The Quarterly Journal of Experimental Psychology

ISSN: 1747-0218 (Print) 1747-0226 (Online) Journal homepage: http://www.tandfonline.com/loi/pqje20

Estimating Group Size From Human Speech: Three's a Conversation, but Four's a Crowd

Michael S. Vitevitch & Cynthia S. Q. Siew

To cite this article: Michael S. Vitevitch & Cynthia S. Q. Siew (2015): Estimating Group Size From Human Speech: Three's a Conversation, but Four's a Crowd, The Quarterly Journal of Experimental Psychology, DOI: 10.1080/17470218.2015.1122070

To link to this article: http://dx.doi.org/10.1080/17470218.2015.1122070

Accepted author version posted online: 23 Nov 2015.

Publisher: Taylor & Francis & The Experimental Psychology Society
Journal: *The Quarterly Journal of Experimental Psychology*
DOI: 10.1080/17470218.2015.1122070

Estimating group size from human speech:

Three's a conversation, but four's a crowd

Michael S. Vitevitch and

Cynthia S. Q. Siew

Department of Psychology

University of Kansas

Lawrence, KS 66045

Correspondence should be addressed to:
Michael S. Vitevitch, Ph.D.
Spoken Language Laboratory
Department of Psychology
1415 Jayhawk Blvd.
University of Kansas
Lawrence, KS 66045
e-mail: mvitevit@ku.edu
ph: 785-864-9312

Abstract

Much previous research has examined various aspects of auditory processing, including the localization of sounds, and the influence of lexical and indexical information on language processing. In the present set of experiments we explored the ability of listeners to estimate the number of speakers in a group solely from the information in an auditory signal. The bound on accurately estimating the number of simultaneous speakers is 3. We suggest that subitization—the ability to estimate numerosity of visual and auditory elements without explicitly counting these elements—rather than the capacity of short-term memory, may underlie this limitation. The cognitive constraint on estimating the number of simultaneous speakers may have implications for a wide variety of seemingly unrelated psychological phenomena.

Keywords: short-term memory, subitization, Beau Geste effect, group size, human speech, auditory processing

The world around us contains a rich soundscape. Various cues in the auditory signal enable humans and other animals to locate in space the source of a sound (Morrongiello & Gotowiec, 1990), determine the material that comprises the object (Gaver, 1993), determine whether an object is moving (Vasilenko & Shestopalova, 2010), and estimate the size of an object (Grassi, Pastore, & Lemaitre, 2013) among other things.

One of the most acoustically complex and culturally significant auditory signals processed by humans is speech. Human speech contains numerous cues that enable listeners to comprehend the *linguistic message* being conveyed (e.g., the phonological, semantic, and syntactic information about the words being spoken), as well as a wide range of *indexical properties* that give the listener information about the speaker (Abercrombie, 1967). That is, certain auditory cues in speech provide the listener with information about the speaker including: gender (Gradol & Swann, 1983), age (Krauss, Freyberg, & Morsella, 2002), height and weight (Krauss et al., 2002), identity (Cook & Wilding, 1997), sexual orientation (Munson, 2007), socio-economic class (Ellis, 1967), country or region of origin (Clopper & Bradlow, 2009), certain neurological pathologies (Velasco García, Cobeta, Martín, Alonso-Navarro, & Jimenez-Jimenez, 2011), emotional state (Williams & Stevens, 1972), and state of intoxication (Chin & Pisoni, 1997), among others. The indexical information contained in the speech signal influences decisions related to linguistic properties (Green & Barber, 1981; Houston & Jusczyk, 2000; Vitevitch, Sereno, Jongman, & Goldstein, 2013), as well as decisions about certain competencies of the speaker (Gill, 2009).

It is widely recognized that non-human primates and humans have similar brains and auditory processing abilities, with the exception of areas related to processing speech and language in humans (Kaas, O'Brian, & Hackett, 2013). Despite this difference between primate and human brains, species-specific vocalizations still seem to hold a privileged place in certain cognitive processes. For example, Ng, Plakke, and Poremba (2009) found that *Macaca mulatta* (commonly known as the rhesus macaque) recalled species-specific vocalization sounds more accurately in a

matching-to-sample task than non-vocalization sounds such as music clips, other sounds found in nature (e.g., water rippling, thunder), and man-made sounds (e.g., engine noise, police siren).

Although a large amount of research has been done to investigate diverse aspects of human speech, little work has examined the ability of a human listener to estimate the number of speakers in a group. The lack of research on this ability in humans is surprising given that there has been research on the quantitative abilities of non-human primates (e.g., Nieder & Dehaene, 2009), and research on the use of vocalizations to estimate group size in non-human animals. For example, Harrington (1989) observed that packs of wolves use chorus howling to signal their presence in a particular territory, and to deter other wolves from entering their territory. If only a few animals appear to occupy a given territory, another animal or pack may pass through or attempt to take over the territory. If many animals appear to occupy a territory, another animal or pack may instead avoid that territory, searching elsewhere for a through-path or for territory of their own.

The ability to estimate group size from vocalizations has adaptive value and may play an important role in survival. In non-human animals the ability to estimate accurately the size of a group based on auditory cues alone is crucial, because temporal or geographic constraints may prevent an animal from using visual observation to estimate the number of animals occupying a given territory. That is, the time required to engage in search of the territory or the large size of the territory to be searched make visual observation difficult, costly, and perhaps even risky to carry out. These (and other) constraints often make auditory signals such as vocalizations the sole criterion used to determine the size of a group that may occupy a given territory.

However, humans may rarely rely on auditory cues alone for any type of processing in modern life. Consider that even in the processing of speech visual cues may take precedence over auditory cues, which may lead to illusory perceptions as occurs in the well-known McGurk effect (e.g., Brancazio & Miller, 2005). Although most modern humans living in industrialized settings may

have few instances in which they rely on auditory cues alone, examining the ability of humans to estimate group size using speech has practical value for military personnel tracking the movement of enemy troops, and in helping us understand the influence of crowd size on decision-making. For example, referees may use crowd size as a cue to decide the winner of kickboxing bouts (Myers & Balmer, 2012), or to decide whether to issue fouls in soccer matches (Unkelbach & Memmert, 2010).

Understanding how humans estimate group size using speech also has implications for legal policy, much like research showing that humans inaccurately identify specific speakers, which casts doubt on the use of "voice line-ups" as primary evidence in the courtroom (Clifford, 1980; Clifford & Bull, 1978; Cook & Wilding, 1997; Yarmey, Yarmey, Yarmey, & Parliament, 2001). Finally, understanding this ability in humans may further the development of auditory information displays in computers (Gamper, Dicke, Billinghurst, & Puolamäki, 2013), and of automatic speech recognition systems in computers by providing insight into the process known as speaker diarization (Kotti, Moschou, & Kotropoulos, 2008).

In the present set of experiments we explored the human ability to estimate group size from speech by presenting to listeners sound files that contained from 1 to 10 simultaneous speakers. In Experiments 1-3, listeners were simply asked to indicate how many speakers they heard. Hearing multiple simultaneous speakers might occur when one is socializing at a party, when attending a sporting event in a crowded stadium, or when exercising one's constitutional right to peaceably assemble. Members of the group, individuals monitoring a group, or individuals performing in front of a group may use the auditory cues present in these situations to estimate the size of the group. Despite the simplicity of the task employed in these experiments, the results of these studies provide important information regarding the limits in humans, and (in Experiments 2 and 3) the underlying mechanism responsible for the limitation on the ability to accurately estimate group

size. Given the lack of previous research in this area, we wish to emphasize that the experiments we report should be considered somewhat exploratory in nature.

