Using spatial release from masking to estimate the magnitude of the familiar-voice intelligibility benefit

Ysabel Domingo, Emma Holmes, Ewan Macpherson, and Ingrid S. Johnsrude

Citation: The Journal of the Acoustical Society of America **146**, 3487 (2019); doi: 10.1121/1.5133628
View online: https://doi.org/10.1121/1.5133628
View Table of Contents: https://asa.scitation.org/toc/jas/146/5
Published by the Acoustical Society of America

Using spatial release from masking to estimate the magnitude of the familiar-voice intelligibility benefit

Ysabel Domingo[a),b)] and Emma Holmes[c)]
Brain and Mind Institute, University of Western Ontario, London, Ontario, Canada

Ewan Macpherson[d)]
School of Communication Sciences and Disorders, University of Western Ontario, London, Ontario, Canada

Ingrid S. Johnsrude[a)]
Brain and Mind Institute, University of Western Ontario, London, Ontario, Canada

(Received 18 May 2019; revised 19 October 2019; accepted 23 October 2019; published online 25 November 2019)

The ability to segregate simultaneous speech streams is crucial for successful communication. Recent studies have demonstrated that participants can report 10%–20% more words spoken by naturally familiar (e.g., friends or spouses) than unfamiliar talkers in two-voice mixtures. This benefit is commensurate with one of the largest benefits to speech intelligibility currently known—that which is gained by spatially separating two talkers. However, because of differences in the methods of these previous studies, the relative benefits of spatial separation and voice familiarity are unclear. Here, the familiar-voice benefit and spatial release from masking are directly compared, and it is examined if and how these two cues interact with one another. Talkers were recorded while speaking sentences from a published closed-set "matrix" task, and then listeners were presented with three different sentences played simultaneously. Each target sentence was played at 0° azimuth, and two masker sentences were symmetrically separated about the target. On average, participants reported 10%–30% more words correctly when the target sentence was spoken in a familiar than unfamiliar voice (collapsed over spatial separation conditions); it was found that participants gain a similar benefit from a familiar target as when an unfamiliar voice is separated from two symmetrical maskers by approximately 15° azimuth.
https://doi.org/10.1121/1.5133628

[AKCL] Pages: 3487–3494

I. INTRODUCTION

Many everyday conversations occur in the presence of background sounds. The ability to separate simultaneous sounds is essential for successful communication, and recognising what one person is saying in the presence of other talkers (termed "the cocktail party problem"; Cherry, 1953) is a perceptual challenge that has received considerable attention. Much of previous work has focused on how similarity or differences in acoustic features—such as spatial location, frequency, timbre, or onset time—contribute to perceptual grouping/segregation of sounds in mixtures (e.g., Brungart *et al.*, 2001; Cusack *et al.*, 2004; Darwin *et al.*, 2003; Kitterick *et al.*, 2010; Singh and Bregman, 1997).

One feature that robustly improves the ability to segregate speech from competing sounds is prior knowledge of the talker's voice (e.g., Holmes *et al.*, 2018; Johnsrude *et al.*, 2013; Kreitewolf *et al.*, 2017; Newman and Evers, 2007; Souza *et al.*, 2013). Recognition of familiar voices is different from the process of talker normalization, in which speech processing is thought to be recalibrated when listening to speech from a new talker in order to resolve acoustic-phonetic ambiguities (Wong *et al.*, 2004). Familiar voice recognition may occur through learning acoustic patterns that are formed from averaging multiple utterances of a single speaker to form a speech prototype (Fontaine *et al.*, 2017). Therefore, if a listener is exposed to a wide variety of utterances in terms of prosody, affect, and linguistic content, the speech prototype developed will be more flexible than one formed from limited input. When a speech prototype is formed, incoming speech is then compared to it to determine if it was produced by a familiar talker.

Benefits of voice familiarity on speech-on-speech listening tasks have been established using training paradigms (Levi *et al.*, 2011; Nygaard and Pisoni, 1998; Nygaard *et al.*, 1994; Yonan and Sommers, 2000). A large benefit has also been shown using naturally familiar voices such as those of the participant's spouse or friend (Holmes *et al.*, 2018; Domingo *et al.*, 2019; Johnsrude *et al.*, 2013). A benefit of 2–9 dB is observed when a familiar voice is masked by a single unfamiliar talker at a signal-to-noise ratio (SNR) of −3 to −6 dB, when using a closed-set matrix task such as the Boston University Gerald (BUG) task (Kidd *et al.*, 2008) in which all

[a)]Also at: Psychology Department, University of Western Ontario, London, Ontario, Canada.
[b)]Electronic mail: bdomingo@uwo.ca
[c)]Current address: Wellcome Centre for Human Neuroimaging, UCL Queen Square Institute of Neurology, University College London, London, UK.
[d)]Also at: National Centre for Audiology, University of Western Ontario, London, Ontario, Canada.

of the sentences are of the form ⟨Name⟩ ⟨past tense verb⟩ ⟨number⟩ ⟨adjective⟩ ⟨noun⟩, where all the words are monosyllables (e.g. "Pat bought five old gloves.").

Despite differences in testing paradigms, the considerable improvement in intelligibility from voice familiarity is commensurate with one of the most thoroughly researched cues known to improve speech intelligibility in multitalker situations—spatial release from masking (Arbogast *et al.*, 2005; Best *et al.*, 2006; Best *et al.*, 2011; Glyde *et al.*, 2015; Kidd *et al.*, 2010; Singh *et al.*, 2008). Spatial release from masking is the improvement in word report when one or more masker talkers are presented at different spatial locations than a target talker, compared to when they are collocated.

