Original Paper Vol.27 No.4, 2025

Enhancing of Vocabulary Learning by the Representation of the Meaning of Words with Binaural Audio

Kosuke Shimizu,[*1] Shogo Fukushima,[*2] Hirokazu Doi[*3] and Takeshi Naemura[*4]

Abstract – This study investigates the effectiveness of binaural audio, representing meanings of words, in English vocabulary learning. The Bilingual Dual Coding Theory is utilized as a theoretical framework, with emphasis placed on the interplay between verbal and sensory systems for optimal learning. An expressive sound archive of 40 English words was developed using binaural recording technology, with the intention of providing auditory representations of word meanings. A within-participant experiment was conducted with 20 high school students, who memorized two sets of words under both binaural and monaural audio conditions. Immediate and one-week post-tests using multiple-choice questions revealed significantly higher retention rates when students learned with binaural audio. A subsequent two-way repeated measures ANOVA indicated a main effect for audio condition and time progression, but no significant interaction, suggesting that binaural audio consistently enhances vocabulary retention. Observations of participants' gestures and vocal repetitions supported the idea that interactive, multidimensional sensory input can reinforce memory formation. A follow-up EEG study exploring neural correlates found a non-significant trend linking theta wave ERSP with correct responses, aligning with previous research on theta oscillations and memory. The findings under discussion highlight the potential of binaural audio as a powerful tool in computer-assisted language learning, paving the way for further exploration of its long-term benefits and the mechanisms of neural activity involved in vocabulary acquisition.

Keywords : Binaural Audio, Immersive Learning, Vocabulary Learning

1. Introduction

It is widely acknowledged that vocabulary knowledge constitutes a pivotal element in the acquisition and instruction of a second language (L2). Vocabulary number is crutial for clear understanding in reading passage [1]~[3]. Laufer [4], [5] have introduced the concept of "Lexical Threshold", which refers to the minimum vocabulary size or knowledge required for a person to understand a text or spoken language effectively. This concept discussed it is challenging to comprehend the content in an effective manner if one does not possess a fundamental understanding of at least 95 percent of the lexical items in a text [6], [7], with this figure reaching 98 percentage in numerous subsequent studies [8]. In addition, the correlation between vocabulary size and L2 proficiency is not exclusive to the quantity of words learners possess; the depth of their knowledge of these words is also directly proportional [9].

Following the needs to efficiently support vocablary aquisition, teaching vocabulary have been demonstrated in several approaches. Lexical Approach [10], [11] posits that language is not merely a collection of grammatical rules but rather consists of "grammaticalized lexis." In this view, instruction should focus on "lexis"—including words, collocations, and fixed expressions—to foster more natural communicative abilities. Conversely, Krashen's Input Hypothesis [12], [13] underscores the significance of 'comprehensible input', proposing that learners require a marginally more sophisticated level of language exposure (i+1) to facilitate progress. However, it is crucial to note that inadequate vocabulary impedes the comprehension of input, potentially impeding learning progress. The Involvement Load Hypothesis [14] emphasises the pivotal role of involvement load, defined as the degree of mental engagement or "need" to process new words, in promoting long-term retention evaluating that vocabulary learning is most effective when learners direct substantial attention to words, actively explore their meanings, and apply them in context, leading to deeper processing and enhanced acquisition. Positioned alongside these frameworks,

*1: University of Tsukuba
*2: Kyushu University
*3: Nagaoka University of Technology
*4: University of Tokyo

Bilingual Dual Coding Theory [15], [16] offers a powerful lens for understanding how vocabulary is learned, stored, and retrieved in bilingual contexts. It posits that learners encode information through two interconnected representational systems —verbal-linguistic and nonverbal-imagistic—and can leverage both when acquiring new L2 vocabulary. In bilingual learners, these dual codes may operate across languages, creating multiple mental routes (e.g., linking an L2 word to its L1 equivalent, paired with a visual image) that strengthen memory traces. English Picture Dictionary is a dictionary that stores not only English words but also images representing the meaning of English words.

