JASA ARTICLE

Simultaneous relative cue reliance in speech-on-speech masking

R. A. Lutfi,[a] M. Zandona, and J. Lee

Auditory Behavioral Research Lab, Department of Communication Sciences and Disorders, University of South Florida, Tampa, Florida 33620, USA

ABSTRACT:

Modern hearing research has identified the ability of listeners to segregate simultaneous speech streams with a reliance on three major voice cues, fundamental frequency, level, and location. Few of these studies evaluated reliance for these cues presented simultaneously as occurs in nature, and fewer still considered the listeners' relative reliance on these cues owing to the cues' different units of measure. In the present study trial-by-trial analyses were used to isolate the listener's simultaneous reliance on the three voice cues, with the behavior of an ideal observer [Green and Swets (1966). (Wiley, New York), pp.151–178] serving as a comparison standard for evaluating relative reliance. Listeners heard on each trial a pair of randomly selected, simultaneous recordings of naturally spoken sentences. One of the recordings was always from the same talker, a distracter, and the other, with equal probability, was from one of two target talkers differing in the three voice cues. The listener's task was to identify the target talker. Among 33 clinically normal-hearing adults only one relied predominantly on voice level, the remaining were split between voice fundamental frequency and/or location. The results are discussed regarding their implications for the common practice in studies of using target-distracter level as a dependent measure of speech-on-speech masking.

© *2023 Acoustical Society of America*. https://doi.org/10.1121/10.0021874

(Received 20 July 2023; revised 26 September 2023; accepted 27 September 2023; published online 23 October 2023)

[Editor: Emily Buss] Pages: 2530–2538

I. INTRODUCTION

The ability to "hear out" the speech of one talker separately from others speaking at the same time has long fascinated researchers in the field of hearing science because of the remarkable achievement it represents for a sensory system (Cherry, 1953). The pressure wavefronts corresponding to the speech of different talkers superimpose (add together) in the air before reaching our ears. Hence, for us to hear each talker separately, the auditory system must somehow extract those individual wavefronts from the sum. The problem is akin to solving for x, y, and z in the expression $x + y + z = 42$; there is not one, but an indeterminate number of possible solutions.

We now know that the auditory system solves the problem, in part, by taking advantage of predictable constraints governing the speech of talkers. Three of these constraints, widely established by research, are the harmonic structure of the speech, the location of the talker relative to the position of our head, and the relative level of the different speech streams [see Bronkhorst (2000, 2015), Byrne *et al.* (2022), Kidd and Colburn (2017), and Szabo *et al.* (2016) for reviews]. Harmonic structure in speech is created by the periodicity of glottal pulses in the voicing of speech, which gives rise to the sensation of pitch corresponding to the fundamental frequency (F0) of the glottal pulses. Males, whose vocal folds tend to be longer and heavier than females, tend to have lower F0s. Hence, a male and female speaking simultaneously will often be heard separately based on simultaneously heard differences in the pitch of their voice. The speech of a talker also arrives at our two ears differently depending on the location of the talker relative to our head. Differences in the level and time of arrival of the speech between the two ears are cues for the direction of the sound, which can be used to segregate talkers based on the different locations from which they speak on the azimuthal plane. The relative level of the speech of talkers is also a dominant cue, as anyone can testify who has attended a gathering where one loud voice rises above everyone else in the crowd. Relative speech level, in fact, has a special place in studies of speech-on-speech masking because it also regularly serves as a dependent measure (signal-to-noise level at threshold) of the influence of other cues [Bronkhorst (2000), (2015) and Byrne *et al.* (2022); cf. Ahrens *et al.* (2020) and Ozmeral and Higgins (2022)].

Knowledge of the importance of these and other cues comes largely from studies of their effects on the overall performance of listeners in various multi-talker listening scenarios. Here, evidence for listener reliance on a cue is inferred from any observed reduction in performance in conditions where that cue is either distorted, made uncertain, or eliminated. However, such effects are not always easy to interpret as they can vary considerably from one individual to the next. It is not uncommon for overall performance to vary from near chance to near perfect within the same experimental condition [see Lutfi *et al.* (2020) and (2021) for review]. Such differences in performance have raised the question as to whether there are circumstances for which

[a] Email: rlutfi@usf.edu

https://doi.org/10.1121/10.0021874

individuals may attend differently to different cues, affecting their performance in those cases. Some researchers have speculated that this could be a factor responsible for why many individuals, evaluated to have normal hearing, have unusual difficulty understanding speech in noisy environments (Oberfeld and Klöckner-Nowotny, 2016; Dai and Shinn-Cunningham, 2016; Ruggles and Shinn-Cunningham, 2011; Shinn-Cunningham, 2017).

Work is under way to evaluate such speculation, but there are major challenges involved. First, there are limits to what can be concluded regarding reliance on cues from single metrics of performance accuracy when, as in everyday listening, multiple cues are available to the listener simultaneously. In such cases, eliminating or distorting a cue can have no effect on performance either because the listener originally placed no reliance on that cue, or because they simply switched to a different equally viable cue. The response historically to the preoccupation in psychophysics with single metrics of performance accuracy has been molecular psychophysics [Watson (1973), hearing; Ahumada (2002), vision]. Molecular psychophysics recognizes that the averaging of responses required for measures of performance accuracy can conceal distinctive influences of simultaneous cues and their interaction that are only evident in the relation between the stimulus and the listener's response from trial to trial [see Berg (1990), Lutfi (1995), and Calandruccio and Doherty (2007) for contemporary analyses]. These analyses differ in detail, but all are designed to estimate the degree to which individuals *weigh* different features of the stimulus to arrive at a response. They are principally based on the parsing of trial-by-trial responses into categories defined by features of the stimuli, rather than whether the response is correct or incorrect.