Spatial cues include the "better ear effect" due to head shadow, defined as attending to the ear with a more favourable SNR (Carlile, 2014) and binaural interaction, in which the auditory system leverages interaural time or level differences between target and maskers (Freyman *et al.*, 1999).

The magnitude of spatial release from masking depends, in part, on the spatial relationship between target and masker stimuli. The symmetrical masker paradigm has a stimulus configuration in which two maskers are presented symmetrically (i.e., one on the left and the other the same distance to the right) about a centrally located target (Brungart and Iyer, 2012; Marrone *et al.*, 2008). Unlike other designs that have used asymmetrically configured speech signals (Arbogast *et al.*, 2002; Freyman *et al.*, 1999; Hawley *et al.*, 2004; Johnstone and Litovsky, 2006), this design controls for "better ear" listening and head-shadow effects because the SNR is the same in the left and right ears (Brungart and Iyer, 2012). Using symmetrical maskers placed at 90° about the target, listeners obtained a spatial release from masking of 4 dB in an open-set sentence identification in a modulated-noise task (Bronkhorst and Plomp, 1992), 6 dB in a closed-set word identification in masking speech task (Yost, 2017) and 12 dB in a closed-set coordinate response measure (CRM) speech-in-speech task (Marrone *et al.*, 2008).

The current study aimed to more directly compare the benefits to speech intelligibility of spatial separation and voice familiarity. We also examined whether, and how, these acoustic (spatial) and cognitive (familiarity) cues interact with one another. We used the symmetric masker paradigm with spatial separations ranging from 0° to 90° in order to compare intelligibility of a personally familiar voice to that of an unfamiliar voice in the presence of an unfamiliar masking talker (producing two different sentences). The target voice was either familiar, such as the listener's friend or romantic partner, or unfamiliar (the friend or partner of another listener). The two maskers were always different sentences and were spoken by an unfamiliar voice different from the target voice. We measured the magnitude of the familiar-voice benefit to intelligibility, and cast this in terms of the degrees of spatial separation required to produce a benefit of equal magnitude (relative to the collocated condition) when the target voice was unfamiliar. We compared the benefits of voice familiarity and spatial separation on intelligibility at three different target-to-masker ratios (TMRs; –3, 0, or 6 dB).

II. METHODS

A. Participants

Participants were nine pairs of friends, siblings, roommates, or romantic couples who were naturally familiar with each other's voices. There were three male-female pairs, four female-female pairs, and two male-male pairs. Pairs of participants had known each other for longer than six months [median = 4.7 yr, interquartile range (IQR) = 5.7] and reported that they spoke to each other between 3 and 90 h per week (median = 21 h, IQR = 18.9). The 18 participants (7 male, 11 female) were 18–33 yr of age (median = 20.5 yr, IQR = 6.8). Participants were native Canadian English speakers with no known history of speech or hearing impairments. Participants had four-frequency (0.5, 1, 2, and 4 kHz) average pure-tone hearing thresholds of 20 dB hearing level (HL) or better in each ear.

This experiment was approved by the Non-Medical Research Ethics Board at the University of Western Ontario. Informed consent was obtained from all participants prior to testing.

One pair completed the recording sessions but did not return for the listening task, and one participant's responses were dropped from the analysis due to experimenter error. Data from the remaining 15 participants were analysed.

B. Apparatus

The experiment was conducted in a single-walled sound-attenuating booth (Model CL-13 LP MR, Eckel Industries, Morrisburg, Ontario, Canada). Participants sat in a chair facing a 24-in. liquid-crystal display (LCD) monitor (either ViewSonic VG2433SMH, Brea, CA, or Dell G2410t, Round Rock, TX).

Speech stimuli were recorded using a Sennheiser e845-S microphone (Wedemark, Germany) connected to a Steinberg UR22 mkII sound card (Steinberg Media Technologies, Hamburg, Germany) and delivered binaurally through Grado Labs SR224 headphones (Grado Labs, Brooklyn, NY). Recordings were made and edited using Audacity (version 2.0.3, retrieved from https://audacityteam.org/) software.

C. Stimuli

Stimuli were sentences from the BUG corpus (Kidd *et al.*, 2008). The sentences in this corpus are of the format "⟨Name⟩ ⟨past-tense verb⟩ ⟨number⟩ ⟨adjective⟩ ⟨noun⟩." We used a subset of 480 sentences containing 2 names ("Bob" and "Pat"), 8 verbs ("bought," "sold," "found," "lost," "took," "gave," "held," "saw"), 8 numbers ("two," "three," "four," "five," "six," "eight," "nine," "ten"), 8 adjectives ("blue," "red," "hot," "cold," "big," "small," "old," "new"), and 8 nouns ("hats," "bags," "shoes," "socks," "pens," "gloves," "toys," "cards"). An example is "Pat held three blue hats."

Unlike the original corpus in which individual words were recorded in citation form, our participants were recorded speaking complete sentences (480 in total, recorded in mono sound; 44.1 kHz sampling rate). Participants were shown a sentence on the screen, and a vertical bar moved across the sentence from left to right (Holmes *et al.*, 2018).

Participants were instructed to read the words in the sentence as the bar moved over them in an effort to maintain a consistent speaking rate throughout the recording session. All sentences were normalized to the same root-mean-square (RMS) amplitude and each had a duration of approximately two seconds.

Throughout the experiment, each participant heard sentences spoken by three different talkers. These included one familiar voice—that of the participant's partner—and two unfamiliar voices (who were the familiar voices of other participants). The unfamiliar voices were sex-matched to each participant's familiar voice; we did not attempt to match F_0 between familiar and unfamiliar voices. All voices were presented once as familiar and twice as unfamiliar, except for the three participants whose data were not analysed. Two of the three participants were partners with each other, so their voices were only presented as unfamiliar (twice). The third participant's voice was presented as both familiar and unfamiliar, but his partner's voice only served as an unfamiliar voice (twice).

