Original Article

Trends in Hearing
Volume 26: 1–16
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/23312165221077555
journals.sagepub.com/home/tia
SAGE

The Effects of Uncertainty in Level on Speech-on-Speech Masking

Andrew J. Byrne[1], Christopher Conroy[1] and Gerald Kidd Jr.[1,2]

Abstract

Identification of speech from a "target" talker was measured in a speech-on-speech masking task with two simultaneous "masker" talkers. The overall level of each talker was either fixed or randomized throughout each stimulus presentation to investigate the effectiveness of level as a cue for segregating competing talkers and attending to the target. Experimental manipulations included varying the level difference between talkers and imposing three types of target level uncertainty: 1) fixed target level across trials, 2) random target level across trials, or 3) random target levels on a word-by-word basis within a trial. When the target level was predictable performance was better than corresponding conditions when the target level was uncertain. Masker confusions were consistent with a high degree of informational masking (IM). Furthermore, evidence was found for "tuning" in level and a level "release" from IM. These findings suggest that conforming to listener expectation about relative level, in addition to cues signaling talker identity, facilitates segregation of, and maintaining focus of attention on, a specific talker in multiple-talker communication situations.

Keywords

speech recognition, informational masking, competing talkers, cocktail party

Received 10 May 2021; Revised received 11 January 2022; accepted 13 January 2022

Introduction

The difficulty experienced by a human listener attempting to focus attention on the voice of one particular talker (the "target") and recognize the message conveyed by the target talker's speech in the presence of multiple competing talkers ("maskers") is known as the “cocktail party problem” (CPP; see Middlebrooks et al., 2017, for a series of recent reviews). In typical multiple-talker communication situations, the listener has many different cues that may assist with solving the CPP by enhancing the perceptual segregation of sound sources and aiding in the focus of attention on the target while ignoring competing masker speech or other unwanted sources of environmental "noise." These cues include both auditory and visual information (e.g., lip reading, Woodhouse et al., 2009), as well as higher-level linguistic factors such as syntactic and semantic context (e.g., Brouwer et al., 2012; Kidd et al., 2014; review in Mattys et al., 2012). Spatial separation of sound sources may also afford a substantial improvement in speech recognition performance in multiple-source sound fields compared to conditions where competing sources appear to emanate from the same spatial location (i.e., "colocated" with the target source; e.g., Best et al., 2013a, 2013b; Hawley et al., 2004; Marrone et al., 2008). Furthermore, multiple factors may combine to facilitate the maintenance of attentional focus on the target speech stream under masked conditions (e.g., Rennies et al., 2019; Rodriguez et al., 2021; Swaminathan et al., 2015).

Because natural speech is inherently a dynamic stimulus, following one specific talker among competing talkers involves using the patterns of variation of the target talker's voice to anticipate that talker's speech. Prosodic information, which involves the plausible variation in vocal pitch, intensity, and timing, may be particularly useful in that regard (e.g., Calandruccio et al., 2019; Kim & Sumner, 2017; Zekveld et al., 2014). Of the cues that comprise prosody, variation in level has received less attention as a

[1]Department of Speech, Language, & Hearing Sciences, Boston University, MA, USA
[2]Department of Otolaryngology, Head-Neck Surgery, Medical University of South Carolina, Charleston, SC, USA

Corresponding Author:
Andrew Byrne, Department of Speech, Language, & Hearing Sciences, Boston University, Boston, MA, USA.
Email: ajbyrne@bu.edu

potential cue in CPP communication situations than other factors, such as pitch and intonation. This is due in part to the obvious strength of vocal pitch as a cue: pitch may signal the sex/age of the talker (Hazan & Markham, 2004), indicate declarative/interrogative statements, and aid in talker identification. Furthermore, because speech is intelligible over a wide range of absolute levels, once it is above detection threshold it could be assumed that level differences play a secondary, even insignificant, role in maintaining speech stream integrity because each individual unit (i.e., word) is fully audible/identifiable.

Although the effect of talker intensity or loudness (i.e., talker "level"), has been a variable in the aforementioned work, it most often is used either for measuring psychometric functions for speech recognition, or for quantifying performance when level is adapted to measure a speech reception threshold. In either case, the strength of the relative intensity of the target as a cue in multiple-source, CPP communication situations either is implicit or is easily explained by energetic masking (EM). That is, when the target speech is more intense than the masker(s), its spectral energy dominates the stimulus representation in the auditory system, presumably causing it to be more salient perceptually and easier to segregate and understand than when it is less intense than the masker(s). Furthermore, the relative levels of target and masker speech produce varying "glimpses" of target energy that must be integrated by the listener to accurately identify the message conveyed by the target (e.g., Brungart et al., 2006; Cooke, 2006). Broadly speaking, as the relative intensity of the target increases the extent to which glimpses of target information are accessible tends to increase, allowing it to be segregated from competitors and perceived as perceptually distinct.

There is evidence, however, that recognition is not always a monotonic function of overall level. For instance, with colocated speech-on-speech (SOS) masking, well-trained observers can produce psychometric functions with a relative improvement of recognition on either side of a 0-dB target-to-masker ratio (TMR; e.g., Brungart, 2001; Dirks & Bower, 1969; MacPherson & Akeroyd, 2014). One explanation for this distinct dip in performance of the psychometric function is that the target and masker words are more easily confused when there is no consistent relative level cue available between the speech streams. This does not necessarily reflect a failure to perceive specific words, but rather indicates a failure to focus attention on the correct talker while ignoring others, a characteristic of informational masking (IM; see discussion in Kidd & Colburn, 2017). Brungart (2001) found that the amount of IM could be varied with different types of maskers, and that psychometric functions were the most orderly (monotonic) for low-IM maskers. Conversely, under high-IM conditions the psychometric functions could flatten or even become “U” shaped. When the target and the masker were the same talker (the configuration with the greatest IM), level became the only cue to distinguish the target speech stream, and there was convincing evidence that a listener could effectively attend to the less intense of two talkers. This improvement in speech recognition at negative TMR values suggests that a listener may actively focus attention on a lower-level speech stream, despite the presumably greater EM that it creates, compared to higher TMRs. However, evidence in support of level as an effective segregation cue separate from its relation to EM is quite limited.

Given that relative level differences between a target and masker(s) may exert an influence on the success of selective attention to the target speech stream under conditions high in IM, the present study aimed to evaluate the effectiveness of level as a cue for speech segregation more directly. Specifically, we pose the question of whether a known target level (i.e., low target level uncertainty) can be used to focus attention at that level, ignoring other voices which are either louder or softer than expected. Conversely, any benefit that might be obtained from attentional focus on level could be mitigated by listener uncertainty about where in level attention should be focused.

The present study investigated target speech identification in the presence of competing speech maskers with various amounts of level separation between the talkers, as well as with varying degrees of talker level uncertainty, for sentences having identical syntactic and semantic structure. Moreover, the use of a matrix identification task in which both the target and masker sentences comprised individually concatenated words (see Kidd et al., 2008) mitigated the influence of prosodic information so that the effectiveness of level as a cue could be ascertained more directly. It is hypothesized that *a priori* knowledge and consistency of talker level will result in better speech identification than when level cues are uncertain.

Although there has been little evidence to date supporting the proposition that selective attention in level produces a "tuned" response (i.e., enhanced performance at the point of attentional focus along the given stimulus dimension) in SOS masking, past studies have not examined the issue under different degrees of stimulus uncertainty. Tuning at the focus of attention can be most pronounced when the conditions produce a high degree of uncertainty in the listener, and this idea forms a part of the rationale for the experimental design in the work that follows.

Methods

Subjects

Eleven listeners participated in the experiment (two male and nine female, 20–40 years old, mean age: 25.4), two of which were the first and second authors (designated as S1 and S2, respectively, in a later figure), while the others were students from Boston University who were paid to participate. All participants had normal hearing based on pure-tone thresholds of 20 dB hearing level (HL) or better at octave frequencies from 250 to 8000 Hz. (Listener audiograms were measured in the lab prior to online testing.) Each participant reported U.S. English as their first and primary language, and all had previous experience with other psychoacoustics tasks, including the speech corpus and response method described below. All procedures were approved by the institutional review board of Boston University, and all participants gave written informed consent.

Stimuli and Procedure

The speech identification task used was an established matrix identification design and speech corpus (e.g., Helfer et al., 2020; Holmes et al., 2021; Kidd et al., 2008, 2020; Puschmann et al., 2019) and consisted of three simultaneous co-located talkers. Each talker spoke (artificially constructed) five-word sentences, one word from each of the five syntactic categories of name, verb, number, adjective, and object (in that fixed order), e.g., “Sue took ten small toys”. Natural production of eight exemplar words in each syntactic category of the corpus had been previously recorded from twelve different female talkers (refer to Kidd et al., 2008). Non-individualized head-related transfer functions (HRTFs; Gardner & Martin, 1994) were used to spatialize all speech to be at 0° azimuth and elevation for binaural presentation over headphones.

