Original Article

The Listener Effect in Multitalker Speech Segregation and Talker Identification

Trends in Hearing
Volume 25: 1–11

DOI: 10.1177/23312165211051886
journals.sagepub.com/home/tia
SAGE

Robert A. Lutfi, Briana Rodriguez and Jungmee Lee

Abstract

Over six decades ago, Cherry (1953) drew attention to what he called the "cocktail-party problem"; the challenge of segregating the speech of one talker from others speaking at the same time. The problem has been actively researched ever since but for all this time one observation has eluded explanation. It is the wide variation in performance of individual listeners. That variation was replicated here for four major experimental factors known to impact performance: differences in task (talker segregation vs. identification), differences in the voice features of talkers (pitch vs. location), differences in the voice similarity and uncertainty of talkers (informational masking), and the presence or absence of linguistic cues. The effect of these factors on the segregation of naturally spoken sentences and synthesized vowels was largely eliminated in psychometric functions relating the performance of individual listeners to that of an ideal observer, d'_{ideal}. The effect of listeners remained as differences in the slopes of the functions (fixed effect) with little within-listener variability in the estimates of slope (random effect). The results make a case for considering the listener a factor in multitalker segregation and identification equal in status to any major experimental variable.

Keywords

cocktail-party problem, listener effect

Received 2 June 2021; Revised 17 September 2021; accepted 20 September 2021

Introduction

Individuals often have different ideas than researchers on how they should behave in psychophysical experiments. This is certainly true for research on the cocktail-party problem where wide variation in individual listener behavior frequently complicates the interpretation of results. In a typical experiment, the stimuli are sequences of vowels or words spoken by two or more talkers; one of the talkers is identified as the target and the other talker(s) are identified as nontarget distractors. In one version of the experiment, the listener is asked to judge whether the speech of the target talker is heard separately from that of the nontarget talker(s) (e.g., Lutfi et al., 2020); in another, the listener must ignore the nontarget talker(s) and report on some property of the target: who they are, where they are or what they said (see Kidd & Colburn, 2017 for a review). The participants in these experiments are most often young, healthy, clinically normal-hearing adults who have been well practiced in the task before data collection, yet their performance within conditions often differs quite substantially. In tasks requiring the identification of words spoken by the target performance within conditions has ranged over 40 percentage points (Getzmann et al., 2014; Johnson et al., 1986; Kidd et al., 2007; Oberfeld & Klöckner-Nowotny, 2016; Ruggles & Shinn-Cunningham, 2011; Ruggles et al., 2011). When thresholds have been obtained for constant word identification performance, the differences have been as much as 20 dB (Füllgrabe et al., 2015; Hawley et al., 2004; Kidd et al., 2007; Kubiak et al., 2020; Swaminathan et al., 2015). Even for the relatively simple tasks involving the segregation or identification of talkers, the performance of individual listeners has been observed to range from near chance to perfect within conditions (Best et al., 2018; Lutfi et al., 2018).

Deciding how to deal with the wide variation among listeners has long been a challenge for researchers. The reasons for it are not well understood. Many factors are likely responsible, and they may be different for different listeners. Those

Auditory Behavioral Research Lab, Department of Communication Sciences and Disorders, University of South Florida, Tampa, Florida

Corresponding author:
Robert A. Lutfi, Auditory Behavioral Research Lab, Department of Communication Sciences and Disorders, University of South Florida, Tampa, Florida, 33620.
Email: rlutfi@usf.edu

suggested in the literature include variation in the capacity of working memory (Conway et al., 2001; McLaughlin et al., 2018; Tamati et al., 2013), differences in the ability to selectively attend to targets (Dai & Shinn-Cunningham, 2016; Oberfeld & Klöckner-Nowotny, 2016; Ruggles & Shinn-Cunningham, 2011; Shinn-Cunningham, 2017), lapses in attention (Bidelman & Yoo, 2020; Brungart & Simpson, 2007), variations in hearing sensitivity (Dewey & Dhar, 2017; Lee & Long, 2012; Plack et al., 2014) and cochlear pathology missed by conventional audiometry (Bharadwaj et al., 2015; Kujawa & Liberman, 2009).

Without a clear account or practical way to control for the individual differences, researchers have been forced to either accept them as an inevitable source of unexplained variance in their data or to analyze for them as a factor affecting the results. The former approach has been, by far, the more popular. The focus has been on the main effects of experimental factors with little information provided regarding the individual differences in performance beyond the error bars associated with group means. We lack, therefore, answers to basic questions regarding the individual differences. What is the effect of replications for individuals? Do they have a *fixed effect* on averaged performance, deviating from the average in much the same way each time, or is it a *random effect*, different in each case reflecting what is commonly considered to be measurement error associated with the listener having "good and bad" days? The latter is assumed in the statistical analysis of the main effects of experimental factors, but the exact behavior is virtually never reported in studies. We also do not know to what extent the listener effect may interact with that of major experimental factors. Does it depend critically on these factors or does it reflect more general listener traits, for example, lapses in attention or limits in trace memory that likely have a similar influence across different experimental conditions and studies? This question too is fundamental to understanding how we should treat the individual variation, but identifiers that would allow for the tracking of individual behavior across conditions are rarely reported. Finally, what is the size of the listener effect relative to that of experimental factors? Different experimental factors may be shown to have a statistically significant effect when averaged over the performance of individual listeners. However, how does that compare to the effect of individual listeners averaged over different experimental factors? The answer could have implications for the advisability of treating the listener effect as a source of error variance in statistical analyses.

The present study was undertaken to provide some data relevant to these questions. The goal was to evaluate the behavior of each listener's specific effect on performance relative to the main effect of major experimental factors known to impact performance. We consider the main effects of differences in task, differences in type of segregation cue, differences in talker voice similarity and uncertainty, and the availability of linguistic cues. These factors involve fundamentally different experimental manipulations and units of measure, so to permit comparisons we have adopted an approach specifically developed for this purpose: the analysis of ideal observers from signal detection theory (Green & Swets, 1966). Our results show that, once normalized for the performance of an ideal observer, the one factor having greatest impact on performance in these experiments is the listener.

General Linear Model

To understand our approach, consider the following example. We wish to evaluate the separate effect of the listener on performance for which the main experimental factor is the type of segregation cue. We have performance for the same group of listeners in two different conditions: one in which the target and nontarget talker voices differ in fundamental frequency, the other in which they originate from different locations. Let $k \in \{1,2\}$ identify these two cues and let X_k denote the corresponding effect of each cue (Note that k more generally could refer to the different conditions associated with any main experimental factor). We take performance for the i-th listener on the j-th replication (an average of a block of trials) for the k-th cue to be given by

$$d'_{ijk} = X_k + L_i + L_{i|k} + e_j \quad (1)$$

where L_i is the fixed effect of the listener across cues, $L_{i|k}$ is the effect of the listener that depends on the cue and e_j is the random effect of the listener. We assume for the population of all listeners that L_i, $L_{i|k}$, and e_j are normally distributed random variables with zero mean. This is the common justification for averaging performance across listeners and replications to estimate the main effects of experimental variables. The estimated main effect of cues by this assumption is

$$d'_{..k} = \text{est}(X_k). \quad (2)$$

Here, we adopt a common notation where a period is put in place of the subscript for the variable being averaged, in this case subjects (i) and replications (j).