The recorded sentences were presented binaurally over headphones using virtual spatial cues in the azimuth plane. Binaural stimuli were processed with anechoic head-related transfer functions (HRTFs) measured on a KEMAR mannequin (Knowles Electronics, Itasca, IL; Algazi *et al.*, 2001).

Acoustic stimuli were presented at a comfortable listening level [approximately 67 dB sound pressure level (SPL)]. Across trials, the overall amplitude of the stimuli was roved over a range of 3 dB (in six equally spaced levels) to ensure that participants could not use the amplitude of either the target or the masker sentences as a cue to identify the target sentence.

D. Methods and procedures

On each trial, participants were presented with three simultaneous sentences. The target sentence, which was presented at 0° azimuth (i.e., in front of the participant), was spoken in one voice. The two masker sentences were spoken in a second (always unfamiliar) voice of the same sex as the target. The voice speaking the two masker sentences was always the same for each trial. They were either collocated with the target (i.e., also presented at 0° azimuth) or separated symmetrically about the target at ± 5, 10, 15, 25, 45, or 90° azimuth. A schematic of stimulus configuration is shown in Figs. 1(A) and 1(B). The target sentence always began with a particular name word ("Bob" in one half of the experiment, "Pat" in the other; order counterbalanced across participants). The two masker sentences began with the other name word. The four remaining words were always different in the three sentences. Participants were asked to identify the four words in the target sentence by clicking the words on a screen [Fig. 1(C)].

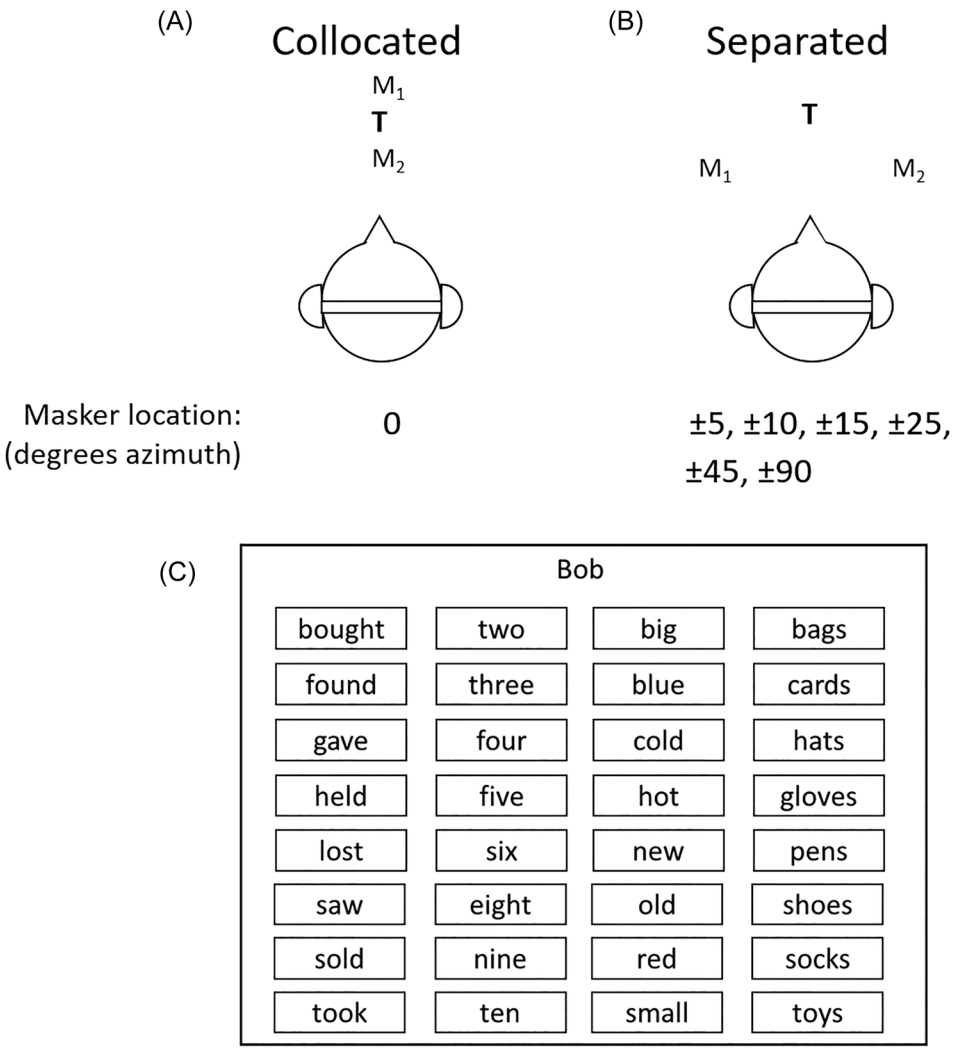

FIG. 1. Procedure used in listening sessions. In the collocated condition (A), the target, T, and masker sentences, M_1 and M_2, were played in virtual auditory space at 0° azimuth. In the spatially separated condition (B), the target was played at 0° azimuth, and the two masker sentences were played symmetrically about the target at ±5, ±10, ±15, ±25, ±45, and ±90° azimuth. Participants tracked the target voice and responded by choosing one word (by a mouse press) from each column on the response screen (C) according to what they had heard in the target sentence, indicated by the target name (in this example, "Bob").

We tested listeners in two familiarity conditions. In the familiar target (FT) condition, the target sentence was spoken in the familiar voice, and the two masker sentences were spoken in one of the two unfamiliar voices (half of trials in each of the two unfamiliar voices). In the both unfamiliar (BU) condition, the target was spoken by one of the unfamiliar voices, and the two masker sentences were spoken by the other unfamiliar voice (each unfamiliar voice was the target on half of the trials).

