Vocal and semantic cues for the segregation of long concurrent speech stimuli in diotic and dichotic listening-The Long-SWoRD test

Moïra-Phoebé Huet, Christophe Micheyl, Etienne Gaudrain, Etienne Parizet

Vocal and semantic cues for the segregation of long concurrent speech stimuli in diotic and dichotic listening—The Long-SWoRD test

Moïra-Phoebé Huet, Christophe Micheyl, Etienne Gaudrain, et al.

Citation: The Journal of the Acoustical Society of America **151**, 1557 (2022); doi: 10.1121/10.0007225
View online: https://doi.org/10.1121/10.0007225
View Table of Contents: https://asa.scitation.org/toc/jas/151/3
Published by the Acoustical Society of America

JASA ARTICLE

Vocal and semantic cues for the segregation of long concurrent speech stimuli in diotic and dichotic listening—The Long-SWoRD test

Moïra-Phoebé Huet,[1,a)] Christophe Micheyl,[2] Etienne Gaudrain,[3,b)] and Etienne Parizet[1]

[1]*Laboratory of Vibration and Acoustics, National Institute of Applied Sciences, University of Lyon, 20 Avenue Albert Einstein, Villeurbanne, 69100, France*

[2]*Starkey France, 23 Rue Claude Nicolas Ledoux, Créteil, 94000, France*

[3]*Lyon Neuroscience Research Center, Auditory Cognition and Psychoacoustics, Centre National de la Recerche Scientifique UMR5292, Institut National de la Santé et de la Recherche Médicale U1028, Université Claude Bernard Lyon 1, Université de Lyon, Centre Hospitalier Le Vinatier, Neurocampus, 95 boulevard Pinel, Bron Cedex, 69675, France*

ABSTRACT:

It is not always easy to follow a conversation in a noisy environment. To distinguish between two speakers, a listener must mobilize many perceptual and cognitive processes to maintain attention on a target voice and avoid shifting attention to the background noise. The development of an intelligibility task with long stimuli—the Long-SWoRD test—is introduced. This protocol allows participants to fully benefit from the cognitive resources, such as semantic knowledge, to separate two talkers in a realistic listening environment. Moreover, this task also provides the experimenters with a means to infer fluctuations in auditory selective attention. Two experiments document the performance of normal-hearing listeners in situations where the perceptual separability of the competing voices ranges from easy to hard using a combination of voice and binaural cues. The results show a strong effect of voice differences when the voices are presented diotically. In addition, analyzing the influence of the semantic context on the pattern of responses indicates that the semantic information induces a response bias in situations where the competing voices are distinguishable and indistinguishable from one another.

 https://doiorg/10.1121/10.0007225

(Received 4 December 2020; revised 18 October 2021; accepted 25 October 2021; published online 9 March 2022)

[Editor: Karen S. Helfer] Pages: 1557–1574

I. INTRODUCTION

Since Cherry (1953) introduced "cocktail party" listening as an influential research topic, our understanding of the ability to selectively listen to a voice among other competing sounds, including other voices, has increased tremendously (see Shinn-Cunningham and Best, 2008, for a review). More recently, a number of studies have started to identify the neurophysiological correlates of selective attention while listening to a voice competing with another (Ding and Simon, 2012; O'Sullivan *et al.*, 2015). This scenario provides a special case of cocktail party listening with a single masker, a situation that often occurs in everyday life and can prove very challenging for hearing-impaired individuals (for a review, see Bronkhorst, 2015). Several studies, in particular, have found that when a listener is attending to one of two speech streams that correspond to two different voices, the cortical responses measured using intracortical recordings, electroencephalography (EEG), or magnetoencephalography (MEG) follow more strongly the temporal envelope of the attended stream than that of the unattended stream (Ding and Simon, 2012; Mesgarani and Chang, 2012; O'Sullivan *et al.*, 2015). This process of uncovering which speech stream is being attended from the cortical temporal response function (TRF) is referred to as "attention decoding" (e.g., Biesmans *et al.*, 2015)—a technique that potentially has important clinical applications (Slaney *et al.*, 2020). In addition, the TRF can be used to unravel the neural correlates of the online processing of speech in situations in which sounds or voices compete with one another (Broderick *et al.*, 2018; Di Liberto *et al.*, 2015).

One shortcoming of these recent neurophysiological studies relates to their assumption regarding relationships between neural responses and the attended speech envelope. With few exceptions (Akram *et al.*, 2017; Miran *et al.*, 2018), most of these studies presume that listeners can maintain an uninterrupted focus on the target speech stream for relatively long periods of time, ranging from at least a few tens of seconds to over 10–20 min.

[a)]Current address: Department of Electrical and Computer Engineering, Laboratory for Computational Audio Perception, Johns Hopkins University, 3400 North Charles Street, Barton Hall Room 323, Baltimore, MD 21218, USA. Electronic mail: mphuet@jhu.edu. ORCID: 0000-0002-0470-0894.

[b)]Also at: Department of Otorhinolaryngology, University Medical Center Groningen, University of Groningen, Groningen, Netherlands. ORCID: 0000-0003-0490-0295.

The key reason for using long stimuli in these studies is that the performance of the algorithms decoding the neural responses to speech usually increases with the amount of training data available. Studies suggest that the total duration of the stimuli should be at least 15 min (Mirkovic *et al.*, 2015), and the stimulus itself should last at least 10 s (O'Sullivan *et al.*, 2015). Informal reports from others in addition to our own experiences as participants in tasks using such long stimuli, strongly suggest that this constant-attention assumption may not be warranted. Rather, it appears that for many listeners, maintaining one's undivided attention on a single speech stream for several tens of seconds (such as listening to a voice telling a story) while another speech stream is being played concurrently at approximately the same sound level places demands on the listener's focus that can lead to attentional shifts from the target to the nontarget voice. In fact, when a behavioral approach is used, experimenters seem to favor relatively short stimuli, such as individual words or single sentences, to avoid this type of pitfall (as also argued by Herrmann and Johnsrude, 2020). Series of short, unrelated stimuli potentially provide more respite for the participant than long, uninterrupted stories. Between short stimuli, the participant may have an opportunity to restore their attentional resources before moving on to the next stimuli. It is worth noting that even in such conditions, participants in this type of experiment demonstrate a fair amount of confusion errors where their response is based on the masker rather than the target (e.g., Rennies *et al.*, 2019). These errors can reflect a failure to identify which voice is the target but could also result from momentary attentional switches from the target to the masker. It can be hypothesized that such confusion errors would occur at least as often when the task uses long stimuli as seen in many of the neurophysiological studies.

Therefore, there is a discrepancy between the bulk of behavioral studies on concurrent speech perception and the neurophysiological approaches that attempt to describe the underlying neural mechanisms. The difference in the length of the stimuli has consequences not only on the supposed sustainability of attention during the task but also on the potential to influence other mechanisms that have been shown to influence competing speech perception. Although the "peripheral" or sensory mechanisms involved in the perception of long and short stimuli may be largely identical, higher-level perceptual or cognitive mechanisms involved in the two situations may differ greatly. For instance, working memory, which has been shown to be involved in selective-attention tasks (Conway *et al.*, 2001), is likely to be engaged to a greater extent when listening to long, story-like concurrent speech stimuli than while listening to shorter stimuli. Brief auditory storage (Darwin *et al.*, 1972)—also referred to as echoic memory—has an estimated capacity of about 5 s in a concurrent speech context (Treisman, 1964). As a result, short speech stimuli (such as disconnected sentences) can be entirely retained in echoic memory until the participant provides their answer. In contrast, retaining information about auditory sequences that exceed the echoic memory capacity requires further cognitive processing of the linguistic context, which involves the working memory (Lewis *et al.*, 2006). Further, whereas verbal working memory has been shown to play a role in speech-on-speech perception (see Besser *et al.*, 2013, for a review), little is known about the effect that long, coherent stimuli may have on the working memory, particularly stimuli that more closely resemble a real-life communication situation vs short, isolated sentences. In such situations, the working memory could also run out of capacity or else the presence of continuous interfering speech could create challenges in storing information in memory.

More generally, linguistic processing has been shown to play an important role in concurrent-speech perception tasks. For example, Clarke *et al.* (2014) showed that semantic context can trump voice cues, as indicated by the listeners' attention being guided by semantic continuity despite changes in the attended voice characteristics over time. Similarly, Kidd *et al.* (2014) showed that the syntactic structure helps listeners bind words together into coherent speech streams. It can, thus, be argued that long, linguistically coherent stimuli may engage these mechanisms to their full extent in a way that more likely closely resembles a real-life situation.

Another psychological mechanism—one that may also differentially influence the performance in selective listening tasks with short or long stimuli and is discussed in a different body of literature—relates to an effect traditionally referred to as the "buildup" of auditory stream segregation (Bregman, 1990). This mechanism comprises the perceptual organization of two concurrent sequences of sounds into separate auditory streams that are not instantaneous but, instead, build up slowly over time. Although this effect was originally demonstrated with tones (Bregman, 1978), it has been shown to extend to speech stimuli (Best *et al.*, 2018). The buildup can be more or less rapid, depending on how perceptually distinguishable the two streams are (Moore and Gockel, 2002); when the two streams are perceptually similar, it can take up to a few tens of seconds to complete (Bregman, 1978). Second, once the two concurrent streams are separated perceptually, the listener may need additional time to selectively focus their attention on the target stream (Shinn-Cunningham and Best, 2008). These rather slow phenomena are at risk of being overlooked when applying research methods that only involve short stimuli. Moreover, following Bregman's approach, primitive auditory scene analysis cues (which are thought to be related to low-level sensory cues) may be the same for long and short stimuli, while schema-based segregation cues (which rely on higher-level information) may be more available in longer, context-rich stimuli.

These considerations highlight the importance of the duration and complexity of the stimuli, leading to the involvement of different cognitive mechanisms than for shorter, simpler stimuli, therefore potentially making behavioral studies somewhat irreconcilable with a number of neurophysiological studies. On the one hand, behavioral studies

https://doi.org/10.1121/10.0007225

offer better control of the attentional focus of the participant by using short stimuli that do not create many opportunities for attentional switches, but are relatively limited in the cognitive processes they may involve. And on the other hand, neurophysiological studies offer little or no control on attention switches, instead using long stimuli that are likely to engage sustained cognitive processes based on an extended context, which are common in real-life communication and are unlikely to be elicited by short sentences.

To mitigate this issue, some investigators in the neurophysiological approach have made attempts to control for such attentional-shift effects by supplementing it with a behavioral task. For instance, O'Sullivan *et al.* (2015) asked their listeners multiple-choice questions following each 1-min stimulus in order to assess whether the listener was paying attention to the target story. One limitation of this approach, however, is that listeners may have been able to answer such questions correctly even without paying close attention to the target story. Crosse *et al.* (2015), for instance, asked participants to press a button whenever they were listening to the target voice. First, it is possible that listeners were unable to precisely track the wanderings of their attention with their button presses as they were listening to the story; second, asking listeners to press buttons according to their attention while they are focused on listening to a story introduces a secondary task, which may disrupt their performance of the primary, selective-attention task. It would be useful for future studies focusing on neural correlates of auditory attention to employ methods that consistently track fluctuations in selective auditory attention without requiring listeners to perform a secondary motor task while engaged in active listening situations.

To address the aforementioned limitations of the previous studies, we have designed a concurrent-speech perception task—*Long-SWoRD*—and have assembled a set of stimuli specifically designed to provide experimenters with a means to infer the fluctuations in auditory selective attention while participants are listening to a short stories presented concurrently, based solely on behavioral data. Our initial goal was to design a behavioral task that uses long stimuli and provides a behavioral account of attention throughout the sequence that can further be combined with EEG TRFs to improve the attention decoding techniques (Huet, 2020). However, because the method described here uses long, coherent stimuli designed to engage the full extent of the cognitive processes that are suspected of being used in real-life communication situations, it could also be useful to explore more specifically the role of these processes when applying behavioral or electrophysiological approaches. In this context, the goal of the present study was to assess whether the effects of two very common cues for speech-on-speech separation—voice and location—were measurable using the Long-SWoRD test and whether signs of the involvement of higher-level cognitive processes could be extracted from the data. In Sec. II, we describe the construction of the new material and task in a way that could be replicated in other languages. In Sec. III, this article presents an experiment that assesses the joint role of voice and binaural cues in the separation of concurrent speech. In Sec. IV, we examine the role of voice cues in further detail. Finally, Sec. V analyzes the role of acoustic and semantic cues present in the test paradigm.