On each experimental trial, one of the twelve talkers was randomly selected to be the target talker, and the first word of the target sentence was always the name "Sue," which served to cue the listener to the target voice. The other four words of the target sentence were randomly selected from the eight exemplars available in each of the remaining four syntactic categories. Two different masker talkers were then randomly selected from the remaining talkers, and two different masker sentences were constructed from additional random selection of words (without replacement) from each category. The sentences were constructed such that the first word of each talker began simultaneously; however, all following words were concatenated without any attempt at time-aligning the words within each category across talkers. This design was chosen to accentuate the “multimasker penalty” (Iyer et al., 2010) with not only a simultaneous multi-talker masker, but also with all maskers being contextually relevant; therefore, increasing the potential for effects (both improvements in performance as well as deficits) to occur.

The task of the listener was to focus on, and respond to, only the words spoken by the target talker on each trial. These responses were obtained by the listener clicking on the target words that were heard, as shown on a graphical user interface (GUI) displaying the matrix of possible words in syntactic columns and exemplar rows. The cue word “Sue” was always pre-selected, and the listener was required to make a selection in each of the other categories in the syntactic order that they were presented. Correct answer feedback (displaying and highlighting the correct words) was provided only after all four responses for that trial were selected.

The controlled variable across conditions was the level of the target and maskers, as defined by the root-mean-square of the waveform of each individual word. Due to remote testing (described in the next section), absolute values of sound pressure level (SPL) are not specified; instead, the relative level of the target to the masker is given. However, as an approximate reference based on the remote calibration routine, a relative level of 0 dB roughly corresponds to 60 dB SPL.

For the "Fixed Level" (FL) conditions, the target and each masker were presented at constant levels within each trial and throughout each block of 12 trials. Across blocks, the level difference (ΔL) between the three talkers varied from 0 to 9 dB, in 3 dB steps, but that spacing in level was fixed for each block of trials. One talker was always at 0 dB, while the other two talkers were at + /-ΔL. Note that the ΔL value refers to adjacent talker levels and thus the total range of levels is twice the value (e.g., a level spacing of 9 dB between adjacent talkers equals an 18-dB range between loudest and softest talkers). The level of the target will be designated as “Loud”, “Mid”, or “Soft” relative to the other talkers, given levels of + ΔL, 0, or -ΔL, respectively. Prior to each block, the designation "Fixed Level", followed by the relative target level designation, was displayed on the GUI, so that the listener could try to focus attention at the appropriate intensity. Figure 1 (upper row) illustrates the FL condition for a “loud” target level.

In another set of conditions, the “Random Sentence Level” (RSL) conditions, ΔL was fixed for a block of trials, but the level of the target relative to the maskers was varied. Level was randomized from trial to trial, but held constant within a trial, and there was a new random permutation order chosen for each block, such that there were always four targets of each relative level in each block. Prior to each block, the designation "Random Sentence Level" was displayed on the GUI, so that the listener would be informed of the inconsistent target level across trials and instead would be required to use the cue word “Sue” to designate the target speech stream. An illustration of the RSL condition is shown in the middle panels of Figure 1. In contrast to the FL condition (upper panels), the target level is uncertain from trial to trial but is consistent within a trial.

The final set of conditions involved changes in level across words within a sentence: the “Random Word Level” (RWL) condition. The level of each word was randomly selected from the set of three relative levels for that block, with the stipulation that no two consecutive target words were ever at the same level, and no sentence contained more than three target words at a given level. Given the randomization strategy utilized, there were 24 different level variations possible for the target sentence on each trial, with 75% of those variations including a subsequent target word being presented at the cue level. For both of the other

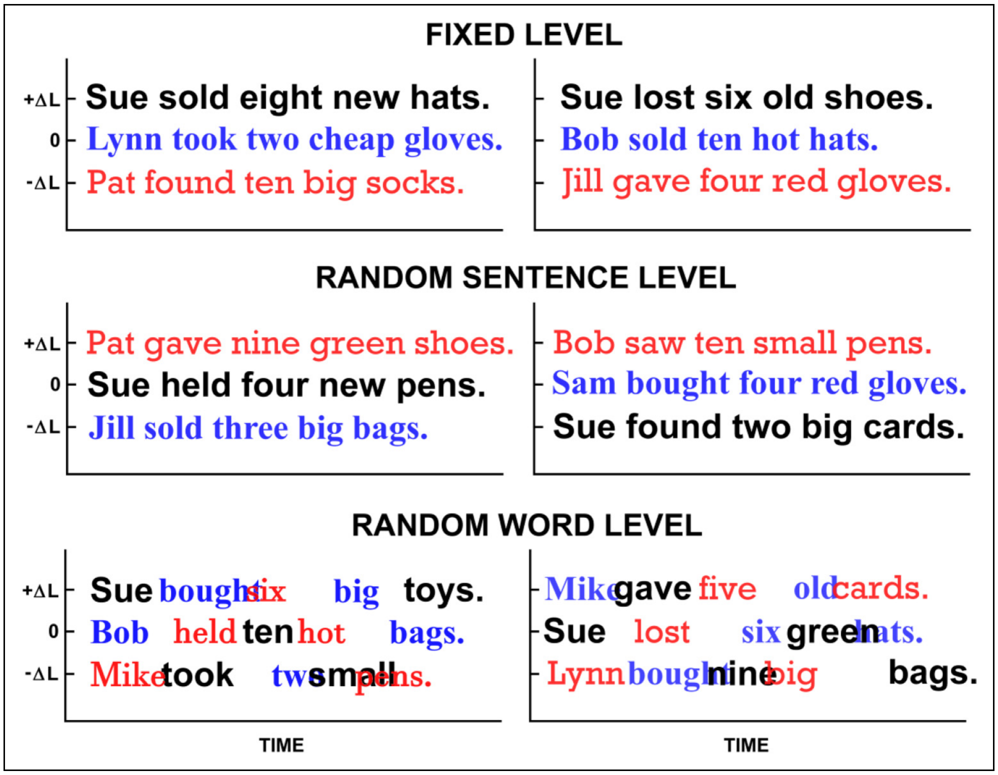

Figure 1. Illustrations of two possible trials (each column) for the Fixed Level (FL; top panels), Random Sentence Level (RSL; middle panels), and Random Word Level (RWL; bottom panels) conditions. Overall level difference by word (y-axis) is shown as a function of time along the x-axis, with different colors/fonts depicting the different simultaneous talkers. The target sentence always begins with the name "Sue".

(non-cue) levels, there was only a 62.5% chance of a variation including a double presentation at the given level. This resulted in the level of the "average" target sentence (across all possible variations) being shifted slightly towards that of the cue.

An illustration of a sample trial of the RWL condition is shown in the lower panels of Figure 1. (Note that the jumbled/overlapping appearance of the words illustrated for this condition is simply due to their display after sorting by level; the words in all three conditions shown in Figure 1 temporally overlap in roughly the same way.) In addition, as with the RSL condition, the level of the cue word for the target sentence was randomized on each trial such that there were equal numbers of trials starting with the cue word at the Loud, Mid, and Soft levels. Again, the listener was informed prior to the experimental block with the label "Random Word Level" and that, for these conditions, the only reasonable strategy would be to listen for the cue word and attempt to follow the vocal characteristics of the target talker while ignoring intensity fluctuations. (The difference between all three types of conditions was emphasized in the instructions to the listeners prior to data collection.)

The full experiment consisted of 112 blocks, with four blocks of each of the four FL conditions (four different ΔL's), and 16 blocks of each RSL and RWL condition at each of the three ΔL possibilities. A one-quarter complete set of conditions (28 blocks) was obtained before the next repetition, and the condition order within each set was randomized. Listeners ran the experiment self-paced, typically with 1-h sessions (including brief breaks), and completed the full experiment in 5–6 hours total.

Remote Testing Considerations

All data included in the present study were gathered remotely.[1] Thus, listeners were allowed to perform the task at home using their own computers and headphones. However, prior to participating in the experiment, each participant was required to sign an online "Attestation of Low-Noise Distraction-Free Testing" form, agreeing to do the experiment under the quietest and most distraction-free testing environment that they could create at home.

A MATLAB (MathWorks, Natick, MA) experiment was built and compiled as a "Web App" and run on a Microsoft (Redmond, WA) Windows 2019 Virtual Machine (VM) with MATLAB Web App Server installed, which was open to the Internet. The GUI was accessed and displayed within an Internet browser window. Audio presentation consisted of a trial-by-trial uncompressed WAV file (44100-Hz sampling rate) generated by the application on the VM and presented through the computer audio hardware of the listener. Each listener used a different model of personal headphone, with both standard and earbud, as well as wired and wireless, headphones being used. The headphones used were: Apple (Cupertino, CA) AirPods Pro (three listeners), Apple EarPods (two listeners), Beats Electronics (Culver City, CA) Flex, Jabra GN (Copenhagen, Denmark) Elite 65t, Samsung (Seoul, South Korea) Level On, Sennheiser HD300, Sennheiser HD380 Pro, and Sony (Tokyo, Japan) MDR-V6. The exact hardware configurations (e.g., sound cards, audio enhancement settings, wireless compression protocols) of each remote setup were unknown.

