Hearing Research 433 (2023) 108770

Contents lists available at ScienceDirect

Hearing Research

journal homepage: www.elsevier.com/locate/heares

Research Paper

The effect of topic familiarity and volatility of auditory scene on selective auditory attention

Jonghwa Jeonglok Park[a,b,#], Seung-Cheol Baek[a,c,#], Myung-Whan Suh[d], Jongsuk Choi[a,e], Sung June Kim[b], Yoonseob Lim[a,f,*]

[a] Center for Intelligent & Interactive Robotics, Artificial Intelligence and Robot Institute, Korea Institute of Science and Technology, Seoul 02792, South Korea
[b] Department of Electrical and Computer Engineering, College of Engineering, Seoul National University, Seoul 08826, South Korea
[c] Research Group Neurocognition of Music and Language, Max Planck Institute for Empirical Aesthetics, Grüneburgweg 14, Frankfurt am Main 60322, Germany
[d] Department of Otorhinolaryngology-Head and Neck Surgery, Seoul National University Hospital, Seoul 03080, South Korea
[e] Department of AI Robotics, KIST School, Korea University of Science and Technology, Seoul 02792, South Korea
[f] Department of HY-KIST Bio-convergence, Hanyang University, Seoul 04763, South Korea

ARTICLE INFO

Article history:
Received 15 December 2022
Revised 6 April 2023
Accepted 15 April 2023
Available online 16 April 2023

Keywords:
Selective auditory attention
Topic familiarity
Listening volatility
Electroencephalography(EEG)
Neural decoding
Auditory attention detection(AAD)

ABSTRACT

Selective auditory attention has been shown to modulate the cortical representation of speech. This effect has been well documented in acoustically more challenging environments. However, the influence of top-down factors, in particular topic familiarity, on this process remains unclear, despite evidence that semantic information can promote speech-in-noise perception. Apart from individual features forming a static listening condition, dynamic and irregular changes of auditory scenes—volatile listening environments—have been less studied. To address these gaps, we explored the influence of topic familiarity and volatile listening on the selective auditory attention process during dichotic listening using electroencephalography. When stories with unfamiliar topics were presented, participants' comprehension was severely degraded. However, their cortical activity selectively tracked the speech of the target story well. This implies that topic familiarity hardly influences the speech tracking neural index, possibly when the bottom-up information is sufficient. However, when the listening environment was volatile and the listeners had to re-engage in new speech whenever auditory scenes altered, the neural correlates of the attended speech were degraded. In particular, the cortical response to the attended speech and the spatial asymmetry of the response to the left and right attention were significantly attenuated around 100–200 ms after the speech onset. These findings suggest that volatile listening environments could adversely affect the modulation effect of selective attention, possibly by hampering proper attention due to increased perceptual load.

1. Introduction

Selective auditory attention has been suggested as underlying the "cocktail party effect," which describes listening to a sound of interest in the presence of competing sounds (Cherry, 1953; Shamma et al., 2011). A growing body of research has observed that information on an attended speech is more strongly represented in the brain than information on an ignored speech (Ding and Simon, 2012a, 2012b; Mesgarani and Chang, 2012; Zion Golumbic et al., 2013). This observation indicates that selective attention could modulate the neural representation of speech, which reflects the cognitive process of selecting the target auditory stream in a cocktail party environment (Obleser and Kayser, 2019). Additionally, attempts to explore how selective attention is engaged in acoustically more challenging environments—for example, where the speech of interest is masked by noise or degraded by reverberation (Ding and Simon, 2013; Fuglsang et al., 2017)—allow us to understand how the modulation effect of selective attention depends on the amount of accessible acoustic information on the target sound.

In contrast to the findings from leveraging bottom-up acoustic features, top-down influences on the neural representation of selectively attended speech are not well documented, even though

Abbreviations: AAD, Auditory attention decoding; EEG, Electroencephalography/Electroencephalogram; GFP, Global field power; TRF, Temporal response function.

* Corresponding author.
E-mail address: yslim@kist.re.kr (Y. Lim).
# JH Park and SC Baek contributed equally to this paper.

https://doi.org/10.1016/j.heares.2023.108770

it has long been argued that high-level cognition is also considerably involved in selective auditory attention (Deutsch and Deutsch, 1963). In fact, various top-down factors have been shown to affect the analysis of a complex auditory scene. For instance, voice familiarity can promote the segregation of speech in a competing-speaker environment, as demonstrated by behavioral and functional neuroimaging studies (Holmes and Johnsrude, 2021; Johnsrude et al., 2013; Newman and Evers, 2007), and musical knowledge—proposed to be processed by the brain network that is partially shared with speech (Patel, 2011)—can be helpful for the perception and cortical tracking of speech when noise or competing sounds are present in an auditory scene (Du and Zatorre, 2017; Puschmann et al., 2019). For language-related factors, semantic information (or priming) has been shown to improve the intelligibility of speech in noise (Bhandari et al., 2021; Chan and Alain, 2021; Warzybok et al., 2021; Zekveld et al., 2011, 2013), and its neural substrates have also been relatively well documented (Obleser, 2014; Obleser and Kotz, 2010, 2011; Rysop et al., 2021). However, the advantage of semantic knowledge was not consistent in the neural representation of selectively attended speech in competing-speaker environments where a target speech was primed (Wang et al., 2019) or when experience in the language of the speech was varied (Reetzke et al., 2021; Zou et al., 2019).

This inconsistency in previous studies is probably because different levels of linguistic factors were confounded from lower segmental features (e.g., consonants, vowels, and phonemes) to higher structural knowledge (e.g., semantics and syntax), which precludes a clear understanding of these top-down influences. As such, it is essential to isolate each top-down factor and examine its influence. However, it is hard to dissociate mingled top-down features, particularly in naturalistic speech, because linguistic information at different hierarchical levels is recursively combined to form higher-level structures and, therefore, covary (e.g., Brodbeck and Simon, 2020). One feature that is worth considering is *topic familiarity*. A topic, which is a higher-level information structure in a text, arguably facilitates the processing of sentences by delimiting the context and, in turn, priming the upcoming words (Brothers et al., 2015; Foss, 1982; Jordan and Thomas, 2002). However, this facilitatory effect has seldom been tackled in speech processing. Therefore, we examine the top-down influence of topic familiarity on the selective attention process by adjusting topic familiarity in naturalistic speech.

The above discussions so far have assumed that selective attention operates in stationary auditory scenes. However, the listening environments that we encounter in everyday life are non-stationary, and dynamically and irregularly changing. It comprises sound sources that varies in locations (for a review, see van der Heijden et al., 2019), speakers (e.g., Shamma et al., 2011), and semantic information (e.g., Gregg and Samuel, 2009; Gregg and Snyder, 2012). Here, we define the listening environments where auditory scenes dynamically and irregularly change as *volatile* listening environments. The volatility of listening environments is important to extend our understanding of the selective auditory attention process because it can affect the engagement of attention to identify and follow an auditory object in a scene (Alain and Arnott, 2000; Best et al., 2008; Shinn-Cunningham, 2008). In contrast to static listening conditions where selective attention could be enhanced over time by focusing on a continuous auditory object (Best et al., 2008), in volatile conditions, selective attention could weaken by additional perceptual loads due to the requirements for re-analyzing an altered auditory scene and for identifying a new object (Lim et al., 2019, 2021; Shinn-Cunningham, 2008). However, note that volatile listening environments should be differentiated from regularly changing listening conditions because, in regularly changing environments, the listeners can have a perceptual benefit from predicting an auditory scene to be changed (Choi and Perrachione, 2019; Winkler et al., 2009; Zhao et al., 2019).

Despite the potential implications for the selective attention process and its underlying neural mechanisms, the effects of volatile listening conditions have not been well documented. Several studies have shown that talker discontinuity could modulate evoked potentials in response to the target speech (Getzmann et al., 2020; Lim et al., 2021; Mehraei et al., 2018). However, the time windows when this modulation effect was observed varied across these studies, and only Mehraei et al. (2018) found the modulation of N1 response that is known to reflect early auditory processing and to be reduced in speech-in-noise perception (Koerner and Zhang, 2015). In addition, all of the previous studies used controlled speech stimuli, for instance, syllable and digit trains or word-and-digit pairs, which made it hard to generalize their findings to naturalistic speech processing. In fact, Teoh and Lalor (2019) could not find evidence that supported the effect of talker discontinuity on the selective attention process using narrative naturalistic speech. Talker discontinuity could be one of the elements that contribute to the volatility of listening environments. However, to introduce talker discontinuity, the previous studies only manipulated the location of sound sources, and other features were kept constant (Getzmann et al., 2020; Mehraei et al., 2018; Teoh and Lalor, 2019), which does not seem to be enough to make auditory scenes dynamic and irregular. Therefore, we altered various features to form volatile listening conditions, such as source locations of the target sounds, speakers, and contents, and examined how this listening volatility could influence the neural representation of attended speech — as a first step to looking for the effect of listening volatility on the selective auditory attention process.