II. PRE-EXPERIMENT: CORPUS AND TASK DEVELOPMENT

This section describes the creation and development of the Selective Word Recognition Discrimination (SWoRD) test associated with long stimuli: the *Long-SWoRD* test. In this test, the participants hear two short stories presented concurrently. Each story is composed of a few sentences. The participants' task is to retrieve three words from different key time points (or keywords) belonging to the *target* story, i.e., the story they have to listen to (see Fig. 1 for an example of a trial). After the initial creation of the corpus, a pre-experiment was run to evaluate the material and prune it to remove items that give abnormally low scores.

A. Methods

1. Stimuli

In selecting a speech corpus for the new test, we used two main criteria for inclusion: first, we were mindful to include sufficiently varied topics for stories that were amusing and/or informative to elicit interest from a diverse

FIG. 1. (Color online) The concurrent waveforms for the final procedure of the Long-SWoRD test with the target story above (in dark gray/blue) and the masker story below (in light gray/yellow). The target and masker keywords (marked with hatches) are scattered throughout both stories. The corresponding text, where the keywords are marked in bold, is loosely aligned with the waveform. The English translation is provided in gray italics (Enders, 2015b).

audience of younger and older listeners. Second, we ensured that the voice of the sole narrator of the audiobook could be credibly modifiable using the STRAIGHT software (Kawahara *et al.*, 1999) to create masker voices that were different from the original voice (e.g., from a female voice to a male voice). The French Audiobook, *Le Charme discret de l'intestin* (*The Inside Story of Our Body's Most Underrated Organ*; Enders *et al.* (2015a; 2016)), narrated by a single female speaker, met these criteria.

2. Speech material

547 stories were extracted from the audiobook according to several criteria, including quantitative criteria, such as duration (between 11 and 18 s per story) and number of words (between 22 and 55 words per story), as well as qualitative criteria, such as whether the stories still made sense after being separated from their broader context (so that they could be used as standalone material). As a result, mostly anecdotes and "fun facts" were selected.

Next, for each story, the three keywords that the participants would later have to identify were also chosen carefully, using the following criteria. First, the three keywords had to occur at different times within the story (see Fig. 1): one keyword near the beginning, one keyword toward the middle, and one keyword toward the end of the story. To reduce the impact of the primacy and recency effects that are classically observed in the serial recall tasks (e.g., Schlittmeier *et al.*, 2008), the very first and last words in the stories were never selected as keywords. Second, the selected keywords could only occur once in the story. Third, potential keywords were selected from verbs, nouns and adjectives, containing 2–15 phonemes. Finally, any word that was a candidate for inclusion was compared against a lexical database containing the occurrence frequencies of words in the French language (Lexique 3.081; New *et al.*, 2001). Words that were unusually rare or frequent were replaced by another word from the same story, which had a more typical occurrence frequency in the language. As it was sometimes not possible to find a word that met all three inclusion criteria, some stories were removed, bringing the total number of stories to 526. The distributions for occurrence frequencies of words in the French language for both the whole lexical database and the keyword set used here are shown in Fig. 2. The selected keywords distribution had a geometrical mean of 6.67 per million occurrences in the French language and 95% of values fell between 0.04 and 532.79 per million occurrences. In addition, the keywords were such that they were representative of the French phoneme inventory as they covered the complete range of phonemes found in French and followed a distribution that is similar to the one found in the Lexique database.[1]

The last step was to match the target and masker stories. To form a pair, the two stories had to be of a similar duration. In addition, the three target keywords associated with one story of a pair (e.g., target) could not appear in the other story (e.g., masker). In addition, three *extraneous* keywords,

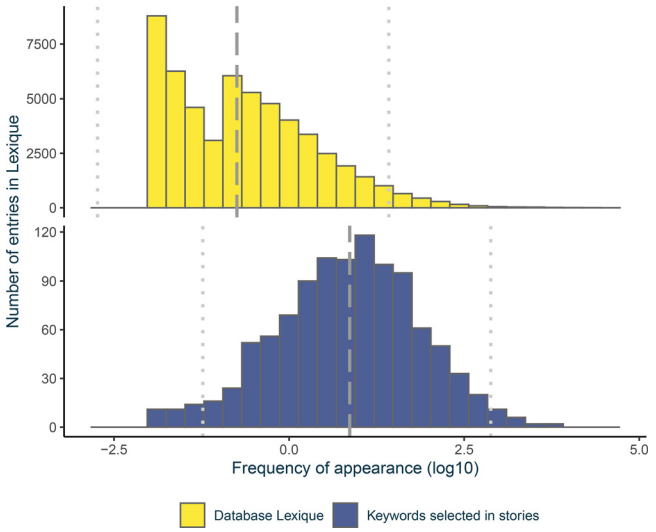

FIG. 2. (Color online) The word frequency distributions [$\log_{10}$(occurrences per million)] in the Lexique database (top) and keyword set (bottom). The dashed lines represent the mean, whereas the dotted lines represent two standard deviations of the mean.

selected from another story, i.e., belonging neither to the target nor to the masker story of the pair, were added to the set of keywords as decoys. These keywords also were randomly assigned to each pair of stories, bringing the total number of keywords to nine for each pair. Adding these extraneous words lowered the chance level down to 33% in the task, but also allowed the determination of whether errors were due to target-masker confusion or if the participants simply did not hear any of the keywords in the mixture. In addition, because the extraneous keywords were coming from another story from the same audiobook, they also revolved around the same general topic and were semantically not too far from the target and masker stories and keywords and could credibly belong to them, thus, making the task significantly harder. Finally, having the extraneous word offered the potential possibility of adjusting the difficulty of the task by adjusting the extraneous words to control that they remained within a semantic distance from the target and masker, which was similar to that between the target and masker. After the evaluation described below, this adjustment step was deemed unnecessary for the material at hand and was, therefore, not applied in this version of the Long-SWoRD. This extra step could, however, be used in subsequent iterations of the test.

3. Procedure

Recalling three keywords in stories that could sometimes last 18 s seemed potentially strenuous for the participants because of the implied cognitive load on the working memory. Given the volume of material to be evaluated—526 stories—it did not seem realistic to expect the participants to engage in an experiment covering all of the items while maintaining a good level of alertness. Instead, we decided to create an experimental setup that would limit the data collection to a small, random subset of the material per

https://doi.org/10.1121/10.0007225

participant but had a large number of participants. To achieve this, we evaluated the material and experimental procedure with an online study, limiting each session to a small subset of the corpus but aiming to recruit a large number of participants.

The online experiment consisted of 20 trials, each corresponding to one story. The 20 stories were selected by randomly choosing 20 pairs out of the 263 available pairs of target and masker and selecting randomly either the target or masker. In each trial, the participants were presented with a screen instructing them to listen to the story while the audio recording was being played in isolation without a concurrent masker. When the story was finished, the participants were shown nine buttons, each displaying a word, arranged in three rows of three buttons. For each row, one of the words was actually part of the story that had just been played, one word came from the other story of the pair (which was not played), and one word was a distractor word that belonged to neither of the two stories. In addition, each row matched one of the chronological keyword positions in the target and masker stories (beginning, middle, and end). The participants were instructed to find, in each row, the word that had been presented in the story. Once finished, the participants were offered the possibility to run the experiment again, and 20 stories were randomly selected once more. This selection was independent from the previous experimental sessions with the same participant, meaning that they could be presented stories that they had already heard. This happened in 84 trials out of 3847 (2%).

4. Participants

231 participants (mean age, 34.1 years old; minimum age, 18 years old; maximum age, 72 years old) took part in this online experiment. Twelve of these participants were removed from this analysis because French was not their native language. The volunteers provided informed consent before participating.

B. Results

1. Pruning the corpus

Because of the random sampling of the corpus, the total number of presentations for each story varied between 1 and 16. The participants' average score was 89.5%.

In the pruning process, we considered individual stories and also individual words. Stories that had remarkably low scores were excluded, and stories containing words that had remarkably low scores were also excluded. To define what qualified as remarkable scores, we considered the width of the score distribution for the stories (standard deviation $\sigma = 9.74$ percentage points) and the width of the score distribution for the individual words ($\sigma = 17.1$ percentage points).

From the 263 pairs of stories, only 178 were kept on the basis of three criteria. First, both target and masker stories had to have been tested at least 3 times each, which led us to exclude 16 pairs of stories. Second, the average scores for the target and masker stories of a pair had to be both higher than 70.0% (the average minus two standard deviations, 89.5% - $2 \times 9.7\%$). Third, neither the keywords from the target story nor the keywords from the masker story had an individual score inferior to 55.3% (89.5% – $2 \times 17.1\%$).

2. Grouping into lists

Lists were created to distribute the semantic topics throughout the experiment. Indeed, the predominance of keywords, such as "bacteria," could influence the participants from one trial to another. Of the available 178 pairs of stories selected based on the previous criteria, 144 pairs were chosen so as to form 12 lists, each list comprising 12 pairs of stories. Within a list, all story pairs contained unique keywords that were not used in any other story. The overall duration of each list was also approximately the same ($\mu = 175.08$ s, $\sigma = 0.9$ s) and each list contained equally short and long stories.[1] With this grouping, the average score for the target alone or masker alone for each list was between 89.3% ($\sigma = 8.40$ percentage points) and 97.3% ($\sigma = 2.89$ percentage points).

The final version of the corpus, along with the selected keywords, can be consulted in Huet (2020, Appendix B, p. 163).

3. Extraneous keywords analysis

As indicated above, the extraneous keywords of a story are target and masker keywords originating from the other stories. Because all of the stories came from the same book and by extension from a similar lexical field, it was necessary to ensure that the extraneous keywords were semantically equidistant from the target and masker keywords to avoid introducing response biases.

A French word2vec model (Mikolov *et al.*, 2013) trained on a lemmatized version of the French Wikipedia corpus (Gaudrain and Crouzet, 2019) was used to estimate the semantic similarity between the target, masker, and extraneous keywords for each story pair (for more details, see Sec. V and Huet, 2020, p. 30). Figure 3 shows the semantic similarities between the target and masker keywords, between the target and extraneous keywords, and between the masker and extraneous keywords. The similarity varied from zero (not semantically similar) to one (synonyms) and there was no difference between the target-masker similarity ($\mu = 0.227$), the target-extraneous similarity ($\mu = 0.229$), and the masker-extraneous similarity ($\mu = 0.228$) [$F(2, 286) = 0.12, p = 0.88$]. Therefore, from a semantic point of view, the extraneous keywords were as close to the target keywords as to the masker keywords and were, thus, not introducing any response bias in favor of the target or masker.

III. EXPERIMENT 1: VOICE CUES IN DIOTIC AND DICHOTIC LISTENING (REF. 2)

The purpose of this experiment was to assess how voice differences and a simple binaural cue contribute to the concurrent speech perception using the Long-SwoRD test.