Calibration for remote testing was done prior to the start of each experimental session. The listener was presented with sample speech (from the experimental stimuli) over a 30-dB range. The audio samples were labeled: “loud voice (75 dB)”, “normal conversation (60 dB)”, and “quiet voice (45 dB)”. The listener then adjusted the overall computer volume until the speech subjectively matched the given labels. By no means should it be assumed that each listener was able to set “60 dB” speech to an overall level of 60 dB SPL, but by setting the softest speech to be both audible and understandable, and the loudest speech to be still “comfortable”, the dynamic range of the stimuli presented in the present task (18 dB, with 0-dB ΔL calibrated as 60 dB “SPL”) was considered to be within an acceptable range of values for testing relative level effects in this speech identification paradigm.

Results

Speech identification performance was defined as the proportion of words that the listener identified correctly in the final four words of the target sentence (i.e., excluding the initial target cue word “Sue”). Broadly, and as expected, performance varied with changes in the level of the target words and as the level difference between the target and maskers was altered. Furthermore, in support of the underlying premise of the study, performance also depended on the degree of uncertainty with respect to the target level.

The mean results across the eleven listeners are shown in Figure 2. Each plotted point represents the average proportion correct of 48 trials for each listener, with the group means across listeners and standard error of the means (SEM) plotted in the figure. A 3-by-3 repeated-measures analysis of variance (ANOVA) performed on the FL data (not including the 0-dB ΔL reference point) revealed that both cue level (Loud, Mid, Soft) [$F(2,20) = 115.12$, $p < 0.001$] and level difference (3, 6, and 9 dB) [$F(2,20) = 49.04$, $p < 0.001$] were significant main effects, as was the interaction of the two factors [$F(2,40) = 18.22$, $p < 0.001$]. For the FL conditions (top panel), performance improved from 0.53 with no level difference across the three talkers (which is the reference in each panel for gauging the effects of relative level) to nearly perfect performance with a 9-dB separation for the Loud Cue condition (solid squares). This result would be expected simply due to the increased salience of the more intense target voice over the masker voices (cf., Brungart, 2001). For both the Mid and Low Cue conditions (shaded and open squares, respectively), group mean performance was relatively constant and near the value obtained for 0 dB (i.e., equal talker level) across the range of talker level differences, with some separation between Mid and Soft values apparent for the 9 dB level difference.

It should be noted that a purely EM-based explanation of the results would not seem to be compatible with the trends in performance found for the Soft cue condition. In that case, performance remained constant as both the target voice was made less intense and one of the maskers was made more intense. In order to evaluate this impression quantitatively, a glimpsing analysis based on "ideal time-frequency segregation" (e.g., Best et al., 2017; Brungart et al., 2006; Conroy et al., 2020; Kidd et al., 2016) was performed on these stimuli and is described more fully in the Appendix. The analysis confirmed that this finding could not be attributed to changes in EM (which is increasing, while performance was stable) and was, instead, evidence of a release from IM, presumably from level cues facilitating attentional focus on the target.

The results for the RSL condition (bottom left panel) were qualitatively quite similar to those from the FL condition and the trends were roughly the same. A 3-by-3 repeated-measures ANOVA for the RSL condition yielded significant main effects of both cue level [$F(2,20) = 150.50$, $p < 0.001$] and level difference [$F(2,20) = 13.55$, $p < 0.001$], as well as a significant interaction [$F(4,40) = 15.56$, $p < 0.001$] of the two factors. The similarity of the findings for FL and RSL conditions suggested that a constant level for each talker throughout the trial interval was generally sufficient for the listener to focus on the target speech.

The three functions for the RWL condition were not that different from each other and all showed a moderate decrease in performance of roughly 0.12 proportion correct relative to the reference condition as the level difference was increased. A 3-by-3 repeated-measures ANOVA for the RWL conditions revealed that the main effect of cue level was significant [$F(2,20) = 4.91$, $p = 0.002$], as well as level difference [$F(2,20) = 20.80$, $p < 0.001$], but with no significant interaction [$F(4,40) = 0.36$, $p = 0.97$]. Recall that the designation of the functions as Loud, Mid, or Soft for this condition refers *only* to the cue word, given that the levels of the remaining target words were chosen pseudo-randomly (see the description of the RWL condition in the Methods above). This explains the similar performance shown for each function. If the salience of the cue word, as implied by its relative level, facilitated extracting the subsequent target words then we would have expected some degree of ordering of the functions from Loud to Soft. Although the Soft Cue function does appear slightly worse than the

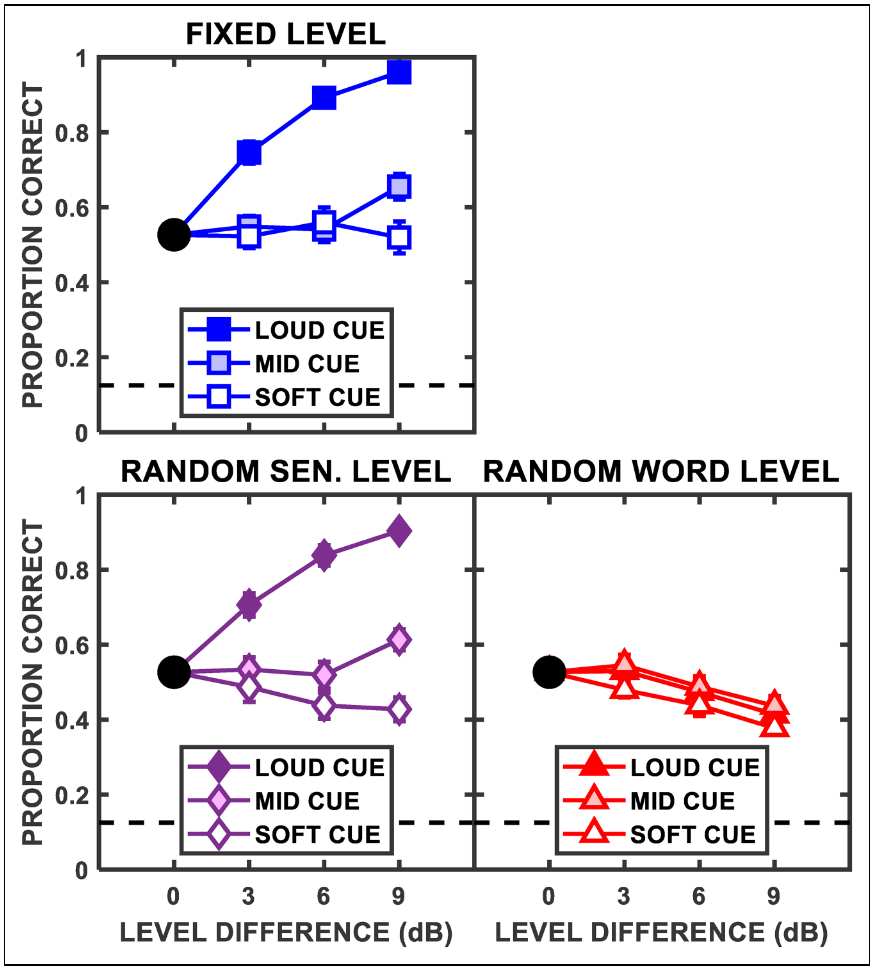

Figure 2. Group mean speech identification results for the three different amounts of target talker level uncertainty, shown in separate panels. The level difference between the three simultaneous talkers forms the x-axis in each panel, while the y-axis is the mean proportion correct performance, with SEM error bars. Within each panel, the separate functions indicate the conditions where the target cue word was at the Loud, Mid, or Soft relative levels (solid, shaded, and open symbols, respectively). For the Random Word Level condition, each function contains target words from all levels with the percent correct values computed based on the cue word level of each trial (see text). The dashed horizontal line represents chance performance.

others, simply providing a more intense cue word was not sufficient to compensate for the level variation of the subsequent words in the target sentence.

To examine the role of target level uncertainty explicitly, the group mean data for the FL, RSL, and RWL conditions were transformed and plotted as psychometric functions in Figure 3. The values along the x-axis are the relative levels of the individual target words while the values on the y-axis are the proportion correct performance. Note that, for the FL and RSL conditions, the relative levels of the target words were the same as the cue word level (as illustrated in Figure 1). However, for the RWL condition, the relative levels of the target words corresponded to the actual levels of the words independent of the level of the cue word. Referring back to the lower panels of Figure 1, to compute proportion correct for the $+\Delta L$ word level for the RWL condition, identification performance for only the highest-level words (upper row) from both of the two trials illustrated would be counted, even though the target cue level was different across trials. The same is true across all trials and for all cue word levels. Thus, the proportion correct represents a word-by-word computation of proportion correct according to target word level regardless of the levels of the cue or the other words in the sentence.