In this study, we investigated how topic familiarity or the volatility of a listening environment influences selective auditory attention and its underlying neural process. To this end, we employed a dichotic listening paradigm with naturalistic narrative speech sounds and observed the listener's neural activity using electroencephalography (EEG). We manipulated topic familiarity by presenting stories that the listener had never known or heard before, and we built a volatile listening environment by randomly varying the listener's spatial attention with different contents and speakers. Since each condition was driven by different (or perhaps even orthogonal) factors, we hypothesized that the engaged neural mechanisms for selective auditory attention were distinct, although both of them might have strongly involved top-down processing. We investigated the cortical representations of the attended speech using a neural decoding approach and temporal response function (TRF) analysis. We tested the effect of each manipulation by comparing the neural decoding result and the TRF with the ones in the control condition, where the listener attended to one of two stories with high topic familiarity and without changing the spatial attention, contents, and voices (i.e., with low volatility).

Furthermore, we used another approach to determine changes in attentional states in different listening conditions. Previous studies have observed asymmetric brain activation patterns for different spatial attention (Das et al., 2016; Ding and Simon, 2012a; Power et al., 2012), which implied the presence of a neural process specialized for attention in each left and right ear. Based on these findings, we compared the asymmetric patterns of the TRFs over the different listening conditions. This approach informed us of the listener's attentional state or of the extent to which the listener was engaged in the target speech during dichotic listening. For example, when the listener was fully engaged in the target speech, we observed clear directional asymmetry of the TRFs because there was no disruption while the listener was paying attention to the sounds either on the left or right ear and thus, the neural activation pattern specific to each spatial attention was well captured.

Otherwise, the directional asymmetry of the TRFs would be weakened due to inadequate attention. With this idea, we hypothesized that the latter case would occur in volatile conditions because the increased perceptual load (through frequent re-engagement with a new auditory object; Best et al., 2008; Lim et al., 2019; 2021; Shinn-Cunningham, 2008) could hamper proper attention to the target speech.

2. Materials and methods

2.1. Participants

Thirty undergraduate students (16 male) were recruited from Seoul National University to participate in this study. Ten participants were randomly assigned to one of three experimental groups. Each group experienced one of the three listening conditions. All the participants were native speakers of Korean with self-reported normal hearing ability and no history of neurological disorders. All the procedures in this study followed the ethical standards of the Declaration of Helsinki and were approved by the Institutional Review Board (IRB) of the Korea Institute of Science and Technology and Seoul National University Hospital (IRB codes: 2017–016 and 1706–137–861, respectively).

2.2. Experimental design and procedure

This study investigated the selective auditory attention process in the following three experimental groups in the different listening environments: a control group, a low topic familiarity group, and a volatile group (Table 1). The experiment was conducted in a competing-talker scenario, in which the participants had to attend to one of two dichotically presented speech sounds while ignoring the other (Fig. 1a).

For the control group, the listening environment was designed by following one of the most basic scenarios, in which selective attention was investigated (O'Sullivan et al., 2015; Power et al., 2012). For the speech stimuli, Korean translations of two stories by Jules Verne—*Twenty Thousand Leagues under the Sea* and *Journey to the Center of the Earth*—were recorded in different male voices (sampled at 44.1 kHz). During the experiment, participants in the control group attended to only one of the two stories. These stories were easy for undergraduate students to understand, and the participants were already familiar with their topics. Moreover, the storylines and each participant's spatial attention to the task were maintained, which made the listening condition less volatile. This group functioned as the control for comparisons with the two other experimental groups. Of the 10 participants in the control group, six attended to the story *Twenty Thousand Leagues under the Sea* on their left ear, and four attended to the other story, *Journey to the Center of the Earth*, on their right ear.

In the second group, we aimed to investigate the top-down influences on the selective attention process driven by low topic familiarity. To do so, we used two stories with unfamiliar and difficult-to-understand topics such that participants could not easily anticipate the upcoming contents. Parts of the text from challenging philosophical books, namely, a translation of *Critique of Judgement* by Immanuel Kant and *Essentials of Neo-Confucianism* by Yulgok Yi-I, were recorded in different male voices (sampled at 44.1 kHz). Of the 10 participants in this low topic familiarity group, five participants attended to *Critique of Judgment* on their left ear, and five participants attended to *Essentials of Neo-Confucianism* on their right ear. As in the control group, in the low topic familiarity group, the story paths of both texts were preserved, and the spatial attention of each participant was fixed, with the texts of each story presented on the same ear throughout the experiment to prevent volatility of the listening environment.

For the third group, the listening environment was manipulated to be volatile. Unlike in the other two groups, where the spatial attention of each participant was fixed, in this volatile group, the direction of attention varied stochastically in each trial. The speech stimuli consisted of distinct excerpts from a Korean-language listening comprehension test for high school students in Korea (recorded in different female voices sampled at 44.1 kHz), each of which had its own topic and speaker. Therefore, not only the script was renewed for each trial but also the voice in the speech segments. These dynamic and irregular changes in the lis-

Table 1
Information about three experimental groups.

Group	Age (SD)	Content difficulty	Spatial attention	Context	Voice
Control ($N = 10$)	23.8 (3.74)	Normal	Fixed	Maintained	Male (Maintained)
Low topic familiarity ($N = 10$)	23.9 (1.67)	Hard	Fixed	Maintained	Male (Maintained)
Volatile ($N = 10$)	23.8 (2.77)	Normal	Randomized	Altered	Female (Altered)

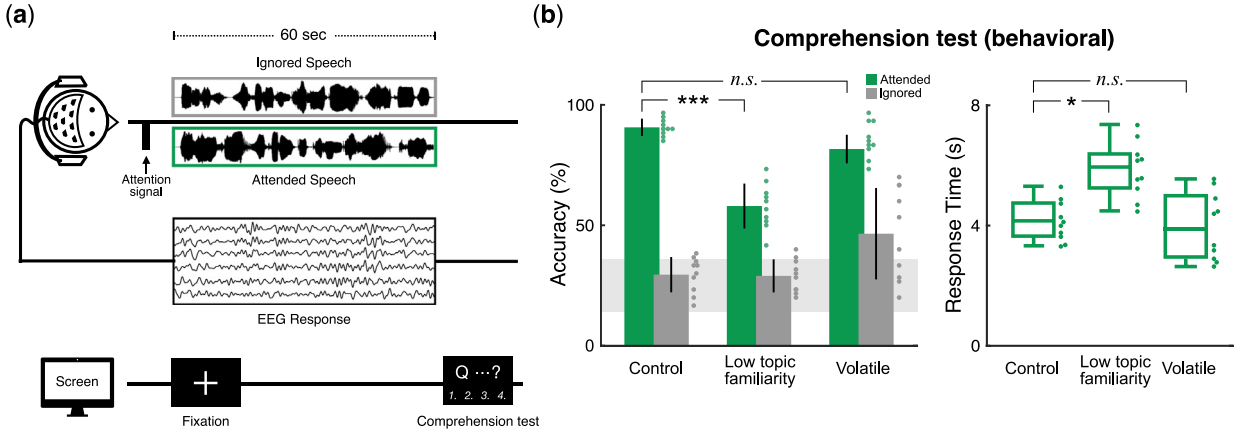

Fig. 1. Experimental paradigm and behavioral results. (a) A dichotic listening task was applied to all experimental groups. (b) Left: accuracy of the comprehension questions on the attended and ignored speech. Error bar denotes 1 standard deviation, and the shaded area stands for chance. Right: response time taken to answer the comprehension questions on the attended speech. Dots represent individual data.

tening environment could have made it difficult for the listeners to predict the sound source location, speaker, and topic of the following trial (Best et al., 2008; McCloy et al., 2017, 2018). However, these stimuli were content-wise not as difficult to comprehend as those in the low topic familiarity group, and the participants in the volatile group were familiar with all of their topics, since they had heard the stimuli in their school days. Thus, the listening environment of the volatile group could be said to have had both high topic familiarity and volatility. In this group, participants attended to a speech sound either on their left or right ear, randomly determined for every trial. However, the number of the trials for each attentional direction was equal for each participant (i.e., 15 trials each).

The experiment was conducted in a soundproof chamber installed at the Department of Otorhinolaryngology at Seoul National University Hospital. The experiment consisted of 30 trials, and in each trial, two distinct one-minute-long speech segments were presented to different sides of the ears, free of noise and without any degradation of the acoustic information. Silent gaps within a speech segment were adjusted to be no longer than 0.5 s so that the onset of speech sounds on each ear would be as similar as possible. The intensity of all the speech segments was normalized to a root mean square level of 0.8. The speech stimuli were presented through a pair of insert earphones at 65 dBL (ER-2 insert earphones, Etymotic Research, IL, USA). The participants were instructed to attend to the speech either on their left or right ear for each trial, and the direction of attention was signaled by a short tone introduced prior to the onset of each speech on the side where the participants should direct their attention. During the speech presentation, the participants were asked to focus their gaze on a fixation mark on the screen and to refrain from bodily movement. After the speech presentation, the participants were asked four multiple-choice questions to test whether they had attended to the target speech (two questions on the attended speech and the two others, on the ignored speech). The entire procedure was administered using the PsychoPy open-source software (v3.1.2; Peirce et al., 2019). An illustration of the overall experimental paradigm is shown in Fig. 1a.