Experiment 1

Experiment 1 was a preliminary assessment of the ability in humans to estimate group size from speech. In this study we created in the laboratory a listening situation that was controlled, but also provided listeners with—what we intuited to be—numerous cues that could be used to perform the task at near-optimal levels. To control the content of what was produced by the speakers we recorded individual speakers (n = 10) reading from the same text (i.e., the text found in Vitevitch (2002)), and then digitally edited the sound files to create stimuli that contained 1 to 10 speakers (see the Methods section for more details of how the stimuli were edited).

To maximize the difference among the speakers, and thereby provide listeners with a number of auditory cues that could be used to distinguish among the speakers, we selected 5 male and 5 female speakers. The speakers varied in age (from ~18 to 40 years of age), speaking rate, and regional dialect (including the Standard Singaporean English accent, and the Inland-North, North-Central, Midland, and West dialects of American-English). In addition, a discrimination task, in which each voice was paired with every other voice, confirmed that the voices used in the present study were easily distinguishable from each other (see the Methods section).

Furthermore, when more than one speaker was presented to a listener in a sound file the speakers were sampled such that different portions of the text were being read. This ensured that in addition to the acoustic cues that could be used to distinguish among the speakers, additional semantic and syntactic cues would be present in the stimulus as well.

Moreover, the sound files ranged in length from .5s to 30s in length, providing listeners with (what we believed to be in the longest duration) ample amounts of time to accurately estimate (or simply count) the number of speakers present in the stimulus. The combination of these cues and ample stimulus presentation time provided listeners with what we intuited to be an opportunity for near-optimal performance in the simple task we asked listeners to engage in: In each sound file you will hear anywhere

from 1 to 10 speakers; please indicate how many speakers you hear in each sound file. As the results of the present study show, however, our intuitions about this task were wrong.

Method

Participants: Twenty-five native English speakers were recruited from the pool of Introductory Psychology students enrolled at the University of Kansas and received partial course credit for their participation. No participants reported a history of speech or hearing disorders.

Materials: Independent recordings of ten speakers (5 males, 5 females) varying in age (from 19 to 41 years of age), regional dialect, and speaking rate were made as each speaker read the same text (i.e., the text found in Vitevitch (2002)). All speakers produced these recordings by speaking at the rate that was normal for each individual into a high-quality microphone in an Industrial Acoustics Company (IAC) sound-attenuated booth. These were digitally recorded at a sampling rate of 44.1 kHz.

Samples of .5s, 1s, 5s, 15s and 30s were extracted for each speaker and combined to create 50 stimuli that contained 1 to 10 speakers, such that all speakers were represented evenly across conditions, and different sections of the text were read. To prevent participants from using overall amplitude as a cue to the number of speakers, sound files of each speaker were normalized in proportion to the number of speakers in each stimulus. No additional modifications were made to the auditory signals.

To verify that the voices used in the experiment could be readily discriminated from each other, ten participants indicated whether 1s stimuli containing just 1 speaker were produced by the same or different speakers for every possible pair of speakers. Accurate performance in this task (M = 0.94, SD = 0.04) indicates that the voices were discriminable, and suggests that performance in the experiment is unlikely to be the result of voices that resembled each other.

Procedure: Participants were tested in groups of 2 or 3. Each participant was seated in front of an iMac running PsyScope 1.2.2 (Cohen, MacWhinney, Flatt, & Provost, 1993), which controlled stimulus presentation and collected the participants' responses.

Participants listened to sound clips containing 1 to 10 simultaneous speakers that varied in duration from .5s to 30s, and indicated the number of speakers in each sound clip. Participants were informed that the number of speakers in each sound clip ranged from 1 to 10 speakers, and used the keyboard to enter their responses (a numerical value from 1 to 10).

Prior to the experiment, participants received 5 practice trials to familiarize themselves with the task. These practice trials were not included in the subsequent analyses. In each trial, the word "READY" was displayed on the screen for 500ms, and one of the 50 sound clips was randomly played through a set of BeyerDynamic DT100 headphones at a comfortable listening volume. Each sound clip was presented only once.

Results

In Figures 1 and 2 the diagonal (blue) line indicates the response that corresponds to correct performance in the task (when presented with a sound file containing 4 speakers, the listener should have responded with 4). The jagged (red) line in Figure 1 shows the mean response (i.e., how many speakers the listeners reported hearing) from 25 participants to the .5s stimuli. The jagged (red) line in Figure 2 shows the mean response from 25 participants to the 30s stimuli. The mean response data from the other durations (1s, 5s, and 15s) are similar to those displayed in Figures 1 and 2. The consistency in performance across stimulus duration was striking.

(Figure 1 about here)

(Figure 2 about here)

What is more striking, however, is that performance is accurate (i.e., the red line overlays the blue line) only when presented with 1-3 speakers in a sound file. When listeners were presented with 4-10 speakers, they did not respond with the correct number of speakers that were present in the stimulus. As

shown in Table 1, the results of one-sample *t*-tests (using the correct response as the population mean, the mean response across durations to compute the sample mean and standard deviation, $n = 25$, and a two-tailed test[1]) confirm that performance is indistinguishable from the correct response for sound files that contained 1-3 speakers, but differed significantly (i.e., the listeners responded incorrectly) for sound files that contained 4-10 speakers.

(Table 1 about here)

We were further struck that the mean response when the stimuli contained 6-10 simultaneous speakers was approximately 5 speakers. Recall that participants were informed prior to the start of the experiment that the stimuli would contain 1 to 10 speakers. We found it curious that listeners did not respond in a biased manner with a value closer to 10 when the stimuli contained 6-10 speakers.

Discussion

In the present study, participants listened to sound clips containing 1 to 10 simultaneous speakers that varied in duration from .5s to 30s. Participants were asked to indicate the number of different voices heard in each sound clip. Our intuition led us to believe that there were ample acoustic, semantic, and syntactic cues available in the speech stimuli to enable listeners to perform at near-optimal levels in this simple task. However, this was not the case.

Across all durations, participants were accurate in identifying 1 to 3 speakers. Performance was inaccurate when the stimuli contained 4 or more speakers. The arguably poor performance of the listeners in this task was, therefore, somewhat surprising. We were further surprised that performance was poor even at the longest duration (30s), where we believed listeners would have more than enough time to distinguish among—and, therefore, explicitly count—the number of speakers present in the stimulus.

Why was performance so poor for the sound files that contained 4-10 speakers? One possible explanation for the poor performance relates to what one might call "external" causes, or problems in the signal being presented to listeners. Perhaps phenomena such as informational masking or energetic masking simply prevented listeners from being able to perceive more than 3 speakers in a group.

In *informational masking* the ability to discriminate or detect an auditory signal becomes worse when the signal is embedded in a context of other similar sounds (Leek, Brown, & Dorman, 1991). Intuitively one might predict that with more speakers present in the signal there would be more informational masking and worse performance. Recall, however, that a separate sample of listeners was quite accurate (94% correct) in discriminating among every possible pair of speakers, suggesting that the voices used in the present experiment (which naturally varied in speaking rate, regional dialect, age, gender, etc.) were not that similar to one another, casting doubt on the possibility that informational masking may be the cause of the poor performance on stimuli that contained 4 or more speakers.

Furthermore, with the addition of more voices to the signal the number of dimensions along which they may be distinguished increases. For example, in the case of two male voices one might simply label them as "faster male" and "slower male." With three male voices one might label them as "faster male," "slower male," and "male with creaky voice," making the voices more, not less, distinct. Perhaps contrary to one's initial intuitions regarding the influence of informational masking on performance, adding more speakers to the signal increases the number of dimensions on which speakers could be discriminated, thereby making each speaker more distinct. With more dimensions to discriminate among speakers, listeners should have done much better on stimuli that contained 4 or more speakers. Given that we observed poor not better performance in the present experiment, further doubt is cast on the possibility that informational masking may underlie the poor performance on stimuli that contained 4 or more speakers.

