

# Voice Familiarization Training Improves Speech Intelligibility and Reduces Listening Effort

Trends in Hearing

Volume 29: 1–18

© The Author(s) 2025

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/23312165251401318

journals.sagepub.com/home/tia



Freja Baxter<sup>1</sup>, Harriet J. Smith<sup>1</sup>  and Emma Holmes<sup>1</sup> 

## Abstract

Understanding speech among competing speech poses a substantial challenge. In these environments, familiar voices—including naturally familiar (e.g., friends, partners) and lab-trained voices—are more intelligible than unfamiliar voices. Yet, whether familiar voices also require less effort to understand is currently unknown. We trained 20 participants to become familiar with three voices, then tested listening effort during a speech intelligibility task. During familiarization and training, participants were exposed to three talkers for different lengths of time, either speaking 88, 166, or 478 sentences (“Least Familiar,” “Moderately Familiar,” or “Most Familiar” voice, respectively). During each trial of the speech intelligibility task, two competing sentences were presented at a target-to-masker ratio (TMR) of  $-6$  or  $+3$  dB. Participants reported target sentences that were spoken by trained or by novel, unfamiliar talkers. We assessed effort using self-reported ratings and physiologically, using pupil dilation. We found that self-report scores were more sensitive than pupil dilation to differences in TMR, with lower self-reported effort at  $+3$  than  $-6$  dB TMR. The two measures may also be differentially sensitive to the extent of training. We found lower self-reported effort for all three trained voices over unfamiliar voices, with no differences among the trained voices, whereas pupil dilation was only lower for the voice that had been trained for the longest. Thus, both self-report scores and pupil dilation showed advantages for the voice that was trained for the longest ( $\sim 1$  h), but self-report scores additionally showed reduced effort even following relatively short durations of training ( $< 10$  min).

## Keywords

voice familiarity, speaker familiarity, speech perception, cocktail party listening, effort, pupillometry

Received: July 30, 2025; revised: November 3, 2025; accepted: November 12, 2025

## Introduction

In many everyday situations, understanding speech is fundamental for successful communication. Speech perception can occur effortlessly in quiet settings, but can be very effortful in adverse listening conditions, such as when competing talkers are present (Koelewijn et al., 2012). Known as the “cocktail party problem” (Cherry, 1953), the challenge of comprehending speech among competing speech is particularly demanding for older adults and for people with hearing loss (Duquesnoy, 1983; Marrone et al., 2008)—and substantial effort can be required to maintain accurate speech intelligibility in these settings (Pichora-Fuller et al., 2016). Previous studies have demonstrated that being familiar with a talker’s voice can substantially improve speech intelligibility when competing speech is present (Barker & Newman, 2004; Buntrock et al., 2021; Case et al., 2018; Domingo et al., 2019, 2020; Holmes et al., 2021; Holmes & Johnsrude, 2020,

2021, 2023; Holmes, Domingo et al., 2018; Johnsrude et al., 2013; Kreitewolf et al., 2017; Levi, 2015; Levi et al., 2011; Magnuson et al., 1995; Newman & Evers, 2007; Nygaard et al., 1994; Nygaard & Pisoni, 1998; Souza et al., 2013; Yonan & Sommers, 2000; Zhu & Holmes, 2025). Yet, whether voice familiarity also reduces effort when people try to understand speech in competing speech has not previously been examined.

<sup>1</sup>Division of Psychology and Language Sciences, Department of Speech Hearing and Phonetic Sciences, University College London, London, UK

### Corresponding author:

Emma Holmes, Division of Psychology and Language Sciences, Department of Speech Hearing and Phonetic Sciences, University College London, Chandler House, 2 Wakefield Street, London WC1N 3PF, UK.  
Email: emma.holmes@ucl.ac.uk



Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (<https://creativecommons.org/licenses/by/4.0/>) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

It is well-documented that familiar voices are more intelligible than unfamiliar voices in noisy environments: Words spoken by friends (Domingo et al., 2019, 2020; Holmes & Johnsrude, 2020, 2021, 2023; Holmes, Domingo et al., 2018), spouses (Domingo et al., 2020; Johnsrude et al., 2013; Souza et al., 2013), family members (Barker & Newman, 2004; Magnuson et al., 1995), and personally familiar university professors (Buntrock et al., 2021; Newman & Evers, 2007) are reported more accurately than words spoken by unfamiliar people. This finding holds when voices are counterbalanced across familiar and unfamiliar conditions, such that one participant's familiar voice is another participant's unfamiliar voice (e.g., Buntrock et al., 2021; Domingo et al., 2019, 2020; Holmes et al., 2021; Holmes & Johnsrude, 2020, 2021, 2023; Holmes, Domingo et al., 2018; Johnsrude et al., 2013)—meaning that the intelligibility benefit can be ascribed to familiarity, rather than voice-specific effects. Interestingly, participants seem to gain a similar-magnitude intelligibility benefit—of approximately 10–15% improvement in sentence report—from a friend they have known for at least 6 months compared to a spouse they have cohabited with for 5 years or longer (Domingo et al., 2019). This finding implies that the familiar-voice benefit to intelligibility develops relatively quickly as people get to know each other naturally, then remains relatively stable across longer durations of time.

Studies that have trained participants to become familiar with new voices in the lab demonstrate that the familiar-voice benefit can occur even with relatively short, artificial exposure (Case et al., 2018; Holmes et al., 2021; Kreitewolf et al., 2017; Levi, 2015; Levi et al., 2011; Nygaard et al., 1994; Nygaard & Pisoni, 1998; Yonan & Sommers, 2000; Zhu & Holmes, 2025). Holmes et al. (2021) compared intelligibility for three trained voices, to which participants had been exposed for different lengths of time. They first familiarized participants with the names of three new talkers, then trained participants to recognize the names of the talkers, providing feedback on each trial. Each participant was exposed to one voice speaking 88 sentences during familiarization and training, another speaking 166 sentences, and a third speaking 478 sentences. Participants then underwent a speech intelligibility test, in which they were asked to report sentences from trained and novel voices when a competing talker was present. The results showed that the three trained voices were all more intelligible than novel voices. The voice that was trained for the longest duration of time was the most intelligible, producing a 10–15% improvement in sentence report when compared to novel unfamiliar voices—which is comparable to the magnitude of the benefit that has been reported for naturally familiar voices (Domingo et al., 2020; Holmes & Johnsrude, 2021, 2023; Holmes, Domingo et al., 2018; Johnsrude et al., 2013). Thus, training participants to recognize a voice has great potential for improving the intelligibility of those voices in noisy, everyday environments.

Understanding speech in noisy environments is particularly challenging for people with hearing loss (Gatehouse &

Noble, 2004), and it has recently been recognized that listeners who achieve the same level of intelligibility might exert different amounts of effort to achieve it (McGarrigle et al., 2014). Listening effort has been defined as the “deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a listening task” (“Framework for Understanding Effortful Listening”; Pichora-Fuller et al., 2016). Exerting a high level of effort throughout the day can lead to excessive fatigue that particularly affects people with hearing loss (Alhanbali et al., 2017; Hornsby et al., 2016). Therefore, identifying ways to alleviate listening effort in noisy environments could help to mitigate this challenge. Potentially, voice familiarity could reduce effort as well as improving intelligibility—for example, by allowing listeners to better resist interference from competing speech (Holmes & Johnsrude, 2020), thereby decreasing cognitive load. Although, how voice familiarity affects effort in noisy environments has not previously been tested.

Studies measuring listening effort have commonly focussed on how acoustic factors contribute to effort. Measures of listening effort include a listener's subjective, self-reported effort score (Alhanbali et al., 2017; Holmes, Folkeard et al., 2018), and physiological measures, such as pupillometry (Winn et al., 2015; Zekveld et al., 2011) and electroencephalography (EEG; e.g., Dimitrijevic et al., 2019; Obleser et al., 2012), which have been suggested as objective markers of effort. These self-reported and physiological measures of listening effort do not always fully align with one another (e.g., Alhanbali et al., 2019; Shields et al., 2023). For example, Alhanbali et al. (2019) took a variety of measures while participants listened to sentences in noise, including self-reported fatigue on a visual analog scale, pupil size, and alpha power during speech perception measured with EEG. They found that each of these measures showed good internal reliability (intraclass correlation coefficients  $> .70$ ), but they were only weakly correlated with one another, and loaded onto different factors in a factor analysis. Nevertheless, studies examining listening effort using various methods have consistently reported that acoustically degrading speech (e.g., Obleser et al., 2012; Winn et al., 2015) and decreasing the signal-to-noise ratio between speech and competing sounds (e.g., Wendt et al., 2016; Wiggins et al., 2025; Zekveld et al., 2011, 2014) increase effort, when tested across conditions in which the stimuli are at least partially intelligible.

Fewer studies have examined how nonacoustic factors contribute to effort. Nevertheless, Borghini and Hazan (2018) demonstrated that language proficiency contributes to effort when participants try to understand speech in background noise: They found lower pupil dilation (indicating less effort) when speech was in a listener's native language compared to when it was in a non-native language. Similarly Brown et al. (2020) found lower pupil dilation for native accents compared to unfamiliar, non-native accents. In addition, Koelewijn et al. (2015) examined how the certainty of a target talker's location affects self-reported effort

ratings and pupillometry measures. Their results showed lower pupil dilation and lower self-reported effort when talker location was fixed than variable, implying that predictability affects listening effort. Semantic context can also reduce listening effort, which has been demonstrated by both lower self-reported effort (Holmes, Folkeard et al., 2018) and lower pupil dilation (Winn, 2016) for conditions with greater predictability. Together, these findings suggest that nonacoustic factors can contribute to the amount of effort that listeners exert to understand speech.

While voice familiarity has not previously been tested as a way to reduce effort when listeners try to understand speech in noisy environments, one previous study (Biçer et al., 2023) examined how voice familiarity affects effort during a voice-cue discrimination task. In this task, participants were asked to determine which of three stimuli was the odd-one-out, based on differences in the fundamental frequency and vocal tract length of the voice. They assessed listening effort during this discrimination task using pupillometry. Biçer et al. (2023) examined how implicit training with a new voice affected voice-cue discrimination, by exposing participants to a new voice through passive listening, and comparing voice-cue discrimination for the trained voice with novel voices. Their results showed no differences in the accuracy of voice-cue discrimination between the trained and unfamiliar voices. Although they found some evidence of lower pupil dilation for familiar voices. Thus, these results imply that voice familiarity has potential to reduce effort, but this has never been studied during a speech intelligibility task.

Here, we aimed to examine how voice familiarity affects the effort that listeners exert during speech perception when competing speech is present, alongside improvements in performance. Similar to Holmes et al. (2021), we explicitly trained participants to become familiar with three new voices for different durations of time (speaking 88, 166, or 478 sentences during familiarization and training), so that we could study performance and effort during the initial stages of voice learning. We compared listening effort for trained and unfamiliar voices during speech-in-speech perception, using both self-reported effort and pupillometry. We hypothesized that trained voices would be associated with better performance, lower self-reported effort, and lower pupil dilation, compared to unfamiliar voices. In addition, we hypothesized that we might find differences among the three trained voices, with longer training leading to better performance and lower effort.

## Methods

### Transparency and Openness

The stimuli ([https://osf.io/xajkq/?view\\_only=0c3cef41e75e4c67bd16c32549e08dc3](https://osf.io/xajkq/?view_only=0c3cef41e75e4c67bd16c32549e08dc3)), data ([https://osf.io/b8gez/?view\\_only=2031708b23214165b08aea84ed108042](https://osf.io/b8gez/?view_only=2031708b23214165b08aea84ed108042)), and analysis scripts ([https://osf.io/b8gez/?view\\_only=2031708b23214165b08aea84ed108042](https://osf.io/b8gez/?view_only=2031708b23214165b08aea84ed108042)) are available on the Open Science Framework. The design and its analyses were not preregistered.