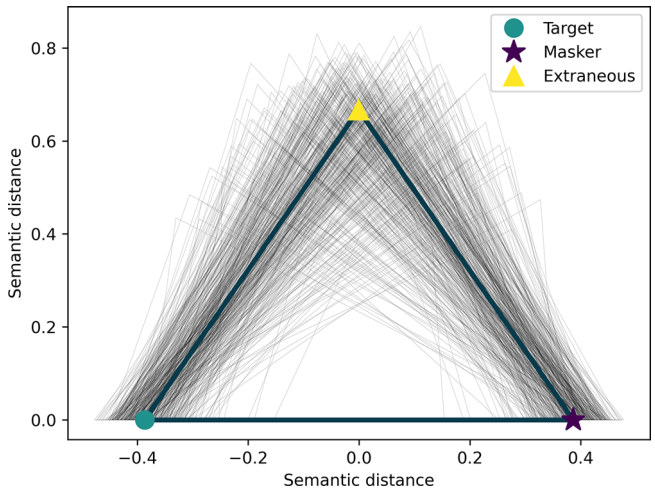

FIG. 3. (Color online) The semantic similarities between the target (circle green), the masker (purple star), and the extraneous (yellow triangle) keywords. The similarities for each story pair and keyword position (i.e., beginning, middle, and end of story) are represented with gray lines. The thick dark line shows the average similarity. For illustration purpose, all of the lines are one minus the semantic similarity, i.e., their lengths represent the dissimilarity: two words that are semantically close will be shown as close on the graph. In addition, the graph is aligned such that the target-masker line is aligned along the x axis and such that the target and masker averages are symmetrically positioned relative to the origin.

Previous studies have shown that the source location and voice cues were both important for speech-on-speech perception (Bronkhorst, 2015). Ives *et al.* (2010) found, using short concurrent syllables, that voice cues were only effective when the sources were co-located and, vice versa, that location cues were most effective when the voices were identical. However, very short stimuli do not reveal the full extent of the benefit voice cues may be able to provide. Using short stimuli structured into longer sequences, Bressler *et al.* (2014) found that voice consistency enhances the segregation of the competing talkers. Samson and Johnsrude (2016) extended these results to short sentences and showed that the consistency of the masker voice across trials also facilitates segregation. Whereas these authors have argued that this effect is largely automatic or stemming from a bottom-up process, others have produced evidence that the voice consistency benefit may also hinge on cognitive processes. Zekveld *et al.* (2014) showed that cognitive load, as estimated through pupil dilation, decreases as the voices become more different from each other, suggesting that benefiting from voice differences requires cognitive resources. In contrast, the same study found that the spatial separation has no effect on pupil dilation, suggesting that the benefit from the binaural cues is primarily bottom-up.

Using the long, meaningful stimuli of the Long-SWoRD, the continuity of the vocal identities of the target and masker talkers, as well as the continuity of the linguistic features spanning across sentences within the same trial, are fully available to the listener. The purpose of this first experiment is to establish that the test is sensitive to voice differences as well as to spatial location differences.

Although some of the previous studies used head-related transfer functions for the spatialization of the stimuli, we used simple dichotic vs diotic presentations that give, respectively, maximal and minimal separation of the sources. In addition, whereas most previous studies used actual different speakers to control the voice parameter, we used a more elaborate approach that allowed us to quantify the voice differences (Başkent and Gaudrain, 2016; Darwin *et al.*, 2003; Vestergaard *et al.*, 2009; Ives *et al.*, 2010). In this approach, the recordings from a single speaker are manipulated to generate a number of different voices by artificially altering their F0 (vocal pitch) and apparent vocal tract length (VTL).

A. Methods

1. Participants

Twenty-two native French speakers, between 20 and 32 years old ($\mu = 24$), took part in the experiment. Pure-tone audiometry (0.25–8 kHz) was performed with an Interacoustics AC40 audiometer (Middlefart, Denmark). All of the participants but one had audiometric thresholds $\leq$20 dB in hearing level (HL) at the test frequencies between 250 Hz and 4 kHz. All of the participants but three also had audiometric thresholds $\leq$20 dB HL at test frequencies between 6 and 8 kHz. The participants provided informed consent before participating and were paid an hourly wage for their participation. The experimental procedure was approved by a local ethics committee (CPP Sud Est II).

2. Procedure

The experiment was composed of 144 trials arranged into 12 blocks of 12 trials. In each trial, the participant heard two competing stories. The participants were instructed to listen to the target story, which was preceded by the word "attention," uttered by the target voice, before the two stories started. Once the stories finished, nine buttons arranged in three rows appeared on the screen with the instructions to identify the three words (one per row) that belonged to the target story. In a random order, each row was occupied by one keyword from the target story and one keyword from the masker story, whereas the remaining "extraneous" keyword was contained neither in the target nor in the masker stories. The rows were chronologically arranged from top to bottom. Once the participant had selected one button per row, the experiment moved on to the next trial.

The audio stimuli in each of these 12 blocks were the same for all of the participants except for the order of presentation of the stories, which was randomized within a block across participants. This within-block randomization scheme was intended to eliminate any systematic "sequential" bias that might have been caused by a particular presentation order while also retaining an ability to analyze the learning effects across blocks despite potential differences in the task difficulty across the different blocks. Each block was randomly associated with a condition for

https://doi.org/10.1121/10.0007225

each participant. Between each block, the subjects were allowed to take a break for as long as they preferred.

Half of the blocks was presented diotically and the other half of the blocks was presented dichotically. In the dichotic condition, the target was presented in the right ear, and the masker was presented in the left ear. For each binaural condition, three different masker voices were used. The same voice difference between the target and masker voices was kept within a block, and the order of the blocks was randomized. The characteristics of these voices are described in Sec. III A 3. Finally, in all of the conditions, the target and masker were presented at the same level, meaning that the target-to-masker ratio (TMR) was set to 0 dB.

The data collection lasted 60–90 min, and the entire procedure was completed within a single session.

3. Stimuli

The audio stimuli were originally recorded by an adult female speaker. This voice, analyzed and resynthesized without modification with the STRAIGHT toolbox (Kawahara *et al.*, 1999) implemented in MATLAB (Natick, MA, USA), was chosen as the target voice. For the masker voices, the voice pitch (F0) and VTL were manipulated during this analysis-resynthesis. The first step of the masker voice creation was to generate a credible male voice (represented by a circle in Fig. 4) by adjusting the F0 and VTL to obtain the direction for the voice manipulation in the F0-VTL plane. The F0 and VTL differences are expressed in semitones (st) relative to the original voice, thereby reflecting a ratio of 2 for 12 st. The second step was to choose the parameters of the masker voices based on the literature. Shifting down the F0 by 8 st and increasing the VTL by 3.04 st allowed enough auditory differentiation to create a "male" voice percept that was very different from the female target voice (as similarly done by Başkent and Gaudrain, 2016). Then, to obtain a very similar voice that was still distinguishable from the target voice, the parameters were adjusted with the values of the just-noticeable difference (JND) as reported by Gaudrain and Başkent (2015). This second voice was, thus, a very similar female voice with a total difference of 1.71 st along the male direction axis. This total distance is calculated in semitones as

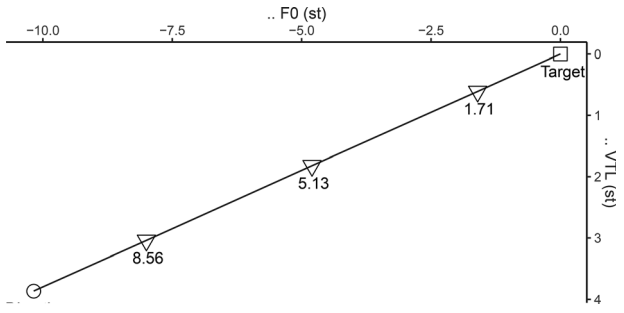

FIG. 4. The distances in semitones (st) between the target and masker voices for experiment 1. The original voice (target) is represented by a square, the credible male voice direction is represented by a circle, and the masker voices are represented by triangles.

TABLE I. The distance between the target and masker voices in semitones for experiment 1.

Voice	ΔF0	ΔVTL	$\sqrt{\Delta F0^2 + \Delta VTL^2}$
Male	−8.00	3.04	8.56
Intermediate	−4.80	1.82	5.13
JND	−1.60	0.61	1.71

$\sqrt{\Delta F0^2 + \Delta VTL^2}$. The last voice was synthesized to be equidistant from the first two voices. Table I shows the parameter values for the three masker voices.

4. Apparatus

Stimuli were presented with OpenSesame (Mathôt *et al.*, 2012). The participants listened to stimuli over Sennheiser HD250 Linear II headphones (Wedemark, Germany) in a sound-attenuated booth. The presentation level was calibrated to 65 dB sound pressure level (SPL) using an AEC101 artificial ear and sound level meter LD824 (Larson Davis, Depew, NY).

5. Statistical analyses

To score the results, we first considered the collected data as binary correct/incorrect for each keyword based on whether the participant identified the target word. We used generalized linear mixed models (gLMMs) that were based on the binomial distribution using logit as the link function. Such models are well suited to minimize the effects of saturation in the binomial data. The models were implemented in *R* using the *lme4* package (Bates *et al.*, 2014, p. 4) and reduced using a top-down strategy for the model selection (Zuur *et al.*, 2009). The final model is reported with the *lme4* syntax, such as

$$\text{BinaryScore} \sim \text{factor}_A * \text{factor}_B + (\text{factor}_A * \text{factor}_B \mid \text{subject}). \quad (1)$$

The full-factorial model is indicated by the fixed effect term factor_A * factor_B and includes the main effects and interactions for these two key conditions. The last term of the equation describes an individual random intercept and slope per subject for the factor_A * factor_B. In some models, the full-factorial structure could not be used as random effect to obtain a good convergence in the model. In that case, the random structure was gradually simplified by eliminating the interaction terms until the convergence became reliable. Note that for modeling and interpreting purposes, the continuous variables (e.g., the voice distance factor) were rescaled, such that the minimum value corresponded to zero and the maximum value corresponded to one, and were then centered on the average. To perform the *post hoc* analyses, we ran new gLMMs on subsets of the data (e.g., dichotic or diotic) with (1 | subject) as well as normalized pairwise comparisons of the proportion with a *false discovery rate correction*. The significance of the

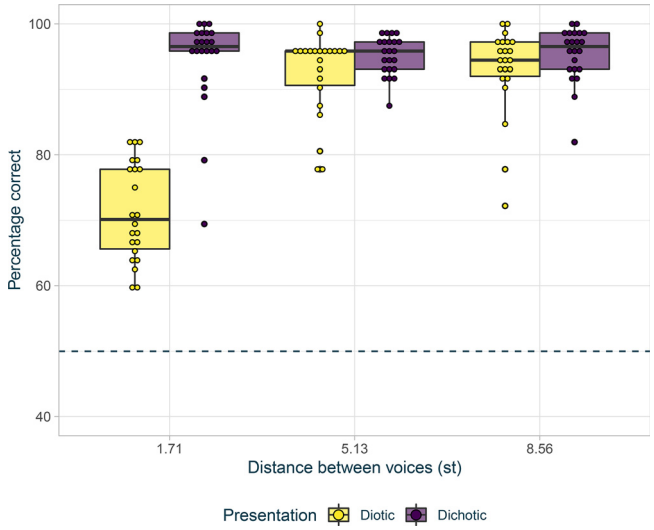

FIG. 5. (Color online) The percentage of correct responses for each voice in both the diotic (bright yellow) and dichotic (dark purple) presentations. The points represent the scores for every participant in each condition. The box extends from the first to the third quartile. The line in the middle of the box is the median. The whiskers extend from the box to the observation furthest from the median and within 1.5 times the interquartile range. The dashed line (50%) indicates the level at which the performance is significantly greater than chance based on a binomial test at the 5% significance level.

factors in the model was evaluated using the analysis of variance (ANOVA) function from the *car* package (Fox and Weisberg, 2019).

B. Results

1. General description

Figure 5 shows the average performance as the percentage of correctly identified target keywords for each condition. Based on a binomial test at the 5% significance level, 50%-correct corresponded to the threshold of significance for a score to be different from the one-in-three chance level. All of the subjects demonstrated high scores, well above chance, for every condition.