In *energetic masking* physical interactions between the sound waves leads to degradation of the signal (Leek et al., 1991). Consider the case of two sine-waves that are 180 degrees out of phase; the result of adding the two waves together is that they cancel each other out (i.e., a flat line). The

phenomenon of the two waves cancelling each other out is also known as *destructive interference*. One might intuitively predict that adding more speakers to the signal would increase the amount of destructive interference, and make performance worse for stimuli that contained 4 or more speakers.

However, there is related phenomenon known as *constructive interference*, in which two waves combine to amplify the signal, which would occur if the two sine waves were in phase (instead of out of phase as in the case of destructive interference). While it is possible that destructive interference could have occurred in the present experiment, given the natural variation in speaking rate, duration of pauses between sentences and clauses, etc., it is perhaps more likely that increasing the number of speakers increased the likelihood (especially in the stimuli with longer durations) that some portion of the signals would be in-phase and lead to constructive interference for at least one of the voices, or for several different voices at several different points in the stimuli. If constructive interference occurred for the stimuli with many voices, it would be easier not more difficult (as was observed in the present experiment) to identify several different speakers in the stimuli that contained 4 or more speakers, casting doubt on the idea that energetic masking might be responsible for the poor performance on stimuli that contained 4 or more speakers.

Having considered several "external" causes for the observed results, we now consider an "internal" cause related to certain cognitive mechanisms or limitations that might account for the poor performance of listeners on stimuli that contained 4 or more speakers. In his work on chorus howling in wolves, Harrington (1989; p. 134) states that: "…it is likely that there is an upper limit to perceived pack size, governed by the capacity of the wolf's short term auditory processing system. For humans, the limits of short-term memory are between five and nine items (Miller, 1956); similar limits would be expected for wolves." We note that Harrington did not elaborate any further on this point, and that we were unable to find any published reports that confirm the short-term memory capacity of the wolf or other species of *Canis*. However, lacking any other obvious account of the present results, we pursued this admittedly speculative hypothesis described briefly in Harrington (1989): the capacity of short-term memory in humans may be responsible for the poor ability of humans to estimate group size from speech observed in

the present study. More recently, however, Cowan (2001) argued for a more conservative estimate of 4 items as the capacity of short-term memory. Given that performance in the present task declined significantly at 4 voices, we decided to explore in Experiment 2 the hypothesis that limits in short-term memory may be responsible for the poor ability to estimate group size from speech.

Experiment 2

In the present experiment we sought to test in humans the admittedly speculative hypothesis proposed by Harrington (1989) that short-term memory is the limiting factor in estimating the number of individuals in a group from vocalizations. Varying the memory load that participants experience while they estimate the number of speakers represents one way to test whether the capacity of short-term memory accounts for the limited ability of humans to accurately determine the number of simultaneous speakers from speech. The load on short-term memory was manipulated by asking participants to remember either 1 or 4 letters of the alphabet while engaging in the same task used in Experiment 1: estimate the number of speakers in each stimulus.

In the low memory load condition, participants were asked to remember 1 letter while estimating the number of speakers in each stimulus. We hypothesized that this low load on short-term memory should do little to affect the performance of listeners estimating the number of speakers in the stimulus. Thus, performance in the low memory load condition should replicate the performance observed in Experiment 1, where there was no additional memory load imposed on the listeners.

In the high memory load condition, participants were asked to remember 4 letters while estimating the number of speakers in each stimulus. We hypothesized that this higher load on short-term memory should reduce the capacity of short-term memory, and therefore cause performance in the estimation task to suffer. That is, when under a high memory load, participants should be overall less accurate in their estimates of group size.

Method

Participants: Forty individuals from the same population that was sampled in Experiment 1 took part in the present experiment. None of the participants who took part in the present experiment took part in the previous study.

Materials: One hundred sound files (10 for each speaker condition of 1 to 10 speakers in the stimulus) that were 1s in duration were sampled from the 30s sound files used in Experiment 1. Half of the sound files were presented in the high memory load condition, and the remaining sound files were presented in the low memory load condition.

In the high load condition, participants had to recall a randomly generated string of 4 letters (e.g., SQTC) whereas in the low load condition, participants had to recall 1 letter (e.g., H). All letters were presented in upper case to prevent the use of letter shape or other visual features of the sequence to aid recognition (Perea & Rosa, 2002). Furthermore, only consonants were presented to prevent the formation of pronounceable nonwords (e.g., BAPE), thereby limiting the use of phonological cues to retrieve the letters (Schweickert, 1993).

Procedure: As in Experiment 1, participants were tested in groups of 2 or 3. Participants were informed that the number of speakers in each sound file ranged from 1 to 10 speakers, and that their task was to identify the number of speakers in each sound file. In addition, participants were informed that they had to remember the letters shown prior to the presentation of the sound file, and that recall of the letters would be tested after the participant estimated the number of speakers in the sound file.

At the start of each trial, either 1 or 4 letters were shown on the screen for 1s, and then a 1s sound file was played. Participants indicated the number of speakers in each sound file, and typed in the letters that they could recall. There were 100 trials in total, and the presentation of high- and low-load trials was randomized.

Results

Responses were coded as correct if participants typed in the correct letters in the order that they were presented. Recall accuracy for the letters was 63.5% (*SD* = 21.8%) for the high load condition and 87.9% (*SD* = 7.6%) for the low load condition, suggesting that the high load condition indeed was cognitively taxing to the listeners.

A two-way within-participants ANOVA was conducted on the number of speakers reported by the participants. The two independent variables were memory load (high vs. low) and number of speakers (1 to 10). This analysis was conducted on (1) the responses from all trials, and (2) only for the responses in which participants managed to correctly recall the letters. As the overall findings were similar for both analyses, we report only the analyses in which all trials were included so as to maximize statistical power in the analysis.

As shown in Table 2, participants reported similar numbers of speakers in the high and low memory load conditions, and this was true across all 10 speaker conditions. The results of the statistical analysis revealed only a significant main effect of number of speakers, such that participants reported hearing more speakers for sound files that indeed contained more speakers (F (9, 351) = 178.3, $p < .001$). Neither the main effect of load (F (1, 39) < 1, $p = .90$) nor the interaction between load and number of speakers (F (9, 351) = 1.35, $p = .21$) were significant.

(Table 2 about here)

Interestingly, despite participants being informed that the stimuli would contain 1 to 10 speakers, the average number of speakers again peaked at about 5 speakers across all speaker conditions, mirroring the results of Experiment 1.

Discussion

If our hypothesis derived from the work of Harrington (1989) regarding the limitations of short-term memory being responsible for the poor ability to accurately estimate the number of speakers was correct, then participants would have been more accurate at estimating the number of speakers in the low memory load condition, where memory resources were not as taxed by the letter recall task, as compared to the high memory load condition. This was not the case. No difference was observed between the high

and low load conditions, suggesting that short-term memory may have little to no influence on the ability to accurately estimate the size of a group from simultaneous speech.

To test if the failure to find a significant effect of memory load in the present experiment was due to a lack of statistical power, a post-hoc power analysis was conducted using the GPower3 software (Faul, Erdfelder, Lang, & Buchner, 2007). Based on a small effect size ($f = 0.1$), alpha level of .05, and a sample size of 40, the estimated power for this experiment was 0.86, which is greater than the recommended power level of 0.80. Therefore, if short-term memory capacity was involved in the ability to estimate the number of speakers, there was sufficient power in the present experiment to detect that influence.