The target and masker sentences were presented at TMRs of −3, 0, and 6 dB, defined as the ratio between the target and each individual masker. TMRs were maintained while the roving overall level of the combined stimuli.

There were 16 trials of each combination of the 2 familiarity conditions, 7 spatial configurations, and 3 TMRs—producing a total of 672 trials for each participant across 42 unique conditions. Trials were presented in 14 blocks of 48; each condition was presented 3 times per block in random order. Participants were given the option to take a short break between blocks.

E. Data analysis

Speech intelligibility was calculated as the proportion of words (out of a possible 64, 4 words in each of the 16 trials) that each participant correctly identified from the target sentence in each condition. Chance performance for each word was 1/8 or 12.5%. These proportions were then normalized into rationalized arcsine units (RAUs; Studebaker, 1985). To determine the effects of voice familiarity and spatial separation on speech intelligibility, we conducted a three-way repeated measures analysis of variance (ANOVA) on RAU-transformed data, with familiarity (two levels: FT, BU), spatial separation (seven levels: 0°, 5°, 10°, 15°, 25°, 45°, 90°), and TMR (three levels: −3, 0, 6 dB) as within-subjects variables. Mauchly's test indicated that the assumption of sphericity was violated for the main effects of TMR [$\chi^2(2) = 36.4$, $p < 0.001$] and spatial separation [$\chi^2(20) = 51.1$, $p < 0.001$], the interactions between familiarity and TMR [$\chi^2(2) = 12.9$, $p = 0.002$], and between familiarity and spatial separation [$\chi^2(20) = 40.71$, $p = 0.005$]. Thus, these effects are reported with Greenhouse-Geisser correction. Pairwise comparisons are reported with Sidak correction for multiple comparisons.

In order to determine the equivalence point (the spatial separation that provides release from the masking equivalent to the familiar-voice benefit), we used the lsqcurvefit function on MATLAB R2014b (The MathWorks Inc., Natick, MA) to fit the data to the following three-parameter exponential function:

$$y = a(e^{bx}) + c,$$

where a, b, and c are free parameters, and x is the spatial separation in degrees.

We then used the function fitted to the BU data to estimate the spatial separation that produced an improvement in accuracy equivalent to the average intelligibility in the FT condition when the maskers were collocated (at 0°). This was done for each TMR separately.

III. RESULTS

A. Familiarity, spatial separation, and TMR affect intelligibility

Figure 2 illustrates the effects of spatial separation and familiarity factors on RAU-transformed proportions of correct

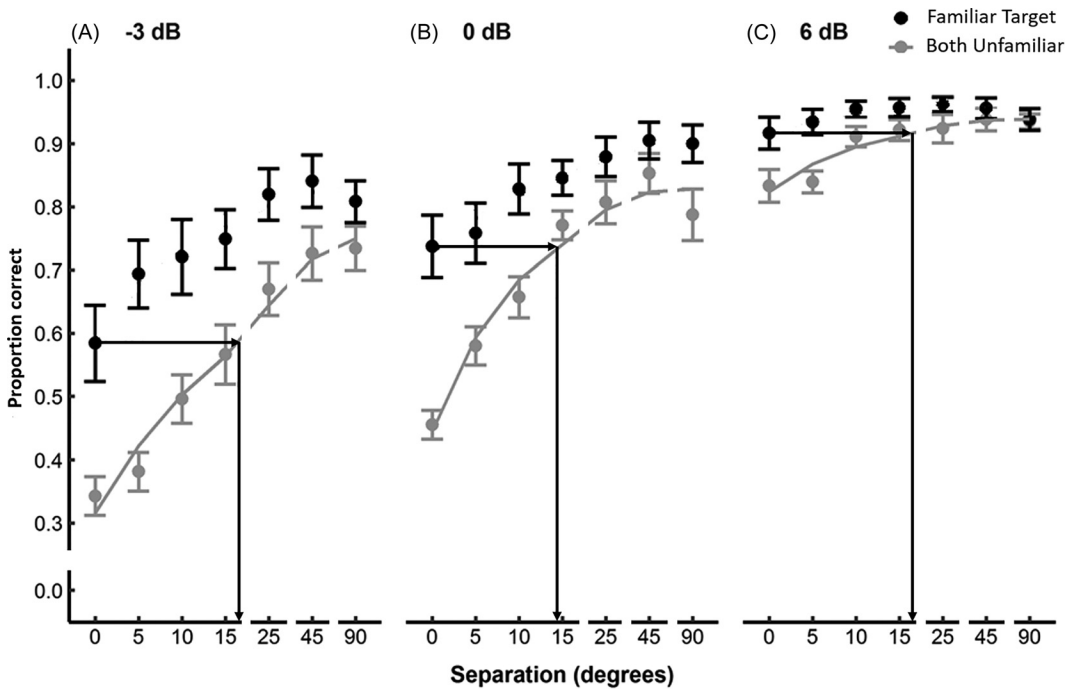

FIG. 2. Proportion of correct words as a function of spatial separation at -3 dB (A), 0 dB (B), and 6 dB (C) TMR. The markers represent averaged raw speech intelligibility data in the FT (black) or BU (grey) condition. The line is the exponential functions fitted to the raw data in the BU condition. The black arrows show the spatial separation on the BU function that has equivalent intelligibility to the FT condition at 0°. Error bars are ±1 standard error of the mean.

words. A repeated measures ANOVA showed that intelligibility was significantly better when the target sentence was spoken in the familiar voice [mean = 86.69%, standard error of the mean (SE) = 3.69%] than when it was spoken in the unfamiliar voice [mean = 72.44%, SE = 2.06%; $F(1,14) = 23.55$, $p < 0.001$, $\omega^2 = 0.58$].