A 3-by-7 repeated-measures ANOVA revealed that both the degree of uncertainty [$F(2,20) = 189.73$, $p < 0.001$] and the relative level [$F(6,60) = 106.88$, $p < 0.001$] were significant main effects, as was the interaction between the two factors [$F(12,120) = 53.85$, $p < 0.001$]. [A post hoc statistical power analysis was performed using G*Power (Faul et al., 2007) indicated that the $N = 11$ used in the present study was sufficiently powered (greater than 0.95, with $\alpha = 0.05$)

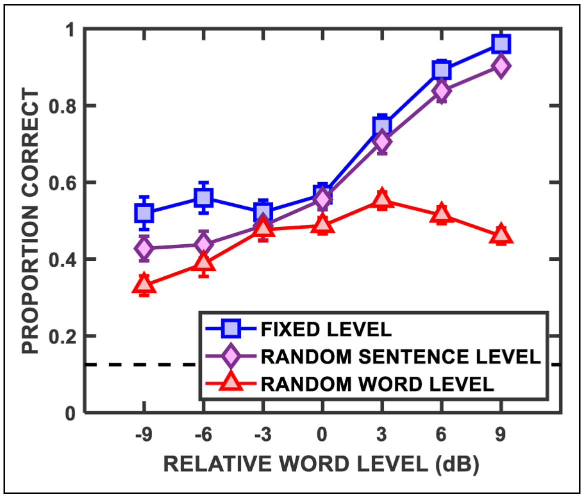

Figure 3. Group mean speech identification performance, as a function of the relative level of the individual target words. The psychometric functions shown are for the three types of level uncertainty: the Fixed Level, Random Sentence Level, and Random Word Level conditions, which are plotted together in each panel (squares, diamonds, and triangles, respectively). The error bars represent the SEM.

to rule out the null hypothesis that the degree of uncertainty (i.e., the main effect between conditions) was not significantly different.] Performance for the FL (blue square) condition improved steadily as the relative level cue was increased above the 0-dB reference but was roughly constant below 0 dB. (Refer to the Appendix for the pattern expected from only energetic masking.) This flattening of the psychometric function below the mid-point was likely due to the ability of listeners to focus on the softest voice during the FL blocks. A similar effect was shown in some of the studies reviewed by MacPherson and Akeroyd (2014) and suggested that prior knowledge of target level can improve speech identification under conditions high in IM. For the RSL condition, the function was quite similar to the FL function above 0 dB, but declined somewhat more below 0 dB. (Post hoc t-tests, with the Bonferroni correction applied, confirmed that the FL function was significantly greater than RSL function at both the −6 and −9 relative levels: −6 dB, $t(10) = 3.30$, $p = 0.004$, −9 dB, $t(10) = 5.79$, $p < 0.001$.) This suggested that, as a group, listeners were less able to use the softer target voice as a cue than was the case in the FL condition. Here, apparently, the reduced variability across trials in the FL condition yielded an additional benefit for focusing on the target stream. For the RWL condition, performance varied less with relative level and, below 0 dB, declined with increasing level difference to 0.33 proportion correct for the largest level difference (but still substantially above chance performance).

To highlight the individual differences found, and given how common across-subject variability can be in IM tasks (see Lutfi et al., 2021), the individual listener data are plotted in Figure 4 in the same manner as Figure 3. Post hoc between-subjects effects were found to be significant: main effect of listener [$F(10,45) = 12.67$, $p < 0.001$], degree of uncertainty by listener interaction [$F(20,120) = 2.97$, $p < 0.001$], and relative level by listener interaction [$F(60,120) = 3.41$, $p < 0.001$]. Although the general trends were broadly similar across subjects, there were a few noteworthy differences. Some subjects (e.g., S1 and S5) showed an improvement at the lower relative levels for the FL condition over the RSL and RWL conditions with the psychometric function appearing flat or even "U"-shaped. In contrast, the performance of other subjects (e.g., S3 and S4) at the low levels continued to decline steadily with decreasing relative level, thus creating more typical monotonic psychometric functions. These individual differences may suggest an inability of some listeners to focus attention on only the softest talker (when explicitly told to), i.e., attentional tuning in level. For the RSL case, the function at lower levels was generally diminished relative to the FL case for most listeners, indicating that the greater uncertainty about the target level on a trial-by-trial basis affected the ability to focus on the softer talker during the trial. There was less variability across subjects in the shapes of the psychometric functions for the RWL condition where all listeners produced shallow, somewhat convex psychometric functions indicating little benefit, and indeed occasionally some cost, for higher or lower relative levels for the target words.

Additional Evidence of Tuning in Level

Level Release from Masking

It was also of interest to examine the extent to which focusing attention at the target's value along the physical dimension of level led to a "tuned" response pattern (e.g., analogous to what has been found for the dimensions of frequency/fundamental frequency and spatial location; cf. Arbogast & Kidd, 2000). The Mid Cue functions shown in Figure 5 (replotted from Figure 2) revealed some support for this proposition: There was an increase in performance for the FL and RSL conditions at the largest level difference across talkers, despite the overall level of the target voice being held constant (at 0 dB relative level) across all ΔL's in the Mid Cue condition, while the two masker talkers varied symmetrically in level above and below the target. Post hoc paired-samples t-tests for the FL and RSL functions confirmed that the improvement in performance from the 0-dB ΔL condition was significant at 9-dB ΔL: FL, $t(10) = -7.91$, $p < 0.001$; RSL, $t(10) = -4.92$, $p = 0.002$.

This effect would be analogous to the improvement in performance seen when increasing the difference between the target and masker along different dimensions using other psychophysical tuning paradigms (e.g., spatial release from masking, Marrone et al., 2008). This pattern could suggest that, with a 9-dB separation in level between talkers, listeners could begin to more easily differentiate all three talkers as separate streams based on relative location along the level

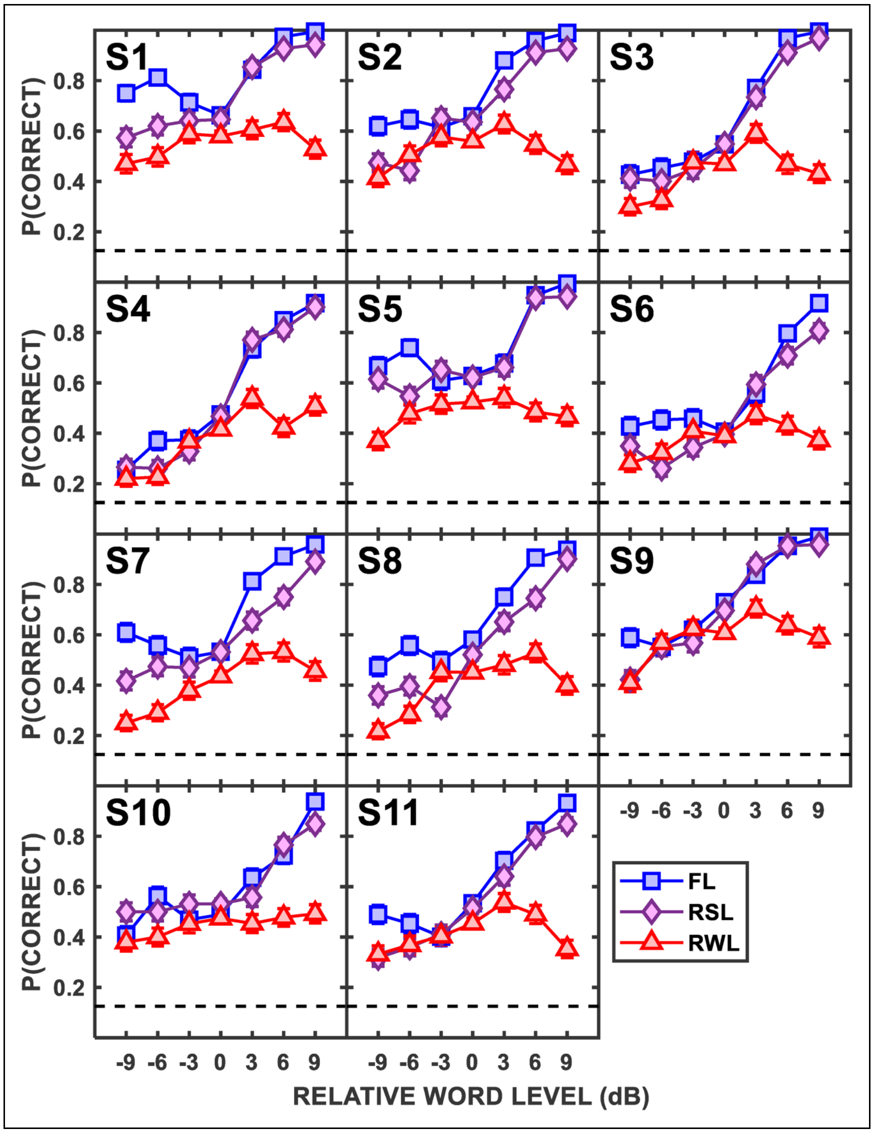

Figure 4. Speech identification performance for each listener (separate panels) is plotted as a function of the relative level of the individual target words, with error bars representing the standard error of the proportions.

dimension; an example of a level "release" from IM. The apparent benefit of a 9-dB level separation was not due to reduced EM, however; a point supported by the glimpsing analysis described in the Appendix. This trend was not seen for the RWL condition, however, as target words were presented at all relative levels and only the cue word was guaranteed to be at the middle level.

