Hearing Research 337 (2016) 80–88

Contents lists available at ScienceDirect

Hearing Research

journal homepage: www.elsevier.com/locate/heares

Research paper

Effect of motion on speech recognition

Timothy J. Davis*, D. Wesley Grantham, René H. Gifford

Vanderbilt University, Department of Hearing and Speech Sciences, Nashville, TN, USA

ARTICLE INFO

Article history:
Received 22 December 2014
Received in revised form
11 April 2016
Accepted 8 May 2016
Available online 27 May 2016

Keywords:
Motion
Speech recognition
Cocktail party
Spatial hearing

ABSTRACT

The benefit of spatial separation for talkers in a multi-talker environment is well documented. However, few studies have examined the effect of talker motion on speech recognition. In the current study, we evaluated the effects of (1) motion of the target or distracters, (2) *a priori* information about the target and distracter spatial configurations, and (3) target and distracter location. In total, seventeen young adults with normal hearing were tested in a large anechoic chamber in two experiments. In Experiment 1, seven stimulus conditions were tested using the Coordinate Response Measure (Bolia et al., 2000) speech corpus, in which subjects were required to report the key words in a target sentence presented simultaneously with two distracter sentences. As in previous studies, there was a significant improvement in key word identification for conditions in which the target and distracters were spatially separated as compared to the co-located conditions. In addition, 1) motion of either talker or distracter resulted in improved performance compared to stationary presentation (talker motion yielded significantly better performance than distracter motion) 2) *a priori* information regarding stimulus configuration was not beneficial, and 3) performance was significantly better with key words at 0° azimuth as compared to −60° (on the listener's left). Experiment 2 included two additional conditions designed to assess whether the benefit of motion observed in Experiment 1 was due to the motion itself or to the fact that the motion conditions introduced small spatial separations in the target and distracter key words. Results showed that small spatial separations (on the order of 5-8°) resulted in improved performance (relative to co-located key words) whether the sentences were moving or stationary. These results suggest that in the presence of distracting messages, motion of either target or distracters and/or small spatial separation of the key words may be beneficial for sound source segregation and thus for improved speech recognition.

1. Introduction

Complex listening environments, such as those in which a listener needs to attend to one talker while ignoring several others, present well-documented difficulties for listeners both with normal hearing and with hearing loss (e.g., Cherry, 1953; Loizou et al., 2009). Since multi-talker environments are common, many studies have focused on isolating the components contributing to speech understanding in complex listening environments. Nearly all of these studies, however, have used stationary sound sources; that is, stimuli were presented from fixed locations. The effect of moving talkers remains to be investigated. Thus the primary goal of this study was to determine if motion of one or more talkers could provide a benefit to the listener in a multi-talker environment.

Using stationary sound sources, several studies have shown that spatially separating talkers can lead to large improvements in speech understanding compared to conditions in which multiple talkers are presented from a single location (Bronkhorst and Plomp, 1988; Kidd et al., 1994; Freyman et al., 2001). This benefit of spatial separation is commonly referred to as spatial release from masking. In addition to using stationary sound sources, these studies have used spatial positions that were consistent and predictable from one trial to the next.

Kidd et al. (2005) investigated whether knowledge of target talker location affected one's ability to identify key words in the presence of two simultaneous competing messages. In that experiment, three simultaneous messages were presented from three locations (−60°, 0°, +60°). Prior to each stimulus presentation, listeners were told the likelihood of the target sentence coming from a specific location [probability was set to 1.0, 0.8, 0.6, or 0.33

* Corresponding author. 1215 21st Ave S, MCE-South Tower, Room 8310, Nashville, TN 37232, USA.
E-mail address: timothy.j.davis@vanderbilt.edu (T.J. Davis).

http://dx.doi.org/10.1016/j.heares.2016.05.011
0378-5955/

(chance)]. Performance was always highest when the listener was certain (probability = 1.0) of the location of the target talker.

Brungart and Simpson (2007) found that listener certainty did not affect performance in a two-talker environment, but resulted in a 20-percentage point improvement in three and four-talker environments. They also reported that performance was better when the target phrase originated from either the left or right than when it originated from directly in front of the listener. The authors attributed this finding to a greater signal-to-noise ratio in the ear closest to the target, as well as greater interaural timing difference (ITD) for the target than for the distracters.

Many of the above-cited studies utilized the Coordinate Response Measure (CRM) speech corpus (Bolia et al., 2000). All CRM sentences are structured according to the following formula: "Ready [Call sign], go to [Color] [Number] now." Combinations of eight call signs, four colors, and eight numbers create 256 unique sentences for each of 8 talkers in the corpus. The corpus is designed for simultaneous presentation of two or more talkers. The listener is typically tasked with attending to one talker and repeating the color and number spoken by that talker while ignoring the distracter(s). This corpus was designed to maximize informational masking, as the general structure of each sentence is similar in content and timing, and same-gender talkers can be used as target and distracter(s).