The quantity given by Equation 2 pertains exclusively to the effect of a particular experimental variable on performance (type of cue in this case) and with its error of estimate is most often the only data reported in studies. In rare instances where data from individual listeners are reported, those data are given as the average of the individual's performance across replications, within conditions, $d'_{i.k}$. This measure provides some information about the effect of individual listeners but conflates that effect with the effect of the experimental variable,

$$d'_{i.k} = \text{est}(X_k + L_i + L_{i|k}). \quad (3)$$

In the present analysis, we focus on three different measures that pertain exclusively to the effects of the listener on performance. The first is the effect of the listener that carries over across conditions, the fixed effect of the listener, L_i. In our example, the fixed effect is estimated by averaging the

performance for each listener across the two conditions of the experimental factor and all replications,

$$d'_{i..} = \text{est}(L_i). \quad (4)$$

The second measure is the effect of the listener that depends on the cue, what we call the conditional effect of the listener, $L_{i|k}$. This effect pertains to individual differences in the relative effectiveness of the two cues. It is one of three factors that influence the average performance of the listener for the cue (right side of Equation 3) and so can be isolated by subtracting out the other two. The other two are the effect of the cue (Equation 2) and the fixed effect of the listener (Equation 4),

$$d'_{i|k} = d'_{i.k} - d'_{..k} - d'_{i..} = \text{est}L_{i|k} \quad (5)$$

The third measure is the random effect of the listener and is estimated from the standard deviation of replications about their own mean or, as in the present study, about a regression curve representing a complete psychometric function,

$$\text{S.D.}(d'_{i.k}) = [\text{AVG}\, j(d'^{2}_{i.k} - d'^{2}_{ijk})]^{1/2} = \text{est}(e). \quad (6)$$

These three measures will tell us if there is, in fact, a significant listener effect on performance, whether it is primarily a fixed or random effect, whether it depends for some listeners on the experimental condition and how big it is relative to the main effect of the type of cue as the experimental factor.

There is just one more step. The two cues in our example entail very different manipulations of the stimulus and different units of measure (frequency vs. location). Simple comparisons of performance for the two cues, and more generally other experimental factors involving different manipulations, are therefore not particularly meaningful. Instead, we need to compare using a performance standard common to both cues. In signal detection theory, the gold standard for evaluating perceptual capability across different stimulus conditions and psychophysical tasks is the performance of an ideal observer, d'_{ideal} (see Green & Swets, 1966, 2020). The ideal observer is a *noise-free* observer that optimizes decisions within the constraints imposed by the task and the known statistical properties of signals. In the present example, and for the three other major experimental factors investigated, those statistical properties relate to the differences in the voice fundamental frequency and location of talkers speaking in any trial. We report complete psychometric functions relating listener performance to d'_{ideal} for each experimental factor and use these functions derive our three measures of the listener effect.

General Methods

Stimuli and Tasks

The stimuli and tasks were similar to those used in previous publications by the authors (see Lutfi et al., 2018, 2020). On each trial, two interleaved sequences of vowels, A and B, were played in the pattern ABA_ ABA_ABA_ABA, where the underscore character represents a 100-ms silent interval. The pattern follows that of the ABA tone sequences used in popular stream segregation experiments (cf. Bregman, 1990). The vowels were selected at random on each trial and for each ABA triplet with replacement from a set of 10 exemplars. The only constraint on the random selection was that the first and last vowel within each ABA triplet be acoustically identical. The 10 exemplars, identified by their international phonetic alphabet names, were i, I, ɛ, æ, ʌ, ɑ, ɔ, ʊ, u, and ɜ˞ (nominal frequencies of vowel formants as given by Peterson & Barney, 1952). The vowels were synthesized using the MATLAB program Vowel_Synthesis_GUI25 available on the MATLAB exchange. Each vowel was 100 ms in duration and was gated on and off with 5-ms, cosine-squared ramps. The stimuli were played at a 44,100-Hertz (Hz) sampling rate with 16-bit resolution using an RME Fireface UCX audio interface. They were delivered to listeners seated in a double-wall, sound-attenuation chamber listening over Beyerdynamic DT990 headphones.

In all experiments, the two cues listeners could use to distinguish talkers were differences in voice fundamental frequency Δ_{F0} and location given as a difference in azimuth angle Δ_θ. Specific values for talkers are given in each experiment. The azimuth locations were simulated over headphones using the head-related transfer functions of the Knowles electronics manikin for acoustic research (KEMAR). A small, random perturbation was added independently to the nominal values of F0 and θ assigned to each talker for each vowel on each trial. The perturbation was normally distributed with zero mean and standard deviation σ_{F0} and σ_θ. Specific values, again given in each experiment, were chosen to be within the normal range of human speech. The perturbation served somewhat to simulate natural variation that occurs in these cues, but they also established the different values of d'_{ideal} for the psychometric functions in each experiment.

For all conditions, the ratio of Δ and σ for the two stimulus cues were equated in value, $\Delta_{F0}/\sigma_{F0} = \Delta_\theta/\sigma_\theta$. We will henceforth refer to this as the delta/sigma ratio and denote it as Δ/σ. We compute performance for an ideal observer who, like the listener, does not have complete knowledge of the statistical distributions of F0 and θ. Such an observer takes the difference between sequence A and B in the values of F0 and θ (recall that the two presentations of A within the triplet are identical). There are four such values corresponding to the four ABA triplets for each cue ($n = 8$ observations altogether) and they are statistically independent. Hence, since the standard error of independent observations decreases as the square root of n, the performance of the ideal observer is $d'_{ideal} = 8^{1/2}\Delta/\sigma$. For simplicity and without effect on interpretation, listener d' performance will

be plotted against d'_{ideal} for a single cue and vowel triplet, $d'_{ideal}(1) = \Delta/\sigma$.

There were two tasks. In the segregation task, the B sequence of vowels was always spoken by the same talker; we shall call him Bob. The A sequence of vowels was spoken by Bob or by a different talker, Barb, with equal probability on each trial. The listener's task was to judge on each trial whether the A and B sequences were spoken by Bob alone or by Bob and Barb. The second task was talker identification; the B sequence of vowels was spoken by Bob as before; however, the A sequence was now spoken by either Barb or Ben. The listener's task was to judge on each trial who was speaking, Barb or Ben.

Listeners and Procedures

The listeners were sevens male and 22 female students at the University of South Florida—Tampa, ages 18–27 years, and were paid in cash or gift cards for their participation. Not all listeners participated in all experiments. All, however, had normal hearing as determined by standard audiometric evaluation, which included pure-tone audiometry and tympanometry. Prior to data collection, the listeners received three blocks of 30 trials in an easy condition ($\Delta/\sigma >> 3$) to provide some basic training and to ensure that they understood the task. They then started experimental trials with the easiest condition in each experiment, proceeding with successively harder conditions as the experiment progressed. Our experience has been that this ordering helps to reduce variability in performance for the most difficult conditions. A complete psychometric function relating listener d' to $d'_{ideal}(1) = \Delta/\sigma$ was obtained for one condition before proceeding to the next. The data were collected in eight blocks of 50 trials per block, each within a 1-hr session. Each trial block corresponded to the datum for a single condition k and replication j, replications given by the different points on the psychometric function. Listeners were allowed frequent breaks between trial blocks. Informed consent was obtained from all listeners and all procedures were followed in accordance with University of South Florida internal review board (IRB) approval.