If limits in short-term memory are not responsible for the poor performance in estimating group size from speech, then what is responsible for this (in)ability to estimate group size from speech? A closer examination of Figures 1 and 2, and Tables 1 and 2 shows that listeners were accurate when presented with 1-3 speakers, but were inaccurate when presented with more than 3 speakers. The value of 3 is consistent with the threshold associated with the cognitive phenomenon known as subitization. *Subitization* refers to the ability to quickly and accurately estimate the number of elements in an array without explicitly counting them (Kaufman, Lord, Reese, & Volkmann, 1949). When the number of objects is greater than the subitization threshold of 3 elements, subitization breaks down and people have to count the number of elements in order to report numerosity. Although most work on subitization has studied visually presented stimuli, the limit of 3 elements appears to be true of tactile (Riggs et al., 2006) as well as auditory stimuli (Repp, 2007), and may also play a role in action planning (Gallivan et al., 2011).

Experiment 3

Subitization refers to the ability to quickly and accurately report the number of objects in an array without engaging in serial counting (Kaufman et al., 1949). This phenomenon is typically studied by presenting participants with visual arrays containing 1 to 200 dots (or other shapes) for very short durations ranging from 200ms to 800ms, and asking the participants to report the number of elements present in the display as quickly and accurately as possible (Gallistel & Gelman, 1992; Kaufman et al.,

1949; Mandler & Shebo, 1982). Subitization occurs for small display sizes containing 1 to 3 elements, while counting typically occurs for display sizes with 4 or more elements.

Although most of these studies used visual displays, there are a handful of studies that used auditory stimuli (McLachlan, Marco, & Wilson, 2012; Repp, 2007; Thurlow & Rawlings, 1959). For instance, Repp (2007) investigated the enumeration of sequential auditory events by asking participants to tap in synchrony with every n^{th} tone. Repp found that tapping was most efficient when n could be divided into groups of 2 or 3, suggesting that subitization was occurring for the auditory stimuli and yielding a limit similar to that of visual displays: accurate and rapid numerosity judgments are generated when the number of elements ranges from 1 to 3. Given that participants in the present studies were most accurate at estimating group sizes of 1 to 3 speakers, it is possible that subitization may be responsible for the highly accurate responses for group sizes of 1 to 3 observed in the present experiments.

To investigate if subitization might account for the limited ability to accurately estimate the number of speakers, participants listened to sound files of very short durations (< 1s) in Experiment 3. The very short durations of the sound files used in the present experiment are within the duration range of visually presented stimuli, and would prevent participants from explicitly counting the number of speakers. If the mechanisms related to the process of subitization also influence the ability to estimate the number of speakers from speech, accurate performance for 1-3 speakers would be expected despite the very short durations of stimuli used. However, estimates for 4-10 speakers will again be inaccurate.

Method

Participants: Thirty individuals from the same population that was sampled in the previous experiments took part in the present experiment. None of the participants who took part in the present experiment took part in the other experiments reported here.

Materials: Sound files were created by selecting samples of 100ms, 200ms, 300ms, 400ms and 500ms from the recordings of 10 speakers to create 50 stimuli that contained 1 to 10 speakers. As in Experiment 1, stimuli were created such that all speakers were represented evenly across conditions, and different sections of the text were read.

Procedure: The procedure was similar to that in Experiment 1. Participants were informed that the number of speakers in each sound file ranged from 1 to 10 speakers, and their task was to identify the number of speakers in each stimulus. A response was correct if the number indicated by the participant corresponded to the number of speakers in the stimulus. Prior to the experiment, participants received 5 practice trials to familiarize themselves with the task. These practice trials were not included in the subsequent analyses.

Results

As in Experiment 1 we again used one-sample *t*-tests (using the correct response as the population mean, the mean response across durations to compute the sample mean and standard deviation, $n = 30$, and two-tails) to confirm that performance was indistinguishable from the correct response for sound files that contained 1-3 speakers, but differed significantly (i.e., the listeners responded incorrectly) for sound files that contained 4-10 speakers (see Figure 3). The results of the one-sample *t*-tests are shown in Table 3.

(Table 3 about here)

(Figure 3 about here)

Discussion

Despite the durations of the sound files being less than 1s long in the present experiment, the results of Experiment 3 are very similar to those of Experiment 1: participants could accurately estimate the number of simultaneously presented speakers when presented with 1-3 speakers in the sound file, but were inaccurate when presented with 4-10 speakers in the sound file. Due to the extremely short durations of the sound files in the present experiment, it is unlikely that participants had time to count the number of speakers, suggesting that accurate performance for 1-3 speakers could be due to the process of subitization instead of limitations in short-term memory as suggested by Harrington (1989). In Experiment 4 we examined further the hypothesis that the

process of subitization may underlie listeners' estimation of the number of simultaneous speakers in Experiments 1-3.

Experiment 4

To further examine the hypothesis that subitization may be involved in the estimation of the number of simultaneous speakers, we conducted a simple categorization task in which listeners were asked to indicate if the 1s sound file they heard contained "few" or "many" speakers, instead of asking for a specific value to estimate the number of speakers as in Experiments 1-3. It is understood that subitization is a non-verbal ability, as shown from the well-established ability of infants and children—who have yet to acquire the ability to count explicitly—to subitize (Dehaene, 1992). By removing the need to use verbal, linguistically based representations of numbers to perform the task, the results of Experiment 4 can more directly test if subitization is involved in the estimation of group size from speech.

If subitization is involved in the estimation of group size from speech, we predicted that listeners would categorize sound files with 1-3 speakers as having "few" speakers, because the number of speakers could be directly perceived. In contrast, we predicted that listeners would not be able to subitize the number of speakers when presented with 4-10 simultaneous speakers. Given that sound files containing 4-10 speakers would be indistinguishable from each other, we predicted that listeners would label them as containing "many" speakers.

Method

Participants: Thirty individuals from the same population that was sampled in the previous experiments took part in the present experiment. None of the participants that took part in the present experiment took part in the other experiments reported here.

Materials: Fifty sound files (5 for each speaker condition of 1 to 10 speakers) that were 1s in duration were sampled from the 30s sound files used in Experiment 1. These stimuli were the same stimuli used in Experiment 2.

Procedure: Participants listened to each sound clip and indicated by pressing the appropriately labeled button on a response box whether they heard few or many speakers. Participants were instructed to not explicitly count or identify the number of speakers for each sound clip, but rather to make their decisions based on their own subjective judgments of the stimuli.

Results

One-sample *t*-tests using 0 as the population mean (whereby proportion of responses for "many speakers" was zero), the mean proportions of "many speakers" responses across speakers to compute the sample mean and standard deviation, and $n = 30$ revealed that participants rated sound files that contained 1 or 2 speakers as containing "few speakers", but rated sound files that contained 3-10 speakers as containing "many speakers" (see Figure 4). The results of the one-sample *t*-tests (with two-tails) are shown in Table 4.

(Table 4 about here)

(Figure 4 about here)

Discussion

The results of Experiment 4 are mostly consistent with our prediction that participants would rate sound clips with 1-3 speakers as having "few speakers" and sound files with 4-10 speakers as having "many speakers" (the value of 3 speakers was closer to the "many" category than to the "few" category contrary to what we predicted). This suggests that subitization rather than short-term memory processes may be involved in the estimation of the number of simultaneous speakers. When sound clips contained fewer than 3 speakers, participants may be able to directly engage in subitization to perceive the number of simultaneous speakers, leading them to indicate that "few speakers" were present in the sound clip. When sound clips contained 3 or more speakers, they were unable to subitize or to accurately perceive the number of simultaneous speakers, leading them to categorize these sound clips as having "many speakers."