Allen et al. (2008) examined the effect of changing the spatial configuration of talkers in the middle of a sentence on spatial unmasking. Instead of presenting talkers in three different fixed positions, as Kidd and colleagues had done, Allen et al. (2008) employed conditions in which the talkers were either 1) spatially separated throughout the sentence, 2) co-located throughout the sentence, 3) switched positions in mid-sentence from co-located to separated (±30°) (i.e. start-co-located), or 4) switched positions in mid-sentence from separated (±30°) to co-located (i.e. start-separated). 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. Given the significant advantages conferred by binaural interactions (ITD) and head shadow effects [interaural level differences (ILD)], it was not surprising that there was a significant advantage for conditions involving some amount of spatial separation. The start-separated condition yielded the smallest release from masking, as the key words occurring at the end of the sentences (color and number) were presented from the same location as the maskers, whereas in the separated and start-co-located conditions, the key words were presented from spatially separate locations.

One of the key findings from Allen et al. (2008) was the 3.6-dB advantage in speech reception threshold (SRT) for the start-separated condition compared to the co-located condition. In the start-separated condition, the call signs occurred when the three messages were spatially separated, but the key words occurred when the messages were co-located at 0° azimuth. Their findings suggested two main conclusions. The first conclusion was that listeners were able to identify the unique characteristics of the target talker's voice and maintain focused attention on the target when the talkers changed locations. The second was that spatial separation can be beneficial even when it occurs for only a portion of the sentence.

The location changes in the conditions studied by Allen et al. (2008) were instantaneous, occurring after the call signs but before the color and numbers, and only involved changing location of the distracters. In the present study, we investigated configurations similar to those employed by Allen and colleagues, but in which the location change was continuous motion of targets and distracters. The first goal of the current study was to determine if position change of talkers during an utterance via smooth motion would lead to a similar benefit as seen by Allen et al. (2008). Based on their results (Allen et al., 2008), we hypothesized that the motion of the target or distracter would facilitate performance to the extent that it would enable listeners to preserve focused attention of target sentences that may have initially been spatially separated, but that were co-located at the time of key-word presentation.

The second goal of the current study was to investigate the effect, in an identification task, of prior knowledge of the spatial configuration of an auditory environment involving moving sources. Based on the previous studies cited, our hypothesis was that prior knowledge of the location and type (moving or non-moving) of source would lead to improved performance compared to conditions in which there was no prior knowledge.

The third goal of the current study was to evaluate the effect of presenting key words directly in front of the listener as well as spatially offset to one side. Mills (1958) showed that sensitivity to differences in azimuth is much poorer at 90° to one side compared to directly in front of the listener. Given that, our hypothesis was that listeners would derive less benefit from motion when the key words were presented on their left than when presented from directly in front. This study involved comparisons of moving and stationary talkers presented simultaneously. Since the talkers were only truly co-located for a brief moment in time, at least a small amount of spatial separation was essentially always present. Brungart and Simpson (2005) found that small spatial separations (on the order of ±10°) in the CRM task were sufficient to obtain high levels of speech understanding. To obtain that same amount of benefit on the listener's side, they found that a much larger separation (60°) was required. The extent to which a moving stimulus is discriminable from a stationary stimulus at 0° and to the listener's left is not yet understood.

2. Experiment 1: method

2.1. Participants

Eleven adults with normal hearing participated in this experiment. Participants were all graduate students at Vanderbilt University. Prior to enrollment, pure tone thresholds were screened at octave frequencies 250–8000 Hz with a GSI-61 audiometer and ER-3A insert earphones. Normal hearing was defined as having air conduction thresholds equal to or lower than 20 dB HL at all tested frequencies. Testing was completed in one two-hour session. Participants were paid for their time.

2.2. Stimuli and test environment

We used the coordinate response measure (CRM) corpus (Bolia et al., 2000) which contains sentences spoken by eight different talkers (4 male, 4 female). Only the male talkers were used in this study to serve as both target and distractor in an effort to minimize additional cues such as fundamental frequency differences of talkers which could aid in target identification. Three different male talkers were presented simultaneously from one or more loudspeakers. The listener was asked to attend to only one talker, referred to as the target talker (T). The target sentence always began with "ready [Baron]". The other two talkers, whose sentences began with "ready [non-target call sign (two unique call signs randomly selected on each trial from the remaining seven call signs)]", were referred to as distracters (D). Thus, on each trial there were three sentences presented by three different talkers (one target plus two distracters), containing three different call signs, three different colors, and three different numbers. The listener's task was to respond with the color and number spoken by the