A gLMM was fitted on the binary (correct/incorrect) scores, and the voice distance was treated as a continuous factor. Equation (2) indicates the final model:

$$\text{score} \sim \text{presentation} * \text{voice} + (\text{presentation} + \text{voice}|\text{subject}). \quad (2)$$

The results (see Table II) showed that the participants had better scores when the stimuli were presented dichotically vs diotically $[z = -15.50, p < 0.001]$. The distance between the voices also had an effect on the participants' answers as well as the interaction between the two factors. The *post hoc* analysis showed that there was no voice effect for the dichotic presentation $[\chi^2(1) = 1.36, p = 0.13]$, whereas the participants had higher scores when the distance increased between the target and masker voices in a diotic presentation $[\chi^2(1) = 255.67, p < 0.001]$. In addition, when the distance between the target and masker voice was 1.71 st, the participants obtained lower scores than when the distance was 5.13 st $[z = 13.49, p < 0.001]$ or 8.56 st $[z = 14.1, p < 0.001]$. There was no difference in the performance between voices 5.13 and 8.56 st $[z = 0.69, p = 0.49]$.

TABLE II. Equation (2) statistics.

Fixed effect	Statistics
Presentation	$\chi^2(1) = 76.62$, $p < 0.001$
Voice	$\chi^2(1) = 103.23$, $p < 0.001$
Voice × presentation	$\chi^2(1) = 72.99$, $p < 0.001$

2. Keyword position analysis

The position of the keyword in the story (treated as a categorical factor with the levels beginning, middle, and end of the story) was also analyzed and Fig. 6 shows the data. Equation (3) indicates the final model:

$$\text{score} \sim \text{position} * \text{voice} * \text{presentation} + (\text{voice}\,|\,\text{subject}). \quad (3)$$

The results are presented in Table III. Regarding the stimulus presentation and distance between voices, the results were similar to those of the previous analysis. The position of the keyword in the story had an effect on the participants' scores with an advantage for the keywords located at the end over the middle $[z = -7.97, p < 0.001]$, as well as the beginning keywords $[z = -6.97, p < 0.001]$, which characterizes a recency effect. In contrast, no primacy effect, characterized by a better score for the beginning keyword, was observed $[z = -1.02, p = 0.31]$. Because of the three-way interaction, we reran this analysis on subsets of the data to better understand where this recency effect was present. In the diotic condition, this recency effect interacted with the voice difference $[\chi^2(2) = 25.66, p < 0.001]$: the recency effect was present for voice 8.56 st $[\chi^2(2) = 26.9, p < 0.0001]$ and voice 5.13 st $[\chi^2(2) = 18.3, p < 0.001]$ but not for voice 1.71 st $[\chi^2(2) = 4.76, p = 0.09]$. In contrast, this interaction between the keyword position and voice was not present when the stimuli were presented dichotically $[\chi^2(2) = 0.17, p = 0.91]$. In this condition, the word position had a significant effect for all of the voices $[\chi^2(2) > 13.8, p < 0.001]$.

3. Error analysis

Analyzing the nature of the errors was necessary to infer whether an error was likely caused by the participant listening to the "wrong" (masker) story. Figure 7 illustrates the error distribution, whereas Eq. (4) represents the final model of a top-down strategy modeling gLMM on the binary (masker or extraneous) data from a subset of answers where the participant did not choose the target keyword,

$$\text{error type} \sim \text{presentation} * \text{voice} + (1|\text{subject}). \quad (4)$$

Because the error type was either "masker" or extraneous, the results of this analysis (see Table IV) indicated how the

https://doi.org/10.1121/10.0007225

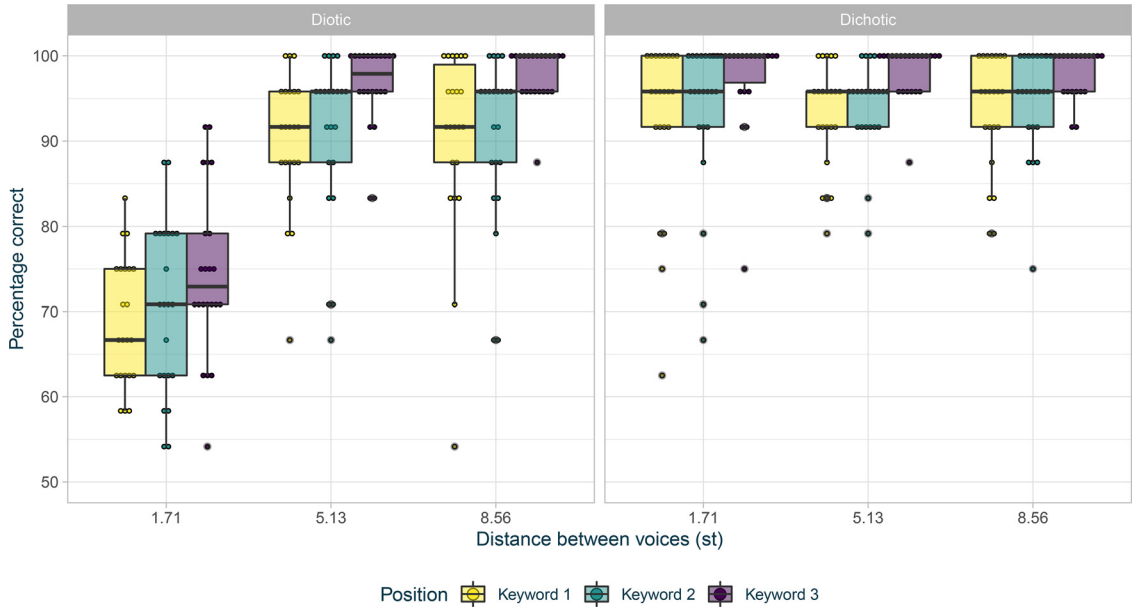

FIG. 6. (Color online) The percentage of correct responses for each voice in both the diotic and dichotic presentations for each keyword position. The details of the boxplot are the same as those in Fig. 5.

relative balance between these two types of errors depends on the presentation mode and voice distance. Because the two-way interaction was significant, we again separated the data according to the mode of presentation. In the dichotic presentation mode, the voice effect was not significant $[\chi^2(1) = 1.56, p = 0.21]$. By contrast, in the diotic presentation mode, the participants made relatively more masker errors when the voice distance decreased between the target and masker $[\chi^2(1) = 13.66, p < 0.001]$. The *post hoc* comparisons in this condition showed that when the distance between the target and masker voice was 1.71 st, the participants selected the masker keyword more than when the distance was 5.13 st $[z = 3.43, p < 0.01]$ or 8.56 st $[z = 3.07, p < 0.01]$. There was no difference between the voices 5.13 and 8.56 st $[z = -0.20, p = 0.84]$.

Finally, there were significantly more masker responses than extraneous responses when stimuli were presented diotically with the 1.71-st voice $[z = 5.34, p < 0.001]$ but not in all of the other conditions $[p > 0.23]$. These results indicated that the participants were listening, at least partially, to the masker voice instead of the target voice in a difficult condition, such as a diotic presentation, with a small distance between the masker and target voice.

TABLE III. Equation (3) statistics.

Fixed effect	Statistics
Presentation	$\chi^2(1) = 182.53$, $p < 0.001$
Voice	$\chi^2(1) = 95.99$, $p < 0.001$
Position	$\chi^2(2) = 55.53$, $p < 0.001$
Presentation × voice	$\chi^2(1) = 68.18$, $p < 0.001$
Presentation × position	$\chi^2(2) = 3.28$, $p = 0.19$
Voice × Position	$\chi^2(2) = 18.60$, $p < 0.001$
Presentation × voice × position	$\chi^2(2) = 7.95$, $p < 0.05$

C. Discussion

Stimuli from the Long-SWoRD test, which are longer than those used in the behavioral speech-on-speech studies found in the literature, allow the contribution of perceptual mechanisms such as the participants' knowledge of the language (e.g., Warzybok *et al.*, 2015). In general, however, the results of our study were consistent with those obtained with shorter stimuli reported in the literature. The higher performance for dichotic than for diotic presentation observed in the current study is consistent with and can be explained by the spatial-separation advantage observed in earlier studies involving concurrent speech listening tasks (e.g., Broadbent, 1954; Cherry, 1953; Ericson and McKinley, 2001). The observed decrease in the performance with a decreasing F0/VTL distance between the target and masker voices was also in line with previous findings (e.g., Başkent and Gaudrain, 2016; Darwin *et al.*, 2003). The present results extend these previous findings to different stimuli that cover another type of real-life situation.

Unlike most previous studies, the Long-SWoRD approach requires a substantial role of memory over a rather long period of time. This is highlighted by the effect of the word position on the scores of the participants: the last presented keyword yielded higher scores than the previous two. Given that the time gap between the keyword positions was, on average, 4.8 s, this effect could not have been captured using stimuli that are typically shorter than 5 s. Interestingly, this recency effect was not affected by the binaural mode of presentation but instead was affected by the voice differences between the target and masker: when the task became more difficult, in the diotic-1.71-st condition, the recency effect vanished. Remarkably, this condition was also the condition in which the proportion of masker vs extraneous errors increased. These two results together indicate that in

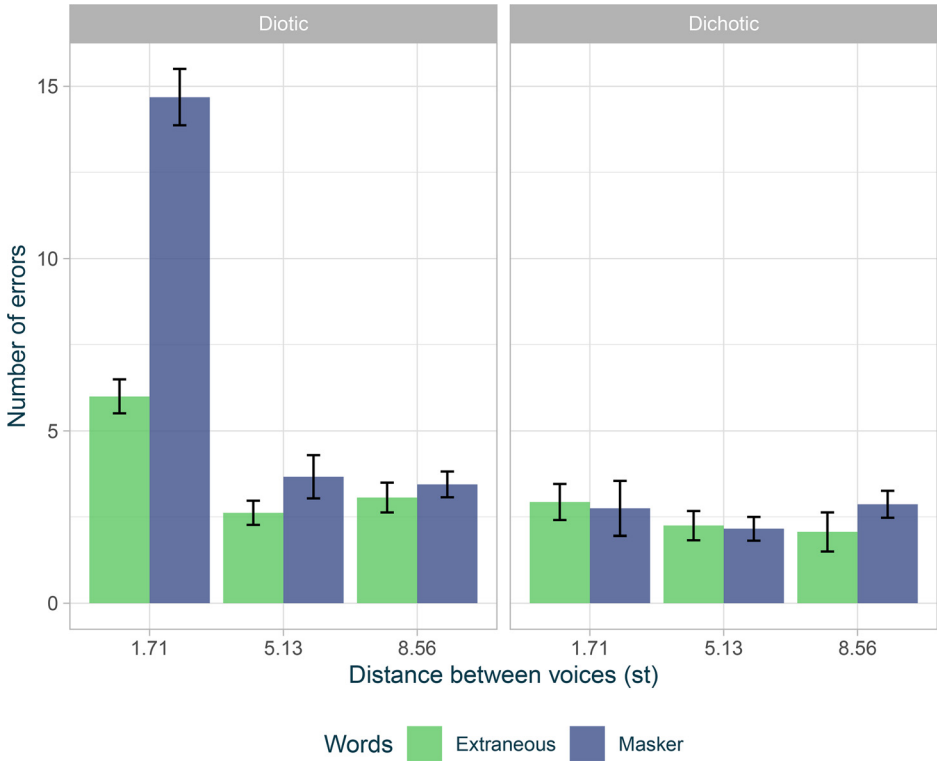

FIG. 7. (Color online) The average number of errors per participant for each condition out of 72 observations. The bars represent the masker answers (in dark blue) and extraneous answers (in light green). The error bars are the standard error of the mean across the participants.

this adverse listening condition, the participants may be confusing the two speakers. This interference may also lead them to change strategy, which could jeopardize some cognitive functions such as the working memory. When the two competing voices became more similar, the listeners seemed to be unable to maintain a selective attention listening strategy and were forced into divided attention. Koelewijn *et al.* (2014) showed that in the case of divided attention, the listening effort, as captured by the pupil dilation, increases compared to selective attention. In the present experiment, in the diotic-1.71-st condition, if the participants struggle to separate the two voices—which is suggested by the fact that they seem to be confusing the target and masker more frequently—they are effectively monitoring twice as much speech material as when the voice separation allows them to inhibit and ignore the masker voice. The recency effect could, thus, be disappearing in this condition because maintaining two sentences exceeds the capacity of the memory resources dedicated to this task.