The second challenge in evaluating listener reliance on cues is that their effects, by virtue of their different physical units (Hz, dB, azimuthal angle), are not directly comparable. This means that without some standard for equating the relative information provided by each cue one can never be confident that the values chosen did not bias listeners to favor one cue over another. This is of foremost concern where the goal is to evaluate individual differences in the listener's reliance on cues. The reaction historically to this kind of problem has been signal-detection theory (SDT) (Green and Swets, 1966). SDT permits comparisons of the effects of stimulus manipulations involving different physical units by expressing listener performance in each case relative to that of an ideal observer, an observer who bases decisions on the likelihood ratio of signal-to-noise. It serves to identify the relative information provided by each cue so that a listener's true preference for cues can be distinguished from a forced reliance based on an arbitrary selection of stimulus values.

The present study combined molecular analyses with elements of SDT to evaluate the *simultaneous*, *relative* reliance listeners placed on voice fundamental frequency, location, and level cues for a speech-on-speech masking task. Goals were to (1) document the simultaneous relative reliance on these cues for a group of clinically normal-hearing adults, (2) identify any individual differences among listeners in the relative reliance placed on these cues, and (3) determine the impact of those differences, if any, on overall performance accuracy.

II. GENERAL METHODS

Listeners in the study performed a talker identification task in which the stimuli were recordings of spoken sentences processed to differ for talkers in voice F0, level, and location, all available simultaneously to the listener. The three cues were perturbed slightly from trial to trial to simulate small changes in voice F0, level, and location that occur naturally from one moment to the next. The perturbations also served two important methodological functions. First, they leveled the playing field for the three cues by making them equally viable for the task. This was done by selecting the variances σ^2 of the perturbations to equate the normalized differences in the mean values Δ of the cues (that is to equate Δ/σ for the cues). In signal-detection theory (SDT) this is equivalent to equating conditions for the performance of an ideal observer (Green and Swets, 1966). Second, the perturbations served as predictor variables in a discriminant analysis of the listener's trial-by-trial response, wherein the regression coefficients of the analysis served as estimates of the relative reliance (decision weight) listeners placed on the three cues.

A. Procedure

A single-interval, two-talker identification task was used [see Lutfi *et al*. (2020)]. Listeners were read the following instructions at the beginning of the experiment: "There will be a series of trials in which you will listen over headphones to two talkers speaking sentences. One of the two talkers will always be Pat, the other talker speaking at the same time is equally likely to be Jon or Jen. Jon has a soft, low-pitch voice and is located on your left. Jen has a loud, high-pitch voice and is located on your right. Pat has a voice with pitch and loudness intermediate between Jon and Jen and is located center/front. Your task on each trial is to ignore Pat and identify by button press whether you heard Jon or Jen."

The listeners then underwent a three-step voice familiarization routine. In step 1, they listened passively to a block of 20 trials in which Jon and Jen alternately spoke a random sentence. The name of the talker appeared on the listener's monitor as the sentence was being played. In step 2, the 20 trials were repeated in random order and the listener identified on each trial which of the two talkers was speaking. Listeners were given feedback after each response. Finally, in step 3, the listeners repeated the task they were given in step 2, but this time the to-be-ignored Pat also spoke a random sentence on each trial. Listeners had little difficulty with the familiarization tasks as the difference in loudness, pitch, and location of talkers were fixed from trial to trial and were quite easy to discern (see stimuli below).

Once familiarized with the talkers' voices, the listeners proceeded to experimental trials. On experimental trials, the listeners were told that this time "the loudness, pitch and location of the talkers' voices will vary a little from trial to trial, as happens from moment to moment in natural speaking. For example, Jon on any given trial might raise the pitch of his voice to be closer to that of Jen. Similarly, Jen on any given trial might change her position to be closer to that of Jon. On these trials you might confuse one with the other talker and make an error, but this is ok as errors are expected. Use the three cues (loudness, pitch and location) individually or in any combination to identify Jon and Jen; whatever works best for you." Listeners were given correct feedback after each trial and an overall percent correct score at the end of each trial block. A sign was posted in the booth to remind listeners of the nominal differences in the Jon and Jen's voices.

The data were collected in 10 blocks of 50 trials per block within a one-hour session per day. Listeners were allowed breaks at their discretion between trial blocks. At least one week after collecting the data for the first session, the listeners returned to repeat the entire procedure, both familiarization routine and experimental trials, for different stimulus values that allowed comparisons of the decision weights for two different levels of performance (see Sec. II B below). Listeners also repeated the entire procedure for three additional conditions designed to measure relative sensitivity for each cue. These conditions were the same as the previous all-cue conditions except only one cue distinguished the target talkers, the mean difference Δ for the remaining two cues were zeroed. Before each of these conditions the listener was instructed as to which cue distinguished the target talkers.

B. Stimuli

The stimuli were recordings of naturally spoken, grammatically correct, English sentences selected at random on each trial for each talker from 200 neutral exemplars (folder 0019) of the Emotional Speech Dataset (Zhou *et al.*, 2021). The sentences were selected without replacement so that the two talkers never spoke the same sentence within a trial, the same sentence may, however, been spoken more than once within the session. The duration of the sentences ranged from 1.6 to 2.5 s and had slightly different starting times; because of this the sentences did not exactly temporally align; the two talkers on each trial had an equal probability of beginning first and/or ending last. This, of course, happens in natural listening and so no attempt was made to synchronize the onsets and offsets of the sentences.