Several items to support vocabulary learning material such as multimedia learning contents are actively discussed in the field of technology-enhanced learning, and some application are implemented to educational field [17], [18]. Drawing from Bilingual dual coding theory, multimedia learning contents are actively applied. For instance, Zhu et al. developed Vivo, a system that employs video-based explanations of word meanings as an alternative to traditional paper dictionaries [19]. In contrast to visual stimuli—which can express virtually anything simply by displaying a photograph—it is quite difficult to represent the meaning of English words through auditory stimuli because it is challenging to conceive methods beyond, for example, using sound effects. Nonetheless, there are still possibilities in the auditory domain. Moetan employs the voices of game characters to aid in vocabulary learning [20], and there are applications that engage professional voice actors to produce stock voices for language learning [21], [22]. In addition, Fukushima [23] have developed "EmoTan," a system that uses emotionally expressive sounds and realistic sound effects within a broader context of storytelling. While Fukushima's [23] work utilized binaural recording, its focus was on learning through a 10-20 second audio story combining narration and pronunciation, and its primary goal was not to isolate the specific cognitive effect of spatial audio itself. Our research, in contrast, aims to specifically isolate and independently examine how representing a word's meaning spatially through binaural audio affects vocabulary acquisition and retention. This focus on the spatial dimension as the key independent variable is the central contribution of our paper. While our preliminary study [24] provided initial evidence, the present study offers a more robust analysis by expanding on our previous work with a larger cohort of participants.

2. Literature Review

In this section, we examine three major strands of Technology-Enhanced Language Learning that inform our research focus: *(1) technology-supported approaches to vocabulary instruction, (2) immersive and embodied methods, and (3) audio-centered innovations, especially binaural solutions.* By reviewing these areas, we highlight how emerging tools and modalities contribute to vocabulary acquisition, while clarifying the gaps that our study aims to address.

2.1 Technology-Supported Approaches for Vocabulary Instruction

Early computer-assisted and mobile-assisted language learning systems sought to simplify or automate common vocabulary tasks such as look-up, flashcard repetition, and spaced practice. For instance, online dictionaries [25]~[28] and dedicated mobile applications [20], [29], [30] have long offered on-demand word references, often accompanied by definitions or example sentences. Yet these resources primarily deliver textual data, lacking the contextual richness found in more advanced multimedia tools. Recent reviews of mobile language learning [31], [32] report that a large majority of studies emphasize reading and listening comprehension, but few thoroughly address strategies for deep vocabulary acquisition. Similarly, e-readers and interactive applications have demonstrated potential in improving lexical gains [33], although the modal focus remains on text or static images. Zhu et al. [19] introduce a video-augmented dictionary to immerse learners in real-life contexts, while Hong et al. [34], Hu et al. [35], and Brown et al. [36] propose dynamic subtitling systems to connect words with meaningful visual and situational cues. These advancements highlight how layering multiple representations can support vocabulary learning and retention. Nonetheless, a strong reliance on visual aids persists, and the potential benefits of immersive audio remain comparatively underexplored.

Attempts to contextualize language learning with sensor-enriched or location-based environments have

also emerged [37], [38]. Such systems show that embedding vocabulary in spatial settings fosters better contextual recall, supporting the notion that vocabulary gains can increase when words are consistently linked to meaningful or authentic contexts [39]~[42]. While these studies vary in platform and approach, a unifying finding is that an enriched learning environment—whether physical or virtual—can deepen engagement and facilitate memory processes for second language vocabulary.

2. 2 Immersive and Embodied Methods in Language Learning

A parallel line of research has explored immersive media and embodied interaction. Virtual Reality (VR) and Augmented Reality (AR) systems offer learners interactive spaces in which verbal and situational cues co-occur [43], [44]. For example, Vázquez et al. [45] introduce kinesthetic language learning in VR, while Ratcliffe et al. [46] investigate how embodied controller interactions affect user engagement and retention. Beyond VR, XR-based language platforms increasingly experiment with contextual cues designed to simulate real-life communicative scenarios [47]~[49]. Embodiment and gesture have also gained traction as potent enhancers of vocabulary memorization. Empirical studies [50] illustrate that coupling newly encountered words with physical movements facilitates deeper encoding, likely due to the multi-modal activation of motor and linguistic neural networks. Outside VR, body-based or movement-based tasks in real classrooms [51], [52] validate similar benefits. The growing momentum behind immersive and embodied approaches signals a broader interest in transcending traditional text-and-audio instruction, yet most such studies concentrate on visual or kinesthetic immersion rather than exploiting advanced auditory experiences.