One issue with the design of this experiment, however, which is made apparent in retrospect by the results shown in Fig. 5, relates to the fact that for a number of the conditions tested, many of the participants' scores were close to the ceiling. The Long-SWoRD method seems to provide the most interesting insight for conditions that are away from saturation as this is where the error patterns can be best analyzed. We, therefore, conducted a follow-up experiment, which focused on conditions that remain away from the ceiling by excluding the dichotic presentation and including a range of voice differences whose results would remain away from the ceiling.

TABLE IV. Equation (4) statistics.

Fixed effect	Statistics
Presentation	$\chi^2(1) = 6.58$, $p < 0.05$
Voice	$\chi^2(1) = 5.48$, $p < 0.05$
Voice × presentation	$\chi^2(1) = 7.87$, $p < 0.01$

IV. EXPERIMENT 2: SMALL TO LARGE VOICE DIFFERENCES

A. Method

1. Procedure and apparatus

In experiment 2, the procedure and material content were similar to those in experiment 1. However, the stimuli were presented only diotically, and six voice distances were presented. The data collection lasted 60–100 min, and the entire procedure was completed in a single session.

The apparatus was identical to that of experiment 1.

2. Stimuli

The masker voices of experiment 2 were created with the same analysis-synthesis used in experiment 1. To be able to compare the two experiments, the JND voice was kept (Gaudrain and Başkent, 2015). In addition to this voice, five new equidistant voices were synthesized. Because of the ceiling effect observed in experiment 1, it was decided that the largest distance between the target and masker voices should be 3.42 st, which was equidistant between 1.71 and 5.13 st. The parameter values for the six masker voices are displayed in Fig. 8 and Table V.

https://doi.org/10.1121/10.0007225

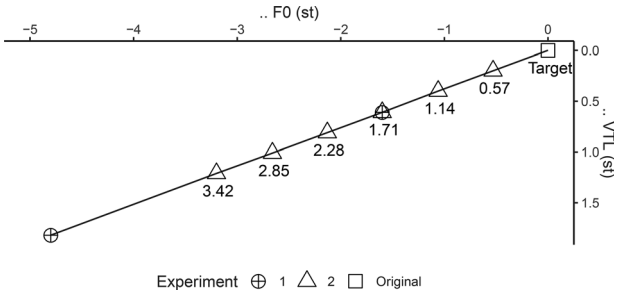

FIG. 8. The distances in semitones (st) between the target and masker voices for experiment 2. The experiment 1 masker voice is represented by circles and experiment 2 masker voices are represented by triangles. The voice 1.71 st is present in both experiments.

3. Participants

Thirty new participants (different from experiment 1), between 20 and 26 years old ($\mu = 21$), participated in this second experiment. All of them were native French speakers. Twenty-four participants had audiometric thresholds $\leq$20 dB HL and six participants had thresholds $\leq$25 dB HL at test frequencies between 250 Hz and 4 kHz. Twenty-three participants also had audiometric thresholds $\leq$20 dB HL and seven participants had thresholds $\leq$25 dB HL at test frequencies between 6 and 8 kHz. The participants provided informed consent before participating and were paid an hourly wage for their participation.

B. Results

1. General description

A gLMM was fitted on the binary (correct/incorrect) scores. The analysis methodology of experiment 2 is similar to that of experiment 1. Equation (5) shows the final model with a top-down strategy modeling,

$$\text{score} \sim \text{voice} + (\text{voice}|\text{subject}). \quad (5)$$

The participants had better scores when the distance between the target and masker voices increased $[\chi^2(1) = 244.46, p < 0.001]$ at a rate such that the odds of obtaining a correct response $[p/(1-p)]$ doubled for every 0.9 st of voice difference (See Fig. 9). The *post hoc* analysis with a *false discovery rate* correction showed that the performance was different for each pair of consecutive voices (see Table VI).

TABLE V. The distance between the target and masker voices in semitones for experiment 2.

Voice	ΔF0	ΔVTL	$\sqrt{\Delta F0^2 + \Delta VTL^2}$
1	−0.53	0.20	0.57
2	−1.06	0.40	1.14
3 (JND)	−1.60	0.61	1.71
4	−2.13	0.81	2.28
5	−2.67	1.01	2.85
6	−3.20	1.21	3.42

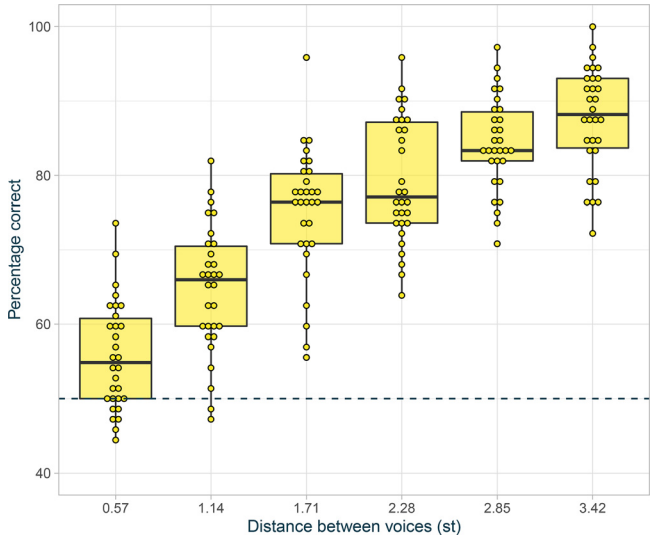

FIG. 9. (Color online) The percentage of correct responses in diotic listening as a function of the distance between the voices. The dots represent the scores for every participant in each condition. See Fig. 5 for a description of the boxplot characteristics. The dashed line (50%) indicates the level at which the performance is significantly greater than chance based on a binomial test at the 5% significance level.

2. Keyword position analysis

Equation (6) shows the final gLMM model when the position of the keyword in the story (beginning, middle, and end of the story) was added to the model,

$$\text{score} \sim \text{position} * \text{voice} + (1|\text{subject}). \quad (6)$$

The position of the keyword in the story had an effect on the participants' scores $[\chi^2(2) = 47.52, p < 0.001]$ as well as the voice $[\chi^2(1) = 755.66, p < 0.001]$ and interaction $[\chi^2(2) = 34.20, p < 0.001]$ (see Fig. 10). The *post hoc* analysis showed that there was no performance difference for the three keywords in the voice 0.57 $[\chi^2(2) = 0.68, p = 0.71]$ and voice 1.14 $[\chi^2(2) = 2.46, p = 0.29]$. There was, however, a recency effect for the four other voices as the participants had better scores for the end-keyword than for keywords in the beginning and middle of the story.

3. Error analysis

Figure 11 illustrates the error distribution, and Eq. (7) represents the final gLMM on the binary (masker-extraneous) data for experiment 2. The participants answered with the

TABLE VI. The *post hoc* comparisons for the voice factor.

Voice distance comparison	Statistics
0.57 vs 1.14	$z = -6.03$, $p < 0.001$
1.14 vs 1.71	$z = -7.47$, $p < 0.001$
1.71 vs 2.28	$z = -3.30$, $p < 0.01$
2.28 vs 2.85	$z = -4.22$, $p < 0.001$
2.85 vs 3.42	$z = -3.20$, $p < 0.01$

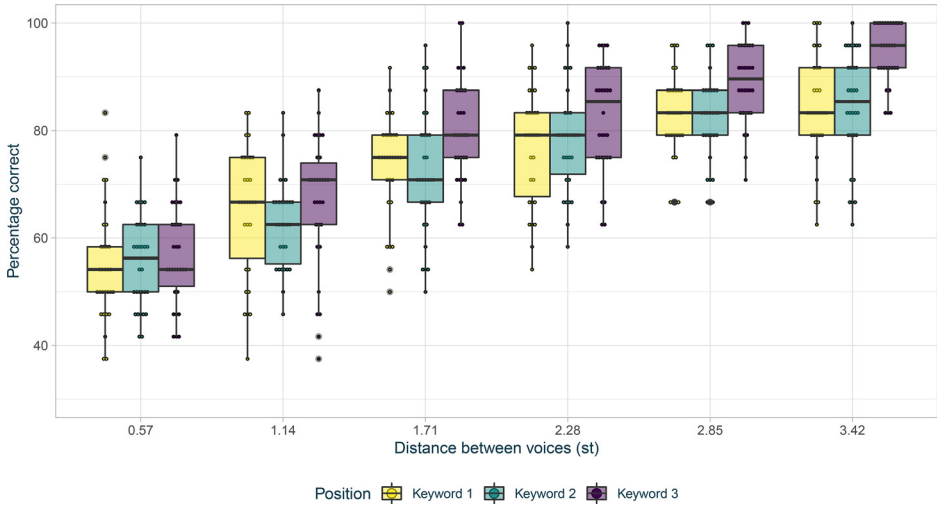

FIG. 10. (Color online) The percentage of correct responses for each voice for every keyword position. The details are identical to those in Fig. 5.

masker keyword over the extraneous keyword when the distance between the target and masker voices decreased [$\beta = -1.28$, standard error (SE) $= 0.16$; $z = -8.12$, $p < 0.001$],

$$\text{error type} \sim \text{voice} + (\text{voice}|\text{subject}). \quad (7)$$

The masker-extraneous ratio was above chance for all of the voices except when the distance between voices was 3.42 st (see Table VII). The chance level was computed for each voice with a binomial test.

C. Discussion

This second experiment investigated the participants' performance and voice distance relationship in the LongSWoRD test in diotic adverse listening conditions. First, the 1.71-st voice distance condition was present both in experiment 1 and in this experiment. We found that the participants' scores in this condition were not significantly different across the two experiments $[t(50) = -1.75, p = 0.09]$, indicating good test/retest repeatability. As in the diotic condition of experiment 1, the scores increased with the voice distance, but this effect—which manifested only in a single condition in experiment 1—is here demonstrated systematically across all of the conditions. We found, here, that the odds of being correct would double for about every semitone of the voice difference. This is a stronger voice difference dependency than what was observed by Başkent and Gaudrain (2016), as they found that 4.8 st in F0 or 2.0 st in VTL were needed to double the odds, which can be combined into 2.6 st needed to double the odds when increasing the voice distance along a diagonal combining F0 and VTL. This is particularly striking because Başkent and Gaudrain (2016) used an average TMR of −6 dB, which has been shown to be more favorable to voice difference effects than the 0 dB TMR that we have used in the present study (Nagels *et al.*, 2021; Brungart, 2001; Darwin *et al.*, 2003). Our result could be due to the length and linguistic richness of the stimuli used in the present study. Bricker and Pruzansky (1966) observed that talker identification improved with the number of phonemes available in the stimuli. More recently, Meister *et al.* (2016) found that listeners were more sensitive to voice gender differences in sentences than in words. These effects could be related to the fact that a longer speech context may provide enhanced voice consistency, which was shown to favor stream segregation (Samson and Johnsrude, 2016). Although none of these

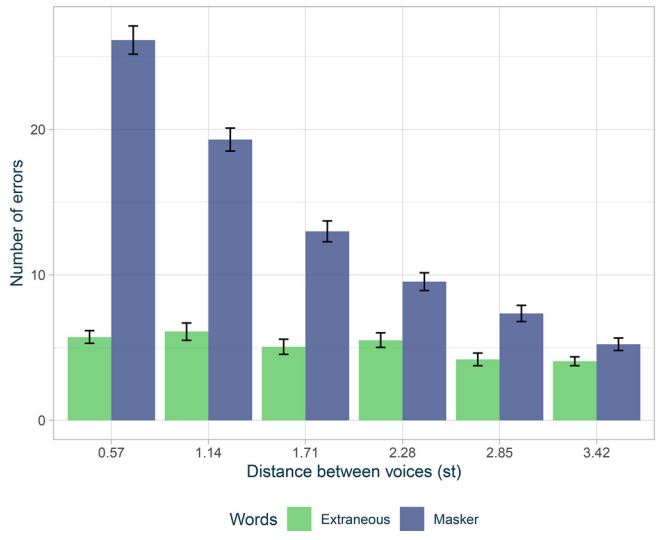

FIG. 11. (Color online) The average number of errors per participant for each voice condition out of 72 observations. The bars represent the masker answers (in dark blue) and extraneous answers (in light green). The error bars are the standard error of the mean across the participants.