Results

Experiment 1: Talker Identification versus Segregation

The first experiment investigated the effect of task, talker segregation versus identification, on individual differences in listener performance. For both tasks, the nominal values of F0 and θ for Bob were 130 Hz and 0°. For the segregation task, Barb differed in F0 from Bob by a positive $\Delta = 10$, 15, 20, or 25 Hz and by an equivalent number of degrees. The corresponding perturbation values σ of F0 and θ were fixed, respectively, at 10 Hz and 10°, again with F0 and θ being sampled independently at random for each presentation. Note here that the F0 and θ cues are equated in $d'_{ideal}(1) = \Delta/\sigma$; they just so happen to also have the same numerical values. For the talker identification task Barb differed from Bob by $\Delta = 5$, 10, 15, or 20 Hz and by an equivalent number of degrees. Ben differed from Bob by $\Delta = -5$, -10, -15 or -20 Hz and by an equivalent number of degrees; $d'_{ideal}(1)$ was computed based on the difference between Barb and Ben. To maintain performance over a range comparable to the segregation task, the perturbation values σ were increased to 15 Hz and 15°.

Figure 1 shows in different panels the individual psychometric functions, d' versus $d'_{ideal}(1) = \Delta/\sigma$, for 11 listeners. Each datum represents the average of 400 trials with replications given by the different points within a condition. Filled symbols are the data for the talker segregation task, unfilled symbols are the data for the talker identification task. The continuous lines drawn through these data are linear least-squares fits. The intercepts were a free parameter in the fits, but as expected are all close to 0 (chance performance) at $\Delta/\sigma = 0$. The dashed line gives $d'_{ideal}(1) = \Delta/\sigma$, the performance of the ideal observer for a single A–B difference in F0 or θ. Note that since the intercepts of the functions converge to 0, we can talk about the effect of experimental and listener factors exclusively in terms of their impact on the slopes of these functions. Likewise, we can take each datum as an independent estimate (replication) of slope.

The first point to note regarding these functions is that the slopes for each listener are near identical for the two tasks. That is, the change in task has little or no effect on observed performance when expressed relative to that of the ideal observer. This type of behavior has been reported previously by Lutfi et al. (2013) when comparing conditions of target-masker similarity and masker uncertainty resulting in informational masking (IM). These authors report identical psychometric functions in both conditions when d' is plotted against Simpson–Fitter's d_a, a statistic equivalent to d'_{ideal} for the conditions of the present study. Lutfi et al. (2013) suggest that the results are an example of a more general behavior in IM tasks wherein performance is related to the information divergence of the target and masker. The high degree of listener uncertainty regarding signals that exist in IM tasks as in the present conditions is a factor responsible for this behavior. We will expand on this point in the discussion.

Consider next the effect of listeners. The random effect of listeners, given by the deviations of the data about the regression curves, S.D.$(d'_{i.k})$, is quite small and differs hardly at all across listeners. So too, the conditional effect of listeners d'_{ilk} is quite small, again the slopes of the functions for the two tasks being near identical for each listener. This leaves the fixed effect of listeners $d'_{i..}$ given by the difference in slopes across listeners as the largest, one might say the only effect in the data, and it is huge. Across listeners the

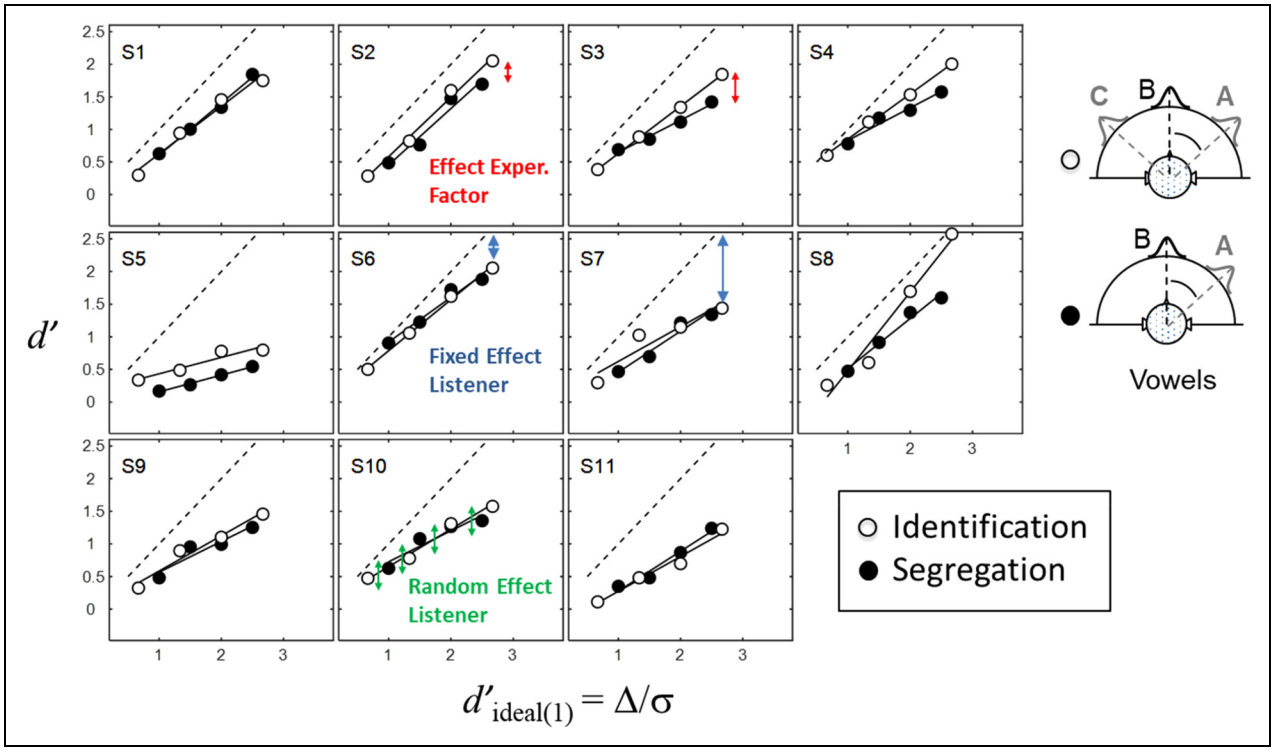

Figure 1. Listener d' performance is plotted against Δ/σ for 11 listeners (panels) for both the talker segregation (filled symbols) and identification (unfilled symbols) tasks of Experiment 1. The continuous lines drawn through the data are linear least-squares fits. The dashed line is the performance of an ideal observer for a single A–B difference in F0 or θ (see text for explanation).

slopes correspond to performance levels ranging anywhere from near $d'_{ideal}(1)$ for listener S8 to near chance for listener S5.