### Participants

We did not know in advance the sizes of expected voice familiarity effects on self-reported effort or pupil dilation. Therefore, we targeted a final sample size of 20, because this sample size was used by Winn (2016) to examine how pupil dilation is affected by semantic context in participants with normal hearing.

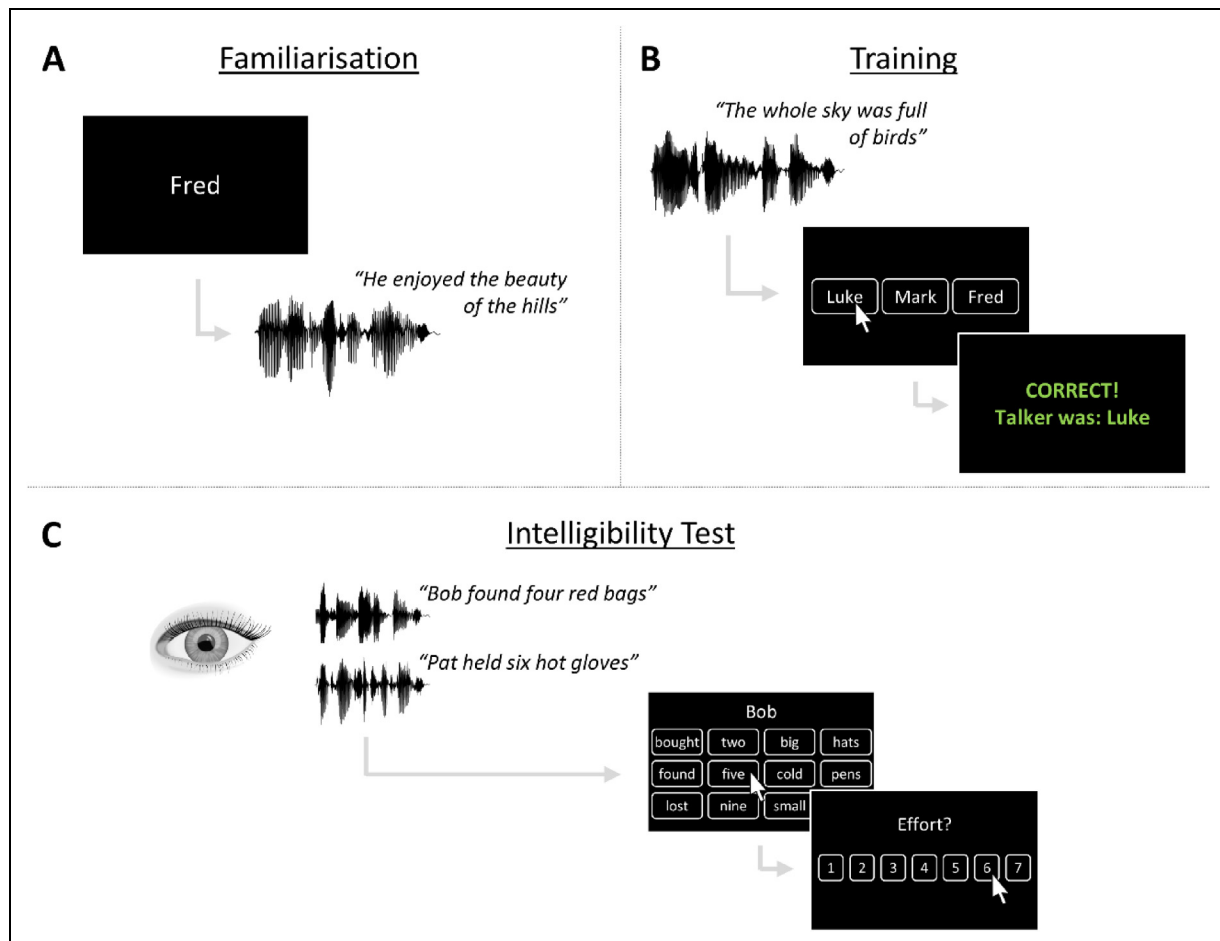
To examine the sensitivity of a sample size of 20, we conducted power analyses using G\*Power (version 3.1.9.2; Faul et al., 2007) with an alpha of .05. When using a two-tailed *t*-test to test a difference between two dependent means (such as a comparison between a familiar and unfamiliar voice), 20 participants provides 80% power to detect effect sizes of  $d_z \geq .66$  (i.e., moderate-sized effects). Familiar-voice effects on *speech intelligibility* have been reported to be much larger than this: for example, the familiar-voice intelligibility benefit had a size of  $d_z = .94$  in Holmes et al. (2021), and effects of  $d_z = .94$  are detectable with 98% power in a sample size of 20 (using a *t*-test to detect a difference between two dependent means). We thought it was plausible that effects of voice familiarity on self-reported effort and pupil dilation could be smaller than effects on intelligibility, but we anticipated that 20 participants would be an appropriate sample size, given our expectation of medium-to-large effect sizes.

As planned, we analyzed data from 20 participants. We recruited two additional participants who did not complete the study because the eye-tracker calibration failed. Participants were aged 18–34 years (median = 23.5 years, interquartile range = 5.5). Ten were male, nine were female, and one identified as nonbinary. All participants were native English speakers and had normal or corrected-to-normal vision. Two participants wore glasses throughout the experiment and one participant wore contact lenses. Participants reported no history of neurological issues, hearing loss, or hearing difficulties. They had average pure-tone thresholds (measured at octave frequencies between 5 and 8 kHz using a Resonance R07A Portable Screening Audiometer) better than 15 dB HL in each ear (mean = 3.3 dB HL, standard deviation = 3.3). Participants were instructed not to consume caffeine for a period of 6 h before starting the study, because caffeine can affect pupil responses (Winn et al., 2018).

The study gained ethical approval from the UCL Psychology and Language Sciences Local Research Ethics Committee. Participants provided written informed consent and were compensated for their time.

### Design

The study involved three main phases: (i) familiarization with three novel voices (10 sentences for each voice), (ii) voice-identification training with the same three voices (“Least



**Figure 1.** Schematics of the three phases of the study: (A) Familiarization, (B) training, and (C) speech-intelligibility test. Note that the first response screen for the intelligibility test contains only three rows of words for clarity, but eight rows were shown (i.e., displaying all word options) during the study.

Familiar": 78 sentences; "Moderately Familiar": 156 sentences; "Most Familiar": 468 sentences), and (iii) a speech-intelligibility test containing the three trained voices and two novel voices (Figure 1). The familiarization and training phases exposed participants to the same number of sentences as in Holmes et al. (2021)—for a total of 88, 166, and 478 sentences. Across participants, the assignment of the five voices to the familiar and unfamiliar conditions was counterbalanced, as was the assignment of sentences to voices. The study lasted approximately 3 h in total, including breaks. Participants were encouraged to take self-paced breaks between tasks and also between blocks of trials within each task.

In the intelligibility test, participants reported words from a target sentence in the presence of a competing sentence spoken by a different talker. Two factors were manipulated in the intelligibility test: the familiarity of the voice that spoke the target sentence (Most Familiar, Moderately Familiar, Least Familiar, or Unfamiliar), and the target-to-masker ratio (TMR). As in previous voice familiarity studies (Holmes et al., 2021; Holmes, Domingo et al., 2018; Zhu & Holmes,

2025), the TMR for each trial was either  $-6$  or  $+3$  dB, reflecting different levels of acoustic challenge. Based on previous studies (Domingo et al., 2020; Holmes et al., 2021; Holmes & Johnsrude, 2023; Holmes, Domingo et al., 2018),  $-6$  and  $+3$  dB TMR avoid ceiling and floor effects; in addition, having one positive and one negative TMR means that participants cannot use TMR as a cue to determine the target sentence.

### Apparatus

The acoustic stimuli were recorded in a sound-attenuating booth (IAC Acoustics UK Ltd), using a RØDE NT1-A microphone (The Freedman Group, Sydney) connected to a FireFace UC sound card (RME Audio Interfaces, Germany).

The experiment took place in a double-walled audiometric booth (120a Series, IAC Acoustics UK Ltd). Acoustic stimuli were delivered diotically through Sennheiser HD200 pro headphones. Participants were seated in a comfortable, adjustable chair, facing a Dell U2715H LCD monitor.

During the speech-intelligibility test, pupillometry was conducted with an Eyelink 1000 Plus (SR Research Ltd) eye-tracker, and participants were required to use a headrest. The headrest was positioned for each participant so that looking at the entire screen was comfortable. The illumination inside the audiometric booth was kept constant across participants.

The left eye was selected as the tracked eye for all participants. The pupil and corneal reflection thresholds were set manually for each participant, ensuring that no shadows, noise, or reflections were erroneously detected. The eye-tracker was calibrated for each participant at the beginning of the experiment using a 9-point sequence. Gaze fixations were manually accepted when they were stable. The calibrations were then validated. The calibration was only accepted when the maximum degree of spatial error during validation was less than .5 degrees. Otherwise, we repeated the calibration process until this criterion was met. The eye-tracker was set to measure pupil diameter. Recordings were made at either 1,000 Hz ( $N = 16$ ) or 500 Hz ( $N = 4$ ).

### Stimuli

The sentences were the same as those used in Holmes et al. (2021). Different sets of sentences were used for the familiarization and training phases compared to the test phase. For the familiarization and training phases, we used 354 meaningful sentences based on the sentence corpus developed by Rodd et al. (2005); for example, “The daisies began to grow quite soon.” For the speech-intelligibility test, sentences were taken from the Boston University Gerald (BUG) corpus (Kidd et al., 2008). Sentences from the BUG corpus each have five words, which follow the structure, “*Name verb number adjective noun*”; for example, “Pat lost four red bags.” The name was always fixed as either “Bob” or “Pat,” and each of the four remaining words had eight possible options (see Table 1). We selected 384 sentences from this corpus for the speech-intelligibility test. An advantage of using these closed-set matrix sentences for the speech-intelligibility test, rather than open-set sentences, is that participants are unable to guess the correct response based on semantic probability or the absence of semantic neighbors. In addition, the test requires participants to make four separate and unrelated responses, and to guess if uncertain, which eliminates the possibility that familiar voices are more intelligible because participants are more willing to guess words for familiar voices than for unfamiliar voices (i.e., a change in criterion or bias). In contrast, open-set tests that ask participants to report as many target words as they heard could result in different estimates of performance between familiar and unfamiliar voices, regardless of sensitivity, if participants feel less confident reporting words for unfamiliar voices, so report fewer words overall.

The sentences were recorded by five male, native English speakers aged 22–37 years (median = 31 years, interquartile range = 12). While recording the sentences, the speakers

**Table 1.** Words from the BUG Corpus That Were Used in the Speech-Intelligibility Test.

Name	Verb	Number	Adjective	Noun
Bob	bought	two	big	bags
Pat	found	three	blue	cards
	gave	four	cold	gloves
	held	five	hot	hats
	lost	six	new	pens
	saw	eight	old	shoes
	sold	nine	red	socks
	took	ten	small	toys

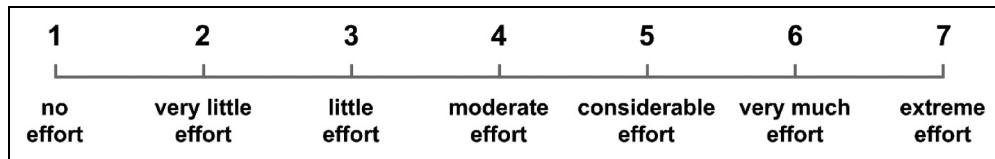
Each sentence contained five words: one name, verb, number, adjective, and noun.

were encouraged to use their natural speaking style, while keeping to a desired pace that was indicated by a video for each sentence (Holmes, 2018). The sentences were recorded in stereo at a sample rate of 44100 Hz, then converted to mono. After conversion, sentences were normalized to the same root-mean-square amplitude.