The sentence exemplars were from a single native speaker of English whose voice fundamental frequency was roughly midway between that of an average male and female voice. This talker was identified as Pat. The original sentences associated with Pat were then processed to produce the sentences of Jon and Jen. This ensured that only differences in the level (L), fundamental frequency (F0), and location (azimuthal angle θ) of the talkers would serve as viable identification cues. For Pat the nominal values of the parameters were, respectively, $\mathrm{L} = 62$ dB SPL (calibrated at the headphones), F0 of the original sentences, and $\theta = 0°$. For Jon and Jen, the nominal values were arithmetically centered about the values for Pat, differing from each other, respectively, by $\Delta = 8$ dB, 28 Hz, and 28°. These values were selected based on previous experience targeting performance within a range of 75%–95% correct. In a second condition the values were reduced to $\Delta = 6$ dB, 21 Hz, and 21° to target performance within a range of 65%–85% correct. All values of Δ were chosen to be clearly discriminable from one another in the absence of perturbations or masking (Jesteadt *et al.*, 1977; Houtsma and Smurzynszki, 1990, Perrot and Saberi, 1990). For all talkers, the trial-by-trial jitter added to these parameters was linearly normally distributed with standard deviation $\sigma = 2$ dB, 7 Hz, and 7°, respectively. (Extreme values were avoided by resampling anything 2.5 standard deviations above or below the mean.) Values of σ were selected to represent normal variation in natural speech [see Horii (1975) for the distributional statistics of F0], but also to make the three cues equally diagnostic for the task, that is to equate $d'_{\mathrm{ideal}} = \Delta/\sigma$ for each cue. The optimal listening strategy in this case is to place equal reliance on all three cues.

As in past studies [e.g., Lutfi *et al.* (2020) and Lutfi *et al.* (2021)], the trial-by-trial values of θ were achieved through filtering using Knowles Electronics Manikin for Acoustic Research (KEMAR) head-related, transfer functions (HRTFs) (Gardner and Martin, 1995).[1] Perturbations in θ were interpolated within a 5° resolution for the HRTFs [see Wightman and Kistler (1989) for details]. The trial-by-trial values of F0 were achieved using the overlap and add method to maintain the original duration of the sentences (Hejna and Musicus, 1991). The method was implemented by the function "solaf" on the MATLAB exchange. All sentences were played at a 44 100-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.

C. Listeners

Thirty-three students at the University of South Florida–Tampa, 18 male and 15 female, ages 19–38 years, participated as listeners in the study. They were reimbursed with gift cards for their participation. All had normal hearing as determined by standard audiometric evaluation, which included pure-tone audiometry and tympanometry. An initial eight of these listeners, 2 males, 6 females, ages 19–24, participated in all conditions of the experiment. In a follow-up, the remaining 25 listeners participated in just the all-cue condition. Informed consent was obtained from all listeners and all procedures were followed in accordance with University of South Florida internal review board (IRB) approval.

D. Estimates of decision weights

The principle method and justification for estimating the listener decision weights has been described in detail in previous papers and is only briefly reviewed here [see Lutfi *et al*. (2020) and Berg (1990)]. Fundamentally it is a form of multiple regression where the trial-by-trial values of cues are the predictor variables, the listener's trial-by-trial response is the predicted variable, and the normalized regression coefficients are the reliance relative weights. Let **f** denote the 1-by-3 vector representing the z-score sums for the two talkers in L, F0, and θ on any trial.[2] The listener decision weights **w** on these parameters are determined from the coefficients **c** in a logistic regression of the listener's response R across trials contingent on **f**,

$$\text{logit}[\text{P}(\text{R} \equiv \text{Jen})] = c_0 + \mathbf{c} \cdot \mathbf{f}' + err, \tag{1}$$

where c_0 captures the listener's bias to respond Jen, *err* is the residual error associated with the regression, and

$$\mathbf{w} = \mathbf{c} / \sum |\mathbf{c}| \tag{2}$$

are the estimates of the relative weights. Note that the regression, in this case, depends on who the listener identifies with **f**, Jon, or Jen, not whether the response is correct or incorrect. This is important for two reasons. First, the perturbation in cues, although quite rarely for the parameters selected, can cause the listener to make an error even if their weights are optimal for the task. This happens in natural listening because the values of voice parameters of talkers will tend to overlap over the course of many different utterances. Second, and more importantly, our goal is to determine what the listener judges the differences to be between Jon and Jen, not what we have determined those differences to be based on selection of stimulus values. This, as pointed out earlier, is a fundamental difference between the present analysis of cue reliance and those based on metrics of performance accuracy. Finally, it is worth noting that the **w** in some studies are identified with the *attentional* weights given to cues. We attach no such perceptual significance to **w** in this application, recognizing that many factors can influence the estimates of **w**.[3] For this reason, we use the more neutral term "reliance" for **w** taking it merely as an estimate the relative strength of relation of the listener's response to the different cues.

III. RESULTS

A. Listener decision weights

We consider first the decision weights, the effect of the decision weights on listener performance is then presented in Sec. III B. Figure 1 gives the relative weights **w** for the initial eight listeners participating in all conditions of the study. The panels are ternary plots where each of the eight different symbols represents for each of the eight listeners the relative weight on the three stimulus cues. Symbols that converge on the upmost corner of this plot represent predominant weight on talker location (θ), those that converge on the lower right corner predominant weight on relative speech level (L), and those that converge on the lower left corner predominant weight on talker fundamental frequency (F0). The + symbol identifies the center-point of the plots corresponding to equal weight on the three cues, the ideal weights for this task. The left and right plots are for the two conditions where Δ was selected to target performance in the range of 75%–95% and 65%–85% correct, respectively. The corresponding obtained percent correct scores across listeners ranged from 77%–92% and 72%–85%, respectively. The agreement between the two estimates of the decision weights for the two levels of performance is gauged by comparing symbols across plots. An analysis of variance using just these two estimates to measure the within-listener sums of squares revealed listener differences in the relative weights for each cue to be significant at the $\alpha = 0.025$ level ($F_{1,7} = 4799$, 1290, and 2720 for F0, L, and θ, respectively). The result is consistent with other studies

