.06 – 1.89). This group of conditions included both stationary and motion conditions, with the target between and on the end of the distracters. Condition 4 resulted in the best performance among this group, with mean performance of -1.07 dB, while the other conditions ranged from -0.06 to 1.89 dB SNR. The remaining three conditions in Experiment 2 resulted in performance substantially better than all the rest. Conditions 2, 8, and 9 had mean SNRs of just under -4dB. Conditions 8 and 9 were motion conditions where the target crossed in front of the listener from 45 degrees on one side to 45 degrees on the other. Condition 2, repeated from Experiment 1, was a stationary condition with the talkers in front of the listener, but also involved the target on the end of the two distracters. Average performance in condition 2 (-4.26 dB) was not statistically different ($t = 0.295$, $p = 0.776$) from same condition when it was tested in Experiment 1 (-4.72 dB). The other repeated condition in Experiment 2 was condition 1 (condition 3 in Experiment 1). Like the first repeated condition, performance was not statistically different from Experiment 1 to Experiment 2 (mean difference = 0.76 dB, $t = 0.66$, $p = 0.530$).

As previously mentioned, groups of conditions were designed to assess whether a favorable SNR at the beginning or end of a trial would be more helpful. Conditions 1,3,10, and 11 formed one of these groups. In this example, it was hypothesized that the best performance would occur in condition 3 where the target was positioned in the bright spot of the other ear. However, performance in all four conditions was rather poor, with the best performance surprisingly in condition 1 (mean performance = -0.06 dB). In fact, the worst performance in all of Experiment 2 was observed in condition 11 (mean performance 1.89 dB), where the target

started at 90 degrees to one side and moved to directly in front of the listener. Instead of showing evidence for benefit from the target being positioned in the acoustic bright spot in the far ear in condition 3, performance was actually 1.19 dB worse in this condition than in condition 1 where the target was positioned directly in front of the listener. This difference did not reach statistical significance ($t = 1.376$, $p = 0.211$).

A similar comparison could be made among conditions 2, 5, 6, and 7. In this case, the target was not between the distracters, and stayed on one side of the listener. Performance in condition 2 was found to be quite good when it was tested in Experiment 1 and was expected to be the best in this group of 4 conditions. Indeed, condition 2 (mean performance = -4.26 dB) widely outperformed the other three conditions in this comparison (mean performance = 0.98 dB). In condition 2, the target was positioned at one side of the head, while one distracter was directly in front, and the other distracter was on the opposite side of the head from the target. This produced a favorable SNR in the ear nearest the target. It was suspected that one of the two comparable motion conditions would produce similar results to condition 2, but that was not the case. Like in Experiment 1, there was essentially no evidence in this group of conditions that motion aided in speech understanding. Even in a case where the signal to noise ratio of the target improved during motion (e.g. condition 6), performance was essentially equivalent to the stationary control condition where the SNR remained negative and unchanged throughout the trial (e.g. condition 5). From this, it seems that listeners benefitted from a consistent, positive SNR throughout the utterance, which was only present in condition 2. In the other three conditions, the target was on the same side of the head as the two distracters for at least a portion of the utterance, thereby decreasing the SNR for both ears.

The groups of comparisons described up to this point have suggested that motion is not a helpful cue. However, analysis of the third group of conditions revealed a different pattern of results. Conditions 2, 4, 8, and 9 involved the target with asymmetrically positioned distracters, but also involved the target moving across the front of the listener. Of the two stationary conditions, better performance was expected in condition 2, since the target was presented from one side of the head, and distracters from in front and the other side. As previously discussed, this creates a robust head shadow and a favorable SNR. In contrast, condition 4 involved presentation of all three talkers from the same side of the head, and with one of the distracters in the bright spot at the far ear. Unlike the other two groups of comparisons where performance in the motion conditions was in line with the worst of the stationary conditions, in this case the motion conditions (conditions 8 and 9) were among the best in Experiment 2, with average results under -4 dB SNR. While the motion conditions did not outperform the best stationary condition, they were no worse either.