2.3. Data acquisition and preprocessing

A 64-electrode system was used to obtain EEG data with Neuroscan SynAmps RT (64-channel Quik-Cap, Compumedics, Victoria, Australia). Raw EEG data were recorded at a sampling rate of 1000 Hz and re-referenced with a common average reference (except for the vertical and horizontal electro-oculograms). The re-referenced data were bandpass-filtered at the cutoff frequency range of 1–50 Hz, and subsequently, the 2–8 Hz EEG signal band, which is known as a relevant frequency range for the processing of the speech envelope (Hickok and Poeppel, 2016; O'Sullivan et al., 2015; Pasley et al., 2012; Zion Golumbic et al., 2013), was extracted.

To analyze the neural representation of speech under different listening conditions, we calculated the envelopes of the speech stimuli at the frequency range that corresponded to that of the preprocessed EEG data. To do so, we initially applied Hilbert transform to the waveforms of the stimuli and low-pass-filtered the absolute values of the transformed signals at the cutoff frequency of 8 Hz.

To match the sample lengths of the EEG and speech envelope data for further analyses, these data were downsampled to 64 Hz and normalized via z-scoring. All the data analyses in this study, including the preprocessing, were conducted with MATLAB (v9.5.0 R2018b, The MathWorks, Inc., Natick, MA, USA), EEGLAB (v19.1, Delorme and Makeig, 2004), and the mTRF toolbox (v2.0, Crosse et al., 2016). The preprocessed experimental stimuli and EEG data that we used can be found at https://github.com/hbum/AAD_Complexity.

2.4. Decoder model

To investigate how different listening conditions affect the selective attention process and the neural representations of speech, we used a reconstruction-based decoding method. In this approach, a speech envelope is reconstructed based on the EEG signal acquired while listening to actual speech signals, and the similarity between the actual and reconstructed speech envelopes is measured (Alickovic et al., 2019; O'Sullivan et al., 2015; Wong et al., 2018). This similarity reflects how well the actual speech signals are represented at the cortical level (Ding and Simon, 2014; Vanthornhout et al., 2018) and can be modulated by the listener's attention level (Calderone et al., 2014). In the dichotic listening scenario, the similarity of a reconstructed speech envelope to the attended speech is usually compared with that to the ignored speech, and based on this comparison, the direction of the listener's attention can be estimated (O'Sullivan et al., 2015).

This decoding method has been widely used in previous auditory attention detection (AAD) studies because it can detect the listener's attention with reasonable accuracy, based on single-trial EEG data, and therefore, is computationally inexpensive (Crosse et al., 2016; O'Sullivan et al., 2015). Using this approach, we compared the detection accuracy levels of the low topic familiarity and the volatile groups to the control group to understand how low topic familiarity or the volatility of a listening environment affects the selective auditory attention process.

To build a decoder model $w(.)$, ridge regression was applied to determine if there was a linear relationship between the EEG data and the envelope of the attended speech (Crosse et al., 2016), using the following equations:

$$\hat{S}(t) = \sum_{k}\sum_{\tau} w(\tau,k)R(t-\tau,k) \tag{1}$$

$$w = \left(RR^T + \lambda I\right)^{-1} RS^T \tag{2}$$

where $R(.)$ refers to the neural response recorded in an EEG channel, k; $\hat{S}(.)$ denotes the reconstructed speech envelope at a given time t; τ models the time lag between the stimulus presentation and the neural response; and S is the envelope of the actual speech.

The attended speech was estimated by comparing the Pearson's correlation of the reconstructed speech envelope with the target speech and the ignored speech in a given trial. The decoder model was considered to correctly detect the direction of the listener's attention only when the correlation of the reconstructed envelope with the attended speech was higher than that with the ignored speech. The detection accuracy had often been measured with the leave-one-out cross-validation scheme (Alickovic et al., 2019; O'Sullivan et al., 2015), whereby a model is constructed based on the trials after leaving out one trial and is tested in this held-out trial for a total number of given trials. However, we validated the decoder model via bootstrapping to achieve stable detection accuracy despite the relatively small number of our trials. Specifically, we trained a decoder model with 20 sampled trials out of the remaining 29 after leaving out one trial for testing. We repeated this procedure 30 times over a total of 30 trials. Consequently, we evaluated the bootstrapped model 900 times for each participant.

For the model construction, referring to previous studies, we determined the time-lag parameter τ as a range, 0–250 ms (O'Sullivan et al., 2015), and set the L2 regularization parameter λ at 10^3 by choosing the value that showed the maximum detection accuracy after the iterative search (Hoerl and Kennard, 1970; Wong et al., 2018).

2.5. Temporal response function (TRF)

TRF analysis was performed to examine the underlying neural processes of selective auditory attention in different listening conditions. TRF—also called the *forward* or *encoder model*—linearly maps the features of external stimuli onto neural activity (Alickovic et al., 2019; Wong et al., 2018), and its coefficients could reflect how the neural activity in a certain region responds to given stimuli over time (Ding and Simon, 2012b).

Similar to the decoding method, Eq. (3) below showed that given the envelope(s) $S(.)$ of the actual speech at a given time t, $\hat{R}(t,k)$—the estimated EEG response at a channel, k—was computed by convolving $TRF(.)$ with $S(.)$, in the encoder modeling. As shown in Eq. (4), $TRF(.)$ was calculated using the least squares method with an L2 regularizer, λ (also set at 10^3), based on $S(.)$ and the actual neural response acquired while listening, $R(.)$.

$$\hat{R}(t,k) = \sum_{\tau} TRF(\tau,k)S(t-\tau) \tag{3}$$

$$TRF = \left(SS^T + \lambda I\right)^{-1} SR^T \tag{4}$$

The parameter τ in Eq. (3) also modeled the time lag between the speech onset and the EEG response. However, in contrast to the decoder model, it was set with a wider range of −125–500 ms to inspect the cortical activity in response to the speech stimuli for a longer period.

2.6. Comparison of TRFs in the different directions of attention

To explore how the underlying neural state in the selective auditory attention process varies depending on the listening condition, we compared the TRFs in the different directions of attention. In each experimental group, the TRFs of the left and right attended speech were measured for each trial, and these trial-level TRFs for each direction were compared in a data-driven manner by applying cluster-based non-parametric permutation testing. Cluster-based non-parametric permutation testing can be a useful method of identifying the directional asymmetry of TRFs in time and in the electrode space because it can correct the increase in Type I errors caused by multiple comparisons (Maris and Oostenveld, 2007). To do so, the TRF coefficients of the left and right attended speeches were compared at each point in time and within the electrode space via a t-test. Note that an independent two-sample t-test was applied in the control and low topic familiarity groups because the left and right TRF coefficients were obtained from different participants. However, in the volatile group, a paired sample t-test was performed because the TRF coefficients were obtained from the same participants. Then, clusters were built by aggregating the neighboring points with p-values lower than the threshold of 0.05. For this, we defined two electrodes as neighbors when the Euclidean distance between them was less than 0.4 (assuming that the head was a sphere and its radius equaled one). After extracting the clusters in time and in the electrode space, cluster-level t-statistics were assigned by summing up the point-level t-statistics within the same clusters.

Subsequently, the distribution of the cluster-level t-statistics was estimated under the null hypothesis, which assumed that the TRFs had no directional asymmetry for the attended speech. For each trial, a fake direction of attention was randomly assigned. Based on this random assignment, the left and right TRF coefficients were compared at each point in time and in the electrode space. In the same way as that described above, the clusters and the cluster-level t-statistics were computed. This procedure was repeated 10,000 times while keeping the maximum and minimum cluster-level t-values that were obtained from each repetition. A total of 20,000 cluster-level t-values approximated the distribution of the cluster-level t-statistics under the null hypothesis. Among the clusters that were computed under the real spatial attention, only those with a cluster-level t-statistic either above the upper or below the lower 2.5 percentile of the distribution were considered significant.

2.7. Responsiveness analysis

As mentioned, in contrast to the volatile group, in which the TRFs of the left and right attended speeches were from the same participants, in the control and low topic familiarity groups, the TRFs of the attended speech for each direction were computed from the neural response of different participants. Therefore, in these two experimental groups, even if the TRFs of the attended speech for each spatial attention showed different patterns, these differences could not be simply explained as the effect of the spatial attention because they might have been influenced by the individual variability of the neural responses to the speech stimuli. To rule out this possibility, we additionally conducted a responsiveness analysis to investigate whether the extent of the overall neural response to the speech stimuli remained the same regardless of the attention direction.

To perform the responsiveness analysis, trial-wise TRFs were computed based on both the attended and ignored speeches at all the electrode sites. Subsequently, the global field power (GFP) was measured by calculating the standard deviations of the TRF coefficients in the electrode dimension. The GFP quantifies the amount of neural activity in a given field (Skrandies, 1990). Thus, it allowed us to measure the degree of the overall neural activity in response to auditory stimuli. We therefore averaged the GFP over time to obtain a scalar value that would represent the overall neural responsiveness to the speech stimuli for each trial (Zou et al., 2019).