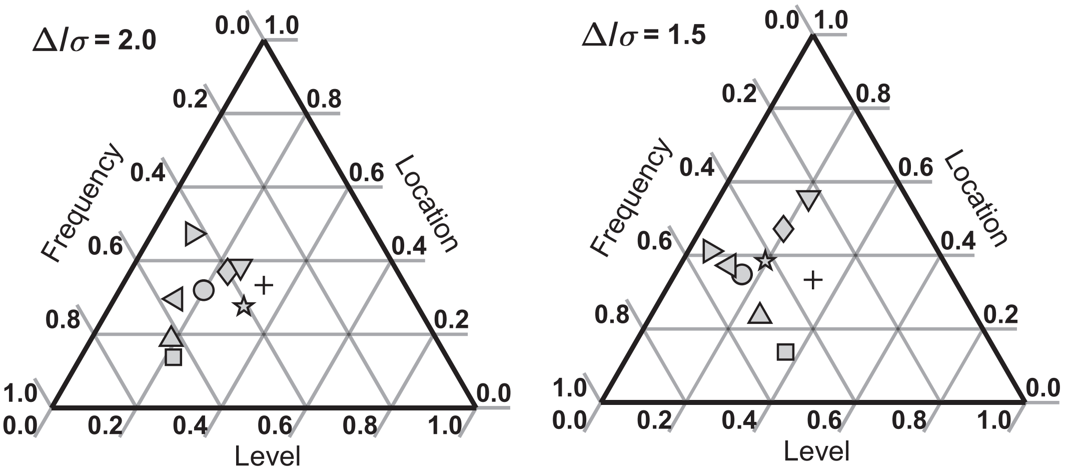

FIG. 1. Ternary plots showing the estimates of relative decision weights w on fundamental frequency, Fo, location, θ, and level, L, for the first group of eight listeners, different symbols identifying the different listeners. The + symbol denotes the optimal relative decision weights for the task. Left and right panels are for the conditions targeting two different levels of performance.

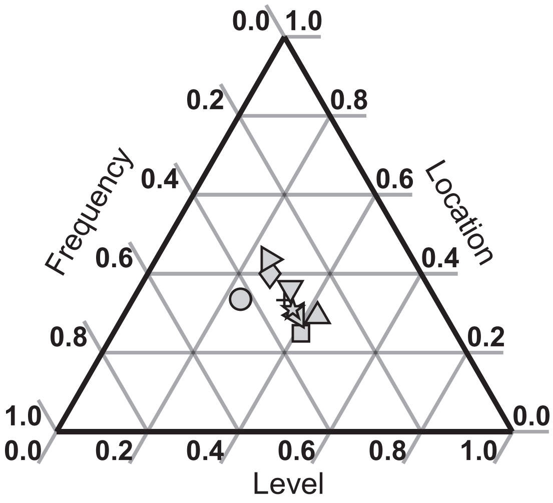

FIG. 2. Ternary plot giving relative sensitivity d' for the three cues for the first group of listeners.

showing reliable individual differences in the estimates of decision weights across different performance levels and on different days (Doherty and Lutfi, 1996; Berg, 1990; Lutfi and Liu, 2007).

Comparing listeners, five give predominant weight to talker F0, while the remaining three give equal or predominant weight to talker location. Talker speech level by comparison is largely ignored. The data from the single cue conditions make clear that this outcome was not due to a failure of listeners to hear the differences in talker speech level. Figure 2 shows in a ternary plot equal relative sensitivity d' of listeners for detecting the differences in the three cues when only one of the three distinguished talkers. Figure 3, moreover, shows that listeners could reliably switch their reliance on cues to give predominant weight to the single cue that distinguished talkers. Single cue weights can be obtained in this case because the perturbation in non-informative cues continues to be present across trials and enters into the regression model of Eq. (1).

Taken together, the data of Figs. 1–3 indicate that, for the equally diagnostic simultaneous cues available for this masking task, listeners show a predominant reliance on the F0 and/or location cues at the expense of relative speech level. This is a new result, notable considering the common practice of using relative speech level as a dependent measure to assess the reliance listeners place on other stimulus cues, most commonly F0 and location (Bronkhorst, 2000,