Attentional Tuning to the Cue Level

In order to solve either of the random target speech identification (RSL and RWL) conditions, the listener must have recognized the target cue word ("Sue") at the beginning of the sentence and then followed that talker's voice throughout the remainder of the stimulus while ignoring the competing speech streams. The FL and RSL functions discussed above and shown in Figure 5 suggested that, in an attempt to solve the speech identification task, the listeners used the cue word on each trial to form an attentional filter centered at the location (in level) of the cue and reported the words on that trial that were consistent with that level. An additional analysis of the listener's responses relative to the cue level in the RWL condition suggested that the listeners did indeed employ such a strategy, despite level being uninformative for that condition.

In Figure 6, the RWL condition was analyzed with each target word in the sentence being compared to the cue word level on a given trial. That is, the target words were sorted according to the difference in level from the cue word and proportion correct performance was computed for each level difference separately. Thus, the x-axis in

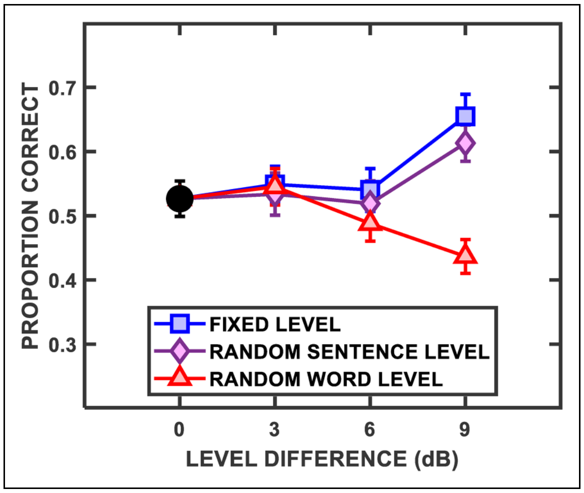

Figure 5. Mid Cue functions, replotted from Figure 2, highlighting the increase in performance at the largest level difference for the Fixed Level and Random Sentence Level conditions. The error bars represent the SEM.

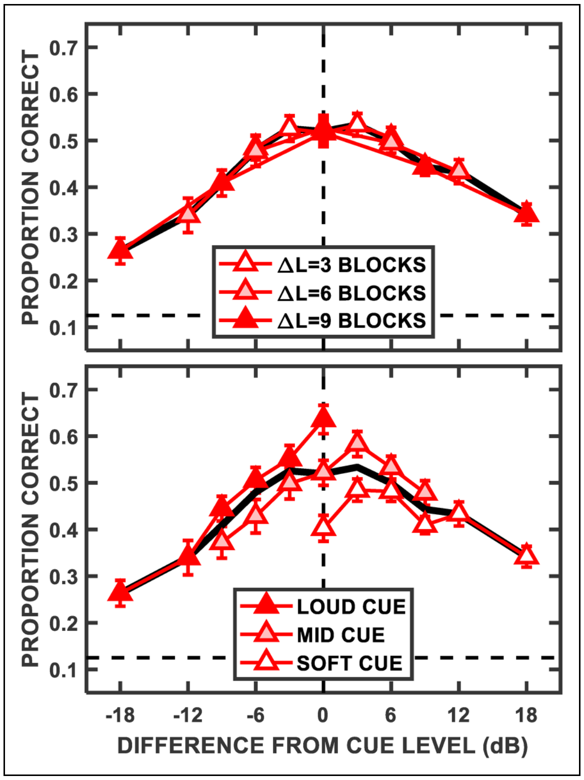

Figure 6. Group mean performance based on the difference in target word level from the current trial's cue level for the Random Word Level condition. The different functions indicate either the difference in level between the three talkers (top panel) which was fixed for a block of trials, or the cue level (bottom panel) on a given individual trial. The solid black line is the average across all trials, and the error bars represent the SEM.

Figure 6 is the difference in level between subsequent words in the target stream and the target cue word, computed over all RWL trials (a positive value indicates the target word was more intense than the cue word). The top panel plots functions obtained for the three relative level separations (i.e., ΔL's of 3, 6, and 9 dB), each presented in separate blocks of trials, with the heavy black line as a fit to all trials (the mean across functions). The results strongly suggested a tuned pattern of responses for relative level: listeners performed the best (53% correct) when the word was presented at, or very near, the same level as the initial cue word and performance declined (down to 34% and 26%, at + /-18 dB, respectively) as the level was changed from that of the cue. Post hoc paired-samples *t*-tests comparing the 0-dB level difference to the + /-18 dB difference endpoints confirmed that the decrease in performance was significant: −18 dB, $t(10) = 9.26$, $p < 0.001$, + 18 dB, $t(10) = 5.63$, $p < 0.001$.

The bottom panel of Figure 6 sorts the same data set by the level of the cue (as “Loud," "Mid," or "Soft”) on each trial to show any effect that the salience of the cue may have had, something which was not clearly evident in Figure 2. With no level difference (the 0-dB points, the only value all three cue levels share), the effect of cue level was quite clear and ordered as would be predicted. For the Loud Cue function (solid symbols), performance simply improved as the target word approached the cue level, likely due to the increased intensity/salience of the target words; however, both the Mid and Soft Cue functions (shaded and open symbols, respectively) were non-monotonic. In both cases, the function does not simply increase with the intensity of the target word, another trend which does not follow an EM pattern (see the Appendix).

Masker Confusions

Analyzing “masker confusions” for matrix identification speech-on-speech tasks can also provide considerable insight into the factors underlying the incorrect responses that listeners make (e.g., Brungart, 2001; Brungart et al., 2001; Iyer et al., 2010; Kidd et al., 2005). High proportions of masker confusions (i.e., selecting words from masker sentences rather than words from the target sentence) is interpreted as evidence for IM, whereas selection of words not present in the stimulus (i.e., random choices) is consistent with performance limited by EM. This is because listeners tend to choose more masker words when the competing masker streams are both intelligible and similar (e.g., same gender talker) to the target, rather than making random word selections (Brungart, 2001). This can be due to weak segregation cues and/or because the masker will frequently intrude into a listener's focus of attention. Thus, masker confusions can provide insights into the types of cues that listeners use to segregate competing talkers and attend to a target under high IM conditions. If listeners rely on level cues, situations in which level differences among talkers are minimal should produce more masker confusions. Moreover, if listeners are attending to a range of levels surrounding the target, masker confusions should be more frequent for maskers that fall within that particular range.

An analysis of masker confusions for the current study is presented in Figure 7 according to the relative level of the target cue for all three conditions: FL, RSL and RWL. To obtain the data plotted in Figure 7, incorrect responses were compared to the two masker sentences that were presented on each trial. Incorrect responses that corresponded to the louder (more intense) masker sentence (on a word-by-word basis) are plotted as the (blue) upward triangles, while responses that corresponded to the softer (less intense) masker are plotted as the (red) downward triangles (in this context "louder" and "softer" refer to the masker talkers relative to each other, not relative to the target word level; the 0-dB ΔL condition was not included). An incorrect response that was consistent with neither masker word was labeled as a random choice.

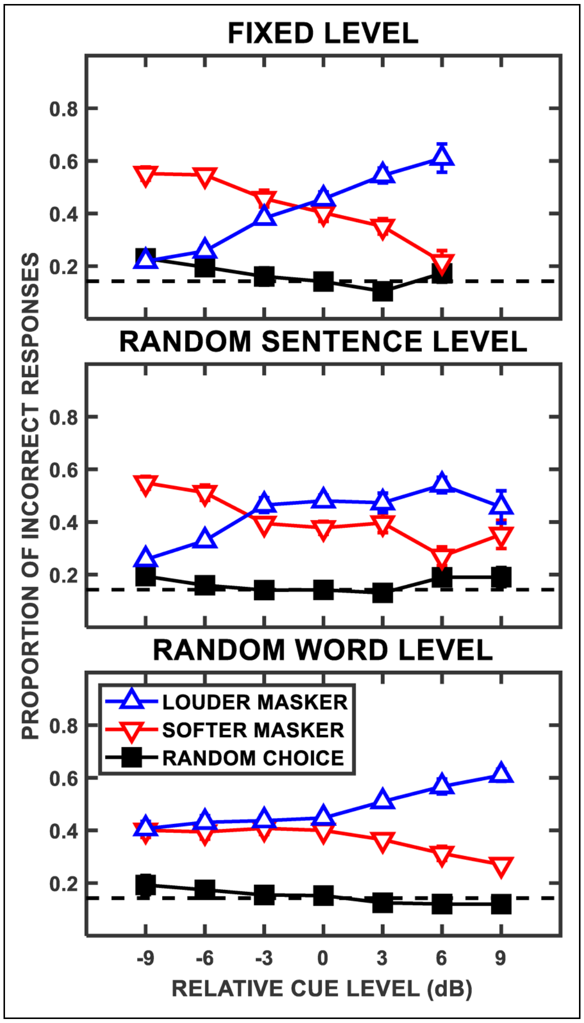

Figure 7. Group mean masker confusion functions for each type of condition (different panels), with SEM error bars. The proportion of incorrect responses that were either consistent with the louder masker word (blue upward triangles), the softer masker word (red downward triangles), or a random word choice that was not presented (black squares), as a function of the relative level of the target cue word.