The trial-wise responsiveness values were separated into two sets according to the spatial attention, and for each experimental group, the responsiveness values from the left- and right-attention trials were compared via a t-test. Also, in this case, an independent two-sample t-test was performed in the control and low topic familiarity groups. However, in the volatile group, the responsiveness values from the left- and right-attention trials were compared using a paired sample t-test.

3. Results

3.1. Behavioral results

The behavioral results are illustrated in Fig. 1b. For each experimental group, the average accuracy of such results for the questions on the attended speech was above the statistical-chance level (upper bound of 95%, significance level = 35%), which indicates that in general, the participants successfully attended to the target speech (see the left panel in Fig. 1b). For the questions on the ignored speech, the average accuracy was above the statistical-chance level in the volatile group, but not in the control and low topic familiarity groups. Then, we compared the average accuracy of the results for the questions on the attended speech in the low topic familiarity and volatile groups to that in the control group (Table 2). The accuracy in the low topic familiarity group was significantly lower than that in the control group. However, no evidence was observed for the difference in the accuracy between the control and volatile group (Mann–Whitney U test, control vs. low topic familiarity: $U = 0$, $p < 0.001^{***}$; control vs. volatile: $U = 26.5$, $p > 0.05$, n.s.; Bonferroni-corrected).

During the experiment, we also measured the response time to each question on the target speech (see the right panel in Fig. 1b). The response time in the low topic familiarity group was significantly longer than in the control ($U = 7$, $p = 0.001^{**}$, Bonferroni-

Table 2
Auditory attention detection (AAD) accuracy and comprehension test results (accuracy and response time) for the attended and ignored speech.

Group	Comprehension test results				Auditory attention detection accuracy (SD)
	Attended speech		Ignored speech		
	Accuracy (SD)	R.T. (SD)	Accuracy (SD)	R.T. (SD)	
Control	90.67% (3.62)	4.17 s. (0.65)	29.50% (7.33)	2.71 s. (1.34)	96.92% (3.92)
Low topic familiarity	58.00% (9.32)	5.91 s. (0.97)	29.00% (6.86)	3.11 s. (1.58)	95.94% (4.03)
Volatile	85.33% (7.45)	3.99 s. (1.12)	46.50% (1.89)	3.13 s. (1.05)	84.76% (9.00)

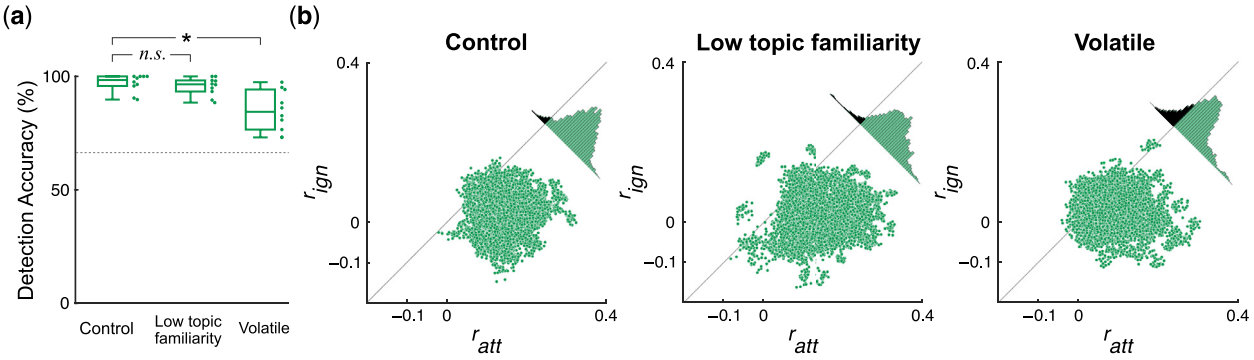

Fig. 2. Auditory attention detection (AAD) results. (a) AAD accuracy for each experimental group. Dots next to boxes indicate average AAD accuracy of each participant. A dashed line stands for chance. (b) Trial-wise AAD results in the different experimental groups. Correct trials, where r_{att} is bigger than r_{ign}, are positioned below the diagonal line. Histograms illustrate the distribution of the difference between r_{att} and r_{ign}. Regions marked in black represent the trials in which AAD failed.

corrected). However, the difference in the response time between the control and volatile groups was not significant ($U = 44$, $p > 0.05$).

3.2. Auditory attention detection results

As shown in Fig. 2a, the average AAD accuracy was above the statistical-chance level in all the experimental groups, and no participant had a below-chance detection accuracy (upper bound of 95%, significance level = 66.67%). However, a difference in AAD accuracy was observed among the experimental groups (Table 2). The average detection accuracy in the control group was significantly higher than that in the volatile group [$t(18) = 3.92$, $p = 0.001^{**}$; Bonferroni-corrected] and was comparable to that in the low topic familiarity group [$t(18) = 0.77$, $p > 0.05$].

A similar pattern was observed in the trial-wise detection results. The scatter plots in Fig. 2b show that the attended speech could be detected by a decoder model in the three experimental groups. In other words, in most of the trials, the correlation between the reconstructed envelope and the target envelope (r_{att}) was greater than the correlation between the reconstructed envelope and the ignored envelope (r_{ign}). However, the volatile group had more incorrect trials, where the decoder model failed to detect the target speech, than the control and low topic familiarity groups (see the black region in each histogram in Fig. 2b). Moreover, the difference between r_{att} and r_{ign} (Δr) was smaller in the volatile group (median $\Delta r = 0.0597$) than in the control (median $\Delta r = 0.1243$) and low topic familiarity (median $\Delta r = 0.1202$) groups.

3.3. TRF analysis

To investigate the differences in the underlying neural state in the selective auditory attention process, we compared the TRFs when the participants directed their attention to the left versus right speech streams. Research has shown that different underlying neural processes can support the modulation effect of left and right attention (Das et al., 2016; Ding and Simon, 2012a; Power et al., 2012). However, it remains unclear whether this also holds when the listener is less familiar with the presented speech streams or when the listening environment is volatile. As these listening conditions are expected to be perceptually more demanding than the one in the control group (Best et al., 2008; Zekveld et al., 2012), we hypothesized that the directional asymmetry of the TRFs could weaken or could disappear because the increased perceptual demands might adversely affect the attention modulation.

Based on cluster-based non-parametric permutation testing, the TRFs of the left and right attention were compared in time and in the electrode space in each experimental group (Fig. 3). Significant clusters were observed in the control and low topic familiarity groups, whereas in the volatile group, no significant cluster was identified (see the middle row in Fig. 3). These results imply that the underlying neural process when listening to speech presented to one side of the ear may differ from that on the other side, as in previous findings (Das et al., 2016; Ding and Simon, 2012a; Power et al., 2012), but as the listening environment becomes volatile, this asymmetry could weaken.

Although the numbers and regions of significant clusters differed between the control group (four clusters) and the low topic familiarity group (two clusters), both groups showed directional asymmetry at around 100–200 ms. In this time range, the dominance of the right TRF was observed in the left frontotemporal region in the control group ($p = 0.014^{*}$), whereas the dominance of the left TRF was seen in the right frontotemporal region in the low topic familiarity group ($p < 0.01^{**}$).

Then, we focused on the TRFs for each attention direction at the frontal electrodes, where prominent cortical activation has typically been observed while listening to speech (Crosse et al., 2015;