General Discussion

In the present set of experiments we explored the less-studied ability of listeners to estimate the number of speakers in a group solely from auditory input (i.e., speech). In Experiments 1-3 we found that listeners could accurately indicate that up to 3 simultaneous speakers were present in an auditory signal, but that listeners were not able to accurately estimate the number of speakers in a group that contained 4-10 speakers. We then focused this preliminary investigation on the underlying cause of the 3-speaker limit, testing and rejecting the hypothesis that the capacity of short-term memory underlies the limit, but finding some evidence to support the hypothesis that subitization may underlie the observed limit. Given that several species of vertebrate and non-vertebrate, such as monkeys, fish, and bees, are able to subitize, or quickly and accurately estimate and discriminate between small sets of items (Agrillo, Piffer, Bisazza, & Butterworth, 2012; Beran, Decker, Schwartz, & Schultz, 2011; Dacke & Srinivasan, 2008), it is perhaps not surprising that the limit observed in the present set of experiments is 3 speakers.

Interestingly, Dunbar (2004) has suggested that cognitive limitations may underlie certain patterns found in human social interaction. As observed in Experiment 4 of the present study, the subitization-related limit of 3 speakers was at the boundary between a distinguishable number of individuals (i.e., a "few") and "many" speakers. It may not be a coincidence that the same minimum number of people is required to exert pressure on an individual to conform to the behavior of the group (Bond, 2005). For example, Milgram, Bickman, and Berkowitz (1969; see also Gallup et al., 2012) had an increasing number of actors stop on a city street and stare at a designated point on a building. When the "crowd" contained more than 3 people, the probability that an unsuspecting passer-by would join the "crowd" in staring at the building increased dramatically.

Similarly, Linguistics has documented the existence of languages, like English, that make a distinction between singular (e.g., cat) and plural (e.g., cats). There are also languages, like Arabic, that distinguish between singular, dual (i.e., two items), and plural (more than 2). Some Austronesian languages distinguish between singular, dual (i.e., two items), trial (i.e., three items),

and quadral (4 or more). However, there is some debate about the existence of languages that distinguish between 1, 2, 3, 4, and 5 or more items (Corbett, 2000). Thus, the limited cognitive ability to distinguish 3 speakers/people from a "crowd" of four or more speakers/people may be related to a number of other psychological phenomena, opening the door for several new research directions.

One new research direction may intersect with the large body of literature on voice perception, which refers broadly to abilities associated with perceiving the paralinguistic information in voices that indicate the identity of a speaker. Belin, Fecteau, and Bédard (2004) suggest that voice perception may be similar to face perception, because a series of analogous, complex cognitive and neurological processes are involved in both voice and face perception; indeed, they go so far as to call the human voice an "auditory face." If voice and face perception are similar, then the present studies make a unique and testable prediction regarding face perception: viewers are likely to be able to identify a maximum of 3 faces in a "crowd" that contains 4-10 faces.

The poor ability to correctly estimate the number of speakers could also have key implications for legal policy and judicial proceedings. For example, the tendency to underestimate the number of speakers when there are actually more speakers should be considered when a witness states, "I heard X-number of people talking in the alley," as testimony in court.

The present results might also have implications for automatic speech recognition systems. A better understanding of the human cognitive limitations associated with estimating the number of speakers could ultimately lead to the development of better algorithms for successful speaker diarization—that is, partitioning auditory input according to speaker identity—as well as other automatic speech recognition related processes.

Further, the present studies focused on the accurate performance of listeners in the range of 1-3 speakers, leaving much to explore about the range containing 4-10 speakers. Specifically, why

did listeners greatly *under*estimate the number of speakers, especially when they were informed prior to the start of the experiment that they would hear 1-10 speakers? In his work on the chorus howling of wolves, Harrington (1989) described a phenomenon known as the *Beau Geste effect*, where wolves and birds modulate their vocalizations to give the illusory appearance of more conspecifics in a given pack or territory than there actually are (Harrington, 1989; Krebs, 1977; cf., Yasukawa & Searcy, 1985). For example, certain bird species employ a large repertoire of songs and sing from several different trees to give the false impression of a densely "occupied" habitat, deterring others from settling there. Likewise, the chorus howling of two wolves is indistinguishable from the chorus howling of a pack containing five or more wolves, enabling small wolf packs to deceive competing packs into overestimating the true size of the pack (Harrington, 1989). Why did the human listeners in our experiments *under*estimate instead of *over*estimate the number of speakers as in the Beau Geste effect?

Although there has been much research on various types of auditory processing, and a great deal of research on the processing of speech and voice information, the present set of exploratory studies points to a new area of research on the (in)ability of humans to accurately estimate the number of speakers in a group. The simple findings reported here have implications for a number of topics in psychophysics, psychology, and cross-species comparisons, as well as practical implications for a number of areas. There is still much to learn about the very human ability of speech.

References

Abercrombie, D. (1967). *Elements of general phonetics.* Chicago, IL: Aldine.

Agrillo, C., Piffer, L., Bisazza, A., & Butterworth, B. (2012). Evidence for two numerical systems that are similar in humans and guppies. *PloS one, 7*(2), e31923.

Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences, 8*, 129-135.

Beran, M. J., Decker, S., Schwartz, A., & Schultz, N. (2011). Monkeys (Macaca mulatta and Cebus apella) and human adults and children (Homo sapiens) compare subsets of moving stimuli based on numerosity. *Frontiers in Psychology, 2*.

Bond, R. (2005). Group size and conformity. *Group Processes & Intergroup Relations, 8*, 331-354.

Brancazio, L., & Miller, J. L. (2005). Use of visual information in speech perception: Evidence for a visual rate effect both with and without a McGurk effect. *Perception & Psychophysics, 67*, 759-769.

Chin, S. B., & Pisoni, D. B. (1997). *Alcohol and Speech*. Brill Academic Publishers, Inc.

Clifford, B. R. (1980). Voice identification by human listeners: On earwitness reliability. *Law and Human Behavior, 4*, 373-394.

Clifford, B. R., & Bull, R. (1978). T*he psychology of person identification*. Routledge & Kegan Paul London.

Clopper, C. G., & Bradlow, A. R. (2009). Free classification of American English dialects by native and non-native listeners. *Journal of Phonetics, 37*, 436-451.

Cohen, J., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, 25*, 257-271.

Cook, S., & Wilding, J. (1997). Earwitness testimony: Never mind the variety, hear the length. *Applied Cognitive Psychology, 11*, 95-111.

Corbett, G. (2000). *Number*. Cambridge University Press.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Beh*avioral and Brain Sciences, 24*, 87-114.

Dacke, M., & Srinivasan, M. V. (2008). Evidence for counting in insects. *Animal Cognition, 11*, 683-689.

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition, 44*, 1-42.

Dunbar, R. I. M. (2004). Gossip in evolutionary perspective. *Review of General Psychology, 8*, 100-110.

Ellis, D. S. (1967). Speech and social status in America. *Social Forces, 45*, 431–437.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191.

Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition, 44*, 43-74.

Gallivan, J. P., Chapman, C. S., Wood, D. K., Milne, J. L., Ansari, D., Culham, J. C., & Goodale, M. A. (2011). One to four, and nothing more nonconscious parallel individuation of objects during action planning. *Psychological Science, 22*, 803-811.

Gallup, A. C, Hale, J. J., Sumpter, D. J. T., Garnier, S., Kacelnik, A., Krebs, J. R., & Couzin, I. D. (2012). Visual attention and the acquisition of information in human crowds. *Proceedings of the National Academy of Science, 109*, 7245-7250.

Gamper, H., Dicke, C., Billinghurst, M., & Puolamäki, K. (2013). Sound sample detection and numerosity estimation using auditory display. *Association for Computing Machinery Transactions on Applied Perception, 10*, 1, Article 4.

Gaver, W. W. (1993). What in the world do we hear? An ecological approach to auditory event perception. *Ecological Psychology, 5*, 1-29.

Gill, M. M. (1994). Accent and stereotypes: Their effect on perceptions of teachers and lecture comprehension. *Journal of Applied Communication Research, 22*, 348-361.

Gradol, D., & Swann, J. (1983). Speaking fundamental frequency: Some physical and social correlates. *Language and Speech, 26*, 351–366.