TABLE VII. The comparisons with chance level for the masker-extraneous ratio.

Voice distance	Statistics
0.57	$z = 13.7$, $p < 0.001$
1.14	$z = 9.37$, $p < 0.001$
1.71	$z = 6.45$, $p < 0.001$
2.28	$z = 3.13$, $p < 0.01$
2.85	$z = 2.18$, $p < 0.05$
3.42	$z = 0.52$, $p = 0.6$

https://doi.org/10.1121/10.0007225

studies used material that was as long as the stories of the Long-SWoRD used here, the effect of the duration/richness that they report could perhaps explain the greater sensitivity to the voice differences that we observed in the present study. Overall, these considerations support the suggestion that allowing more time for the voice cue differences to build up auditory streams perhaps provides a larger voice difference benefit than with shorter stimuli.

Another potential explanation for this strong role of voice cues would be that the nature of the task is also different from the other speech-on-speech studies listed above. The Long-SWoRD task heavily relies on memory, as is illustrated by the recency effect that we observed. The fact that the recency effect disappears under the most challenging conditions could be interpreted as a sign that some cognitive processes involved in the task collapse in this condition. Notably, the reliance on the semantic context, which facilitates memory encoding, could be affected in situations where the voices become very similar, leading to a steeper decline in the performance than in experiments where semantic context would not be expected to play a strong role. To evaluate this effect, we have performed a computational analysis of the influence of the semantic context on the responses provided by the participants, which is presented in Sec. V.

V. ROLE OF ACOUSTIC AND LINGUISTIC PROPERTIES: LOCAL TMR AND SEMANTIC CONTEXT

In this section, we are assessing whether some acoustic and linguistic properties of the stimuli are able to predict some of the errors made by the participants. From an acoustics point of view, although the average TMR was fixed to 0 dB, there are momentary fluctuations that could lead the keywords from the masker to be more audible than the words of the target (e.g., Gaudrain and Carlyon, 2013).

A parallel can be drawn for the use of the semantic context: listeners may be more prone to answer with a keyword that belongs more strongly to the semantic context they were listening to than a keyword that is less strongly associated with the context. In other words, when the masker keyword is close to the target's semantic context, the listeners may be more likely to make a mistake and choose the masker keyword.

Our hypothesis was that these acoustic and linguistic predictors would only influence the results when the separability of the sources was not obvious. In other words, in line with the conclusions of Zekveld *et al.* (2014), we are expecting to find an effect of these factors only in the diotic condition, where binaural cues were not present to support selective attention. We expected that local TMR biases may occur for all of the voice differences, as it primarily operates at the peripheral level, i.e., before or at the same time as the voice cues are processed. In contrast, we expected that the influence of the semantic context would only show in the most challenging voice conditions, i.e., when peripheral cues have not been useful for stream segregation.

For each of these two cues—local TMR and semantic context—we first built a numerical predictor that we then introduced in our statistical model to evaluate their contribution to the observed scores. Our main goal was to perform simple complementary analyses. As such, some modeling decisions are arbitrary as there could be numerous ways of calculating the influence of the local TMR and semantic context.

A. Local TMR

1. Methodology

In experiments 1 and 2, the distance between the target and masker voices was controlled. However, the level of the voice was not constant throughout the trials either because of the natural modulations occurring in speech or due to the liveliness of the narrator. As a consequence, it is possible that for short periods of time, the TMR was more favorable to the masker than to the target. As illustrated in Fig. 1, the target and masker did not necessarily overlap in time (less than 20% of the keyword pairs are overlapping in time). To address all of the situations for each target-masker keyword pair, we calculated the TMR over a time interval that started at the beginning of the keyword that appeared first and ended at the end of the keyword that ended last. This local TMR over the keywords was used directly as a numerical predictor of the individual responses of the participants for each keyword position.

2. Results

Experiment 1—A gLMM was fitted on the binary data (target/masker) with the local TMR, the distance between voices, and the stimulus presentation. Equation (8) shows the final model with a top-down modeling,

$$\text{score} \sim \text{presentation} * (\text{voice} + \text{TMR}) + (\text{voice}|\text{subject}). \quad (8)$$

Regarding the stimulus presentation and distance between voices, the results were similar to those of the previous analysis (see Table VIII). The participants obtained better scores when the TMR was higher but only when the stimuli were presented diotically $[\beta = 4.90, \text{SE} = 0.75; z = 6.42, p < 0.001]$ in contrast to being presented dichotically $[\beta = 1.14, \text{SE} = 1.31; z = 0.88, p = 0.38]$.

Experiment 2—Similar to experiment 1, Eq. (9) shows the final model for experiment 2,

TABLE VIII. Equation (8) statistics.

Fixed effect	Statistics
Presentation	$\chi^2(1) = 135.06$, $p < 0.001$
Voice	$\chi^2(1) = 102.98$, $p < 0.001$
TMR	$\chi^2(1) = 37.30$, $p < 0.001$
Presentation × TMR	$\chi^2(1) = 7.61$, $p < 0.01$
Presentation × voice	$\chi^2(1) = 73.7$, $p < 0.001$

$$\text{score} \sim \text{voice} + \text{TMR} + (\text{voice}|\text{subject}). \quad (9)$$

The participants had better scores when the distance between the voices increased [$\beta = 2.31, \text{SE} = 0.15; z = 15.72, p < 0.001$], as well as when the local TMR was higher [$\beta = 3.50, \text{SE} = 0.38; z = 9.22, p < 0.001$]. The interaction between these two factors, however, was not significant [$\beta = 0.94, \text{SE} = 1.21; z = 0.77, p = 0.44$].

3. Discussion

Following our expectations, the results of this additional analysis indicate that the local TMR does influence the participants' errors but only when the stimuli are presented diotically. In the dichotic condition (experiment 1), it appears that the binaural cue is so strong that the small local TMR fluctuations are irrelevant. Thus, in this context, even when the two voices are very close to each other, the local TMR does not influence the participants' responses.

B. Semantic context

1. Methodology

Although all of the stories used in the experiments described above came from the same audiobook and, therefore, belonged broadly to the same lexical field, the participants may have been able to use the local semantic context cues to identify the correct answer. The additional analysis described below sought to test this hypothesis by using a quantified measure of the semantic context for each story. The semantic-context measure used here was inspired by the earlier work of Broderick *et al.* (2018). The semantic similarity was evaluated using a metric derived from the word2vec algorithm (Mikolov *et al.*, 2013). The word2vec process analyzes a corpus to yield a word representation that consists of (relatively) low-dimensional vectors, derived from the context in which the word is used in the corpus. The words that are used interchangeably in the same context yield the same vector representation and can be considered synonyms. Therefore, for a word A and a word B, a semantic distance metric—thereafter noted $\varphi(A,B)$—can be calculated from this vector representation such that zero represents the synonyms and one represents two words that are semantically unrelated.

The word2vec model requires training on a corpus. For that purpose, we could have trained it on the Long-SWoRD material or the whole book from which the material was extracted. However, we were not interested in the specific semantic relationships that exist in the material but were, instead, interested in the semantic relationships that the participants have inferred through their exposure to the language. To estimate the semantic knowledge, we used a model provided by Gaudrain and Crouzet (2019), which was trained on a much larger database: the entirety of the French Wikipedia.

To assess whether the participants relied on the semantic information in the task, in a trial, we estimated whether the target keyword or masker keyword was closer semantically to the target story. In each trial, we then compared the semantic distance between the target keyword and target story to the semantic distance between the corresponding masker keyword and target story. Both were compared to the target story as we assumed that the participants were trying to listen to the target from the beginning until the end of the trial.

Mathematically, each target story is a sequence of n words T_i such that the target story is represented as $\{T_{1,}, T_2, T_3, \ldots, T_n\}$. Similarly, the masker story, composed of p words M_j, is represented as $\{M_{1,}, M_2, M_3, \ldots, M_p\}$. Within each target story, i.e., among the words $(i \in 1, \ldots, n)$, the participants have to find three target keywords, $i \in [t_1, t_2, t_3]$, corresponding to the beginning, middle, and end positions. The three target keyword positions, t_1–t_3, can vary from 2 to $n - 1$ for the target story. Additionally, in the masker story, the indices of the three masker keywords, m_1, m_2, m_3, can vary from 2 to $p - 1$. The distance between a target keyword, indexed k, and a target story is estimated as the average distance between the keyword and individual words constituting the context

$$\bar{\varphi}_k^T = \frac{1}{n-1} \sum_{i \neq t_k} \varphi(T_i, T_{t_k}).$$

Similarly, the distance between the kth masker keyword and the target story can be calculated as

$$\bar{\varphi}_k^M = \frac{1}{n-1} \sum_{i \neq t_k} \varphi(T_i, M_{m_k}).$$

We then compare these two distances to obtain the semantic effect predictor,

$$\begin{aligned} \Phi_k &= \bar{\varphi}_k^T - \bar{\varphi}_k^M \\ &= \frac{1}{n-1} \sum_{i \neq t_k} \varphi(T_i, T_{t_k}) - \frac{1}{n-1} \sum_{i \neq t_k} \varphi(T_i, M_{m_k}). \end{aligned} \quad (10)$$

Because φ varies from zero to one, Φ_k can vary from -1 to one. If the semantic effect predictor Φ_k is positive, it shows that the target keyword is closer to the target story than the masker keyword. If participants are getting help from the semantic context, they would then select the keyword from the target story more frequently. On the other hand, if Φ_k is negative, it may show that the masker keyword is closer to the target story than the target keyword, and the participant may be biased toward the masker keyword. In practice, in the present implementation of the Long-SWoRD test, Φ_k varied from -0.2 to 0.47. The fact that the semantic effect predictor distribution was not symmetrical illustrated that most of the target keywords were more semantically related to the target story than the masker keywords.

2. Results

Experiment 1—A gLMM was fitted on the binary data (target/masker) with the semantic context, the distance

https://doi.org/10.1121/10.0007225

TABLE IX. Equation (11) statistics.

Fixed effect	Statistics
Presentation	$\chi^2(1) = 132.67$, $p < 0.001$
Voice	$\chi^2(1) = 89.53$, $p < 0.001$
Semantic	$\chi^2(1) = 16.23$, $p < 0.001$
Voice × semantic	$\chi^2(1) = 4.71$, $p < 0.05$
Presentation × voice	$\chi^2(1) = 70.28$, $p < 0.001$

between voices, and the stimulus presentation. Equation (11) shows the final model with a top-down modeling,

$$\text{score} \sim \text{voice} * (\text{presentation} + \Phi_k) + (\text{voice}|\text{subject}). \quad (11)$$

Regarding the stimulus presentation and distance between the voices, the results were similar to the previous analysis (see Table IX). The semantic context influenced the participants' answers but not in the same way across the voice conditions. The *post hoc* analysis showed that the participants had better scores when the target keyword was semantically closer to the target story than the masker keyword only for voice 1.71 st $[z = 3.05, p < 0.01]$ and voice 5.13 st $[z = 2.48, p < 0.05]$ but not for voice 8.56 $[z = -0.20, p = 0.80]$.