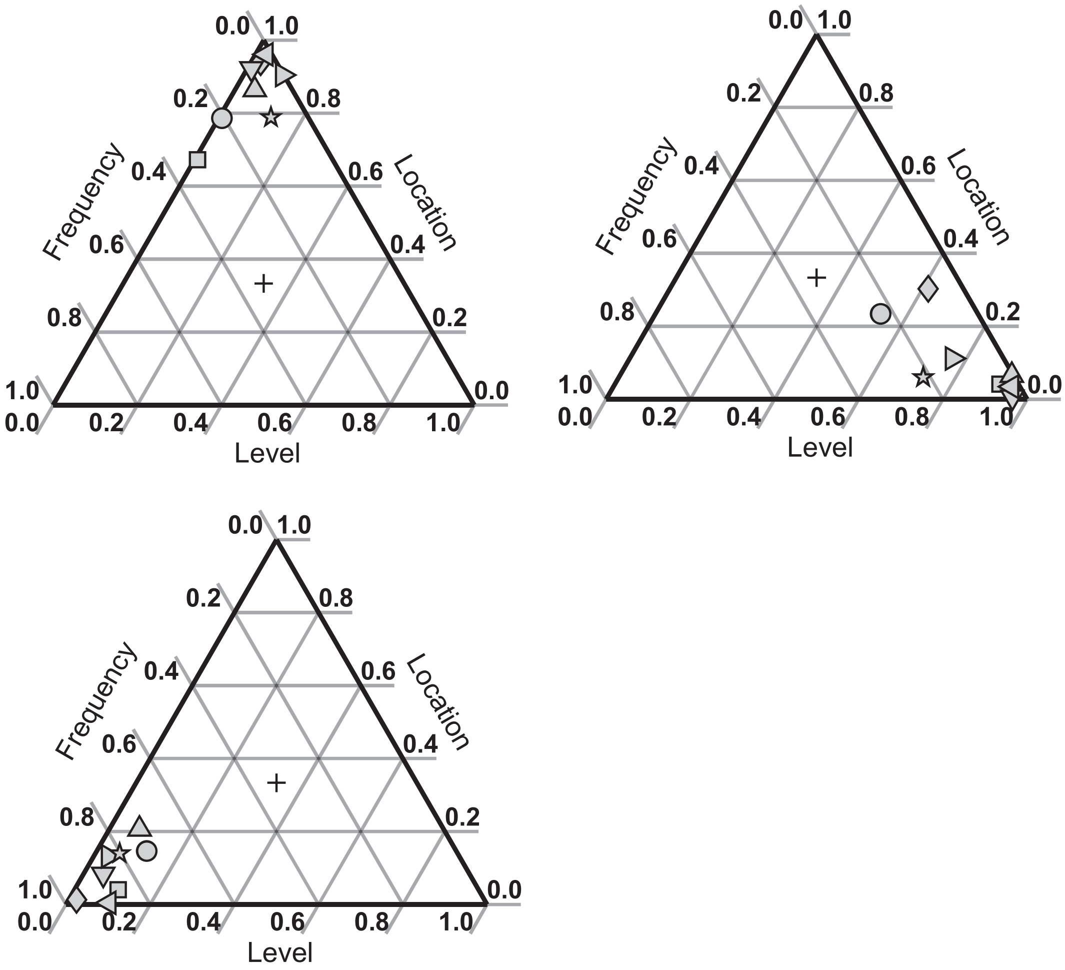

FIG. 3. Ternary plots as before except now giving the relative decision weights for the first group of listeners in the single cue conditions (panels). Reading clockwise from upper left panel, single cues are location, θ, level, L, and fundamental frequency, F0.

2015; Kidd and Colburn, 2017). To evaluate the generality of the result, we recruited an additional 25 USF students (see Sec. II C) to participate as listeners in the all-cue condition, either the higher or lower level performance condition. The data from those listeners together with the data from the original eight (for both values of Δ/σ) are shown in Fig. 4. The data confirm a general listener preference for the F0 and/or location cues, with only one of the 33 listeners giving predominant weight to the level cue.

B. Effect of decision weights on performance

Last, we consider the effect of the decision weights on overall performance in the all-cue condition. Recall that the optimal listening strategy for the all-cue condition is to give equal weight to the three cues. In the absence of any other limiting factors, the optimal strategy yields a prediction of essentially perfect performance for the task.[4] Few listeners approached the optimal listening strategy, as evident from Fig. 4, so we expect the less than optimal decision weights to have some adverse effect on performance. To estimate that effect, we computed for each listener from their decision weights in the all-cue condition what their performance PC_{wgt} would have been if the decision weights were the only factor limiting their performance. In practice this was achieved by "reverse engineering" of the regression Eq. (1) without the error term *err*. We substitute the obtained values of **c** in Eq. (1) for each listener to make a prediction for each listener's trial-by-trial response to the stimuli, and then calculate the percentage of those predicted responses that were correct to arrive at PC_{wgt}. The results of this analysis are shown in Fig. 5, where obtained overall performance PC_{obt} is plotted against the predicted performance PC_{wgt} based on the individual estimates of the decision weights. As expected, the predicted values of PC_{wgt} for all listeners is less than perfect, corresponding to a roughly constant reduction in performance of 8% points, given by the vertical dashed line in the figure. However, Fig. 5 also shows the obtained percent correct scores to vary considerably across the 33 listeners, PC_{obt} ranging from 67% to 97%. If the individual decision weights were responsible for this variability, we should expect the PC_{wgt} values to extend across the same range with symbols falling along the diagonal. Instead, the PC_{wgt} values are distributed almost uniformly from 84% to 95% centered about the dashed line corresponding to a constant PC_{wgt}. So, while Fig. 4 shows that listeners rely differently on the three cues, the present analysis suggests that these differences contribute little to the individual difference in overall performance.

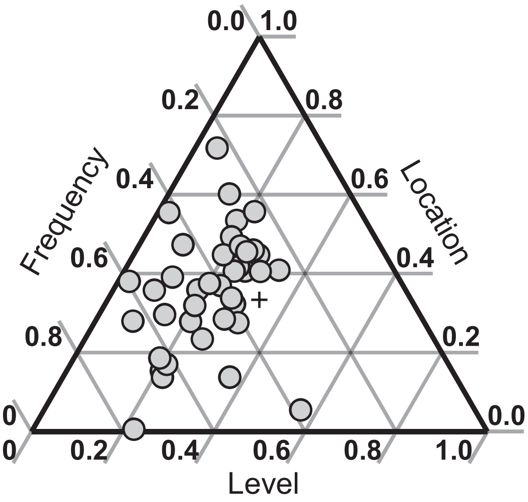

FIG. 4. Ternary plots as before giving the relative decision weights for all listeners participating in the all-cue condition.