Grassi, M., Pastore, M., & Lemaitre, G. (2013). Looking at the world with your ears: How do we get the size of an object from its sound? *Acta Psychologica, 143*, 96-104.

Green, E. J., & Barber, P. J. (1981). An auditory Stroop effect with judgments of speaker gender. *Perception & Psychophysics, 30,* 459-466.

Harrington, F. H. (1989). Chorus howling by wolves: acoustic structure, pack size and the Beau Geste effect. *Bioacoustics, 2*, 117-136.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology-Human Perception and Performance, 26*, 1570-1582.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American Journal of Psychology, 62*, 498-525.

Kotti, M., Moschou, V. & Kotropoulos, C. (2008). Review: Speaker segmentation and clustering. Si*gnal Processing, 88*, 1091-1124.

Krauss, R. M., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology, 38*, 618–625.

Krebs, J. R. (1977). The significance of song repertoires: the Beau Geste hypothesis. *Animal Behaviour, 25*, 475-478.

Leek, M. R., Brown, M. E., & Dorman, M. F. (1991). Informational masking and auditory attention. *Perception & Psychophysics, 50*, 205-214.

Mandler, G., & Shebo, B. J. (1982). Subitizing: an analysis of its component processes. *Journal of Experimental Psychology: General, 111*, 1.

McLachlan, N. M., Marco, D. J. T., & Wilson, S. J. (2012). Pitch Enumeration: Failure to Subitize in Audition. *PloS one, 7(4)*, e33661.

Milgram, S., Bickman, L., & Berkowitz, L. (1969). Note on the drawing power of crowds of different size. *Journal of Personality and Social Psychology, 13*, 79-82.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*, 81.

Morrongiello, B. A., & Gotowiec, A. (1990). Recent advances in the behavioral study of infant audition: The development of sound localization skills. *Journal of Speech-Language Pathology and Audiology, 14*, 51-63.

Munson, B. (2007). The acoustic correlates of perceived masculinity, perceived femininity, and perceived sexual orientation. *Language and Speech, 50*, 125-142.

Myers, T. D., & Balmer, N. J. (2012). The Impact of Crowd Noise on Officiating in Muay Thai: Achieving External Validity in an Experimental Setting. *Frontiers in Psychology, 3*, 346.

Nieder, A., & Dehaene, S. (2009). Representation of number in the brain. *Annual Review of Neuroscience, 32,* 185-208.

Ng, C.-W., Plakke, B., & Poremba, A. (2009). Primate auditory recognition memory performance varies with sound type. *Hearing Research, 256*, 64-74.

Perea, M., & Rosa, E. (2002). Does 'whole-word shape' play a role in visual word recognition? *Perception & Psychophysics, 64*, 785-794.

Repp, B. H. (2007). Perceiving the numerosity of rapidly occurring auditory events in metrical and nonmetrical contexts. *Attention, Perception, & Psychophysics, 69*, 529-543.

Riggs, K. J., Ferrand, L., Lancelin, D., Fryziel, L., Dumur, G., & Simpson, A. (2006). Subitizing in tactile perception. *Psychological Science, 17,* 271–272.

Schweickert, R. (1993). A multinomial processing tree model for degradation and redintegration in immediate recall. *Memory & Cognition, 21*, 168-175.

Thurlow, W. R., & Rawlings, I. L. (1959). Discrimination of number of simultaneously sounding tones. *Journal of the Acoustical Society of America, 31*, 1332-1336.

Unkelbach, C., & Memmert, D. (2010). Crowd noise as a cue in referee decisions contributes to the home advantage. *Journal of Sport & Exercise Psychology, 32*, 483-498.

Vasilenko, Y. A., & Shestopalova, L. B. (2010). Moving sound source discrimination in humans: Mismatch negativity and psychophysics. *Human Physiology, 36*, 139-146.

Velasco García, M. J., Cobeta, I., Martín, G., Alonso-Navarro, H., & Jimenez-Jimenez, F. J. (2011). Acoustic analysis of voice in Huntington's disease patients. *Journal of Voice, 25*, 208-217.

Vitevitch, M. S. (2002). Naturalistic and Experimental Analyses of Word Frequency and Neighborhood Density Effects in Slips of the Ear. *Language and Speech, 45*, 407-434.

Vitevitch, M. S., Sereno, J., Jongman, A., & Goldstein, R. (2013). Speaker Sex Influences Processing of Grammatical Gender. *PLOS ONE 8(11): e79701.*

Williams, C. E. & Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *Journal of the Acoustical Society of America, 52*, 233–248.

Yarmey, A. D., Yarmey, A. L., Yarmey, M. J., & Parliament, L. (2001). Commonsense beliefs and the identification of familiar voices. *Applied Cognitive Psychology, 15*, 283-299.

Yasukawa, K., & Searcy, W. A. (1985). Song repertoires and density assessment in red-winged blackbirds: further tests of the Beau Geste hypothesis. *Behavioral Ecology and Sociobiology, 16*, 171-175.

Figure Captions

Figure 1. Mean response (and standard error) for each speaker condition at .5s duration.

Figure 2. Mean response (and standard error) for each speaker condition at 30s duration.

Figure 3. Mean response (and standard error) for each speaker condition at 100ms duration.

Figure 4. Mean proportion (and standard error) of responses indicating "many speakers" for each speaker condition.

Table 1. Mean responses (collapsed across duration) compared to the correct answer.

	Speakers									
	1	2	3	4	5	6	7	8	9	10
Mean response	1.01	2.05	2.98	3.38	4.18	4.56	4.80	5.40	6.03	5.42
SD	0.04	0.22	0.75	0.88	1.20	1.47	1.72	1.75	1.99	1.76
t-value	1.00	1.10	0.16	3.56	3.42	4.89	6.38	7.44	7.45	13.01
p-value	.33	.28	.87	.002	.002	<.001	<.001	<.001	<.001	<.001

Note: The mean responses in Experiment 1 for speakers 1-3 were indistinguishable from the correct answer (i.e., they were correct), but for speakers 4-10 the mean responses were significantly different from the correct answer (i.e., they were wrong).

Table 2. Mean number of speakers reported by participants across high and low memory load conditions for each speaker condition in Experiment 2.

	Number of speakers									
	1	2	3	4	5	6	7	8	9	10
High load	1.00 (0.00)	2.14 (0.26)	2.98 (0.51)	3.32 (0.75)	3.52 (0.96)	4.46 (1.32)	4.37 (1.17)	4.59 (1.38)	4.93 (1.36)	5.15 (1.54)
Low load	1.00 (0.00)	2.06 (0.19)	2.89 (0.65)	3.27 (0.76)	3.45 (0.77)	4.41 (1.23)	4.50 (1.53)	4.75 (1.43)	4.77 (1.34)	5.32 (1.69)

Note: Standard deviations are in parentheses.

Table 3. Mean responses (collapsed across duration) compared to the correct answer.