### Procedure

During the familiarization phase, participants were presented with 30 sentences spoken in three voices. We presented 10 sentences for each voice, which were randomly interleaved across trials. While each sentence was playing, a name was displayed on the screen, which participants were told to associate with the voice (Figure 1A). The three names were “Mark,” “Fred,” and “Luke,” which were pseudo-randomly paired with the three voices and counterbalanced across participants.

During the training phase, participants heard sentences from the same three voices as the familiarization phase. On each trial, participants heard a sentence and saw the three name options on the screen (Figure 1B). They were asked to identify the name that corresponded to the voice. They could respond at any time after the sentence had begun; although, the sentence always played in full, even if participants responded before the end of the sentence. This ensured that all participants received the same duration of training across the three voices. After the sentence had finished, participants received feedback about their response (“Correct” displayed in green font or “Incorrect” displayed in red font) and were presented with the correct name (e.g., “Talker was Fred”) regardless of whether they correctly identified the name. Of the three voices, one spoke 78 sentences (“Least Familiar”), another spoke 156 sentences (“Moderately Familiar”), and the third spoke 468 sentences (“Most Familiar”), corresponding to approximately 10, 20, and 60 min of training. The talker was selected pseudo-randomly on each trial. During the training phase, 351 sentences were used, which were each presented twice across the training phase, but in a different voice each time they were presented.



**Figure 2.** Visual depiction of the scale used for participants' self-reported effort scores.

Finally, participants completed the speech-intelligibility test (Figure 1C). On each trial, participants heard two BUG sentences simultaneously. One sentence started with “Bob” and the other started with “Pat.” The target sentence was defined by the first word of the sentence, which was “Bob” for one half of the test and “Pat” for the other half of the test, the order of which was counterbalanced across participants. The participant was reminded of the target word at the beginning of the speech-intelligibility test and at the half-way point, when the target word switched. The target voice could be one of the three trained (“familiar”) voices or one of two novel (“unfamiliar”) voices that participants had not heard during training. The masker voice was always one of the two novel voices: when the target was a trained voice, the masker was one novel voice for half of trials, and the other novel voice for the other half of trials (pseudo-randomly selected); when the target was a novel voice, the masker was the other novel voice. This gave rise to four different target-familiarity conditions: Unfamiliar, Least Familiar, Moderately Familiar, and Most Familiar. In total, the speech-intelligibility test contained 192 trials: 48 for each familiarity condition. Within each familiarity condition, the stimuli were presented equally often (24 trials each) at +3 dB and –6 dB TMR.

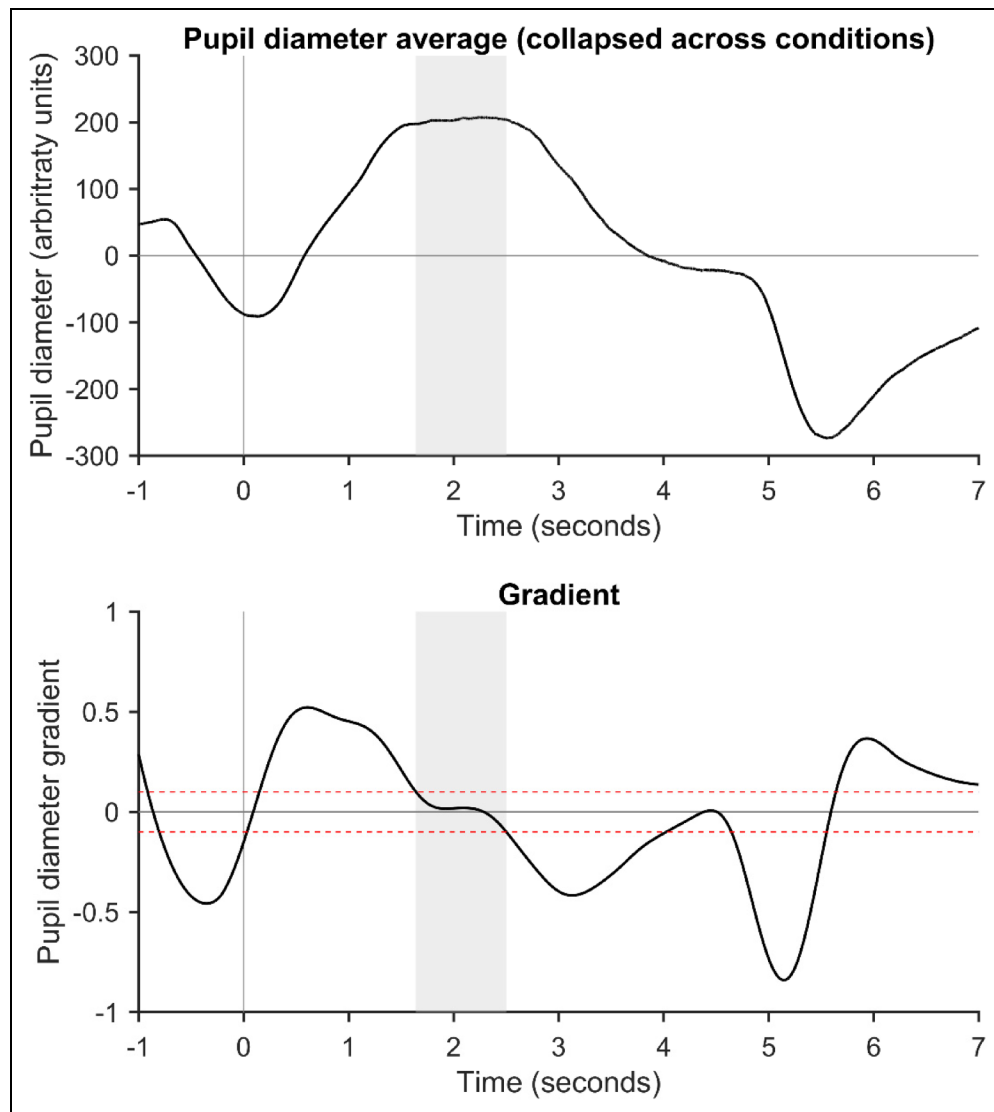
Participants' pupil responses were recorded throughout the speech-intelligibility test. At the start of each trial of the speech-intelligibility test, a fixation cross was presented for 1 s to stabilize the participant's gaze and establish a reliable baseline period. Next, the acoustic stimuli began while the fixation cross remained on the screen. The acoustic stimuli lasted approximately 1.5 s, and the first response screen appeared 2 s after the offset of the acoustic stimuli. This delay was implemented to separate pupil responses to acoustic stimuli from participants' responses. Participants were instructed to report the four words from the target sentence (which began with the target word, “Bob” or “Pat”) by choosing a word from each of four columns that were presented on the screen (Figure 1C). Each column contained eight words. Participants were required to select the four words from the target sentence by clicking one word from each column, in any order. After the participant had reported the words from the target sentence, they were presented with another response screen, which asked them to report the amount of effort that they had exerted during that trial. The scale ranged from 1 to 7, where 1 denoted “no effort” and 7 denoted “extreme effort” (as in Holmes, Folkeard et al., 2018). On each trial, participants were required to select a number from 1 to 7 (without

effort labels). At the beginning of the speech-intelligibility test, the experimenter gave participants a handout that visually depicted the scale and its associated labels, which is shown in Figure 2. In addition, the handout encouraged participants to use the full scale and explained the distinction between effort and difficulty (see Supplemental Material). Participants were encouraged to ask the experimenter questions about the effort scale, to ensure appropriate self-report. Participants were also given guidance about how to best-time their blinks during the test, with the aim of obtaining clean pupil data. There was a variable intertrial interval of 3.5–4.5 s.

Prior to the speech-intelligibility test, participants completed a practice session containing eight trials. The practice trials were all presented at a TMR of 0 dB, to ensure that participants learnt to identify the target sentence based on its first word, rather than using the level of the sentence to distinguish the target and masker sentences. Otherwise, practice trials were identical to those in the speech-intelligibility test.

### Preprocessing

The EyeLink files were converted with the Edf2Mat MATLAB Toolbox (designed and developed by Adrian Etter and Marc Biedermann at the University of Zurich), then were preprocessed using the Pupillometry Pipeliner toolbox (Pupl; Kinley & Levy, 2022) in MATLAB (version 2024b). Pupil data that were recorded at a sample rate of 1,000 Hz ( $N=16$ ) were first downsampled to 500 Hz, so that all data had a sample rate of 500 Hz. Blinks were identified using the pupillometry noise method (Hershman et al., 2018). Each blink, and a 50-ms window either side of the blink, was removed from the data. We then applied linear interpolation (using a maximum duration of 500 ms and a maximum gap of 1 standard deviation) to reduce the amount of missing data. To deal with high-frequency artifacts, we applied a Hann window moving average filter (with a width of 150 ms) to the continuous data before epoching. We applied baseline correction to the epoched data by subtracting the mean pupil diameter across the 1,000-ms prestimulus baseline period for each trial (corresponding to the time during which the fixation cross was displayed). Epochs were rejected if the proportion of missing data exceeded 18% (Lemerrier et al., 2014). On average, 9.74% of epochs were rejected per participant. Epochs were separated into 8 conditions (4 target-familiarity conditions  $\times$  2 TMR conditions).



**Figure 3.** Upper panel: Grand average pupil diameter ( $N = 20$ ), collapsed across target familiarity and TMR conditions. Only Unrejected epochs are included in the average. Lower panel: Gradient of the pupil diameter time course. Red horizontal lines show  $\pm .1$ , which were the boundaries used to identify parts of the time course across which there was little change in pupil diameter. The peak time window was defined as the period where pupil diameter was greatest (upper panel) and the gradient was within  $\pm .1$  (lower panel). The gray shaded area shows the selected peak time window, which was used for subsequent statistical analyses.

To compare pupil dilation across conditions, we calculated the mean pupil diameter across trials, which has been commonly used as a dependent variable in previous pupillometry studies (Kinley & Levy, 2022; Kret & Sjak-Shie, 2019; Winn et al., 2018). We aimed to measure pupil diameter over the portion of the trial for which pupil dilation was maximal, although we wanted to avoid picking a single peak value, given that the exact peak is highly sensitive to noise. To this end, we defined a time window around the peak using a novel method. Specifically, we looked at the gradient of the average pupil response (i.e., the change in pupil diameter) across all epochs, averaged across participants (Figure 3). We defined the maximal pupil dilation time window as the time during the trial at which pupil diameter

was at its greatest, and during which the gradient (i.e., the change in diameter) was within  $\pm .1$ . Under this calculation, the difference in pupil diameter between adjacent samples was  $< .1$ . This provided a time window for pupil analysis at 1.6407–2.5011 s after the onset of acoustic stimuli. We calculated the mean pupil diameter for each condition across this time window.

### Analyses

We conducted analyses using RStudio (version 2024.12.1; Build 563) with R version 4.5.0.

To evaluate accuracy during the training phase, we calculated the proportion of trials for which participants correctly

identified the voice. Shapiro–Wilk tests showed that the training data were not normally distributed ( $p \geq .027$ ), so we used a Friedman rank sum test to compare accuracy across the three voices that were trained for different lengths of time (Least Familiar, Moderately Familiar, and Most Familiar). We used Wilcoxon signed rank tests to compare pairs of adjacent Familiarity conditions (i.e., Least Familiar compared with Moderately Familiar, and Moderately Familiar compared with Most Familiar). To carry out these analyses, we used the `friedman.test()` and `wilcox.test()` functions from the R “stats” package.