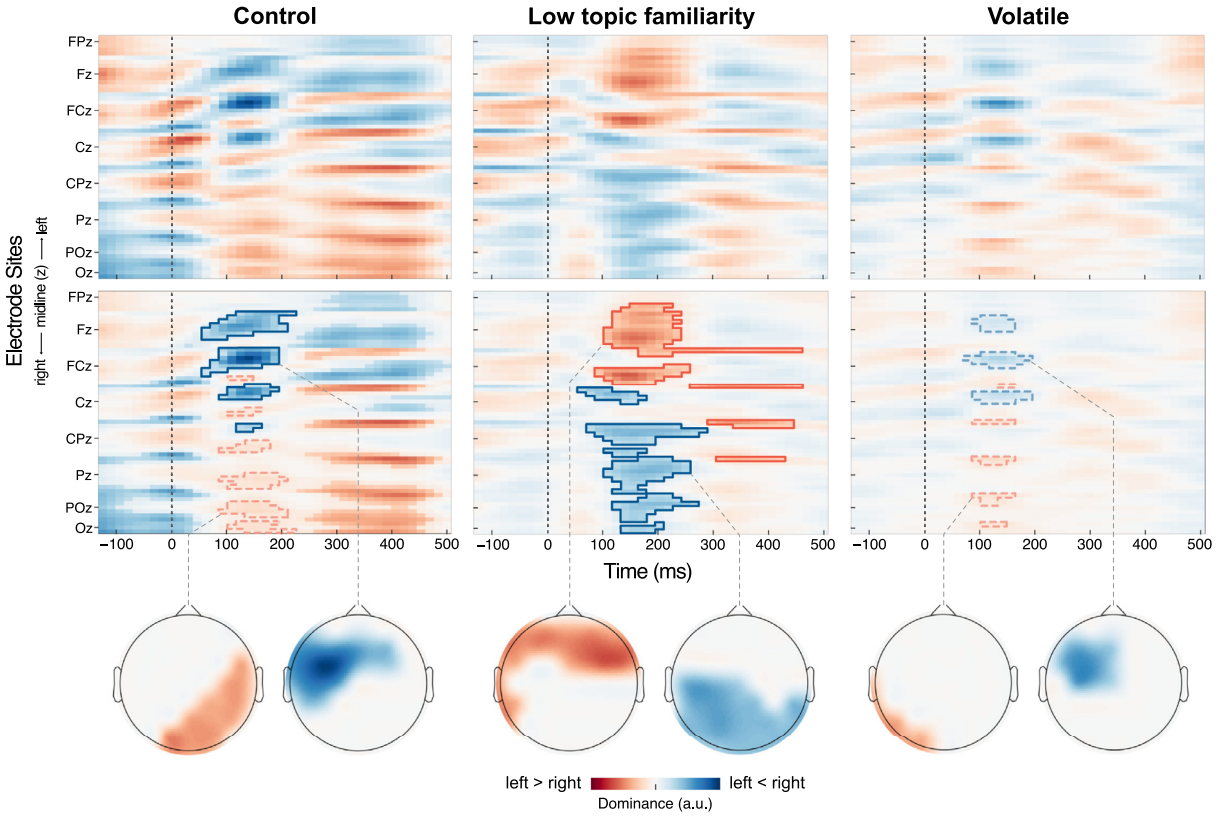

Fig. 3. Difference in temporal response functions (TRF) of the left and right attention for each experimental group. Top row: Raw difference in the TRF coefficients of each attention direction. Middle row: Significant clusters identified based on non-parametric permutation testing. Color intensity in the regions outside the significant clusters are attenuated. Of the significant clusters, those found in the time range of 100–200 ms are highlighted with bold lines. For comparisons, clusters found in the same time range, but statistically insignificant, are demarcated by dashed lines. Bottom row: Topographic plots of the TRF differences corresponding to the (in)significant clusters in 100–200 ms.

Das et al., 2016; Etard and Reichenbach, 2019; Fuglsang et al., 2017). Fig. 4a illustrates the time traces of the TRF at the frontal region in each experimental group (averages over 10 frontal and frontocentral electrodes: F5, F3, Fz, F4, F6, FC5, FC3, FCz, FC4, and FC6). Again, the left- and right-attention TRFs showed different temporal profiles in the control and low topic familiarity groups, whereas the TRFs of the volatile group looked similar regardless of the attentional direction. The control group showed the differences in the early responses (93–188 ms) and late responses (296–453 ms), whereas the low topic familiarity group exhibited a significant difference only at 125–219 ms (see the data with asterisks at the top and middle panels in Fig. 4a; FDR-corrected $q < 0.05$). However, no difference was observed in the volatile group (see the bottom panel in Fig. 4a).

The different directional asymmetry of the frontal TRFs across the three experimental groups indicate that the underlying neural states were different. However, it remains unclear whether these different directional asymmetries could be attributed to variant attention levels due to the different perceptual demands of the listening conditions, as we hypothesized. Motivated by the possibility that an early TRF response (at 100–200 ms) could be modulated by attention (Ding and Simon, 2012a, 2012b; Fuglsang et al., 2017; Horton et al., 2013; Vanthornhout et al., 2019), we also compared the frontal TRF coefficients at around 156 ms across the experimental groups, where both the control and low topic familiarity groups showed clear directional asymmetry (see the green-shaded areas in Fig. 4a). To do so, the TRF coefficients were compared after collapsing the attentional directions over 125–188 ms. As shown in the bar graphs in the inset boxes in Fig. 4a, the TRF responses in the control and low topic familiarity groups did not differ in this time range [$t(598) = 0.93$, $p > 0.05$], whereas the TRF response in the volatile group was significantly lower than that in the control group [$t(598) = 3.26$, $p = 0.012^{*}$].

Finally, the topography of the TRF from the left and right attention was explored when the frontal TRF showed peaks or troughs after the onset of the speech stimuli (Fig. 4b). While the topographic plots at 78 ms show similar distributions of neural activation for both attentional directions in each experimental group, the plots at 156 ms demonstrate that the left and right attention had different neural activation patterns on the scalp in the control and low topic familiarity groups. Similar to previous findings (Das et al., 2016; Power et al., 2012), the right attention led to activation in the left frontal region, whereas the left attention resulted in bilateral activation in the frontal region. The volatile group showed a similar shape of activation for each attentional direction, but the degree of this asymmetry was less conspicuous than in the two other groups. For the plots at 296 ms, the TRF activations exhibited different topographic distributions for the left and right attention in the control group, whereas the topographic distributions looked similar regardless of the direction of attention in the low topic familiarity and volatile groups.

3.4. Responsiveness analysis

Responsiveness analysis was additionally performed to rule out the possibility that the directional asymmetry of the TRFs in the control and low topic familiarity groups was due to individual differences in the neural responsiveness to the speech stimuli. Except in the volatile group, the TRFs of the attended speech for each direction were obtained from different participants. This could have

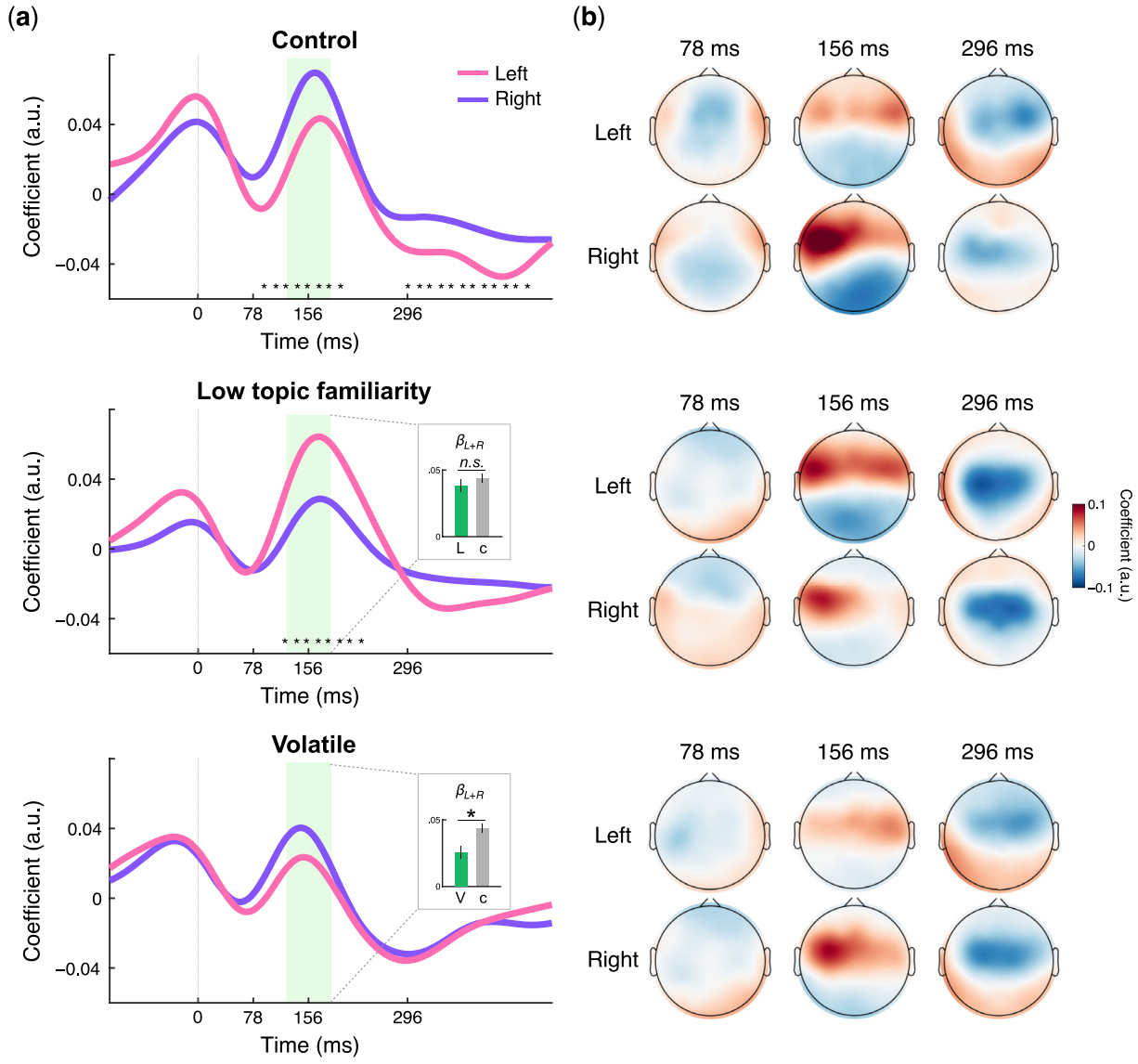

Fig. 4. Frontal TRFs and topography for the left and right attention. (a) TRFs for each attention direction at the frontal region (10 frontal and frontocentral electrodes averaged). Asterisks indicate when the TRF coefficients are significantly different over the directions of attention (FDR-corrected $q < 0.05$). Inset boxes denote the comparisons in the TRF coefficient of main peak (green shaded area) from each (corresponding) group with the control (c: control; L: low topic familiarity; V: volatile). The coefficients are averaged by collapsing the attention directions. Error bars denote S.E.M. (b) Topographic distributions of the TRF for each attention direction at peak or trough times of the frontal TRF.

contributed to the directional asymmetry of the TRFs in the control and low topic familiarity groups.