Experiment 2: Voice Fundamental Frequency versus Spatial Location

Experiment 2 next investigated individual differences in performance for two different talker segregation cues. The experiment was identical to that of Experiment 1, segregation task, except that under different conditions each segregation cue, F0 and θ, was presented in isolation. These two cues were chosen as they are the ones that have had the largest effects in the literature and have been given most attention (see Bronkhorst, 2000, 2015 and Kidd & Colburn, 2017 for reviews). To obtain the psychometric functions, the nominal differences Δ in F0 and θ for the two talkers were incremented in 5 unit steps from 5 to 35 in both Hz and degrees. The perturbation values σ for the two cues were fixed respectively at 10 Hz and 10°. A total of 12 listeners participated in the experiment, seven of whom, 1, 2, and 4–8, had previously participated in Experiment 1.

Figure 2 gives the psychometric functions for each listener (panels) for the fundamental frequency (F0) cue alone (unfilled symbols) and for the azimuth (θ) cue alone (filled symbols). As before, the continuous lines are linear least-squares fits to the data and the dashed line is $d'_{ideal}(1)$. The pattern of results is quite similar to that of Exp. 1. The random effect of listeners on the slopes of the functions is relatively small and uniform across listeners (again each datum in the figure yielding an estimate of slope). The fixed effect of listeners on the slopes is large, and corresponding performance ranges from near chance for listener S5 to that approaching $d'_{ideal}(1)$ for listener S14. Listener 5 who showed the largest fixed effect for task in Experiment1 also shows the largest fixed effect here for talker cues.

An important difference in results from Experiment 1 is the clear evidence for an interaction between the effect of listeners and type of cue, that is, a conditional effect of listeners. Whereas most listeners make equally effective use of the two cues, four listeners (bottom panels of Fig. 2) make considerably more effective use of the location cue. Were we to average over all listeners, these four would make it appear that there was a small main effect of cue. However, this is only because of the size of the effect for these listeners is quite large. These data are an example of how the analysis of main effects of factors can disguise what is actually an effect of a few individual listeners.

That the F0 cue would prove most challenging for some listeners is perhaps not surprising. Estimates are that four percent of the population experience dysmelodia (tone

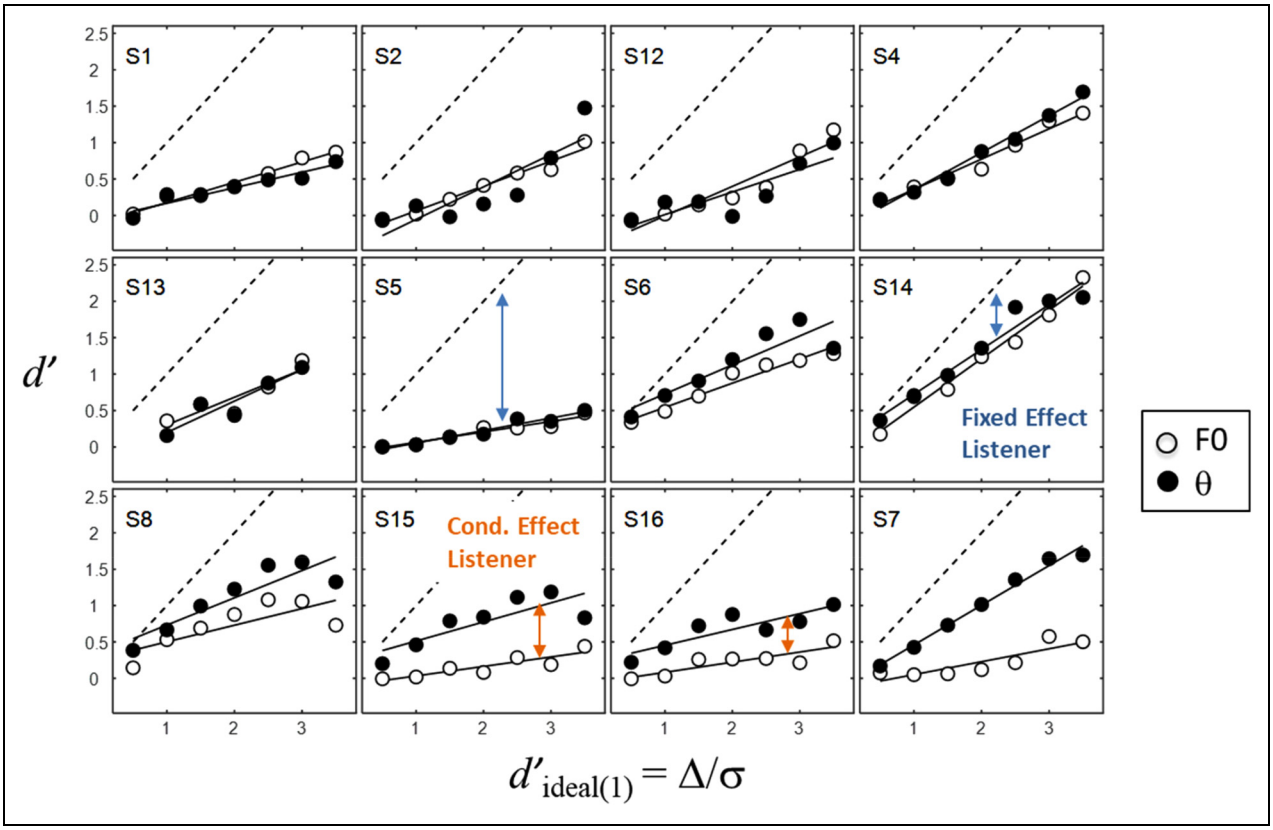

Figure 2. Same as Figure 1 except listener d' performance is plotted against Δ/σ for 12 listeners (panels) for the voice pitch cue presented alone (unfilled symbols) and the location cue presented alone (filled symbols) in Experiment 2.

deafness), an inability to order the pitch of sounds (Kalmus & Fry, 1980). Frequency discrimination of tones is also known to be highly variable across listeners (Wier et al., 1977). Whatever the reason for this difference, there remains a clear fixed effect of listeners on performance and as before the effect is considerably larger than the main effect of the experimental factor (difference in cues) when considered relative to the performance of an ideal observer.

Experiment 3: Voice Similarity versus Uncertainty

The literature on the cocktail-party effect attributes failures to segregate talkers to a combination of two fundamentally different types of masking (Durlach, 2006). Energetic masking (EM) is identified with processes occurring at a peripheral level of the auditory system, in the cochlea and/or auditory nerve, and arises only when there is some degree of overlap or close proximity of target and nontargets in both frequency and time. Informational masking is identified with processes occurring at a more central level of the auditory system and arises when target and nontargets are made perceptually similar to one another or when uncertainty is introduced regarding their acoustic properties. Notably IM, unlike EM, varies dramatically across individual listeners and can occur even when target and nontargets are widely separated in frequency and time (see Kidd et al., 2008 for review). It is for these reasons that IM is thought to be largely responsible for the individual differences in performance in many cocktail-party listening studies (Arbogast et al., 2002; Brungart, 2001; Kidd & Colburn, 2017; Kidd et al., 2016). Experiment 3 investigated the effect of IM on individual differences in the present segregation task. Target and nontarget vowels were made perceptually more or less similar by varying the difference Δ in the nominal values of F0 and θ of the two talkers. The vowels were made more or less uncertain by varying the magnitude of the random perturbation σ imposed on F0 and θ. In the first case, σ was fixed at 10 and Δ took on values of 10, 15, 20, and 25, both in Hz and degrees. In the second case, Δ was fixed at 15 and σ took on values of 5, 10, 15, and 20, again both in Hz and degrees. All other conditions were identical to those of Experiment 2.