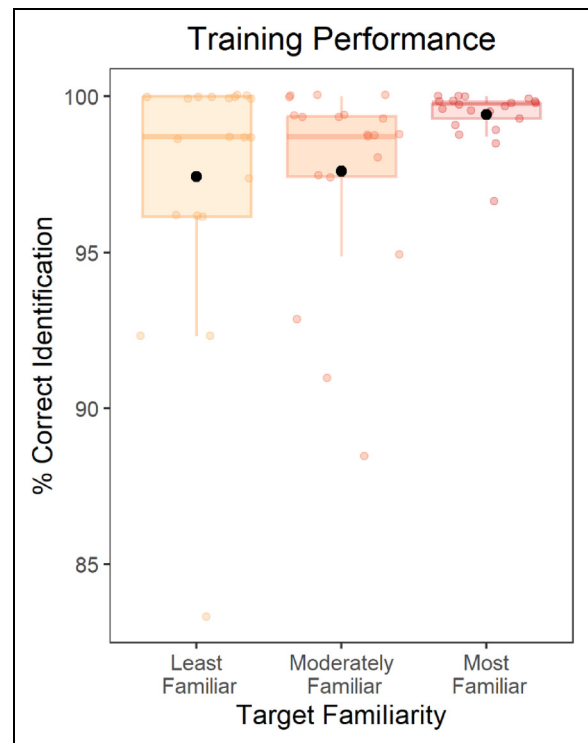
For the speech-intelligibility test, we were interested in behavioral and pupillometry measures. We calculated speech intelligibility as the percentage of trials for which participants reported all four words from the target sentence correctly. We calculated self-reported effort as the mean of participants’ effort ratings across trials. For pupil dilation, we calculated the mean pupil diameter across the 1.6407–2.5011 s time window. None of the data from the speech-intelligibility test violated normality assumptions, as assessed by Shapiro–Wilk tests and visual inspection of the data. We evaluated the effects of Target Familiarity (Unfamiliar, Least Familiar, Moderately Familiar, Most Familiar) and TMR (–6 dB, +3 dB) on each measure using two-way repeated-measures ANOVAs. We ran five planned contrasts, to compare each of the familiar-target conditions (Least Familiar, Moderately Familiar, and Most Familiar) with the unfamiliar condition, and to compare pairs of adjacent familiar-target conditions (i.e., Least Familiar compared with Moderately Familiar, and Moderately Familiar compared with Most Familiar). We used the `aov_ez()` function from the R “afex” package for the ANOVAs, and we used the `emmeans()` and `contrast()` functions from the R “emmeans” package for planned contrasts. Where Mauchley’s test of sphericity gave a significant result, we report Greenhouse–Geisser corrected  $p$ -values.

To examine whether self-report and pupillometry measures of effort were differentially sensitive to effects of TMR and Target Familiarity, we converted the mean effort ratings and mean pupil diameter for each participant into  $z$ -scores, and entered them into a three-way repeated-measures ANOVA. The three factors were Measurement (Self-report or Pupil), Target Familiarity, and TMR. For this analysis, we were primarily interested in interactions with Measurement.

## Results

### Training

Accuracy in the training phase was high overall (Figure 4) and participants performed close to ceiling. Numerically, accuracy was highest for the Most Familiar voice (mean = .99, standard deviation [sd] = .08), followed by the Moderately Familiar voice (mean = .98, sd = .15), and was lowest for the Least Familiar voice (mean = .97, sd = .16).

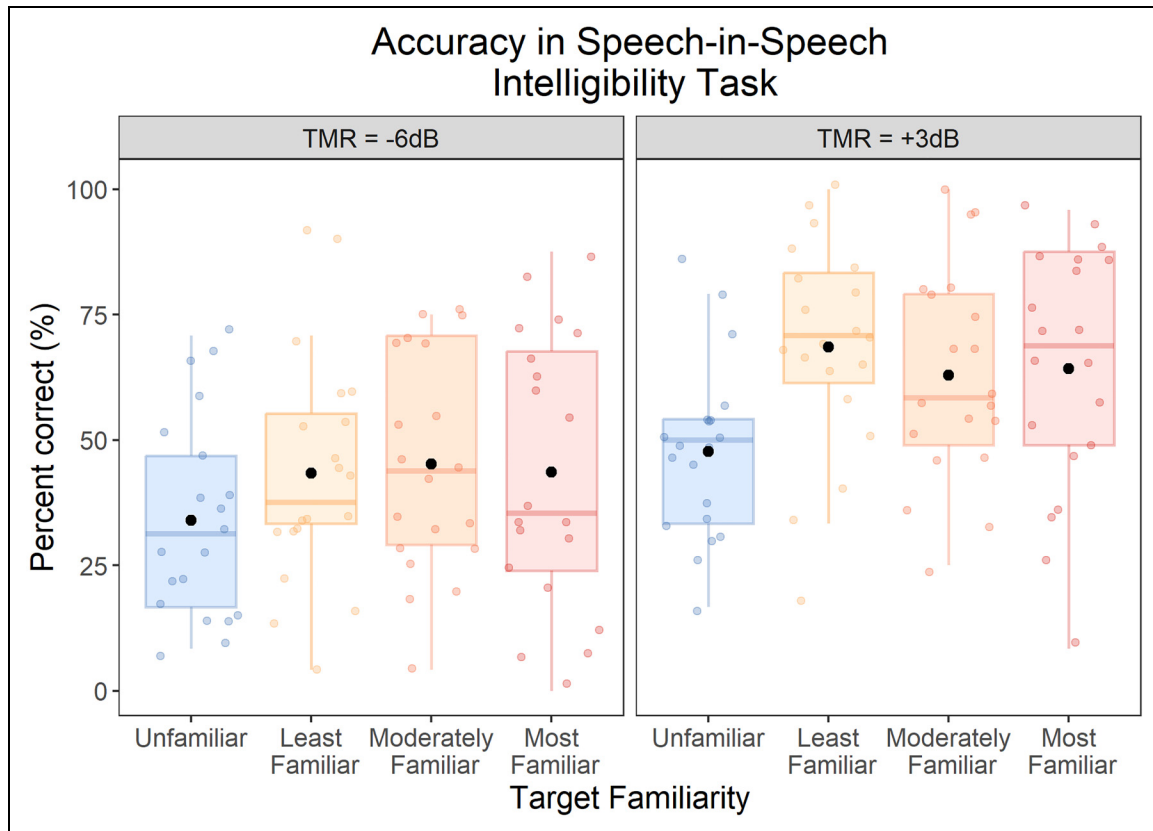


**Figure 4.** Accuracy of voice identification for the three talkers in the training phase. Each box displays the interquartile range, and the thick horizontal lines display the medians. Colored dots illustrate data from individual participants, and the black dots illustrate the means across participants.

A Friedman rank sum test showed a significant effect of familiarity on accuracy,  $\chi^2(2) = 6.77$ ,  $p = .034$ . Wilcoxon signed rank tests showed a significant difference between the Most Familiar and Moderately Familiar conditions ( $V = 13.0$ ,  $p = .001$ ), and no significant difference between the moderately familiar and least familiar conditions ( $V = 74.5$ ,  $p = .65$ ). Thus, accuracy was high for all three trained voices, but was better for the Most Familiar voice than for the two voices that participants heard for shorter durations during training.

### Speech-Intelligibility Test: Accuracy

Figure 5 displays the accuracy results. We found a significant main effect of TMR,  $F(1, 19) = 81.17$ ,  $p < .001$ ,  $\eta_p^2 = .81$ , with better performance at +3 than –6 dB TMR. We also found a significant main effect of Target Familiarity,  $F(3, 57) = 4.79$ ,  $p = .005$ ,  $\eta_p^2 = .20$ . Planned comparisons, comparing each of the familiar conditions with the unfamiliar conditions, revealed a significant familiar-voice benefit to intelligibility when the target was the Least Familiar voice (mean benefit of 15.1%, SE = 4.45),  $t(19) = 3.39$ ,  $p = .003$ ,  $d_z = .76$ , 95% CI: [.25, 1.25], the Moderately Familiar voice (mean benefit of 13.2%, SE = 4.69),  $t(19) = 2.82$ ,  $p = .011$ ,  $d_z = .63$ , 95% CI: [.14, 1.11], and the Most Familiar voice (mean benefit



**Figure 5.** Accuracy for the speech-intelligibility test, by target familiarity and TMR conditions. Each box displays the interquartile range, and the thick horizontal lines display the medians. Colored dots illustrate data from individual participants, and the black dots illustrate the means across participants.

of 13.0%,  $SE = 4.64$ ,  $t(19) = 2.80$ ,  $p = .011$ ,  $d_z = .63$ , 95% CI: [.14, 1.10]. There were no significant differences between the Least Familiar and Moderately Familiar conditions,  $t(19) = .41$ ,  $p = .69$ ,  $d_z = .09$ , 95% CI: [−.35, .53], or between the Moderately Familiar and Most Familiar conditions  $t(19) = .05$ ,  $p = .96$ ,  $d_z = .01$ , 95% CI: [−.43, .45].

There was a significant interaction between familiarity and TMR,  $F(3, 57) = 3.28$ ,  $p = .027$ ,  $\eta_p^2 = .15$ . To examine this interaction, we conducted a one-way ANOVA to examine the effect of Target Familiarity at each TMR. We found a significant effect of Target Familiarity at +3 dB TMR,  $F(3, 57) = 7.41$ ,  $p < .001$ ,  $\eta_p^2 = .28$ , but not at −6 dB TMR,  $F(3, 57) = 2.06$ ,  $p = .12$ ,  $\eta_p^2 = .10$ . Thus, the effect of familiarity was confined to the more favorable TMR condition.