2. 3 Audio-Centered Innovations and Binaural Solutions

While multimedia language applications commonly provide standard speech playback or simple sound effects, in-depth investigations of advanced audio techniques remain relatively sparse. A small subset of work addresses the role of emotional or expressive narration [23], [53], showing that carefully designed auditory cues can amplify engagement and recall by eliciting affective responses [54], [55]. Binaural audio, in particular, introduces three-dimensional spatial effects, enabling learners to perceive sounds as if they originated from distinct directions in a virtual space [56]. In amusement settings, 3D soundscapes have been employed for immersive storytelling [57], but their systematic application to language education is at an early stage. Preliminary research findings on binaural narration implementations indicate that the provision of more realistic audio cues can facilitate learners' ability to anchor new words to perceived physical or narrative contexts. This approach potentially emulates the benefits of visual immersion without necessitating the use of extensive VR hardware [23]. Furthermore, research on "emotional arousal" in memory [58], [59] implies that realistic sound stimuli, especially those conveying mood or environment, can strengthen the encoding process in vocabulary learning. However, the extant literature on binaural audio has largely focused on entertainment, health, or ASMR-related experiences [56], resulting in a paucity of research on the precise manner in which spatial audio might influence language acquisition outcomes over time.

In conclusion, an examination of the three strands of research into Technology-Enhanced Language Learning — that is, technology solutions for vocabulary development, immersive/embodied learning, and advanced audio-based methods — collectively demonstrates avenues that, although promising, have received insufficient exploration. Despite the gains made in immersive Extended Reality and contextualized instruction, comparatively little research has been conducted on the potential of spatial or 3D audio to independently support vocabulary retention. As binaural audio becomes more accessible, examining its pedagogical viability can expand the repertoire of multimodal approaches that foster richer learning experiences in L2 vocabulary acquisition.

3. Method

3. 1 Preparation for Learning Material

The binaural audio stimuli used in this study were developed based on the methodology established by Fukushima [23], which aims to foster deep, episodic memory of word meanings. Professional voice actors were hired to perform short, immersive stories for each word. These recordings were conducted in a studio using a dummy head microphone to cre-

ate a three-dimensional auditory experience, allowing learners to perceive sounds as if they were originating from specific locations in space. For example, for the word "prank," the actor would approach the microphone from behind and whisper to simulate being surprised. The emotional arousal of each narration was validated during production using skin conductance response (SCR) measurements to ensure the stimuli were sufficiently engaging [23]. From these recordings, only the segments containing the target words were extracted and edited for use in the current experiment.

Following the word selection process described in Section 3, the final 40 words were divided into two lists of 20 (Word Set 1 and Word Set 2). The lists were carefully balanced to ensure equivalency in linguistic characteristics, as shown in Table 1. Both lists contained a similar distribution of parts of speech, and there were no significant differences in mean word length or syllable count. The average CEFR level for both lists was approximately B2+, indicating a comparable level of difficulty. The outcome variable for our analysis was defined as the raw number of correctly identified words (out of 20) in each test.

Table 1 Linguistic characteristics of the two word lists.

Characteristic	Word Set 1	Word Set 2
Number of Words	20	20
Parts of Speech	10 V, 9 N, 1 Adj	10 V, 9 N, 1 Adj
Mean Word Length (SD)	5.55 (0.94)	5.65 (1.04)
Mean Syllable Count (SD)	1.60 (0.68)	1.65 (0.67)
Average CEFR Level	Approx. B2+	Approx. B2+

Words	Audio expression
whirl	Circulate: pronounce while circling around the dummy head microphone.
unleash	From bottom to top: pronounce while moving from the bottom of the dummy head microphone to the top.
imminent	Approach: pronounce while approaching the dummy head microphone.
fetch	Left to Right: pronounce while moving to the left or right of the dummy head.
prairie	Facing the wall: pronounce while facing the wall.
aerate	Whispering: pronounce as if whispering near the dummy head microphone