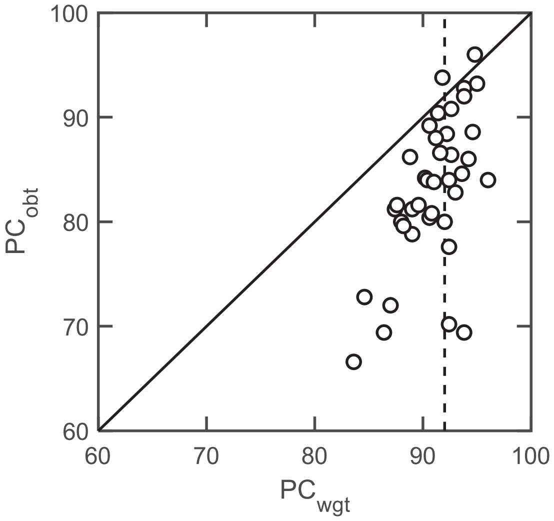

FIG. 5. Percent correct obtained PC_{obt} vs percent correct predicted PC_{wgt} based on the individual estimates of the listeners' decision weights, all listeners. Dashed line indicates a fixed effect of decision weights on performance across listeners.

This result has been reported in two other studies where the tasks were talker search using single recordings of words (Lutfi *et al.*, 2022) and talker detection using synthesized vowels (Lutfi *et al.*, 2020). The present study now documents the result for talker identification task involving recordings of naturally spoken sentences. The result has implications for classes of models intended to account for the wide variation in listener performance commonly observed in studies of speech-on-speech masking. It suggests that variation is due to largely stimulus-independent, stochastic processes (internal noise) that cause information loss at different stages of auditory processing, as opposed to stimulus-dependent failures of selective attention. The topic is beyond the scope of this work, but the interested reader is referred to the cited papers where the topic is discussed at length.

IV. DISCUSSION

The trial-by-trial analyses of this study showed that, for conditions in which multiple stimulus cues were presented simultaneously and were equated in sensitivity for an ideal observer ($d'_{ideal} = \Delta/\sigma$), listeners mostly relied on F0 and/or location to segregate talkers at the expense of relative speech level. Listeners did this despite a demonstrated ability to switch reliance to the level cue and a potential benefit to do so. As in past studies, there were statistically significant differences among listeners in the relative reliance placed on cues, but these differences could not account for the large individual differences in performance observed (Lutfi *et al.*, 2020; Lutfi *et al.*, 2022).

The preferential reliance on F0 and location cues observed here is a new result and is somewhat unexpected given the data also indicate that listeners were equally sensitive to changes in all three cues. We consider several possible contributing factors, none entirely conclusive. The first has to do with possible differences in the perceived magnitude of the values of Δ chosen for the three cues. We refer to this as saliency to avoid confusion with sensitivity, which again was the same for listeners for the three cues. The speculation is that the most salient cue, the one that "stands out" above the rest, is given greatest weight. One way to test this idea is to quantify the saliency of the Δ for each cue as the sum of just-noticeable differences (jnds) in Δ, this approach has early historical precedence (Fechner, 1860). The jnd for a change in level is 1 dB (Jesteadt *et al.*, 1977), for the frequencies tested approximately 3–5 Hz (Wier *et al.*, 1977), and for the change in azimuthal location about 1° (Perrot and Saberi, 1990). Hence, the $\Delta = 8$ dB difference between Jon and Jen corresponds to an 8 jnd perceived difference, the 28 Hz differences in F0 corresponds to a 6–9 jnd perceived difference, and the 28° different in location corresponds to a 28 jnd perceived difference. This might explain the preferential reliance on location except that jnd estimates of stimulus magnitude are indirect estimates that diverge widely from what listeners in practice report as perceived differences in stimulus magnitude (Stevens, 1961). From direct estimates, Jen would be heard as roughly twice as loud as Jon, but only a fraction higher in pitch (Parker and Schneider, 1974). One thus arrives at different conclusions regarding the role of saliency depending on which metric is chosen to measure it. Another factor to consider is the impact the level differences might have had on the audibility of cues due to energetic masking (Brungart, 2001; Kidd *et al.*, 2016). While some amount of energetic masking is unavoidable in these conditions, the fact that talkers always spoke different sentences separated both spatially and by F0, and the fact that listeners had equivalent sensitivity for the three cues is strong evidence that energetic masking did not play a significant role in affecting the outcome. Also, if level differences were masking the other two cues, one might also have expected that level would be the dominant cue. A third possibility is that the within-sentence modulations in level associated with prosodic speech may have served as an additional source of variation discouraging listeners to rely on the level cue. This also could be said, however, for the within sentence variations in F0. A fourth possibility is that the overlearned association of gender with F0 could have biased a reliance on F0. These last two considerations involve natural properties of speech to be considered among factors that influence the relative reliance on cues. Finally, a nonprocedural consideration is the possibility that the relative reliance reflects the heuristic value of these cues in natural listening environments. The F0 of a talker's voice and the location from which she speaks are intrinsic properties of the talker that afford reliable cues for distinguishing that talker from other talkers speaking at the same time (Lutfi, 2008). This is less true for speech level, it can change with the acoustics of the room, a turn of the talker's head, or the sudden movement of someone in the crowd blocking the path of the talker's speech to the listener's head. This idea has perhaps face validity, but it also has the shortcoming of being rather difficult to test.