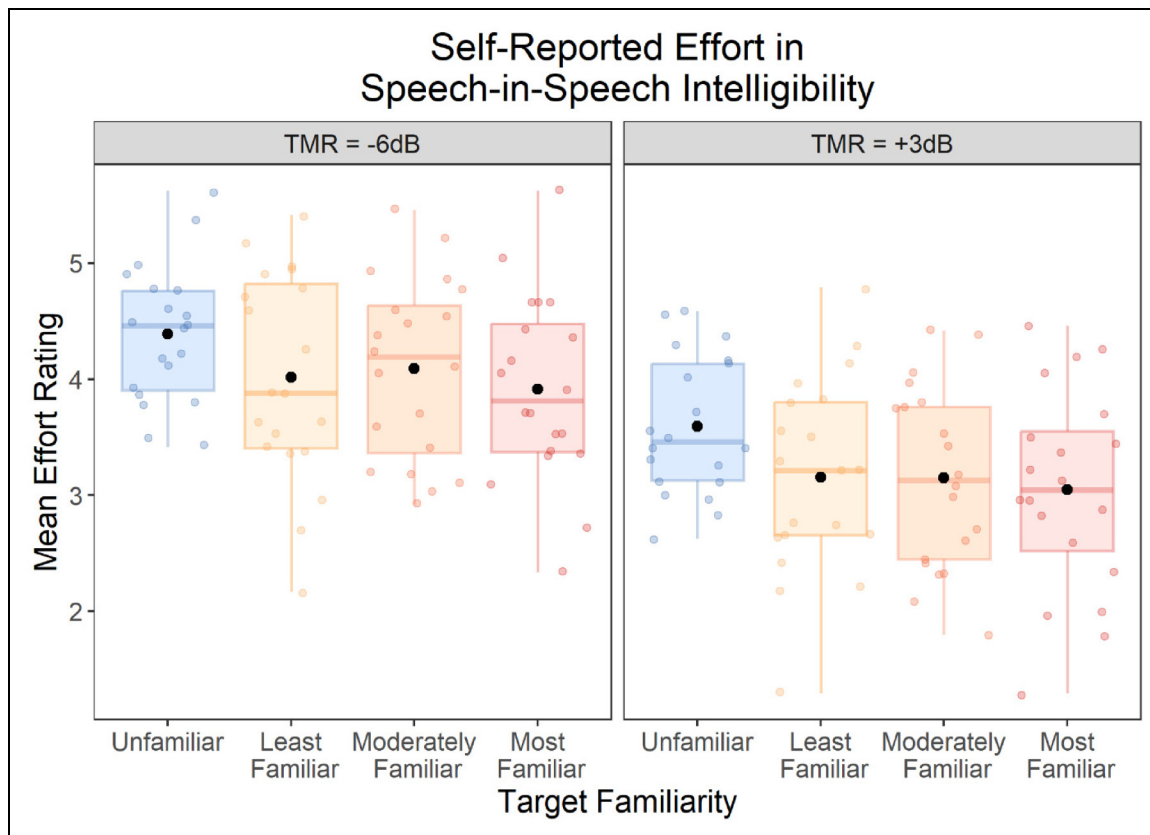
To follow-up on the significant difference among Target Familiarity conditions at +3 dB TMR, we ran post hoc tests with Bonferroni correction. All three familiar conditions differed from the unfamiliar condition [Most Familiar:  $t(19) = 3.38$ ,  $p = .016$ ,  $d_z = .76$ , 95% CI: .25, 1.25; Moderately Familiar:  $t(19) = 3.17$ ,  $p = .025$ ,  $d_z = .71$ , 95% CI: .21, 1.19; Least Familiar:  $t(19) = 4.30$ ,  $p = .002$ ,  $d_z = .96$ , 95% CI: .42, 1.49]. There were no significant differences between the Least Familiar and Moderately Familiar conditions,  $t(19) = 1.27$ ,  $p = 1.00$ ,  $d_z = .28$ , 95% CI: [−.17, .73], or between the

Moderately Familiar and Most Familiar conditions  $t(19) = .28$ ,  $p = 1.00$ ,  $d_z = .06$ , 95% CI: [−.38, .50].

### Speech-Intelligibility Test: Self-Reported Effort

Self-reported effort ratings during the speech-intelligibility test are displayed in Figure 6. We found a significant main effect of TMR,  $F(1, 19) = 62.01$ ,  $p < .001$ ,  $\eta_p^2 = .77$ , with lower self-reported effort at +3 than −6 dB TMR. There was also a significant main effect of Target Familiarity,  $F(3, 57) = 3.88$ ,  $p = .014$ ,  $\eta_p^2 = .17$ . Planned contrasts revealed significantly less self-reported effort in all three familiar-target conditions when compared to the Unfamiliar condition [Least Familiar:  $t(19) = 2.78$ ,  $p = .012$ ,  $d_z = .62$ , 95% CI: [.13, 1.09]; Moderately Familiar:  $t(19) = 2.34$ ,  $p = .030$ ,  $d_z = .52$ , 95% CI: [.05, .99], and Most Familiar:  $t(19) = 2.93$ ,  $p = .009$ ,  $d_z = .66$ , 95% CI: [.16, 1.13]]. However, there were no significant differences between the Least Familiar and Moderately Familiar conditions,  $t(19) = .25$ ,  $p = .81$ ,  $d_z = .05$ , 95% CI: [−.38, .49], or between the Moderately Familiar and Most Familiar conditions  $t(19) = .90$ ,  $p = .38$ ,  $d_z = .20$ , 95% CI: [−.24, .64].

There was no significant interaction between TMR and Target Familiarity,  $F(3, 57) = .29$ ,  $p = .83$ ,  $\eta_p^2 = .02$ , indicating



**Figure 6.** Self-reported effort scores (possible range: 1–7) on the speech-intelligibility test, by target familiarity and TMR. Each box displays the interquartile range, and the thick horizontal lines display the medians. Colored dots illustrate data from individual participants, and the black dots illustrate the means across participants.

that the effect of Target Familiarity on self-reported effort did not differ between the two TMR conditions.

### Speech-Intelligibility Test: Pupillometry

Pupil time courses, separated by TMR and Target Familiarity conditions, are displayed in Figure 7. Figure 8 shows the averages across the peak time window.

Comparing mean pupil diameter across conditions, there was a significant main effect of Target Familiarity,  $F(3, 57) = 3.47$ ,  $p = .030$ ,  $\eta_p^2 = .15$ . Planned contrasts revealed significantly smaller pupil diameter (consistent with less effort) in the Most Familiar condition than the Unfamiliar condition,  $t(19) = 2.23$ ,  $p = .038$ ,  $d_z = .50$ , 95% CI: [.03, .96]. However, there were no significant differences in pupil diameter between the Unfamiliar condition and either the Moderately Familiar,  $t(19) = .82$ ,  $p = .421$ ,  $d_z = .18$ , 95% CI: [−.26, .62], or the Least Familiar,  $t(19) = .43$ ,  $p = .671$ ,  $d_z = .10$ , 95% CI: [−.34, .53], conditions. In addition, pupil diameter was significantly smaller in the Most Familiar than the Moderately Familiar condition,  $t(19) = 2.22$ ,  $p = .039$ ,  $d_z = .50$ , 95% CI: [.02, .96], but there was no significant difference between the

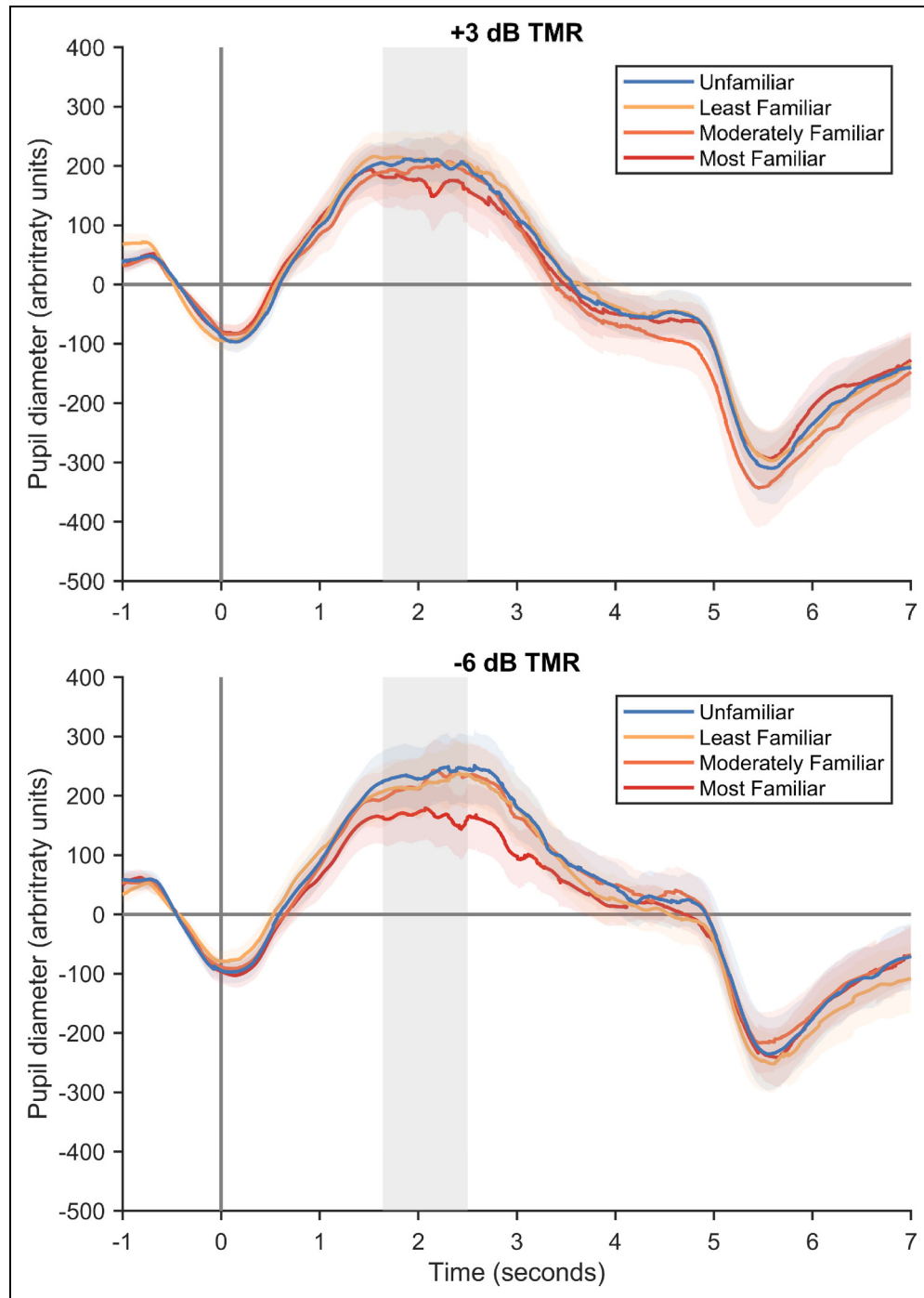
Moderately Familiar and Least Familiar conditions,  $t(19) = .42$ ,  $p = .677$ ,  $d_z = .09$ , 95% CI: [−.35, .53].

The main effect of TMR on mean pupil diameter was not significant,  $F(1, 19) = 2.22$ ,  $p = .153$ ,  $\eta_p^2 = .11$ , and there was no significant interaction between TMR and Target Familiarity,  $F(3, 57) = .39$ ,  $p = .758$ ,  $\eta_p^2 = .02$ .

### Relationships among Effort Measures

When we entered both effort measures (self-reported effort and pupil diameter) into an analysis together, we found significant main effects of TMR,  $F(1, 19) = 70.2$ ,  $p < .001$ ,  $\eta_p^2 = .79$ , and Target Familiarity,  $F(3, 57) = 5.07$ ,  $p = .003$ ,  $\eta_p^2 = .21$ . There was no significant interaction between TMR and Target Familiarity,  $F(3, 57) = .26$ ,  $p = .851$ ,  $\eta_p^2 = .01$ .