The results for 10 new listeners and listeners 9 and 10 from Experiment 1 are shown in Figure 3, where filled symbols give the data for the voice similarity condition and unfilled symbols give the data for the voice uncertainty condition. The general pattern of results is the same as for Experiment 2, a very

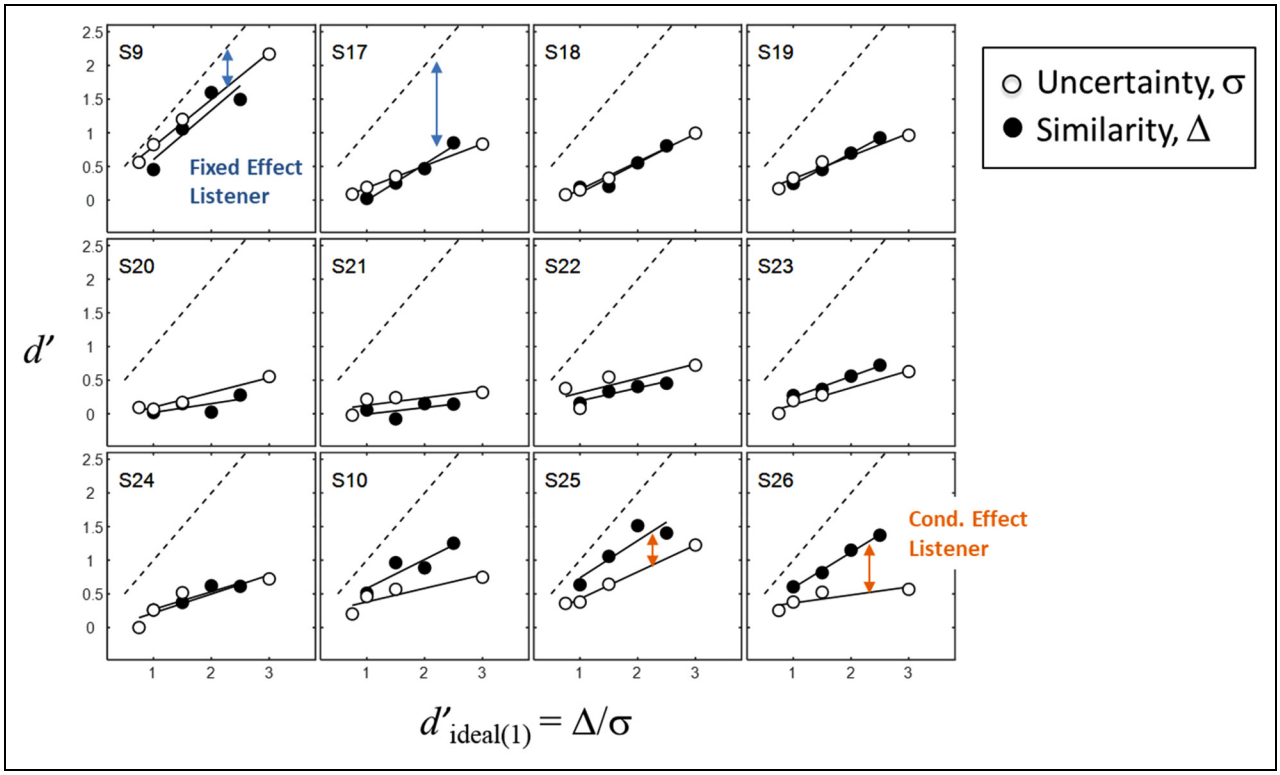

Figure 3. Same as Figure 1 except listener d' performance is plotted against Δ/σ for 12 listeners (panels) for the voice similarity (filled symbols) and voice uncertainty (unfilled symbols) conditions of Experiment. 3.

small random effect of listeners, a very large fixed effect of listeners, and a small main effect of the experimental factor (IM) that is due entirely to a conditional effect of three listeners, S10, S25, and S26, who are more adversely affected by voice uncertainty than voice similarity.

Experiment 4: Role of Linguistic Cues

Experiment 4 was undertaken to evaluate the generality of the conclusions drawn from Experiments 1–3 and to consider the effect of one last factor known to impact performance in multitalker segregation, that of the linguistic cues in speech. Linguistic information is not required for the large individual differences in performance we and others have observed, but there are reasons to consider its possible role in mitigating these differences. Linguistic cues are, of course, commonly available to listeners in natural speech and they have been shown when available to substantially improve performance in multitalker segregation (Kidd and Colburn, 2017; Kidd et al., 2016). They have not, however, been widely investigated as a factor affecting individual differences in performance, except for differences related to one's native language (Brouwer et al., 2012; Calandruccio et al., 2014, 2017). The question then for Experiment 4 was whether linguistic information might serve to mitigate individual differences by allowing individuals who have difficulty perceptually segregating talkers to compensate by using the syntactic structure and semantics of natural speech to distinguish speech streams.

The experiment was identical in all respects to Experiment 1 for the talker identification task. The vowel sequences, however, were now replaced by recordings of naturally spoken English sentences spoken simultaneously by the two talkers. The sentences were selected at random on each trial from the 10 exemplars of the Texas Instrument/Massachusetts Institute of Technology (TIMIT) training set FSLSO DR3. The talkers, as before, were distinguished by differences in their nominal voice fundamental frequency (F0) and azimuthal locations (θ). The differences were produced by taking the original 10 sentences, which were spoken by a single talker, and shifting their F0 and θ while maintaining the original duration of the sentences; the differences in duration were typically <1 s (Note that because of differences in duration and starting time, the sentences were not temporally aligned, target and nontarget talkers had an equal probability of beginning first and/or ending last, as happens at cocktail parties). The θ shift was achieved using the KEMAR transfer functions, as before. The F0 shift was achieved using the overlap and add method (Hejna & Musicus, 1991) implemented by the

function 'solaf' available on the MATLAB exchange. For Bob, the shift in both Hz and degrees was 0, for Ben it was −5, −10, −15, or −20, and for Barb it was + 5, + 10, + 15 or + 20. For all talkers, the random perturbation in the two cues was fixed at 10 Hz and 10 degrees. In the comparison condition, the spoken sentences were simply played in reverse (cf. Kidd et al., 2016). Reversing the sentence eliminates all linguistic information but maintains the key acoustic information for talker identification given by the talker differences in F0 and θ; it thus has as no effect on the computation of d'_{ideal}.

The procedure was as follows: On each trial, two of the 10 sentences were selected at random. One was assigned to the nontarget talker Bob and the other, with equal probability, was assigned to either of the two target talkers, Ben or Barb, whoever was to be speaking on that trial. The sentences were selected with replacement, so it was possible on some trials that they were the same for both talkers. After processing to impose the appropriate values of F0, θ and perturbation for each talker, the sentences were presented concurrently. The listener's task was to report whether the target talker was Barb or Ben.