Listeners clearly only benefitted from the target being on the end of the distracters when it was on the end closer to the front of the listener, rather than behind. Conditions 5-7 all involved the target on the end of the distracters, but the target never passed in front of the listener. Instead, listeners reported feeling that the target “got lost” behind them. Indeed, performance in these conditions was approximately 5dB worse than conditions 2, 8, and 9 where the target was on the other end of the distracters. This was a particularly interesting finding given that this study was conducted in an anechoic chamber where minimal other acoustic cues should differentiate these groups of conditions, save for pinna shadow cues, and possibly poorer spatial resolution around the 135 degree azimuth than at 45 degrees (Mills,

1958). In contrast to Experiment 1, the results from Experiment 2 seemed to suggest that listeners may indeed be better able to attend to a target talker that is directly in front of them, at least when the distracters are both positioned on the same side as the target. This finding may have also been brought on by another difference between the two experiments. In Experiment 1, talker locations were limited to the range of ±45 degrees (plus up to 20 degrees of jitter), whereas in Experiment 2, talker locations expanded to ±135 degrees (plus up to 10 degrees of jitter). Since the talker was actually located at the ±135 degree location only a minority of the time, perhaps listeners tended to neglect this area and instead focus on talkers more or less in front of them.

Finally, there have been several reports in the acoustic literature about the so-called bright spot effect, wherein sound passing around a head (or as is often the case in the literature, a spherical head model) converges on the side of the head opposite the sound source such that rather than a head shadow, a local convergence of sound energy results in an increase in amplitude in the higher frequencies. In studies of this phenomenon using light instead of sound, a literal “bright” spot is observed in the middle of the shadow. There have not been any previous studies documenting the potential impact of the acoustic bright spot on speech understanding in a multi-talker environment. Our recordings showed that for stationary talkers with the target directly in front of the listener and surrounded by two distracters, the long-term high-frequency SNR is about -3.21 dB, and the low-frequency SNR is -3.90 dB. But if all three talkers are rotated by 90 degrees such that the target is presented from 90 degrees azimuth, the target falls into the acoustic bright spot in the far ear while the two distracters are at near maximal head shadow. However, the observed bright spot effect in these

measurements was minimal, and frequency dependent. The high-frequency SNR in the far ear gets slightly worse (-3.37 dB), while the low-frequency SNR improves to -2.58. The largest effect was observed at 1,000Hz where a 5 dB improvement in SNR was observed when the target was presented from 90 degrees, compared to directly in front of the listener. But because this effect was limited to 1,000Hz, there was little impact on the overall SNR or speech understanding.

Discussion

The primary purpose of this study was to determine if motion could be a helpful cue for speech understanding. It is well established that the human auditory system is incredibly sensitive to small changes in location of stationary sound sources. And while it is less sensitive to moving stimuli, there is no doubt that we possess a strong ability to perceive motion of auditory objects. Despite the extensive research literature on these topics, there remains a rather small number of studies that have examined the role of motion on perhaps the most valuable ability of our auditory systems, speech understanding.

In this study, the potential effects of motion were examined with two approaches. First, a pop-out effect of motion was tested by moving one talker against a background of two stationary talkers. Consistent with previous studies (Davis et al., 2016; Pastore & Yost, 2017), no pop-out effect was observed in Experiment 1. Pilot testing had shown that motion of one talker could be a useful cue if it was known ahead of time that only the target would be moving. In instances when the target callsign could not be identified, the motion of the target talker against a background of stationary distracters allowed listeners a second glance at the target and the correct color-number combination. However, once other conditions were added that

included distracter motion, that cue became much less helpful. While these pilot data were collected on individuals with high levels of familiarity with the study design, it does seem possible that this is an example of auditory pop-out effect. Strict conclusions cannot be made here, however, since this was not tested at the group level.

The fact that no effect of motion was observed at the group level should perhaps not be surprising. Since this study design included both moving targets and distracters, motion was not a cue that led a listener only to the target. Listeners may have been able to detect the motion on each trial but may have allocated less attention to it since it was often misleading. One hypothesis from this study was that the longer stimuli (~ 2 seconds) of the CRM compared to the ~750ms stimuli used by Pastore and Yost (2017) would allow listeners enough time to take advantage of motion to better identify the target. Listeners in both studies were able to identify the presence of motion but could not take advantage of it. Additional work is needed to identify whether longer stimuli (e.g. running speech) would allow listeners to use motion to identify and understand a target talker.