Experiment 2—Similar to experiment 1, Eq. (12) shows the final model for experiment 2,

$$\text{score} \sim \text{voice} * \Phi_k + (\text{voice} * \Phi_k|\text{subject}). \quad (12)$$

The participants had better scores when the distance between the voices increased $[\beta = 2.34, \text{SE} = 0.15; z = 15.27, p < 0.001]$ as well as when the semantic context predictor was higher $[\beta = 1.24, \text{SE} = 0.23; z = 5.5, p < 0.001]$. The interaction between these two factors was also significant $[\beta = 1.48, \text{SE} = 0.6; z = 2.46, p < 0.05]$. The *post hoc* analyses showed that the semantic context significantly influenced the performance only for voice 1.71 $[z = 2.89, p < 0.05]$, voice 2.28 $[z = 2.28, p < 0.05]$, voice 2.85 $[z = 2.82, p < 0.05]$, and voice 3.42 $[z = 2.67, p < 0.05]$ but not for voices 0.57 $[z = 2.02, p = 0.052]$ and 1.14 $[z = 0.91, p = 0.36]$.

3. Discussion

The results of this additional analysis indicate that the participants can use a general semantic context to find the keywords belonging to the target story. However, the participants do not seem to use this information in two cases. First, they did not benefit from this information when the distance between the voices was greater than 5.13 st. This outcome could be derived from the fact that the voice difference in this condition was sufficient to create two clearly distinct streams, and using the semantic context to resolve ambiguities is not necessary to reach the ceiling performance. A related, potential explanation is that there were simply too few errors to be able to capture the variations that could be attributed to the embedding of the semantic context. Either way, the scores are too high in this condition for the effect to manifest itself.

Second, the participants did not use the semantic context when the target and masker voices were very close to each other, i.e., when the voice distance was smaller than or equal to 1.14 st. Although this may appear counterintuitive at first sight, one potential explanation could be that the difficulty of these conditions makes it impossible for the participants to access the semantic context. Indeed, these voice differences are below the JND reported by Gaudrain and Başkent (2015) for a similar voice manipulation (from female to male) and are, therefore, barely noticeable. It may be that under these conditions, the target and masker voices are not separated enough to yield two distinct speech streams. As such, the words that are perceived within the mix cannot be attributed easily to one speaker or the other. As a result, the semantic context—which is only relevant within a stream—cannot build up as efficiently and cannot be exploited by the participant.

In conclusion, when the task becomes difficult for the subjects, the participants can benefit from the semantic information to find the target keywords but only if they have managed to access the semantic context of the target story in the first place. These findings are reminiscent of the studies on "phonemic restoration," another phenomenon in which the semantic context has been shown to be instrumental. Phonemic restoration, also called "top-down repair," is thought to represent the intervention of top-down, cognitive processes in filling in missing information on the basis of linguistic knowledge. It was shown that phonemic restoration is hindered when the words of a sentence are presented in reverse order, thus, preserving the lexical content but disrupting the syntax and semantic context (Bashford and Warren, 1987). In line with our results, phonemic restoration seems to appear only when the performance is neither exceptionally good nor exceptionally bad. For instance, Bhargava *et al.* (2014) observed that phonemic restoration occurred for normal-hearing (NH) participants when half of the signal was obliterated (50% duty cycle) but not when only a quarter of the signal was removed (75% duty cycle). At the other end, at 50% duty cycle, cochlear implant (CI) users showed phonemic restoration only if their baseline speech scores were sufficiently high, whereas at 75% duty cycle, all of the CI participants showed phonemic restoration. Our current results align with these observations: the benefit of semantic context is only observable when the context has the potential to build up and there is space for this build up to contribute to an improved performance.

VI. GENERAL DISCUSSION AND CONCLUSION

One of the main purposes of this study was to introduce a new behavioral paradigm—the *Long-SWoRD* test—enabling one to retrospectively infer the attentional fluctuations in an auditory selective attention task with concurrent voices, which will be useful for the study of the neural correlates of selective attention. This task allows the tracking

of the temporal fluctuations of the voice that participants are listening to at three key points in the story (beginning, middle, and end). We showed that, despite the apparent difficulty of the proposed task, the participants obtained rather high scores and proficiency in the task can be manipulated by systematically modifying the voice difference between the two competing speakers. These characteristics make the task particularly suitable for neurophysiological studies.

Another aim of this paper was to examine how listeners segregate two speech streams in the context of longer stimuli in behavioral studies where both the streaming build-up effect and participants' cognitive mechanisms, such as linguistic knowledge and working memory, are allowed to fully contribute. By and large, the results of the present research are consistent with those of earlier studies in the literature. Primitive segregation cues, such as spatialization and vocal characteristics, clearly influenced the participants' performances. The advantage of a dichotic listening condition over a diotic listening condition is consistent with previous studies (Broadbent, 1954; Cherry, 1953; Ericson and McKinley, 2001). The advantage of a large distance between voices over a small distance is also in accordance with previous studies (Başkent and Gaudrain, 2016; Darwin *et al.*, 2003; Ives *et al.*, 2010; Vestergaard *et al.*, 2009). In addition, we found that the participants can also benefit from the semantic information, which is also consistent with the results of previous studies (Aydelott and Bates, 2004; Clarke *et al.*, 2014; Dekerle *et al.*, 2014; Freyman *et al.*, 2001; Helfer and Freyman, 2009; Hoen *et al.*, 2007; Iyer *et al.*, 2010). However, here we showed that this effect is limited to conditions where the voice differences are large enough to let the listeners access the semantic context of the target story. Additional analyses[1] also show that the participants can use other linguistic properties, such as the neighborhood density and word frequency (Luce and Pisoni, 1998), to direct their answers.

Masker errors, which reflect the difficulty ignoring the masker, may stem from difficulties in segregating the two voices, selecting the target voice and suppressing the masker voice, remembering the words, and whether they were uttered by the target or masker voice, or a combination of these. Szalárdy *et al.* (2021) showed that the listeners balance these subtasks differently depending on the tempo of the speech material. In their study, when a faster tempo was used, the listeners seemed to favor a strategy where they merely detected the disconnected target words rather than continuously following the target stream. Although this study used slowed-down and sped-up speech to modulate the difficulty of the task, a parallel can be drawn with our results. Through our semantic context analysis, we showed that in the most challenging conditions, the participants did not rely as much on the semantic context as in the more mildly challenging conditions. It is possible that, instead, they focused on catching individual words. Moreover, in the study by Szalárdy *et al.* (2021), they observed that in the sped-up conditions, the participants also showed a reduced recognition memory performance. Similarly, we found that in our most difficult conditions, the participants did not display any recency effect, which is in line with the idea that the memory capacity may be reduced in such adverse conditions.

In our task, the participants do not know where the keywords will be located, hence, they are tasked with maintaining a list of the potential keywords in memory, all while listening to new incoming items from the target and ignoring the speech coming from the masker. Such a situation in which a listener is trying to maintain a sequence of items in memory while other sounds to be ignored are presented has been largely studied under the name of irrelevant speech effect (ISE; e.g., Schlittmeier *et al.*, 2008). One way to measure the ISE is to present a sequence of verbal elements, followed by an "irrelevant" distractor, before asking the participants to recall the original sequence. The error rate in such experiments increases with the serial position of the item in the sequence. Nevertheless, the last item benefits from a recency effect, which is similar to what we have observed in most of the conditions in our study. In the Long-SWoRD, the different keyword pairs are followed by a relatively long sequence of material. The first and second keyword pairs are followed by more than 10 and 5 s of stimulus, respectively, and even the last keyword is followed by about 2 s of concurrent speech. In the stimuli, one of the streams, the target, is meant to be followed by the listener and would constitute the sequence that needs to be recalled, whereas the masker is meant to be ignored, constituting the irrelevant speech. Schlittmeier *et al.* (2008) found that voice similarity between the relevant and irrelevant sound has no effect on the recall abilities. Yet, in our data, we found that the recency effect disappeared when the voices became very similar. This effect could result from the mechanisms captured by the ISE paradigm not being entirely relevant to the present situation—for instance, because the segment of the target that is presented at the same time as the irrelevant sound still needs to be monitored. But another potential explanation is that in our experiments, the participants have to recall not only the keywords they heard but also who uttered them. Neely and LeCompte (1999) showed that unlike voice similarity, the semantic similarity of the irrelevant sound does interfere with the recall. It is possible that when the information that needs to be recalled—whether it be semantic or vocal identity—is similar between the sequence and irrelevant distractor, it creates interference that hinders recall. In other words, in our experiment, when the voices became more similar, the participants may have become less able to recall whether a word they remembered was uttered by the target or masker voice. This would yield more masker errors, which is precisely what we have also observed in these challenging conditions.

Overall, our results seem to confirm that using long, coherent stimuli does indeed engage cognitive mechanisms that may not be observable otherwise. In their meta-analysis, Dryden *et al.* (2017) note that the association between the working memory and speech in noise performance becomes stronger when there is an increasing

difficulty of the task. This difficulty can be reached either with an increased amount of informational masking or by using longer, linguistically meaningful speech stimuli as the target. Each of these elements place the Long-SWoRD test at the most difficult extreme. Although the Long-SWoRD test was primarily designed for attention decoding in neurophysiological studies, it may also be valuable for the purpose of evaluating the interaction between the acoustic cues and cognitive mechanisms involved in the speech-on-speech perception. This would be particularly useful, for instance, to study the combined effects of aging and hearing loss on the concurrent speech perception.

To conclude, the test method and stimuli used in this study provide a tool for researchers to infer which of the two competing speakers the participants listened to at different points in time. The method could be particularly useful in the context of neurophysiological studies of selective auditory attention to speech; for example, TRF calculations could be improved by taking into account the information about the time points during which the participant is listening to the target voice, the masker voice, or neither of the voices (Huet *et al.*, 2021). Moreover, the analyses of the response patterns while listening to the stories combined with the analyses of the error patterns as well as the acoustic and linguistic context effects provide a new insight into the complexity of speech-on-speech perception while offering the opportunity to highlight the relations between primitive auditory scene analysis and cognitive aspects of speech perception.

ACKNOWLEDGMENTS

The authors thank Alexandra Corneyllie for her technical support; Fanny Meunier, Carolyn McGettingan, Deniz Başkent, Thomas Koelewijn and Amy Braun for comments on earlier versions of this manuscript; and Lucie Vallet and Alice Witkowski for their assistance with the data collection. This work was supported by the LabEx CeLyA ("Centre Lyonnais d'Acoustique," Grant No. ANR-10-LABX-0060) operated by the French National Research Agency.

[1]See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0007225 for the more detailed analyses.

[2]This section contains a new analysis of the data that were partially published in Huet *et al.* (2018).

Akram, S., Simon, J. Z., and Babadi, B. (**2017**). "Dynamic estimation of the auditory temporal response function from MEG in competing-speaker environments," IEEE Trans. Biomed. Eng. **64**, 1896–1905.

Aydelott, J., and Bates, E. (**2004**). "Effects of acoustic distortion and semantic context on lexical access," Lang. Cogn. Process. **19**, 29–56.

Bashford, J. A., and Warren, R. M. (**1987**). "Multiple phonemic restorations follow the rules for auditory induction," Percept. Psychophys. **42**, 114–121.

Başkent, D., and Gaudrain, E. (**2016**). "Musician advantage for speech-on-speech perception," J. Acoust. Soc. Am. **139**, EL51–EL56.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2014**). "Fitting linear mixed-effects models using lme4," arXiv:1406.5823.

Besser, J., Koelewijn, T., Zekveld, A. A., Kramer, S. E., and Festen, J. M. (**2013**). "How linguistic closure and verbal working memory relate to speech recognition in noise—A review," Trends Amplif. **17**, 75–93.

Best, V., Swaminathan, J., Kopčo, N., Roverud, E., and Shinn-Cunningham, B. (**2018**). "A 'buildup' of speech intelligibility in listeners with normal hearing and hearing loss," Trends Hear. **22**, 233121651880751.

Bhargava, P., Gaudrain, E., and Başkent, D. (**2014**). "Top–down restoration of speech in cochlear-implant users," Hearing Res. **309**, 113–123.

Biesmans, W., Vanthornhout, J., Wouters, J., Moonen, M., Francart, T., and Bertrand, A. (**2015**). "Comparison of speech envelope extraction methods for EEG-based auditory attention detection in a cocktail party scenario," in *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 5155–5158.