The data from 11 listeners are given in Figure 4. Native speakers of English were selected to run in this experiment to ensure normal processing of linguistic cues. The overall performance levels are not directly comparable to the previous experiments as the sentences on each trial represent only one observation of the difference in F0 or θ (i.e., only one value of each was selected for each sentence on each trial). Nonetheless, the results can be summarized as follows: no main effect of the linguistic factor, no conditional effect of listeners, a very small random effect of listeners, and a fixed effect of listeners yielding near-chance levels of performance for listener S23 to performance close to $d'_{\text{ideal}}(1)$ for listener S29. The results are generally consistent with those of Experiments 1–3 in showing only a fixed effect of listeners when performance is expressed relative to that of an ideal observer.

Discussion

The authors are aware of three studies that reveal similar effects of listeners in multitalker segregation to those presented here, although such effects were not the specific focus of those studies. Kidd et al. (2016) measured the effect on speech reception thresholds in multitalker masking of three experimental factors: differences in the gender of talkers, differences in the spatial separation of talkers, and time-reversal of the interfering speech. For the same group of six listeners participating in each condition, the main effect of the factors on thresholds was never <9 dB, but then so too was the range of individual thresholds for each factor. The individual differences in thresholds were also highly correlated across conditions, suggesting a predominant fixed effect of listeners. Arbogast et al. (2002) measured the effect of spatial separation of target and masker on word identification for comb-filtered speech and noise maskers. They ran only four listeners but obtained complete psychometric functions for each listener relating performance to the level of the target sentence. Their data, like ours, show a fixed effect of listeners on the slope of the psychometric functions for the speech masker. Also like our data, theirs show a negligible random effect of the listener on the slope. Finally, Lutfi et al. (2013) report conditions paralleling those of the present study but focusing on word recognition performance. They compared the effects of F0 similarity and F0 uncertainty, varying Δ and σ for F0 and equating d'_{ideal} across conditions in the same manner as the present study. For the same group of nine listeners, individual performance ranged over 40 percentage points within each condition and was highly correlated across conditions. As in the present study, no significant difference in the effect of F0 similarity and uncertainty was found for the conditions equated in d'_{ideal}.

The last of these results is noteworthy as it demonstrates, consistent with the present results, that relative to the performance of an ideal observer, the predominant effect on performance across different experimental factors is that of the listener. This finding is new, studies have not generally controlled for or reported on the statistics of segregation cues in such a way that would allow for listener performance to be expressed relative to that of an ideal observer, nor have they attempted a different gold standard for performance comparisons required to evaluate the listener effect across different experimental factors. The outcome has implications broadly for understanding the wide individual variation in performance observed in multitalker segregation studies. In all of these studies, as in real-world, cocktail-party listening, the listener has some degree of uncertainty regarding the location and/or fundamental frequency of the individuals speaking at any given moment in time (or on any trial). This uncertainty is caused (either experimentally or in natural recordings) by variation in these parameters that the listeners cannot anticipate from trial to trial. Such variation is expected to be a factor accounting for the large individual differences in performance observed (Kidd et al., 2008), but again the statistics of the variation are rarely reported in a way that would allow a direct link to be made. Instead, accounts have been more often sought in terms of individual differences in audibility and/or the ability to "resolve" changes in the spectro/temporal and spatial properties of signals (e.g., Bernstein et al., 2013; Humes & Christopherson, 1991; and Humes et al., 2013). The present results suggest that sensitivity to the statistical properties of signals may be far more important than the acoustics, or for that matter linguistic cues, task, or degree of perceptual similarity and uncertainty of talker voices. Lutfi et al. (2013) have argued a similar point for their experiments on informational masking, IM. They report for quite different conditions involving different stimuli and tasks (multitone pattern discrimination, sound-source identification, sound localization, and multitalker word recognition) a strong dependence of

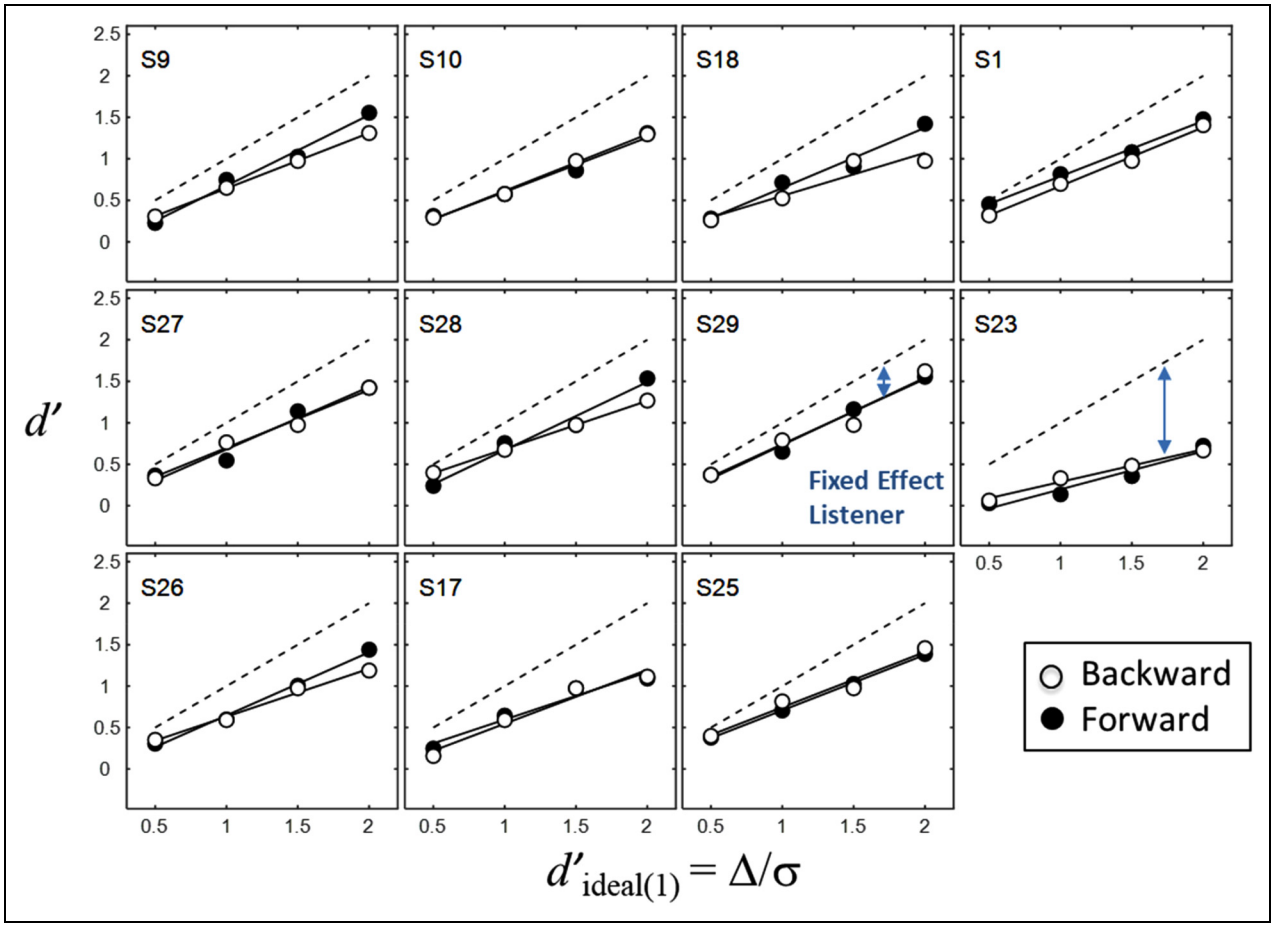

Figure 4. Same as Figure 1 except listener *d* performance is plotted against Δ/σ for 11 native speakers of English (panels) for the linguistic cues absent (unfilled symbols) and present (filled symbols) conditions of Experiment 4.

performance on Simpson-Fitter's d_a for signals, a statistic closely related to d'_{ideal} in the present experiments. They take the result to be an instance of a general principle of perception that relies heavily on differences in the statistical properties of signals for segregation, more so than their acoustic properties. This account could easily generalized to the present results given the large role IM is expected to play in multitalker segregation.