When the relative cue level was positive for the FL condition, the listeners tended to have a higher proportion of confusions with the more intense masker. (For the largest relative cue level, 9 dB, performance was at or near perfect for all listeners, so there were insufficient incorrect responses for the masker confusion analysis.) Conversely, confusions with the less intense masker were made at negative target cue levels. (Random choice confusions were not substantially different from the horizontal chance line.) The functions for these two types of errors were uniformly rising/falling over the range of target word levels; thus, there was a clear level proximity effect between the masker confusions and the levels of the target cue words. A post hoc 2-by-6 repeated-measures ANOVA performed for only the louder- and softer-level masker confusion functions revealed that there was no main effect of Loud versus Soft confusions [$F(1,10) = 0.08$, $p = 0.78$] (this is not unexpected given the symmetry of the functions), but relative cue level was significant [$F(5,50) = 4.16$, $p = 0.03$], and there was a significant interaction [$F(5,50) = 35.35$, $p < 0.001$]. For the lowest relative level (-9 dB), the less intense masker was in fact selected significantly more often than the more intense masker [$t(10) = -12.95$, $p < 0.001$]. This finding may be considered as yet another indication of tuning to level because a listener was more likely to respond to a softer masker after a soft cue, rather than the (likely more salient) louder masker.

A similar trend was seen for the RSL condition (middle panel), although the functions were somewhat flatter, possibly due to the listener needing to initially focus on a broader range of levels within a trial, rather than having *a priori* knowledge of target level before the block began. The post hoc 2-by-7 repeated-measures ANOVA performed for the RSL condition had no significant main effects, but a significant interaction was found [Loud versus Soft: $F(1,10) = 0.64$, $p = 0.44$, relative cue level: $F(6,60) = 1.64$, $p = 0.15$, interaction $F(6,60) = 10.48$, $p < 0.001$]. Similar to the FL results at a −9 dB relative level, the less intense masker of the RSL condition was selected significantly more often than the more intense masker [$t(10) = -8.50$, $p < 0.001$].

Finally, for the RWL condition (bottom panel), there was a pronounced bias toward the louder masker during loud cue trials, while at negative relative cue levels, masker confusions were about equal between the two maskers. For the RWL condition, both main effects and the interaction were significant [Loud versus Soft: $F(1,10) = 34.00$, $p < 0.001$, relative level: $F(6,60) = 4.96$, $p < 0.001$, interaction $F(6,60) = 15.53$, $p < 0.001$]. This trend may have been caused by a soft cue being less salient and the listener not always knowing which talker to focus on, and thus arbitrarily choosing a voice to follow. (Note that the masker confusions were sorted according to the level of the target cue only, since the levels of the words following the cue were uncertain.)

DISCUSSION

The present study manipulated the degree of *a priori* knowledge about the relative level of a target talker in a speech-on-speech masking task. The goal was to evaluate the benefit of certainty about target level over time and the effectiveness of level as a cue in overcoming IM, while controlling the influence of other source segregation cues which are known to affect listener performance in CPP situations.

Overall speech identification performance was much better when the level of a talker's voice was held constant over the duration of a sentence compared to when the level changed unpredictably from word to word. This benefit of level certainty is clearly shown in Figure 8, where performance for the FL and RSL conditions is compared to performance for the RWL condition (derived from the psychometric functions from Figure 3 and 4). It is apparent from Figure 8 that increasing certainty about target level exerts a strong effect on listener performance with advantages of 0.5 in proportion correct at high relative levels. Also, it is of interest to note that group mean performance improvement was similar for FL and RSL conditions for relative levels of −3 dB and above, with only a slight FL benefit. At −6 and −9 dB, however, the benefit of the increased certainty associated with the FL condition was significantly greater than for the RSL condition. This implies that the ability to focus on the softer target varies substantially across subjects (as indicated by the error bars) but can be improved when the target level is stable throughout a block of trials.

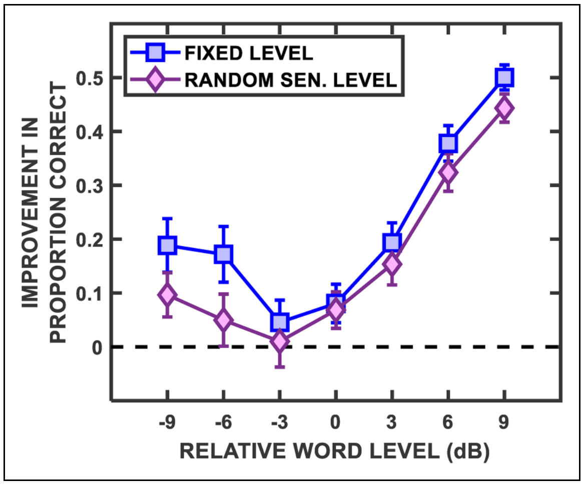

Figure 8. The improvement in proportion correct for the Fixed Level and Random Sentence Level conditions relative to the Random Word Level condition is plotted as a function of the relative level of the target speech. The values are the group mean differences in proportion correct and SEM error bars.

Although performance was still above chance for the RWL condition, in general, (from Figures 3 and 4), it was substantially poorer than for either of the stable speech level conditions, suggesting that listeners benefitted significantly from a known and constant level even when the consistency was only within a trial (e.g., the similarity of results between FL and RSL conditions). This highlights the importance of speech streams conforming to expectations over time for target "stream maintenance" under competing speech (see Kidd et al., 2014, for analogous evidence of the benefit of conforming to a known syntax).

For the stable sentence level conditions, tuning to level and some effect of level uncertainty also were apparent. The ability of (at least some) listeners to obtain an improvement in target speech identification for negative TMRs (cf., Brungart, 2001), suggests that *a priori* knowledge of level can foster attentional focus on a speech stream to reduce the IM that is present, a possible level cueing effect similar to the spatial cueing reported by Kidd et al. (2005). This is consistent with previous research (e.g., Brungart, 2001; Brungart et al., 2001; Dirks & Bower, 1969; MacPherson & Akeroyd, 2014) that has suggested that, under high-IM situations, having a level difference between multiple talkers can improve a listener's ability to segregate and attend to a target, even if that level difference creates greater EM of the target. In addition, masker confusions also suggested tuning to level as more incorrect responses were consistent with the masker talker that was closer in level to the target cue, not simply which talker was louder and likely more salient.

Varying the level of two symmetrically placed maskers around a center target level also produced some evidence for tuning and a level release from masking, but only at the widest level range (9 dB separation between talkers) and only for the conditions with sentence-level certainty about target level (Figure 5). This evidence nonetheless was consistent with the view that listeners may concentrate processing resources at a point along the level continuum, resulting in better performance at the point of focus. This was apparent for both FL and RSL conditions where there was a high degree of target level certainty within trials, but was not apparent for the RWL condition where there was little certainty about target word levels following the cue. The finding that there is some minimum separation in level needed to yield an enhancement in identification of a stream of target speech flanked by competing speech streams is reminiscent of past findings for streams of speech that are separated in spatial location rather than in level (e.g., Marrone et al., 2008; Srinivasan et al., 2016). Much like the benefit of binaural cues that occur from certain separations in spatial location of speech streams, the overall level differences in the present experiment were sufficient to reduce the IM caused by the masker speech and

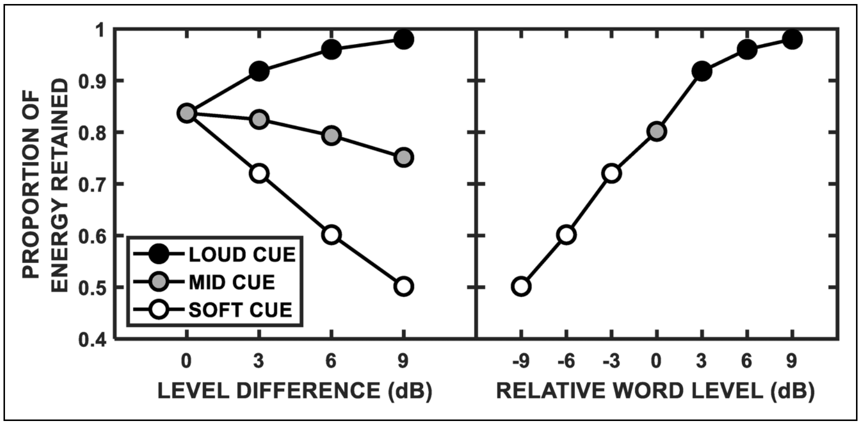

Figure 9. The average proportion of target energy retained following the glimpsed speech analysis plotted as a function of the level difference between speech streams (left panel), as well as the relative level of the target word (right panel). Different symbols represent the different cue levels.

enhance the perceptual segregation of sources, allowing better focus of the listener's attention on the target source. This effect may also be related to a reduction of the multimasker penalty (Iyer et al., 2010), as the soft masker may no longer be fully extracted from the speech mixture, or be more easily ignored, with the largest separation of levels, thus allowing performance to improve to that which might occur with only the loud masker present.