The responsiveness analysis showed that the overall neural responsiveness to the speech stimuli did not differ between the attentional directions nor among the three experimental groups [control: $t(298) = 0.63$, $p > 0.05$; low topic familiarity: $t(298) = -0.85$, $p > 0.05$; and volatile: $t(149) = -0.80$, $p > 0.05$]. This means that the directional asymmetry of the TRFs in the control and low topic familiarity groups was not due to individual differences.

4. Discussion

In this study, we explored how topic familiarity and the volatility of a listening environment affect selective auditory attention and its underlying neural process based on a dichotic listening task with naturalistic narrative speech sounds. In the low topic familiarity group, the participants listened to two stories about sophisticated philosophical topics that they hardly knew or had heard of. However, in the volatile group, not only the spatial attention was randomly changed in each trial but also the storylines and speakers, to create an environment where it would be more perceptually demanding for the participants to re-engage with a new auditory object, although they were generally familiar with the topics. The AAD results based on the EEG decoding approach showed that the detection accuracy significantly decreased in the volatile group compared to the control group (which had topic-wise familiar stories and a constant listening environment; Table 1), while the behavioral performance was more strongly degraded in the low topic familiarity group (Figs. 1b and 2 and Table 2). In addition, the TRFs for the left and right spatial attention showed clear asymmetry in both the control and low topic familiarity groups, but such asymmetry was significantly attenuated in the volatile group (Figs. 3 and 4), which implies that the participants in the volatile group paid less attention to the target speech than the participants in the two other groups.

4.1. Effect of topic familiarity and top-down influences

The AAD accuracy in the low topic familiarity group was comparable to the AAD accuracy in the control group. This result indicates that low familiarity with the topic hardly influenced the modulation effect of selective attention on the processing of the target speech. This finding was also supported by the TRF analysis, which revealed that the TRF to the attended speech did not differ between the control and low topic familiarity groups particularly at around 156 ms after the speech onset (Fig. 4). Given that the TRF at 100–200 ms could reflect auditory processing and could be modulated by attention (Ding and Simon, 2012a, 2012b; Fuglsang et al., 2017; Horton et al., 2013; Vanthornhout et al., 2019; Zion Golumbic et al., 2013), it implies that selective auditory attention was similarly engaged in these two listening conditions.

Moreover, the analogous asymmetric pattern of the TRFs from the left and right attention and of the topography at around 156 ms demonstrated similar engagement of selective attention in both groups, although the directional dominance of the TRFs was different. Notably, only the control group showed a TRF asymmetry at around 300 ms after the speech onset. The TRF response at this time is known to reflect higher-order processing, such as the semantic process (Broderick et al., 2019). This and the absence of TRF asymmetry at around 300 ms in the low topic familiarity group may reflect hampered anticipatory processing of the upcoming semantic information because the participants were probably unable to benefit from the unfamiliar and complex topics (Brothers et al., 2015; Foss, 1982; Jordan and Thomas, 2002).

In line with this idea, we found that the participants' comprehension of the attended speech was severely degraded in the low topic familiarity group (Fig. 1b). However, this finding is incompatible with the observation in previous studies that the degree of cortical envelope tracking or AAD accuracy was positively correlated with speech comprehension (Decruy et al., 2020; O'Sullivan et al., 2015; Peelle et al., 2013). This finding also conflicts with that of other studies that argued the gain of speech neural representation by top-down influences (Baltzell et al., 2017; Constantino and Simon, 2018; Etard and Reichenbach, 2019; Wang et al., 2019).

To account for these contradictory results, we focus on the fact that the amount of the bottom-up acoustic information varied across different studies. In this study, we used clean speech to localize the top-down influence of topic familiarity on the attentional modulation effect and to minimize the intervention of bottom-up factors. On the contrary, most of the studies that reported the gain of top-down influences used acoustically degraded speech by masking the target speech with noise or competing speech (Constantino and Simon, 2018; Etard and Reichenbach, 2019; Wang et al., 2019), removing a part of the speech (Constantino and Simon, 2018), or vocoding the speech (Baltzell et al., 2017). This suggests that when acoustic information is sparse, top-down factors may compensate for the lack of bottom-up information. However, the advantage of top-down factors may be diminished in cases where, as in the low topic familiarity group, the object of interest is already distinct in the auditory scene due to the abundance of bottom-up information. This interaction of top-down and bottom-up factors could also be supported by the study of Zou et al. (2019), which showed that the cortical activity tracked non-native speech similarly to (or even more strongly than) native speech especially at a relatively higher signal-to-noise ratio.

4.2. Effect of volatile listening

Despite the familiar topics of the presented stories in the volatile group, significantly lower AAD accuracy was exhibited therein than in the control group (Fig. 2). This suggests that the volatility of the listening environment could adversely affect the attentional state of the participants, which would result in less engagement in the target speech. In the volatile group, the listeners had to identify a new target auditory object in each trial, since the contexts, speakers, and spatial attention randomly varied. This could increase the listeners' perceptual load (Best et al., 2008; McCloy et al., 2017; Shinn-Cunningham and Best, 2008), which would decrease their overall attention level. This overall decrease in participants' attention level in the volatile group is reflected by the significantly low TRF response of around 156 ms in the volatile group (Fig. 4a, inset box). In addition, the evidence from the behavioral results in which only participants in the volatile group showed the above-chance-level accuracy for the questions on the ignored speech (Fig. 1b) demonstrated that they could not fully engage in the target speech as much as the participants in the control group.

On the other hand, there is an alternative explanation for the degradation of the detection accuracy in the volatile group. Along with the directional asymmetry of the TRFs observed in the control and low topic familiarity groups, previous studies have shown that natural speech processing could be supported by distinct neural processes depending on the attention direction (Das et al., 2016; Ding and Simon, 2012a; Power et al., 2012). Idiosyncratic neural features for the left or right attention could be advantageous for detecting the attended speech on the same side. Das et al. (2016) exhibited this directional effect by showing that a decoder model performed better when all the training and test data consisted of the same spatial attention than when they did not. In contrast to the control and low topic familiarity groups, the spatial attention varied in the volatile group, which might be responsible for the decrease in the AAD accuracy.

However, the finding from the comparison of the left- and right-attention TRFs indicates that the degradation of the AAD accuracy in the volatile group could not be simply attributed to the directional effect. If it were largely due to the heterogeneity of the attention directions across the trials, as Das et al. (2016) argued, clear asymmetry of the left and right TRFs would have been observed in the volatile group as well. However, the volatile group showed significant attenuation of the directional asymmetry of the TRFs (Fig. 3). This implies, rather, that the attentional state in the volatile group might have been different from that in the other groups, which was also reflected by the relatively low TRF response at around 156 ms (Fig. 4) and the behavioral results of the questions on the ignored speech. To properly pay attention to the target speech in the volatile condition, it would have been necessary for the listeners to adaptively recruit the neural process that corresponded to each left and right spatial attention, which could have required additional perceptual resources (Koelewijn et al., 2014, 2015; McCloy et al., 2017, 2018) and could have resulted in an overall decrease in the attention level. In line with this idea, Baek et al. (2021) showed that an online decoder model performed poorly in attention-switching trials in which the listeners had to change their attention direction in each trial unlike in attention-fixed trials, in which the attentional direction was maintained.

4.3. Limitations and further study

Our observations in the low topic familiarity group allow us to examine the top-down influences on selective auditory attention and their underlying neural process. However, these influences should not be generalized as the overall effect of top-down factors. As mentioned, top-down factors are not homogenous and have various aspects and levels. For example, the topic familiarity manipulated in this study can be connected to the abstract or semantic knowledge reached in relatively later stages of speech process-

ing (Friederici, 2002; Hickok and Poeppel, 2007). However, some other top-down factors, such as voice familiarity (Johnsrude et al., 2013; Newman and Evers, 2007) and phoneme sequence statistics (Brodbeck et al., 2018), are known to be involved in the early processing stages (Hickok and Poeppel, 2007; Koelsch, 2011; Sjerps and Chang, 2019).

These various aspects of top-down knowledge can be engaged differently in auditory processing. Lower-level top-down features have been shown to modulate the early (∼ 150 or 200 ms) neural response to auditory stimuli (representatively, mismatch negativity; for a review, see Näätänen et al., 2007), whereas it is still controversial whether abstract and semantic knowledge could affect auditory processing in this earlier stage (Broderick et al., 2019; Norris et al., 2000; Travis et al., 2013). In the context of selective attention, Wang et al. (2019) partly showed that prior knowledge of the contents of upcoming speech could facilitate the neural representation of selectively attended speech. However, this was not the case or was even the opposite case when native and non-native speech were compared (Reetzke et al., 2021; Zou et al., 2019). Considering that the non-native listeners were also devoid of structural knowledge, such as phonology or phonotactics and syntax, these inconsistent findings might have been driven by the lack of knowledge other than semantics. Therefore, future studies should delineate how different aspects of linguistic knowledge affect the selective auditory attention process. In addition, it would be interesting to see whether and how (transferable) knowledge from other domains, such as music, modulates the neural tracking of speech during selective listening (e.g., Puschmann et al., 2019).