One other strategy that was explored in pilot testing was motion that changed direction. It was initially hypothesized that an auditory object that changed direction draw a listener's attention and lead to improved speech understanding. We investigated a motion path that was unidirectional (as in Experiment 1), changed direction 1 time (i.e. back and forth), and changed direction 2 times. In order to complete these changes in direction while keeping the angular extents of motion constant, the velocity of the moving talker had to increase as a multiple of the number of changes in direction. We found that the motion paths that changed direction even one time were somewhat hard to follow as it seemed our attention lagged behind the

stimulus. For this reason, we limited the motion in the study to a single direction and a relatively low velocity of about 20 degrees/second for Experiment 1 and 45 degrees/second for Experiment 2.

Another interesting finding from Experiment 1 was a lack of effect of positioning the target slightly closer to or farther away from the distracters (e.g. condition 7 vs 8, condition 3 vs 9). Davis et al. (2016) found that placing the target just 10 degrees away from the distracters produced a substantial release from masking. It was hypothesized that some additional effect may be observed at greater separations, but none was observed here. Significant spatial release from masking has been demonstrated in other studies with as little as 15 degrees of separation between a target and symmetrically-placed distracters (G. Kidd, Jr., Mason, Best, & Marrone, 2010; Marrone, Mason, & Kidd, 2008). Srinivasan et al. (2016) found that spatial separations as small as 2 degrees can produce significant spatial release from masking for young adults with normal hearing. Perhaps there is little remaining informational masking effect at the 20-degree separations used in the current study. Our SNR calculations also supported that there was little change in these conditions.

The main effect observed in Experiment 1 was that of target position relative to the distracters. On average, performance was 3dB worse when the target was positioned between the two distracters than when it was on either end. It was initially thought that this observed effect was caused by a perceptual impact of having two distracters relatively near the target in the case of a target between the distracters compared to when it is on the end and one of the distracters was 90 degrees away. Indeed, there are no previous studies that suggest any amount of available spatial release from masking with 90 degrees of separation between

talkers. But if the distance argument was valid, then we should have also observed some effect in the offset target conditions (conditions 7-9). The SNR calculations shed some light on the lack of effect of the offset target conditions.

Consider the case where the target is positioned at 45 degrees to one side, and the distracters are directly in front of the listener and 45 degrees to the other side. The high-frequency average SPL in the ear canal will be used for this example. The long-term SNR in ear nearest the target is 2.1 dB, and -12.09 dB in the far ear. Clearly, the listener must rely on the SNR in the better ear to do this task. If the target is then shifted closer to the distracters, it is essentially moving away from the near ear and closer to the far ear. As such, the SNR decreases slightly in the near ear to 1.49 dB and improves slightly in the far ear to -9.2 dB. If instead the target is shifted away from the two distracters even farther, the SNR in the near ear drops to 0.98 dB. The far ear is essentially worthless for speech understanding in this example as the SNR drops to -16.45 dB in the high frequencies. As can be seen from these examples, the SNR in the far ear remains so poor that it can likely contribute little to understanding what is being said. With regard to the near ear, the SNR change across shifted positions was less than 1 dB. This finding agrees with the observed results from Experiment 1 in conditions 2, 7, and 8.

However, it was also clear from the SNR calculations that there is a significant effect of talker position relative to the distracters. The only time a positive SNR existed in these experiments was when the target was positioned on one side of the head, and the distracters on the other. Additionally, we observed a strong relationship between the mean threshold on a given condition in Experiment 1 and the estimated SNR in the better ear. This finding suggested that the SNR in the better ear was a strong driver of speech understanding in that experiment.

Based on these results, additional conditions were created to test for possible SNR effects at different target positions, while still including motion as another variable. It had been hypothesized that performance would be best when the target was positioned such that either the acoustic bright spot (in the case of a target between distracters) occurred at one ear, or asymmetrically positioned talkers (i.e. condition 2) such that the target was on one side of the head and the distracters on the other.

In Experiment 2, additional talker locations were used in order to explore the possibility that target location was a more important variable than motion. A strong effect of talker position was observed, where the best performance was observed when the target was on the end of the two distracters and passed in front of the listener. In contrast, there was again no observed effect of motion type, and no effect of talker position (front vs side). These results suggest that listeners had a directional bias for sound sources in front of them and that the simulated head turn in this study was not helpful (or harmful) for listeners. There were undoubtedly several potential reasons for these findings. First, the amount of motion used here was potentially greater than a natural head turn. The number of interleaved conditions and the wide listening area made the predictability of the target location extremely difficult for listeners, thereby forcing them to attend to a potentially restricted area on any given trial. Given our natural tendency to look at people we are trying to listen to, it seems plausible that listeners tended to pay particularly close attention to talkers positioned in front of them. This conclusion is supported by the relatively poor performance when the target was located at the ±135 degree position (conditions 5-7). Regardless of whether the target was in that position at the beginning, end, or for the entire duration of the trial, listeners were not able to follow the