In addition, when the data from the RWL condition were sorted by level relative to the target cue (based on the premise that the listener would focus attention on the level of the cue word and maintain it at that point for the remainder of the sentence), a tuned response was found (Figure 6). This too was taken as evidence for attentional tuning along the level dimension and suggested that the listener focused on the cue word's level, enhancing identification performance for subsequent words near to the cue level, while effectively attenuating off-axis target words. Thus, a tuned response was found only when performance was calculated with respect to a marker in level (the cue) to which they were obliged to attend in order to solve the task. Given how generally predictable and relatively constant the overall level of speech tends to be in normal conversation (Byrne et al., 1994), it may be very difficult for listeners to suppress level as a contributing factor in the focus of attention, even when it is unreliable. It should be noted that, although word by word level variation is an inherent component of prosody in natural speech, the RWL condition violates our expectations about normal prosodic relationships. Even in the FL and RSL conditions, the artificially constructed sentences, with equal overall levels for each word, produces unnatural prosody. Co-articulated sentences, however, have been shown to only produce a small improvement in speech recognition, compared to concatenated sentences (Jett et al., 2021).

The present pattern of results, which is consistent with attention-based tuning in level, fits within a substantial body of work in the auditory domain revealing similar effects for a variety of stimuli and tasks. Among the first and most influential of these studies was Greenberg and Larkin (1968), who manipulated attention along the frequency dimension by varying the probability of occurrence of "probe" tones masked by noise. In the probe-signal experiment, detectability was higher at the most likely frequency of a pure-tone signal, which presumably was where the listener chose to focus attention, and decreased above and below that frequency for less-likely signals resulting in a bandpass filter-like response. Other studies using adaptations of the probe-signal method also have revealed selective responses for a variety of stimuli and tasks including amplitude modulation detection (Wright & Dai, 1998), detection of a tone of uncertain duration (Wright & Dai, 1994), spatial separation of sources for frequency sweep discrimination (Arbogast & Kidd, 2000), among others. In the present context, tuning to level, or "attentional bands" in the amplitude domain, have been previously studied psychophysically (Luce & Green, 1978) as well as physiologically, with evidence of neurons tuned to level in the marmoset primary auditory cortex (Watkins & Barbour, 2011).

The implementation of a remote testing procedure for the present study, with different hardware used by each listener, should not have affected the results. If unintended level compression did occur, it would have likely reduced the magnitude of the effects reported, not affected overall trends across conditions. The 30-dB range presented during the calibration procedure, and certainly the smaller 18-dB range of the experimental stimuli itself, should have been reasonably produced by personal computer/headphone setups. There were also no individual listener results that would lead one to believe that accurate (relative) levels were not being presented.

Summary

1) Varying the degree of certainty about the target level had a significant effect on speech identification performance. This effect was confined to negative relative target levels for sentence-level variation, but was present (and much larger) at both positive and negative relative levels for word-to-word level variation *within* sentences.

2) Higher certainty about target sentence level *across* trials enhanced the ability of listeners to segregate the target stream when it was at negative relative levels, compared to trial-by-trial variation in target sentence level (although both were superior to word-by-word level variation). This finding indicates that focusing attention on the lower-level speech source in a mixture requires a high degree of *a priori* knowledge and, based on the large inter-subject differences observed, varies significantly across listeners.

3) Evidence for a tuned response in level, and a subsequent level release from IM with a sufficient separation in level, was found in multiple analyses and depended crucially on both *a priori* knowledge provided to the listener and the context in which variability occurred (e.g., across sentences or words, or relative to the level of the cue). This evidence comprised both differences across conditions in speech identification performance, masker confusions, and a pattern of results similar to a tuning curve for the within-sentence variability condition where uncertainty was the greatest.

4) In conclusion, the predictability of speech, whether that be syntactic structure, logical semantic meaning, or the spectral and spatial properties of the speech sounds themselves, assists with our ability to understand one talker among many. The intensity of a person's voice, and consistency of that level, is important in more ways than simply overcoming the EM of a noisy environment. The relative talker level, and spacing between talker levels, must also be considered.

Acknowledgments

The authors would like to thank Dr. Torsten Dau (the Associate Editor), and three anonymous reviewers for their helpful comments on earlier versions of this manuscript.

Data Accessibility Statement

The data described in this manuscript will be made available upon reasonable request.

Declaration of Conflicting Interests

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

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by National Institutes of Health, National Institute on Deafness and Other Communication Disorders (R01 DC004545).

ORCID iD

Andrew J. Byrne https://orcid.org/0000-0001-9677-6927

Note

1. Additional data (not included here) collected on the first author using the lab setup [an Industrial Acoustics Company (North Aurora, IL) double-walled sound-attenuating chamber, RME (Haimhausen, Germany) Digiface USB sound card, and Sennheiser (Wedemark, Germany) HD280 Pro headphones] indicated that there were no noteworthy differences between remote and in-lab results, both in terms of overall speech identification performance and trends across conditions.

References

Arbogast T. L., & Kidd G. Jr. (2000). Evidence for spatial tuning in informational masking using the probe-signal method. *Journal of the Acoustical Society of America*, *108*(4), 1803–1810. https://doi.org/10.1121/1.1289366

Best V., Mason C. R., Swaminathan J., Roverud E., & Kidd G. Jr. (2017). Use of a glimpsing model to understand the performance of listeners with and without hearing loss in spatialized speech mixtures. *Journal of the Acoustical Society of America*, *141*(1), 81–91. https://doi.org/10.1121/1.4973620

Best V., Thompson E. R., Mason C. R., & Kidd G. Jr. (2013a). An energetic limit on spatial release from masking. *J. Assoc. Res. Otolaryn*, *14*, 603–610. https://doi.org/10.1007%2Fs10162-013-0392-1

Best V., Thompson E. R., Mason C. R., & Kidd G. Jr.. (2013b) Spatial release from masking as a function of the temporal overlap of competing maskers. *Journal of the Acoustical Society of America*, *133*(6), 3677–3680. https://doi.org/10.1121/1.4803517

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., Chang P. S., Simpson B. D., & Wang D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *Journal of the Acoustical Society of America*, *120*(6), 4007–4018. https://doi.org/10.1121/1.2363929

Brungart D. S., Simpson B. D., Ericson M. A., & Scott K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *Journal of the Acoustical Society of America*, *110*(5), 2527–2538. https://doi.org/10.1121/1.1408946

Byrne D., Dillon H., Tran K., Arlinger S., Wilbraham K., Cox R., Hagerman B., Hetu R., Kei J., Lui C., Kiessling J., Kotby M. N., Nasser N. H. A., El Kholy W. A. H., Nakanishi Y., Oyer H., Powell R., Stephens D., & Meredith R., … ,& Ludvigsen C. (1994). An international comparison of long–term average speech spectra. *Journal of the Acoustical Society of America*, *96*(4), 2108–2120. https://doi.org/10.1121/1.410152

Calandruccio L., Wasiuk P. A., Buss E., Leibold L. J., Kong J., Holmes A., & Oleson J. (2019). The effect of target/masker fundamental frequency contour similarity on masked-speech recognition. *Journal of the Acoustical Society of America*, *146*(2), 1065–1076. https://doi.org/10.1121/1.5121314

Conroy C., Best V., Jennings T. R., & Kidd G. Jr. (2020). The importance of processing resolution in 'ideal time-frequency segregation' of masked speech and the implications for predicting speech intelligibility. *Journal of the Acoustical Society of America*, *147*(3), 1648–1660. https://doi.org/10.1121/10.0000893

Cooke M. (2006). A glimpsing model of speech perception in noise. *Journal of the Acoustical Society of America*, *119*(3), 1562–1573. https://doi.org/10.1121/1.2166600

Dirks D. D., & Bower D. (1969). Masking effects of speech competing messages. *Journal of Speech and Hearing Research*, *12*(2), 229–245. https://doi.org/10.1044/jshr.1202.229

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*, 175–191. https://doi.org/10.3758/bf03193146

Gardner B., & Martin K. (1994). HRTF measurements of a KEMAR dummy-head microphone. MIT Media Labs. Retrieved from January 3, 2020, from https://sound.media.mit.edu/resources/KEMAR.html

Greenberg G. Z., & Larkin W. D. (1968). Frequency–response characteristic of auditory observers detecting signals of a single frequency in noise: The probe–signal method. *Journal of the Acoustical Society of America*, *44*(6), 1513–1523. https://doi.org/10.1121/1.1911290

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

Hazan V., & Markham D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. *Journal of the Acoustical Society of America*, *116*(5), 3108–3118. https://doi.org/10.1121/1.1806826

Helfer K. S., Poissant S. F., & Merchant G. R. (2020). Word identification with temporally interleaved competing sounds by younger and older adult listeners. *Ear and Hearing*, *41*(3), 603–614. https://doi.org/10.1097/AUD.0000000000000786

Holmes E., To G., & Johnsrude I. S. (2021). How long does it take for a voice to become familiar? Speech intelligibility and voice recognition are differentially sensitive to voice training. *Psychological Science*, *32*(6), 903–915. https://doi.org/10.1177/0956797621991137

Iyer N., Brungart D. S., & Simpson B. D. (2010). Effects of target-masker contextual similarity on the multimasker penalty in a three talker diotic listening task. *Journal of the Acoustical Society of America*, *128*(5), 2998–3010. https://doi.org/10.1121/1.3479547

Jett B., Buss E., Best V., Oleson J., & Calandruccio L. (2021). Does sentence-level coarticulation affect speech recognition in noise or a speech masker? *Journal of Speech, Language, and Hearing Research*, *64*(4), 1390–1403. https://doi.org/10.1044/2021_JSLHR-20-00450

Kidd G., Arbogast T. L., Mason C. R., & Gallun F. J. (2005). The advantage of knowing where to listen. *Journal of the Acoustical Society of America*, *118*(6), 3804–3815. https://doi.org/10.1121/1.2109187

Kidd G. Jr., Best V., & Mason C. R. (2008). Listening to every other word: Examining the strength of linkage variables in forming streams of speech. *Journal of the Acoustical Society of America*, *124*(6), 3793–3802. https://doi.org/10.1121/1.2998980

Kidd G. Jr., & Colburn H. S. (2017). Informational masking in speech recognition. In J. C. Middlebrooks, J. Z. Simon, A. N. Popper, & R. R. Fay (Eds.), *The auditory system at the cocktail party* (pp. 75–109). Springer Nature.