Moreover, we constructed the volatile listening condition by manipulating various components that might demand more perceptual resources, such that the listeners could not pay full attention to an auditory object of interest. These components are spatial attention, speakers, and contexts. However, it was impossible for us to know which components actually contributed to the formation of the volatile listening environment. Thus, there is a need for further investigation of the systematic relationship between these components and the volatility of a listening environment. This investigation would be essential for modeling the attentional states of listeners as a function of the volatility.

Finally, the limitations of the experiment design that we used in this study must be mentioned. We chose a between-subject design to minimize the fatigue effect that could occur from multiple participant visits for the EEG recordings. However, the statistical power of the between-subject design could be lower than that of the within-subject design, as we included only 10 participants in each experimental group. Further experiments are necessary with larger sample sizes or new experiment designs.

5. Conclusion

In this study, we investigated the modulation effect of selective auditory attention in various listening conditions. The low topic familiarity condition driven by presenting narrative speech with unfamiliar topics hardly influenced the selective auditory attention process. This might have been because sufficient acoustic information in noise-free stimuli was prioritized over topic information in the brain to track the relatively slow energy changes in the attended speech. However, the modulation effect of selective attention was degraded in the volatile condition that was induced by frequent and stochastic changes in spatial attention, speakers, and contexts. This suggests that an additional dimension—the volatility of a listening environment—should be considered in the selective auditory attention process. Understanding the relationship between the volatility and the effect of attention modulation may be crucial for modeling the neural processes underlying the cocktail party effect in everyday life or for successfully applying AAD techniques to the real world.

Declaration of Competing Interest

The authors declare no competing financial interests.

CRediT authorship contribution statement

Jonghwa Jeonglok Park: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Seung-Cheol Baek:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Myung-Whan Suh:** Conceptualization, Investigation, Resources. **Jongsuk Choi:** Funding acquisition, Supervision. **Sung June Kim:** Supervision, Resources. **Yoonseob Lim:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Data availability

The preprocessed experimental stimuli and EEG data used in this study can be found in the following link: https://github.com/hbum/AAD_Complexity.

Funding

This work was supported in part by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea Government (MSIT) (2017-0-00432, Development of non-invasive integrated BCI SW platform to control home appliances and external devices by user's thought via AR/VR interface), in part by the KIST Institutional Program (2E31084) and in part by the National Research Council of Science & Technology (NST) grant by the Korea Government (MSIT) (No. CAP21052- 200).

References

Alain, C., Arnott, S.R., 2000. Selectively attending to auditory objects. Front. Biosci.-Landmark 5 (3), 202–212.
Alickovic, E., Lunner, T., Gustafsson, F., Ljung, L., 2019. A tutorial on auditory attention identification methods. Front. Neurosci. 13, 153.
Baek, S.-.C., Chung, J.H., Lim, Y., 2021. Implementation of an online auditory attention detection model with electroencephalography in a dichotomous listening experiment. Sensors 21 (2).
Bhandari, P., Demberg, V., Kray, J., 2021. Semantic predictability facilitates comprehension of degraded speech in a graded manner. Front. Psychol. 12, 714485.
Baltzell, L.S., Srinivasan, R., Richards, V.M., 2017. The effect of prior knowledge and intelligibility on the cortical entrainment response to speech. J. Neurophysiol. 118 (6), 3144–3151.
Best, V., Ozmeral, E.J., Kopco, N., Shinn-Cunningham, B.G., 2008. Object continuity enhances selective auditory attention. Proc. Natl. Acad. Sci. U.S.A. 105 (35), 13174–13178.
Brodbeck, C., Elliot Hong, L., Simon, J.Z., 2018. Rapid transformation from auditory to linguistic representations of continuous speech. Curr. Biol. 28 (24), 3976–3983 e5.
Brodbeck, C., Simon, J.Z., 2020. Continuous speech processing. Curr. Opin. Physiol. 18, 25–31.
Broderick, M.P., Anderson, A.J., Lalor, E.C., 2019. Semantic context enhances the early auditory encoding of natural speech. J. Neurosci. 39 (38), 7564–7575.
Brothers, T., Swaab, T.Y., Traxler, M.J., 2015. Effects of prediction and contextual support on lexical processing: prediction takes precedence. Cognition 136, 135–149.
Calderone, D.J., Lakatos, P., Butler, P.D., Xavier Castellanos, F., 2014. Entrainment of neural oscillations as a modifiable substrate of attention. Trends Cogn. Sci. (Regul. Ed.) 18 (6), 300–309.
Cherry, E.C., Colin Cherry, E., 1953. Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Am. 25 (5), 975–979.
Cervantes Constantino, F., Simon, J.Z., 2018. Restoration and efficiency of the neural processing of continuous speech are promoted by prior knowledge. Front. Syst. Neurosci. 12, 56.