We can offer for now no compelling reason why listeners would largely ignore speech level as a cue in this study, but it does have a potential methodological implication. This regards the common practice of using speech level of the target corresponding to some threshold level of performance as a dependent measure of cue reliance (Bronkhorst, 2000, 2015; Kidd and Colburn, 2017). The practice evolved from early noise masking studies where the very definition of masking depends on the level of the signal at threshold for detection (Moore, 2004). Taking the interference between the speech of talkers as a form of masking, the same dependent measure was naturally adopted as a measure of speech interference. However, the situation is different when the practice is used to measure cue reliance. The speech level of the target is itself a cue for the target speech and competes for the attention of listeners like any other cue. So, when chosen as a dependent measure of cue reliance it takes on special status in determining the outcome. The question becomes whether the outcome would be different if that special status were assigned to a different cue. Principally any major cue could serve this role [see Ahrens *et al.* (2020) and Ozmeral and Higgins (2022)], so it is curious that when the playing field is leveled for the three major cues of this study, the one most often chosen by researchers to measure listener reliance is the one listeners rely on least. The result, at a minimum, suggests another potential complication when inferring cue reliance from metrics based on performance accuracy.

Finally, a caveat. Choices were made in this study to approximate as close as possible a natural listening situation without compromising the goals of the study. Recordings of grammatical, naturally spoken sentences were used and were filtered by HRTFs an attempt to simulate an out-of-head sound image over headphones. Cues were made available simultaneously to listeners and were perturbed to mimic natural variation in speech. A talker identification task was used rather than a speech identification task only because of a concern that the listener's facility with English might influence the results. Still, to allow meaningful statements regarding the listeners relative reliance on cues, the relative information provided by cues had to be fixed within each trial block. Listeners were told of this at the start of the study and were told before each trial block which cue or combination of cues would provide correct information for the task. In real-world listening, of course, the information cues provide is much more fluid, there are more cues to attend to (timbre, speaking rate), and the listener is not afforded beforehand knowledge of precisely which cues to listen for. The results could certainly turn out differently under such circumstances (Brandiwie and Zahorik, 2010; Getzmann *et al.*, 2014; Aspeslagh *et al.*, 2014). A possible direction for future research would be to use the methods adopted here to track from trial to trial how listeners' decision weights change with changes in the information provided by cues without the listeners knowledge. Such a study would have greater bearing on the question regarding to the extent to which the results observed here or elsewhere generalize to natural settings outside the lab.

ACKNOWLEDGMENTS

The authors would like to thank lab mates Lindsey Kummerer and Briana Rodriguez, Associate Editor Dr. Emily Buss, and two anonymous reviewers for helpful comments on an earlier version of this manuscript. This research was supported by NIH Grant No. NIDCD R01-DC001262.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have not conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

[1] The perception of externalized images resulting from HRTF filtering depends in large measure on the listener. We can make no claims here regarding the extent to which images were externalized for individual listeners.

[2] Note that when the listener does not have previous knowledge of which of two speech streams corresponds to the target, as in the present study, the best they can do is to base decisions on the sum of the pair values of L, F0, and θ on each trial.

[3] The regression model of Eq. (1) assumes a linear combination of cues in **f**. Richards (2002) has considered the effect on estimates of c_i of a class nonlinearities of the form $\Sigma((c_i\ f_i^n)^k)^m$. She shows that the estimates of c_i in this case depend on n and k, but not on m. Other nonlinear transformations of f_i are likely also to affect estimates of c_i (cf. Eddins and Liu, 2012). We acknowledge such possible influences, but in the absence of knowledge of these nonlinearities and for the purposes of this study, they are treated as simply additional factors affecting reliance, the strength of the relation between f_i and the listeners response.

[4] The values of the three cues are sampled independently and randomly from trial to trial and have the same mean Δ/σ within the all-cue conditions ($\Delta/\sigma = 1.5$ or 2.0), hence, for equal weight given to the three cues $d'_{\text{ideal}} = 3^{1/2}\Delta/\sigma$ (=2.6 or 3.5) corresponding effectively to perfect percent correct performance.

Ahrens, A., Marschall, M., and Dau, T. (**2020**). "The effect of spatial energy spread on sound image size and speech intelligibility," J. Acoust. Soc. Am. **147**, 1368–1378.

Ahumada, A., Jr. (**2002**). "Classification image weights and internal noise level estimation," J. Vision **2**, 8–131.

Aspeslagh, S., Clark, D. F., Ackeroyd, M. A., and Brimijoin, W. O. (**2014**). "Speech intelligibility can improve rapidly during exposure to a novel acoustic environment," J. Acoust. Soc. Am. **135**, 2227.

Berg, B. G. (**1990**). "Observer efficiency and weights in a multiple observation task," J. Acoust. Soc. Am. **88**, 149–158.

Brandiwie, E., and Zahorik, P. (**2010**). "Prior listening in rooms improves speech intelligibility," J. Acoust. Soc. Am. **128**, 291–299.

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

Bronkhorst, A. W. (**2015**). "The cocktail-party problem revisited: Early processing and selection of multi-talker speech," Atten. Percept. Psychophys. **77**, 1465–1487.

Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.

Byrne, A. J., Conroy, C., and Kidd, G., Jr. (**2022**). "The effects of uncertainty in level on speech-on-speech masking," Trends Hear. **26**, 233121652210775.

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and two ears," J. Acoust. Soc. Am. **25**, 975–979.

Calandruccio, L., and Doherty, K. A. (**2007**). "Spectral weighting strategies for sentences measured by a correlational method," J. Acoust. Soc. Am. **121**(6), 3827–3836.

Dai, L., and Shinn-Cunningham, B. G. (**2016**). "Contributions of sensory coding and attentional control to individual differences in performance in spatial auditory selective attention tasks," Front. Hum. Neurosci. **10**(530), 530.

Doherty, K. A., and Lutfi, R. A. (**1996**). "Spectral weights for level discrimination in listeners with sensorineural hearing loss," J. Acoust. Soc. Am. **99**, 1053–1058.

Eddins, D. A., and Liu, C. (**2012**). "Psychometric properties of the coordinate response measure corpus with various types of background interference," J. Acoust. Soc. Am. **131**, EL177–EL183.