target to a location that was somewhat behind them. Finally, our SNR measurements support the finding that performance was best when the target was positioned on one side of the head, and the distracters in front and on the other side of the head. This configuration was present at least part of the time in both conditions 8 and 9, and for the entirety of condition 2. Further, the target passed directly in front of the listener in these two moving conditions, giving ample opportunity for the listener to identify and follow the target. For many listeners, these conditions were judged to be the easiest. A potential bias for talkers in front of the listener may explain why performance was not as good in conditions 6 and 7, where at least part of these motion paths involved a favorable SNR in one ear.

Through these experiments, our perspective on motion perception as a cue for speech understanding has shifted. Prior to these experiments, motion was viewed as a possible means for release from informational masking. That is to say, it was suspected that motion of one talker would create a perceptual difference from a stationary background sufficient to improve speech understanding. As has been discussed, we found no evidence to support this hypothesis. Instead, we observed a strong relationship between the estimated SNR in the better ear and the mean threshold on each condition. This was true for all conditions in Experiment 1, and the five stationary conditions in Experiment 2. In this sense, we searched for an effect of informational masking and instead observed a strong effect of energetic masking. This should not have been surprising, given what is known about how head shadow changes the level at the ear with changes in azimuth. Further, we made few efforts to isolate the effects of informational masking from energetic. Our use of same-sex talkers ensured that a fairly high level of energetic masking would be present throughout the study. Nevertheless, perhaps the

results from this study suggest that we should take a more conservative approach when claiming to be evaluating the effects of informational masking on speech understanding. Whenever the location of a talker is changed, there are going to be changes in the SNR at each ear which, according to our findings, can be strong drivers of speech understanding.

The body of literature on our ability to understand speech in background noise is robust, to say the least. Understanding speech in background noise can be difficult for individuals with normal hearing, and even more so for individuals with hearing impairment. In the last 70 years, we have learned a great deal about the auditory cues we use to perceptually isolate one talker from one or more others. While the auditory system is precisely tuned to differentiate between very small differences in sound source location, the auditory system is less precise at detecting motion. While a great number of studies have explored the limits of the auditory system's ability to perceive motion, surprisingly few have looked at how speech motion might affect its ability to stand out from a background. Of those few studies, the consistent conclusion is that motion is not a helpful cue for speech understanding. In the capacity it has been tested to date, it also does not prove to be a harmful cue. This certainly leaves open the question of whether other motion parameters may prove to be helpful to speech understanding. A few studies have explored how individuals with normal hearing and hearing impairment can use a head turn to improve the SNR of a speech signal (Grange & Culling, 2016a, 2016b; Grange et al., 2018; Kock, 1950). It seems there is certainly a benefit to be had from using self-motion to orient a listener to the most favorable SNR. In a realistic listening environment, a sighted person will also be able to use the power of vision to both guide direction of attention, and to reinforce speech

understanding with visual cues. A multisensory approach to investigating motion perception could also yield interesting results.

CHAPTER 5

Conclusion

This study evaluated the effect of two different types of motion on speech understanding in multi-talker environments. In the first experiment, motion of either the target or one distracter was compared against several stationary conditions that mimicked the various spatial separations involved in the motion condition. There was no observed effect of talker motion in either direction. Instead, it was observed that speech understanding was the best when the target was positioned with both distracters off to one side. Suspecting these results were caused by an SNR effect induced by head shadow, the second experiment focused more on overall changes in SNR during simultaneous motion of all three talkers in one direction. Like the first experiment, no effect of motion was observed. Listeners simply favored the three conditions with the best SNR, specifically the conditions which included periods with the target on one side of the head and the distracters on the other. Of note, these three conditions included both stationary and motion conditions, suggesting that motion did not hinder speech understanding in this task.

Evaluating the effect of motion on speech understanding poses many potential pitfalls. Motion of one talker will inherently also change the SNR. If a change in performance is noted, how does one untangle the contribution of each parameter? And with so many possible sources of informational masking, it is difficult to create a realistic listening environment with unfiltered speech that leaves room for motion to provide a release from informational masking. While most of the research to date does not support the theory that motion can be a helpful cue,

future studies of other motion parameters may come to different conclusions. But based on the results of these experiments, it seems motion does not provide any useful information for speech understanding.