	Speakers									
	1	2	3	4	5	6	7	8	9	10
Mean response	1.05	2.01	2.71	2.99	3.29	3.52	3.75	3.68	3.93	4.11
SD	0.15	0.84	0.93	1.14	1.32	1.41	1.68	1.59	1.73	1.81
t-value	1.83	0.00	1.71	4.85	7.10	9.63	10.60	14.88	16.05	17.82
p-value	.08	1.00	.10	<.001	<.001	<.001	<.001	<.001	<.001	<.001

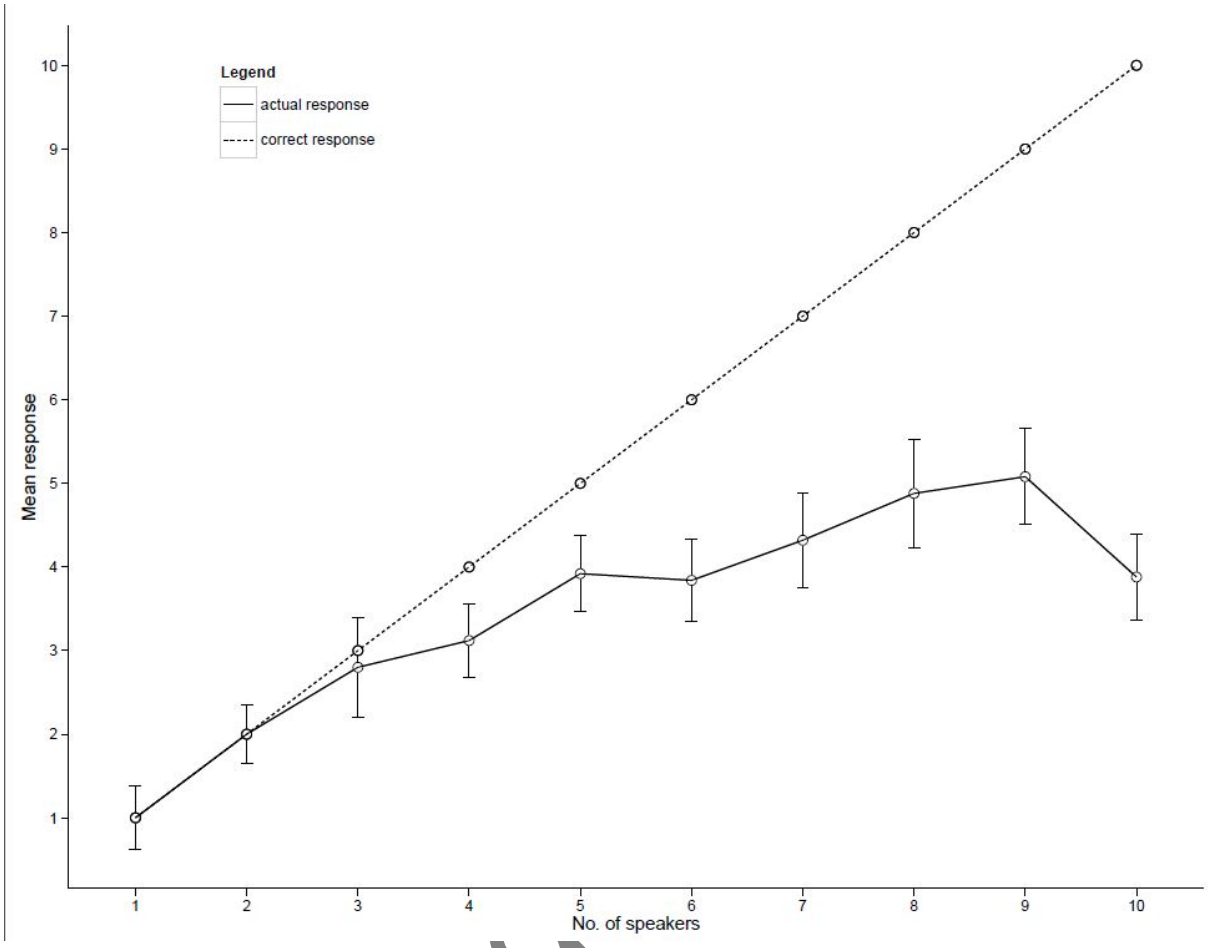
Note: The mean responses in Experiment 3 for speakers 1-3 were indistinguishable from the correct answer (i.e., they were correct), but for speakers 4-10 the mean responses were significantly different from the correct answer (i.e., they were wrong).

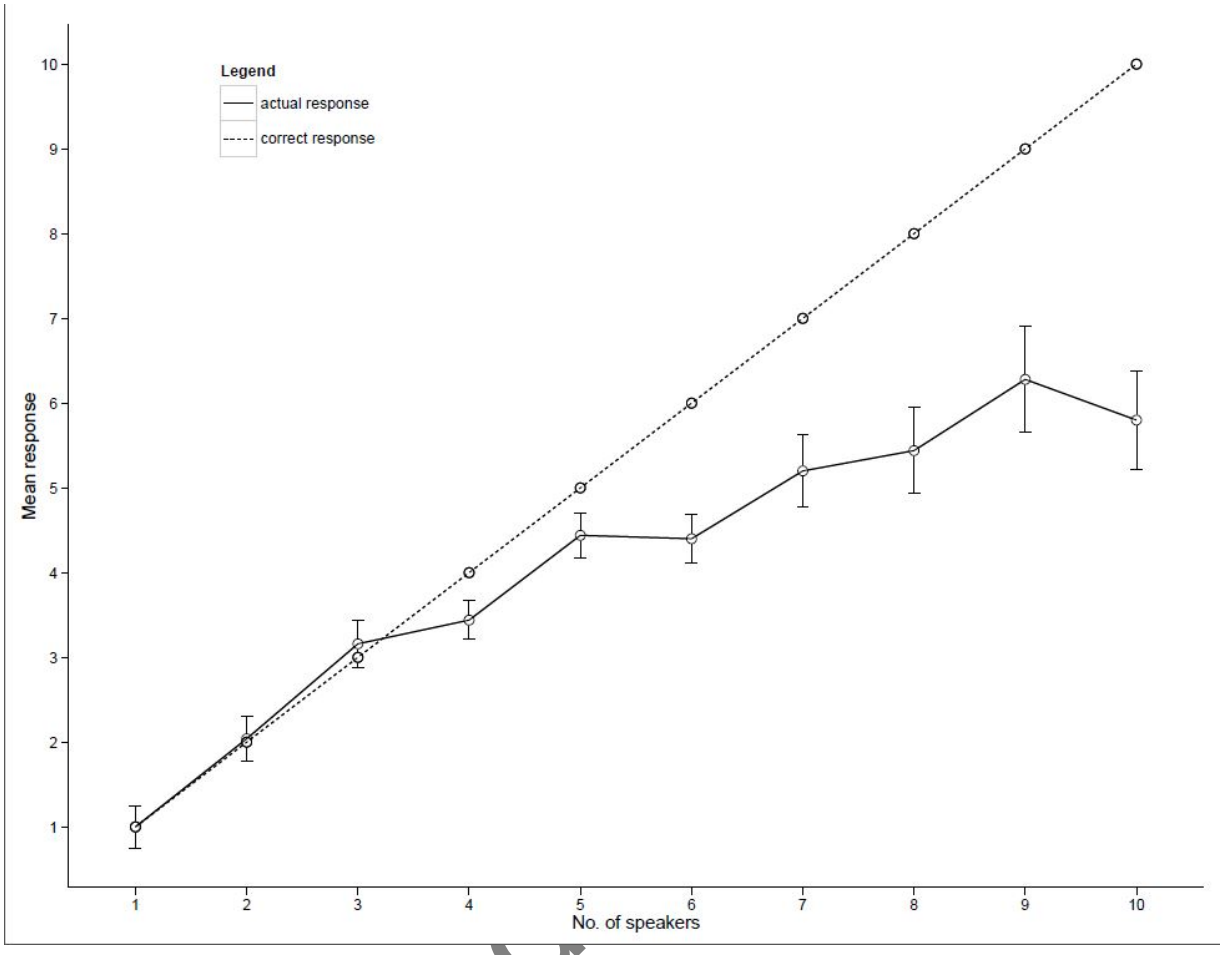
Table 4. Mean proportions of "many speakers" responses compared to the total number of possible responses for each speaker condition.