Kidd G. Jr., Jennings T. R., & Byrne A. J. (2020). Enhancing the perceptual segregation and localization of sound sources with a triple beamformer. *Journal of the Acoustical Society of America*, *148*(6), 3598–3611. https://doi.org/10.1121/10.0002779

Kidd G. Jr., Mason C. R., & Best V. (2014). The role of syntax in maintaining the integrity of streams of speech. *Journal of the Acoustical Society of America*, *135*(2), 766–777. https://doi.org/10.1121/1.4861354

Kidd G. Jr., Mason C. R., Best V., Swaminathan J., Roverud E., & Clayton K. (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

Kim S. K., & Sumner M. (2017). Beyond lexical meaning: The effect of emotional prosody on spoken word recognition. *Journal of the Acoustical Society of America*, *142*(1), EL49–EL55. https://doi.org/10.1121/1.4991328

Luce R. D., & Green D. M. (1978). Two tests of a neural attention hypothesis for auditory psychophysics. *Percep. Psychophys*, *23*, 363–371. https://doi.org/10.3758/BF03204138

Lutfi R. A., Rodriguez B., & Lee J. (2021). The listener effect in multitalker speech segregation and talker identification. *Trends in Hearing*, *25*, 1–11. https://doi.org/10.1177/23312165211051886.

MacPherson A., & Akeroyd M. A. (2014). Variations in the slope of the psychometric functions for speech intelligibility: A systematic survey. *Trends in Hearing*, *18*, 1–26. https://doi.org/10.1177/2331216514537722

Marrone N., Mason C. R., & Kidd G. (2008). Tuning in the spatial dimension: Evidence from a masked speech identification task. *Journal of the Acoustical Society of America*, *124*(2), 1146–1158. https://doi.org/10.1121/1.2945710

Mattys S. L., Davis M. H., Bradlow A. R., & Scott S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*, 953–978. https://doi.org/10.1080/01690965.2012.705006

Middlebrooks J. C., Simon J. Z., Popper A. N., & Fay R. R. (Eds.). (2017). *The auditory system at the cocktail party*. Springer Nature.

Puschmann S., Baillet S., & Zatorre R. J. (2019). Musicians at the cocktail party: Neural substrates of musical training during selective listening in multispeaker situations. *Cerebral Cortex*, *29*(8), 3253–3265. https://doi.org/10.1093/cercor/bhy193

Rennies J., Best V., Roverud E., & Kidd G. Jr. (2019). Energetic and informational components of speech-on-speech masking in binaural speech intelligibility and perceived listening effort.

Trends in Hearing, *23*, 1–21. https://doi.org/10.1177/2331216519854597

Rodriguez B., Lee J., & Lutfi R. (2021). Additivity of segregation cues in simulated cocktail-party listening. *Journal of the Acoustical Society of America*, *149*(1), 82–86. https://doi.org/10.1121/10.0002991

Srinivasan N. K., Jakien K. M., & Gallun F. J. (2016). Release from masking for small spatial separations: Effects of age and hearing loss. *Journal of the Acoustical Society of America*, *140*(1), EL73–EL78. https://doi.org/10.1121/1.4954386

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

Watkins P. V., & Barbour D. L. (2011). Level-tuned neurons in primary auditory cortex adapt differently to loud versus soft sounds. *Cerebral Cortex*, *21*(1), 178–190. https://doi.org/10.1093/cercor/bhq079

Woodhouse L., Hickson L., & Dodd B. (2009). Review of visual speech perception by hearing and hearing-impaired people: clinical implications. *International Journal of Language and Communication Disorders*, *44*(3), 253–270. https://doi.org/10.1080/13682820802090281

Wright B. A., & Dai H. (1994). Detection of unexpected tones with short and long durations. *Journal of the Acoustical Society of America*, *95*(2), 931–938. https://doi.org/10.1121/1.410010

Wright B. A., & Dai H. (1998). Detection of sinusoidal amplitude modulation at unexpected rates. *Journal of the Acoustical Society of America*, *104*(5), 2991–2996. https://doi.org/10.1121/1.423881

Zekveld A., Rudner M., Kramer S. E., Lyzenga J., & Rönnberg J. (2014). Cognitive processing load during listening is reduced more by decreasing voice similarity than by increasing spatial separation between target and masker speech. *Frontiers in Neuroscience*, *8*, 88. https://doi.org/10.3389/fnins.2014.00088

Appendix

Glimpsed Speech Analysis

In order to compare the empirical trends found in the present results to what would be expected based purely on energetic masking, a glimpsed speech analysis (GSA) similar to "ideal time-frequency segregation" (e.g., Best et al., 2017; Brungart et al., 2006; Kidd et al., 2016) was performed. Simulated target and masker stimuli were broken down into spectrotemporal tiles defined by 128 frequency channels logarithmically spaced between 80 Hz and 8 kHz, and 20-ms time windows with 50% overlap. Based on prior knowledge of the stimulus waveforms, computation of TMR within tiles was performed with a level criterion of 0 dB applied (cf., Conroy et al., 2020). The ratio of the sum of squares of the pre- and post-glimpsing target was used to quantify the output of the model, which was the proportion of glimpsed target energy that was retained in the tiles where the target energy was greater than the masker energy. This GSA was performed for 480 simulated trials, using stimuli and randomization identical to that which was employed in the actual experiment.

The results of these simulations are shown in Figure 9. As expected, the Loud Cue trials improved (in terms of energy retained after glimpsing) as they increased in level relative to the other talkers (i.e., as the level difference became +9 and +18 dB relative to the two masker talkers). Conversely, the amount of energy retained decreased (i.e., EM increased) as the level difference increased for the Soft Cue conditions, thus making the target talker −9 and −18 dB relative to the maskers. If only EM were a factor for speech identification during the present experiment, one would expect the empirical results (Figure 2) to follow a pattern/order similar to the functions of Figure 9 (left panel). Since the results of Figure 2 did not match this pattern (i.e., Soft Cue empirical results were relatively stable, rather than decreasing as in the GSA), whatever degradation that EM had on the Soft Cue trials was offset by the increased salience from attending to only the softest voice that was heard.

The Mid Cue condition results (for FL and RSL) shown in Figure 5 supported the view of tuning in level for SOS masking, because performance improved as the masker speech streams were substantially separated from the target. The interpretation that this apparent tuning effect (and level release) was based on the focus of attention along the level dimension assumed that the improvement in performance at the widest separation was not simply a consequence of decreased EM. In Figure 9 (left panel), the Mid Cue function (shaded circles) slightly decreased indicating more, not less, EM as the level difference was increased, the opposite of the trend seen in the empirical results. Thus, this analysis supports the conclusion that the improved performance at the widest separation of masker speech streams was indeed due to factors other than EM; most likely, attentional tuning in level and improved segration of the mid-level speech stream.

The right panel of Figure 9 re-plots the GSA results as a function of the relative target word level (cf., Figures 3 and 4) and produced a monotonically increasing function, an aspect which was also inconsistent with the significant difference between the degree of uncertainty in the emprical results, as the single (composite) function of Figure 9 would apply to all three uncertainty conditions. The x-axis could also be easily converted to the ΔL from cue level value used for the RWL condition in Figure 6 by simply doubling the scale (i.e., for a Loud Cue trial, a relative word level of −9 dB would be equivalent to a ΔL from cue level of −18). When compared to the concave pattern of results in Figure 6 (with statistically significant non-monotonicity), the GSA function clearly did not match. When target words in the RWL condition fell below the level of the cue, both the available target energy (as inferred from the GSA analysis) and speech identification performance decreased as the level was reduced. Above 0 dB, however, the patterns diverge, with the proportion of target energy increasing to a value near 1.0, while the speech identification performance decreased comparable to that of the negative relative levels. Again, this finding was interpreted as strong evidence that varying degrees of EM was not the explanation for the tuned pattern of performance found in the emprical results.