Given that the results from each measure were converted to z-scores, there was no significant main effect of Measurement,  $F(1, 19) < .01$ ,  $p = 1.0$ ,  $\eta_p^2 = 0.0$ , as expected. There was no significant interaction between Measurement and Target Familiarity,  $F(3, 57) = 2.14$ ,  $p = .106$ ,  $\eta_p^2 = .10$ . However, we found a significant interaction between Measurement and TMR,  $F(1, 19) = 40.75$ ,  $p < .001$ ,  $\eta_p^2 = .68$ ,

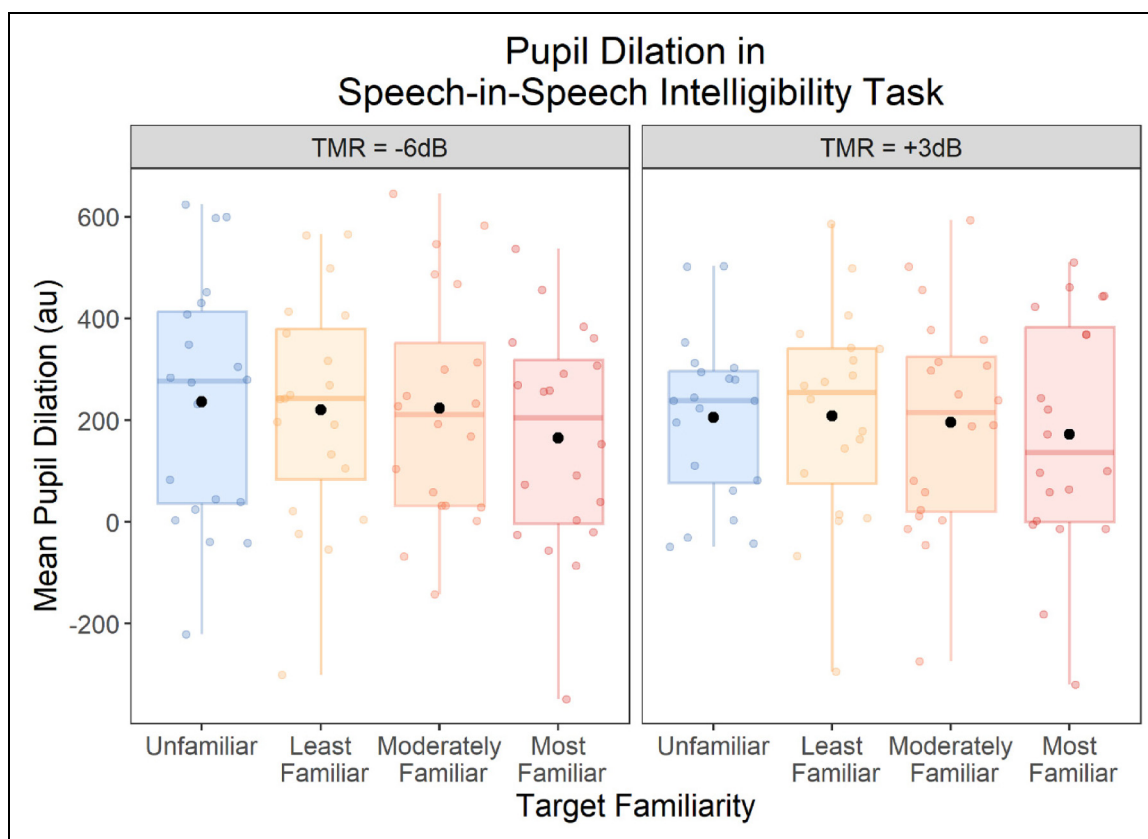


**Figure 7.** Grand average pupil diameter ( $N = 20$ ) across target familiarity and TMR (+3 and  $-6$  dB) conditions. Only unrejected epochs are included in the averages. Error bars show the standard error of the mean. Gray shaded areas indicate the peak time window that was used for statistical analyses.

showing that self-reported effort ratings were more sensitive to the difference between the two TMRs than was pupil dilation. The three-way interaction (Measurement  $\times$  TMR  $\times$  Target Familiarity) was not significant,  $F(3, 57) = .47$ ,  $p = .702$ ,  $\eta_p^2 = .02$ .

## Discussion

Overall, we found that training participants to become familiar with new voices improves speech intelligibility and reduces the amount of effort exerted during speech-in-speech perception. Both effort measures (self-report and pupil



**Figure 8.** Mean pupil diameter across the peak time window (1.6407–2.5011 s) during the speech-intelligibility test. Data are plotted by target familiarity and TMR. Each box displays the interquartile range, and the thick horizontal lines display the medians. Colored dots illustrate data from individual participants, and the black dots illustrate the means across participants.

dilation) indicated lower effort for the Most Familiar voices (that were trained for the longest) compared to unfamiliar voices, demonstrating that ~1 h of explicit training is sufficient to reliably reduce effort.

There are several possible reasons why voice familiarization training led to lower effort. First, a familiar talker's voice may be more predictable than an unfamiliar talker's voice, which may result in better intelligibility and lower effort. The finding that greater predictability results in lower pupil dilation has been reported for simple tone sequences (Milne et al., 2021; Qiyuan et al., 1985) and for speech stimuli (Koelewijn et al., 2015). For example, Koelewijn et al. (2015) found lower pupil dilation and lower self-reported effort when a talker's location was fixed compared to variable across trials. They also found that fixing the talker across trials reduced self-reported effort, although they found no effects of talker variability on pupil dilation. Possibly, the effect of talker familiarization that we studied here generates stronger predictions than does trial-to-trial constancy in talker identity, which could explain why we found a significant effect on pupil dilation, whereas Koelewijn et al. (2015) did not. This explanation is consistent with our finding that pupil dilation was only significantly lower for the Most Familiar voices compared to unfamiliar voices, whereas the two familiar voices

to which participants were exposed for shorter lengths of time (Moderately Familiar and Least Familiar) showed no significant differences in pupil dilation when compared to unfamiliar voices.

A second possible explanation for why voice familiarization training led to lower effort is that it allows listeners to better resist interference from a competing talker, which could reduce the cognitive load of listening to speech in competing speech. Under this explanation, familiar voices do not need to be more predictable than unfamiliar voices. Instead, reduced interference could be explained by quicker or more efficient processing of speech spoken by familiar talkers, leaving more cognitive resources for processing competing speech. This explanation is more consistent with the behavioral findings of Holmes and Johnsrude (2020), who showed that the familiar-voice benefit for naturally familiar voices scales with the extent of linguistic similarity of the masker.

It is worth noting that the effect of familiarity on pupil dilation we observed here is in the opposite direction than would be expected based on recollection (e.g., Papesh et al., 2012; Võ et al., 2008) or emotional arousal (e.g., Bradley et al., 2008) accounts, which would predict *greater* pupil dilation for familiar than unfamiliar voices. For example, familiar pieces of music have been associated with greater pupil

dilation than unfamiliar pieces of music during passive listening (Jagiello et al., 2019). In the current task, participants' focus was on the content of speech spoken by familiar and unfamiliar talkers, rather than the voices themselves, which may be why we observed a different pattern of pupil responses. Given that the Most Familiar voice elicited significantly lower pupil responses than unfamiliar voices, our results are most consistent with an explanation based on reduced listening effort during challenging speech perception. This pattern of results is consistent with Biçer et al.'s (2023) finding that familiar voices are associated with lower pupil dilation during a voice-cue discrimination task. It is also consistent with findings that listening to a native accent is associated with lower pupil dilation than listening to an unfamiliar, non-native accent during speech perception (Brown et al., 2020).

The degree of familiarity appeared to have different effects on self-report and pupil dilation measures. Self-reported effort was lower for all three trained voices compared to unfamiliar voices, with no difference between the three voices that participants heard for different lengths of time during training. These findings indicate that even relatively short (~10 min) exposure to a voice can provide measurable benefits for self-reported effort. Whereas voice familiarity only affected pupil dilation for the voice that was trained for the longest, and we found no evidence for reduced pupil dilation for the two voices that were trained for shorter lengths of time, when compared to unfamiliar voices. These results appear to suggest that the two putative measures of effort—self-report and physiological—may be differentially sensitive to the extent of voice familiarity, with listeners being able to notice that they are exerting less effort even for voices they have been trained on for approximately 10 min, whereas physiological measures do not show differences unless a voice has been trained for approximately 1 h. Although, it is important to note that the interaction between Target Familiarity and Measurement was not significant. Therefore, rather than reflecting a real difference in how voice familiarity is reflected in the two measures, the apparent differences could instead reflect greater noise for the effect of voice familiarity on pupil recordings compared to the effect of voice familiarity on self-report scores. Alternatively, there may be a real difference in how voice familiarity is reflected in the two measures, but it has a sufficiently small effect size that is difficult to detect with a sample size of 20 participants.

Despite finding no significant interaction between Familiarity and Measurement, we did find a significant interaction between TMR and Measurement, which indicates that the two effort measures are differentially sensitive to TMR. TMR had a significant effect on self-reported effort, but not on pupil diameter. We found a trend in the expected direction for pupil diameter (Zekveld et al., 2018), but the difference between the two TMRs was not significant. Thus, it is possible that self-reported effort is more sensitive than pupil dilation to differences in TMR. Although, another possible

explanation is that effects on self-reported effort could be affected by response bias—for example, due to demand characteristics (Orne, 1962). In other words, participants may not have experienced less effort in these conditions, but reported less effort because they noticed differences between conditions and believed they should report effort differently across conditions. Or, participants could have been reporting difficulty or perceived performance, rather than effort (McGarrigle et al., 2014). In our study, the researcher explained to participants the differences between effort, difficulty, and performance at the beginning of the speech-intelligibility test, thereby reducing the likelihood that listeners' ratings of effort were conflated with difficulty or performance. Nevertheless, the inherent nature of a self-report measure makes it difficult to fully rule out these possibilities.

A similar pattern of results, that self-reported effort is more sensitive than pupil diameter to TMR, was also reported by Wendt et al. (2016); although, the opposite pattern was found by McGarrigle et al. (2017). Therefore, the effects may be sensitive to the specific TMRs and types of competing sounds used (Johns et al., 2024; Zekveld et al., 2018). Notably, the difference between the two TMRs we used here—of +3 dB and −6 dB—is smaller than those used in some previous studies (e.g., McGarrigle et al., 2017), which could explain differences in the results. Regardless, it is not too surprising that these two measures gave different results, given that physiological and self-report measures of listening effort often diverge (McGarrigle et al., 2014). Possibly, these measures may index different underlying cognitive or neurobiological processes, consistent with multidimensional views of listening effort (Alhanbali et al., 2019; Shields et al., 2023). For example, they may index different components associated with listening effort, such as general arousal, engagement in the specific task, evaluating task demands, allocating resources, or the motivation to overcome demands (Pichora-Fuller et al., 2016; Saderi et al., 2021); or, they be differentially sensitive to the type of demand that underpins increased effort—for example, increased effort may arise from demands on attention, working memory, or speed of processing, which have overlapping but dissociable neurobiological substrates (Friedman et al., 2008; Gajardo-Vidal et al., 2024; Mendoza-Halliday et al., 2024). Further work comparing putative listening-effort measures under different scenarios is needed to elucidate the reasons for the differences observed in the current and previous studies.

Our finding that trained voices were more intelligible than unfamiliar voices (Figure 5) replicates previous work (Case et al., 2018; Holmes et al., 2021; Kreitewolf et al., 2017; Levi, 2015; Levi et al., 2011; Nygaard et al., 1994; Nygaard & Pisoni, 1998; Yonan & Sommers, 2000; Zhu & Holmes, 2025). Although, we found different effects of familiarity with the target voice at the two TMRs: We found a significant effect at +3 dB, but not at −6 dB. This finding contrasts with previous assumptions that familiarity has the greatest effect on

speech intelligibility in less favorable acoustic conditions (e.g., Johnsrude et al., 2013; Yonan & Sommers, 2000; Zhu & Holmes, 2025). Yet, it is consistent with the results of another study (in preparation) using the same TMRs, in which we also found the biggest familiar-voice benefit at +3 dB TMR and a smaller (but nevertheless significant) effect of familiarity at -6 dB TMR. The reason why familiar-voice effects are strongest at different TMRs across studies is unclear; although, it is unlikely to be explained by poor performance in the task at the lower TMR, because participants were still performing well above chance (0.02%) for unfamiliar voices at -6 dB TMR (~35%). Another possible explanation is that acoustic masking at the lower TMR obscures cues to familiarity, although this explanation is also unlikely given that previous studies have found familiar-voice benefits at -6 dB TMR (e.g., Domingo et al., 2020; Holmes et al., 2021; Johnsrude et al., 2013; Zhu & Holmes, 2025). Also, interestingly, in contrast to the effects on speech intelligibility, we found no evidence that magnitude of the reduction in self-reported effort or pupil dilation for familiar voices differed between the two TMRs. Ultimately, future studies that test a wider variety of TMRs are needed to examine this question further.