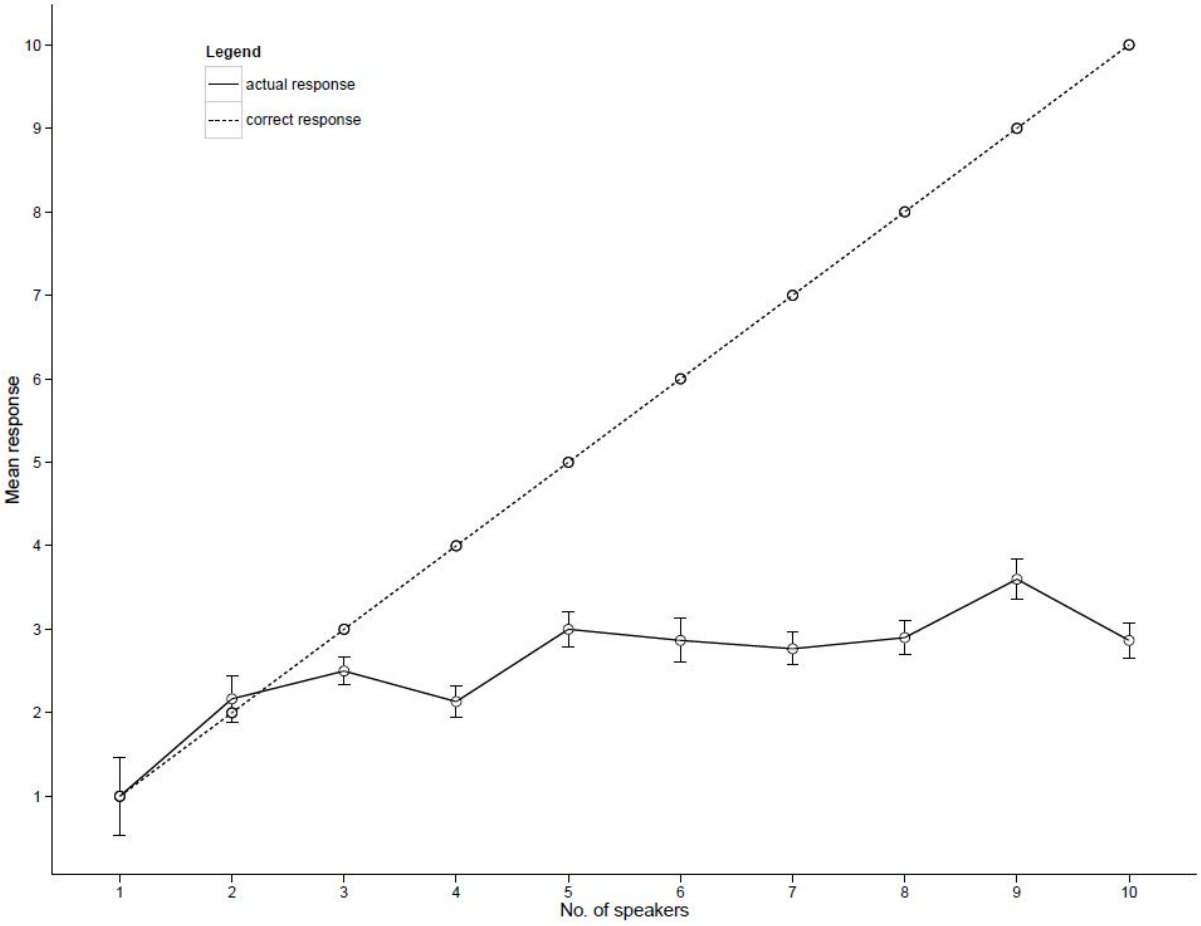
	Speakers									
	1	2	3	4	5	6	7	8	9	10
Mean proportion	0.01	0.06	0.27	0.45	0.71	0.94	0.94	0.98	0.95	0.97
SD	0.04	0.20	0.32	0.34	0.26	0.12	0.13	0.06	0.14	0.09
t-value	1.00	1.66	4.55	7.22	15.13	43.20	39.53	87.96	38.46	57.41
p-value	.33	.11	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001

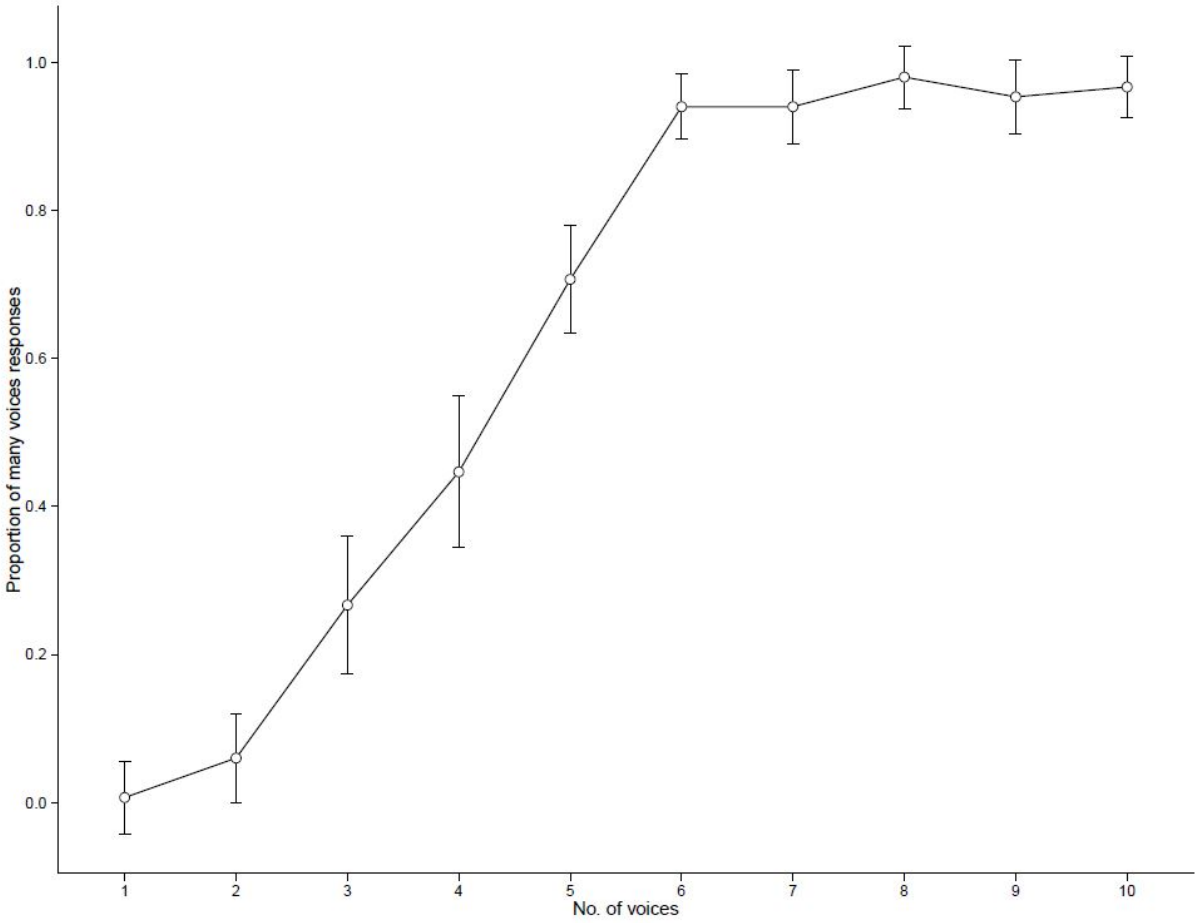
Note: The mean proportions in Experiment 4 for speakers 1-2 were indistinguishable from 0 (i.e., participants rated the stimuli as having "few speakers"), but for speakers 3-10 the mean proportions were significantly different from 0 (i.e., participants rated the stimuli as having "many speakers").

Footnote









[1] Given the exploratory nature of the experiments we report here we have elected to use two-tailed instead of one-tailed t-tests because of their more conservative nature.