The main effect of spatial separation was also significant [$F(2.01,28.14) = 56.43$, $p < 0.001$, $\omega^2 = 0.78$]. Comparing adjacent spatial separation conditions, intelligibility was significantly better for greater spatial separations between 0° and 25° (0°–5°: $p = 0.028$; 5°–10°: $p = 0.04$; 10–15°: $p = 0.035$; 15–25°: $p = 0.011$). However, intelligibility did not improve among 25°, 45°, and 90° (all $p > 0.05$).

Intelligibility improved significantly with increasing TMR [$F(1.07,14.98) = 236.43$, $p < 0.001$, $\omega^2 = 0.94$]. Intelligibility was significantly better at 6 dB (mean = 95.77, SE = 2.04) than at 0 dB (mean = 77.44, SE = 2.82; $p < 0.001$), and better at 0 dB than at −3 dB (mean = 65.34, SE = 3.22; $p < 0.001$).

The interaction between TMR and spatial separation was significant [$F(12,168) = 13.05$, $p < 0.001$, $\omega^2 = 0.44$], probably due to uniformly high performance in the most favourable TMR condition (6 dB). At −3 dB and 0 dB TMR, intelligibility at 0° was worse than at all greater separations, intelligibility at 5° and 10° separation was significantly worse than at 45° and 90°, and performance at 15° was worse than at 45° [all $t(14) \geq 4.24$, all $p \leq 0.017$]. In addition, at −3 dB TMR, intelligibility at 15° was worse than at 90° [$t(14) = 4.81$, $p = 0.006$]. Compared to the lower TMRs, at 6 dB, spatial cues had less of an effect on intelligibility. At 6 dB TMR, intelligibility at 0° was worse than at 10°, 15°, and 45° [all $t(14) \geq 3.83$, all $p \leq 0.038$], intelligibility at 5° was worse than at 45° and 90° [all $t(14) \geq 4.90$, all $p \leq 0.005$], whereas intelligibility at 15° did not differ from any greater spatial separations.

There were also significant interactions between familiarity and spatial separation and between familiarity and TMR. These two-way interactions will be discussed within the context of the significant three-way interaction below.

There was a significant three-way interaction between familiarity, TMR, and spatial separation [$F(12,168) = 2.25$, $p = 0.012$, $\omega^2 = 0.08$]. To reduce this three-way interaction to a two-way interaction (which is more easily interpretable), we computed the difference in intelligibility between the FT and BU conditions (the "familiar-voice benefit") at each TMR and spatial separation. Figure 3 displays the familiar-voice benefit by spatial separation for each TMR. We then conducted a repeated measures ANOVA with the familiar-voice benefit as the dependent measure. The results showed a significant main effect of TMR [$F(1.23,17.19) = 8.04$, $p = 0.008$, $\omega^2 = 0.31$], a significant main effect of separation [$F(2.70,37.85) = 8.36$, $p < 0.001$, $\omega^2 = 0.32$], and a significant two-way interaction [$F(12,168) = 2.25$, $p = 0.036$, $\omega^2 = 0.10$]. To interpret the interaction, we examined the simple main effect of separation at each TMR by conducting within-TMR, across-separation paired comparisons. The results indicated that there was no simple main effect of separation at 6 dB TMR, whereas there was a simple main effect at 0 dB and at −3 dB TMR. At 6 dB TMR, the familiar-voice benefit did not differ across spatial separations (all $p \geq 0.09$), whereas at −3 and 0 dB TMR, the familiar-voice benefit was greater at small separations than larger separations. The familiar-voice benefit at 5° was greater than at 45° and 90° at −3 dB TMR [all $t(14) \geq 4.03$, all $p \leq 0.026$] and greater at 0° compared to 15° and 45° at 0 dB TMR [all $t(14) \geq 3.85$, all $p \leq 0.037$].

In Figs. 2 (and 3) and from the analysis presented above, it is clear that the familiar-voice benefit is smaller at larger spatial separations across TMRs. This could be due to intelligibility of familiar and unfamiliar targets (in the presence of an unfamiliar masker) both reaching a ceiling at large spatial separations, but this is unlikely to be the explanation for the −3 dB TMR condition at least, performance at −3 dB TMR can clearly go higher. Indeed, intelligibility of both familiar and unfamiliar talkers at 6 dB was significantly better than at −3 dB TMR and at 0 dB TMR [−3 dB: all $t(14) \geq 3.76$, all $p \leq 0.006$; 0 dB: all $t(14) \geq 3.51$, all $p \leq 0.009$; trend only for FT at 90° ($t(14) = 2.44$, $p = 0.08$)].

B. Equivalence between familiar-voice benefit and spatial release from masking

Neither benefit from familiarity nor spatial separation is possible when a target that is spoken in an unfamiliar voice, on the midline, is masked by two collocated sentences spoken in another unfamiliar voice. This served as our baseline condition against which to measure benefits from familiarity