References

Ahissar, M., Ahissar, E., Bergman, H., & Vaadia, E. (1992). Encoding of sound-source location and movement: activity of single neurons and interactions between adjacent neurons in the monkey auditory cortex. *J Neurophysiol, 67*(1), 203-215. doi:10.1152/jn.1992.67.1.203

Alink, A., Euler, F., Kriegeskorte, N., Singer, W., & Kohler, A. (2012). Auditory motion direction encoding in auditory cortex and high-level visual cortex. *Hum Brain Mapp, 33*(4), 969-978. doi:10.1002/hbm.21263

Allen, K., Carlile, S., & Alais, D. (2008). Contributions of talker characteristics and spatial location to auditory streaming. *J Acoust Soc Am, 123*(3), 1562-1570. doi:10.1121/1.2831774

Andersen, R. A. (1997). Neural mechanisms of visual motion perception in primates. *Neuron, 18*(6), 865-872. doi:10.1016/s0896-6273(00)80326-8

Arbogast, T. L., Mason, C. R., & Kidd, G., Jr. (2002). The effect of spatial separation on informational and energetic masking of speech. *J Acoust Soc Am, 112*(5 Pt 1), 2086-2098.

Battal, C., Rezk, M., Mattioni, S., Vadlamudi, J., & Collignon, O. (2019). Representation of Auditory Motion Directions and Sound Source Locations in the Human Planum Temporale. *J Neurosci, 39*(12), 2208-2220. doi:10.1523/jneurosci.2289-18.2018

Baumgart, F., Gaschler-Markefski, B., Woldorff, M. G., Heinze, H. J., & Scheich, H. (1999). A movement-sensitive area in auditory cortex. *Nature, 400*(6746), 724-726. doi:10.1038/23390

Bednar, A., & Lalor, E. C. (2020). Where is the cocktail party? Decoding locations of attended and unattended moving sound sources using EEG. *Neuroimage, 205*, 116283. doi:10.1016/j.neuroimage.2019.116283

Bolia, R. S., Nelson, W. T., Ericson, M. A., & Simpson, B. D. (2000). A speech corpus for multitalker communications research. *J Acoust Soc Am, 107*(2), 1065-1066.

Borst, A. (2000). Models of motion detection. *Nat Neurosci, 3 Suppl*, 1168. doi:10.1038/81435

Brimijoin, W. O., & Akeroyd, M. A. (2012). The role of head movements and signal spectrum in an auditory front/back illusion. *Iperception, 3*(3), 179-182. doi:10.1068/i7173sas

Brimijoin, W. O., & Akeroyd, M. A. (2014). The moving minimum audible angle is smaller during self motion than during source motion. *Front Neurosci, 8*, 273. doi:10.3389/fnins.2014.00273

Bronkhorst, A. W., & Plomp, R. (1988). The effect of head-induced interaural time and level differences on speech intelligibility in noise. *J Acoust Soc Am, 83*(4), 1508-1516.

Bronkhorst, A. W., & Plomp, R. (1990). A clinical test for the assessment of binaural speech perception in noise. *Audiology, 29*(5), 275-285. doi:10.3109/00206099009072858

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *J Acoust Soc Am, 109*(3), 1101-1109. doi:10.1121/1.1345696

Brungart, D. S., & Simpson, B. D. (2007). Cocktail party listening in a dynamic multitalker environment. *Percept Psychophys, 69*(1), 79-91.

Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., . . . Ludvigsen, C. (1994). An international comparison of long-term average speech spectra. *J Acoust Soc Am, 96*(4), 2108-2120. doi:10.1121/1.410152

Carlile, S., & Best, V. (2002). Discrimination of sound source velocity in human listeners. *J Acoust Soc Am, 111*(2), 1026-1035.

Carlile, S., & Leung, J. (2016). The Perception of Auditory Motion. *Trends Hear, 20*. doi:10.1177/2331216516644254

Chandler, D. W., & Grantham, D. W. (1992). Minimum audible movement angle in the horizontal plane as a function of stimulus frequency and bandwidth, source azimuth, and velocity. *J Acoust Soc Am, 91*(3), 1624-1636.