Nevertheless, the overall magnitude of the familiar-voice benefit that we observed here (13.8%, on average) is consistent with previous research: the present study replicates the voice training paradigm of Holmes et al. (2021), who found a speech-intelligibility benefit of 7%–15% (9.31%, on average). The magnitude is also comparable to that reported in studies which utilize the same task and type of masker to investigate the familiar talker advantage, but where listeners hear naturally familiar voices (Domingo et al., 2020; Johnsrude et al., 2014). This result reinforces the idea that familiar voices are more intelligible than unfamiliar voices, even following short durations of lab-based training. This effect is not simply due to familiarity with the sentence materials themselves, as we used different sentence materials during the training and test phases of the experiment—meaning that the benefit reflects familiarity with a voice that generalizes to new sentence materials.

Unlike Holmes et al. (2021), we found no significant differences in speech intelligibility between the three voices that were trained for different lengths of time. Holmes et al. (2021) found a similar-magnitude benefit for the two voices that were trained for the shortest lengths of time, and a significantly bigger benefit for the voice that was trained for the longest, but we did not replicate this effect in the current study. One possible reason may be that we had a smaller sample size (20 compared to 50 participants), which would be less sensitive to differences in the magnitude of the familiar-voice benefit among the three voices. Despite not finding a difference in speech intelligibility between the three voices, we did find a difference in pupil diameter, which provides evidence that the voice trained for the longest was perceived differently to the voices that were trained for shorter durations.

The current study is the first, to our knowledge, that has examined whether voice familiarity reduces effort when listeners try to understand speech in competing speech. A previous study by Biçer et al. (2023) studied how voice training effects pupil dilation during a voice-cue discrimination task, in which participants discriminated differences in fundamental frequency and vocal tract length among three-syllable consonant-vowel stimuli. Similar to our results, they found that trained voices elicited less pupil dilation than unfamiliar voices, although their effects were subtle. They did not find a significant effect on mean pupil dilation or other summary measures (e.g., peak dilation or latency) and they only found differences in a nonlinear Generalized Additive Mixed Models analysis. In that analysis, they also only found differences in pupil responses when stimuli were vocoded and no differences when the stimuli were not vocoded. Our results may help to explain this subtle effect, as we only found significant effects on mean pupil dilation for the voice that was explicitly trained for ~1 h and not for voices that were trained for shorter lengths of time. Biçer et al. (2023) trained voices implicitly by asking participants to listen to an audiobook lasting 30 min. Possibly, if they had used longer durations of training, and explicit training with feedback, they may have observed stronger effects on pupil responses.

Some researchers have argued that apparent reductions in effort may be conflated with performance (e.g., Moore & Picou, 2018). In other words, when participants are asked about effort, they are really answering a question about how well they think they performed. Similarly, reduced pupil dilation could reflect better speech intelligibility when performance is not matched between conditions. However, there are now many reported cases where performance and effort measures do not match, and are dissociable (Winn & Teece, 2021). Here, we reduced the likelihood that participants were self-reporting performance by explaining the difference to participants before they began (see Supplemental Material). In addition, we found distinct patterns of results across performance, self-reported effort, and pupil diameter measures: For example, we found a significant effect of TMR on accuracy and self-reported effort, but not on pupil diameter, and we found a significant interaction between TMR and Familiarity for accuracy, but not for self-reported effort or pupil diameter. Together, these results suggest that the measures are not entirely conflated with one another.

Our finding that voice training improves intelligibility and reduces effort is promising for real-world interventions designed to improve speech understanding and reduce fatigue when listening in noisy environments. Noisy environments are challenging for many people, although may be particularly difficult for older adults and people with hearing loss (Duquesnoy, 1983; Marrone et al., 2008), for whom exerting a high level of listening effort throughout the day can lead to excessive fatigue (Alhanbali et al., 2017; Hornsby et al., 2016). If individuals could train themselves on the voices that they are likely to hear in noisy, everyday situations, it



could make listening to speech more accurate and less effortful, allowing them to communicate successfully in noisy environments and reach the end of the day feeling less fatigued.

In conclusion, our results underscore the importance of voice familiarity in speech perception, showing advantages for both speech intelligibility and listening effort after relatively short ( $\leq 1$  h) durations of lab-based training. In addition, our results demonstrate differing sensitivities of two widely used measures of listening effort—self-report scores and pupil dilation.

## Acknowledgements

Emma Holmes received funding from the Experimental Psychology Society (SG0122-04). Harriet J. Smith is funded by Wellcome (226582/Z/22/Z). The authors thank Andrea Caso, Milei Shi, and Carla Van Der Goot for their help preparing stimuli.

## ORCID iDs

Harriet J. Smith  <https://orcid.org/0000-0003-4314-6571>  
Emma Holmes  <https://orcid.org/0000-0002-0314-6588>

## Ethical Considerations

This study was approved by the UCL Psychology and Language Sciences Local Research Ethics Committee (approval no. SHaPS-2022-EH-035) on January 09, 2023.

## Consent to Participate

Participants gave written consent before starting the experiment.

## Author Contributions

Freja Baxter: Conceptualization, investigation, formal analysis, and writing—original draft. Harriet J. Smith: Conceptualization, software, visualization, writing—editing, and supervision. Emma Holmes: Conceptualization, software, formal analysis, visualization, writing—original draft, supervision, and funding acquisition.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Experimental Psychology Society (grant number SG0122-04) and Wellcome (grant number 226582/Z/22/Z). For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

## Declaration of Conflicting Interests

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

## Data Availability Statement

The stimuli ([https://osf.io/xajkq/?view\\_only=0c3cef41e75e4c67bd16c32549e08dc3](https://osf.io/xajkq/?view_only=0c3cef41e75e4c67bd16c32549e08dc3)), data ([https://osf.io/b8gez/?view\\_only=2031708b23214165b08aea84ed108042](https://osf.io/b8gez/?view_only=2031708b23214165b08aea84ed108042)), and analysis scripts ([https://osf.io/b8gez/?view\\_only=2031708b23214165b08aea84ed108042](https://osf.io/b8gez/?view_only=2031708b23214165b08aea84ed108042)) are available on the Open Science Framework.

## Supplemental Material

Supplemental material for this article is available online.

## References

- Alhanbali, S., Dawes, P., Lloyd, S., & Munro, K. J. (2017). Self-reported listening-related effort and fatigue in hearing-impaired adults. *Ear & Hearing, 38*(1), e39–e48. <https://doi.org/10.1097/AUD.0000000000000361>
- Alhanbali, S., Dawes, P., Millman, R. E., & Munro, K. J. (2019). Measures of listening effort are multidimensional. *Ear and Hearing, 40*(5), 1084–1097. <https://doi.org/10.1097/aud.0000000000000697>
- Barker, B. A., & Newman, R. S. (2004). Listen to your mother! The role of talker familiarity in infant streaming. *Cognition, 94*(2), B45–B53. <https://doi.org/10.1016/j.cognition.2004.06.001>
- Biçer, A., Koelewijn, T., & Başkent, D. (2023). Short implicit voice training affects listening effort during a voice cue sensitivity task with vocoder-degraded speech. *Ear and Hearing, 44*(4), 900–916. <https://doi.org/10.1097/AUD.0000000000001335>
- Borghini, G., & Hazan, V. (2018). Listening effort during sentence processing is increased for non-native listeners: A pupillometry study. *Frontiers in Neuroscience, 12*, 1–13. <https://doi.org/10.3389/fnins.2018.00152>
- Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology, 45*(4), 602–607. <https://doi.org/10.1111/J.1469-8986.2008.00654.X>
- Brown, V., McLaughlin, D. J., Strand, J. F., & Van Engen, K. J. (2020). Rapid adaptation to fully intelligible nonnative-accented speech reduces listening effort. *Quarterly Journal of Experimental Psychology, 73*(9), 1431–1443. <https://doi.org/10.1177/1747021820916726>
- Buntrock, M. S., Barker, B. A., Gurries, M. M., & Barrett, T. S. (2021). The effect of talker familiarity on sentence recognition accuracy in complex noise. *Experimental Psychology, 68*(1), 49–55. <https://doi.org/10.1027/1618-3169/a000509>
- Case, J., Seyfarth, S., & Levi, S. V. (2018). Short-term implicit voice-learning leads to a Familiar Talker Advantage: The role of encoding specificity. *The Journal of the Acoustical Society of America, 144*(6), EL497–EL502. <https://doi.org/10.1121/1.5081469>
- Cherry, E. C. C. (1953). Some experiments on the recognition of speech, with one and two ears. *The Journal of the Acoustical Society of America, 25*(5), 975–979. <https://doi.org/10.1121/1.1907229>
- Dimitrijevic, A., Smith, M. L., Kadis, D. S., & Moore, D. R. (2019). Neural indices of listening effort in noisy environments. *Scientific Reports, 9*(1), 11278. <https://doi.org/10.1038/s41598-019-47643-1>
- Domingo, Y., Holmes, E., & Johnsrude, I. S. (2020). The benefit to speech intelligibility of hearing a familiar voice. *Journal of Experimental Psychology: Applied, 26*(2), 236–247. <https://doi.org/10.1037/xap0000247>
- Domingo, Y., Holmes, E., Macpherson, E., & Johnsrude, I. S. (2019). Using spatial release from masking to estimate the