Choi, J.Y., Perrachione, T.K., 2019. Time and information in perceptual adaptation to speech. Cognition 192, 103982.
Crosse, M.J., Butler, J.S., Lalor, E.C., 2015. Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. J. Neurosci. 35 (42), 14195–14204.
Crosse, M.J., Di Liberto, G.M., Bednar, A., Lalor, E.C., 2016. The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. Front. Hum. Neurosci. 10, 604.
Chan, T.V., Alain, C., 2021. Brain indices associated with semantic cues prior to and after a word in noise. Brain Res. 1751, 147206.
Das, N., Biesmans, W., Bertrand, A., Francart, T., 2016. The effect of head-related filtering and ear-specific decoding bias on auditory attention detection. J. Neural. Eng. 13 (5), 056014.
Decruy, L., Lesenfants, D., Vanthornhout, J., Francart, T., 2020. Top-down modulation of neural envelope tracking: the interplay with behavioral, self-report and neural measures of listening effort. Eur. J. Neurosci. 52 (5), 3375–3393.
Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods 134 (1), 9–21.
Deutsch, J.A., Deutsch, D., 1963. Attention: some theoretical considerations. Psychol. Rev. 70 (1), 80.
Ding, N., Simon, J.Z., 2012a. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J. Neurophysiol. 107 (1), 78–89.
Ding, N., Simon, J.Z., 2012b. Emergence of neural encoding of auditory objects while listening to competing speakers. Proc. Natl. Acad. Sci. U.S.A. 109 (29), 11854–11859.
Ding, N., Simon, J.Z., 2013. Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. J. Neurosci. 33 (13), 5728–5735.
Ding, N., Simon, J.Z., 2014. Cortical entrainment to continuous speech: functional roles and interpretations. Front. Hum. Neurosci. 8, 311.
Du, Y., Zatorre, R.J., 2017. Musical training sharpens and bonds ears and tongue to hear speech better. Proc. the Nat. Acad. Sci. 114 (51), 13579–13584.
Etard, O., Reichenbach, T., 2019. Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. J. Neurosci. 39 (29), 5750–5759.
Foss, D.J., 1982. A discourse on semantic priming. Cogn. Psychol. 14 (4), 590–607.
Friederici, A.D., 2002. Towards a neural basis of auditory sentence processing. Trends Cogn. Sci. (Regul. Ed.) 6 (2), 78–84.
Fuglsang, S.A., Dau, T., Hjortkjær, J., 2017. Noise-robust cortical tracking of attended speech in real-world acoustic scenes. Neuroimage 156, 435–444.
Getzmann, S., Klatt, L.I., Schneider, D., Begau, A., Wascher, E., 2020. EEG correlates of spatial shifts of attention in a dynamic multi-talker speech perception scenario in younger and older adults. Hear. Res. 398, 108077.
Gregg, M.K., Samuel, A.G., 2009. The importance of semantics in auditory representations. Atten. Percept. Psychophys. 71, 607–619.
Gregg, M.K., Snyder, J.S., 2012. Enhanced sensory processing accompanies successful detection of change for real-world sounds. Neuroimage 62 (1), 113–119.
Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nat. Rev. Neurosci. 8 (5), 393–402.
Hickok, G., Poeppel, D., 2016. Neural basis of speech perception. Neurobiol. Lang. 299–310.
Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67.
Holmes, E., Johnsrude, I.S., 2021. Speech-evoked brain activity is more robust to competing speech when it is spoken by someone familiar. Neuroimage 237, 118107.
Horton, C., D'Zmura, M., Srinivasan, R, 2013. Suppression of competing speech through entrainment of cortical oscillations. J. Neurophysiol. 109 (12), 3082–3093.
Johnsrude, I.S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H.P., Carlyon, R.P., 2013. Swinging at a cocktail party: voice familiarity aids speech perception in the presence of a competing voice. Psychol Sci 24 (10), 1995–2004.
Jordan, T.R., Thomas, S.M., 2002. In search of perceptual influences of sentence context on word recognition. J. Exp. Psychol. 28 (1), 34.
Koelewijn, T., De Kluiver, H., Shinn-Cunningham, B.G., Zekveld, A.A., Kramer, S.E., 2015. The pupil response reveals increased listening effort when it is difficult to focus attention. Hear. Res. 323, 81–90.
Koelewijn, T., Shinn-Cunningham, B.G., Zekveld, A.A., Kramer, S.E., 2014. The pupil response is sensitive to divided attention during speech processing. Hear. Res. 312, 114–120.
Koelsch, S., 2011. Toward a neural basis of music perception: a review and updated model. Front. Psychol. 2.
Koerner, T.K., Zhang, Y., 2015. Effects of background noise on inter-trial phase coherence and auditory N1–P2 responses to speech stimuli. Hear. Res. 328, 113–119.
Lim, S.J., Carter, Y.D., Njoroge, J.M., Shinn-Cunningham, B.G., Perrachione, T.K., 2021. Talker discontinuity disrupts attention to speech: evidence from EEG and pupillometry. Brain Lang. 221, 104996.
Lim, S.J., Shinn-Cunningham, B.G., Perrachione, T.K., 2019. Effects of talker continuity and speech rate on auditory working memory. Atten. Percept. Psychophys. 81, 1167–1177.
Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG–data. J. Neurosci. Methods 164 (1), 177–190.
McCloy, D.R., Larson, E., Lee, A.K.C., 2018. Auditory attention switching with listening difficulty: behavioral and pupillometric measures. J. Acoust. Soc. Am. 144 (5), 2764.
McCloy, D.R., Lau, B.K., Larson, E., Pratt, K.A.I., Lee, A.K.C., 2017. Pupillometry shows the effort of auditory attention switching. J. Acoust. Soc. Am. 141 (4), 2440.
Mehraei, G., Shinn-Cunningham, B., Dau, T., 2018. Influence of talker discontinuity on cortical dynamics of auditory spatial attention. Neuroimage 179, 548–556.
Mesgarani, N., Chang, E.F., 2012. Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485 (7397), 233–236.
Näätänen, R., Paavilainen, P., Rinne, T., Alho, K., 2007. The mismatch negativity (MMN) in basic research of central auditory processing: a review. Clin. Neurophysiol. 118 (12), 2544–2590.
Newman, R.S., Evers, S., 2007. The effect of talker familiarity on stream segregation. J. Phon. 35 (1), 85–103.
Norris, D., McQueen, J.M., Cutler, A., 2000. Merging information in speech recognition: feedback is never necessary. Behav. Brain. Sci. 23 (3), 299–325 discussion 325–370.
Obleser, J., 2014. Putting the listening brain in context. Lang. Linguist. Compass 8 (12), 646–658.
Obleser, J., Kayser, C., 2019. Neural entrainment and attentional selection in the listening brain. Trends Cogn. Sci. (Regul. Ed.) 23 (11), 913–926.
Obleser, J., Kotz, S.A., 2010. Expectancy constraints in degraded speech modulate the language comprehension network. Cereb. Cortex 20 (3), 633–640.
Obleser, J., Kotz, S.A., 2011. Multiple brain signatures of integration in the comprehension of degraded speech. Neuroimage 55 (2), 713–723.
O'Sullivan, J.A., Power, A.J., Mesgarani, N., Rajaram, S., Foxe, J.J., Shinn-Cunningham, B.G., Slaney, M., Shamma, S.A., Lalor, E.C., 2015. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cereb. Cortex 25 (7), 1697–1706.
Pasley, B.N., David, S.V., Mesgarani, N., Flinker, A., Shamma, S.A., Crone, N.E., Knight, R.T., Chang, E.F., 2012. Reconstructing speech from human auditory cortex. PLoS Biol. 10 (1), e1001251.
Patel, A.D., 2011. Why would musical training benefit the neural encoding of speech? The OPERA hypothesis. Front. Psychol. 2, 142.
Peelle, J.E., Gross, J., Davis, M.H., 2013. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb. Cortex 23 (6), 1378–1387.
Peirce, J., Gray, J.R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J.K., 2019. PsychoPy2: experiments in behavior made easy. Behav. Res. Methods 51 (1), 195–203.
Power, A.J., Foxe, J.J., Forde, E.-J., Reilly, R.B., Lalor, E.C., 2012. At what time is the cocktail party? A late locus of selective attention to natural speech. Eur. J. Neurosci. 35 (9), 1497–1503.
Puschmann, S., Baillet, S., Zatorre, R.J., 2019. Musicians at the cocktail party: neural substrates of musical training during selective listening in multispeaker situations. Cereb. Cortex 29 (8), 3253–3265.
Reetzke, R., Gnanateja, G.N., Chandrasekaran, B., 2021. Neural tracking of the speech envelope is differentially modulated by attention and language experience. Brain Lang. 213, 104891.
Rysop, A.U., Schmitt, L.M., Obleser, J., Hartwigsen, G., 2021. Neural modelling of the semantic predictability gain under challenging listening conditions. Hum. Brain Mapp. 42 (1), 110–127.
Shamma, S.A., Elhilali, M., Micheyl, C., 2011. Temporal coherence and attention in auditory scene analysis. Trends Neurosci. 34 (3), 114–123.
Shinn-Cunningham, B.G., Best, V., 2008. Selective attention in normal and impaired hearing. Trends Amplif. 12 (4), 283–299.
Sjerps, M.J., Chang, E.F., 2019. The cortical processing of speech sounds in the temporal lobe. In: Human Language: From Genes and Brain to Behavior. MIT Press, pp. 361–379.
Skrandies, W., 1990. Global field power and topographic similarity. Brain Topogr. 3 (1), 137–141.
Travis, K.E., Leonard, M.K., Chan, A.M., Torres, C., Sizemore, M.L., Qu, Z., Eskandar, E., Dale, A.M., Elman, J.L., Cash, S.S., Halgren, E., 2013. Independence of early speech processing from word meaning. Cereb. Cortex 23 (10), 2370–2379.
Teoh, E.S., Lalor, E.C., 2019. EEG decoding of the target speaker in a cocktail party scenario: considerations regarding dynamic switching of talker location. J. Neural. Eng. 16 (3), 036017.
Vanthornhout, J., Decruy, L., Francart, T., 2019. Effect of task and attention on neural tracking of speech. Front. Neurosci. 13, 977.
Vanthornhout, J., Decruy, L., Wouters, J., Simon, J.Z., Francart, T., 2018. Speech intelligibility predicted from neural entrainment of the speech envelope. J. Assoc. Res. Otolaryngol. 19 (2), 181–191.
van der Heijden, K., Rauschecker, J.P., de Gelder, B., Formisano, E., 2019. Cortical mechanisms of spatial hearing. Nat. Rev. Neurosci..
Wang, Y., Zhang, J., Zou, J., Luo, H., Ding, N., 2019. Prior knowledge guides speech segregation in human auditory cortex. Cereb. Cortex 29 (4), 1561–1571.
Warzybok, A., Rennies, J., Kollmeier, B., 2021. Contribution of low-level acoustic and higher-level lexical-semantic cues to speech recognition in noise and reverberation. Front. Built Environ. 7, 689388.
Winkler, I., Denham, S.L., Nelken, I., 2009. Modeling the auditory scene: predictive regularity representations and perceptual objects. Trends Cogn. Sci. (Regul. Ed.) 13 (12), 532–540.
Wong, D.D.E., Fuglsang, S.A., Hjortkjær, J., Ceolini, E., Slaney, M., De Cheveigné, A., 2018. A comparison of regularization methods in forward and backward models for auditory attention decoding. Front. Neurosci. 12, 531.

Zekveld, A.A., Rudner, M., Johnsrude, I.S., Festen, J.M., Van Beek, J.H.M., Rönnberg, J., 2011. The influence of semantically related and unrelated text cues on the intelligibility of sentences in noise. Ear. Hear. 32 (6), e16–e25.
Zekveld, A.A., Rudner, M., Johnsrude, I.S., Heslenfeld, D.J., Rönnberg, J., 2012. Behavioral and fMRI evidence that cognitive ability modulates the effect of semantic context on speech intelligibility. Brain Lang. 122 (2), 103–113.
Zekveld, A.A., Rudner, M., Johnsrude, I.S., Rönnberg, J., 2013. The effects of working memory capacity and semantic cues on the intelligibility of speech in noise. J. Acoust. Soc. Am. 134 (3), 2225–2234.
Zhao, S., Chait, M., Dick, F., Dayan, P., Furukawa, S., Liao, H.I., 2019. Pupil-linked phasic arousal evoked by violation but not emergence of regularity within rapid sound sequences. Nat. Commun. 10 (1), 4030.
Zion Golumbic, E.M., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R., Mehta, A.D., Simon, J.Z., Poeppel, D., Schroeder, C.E, 2013. Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party. Neuron 77 (5), 980–991.
Zou, J., Feng, J., Xu, T., Jin, P., Luo, C., Zhang, J., Pan, X., Chen, F., Zheng, J., Ding, N., 2019. Auditory and language contributions to neural encoding of speech features in noisy environments. Neuroimage 192, 66–75.