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *J Acoust Soc Am, 25*(5), 975-979. doi:doi:http://dx.doi.org/10.1121/1.1907229
Christ, S. E., & Abrams, R. A. (2008). The attentional influence of new objects and new motion. *J Vis, 8*(3), 27.21-28. doi:10.1167/8.3.27
Davis, T. J., Grantham, D. W., & Gifford, R. H. (2016). Effect of motion on speech recognition. *Hear Res, 337*, 80-88. doi:10.1016/j.heares.2016.05.011
Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychol Rev, 96*(3), 433-458. doi:10.1037/0033-295x.96.3.433
Fletcher, H. (1940). Auditory patterns. *Reviews of modern physics, 12*(1), 47.
Freeman, T. C., Leung, J., Wufong, E., Orchard-Mills, E., Carlile, S., & Alais, D. (2014). Discrimination contours for moving sounds reveal duration and distance cues dominate auditory speed perception. *PLoS One, 9*(7), e102864. doi:10.1371/journal.pone.0102864
Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2001). Spatial release from informational masking in speech recognition. *J Acoust Soc Am, 109*(5 Pt 1), 2112-2122.
Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *J Acoust Soc Am, 106*(6), 3578-3588.
Grange, J. A., & Culling, J. F. (2016a). The benefit of head orientation to speech intelligibility in noise. *J Acoust Soc Am, 139*(2), 703-712. doi:10.1121/1.4941655
Grange, J. A., & Culling, J. F. (2016b). Head orientation benefit to speech intelligibility in noise for cochlear implant users and in realistic listening conditions. *J Acoust Soc Am, 140*(6), 4061. doi:10.1121/1.4968515
Grange, J. A., Culling, J. F., Bardsley, B., Mackinney, L. I., Hughes, S. E., & Backhouse, S. S. (2018). Turn an Ear to Hear: How Hearing-Impaired Listeners Can Exploit Head Orientation to Enhance Their Speech Intelligibility in Noisy Social Settings. *Trends Hear, 22*, 2331216518802701. doi:10.1177/2331216518802701
Grantham, D. W. (1986). Detection and discrimination of simulated motion of auditory targets in the horizontal plane. *J Acoust Soc Am, 79*(6), 1939-1949.
Griffiths, T., Bench, C., & Frackowiak, R. (1994). Human cortical areas selectively activated by apparent sound movement. *Current Biology, 4*(10), 892-895.
Griffiths, T. D., & Green, G. G. (1999). Cortical activation during perception of a rotating wide-field acoustic stimulus. *Neuroimage, 10*(1), 84-90. doi:10.1006/nimg.1999.0464
Griffiths, T. D., Rees, A., Witton, C., Shakir, R. A., Henning, G. B., & Green, G. G. (1996). Evidence for a sound movement area in the human cerebral cortex. *Nature, 383*(6599), 425-427. doi:10.1038/383425a0
Grothe, B., Pecka, M., & McAlpine, D. (2010). Mechanisms of Sound Localization in Mammals. *Physiological Reviews, 90*(3), 983-1012. doi:10.1152/physrev.00026.2009
Harris, J. D. (1972). A florilegium of experiments on directional hearing. *Acta Otolaryngol Suppl, 298*, 1-26.
Harris, J. D., & Sergeant, R. L. (1971). Monaural/binaural minimum audible angles for a moving sound source. *J Speech Hear Res, 14*(3), 618-629.
Hawley, M. L., Litovsky, R. Y., & Culling, J. F. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *J Acoust Soc Am, 115*(2), 833. doi:10.1121/1.1639908
Hillstrom, A. P., & Yantis, S. (1994). Visual motion and attentional capture. *Percept Psychophys, 55*(4), 399-411. doi:10.3758/bf03205298
Hirsh, I. J. (1950). The Relation between Localization and Intelligibility. *J Acoust Soc Am, 22*(2), 196-200. doi:doi:http://dx.doi.org/10.1121/1.1906588
Kaczmarek, T. (2005). Auditory perception of sound source velocity. *J Acoust Soc Am, 117*(5), 3149-3156.