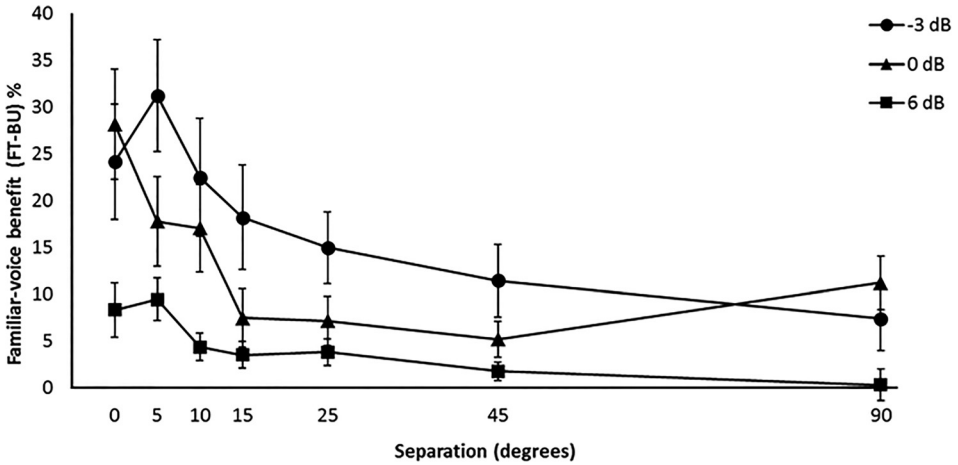

FIG. 3. Familiar-voice benefit (difference percentage of correct words identified between the FT and BU conditions) at each spatial separation and TMR (−3 dB TMR = circles, 0 dB TMR = triangles, 6 dB TMR = squares). Error bars are ±1 standard error of the mean. Statistical analyses were based on RAU-transformed data of each condition.

and spatial separation. The benefit of a familiar voice was calculated by subtracting intelligibility in the baseline condition from intelligibility in the condition in which the maskers were collocated but the target was familiar.

We then fitted the three-parameter exponential function to averaged BU data; see Fig. 2. The functions provided good fits to the data with residuals (i.e., differences between the fitted functions and the data, in terms of proportion correct) smaller than 0.045 for each data point, where the possible range of values is between 0 and 1. Using the function fitted to the BU data, we then determined the spatial separation that yielded the benefit equivalent in magnitude to the familiar-voice benefit (the "equivalence point"), separately at each TMR. At −3 dB TMR, the equivalence point was ±17.1°, at 0 dB TMR, the equivalence point was ±14.6°, and at 6 dB TMR, the equivalence point was ±17.0°.

Next, we quantified the familiar-voice benefit in terms of TMR. When maskers and target were collocated on the midline, participants were 20% more accurate in reporting words spoken by a familiar voice than an unfamiliar voice (averaged across TMRs). In order to quantify this benefit in dB, we fit a linear regression line to the BU condition when target and masker were collocated at 0° and interpolated the TMR that yields the same accuracy as that in the FT at −3 dB (collocated). Figure 4 shows the intelligibility in the FT and BU conditions at each TMR for collocated and ±90° separated data. Since we only used three TMRs, this is necessarily a rather gross estimate. This is equal to a release from masking of 5.1 dB. When target and maskers were separated by 90°, participants were only 6% more accurate when the target voice was familiar, collapsing across TMRs. This is equal to release from masking of 4.4 dB.

IV. DISCUSSION

Our results replicate the familiar-target benefit to intelligibility, consistent with previous studies (Holmes *et al.*, 2018; Johnsrude *et al.*, 2013; Nygaard and Pisoni, 1998), and extend this by showing a familiar-target benefit in a three-sentence mixture produced by two voices. When stimuli were spatially collocated (at 0°) participants reported an average of 20% more words correctly in the FT than in the BU condition. These results are highly consistent with previous studies from our laboratory on demographically similar participants, which have found an average improvement in intelligibility of approximately 15% when a familiar, compared to unfamiliar, voice is the target (Domingo *et al.*, 2019; Holmes *et al.*, 2018; Johnsrude *et al.*, 2013).

Here, we measured the improvement in intelligibility from a familiar voice to be equivalent to the benefit provided by 14°–17°, depending on TMR. Intelligibility scores at larger separations (25°, 45°, and 90°) were not significantly different from each other (84.4%, 87.0%, and 85.1%, respectively), although they were all significantly better than at 15° (80.2%). This shows that intelligibility improvement from a familiar voice is almost as effective as the largest improvement from spatial separation.

Our findings are broadly consistent with previous studies showing that spatial release from masking plateaus at large spatial separations. When comparing the intelligibility of a target at 0° in the presence of symmetrically separated speech maskers, the benefit of increasing spatial separation from ±30° to ±90° was only ∼0.8 dB (Noble and Perrett, 2002) and 1.5 dB (Yost, 2017). These results are similar to those of Jones and Litovksy (2008), who found that spatial release from masking at 45° accounted for the majority of the spatial release from masking observed at 90°, reinforcing the idea that spatial release from masking does not have a linear relationship with spatial separation.

The familiar-voice benefit at smaller spatial separations was significantly larger than at bigger spatial separations (see Fig. 3), particularly at low TMRs (−3 dB and 0 dB). This effect cannot be solely attributed to ceiling effects at large spatial separations because we observed the same pattern at the lowest TMR (−3 dB); at this TMR, intelligibility did not exceed 85%. Furthermore, intelligibility at each spatial separation and familiarity condition generally increased with TMR, providing more evidence that the smaller familiar-voice benefit at bigger spatial separations was not simply because performance was at ceiling at these larger separations. These results suggest that listeners use voice familiarity to improve intelligibility in challenging listening conditions (i.e., at low spatial separations) but perhaps not as much at higher spatial separations when acoustic cues are sufficient to identify words in the target sentence.