- magnitude of the familiar-voice intelligibility benefit. *The Journal of the Acoustical Society of America*, 146(5), 3487–3494. <https://doi.org/10.1121/1.5133628>
- Duquesnoy, A. J. (1983). Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons. *The Journal of the Acoustical Society of America*, 74(3), 739–743. <http://www.ncbi.nlm.nih.gov/pubmed/6630729>. <https://doi.org/10.1121/1.389859>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Friedman, N. P., Miyake, A., Young, S. E., DeFries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology: General*, 137(2), 201–225. <https://doi.org/10.1037/0096-3445.137.2.201>
- Gajardo-Vidal, A., Montembeault, M., Lorca-Puls, D. L., Licata, A. E., Bogley, R., Erhoff, S., Ratnasiri, B., Ezze, Z., Battistella, G., Tsoy, E., Pereira, C. W., DeLeon, J., Tee, B. L., Henry, M. L., Miller, Z. A., Rankin, K. P., Mandelli, M. L., Possin, K. L., & Gorno-Tempini, M. L. (2024). Assessing processing speed and its neural correlates in the three variants of primary progressive aphasia with a non-verbal tablet-based task. *Cortex*, 171, 165–177. <https://doi.org/10.1016/J.CORTEX.2023.10.011>
- Gatehouse, S., & Noble, W. (2004). The Speech, Spatial and Qualities of Hearing scale (SSQ). *International Journal of Audiology*, 43(2), 85–99. <https://doi.org/10.1080/14992020400050014>
- Hershman, R., Henik, A., & Cohen, N. (2018). A novel blink detection method based on pupillometry noise. *Behavior Research Methods*, 50(1), 107–114. <https://doi.org/10.3758/s13428-017-1008-1>
- Holmes, E. (2018). *Speech recording videos* (v1.0.0). <https://doi.org/10.5281/zenodo.1165402>
- Holmes, E., Domingo, Y., & Johnsrude, I. S. (2018). Familiar voices are more intelligible, even if they are not recognized as familiar. *Psychological Science*, 29(10), 1575–1583. <https://doi.org/10.1177/0956797618779083>
- Holmes, E., Folkeard, P., Johnsrude, I. S., & Scollie, S. (2018). Semantic context improves speech intelligibility and reduces listening effort for listeners with hearing impairment. *International Journal of Audiology*, 57, 483–492. <https://doi.org/10.1080/14992027.2018.1432901>
- Holmes, E., & Johnsrude, I. S. (2020). Speech spoken by familiar people is more resistant to interference by linguistically similar speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(8), 1465–1476. <https://doi.org/10.1037/xlm0000823>
- Holmes, E., & Johnsrude, I. S. (2021). Speech-evoked brain activity is more robust to competing speech when it is spoken by someone familiar. *NeuroImage*, 237, 118107. <https://doi.org/10.1016/j.neuroimage.2021.118107>
- Holmes, E., & Johnsrude, I. S. (2023). Intelligibility benefit for familiar voices is not accompanied by better discrimination of fundamental frequency or vocal tract length. *Hearing Research*, 429, 108704. <https://doi.org/10.1016/j.heares.2023.108704>
- 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>
- Hornsby, B. W. Y., Naylor, G., & Bess, F. H. (2016). A taxonomy of fatigue concepts and their relation to hearing loss. *Ear and Hearing*, 37, 136S–144S. <https://doi.org/10.1097/AUD.0000000000000289>
- Jagiello, R., Pomper, U., Yoneya, M., Zhao, S., & Chait, M. (2019). Rapid brain responses to familiar vs. Unfamiliar music—an EEG and pupillometry study. *Scientific Reports*, 9(1), 15570. <https://doi.org/10.1038/S41598-019-51759-9>
- Johns, M. A., Calloway, R. C., Karunathilake, I. M. D., Decruy, L. P., Anderson, S., Simon, J. Z., & Kuchinsky, S. E. (2024). Attention mobilization as a modulator of listening effort: Evidence from pupillometry. *Trends in Hearing*, 28, 1–20. <https://doi.org/10.1177/23312165241245240>
- Johnsrude, I. S., Casey, E., & Carlyon, R. P. (2014). Listen to your mother: Highly familiar voices facilitate perceptual segregation. *The Journal of the Acoustical Society of America*, 135(4), 2423. <https://doi.org/10.1121/1.4878052>
- Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science*, 24(10), 1995–2004. <https://doi.org/10.1177/0956797613482467>
- Kidd, G., Best, V., & Mason, C. R. (2008). Listening to every other word: Examining the strength of linkage variables in forming streams of speech. *The Journal of the Acoustical Society of America*, 124(6), 3793–3802. <https://doi.org/10.1121/1.2998980>
- Kinley, I., & Levy, Y. (2022). Pupl: An open-source tool for processing pupillometry data. *Behavior Research Methods*, 54(4), 2046–2069. <https://doi.org/10.3758/S13428-021-01717-Z/FIGURES/23>
- Koelewijn, T., de Kluiver, H., Shinn-Cunningham, B. G., Zekveld, A. A., & Kramer, S. E. (2015). The pupil response reveals increased listening effort when it is difficult to focus attention. *Hearing Research*, 323, 81–90. <https://doi.org/10.1016/j.heares.2015.02.004>
- Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing*, 33(2), 291–300. <https://doi.org/10.1097/AUD.0B013E3182310019>
- Kreitewolf, J., Mathias, S. R., & von Kriegstein, K. (2017). Implicit talker training improves comprehension of auditory speech in noise. *Frontiers in Psychology*, 8, 1584. <https://doi.org/10.3389/fpsyg.2017.01584>
- Kret, M. E., & Sjak-Shie, E. E. (2019). Preprocessing pupil size data: Guidelines and code. *Behavior Research Methods*, 51(3), 1336–1342. <https://doi.org/10.3758/S13428-018-1075-Y/FIGURES/4>
- Lemercier, A., Guillot, G., Courcoux, P., Garrel, C., Baccino, T., & Schlich, P. (2014). Pupillometry of taste: Methodological guide—from acquisition to data processing-and toolbox for MATLAB. *Quantitative Methods for Psychology*, 10(2), 179–195. <https://doi.org/10.20982/TQMP.10.2.P179>

- Levi, S. V. (2015). Talker familiarity and spoken word recognition in school-age children. *Journal of Child Language*, 42(4), 843–872. <https://doi.org/10.1017/S0305000914000506>
- Levi, S. V., Winters, S. J., & Pisoni, D. B. (2011). Effects of cross-language voice training on speech perception: Whose familiar voices are more intelligible? *The Journal of the Acoustical Society of America*, 130(6), 4053–4062. <https://doi.org/10.1121/1.3651816>
- Magnuson, J. S., Yamada, R. A., & Nusbaum, H. C. (1995). The effects of familiarity with a voice on speech perception. Proceedings of the 1995 Spring Meeting of the Acoustical Society of Japan, 391–392.
- Marrone, N., Mason, C. R., & Kidd, G. (2008). Evaluating the benefit of hearing aids in solving the cocktail party problem. *Trends in Amplification*, 12(4), 300–315. <https://doi.org/10.1177/1084713808325880>. Evaluating
- McGarrigle, R., Dawes, P., Stewart, A. J., Kuchinsky, S. E., & Munro, K. J. (2017). Pupillometry reveals changes in physiological arousal during a sustained listening task. *Psychophysiology*, 54(2), 193–203. <https://doi.org/10.1111/PSYP.12772>
- McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group “white paper.”. *International Journal of Audiology*, 53(7), 433–445. <https://doi.org/10.3109/14992027.2014.890296>
- Mendoza-Halliday, D., Xu, H., Azevedo, F. A. C., & Desimone, R. (2024). Dissociable neuronal substrates of visual feature attention and working memory. *Neuron*, 112(5), 850–863.e6. <https://doi.org/10.1016/J.NEURON.2023.12.007/ATTACHMENT/9512A049-8F77-47F3-8428-358A15F89B7C/MMC2.PDF>
- Milne, A. E., Zhao, S., Tampakaki, C., Bury, G., & Chait, M. (2021). Sustained pupil responses are modulated by predictability of auditory sequences. *Journal of Neuroscience*, 41(28), 6116–6127. <https://doi.org/10.1523/JNEUROSCI.2879-20.2021>
- Moore, T. M., & Picou, E. M. (2018). A potential bias in subjective ratings of mental effort. *Journal of Speech, Language, and Hearing Research*, 61(9), 2405–2421. [https://doi.org/10.1044/2018\\_JSLHR-H-17-0451](https://doi.org/10.1044/2018_JSLHR-H-17-0451)
- Newman, R. S., & Evers, S. (2007). The effect of talker familiarity on stream segregation. *Journal of Phonetics*, 35(1), 85–103. <https://doi.org/10.1016/j.wocn.2005.10.004>
- Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, 60(3), 355–376. <https://doi.org/10.3758/BF03206860>
- Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, 5(1), 42–46. <https://doi.org/10.1111/j.1467-9280.1994.tb00612.x>
- Obleser, J., Wöstmann, M., Hellbernd, N., Wilsch, A., & Maess, B. (2012). Adverse listening conditions and memory load drive a common alpha oscillatory network. *Journal of Neuroscience*, 32(36), 12376–12383. <https://doi.org/10.1523/JNEUROSCI.4908-11.2012>
- Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17(11), 776–783. <https://doi.org/10.1037/H0043424>
- Papesh, M. H., Goldinger, S. D., & Hout, M. C. (2012). Memory strength and specificity revealed by pupillometry. *International Journal of Psychophysiology*, 83(1), 56–64. <https://doi.org/10.1016/J.IJPSYCHO.2011.10.002>
- Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., & Wingfield, A. (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear and Hearing*, 37, 5S–27S. <https://doi.org/10.1097/AUD.0000000000000312>
- Qiyuan, J., Richer, F., Wagoner, B. L., & Beatty, J. (1985). The pupil and stimulus probability. *Psychophysiology*, 22(5), 530–534. <https://doi.org/10.1111/j.1469-8986.1985.tb01645.x>
- Rodd, J. M., Davis, M. H., & Johnsrude, I. S. (2005). The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cerebral Cortex*, 15(8), 1261–1269. <https://doi.org/10.1093/CERCOR/BHI009>
- Saderi, D., Schwartz, Z. P., Heller, C. R., Pennington, J. R., & David, S. V. (2021). Dissociation of task engagement and arousal effects in auditory cortex and midbrain. *ELife*, 10, 1–25. <https://doi.org/10.7554/eLife.60153>
- Shields, C., Sladen, M., Bruce, I. A., Kluk, K., & Nichani, J. (2023). Exploring the correlations between measures of listening effort in adults and children: A systematic review with narrative synthesis. *Trends in Hearing*, 27, 1–20. [https://doi.org/10.1177/23312165221137116/ASSET/IMAGES/LARGE/10.1177\\_23312165221137116-FIG4.JPEG](https://doi.org/10.1177/23312165221137116/ASSET/IMAGES/LARGE/10.1177_23312165221137116-FIG4.JPEG)
- Souza, P. E., Gehani, N., Wright, R., & McCloy, D. (2013). The advantage of knowing the talker. *Journal of the American Academy of Audiology*, 24(8), 689–700. <https://doi.org/10.3766/jaaa.24.8.6>
- Vö, M. L. H., Jacobs, A. M., Kuchinke, L., Hofmann, M., Conrad, M., Schacht, A., & Hutzler, F. (2008). The coupling of emotion and cognition in the eye: Introducing the pupil old/new effect. *Psychophysiology*, 45(1), 130–140. <https://doi.org/10.1111/J.1469-8986.2007.00606.X>
- Wendt, D., Dau, T., & Hjortkjær, J. (2016). Impact of background noise and sentence complexity on processing demands during sentence comprehension. *Frontiers in Psychology*, 7(MAR), 179420. <https://doi.org/10.3389/FPSYG.2016.00345/BIBTEX>
- Wiggins, I. M., Stacey, J. E., Naylor, G., & Saunders, G. H. (2025). Relationships between subjective and objective measures of listening accuracy and effort in an online speech-in-noise study. *Ear and Hearing*, 46(5), 1197–1209. <https://doi.org/10.1097/AUD.0000000000001662>
- Winn, M. B. (2016). Rapid release from listening effort resulting from semantic context, and effects of spectral degradation and cochlear implants. *Trends in Hearing*, 20, 1–17. <https://doi.org/10.1177/2331216516669723>

- Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2015). The impact of auditory spectral resolution on listening effort revealed by pupil dilation. *Ear and Hearing*, 36(4), e135–e165. <https://doi.org/10.1097/AUD.0000000000000145>.The
- Winn, M. B., & Teece, K. H. (2021). Listening effort is not the same as speech intelligibility score. *Trends in Hearing*, 25, 1–26. <https://doi.org/10.1177/23312165211027688>
- Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky, S. E. (2018). Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started. *Trends in Hearing*, 22, 1–32. <https://doi.org/10.1177/2331216518800869>
- Yonan, C. A., & Sommers, M. S. (2000). The effects of talker familiarity on spoken word identification in younger and older listeners. *Psychology and Aging*, 15(1), 88–99. <https://doi.org/10.1037/0882-7974.15.1.88>
- Zekveld, A. A., Heslenfeld, D. J., Johnsrude, I. S., Versfeld, N. J., & Kramer, S. E. (2014). The eye as a window to the listening brain: Neural correlates of pupil size as a measure of cognitive listening load. *NeuroImage*, 101, 76–86. <https://doi.org/10.1016/j.neuroimage.2014.06.069>
- Zekveld, A. A., Koelewijn, T., & Kramer, S. E. (2018). The pupil dilation response to auditory stimuli: Current state of knowledge. *Trends in Hearing*, 22, 1–25. <https://doi.org/10.1177/2331216518777174>
- Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2011). Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear and Hearing*, 32(4), 498–510. <https://doi.org/10.1097/AUD.0b013e31820512bb>
- Zhu, W., & Holmes, E. (2025). Computer-based voice familiarization, delivered remotely using an online platform, improves speech intelligibility for older and younger adults. *Journal of Experimental Psychology: Applied*, 31(4), 314–326. <https://doi.org/10.1037/XAP0000536>