Bregman, A. S. (**1978**). "Auditory streaming is cumulative," J. Exp. Psychol. Hum. Percept. Perform. **4**, 380–387.

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (The MIT Press, Cambridge, MA), 773 pp.

Bressler, S., Masud, S., Bharadwaj, H., and Shinn-Cunningham, B. (**2014**). "Bottom-up influences of voice continuity in focusing selective auditory attention," Psychol. Res. **78**, 349–360.

Bricker, P. D. and Pruzansky, S. (**1966**). "Effects of stimulus content and duration on talker identification," J. Acoust. Soc. Am. **40**(6), 1441–1449.

Broadbent, D. (**1954**). "The role of auditory localization in attention and memory span," J. Exp. Psychol. **47**, 191–196.

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., and Lalor, E. C. (**2018**). "Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech," Curr. Biol. **28**, 803–809.

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

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

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

Clarke, J., Gaudrain, E., Chatterjee, M., and Başkent, D. (**2014**). "T'ain't the way you say it, it's what you say—Perceptual continuity of voice and top–down restoration of speech," Hear. Res. **315**, 80–87.

Conway, A. R. A., Cowan, N., and Bunting, M. F. (**2001**). "The cocktail party phenomenon revisited: The importance of working memory capacity," Psychon. Bull. Rev. **8**, 331–335.

Crosse, M. J., Butler, J. S., and Lalor, E. C. (**2015**). "Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions," J. Neurosci. **35**, 14195–14204.

Darwin, C. J., Turvey, M. T., and Crowder, R. G. (**1972**). "An auditory analogue of the Sperling partial report procedure: Evidence for brief auditory storage," Cognitive Psychology **3**(2), 255–267.

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**, 2913–2922.

Dekerle, M., Boulenger, V., Hoen, M., and Meunier, F. (**2014**). "Multi-talker background and semantic priming effect," Front. Hum. Neurosci. **8**, 878.

Di Liberto, G. M., O'Sullivan, J. A., and Lalor, E. C. (**2015**). "Low-frequency cortical entrainment to speech reflects phoneme-level processing," Curr. Biol. **25**(19), 2457–2465.

Ding, N., and Simon, J. Z. (**2012**). "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," J. Neurophysiol. **107**, 78–89.

Dryden, A., Allen, H. A., Henshaw, H., and Heinrich, A. (**2017**). "The association between cognitive performance and speech-in-noise perception for adult listeners: A systematic literature review and meta-analysis," Trends Hear. **21**, 2331216517744675.

Enders, G., Enders, J., and Liber, I. (**2015a**). *Le Charme Discret de L'intestin: Tout Sur un Organe Mal Aimé* (Gut: The inside Story of Our Body's Most Under-Rated Organ) (Éditions de Noyelles, Paris).

Enders, G., Enders, J., and Shaw, D. (**2015b**). *Gut: The inside Story of Our Body's Most Under-Rated Organ* (Greystone Books, Vancouver, Canada).

Enders, G., Monceau, J., and Liber, I. (**2016**). *Le Charme Discret de L'intestin: Livre Audio (Gut: The inside Story of Our Body's Most Under-Rated Organ)* (Audiolib, Paris).
Ericson, M. A., and McKinley, R. L. (**2001**). "The intelligibility of multiple talkers separated spatially in noise" (No. AFRL-HE-WP-SR-2001-0009), Air Force Research Laboratory Wright-Patterson AFB OH Human Effectiveness Directorate, available at https://apps.dtic.mil/docs/citations/ADA395035 (Last viewed 12/05/2021).
Fox, J., and Weisberg, S. (**2019**). *An R Companion to Applied Regression*, Third ed. (Sage, Thousand Oaks, CA), available at https://socialsciences.mcmaster.ca/jfox/Books/Companion/ (Last viewed 12/05/2021).
Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2001**). "Spatial release from informational masking in speech recognition," J. Acoust. Soc. Am. **109**, 2112–2122.
Gaudrain, E., and Başkent, D. (**2015**). "Factors limiting vocal-tract length discrimination in cochlear implant simulations," J. Acoust. Soc. Am. **137**, 1298–1308.
Gaudrain, E., and Carlyon, R. P. (**2013**). "Using Zebra-speech to study sequential and simultaneous speech segregation in a cochlear-implant simulation," J. Acoust. Soc. Am. **133**, 502–518.
Gaudrain, E., and Crouzet, O. (**2019**). "word2vec model trained on lemmatized French Wikipedia 2018," Zenodo. https://doi.org/10.5281/zenodo.3241447
Helfer, K. S., and Freyman, R. L. (**2009**). "Lexical and indexical cues in masking by competing speech," J. Acoust. Soc. Am. **125**, 447–456.
Herrmann, B., and Johnsrude, I. S. (**2020**). "Absorption and enjoyment during listening to acoustically masked stories," Trends Hear. **24**, 233121652096785.
Hoen, M., Meunier, F., Grataloup, C.-L., Pellegrino, F., Grimault, N., Perrin, F., Perrot, X., and Collet, L. (**2007**). "Phonetic and lexical interferences in informational masking during speech-in-speech comprehension," Speech Commun. **49**, 905–916.
Huet, M.-P. (**2020**). "Voice mixology at a cocktail party: Combining behavioural and neural tracking for speech segregation," Ph.D. thesis, INSA Lyon, Lyon, France, available at https://tel.archives-ouvertes.fr/tel-03178835 (Last viewed 12/05/2021).
Huet, M.-P., Micheyl, C., Gaudrain, E., and Parizet, E. (**2018**). "Who are you listening to? Towards a dynamic measure of auditory attention to speech-on-speech," in *Proceedings of Interspeech 2018*, Hyderabad, India, pp. 2272–2275.
Huet, M.-P., Micheyl, C., Parizet, E., and Gaudrain, E. (**in press**). "Behavioral account of attended stream enhances neural tracking," Front. Neurosci. (in press).
Ives, D. T., Vestergaard, M. D., Kistler, D. J., and Patterson, R. D. (**2010**). "Location and acoustic scale cues in concurrent speech recognition," J. Acoust. Soc. Am. **127**, 3729–3737.
Iyer, N., Brungart, D. S., and Simpson, B. D. (**2010**). "Effects of target-masker contextual similarity on the multimasker penalty in a three-talker diotic listening task," J. Acoust. Soc. Am. **128**, 2998–3010.
Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (**1999**). "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Commun. **27**, 187–207.
Kidd, G., Mason, C. R., and Best, V. (**2014**). "The role of syntax in maintaining the integrity of streams of speech," J. Acoust. Soc. Am. **135**, 766–777.
Koelewijn, T., Shinn-Cunningham, B. G., Zekveld, A. A., and Kramer, S. E. (**2014**). "The pupil response is sensitive to divided attention during speech processing," Hear. Res. **312**, 114–120.
Lewis, R. L., Vasishth, S., and Van Dyke, J. A. (**2006**). "Computational principles of working memory in sentence comprehension," Trends Cognit. Sci. **10**, 447–454.
Luce, P. A., and Pisoni, D. B. (**1998**). "Recognizing spoken words: The neighborhood activation model," Ear Hear. **19**(1), 1–36.
Mathôt, S., Schreij, D., and Theeuwes, J. (**2012**). "OpenSesame: An open-source, graphical experiment builder for the social sciences," Behav. Res. Methods **44**, 314–324.
Meister, H., Fürsen, K., Streicher, B., Lang-Roth, R., and Walger, M. (**2016**). "The use of voice cues for speaker gender recognition in cochlear implant recipients," J. Speech Lang. Hear. Res. **59**(3), 546–556.
Mesgarani, N., and Chang, E. F. (**2012**). "Selective cortical representation of attended speaker in multi-talker speech perception," Nature **485**, 233–236.
Meister, H., Fürsen, K., Streicher, B., Lang-Roth, R., and Walger, M. (**2016**). "The use of voice cues for speaker gender recognition in cochlear implant recipients,," J. Speech Lang. Hear. Res. **59**(3), 546–556.
Mikolov, T., Chen, K., Corrado, G., and Dean, J. (**2013**). "Efficient estimation of word representations in vector space," arXiv:1301.3781.
Miran, S., Akram, S., Sheikhattar, A., Simon, J. Z., Zhang, T., and Babadi, B. (**2018**). "Real-time tracking of selective auditory attention from M/EEG: A Bayesian filtering approach," Front. Neurosci. **12**, 262.
Mirkovic, B., Debener, S., Jaeger, M., and Vos, M. D. (**2015**). "Decoding the attended speech stream with multi-channel EEG: Implications for online, daily-life applications," J. Neural Eng. **12**, 046007.
Moore, B. C. J., and Gockel, H. (**2002**). "Factors influencing sequential stream segregation," Acta Acustica United with Acustica **88**(3), 320–333.
Nagels, L., Gaudrain, E., Vickers, D., Hendriks, P., and Başkent, D. (**2021**). "School-age children benefit from voice gender cue differences for the perception of speech in competing speech," J. Acoust. Soc. Am. **149**, 3328–3344.
Neely, C. B., and LeCompte, D. C. (**1999**). "The importance of semantic similarity to the irrelevant speech effect," Mem. Cogn. **27**, 37–44.
New, B., Pallier, C., Ferrand, L., and Matos, R. (**2001**). "Une base de données lexicales du Français contemporain sur internet: LEXIQUE™," ("A lexical database for contemporary French: LEXIQUE™), Année Psychol. **101**, 447–462.
O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., and Slaney, M. (**2015**). "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," Cereb. Cortex **25**, 1697–1706.
Rennies, J., Best, V., Roverud, E., and Kidd, G. (**2019**). "Energetic and informational components of speech-on-speech masking in binaural speech intelligibility and perceived listening effort," Trends Hear. **23**, 233121651985459.
Samson, F., and Johnsrude, I. S. (**2016**). "Effects of a consistent target or masker voice on target speech intelligibility in two- and three-talker mixtures," J. Acoust. Soc. Am. **139**, 1037–1046.
Schlittmeier, S. J., Hellbrück, J., and Klatte, M. (**2008**). "Can the irrelevant speech effect turn into a stimulus suffix effect?," Quart. J. Exp. Psychol. **61**(5), 665–673.
Shinn-Cunningham, B. G., and Best, V. (**2008**). "Selective attention in normal and impaired hearing," Trends Amplif. **12**, 283–299.
Slaney, M., Lyon, R. F., Garcia, R., Kemler, B., Gnegy, C., Wilson, K., Kanevsky, D., Savla, S., and Cerf, V. G. (**2020**). "Auditory measures for the next billion users," Ear Hear. **41**, 131S–139S.
Szalárdy, O., Tóth, B., Farkas, D., Hajdu, B., Orosz, G., and Winkler, I. (**2021**). "Who said what? The effects of speech tempo on target detection and information extraction in a multi-talker situation: An ERP and functional connectivity study," Psychophysiology **58**(3), e13747.
Treisman, A. (**1964**). "Monitoring and storage of irrelevant messages in selective attention," J. Verbal Learning Verbal **3**(6), 449–459.
Vestergaard, M. D., Ives, D. T., and Patterson, R. D. (**2009**). "The advantage of spatial and vocal characteristics in the recognition of competing speech," in *Proceedings of the International Symposium on Auditory and Audiological Research*, Vol. 2, pp. 535–544.
Warzybok, A., Brand, T., Wagener, K. C., and Kollmeier, B. (**2015**). "How much does language proficiency by non-native listeners influence speech audiometric tests in noise,," Int. J. Audiol **54**, 88–89.
Zekveld, A. A., Rudner, M., Kramer, S. E., Lyzenga, J., and Rönnberg, J. (**2014**). "Cognitive processing load during listening is reduced more by decreasing voice similarity than by increasing spatial separation between target and masker speech," Front. Neurosci **8**, 88.
Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A., and Smith, G. M. (**2009**). *Mixed Effects Models and Extensions in Ecology with R* (Springer, New York).