Spatial separations of ±90° have been shown to provide a release from masking up to approximately 4 dB

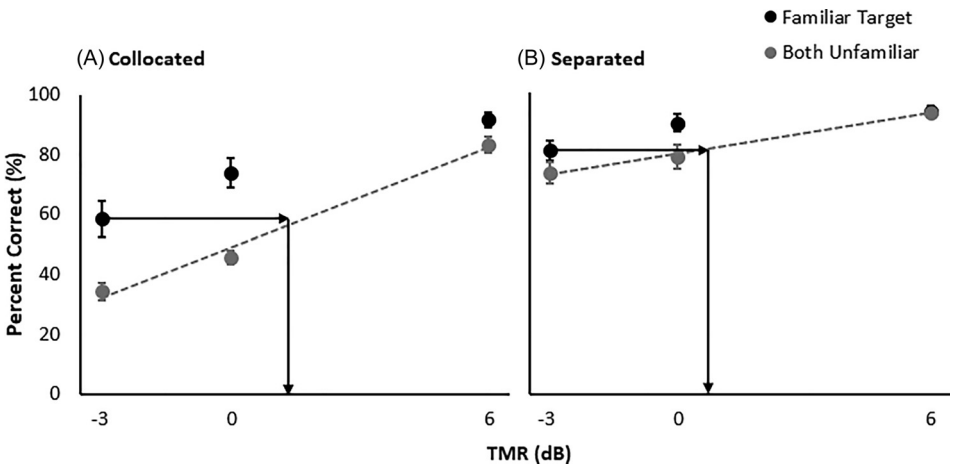

FIG. 4. Intelligibility of the FT (black) and BU (grey) conditions for (A) collocated and (B) spatially separated data at ±90° as a function of TMR. Dashed lines are the linear regression for the BU condition. The black arrows show the TMR in the BU condition that has equivalent intelligibility to the FT condition at -3 dB. Error bars are ±1 standard error of the mean.

(Bronkhorst and Plomp, 1992), 6 dB (Yost, 2017), and 12 dB (Marrone *et al*., 2008). Findings were influenced by task differences, particularly the number of words participants were required to report. The studies in which listeners reported one word (Yost, 2017) or two words (Marrone *et al*., 2008) showed higher spatial release from masking compared to studies in which listeners were required to report short sentences (Bronkhorst and Plomp, 1992). The current study also required listeners to report words from a short sentence with the exception of the first (i.e., name) word, which was used to identify the target. Using TMRs between −3 and 6 dB, release from masking at ±90° was 4.4 dB, which is highly similar to the findings of Bronkhorst and Plomp (1992).

In a previous study (Marrone *et al*., 2008) that presented symmetric maskers, intelligibility increased with greater spatial separations and reached a maximum at around 45°. Although this is greater than the maximum we found of 25°, 45°, and 90° in the current study, Marrone *et al*. (2008) did not include any spatial separations between 15° and 45° in their study. It is possible that if a condition at around 25° was included, they may have observed a plateau in intelligibility at that condition. Differences could also be due to task, where Marrone *et al*. (2008) used the CRM corpus (Bolia *et al*., 2000) and we used the BUG task (Kidd *et al*., 2008) but recorded as complete sentences from our talker participants. The differences could also be due to differences in TMR: Marrone *et al*. (2008) presented stimuli at −5.7 dB and −9.3 dB TMR for 15° and 45° separations, respectively. These TMRs are lower than any used in the current study.

Taken together, Johnsrude *et al*. (2013) and Domingo *et al*. (2019) found that the release from masking from a collocated FT voice ranges from 2 dB to over 9 dB (approximately 10%–15% improvement in intelligibility) at TMRs of −3 to −6 dB, suggesting that the release from the masking benefit of a familiar voice is commensurate with or even larger than that of a 90° spatial separation reported in previous studies. In the collocated condition of the current study, release from the masking benefit of a familiar voice was 5.1 dB (approximately 20% improvement in intelligibility). These results highlight the effectiveness of voice familiarity as a facilitator of intelligibility.

Voices were counterbalanced so each familiar voice served as the unfamiliar voice for two other participants. At a group level, the acoustics of the voices used as familiar and unfamiliar voices were therefore identical to each other, and so we focus here exclusively on group level data. Acoustics were not matched at the individual level; therefore, investigating individual differences is not possible in the current study. This limitation may be overcome in future research using a training paradigm in which all participants are presented with the same voices and different subsets of these voices are familiar for different participants.

V. CONCLUSION

This paper is the first to directly compare the benefits of voice familiarity and spatial separation on intelligibility. We replicate previous studies showing substantial benefits from both naturally familiar voices and spatial separation. Moreover, we demonstrate that the familiar-voice benefit is equivalent to spatial release from masking provided by 14°–17° of symmetric spatial separation in three-talker listening, and also provide the first data demonstrating a potential trade-off between these cues—our results suggest that individuals rely less on familiar voice information when acoustic cues, such as spatial separation and TMR, are sufficient to segregate simultaneous speech streams.

ACKNOWLEDGMENTS

This work was supported by funding from the Canadian Institutes of Health Research (CIHR; Operating Grant No. MOP 133450) and the Natural Sciences and Engineering Research Council of Canada (NSERC; Discovery Grant No. 327429-2012). The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. (**2001**). "The CIPIC HRTF database," in *Proceedings of the 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,* 24 October 2001, New Paltz NY (IEEE), pp. 99–102.

Arbogast, T. L., Mason, C. R., and Kidd, G. (**2002**). "The effect of spatial separation on informational and energetic masking of speech," J. Acoust. Soc. Am. **112**(5), 2086–2098.

Arbogast, T. L., Mason, C. R., and Kidd, G. (**2005**). "The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **117**(4), 2169–2180.

Best, V., Gallun, F. J., Ihlefeld, A., and Shinn-Cunningham, B. G. (**2006**). "The influence of spatial separation on divided listening," J. Acoust. Soc. Am. **120**(3), 1506–1516.

Best, V., Mason, C. R., and Kidd, G. (**2011**). "Spatial release from masking in normally hearing and hearing-impaired listeners as a function of the temporal overlap of competing talkers," J. Acoust. Soc. Am. **129**(3), 1616–1625.

Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**(2), 1065–1066.

Bronkhorst, A. W., and Plomp, R. (**1992**). "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing," J. Acoust. Soc. Am. **92**(6), 3132–3139.

Brungart, D. S., and Iyer, N. (**2012**). "Better-ear glimpsing efficiency with symmetrically-placed interfering talkers," J. Acoust. Soc. Am. **132**, 2545–2556.

Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (**2001**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **110**(5), 2527–2538.

Carlile, S. (**2014**). "Active listening: Speech intelligibility in noisy environments," Acoust. Aust. **42**(2), 90–96.

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**(5), 975–979.

Cusack, R., Deeks, J., Aikman, G., and Carlyon, R. P. (**2004**). "Effects of location, frequency region, and time course of selective attention on auditory scene analysis," J. Exp. Psychol. Hum. Percept. Perform. **30**(4), 643–656.

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (**2003**). "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," J. Acoust. Soc. Am. **114**(5), 2913–2922.

Domingo, Y., Holmes, E., and Johnsrude, I. S. (**2019**). "The benefit to speech intelligibility of hearing a familiar voice," J. Exp. Psychol. Appl. (published online).

Fontaine, M., Love, S. A., and Latinus, M. (**2017**). "Familiarity and voice representation: From acoustic-based representation to voice averages," Front. Psychol. **8**(1180), 1–9.

Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (**1999**). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**, 3578–3588.

Glyde, H., Buchholz, J. M., Nielsen, L., Best, V., Dillon, H., Cameron, S., and Hickson, L. (**2015**). "Effect of audibility on spatial release from speech-on-speech masking," J. Acoust. Soc. Am. **138**(5), 3311–3319.
Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (**2004**). "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," J. Acoust. Soc. Am. **115**(2), 833.
Holmes, E., Domingo, Y., and Johnsrude, I. S. (**2018**). "Familiar voices are more intelligible, even if they are not recognized as familiar," Psychol. Sci. **29**(10), 1575–1583.
Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., and Carlyon, R. P. (**2013**). "Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice," Psychol. Sci. **24**(10), 1995–2004.
Johnstone, P. M., and Litovsky, R. Y. (**2006**). "Effect of masker type and age on speech intelligibility and spatial release from masking in children and adults," J. Acoust. Soc. Am. **120**(4), 2177–2189.
Jones, G. L., and Litovsky, R. Y. (**2008**). "Role of masker predictability in the cocktail party problem," J. Acoust. Soc. Am. **124**, 3818–3830.
Kidd, G., Best, V., and Mason, C. R. (**2008**). "Listening to every other word: Examining the strength of linkage variables in forming streams of speech," J. Acoust. Soc. Am. **124**(6), 3793–3802.
Kidd, G., Mason, C. R., Best, V., and Marrone, N. (**2010**). "Stimulus factors influencing spatial release from speech-on-speech masking," J. Acoust. Soc. Am. **128**(4), 1965–1978.
Kitterick, P. T., Bailey, P. J., and Summerfield, A. Q. (**2010**). "Benefits of knowing who, where, and when in multi-talker listening," J. Acoust. Soc. Am. **127**(4), 2498–2508.
Kreitewolf, J., Mathias, S. R., and von Kriegstein, K. (**2017**). "Implicit talker training improves comprehension of auditory speech in noise," Front. Psychol. **8**, 1–8.
Levi, S. V., Winters, S. J., and Pisoni, D. B. (**2011**). "Effects of cross-language voice training on speech perception: Whose familiar voices are more intelligible?," J. Acoust. Soc. Am. **130**(6), 4053–4062.
Marrone, N., Mason, C. R., and Kidd, G. (**2008**). "Tuning in the spatial dimension: Evidence from a masked speech identification task," J. Acoust. Soc. Am. **124**, 1146–1158.
Newman, R. S., and Evers, S. (**2007**). "The effect of talker familiarity on stream segregation," J. Phonetics **35**(1), 85–103.
Noble, W., and Perrett, S. (**2002**). "Hearing speech against spatially separate competing speech versus competing noise," Percep. Psychophys. **64**(8), 1325–1336.
Nygaard, L., and Pisoni, D. (**1998**). "Talker-specific learning in speech perception," Percept. Psychophys. **60**, 355–376.
Nygaard, L., Sommers, M. S., and Pisoni, D. B. (**1994**). "Speech perception as a talker-contingent process," Psychol. Sci. **5**(1), 42–46.
Singh, G., Pichora-Fuller, M. K., and Schneider, B. A. (**2008**). "The effect of age on auditory spatial attention in conditions of real and simulated spatial separation," J. Acoust. Soc. Am. **124**(2), 1294–1305.
Singh, P. G., and Bregman, A. S. (**1997**). "The influence of different timbre attributes on the perceptual segregation of complex-tone sequences," J. Acoust. Soc. Am. **102**(4), 1943–1952.
Souza, P., Gehani, N., Wright, R., and McCloy, D. (**2013**). "The advantage of knowing the talker," J. Am. Acad. Audiol. **24**(8), 689–700.
Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Hear. Res. **28**, 455–462.
Wong, P. C. M., Nusbaum, H. C., and Small, S. L. (**2004**). "Neural Bases of Talker Normalization," J. Cogn. Neurosci. **16**, 1173–1184.
Yonan, C. A., and Sommers, M. S. (**2000**). "The effects of talker familiarity on spoken word identification in younger and older listeners," Psychol. Aging **15**(1), 88–99.
Yost, W. A. (**2017**). "Spatial release from masking based on binaural processing for up to six maskers," J. Acoust. Soc. Am. **141**(3), 2093–2106.