Kesten, H. (1958). Accelerated Stochastic Approximation. *Ann. Math. Statist., 29*(1), 41-59. doi:10.1214/aoms/1177706705
Kidd, G., Arbogast, T. L., Mason, C. R., & Gallun, F. J. (2005). The advantage of knowing where to listen. *J Acoust Soc Am, 118*(6), 3804. doi:10.1121/1.2109187
Kidd, G., Jr., Mason, C. R., Best, V., & Marrone, N. (2010). Stimulus factors influencing spatial release from speech-on-speech masking. *J Acoust Soc Am, 128*(4), 1965-1978. doi:10.1121/1.3478781
Kidd, G., Jr., Mason, C. R., Rohtla, T. L., & Deliwala, P. S. (1998). Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns. *J Acoust Soc Am, 104*(1), 422-431.
Kock, W. E. (1950). Binaural Localization and Masking. *J Acoust Soc Am, 22*(6), 801-804. doi:10.1121/1.1906692
Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *J Acoust Soc Am, 49*(2), Suppl 2:467+.
Locke, S. M., Leung, J., & Carlile, S. (2016). Sensitivity to Auditory Velocity Contrast. *Sci Rep, 6*, 27725. doi:10.1038/srep27725
Macaulay, E. J., Hartmann, W. M., & Rakerd, B. (2010). The acoustical bright spot and mislocalization of tones by human listeners. *J Acoust Soc Am, 127*(3), 1440-1449. doi:10.1121/1.3294654
Marrone, N., Mason, C. R., & Kidd, G. (2008). Tuning in the spatial dimension: evidence from a masked speech identification task. *J Acoust Soc Am, 124*(2), 1146-1158. doi:10.1121/1.2945710
McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods, 1*(1), 30.
Mills, A. W. (1958). On the Minimum Audible Angle. *J Acoust Soc Am, 30*(4), 237-246. doi:doi:http://dx.doi.org/10.1121/1.1909553
Noble, W., & Perrett, S. (2002). Hearing speech against spatially separate competing speech versus competing noise. *Percept Psychophys, 64*(8), 1325-1336.
Nothdurft, H. C. (1993). The conspicuousness of orientation and motion contrast. *Spat Vis, 7*(4), 341-363. doi:10.1163/156856893x00487
Pastore, M. T., & Yost, W. A. (2017). Spatial Release from Masking with a Moving Target. *Front Psychol, 8*, 2238. doi:10.3389/fpsyg.2017.02238
Perrott, D. R., & Musicant, A. D. (1977). Minimum auditory movement angle: binaural localization of moving sound sources. *J Acoust Soc Am, 62*(6), 1463-1466.
Perrott, D. R., & Tucker, J. (1988). Minimum audible movement angle as a function of signal frequency and the velocity of the source. *J Acoust Soc Am, 83*(4), 1522-1527.
Raleigh, L. (1907). On our perception of sound direction. *Philosophical Magazine, 13*(6), 214-232.
Saberi, K., & Perrott, D. R. (1990). Minimum audible movement angles as a function of sound source trajectory. *J Acoust Soc Am, 88*(6), 2639-2644.
Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends Cogn Sci, 12*(5), 182-186. doi:10.1016/j.tics.2008.02.003
Smith, K. C., & Abrams, R. A. (2018). Motion onset really does capture attention. *Atten Percept Psychophys, 80*(7), 1775-1784. doi:10.3758/s13414-018-1548-1
Srinivasan, N. K., Jakien, K. M., & Gallun, F. J. (2016). Release from masking for small spatial separations: Effects of age and hearing loss. *J Acoust Soc Am, 140*(1), El73. doi:10.1121/1.4954386
Stumpf, E., Toronchuk, J. M., & Cynader, M. S. (1992). Neurons in cat primary auditory cortex sensitive to correlates of auditory motion in three-dimensional space. *Exp Brain Res, 88*(1), 158-168. doi:10.1007/bf02259137
Sunny, M. M., & von Muhlenen, A. (2011). Motion onset does not capture attention when subsequent motion is "smooth". *Psychon Bull Rev, 18*(6), 1050-1056. doi:10.3758/s13423-011-0152-3

Wightman, F. L., & Kistler, D. J. (1999). Resolution of front-back ambiguity in spatial hearing by listener and source movement. *J Acoust Soc Am, 105*(5), 2841-2853. doi:10.1121/1.426899

Zimmer, U., & Macaluso, E. (2009). Interaural temporal and coherence cues jointly contribute to successful sound movement perception and activation of parietal cortex. *Neuroimage, 46*(4), 1200-1208. doi:10.1016/j.neuroimage.2009.03.022