The data also have implications for model development. The common outcome across conditions suggests that a computational model with relatively few free parameters might well account for individual behavior in these and other experiments. Notable in this regard is that the inverse of the slope of the psychometric function, which varies with the fixed effect of listeners, is a measure of internal noise; the greater the slope, the less the internal noise (Green & Swets, 1966). Internal noise is a construct representing spontaneous activity in the nervous system that results in information loss in the processing of signals. Recent work has implicated internal noise models over feature weighting models (e.g., selective attention) as a class to account for individual differences in multitalker speech segregation (Lutfi et al., 2020). Candidates proposed for the source of the internal noise are lapses in attention and limited trace memory, both of which are likely to have similar effects across different experimental conditions. Models that focus on certain cochlear processes, however, can also be expected to have similar effects across studies. The neural undersampling model of Lopez-Poveda (2014), for example, attributes poor performance in multitalker speech segregation to deafferentation in the cochlea resulting in stochastic undersampling of signals. This model is attractive for two reasons. It has a physiological underpinning based on a known cochlear pathology already implicated as a factor responsible for individual differences (Liberman et al., 2016) and it has clearly identifiable parameters with clearly identifiable effects on performance that can be evaluated within the framework of signal detection theory. Current work in our lab is focused on identifying the source(s) of this internal noise.

Summary

The present study was undertaken to provide some basic data regarding individual differences in the behavior of listeners

in multitalker, speech-segregation studies. Individual differences in performance were replicated for four major experimental factors commonly investigated in these studies. In each case, performance was broken down into four components: the effect of the experimental factor, the fixed effect of listeners, random effect of listeners and conditional effect of the listeners. The effect of experimental factors was largely eliminated in psychometric functions relating d' performance to that of an ideal observer, d'_{ideal}. The random effect of listeners was negligible and only a few listeners showed an effect that depended on the condition. The fixed effect of listeners was the dominant effect in the data, in most cases the only effect. The results make a case for considering the listener a factor in multitalker segregation equal in status to any major experimental variable.

Acknowledgments

The authors would like to thank Dr. David M. Green, the Associate Editor, Dr. Virginia Best, and two anonymous reviewers for their helpful comments on an earlier version of this manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Institute on Deafness and Other Communication Disorders (Grant no. NIDCD R01 DC001262-28).

ORCID iD

Robert A. Lutfi https://orcid.org/0000-0003-0847-4342

References

Arbogast T. L., Mason C. R., & Kidd G. J. (2002). The effect of spatial separation on energetic and informational masking of speech. *Journal of the Acoustical Society of America*, *112*(5), 2086–2098. https://doi.org/10.1121/1.1510141

Bernstein J., Mehrai G., Shamma S., Gallun F., Theodoroff S., & Leek M. (2013). Spectrogram modulation sensitivity as a predictor of speech intelligibility for hearing-impaired listeners. *Journal of the American Academy of Audiology*, *24*(4), 293–306. https://doi.org/10.3766/jaaa.24.4.5

Best V., Ahlstrom J. B., Mason C. R., Roverud E., Perrachione T. K., Kidd G. J., & Dubno J. R. (2018). Talker identification: Effects of masking, hearing loss and age. *Journal of the Acoustical Society of America*, *143*(2), 1085–1092. https://doi.org/10.1121/1.5024333

Bharadwaj H. M., Masud S., Mehrai G., Verhulst S., & Shinn-Cunningham B. G. (2015). Individual differences reveal correlates of hidden hearing deficits. *Journal of Neuroscience*, *35*(5), 2161–2172. https://doi.org/10.1523/JNEUROSCI.3915-14.2015

Bidelman G. M., & Yoo J. (2020). Musicians show improved speech segregation in competitive, multi-talker cocktail party scenarios. *Frontiers in Psychology, 11*, 1927. https://doi.org/10.3389/fpsyg.2020.01927

Bregman A. S. (1990). *Auditory scene analysis*. M.I.T. Press.

Bronkhorst A. W. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica United with Acustica*, *86*(1), 117–128.

Bronkhorst A. W. (2015). The cocktail-party problem revisited: Early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics*, *77*(5), 1465–1487. https://doi.org/10.3758/s13414-015-0882-9

Brouwer S., Van Engen K., Calandruccio L., & Bradlow A. R. (2012). Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content. *Journal of the Acoustical Society of America*, *131*(2), 1449–1464. https://doi.org/10.1121/1.3675943

Brungart D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *Journal of the Acoustical Society of America*, *109*(3), 1101–1109. https://doi.org/10.1121/1.1345696

Brungart D. S., & Simpson B. D. (2007). Cocktail party listening in a dynamic multitalker environment. *Perception & Psychophysics*, *69*(1), 79–91. https://doi.org/10.3758/BF03194455

Calandruccio L., Buss E., & Bowdrie K. (2017). Effectiveness of two-talker maskers that differ in talker congruity and perceptual similarity to the target speech. *Trends in Hearing*, *21*. https://doi.org/10.1177/2331216517709385

Calandruccio L., Buss E., & Hall J. W. (2014). Effects of linguistic experience on the ability to benefit from temporal and spectral masker modulation. *Journal of the Acoustical Society of America*, 135(3), 1335–1343. https://doi.org10.1121/1.4864785

Cherry E. C. (1953). Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustical Society of America*, *25*(3), 975–979. https://doi.org/10.1121/1.1907229

Conway A. R., Cowan N., & Bunting M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin & Review*, *8*(2), 331–335. https://doi.org/10.3758/bf03196169

Dai L., & Shinn-Cunningham B. G. (2016). Contributions of sensory coding and attentional control to individual differences in performance in spatial auditory selective attention tasks. *Frontiers in Human Neuroscience*, *10*, 530. https://doi.org/10.3389/fnhum.2016.00530

Dewey J. B., & Dhar S. (2017). A common microstructure in behavioral hearing thresholds and stimulus-frequency otoacoustic emissions. *Journal of the Acoustical Society of America*, *142*(5), 3069–3083. https://doi.org/10.1121/1.5009562

Durlach N. I. (2006). Auditory masking: Need for an improved conceptual structure. *Journal of the Acoustical Society of America*, *120*(4), 1787–1790. https://doi.org/10.1121/1.2335426

Füllgrabe C., Moore B. C., & Stone M. A. (2015). Age-group differences in speech identification despite matched audiometrically normal hearing: Contributions from auditory temporal processing and cognition. *Frontiers in Aging Neuroscience*, *6*(347), 1–25. https://doi.org/10.3389/fnagi.2014.00347

Getzmann S., Lewald J., & Falkenstein M. (2014). Using auditory pre-information to solve the cocktail-party problem: Electrophysiological evidence for age-specific differences. *Frontiers in Neuroscience*, *8*(413), 1–13. https://doi.org/10.3389/fnins.2014.00413

Green D. M. (2020). A homily on signal detection theory. *Journal of the Acoustical Society of America*, *148*(1), 222–225. https://doi.org/10.1121/10.0001525

Green D. M., & Swets J. A. (1966). *Signal detection theory and psychophysics*. Wiley.