Fechner, G. T. (**1860**). *Elemente Der Psychophysik (Elements of Psychophysics)* (Breitkopf und Härtel, Leipzig), Vol. 2.

Gardner, W. G., and Martin, K. D. (**1995**). "HRTF measurements from a KEMAR," J. Acoust. Soc. Am. **97**, 3907–3908.

Getzmann, S., Lewald, J., and Falkenstein, M. (**2014**). "Using auditory pre-information to solve the cocktail-party problem: Electrophysiological evidence for age-specific differences," Front. Neurosci. **8**(413), 413.

Green, D. M., and Swets, J. A. (**1966**). *Signal Detection Theory and Psychophysics* (Wiley, New York).

Hejna, D., and Musicus, B. R. (**1991**). "The SOLAFS time-scale modification algorithm," BBN technical report.

Horii, Y. (**1975**). "Some statistical characteristics of voice fundamental frequency," J. Speech Hear. Res. **18**(1), 192–201.

Houtsma, A. J. M., and Smurzynszki, J. (**1990**). "Pitch identification and discrimination for complex tones with many harmonics," J. Acoust. Soc. Am. **87**, 304–310.

Jesteadt, W., Wier, C. C., and Green, D. M. (**1977**). "Intensity discrimination as a function of frequency and sensation level," J. Acoust. Soc. Am. **61**, 169–177.

Kidd, G., Jr., and Colburn, S. (**2017**). "Informational masking in speech recognition," in *Springer Handbook of Auditory Research: The Auditory System at the Cocktail Party*, edited by J. C. Middlebrooks, J. Z. Simon, A. N. Popper, and R. R. Fay (Springer-Verlag, New York), pp. 75–110.

Kidd, G., Jr., Mason, C. R., Swaminathan, J., Roverud, E., Clayton, K. K., and Best, V. (**2016**). "Determining the energetic and informational components of speech-on-speech masking," J. Acoust. Soc. Am. **140**, 132–144.

Lutfi, R. A. (**1995**). "Correlation coefficients and correlation ratios as estimates of observer weights in multiple-observation tasks," J. Acoust. Soc. Am. **97**, 1333–1334.

Lutfi, R. A. (**2008**). "Sound source identification," in *Springer Handbook of Auditory Research: Auditory Perception of Sound Sources*, edited by W. A. Yost and A. N. Popper (Springer-Verlag, New York).

Lutfi, R. A. and Liu, C. J. (**2007**). "Individual differences in source identification from synthesized impact sounds," J. Acoust. Soc. Am. **122**, 1017–1028.

Lutfi, R. A., Pastore, T., Rodriguez, B., Lee, J., and Yost, W. A. (**2022**). "Molecular analysis of individual differences in talker search at the cocktail party," J. Acoust. Soc. Am. **152**(3), 1804–1813.

Lutfi, R. A., Rodriguez, B., and Lee, J. (**2021**). "The listener effect in multitalker speech segregation and talker identification," Trends Hear. **25**, 233121652110518.

Lutfi, R. A., Rodriguez, B., Lee, J., and Pastore, T. (**2020**). "A test of model classes accounting for individual differences in the cocktail-party effect," J. Acoust. Soc. Am. **148**, 4014–4024.

Moore, B. C. J. (**2004**). *An Introduction to the Psychology of Hearing*, 5th ed. (Elsevier, London).

Oberfeld, D., and Klöckner-Nowotny, F. (**2016**). "Individual differences in selective attention predict speech identification as a cocktail party," eLife **5**, e16747.

Ozmeral, E. J., and Higgins, N. C. (**2022**). "Defining functional spatial boundaries using a spatial release from masking task," JASA Express Lett. **2**, 124402.

Parker, S., and Schneider, B. (**1974**). "Nonmetric scaling of loudness and pitch using similarity and difference estimates," Percept. Psychophys. **15**, 238–242.

Perrot, D. R., and Saberi, K. (**1990**). "Minimum audible angle thresholds for sources varying in both elevation and azimuth," J. Acoust. Soc. Am. **87**, 1728–1731.

Richards, V. M. (**2002**). "Effects of a limited class of nonlinearities on estimates of relative weights," J. Acoust. Soc. Am. **111**, 1012–1017.

Ruggles, D., and Shinn-Cunningham, B. G. (**2011**). "Spatial selective auditory attention in the presence of reverberant energy: Individual differences in normal-hearing listeners," J. Assoc. Res. Otolaryngol. **12**(3), 395–405.

Shinn-Cunningham, B. (**2017**). "Cortical and sensory causes of individual differences in selective attention ability among listeners with normal hearing thresholds," J. Speech. Lang. Hear. Res. **60**(10), 2976–2988.

Stevens, S. S. (**1961**). "To honor Fechner and repeal his law," Sci. New Ser. Am. Assoc. Advanc. Sci. **133**(3446), 80–86.

Szabo, B., Denham, S., and Winkler, I. (**2016**). "Computational models of auditory scene analysis: A review," Front. Neurosci. **10**, 524.

Watson, C. S. (**1973**). "Psychophysics," in *Handbook of General Psychology*, edited by B. Wohlman (Prentice-Hall, Englewood Cliffs, NJ).

Wier, C. C., Jesteadt, W., and Green, D. M. (**1977**). "Frequency discrimination as a function of frequency and sensation level," J. Acoust. Soc. Am. **61**, 178–184.

Wightman, F. L., and Kistler, D. J. (**1989**). "Headphone simulation of free-field listening: I. Stimulus synthesis," J. Acoust. Soc. Am. **85**(2), 858–867.

Zhou, K., Sisman, B., Liu, R., and Li, H. (**2021**). "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 920–924.