Hawley M. L., Litovsky R. Y., & Culling J. R. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *Journal of the Acoustical Society of America*, *115*(2), 833–843. https://doi.org/10.1121/1.1639908

Hejna D., & Musicus B. R. (1991). *The SOLAFS time-scale modification algorithm* (BBN Technical Report).

Humes L., & Christopherson L. (1991). Speech identification difficulties of hearing impaired elderly persons. *Journal of Speech and Hearing Research*, *34*(3), 686–693. https://doi.org/10.1044/jshr.3403.686

Humes L., Kidd G. R., & Lentz J. (2013). Auditory and cognitive factors underlying individual differences in aided speech understanding among older adults. *Frontiers in Systems Neuroscience*, *7*, 55. https://doi.org/10.3389/fnsys.2013.00055

Johnson D. M., Watson C. S., & Jenson J. K. (1986). Individual differences in auditory capabilities. *Journal of the Acoustical Society of America*, *81*(2), 427–438 https://doi.org/10.1121/1.394907

Kalmus H., & Fry D. B. (1980). On tune deafness (dysmelodia): Frequency, development, genetics and musical background. *Annals of Human Genetics*, *43*(4), 369–382. https://doi.org/10.1111/j.1469-1809.1980.tb01571.x

Kidd G. J., & Colburn S. (2017). Informational masking in speech recognition. In Middlebrooks J. C., Simon J. Z., Popper A. N., & Fay R. R. (Eds.), *Springer handbook of auditory research: The auditory system at the cocktail party* (pp. 75–110). Springer-Verlag.

Kidd G. J., Mason C. R., Richards V. M., Gallun F. J., & Durlach N. I. (2008). Informational masking. In Yost W. A., Fay R. R., & Popper A. N. (Eds.), *Springer handbook of auditory research: Auditory perception of sound sources* (pp. 143–190). Springer-Verlag.

Kidd G., Mason C. R., Swaminathan J., Roverud E., Kameron, Clayton K., & Best V. (2016). Determining the energetic and informational components of speech-on-speech masking. *Journal of the Acoustical Society of America*, *140*(1), 132–144. https://doi.org/10.1121/1.4954748

Kidd G., Watson C. S., & Gygi B. (2007). Individual differences in auditory abilities. *Journal of the Acoustical Society of America*, *122*(1), 418–435. https://doi.org/10.1121/1.2743154

Kubiak A. M., Rennies J., Ewert S. D., & Kollmeier B. (2020). Prediction of individual speech recognition performance in complex listening conditions. *Journal of the Acoustical Society of America*, *147*(3), 1379–1391. https://doi.org/10.1121/10.0000759

Kujawa S. G., & Liberman M. C. (2009). Adding insult to injury: Cochlear nerve degeneration after "temporary" noise-induced hearing loss. *Journal of Neuroscience*, *29*(45), 14077–14085. https://doi.org/10.1523/JNEUROSCI.2845-09.2009

Lee J., & Long G. (2012). Stimulus characteristics which lessen the impact of threshold fine structure on estimates of hearing status. *Hearing Research*, *28*(1–2), 24–32. https://doi.org/10.1016/j.heares.2011.11.011

Liberman M. C., Epstein M. J., Cleveland S. S., Wang H., & Maison S. F. (2016). Toward a differential diagnosis of hidden hearing loss in humans. *PLoS ONE*, *11*(9), e0162726. https://doi.org/10.1371/journal.pone.0162726

Lopez-Poveda E. A. (2014). Why do I hear but not understand? Stochastic undersampling as a model of degraded neural encoding of speech. *Frontiers in Neuroscience, 8*, 348, 1–7. https://doi.org/10.3389/fnins.2014.00348

Lutfi R. A., Gilbertson L., Chang A.-C., & Stamas J. (2013). The information-divergence hypothesis of informational masking. *Journal of the Acoustical Society of America*, *134*(3), 2160–2170. https://doi.org/10.1121/1.4817875

Lutfi R. A., Rodriguez B., Lee J., & Pastore T. (2020). A test of model classes accounting for individual differences in the cocktail-party effect. *Journal of the Acoustical Society of America*, *148*(6), 4014–4024. https://doi.org/10.1121/10.0002961

Lutfi R. A., Tan A. Y., & Lee J. (2018). "Modeling individual differences in cocktail-party listening", special issue. *Acta Acustica United with Acustica*, *104*(5), 787–791. https://doi.org/10.3813/AAA.919246

McLaughlin D. J., Baese-Berk M. M., Bent T., Borrie S. A., & Van Engen K. J. (2018). Coping with adversity: Individual differences in the perception of noisy and accented speech. *Attention, Perception & Psychophysics*, *80*(6), 1559–1570. https://doi.org/10.3758/s13414-018-1537-4

Oberfeld D., & Klöckner-Nowotny F. (2016). Individual differences in selective attention predict speech identification as a cocktail party. *eLife*, *5*, e16747. https://doi.org/10.7554/eLife.16747

Peterson G. E., & Barney H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*(2), 175–184. https://doi.org/10.1121/1.1906875

Plack C. J., Barker D., & Prendergast G. (2014). Perceptual consequences of "hidden" hearing loss. *Trends in Hearing*, *18*. https://doi.org/10.1177/2331216514550621

Ruggles D., Bharadwaj H., & Shinn-Cunningham B. G. (2011). Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(37), 15516–15521. https://doi.org/10.1073/pnas.1108912108

Ruggles D., & Shinn-Cunningham B. G. (2011). Spatial selective auditory attention in the presence of reverberant energy: Individual differences in normal-hearing listeners. *Journal of the Association for Research in Otolaryngology, 12*(3), 395–405. https://doi.org/10.1007/s10162-010-0254-z

Swaminathan J., Mason C. R., Streeter T. M., Best V., Kidd G. J., & Patel A. D. (2015). Musical training, individual differences and the cocktail party problem. *Scientific Reports*, *5*, 11628. https://doi.org/10.1038/srep11628

Tamati T. N., Gilbert J. L., & Pisoni D. B. (2013). Some factors underlying individual differences in speech recognition on PRESTO: A first report. *Journal of the American Academy of Audiology*, *24*(7), 616–634. https://doi.org/10.3766/jaaa.24.7.10

Wier C. C., Jesteadt W., & Green D. M. (1977). Frequency discrimination as a function of frequency and sensation level. *Journal of the Acoustical Society of America*, *61*(1), 178–184. https://doi.org/10.1121/1.381251