Fig. 1 Examples of words and their expression

The subsequent step involved the selection of words based on the Common European Framework of Reference for Languages (CEFR), a widely accepted international standard for evaluating language proficiency (Common European Framework of Reference for Languages, 2001; Council of Europe, 2020). The selection of vocabulary was focused on words that were not commonly known to students in middle and high school. Prior research suggests that individuals more easily remember words when the pronunciation of the phonemes within the word aligns with commonly observed orthographic-phonemic correspondences (Gillon, 2004). This understanding stems from the role of speech-sound awareness in reducing reading and spelling difficulties. In accordance with this understanding, the present study opted for words comprising fewer letters to facilitate easier recall. The words selected for the study were chosen to minimise the potential for inferring meaning from prefixes and suffixes, and to ensure that the words were as concise as possible. The words chosen for this investigation were "blast, dizzy, ascend, flock, bawl, eclipse, crave, dismay, bemoan, mirth, whirl, loony, maraud, bovine, peeve, barrow, suckle, yarn, prank, shiver, basin, sniff, buzz, pummel, gust, cajole, sneeze, swig, grotto, implore, honk, snoop, navel, aerate, tumble, rumple, schism, tease, hornet." It's noteworthy that some of these words are not found within the CEFR vocabulary list, indicating they are not often used.

The English vocabulary audio dictionary "Flashcards Deluxe", installed on an iPad and prepared with audio and word data, was utilized. The iPad's "Monaural audio conversion" function was employed to convert binaural audio into monaural audio. Headphones without any frequency correction (SONY MDR-CD900ST) were used to ensure faithful audio reproduction. While the exact decibel level was not recorded, the volume was adjusted to a comfortable listening level for each participant and kept consistent.

3.2 Participant and Experiment Design

Participants were recruited on a volunteer basis via an online Form announcement posted on a bulletin board at Touohgakkan High School in Yamagata Prefecture. The study protocol was conducted in accordance with the ethical guidelines approved by the institutional review board that oversaw prior related work [23]. Written informed consent was obtained from each student participant prior to the ex-

periment. Inclusion criteria for participation were: (1) being a native Japanese speaker, (2) having self-reported normal hearing, and (3) having prior experience taking the EIKEN Test(Japan's widely recognized English language assesment) in Practical English Proficiency. Exclusion criteria were: (1) having lived abroad (i.e., returnees), (2) holding an EIKEN Grade Pre-1(equivalent to CEFR B2 level) or higher, or (3) recognizing the meaning of any of the target words during a pre-screening test. Our sample size was determined prior to the study based on an a priori power analysis. We referred to a previous related study [23], which also validated the use of binaural audio and reported an effect size of d = 0.86 for a similar comparison. Using G*Power, we calculated that for an effect size of d = 0.86 in a two-tailed paired t-test with $\alpha = .05$ and 80% power, a sample of 15 participants would be required. Therefore, we proceeded with N=20, a number consistent with similar exploratory studies in this field and a multiple of four, which suited our four-group counterbalancing design. We adhered to the principle of not adding participants after the experiment began. An initial pool of 25 students participated. Five were excluded from the final analysis: three for self-reporting that they had studied the words between sessions, and two who reported in a post-experiment questionnaire that they had consciously changed their memorization strategy between the two conditions. The final analysis was conducted on the remaining 20 participants (15 male, 5 female; $M_{age} = 16.3$ years). As discussed in Section 5.2, the preliminary EEG data was collected for exploratory purposes from a separate group.

- G1: Memorized Word Set 1 with binaural audio, then Word Set 2 with monaural audio
- G2: Memorized Word Set 1 with monaural audio, then Word Set 2 with binaural audio
- G3: Memorized Word Set 2 with binaural audio, then Word Set 1 with monaural audio
- G4: Memorized Word Set 2 with monaural audio, then Word Set 1 with binaural audio

The overall experimental procedure, including the memorization and testing phases, is illustrated in Figure 2.

Prior to the commencement of the experiment, it was ascertained that the participants were not acquainted with the English words that were to be presented. In order to avoid any potential manipulation of results by altered memorization strategies, observers conducted rigorous monitoring of the techniques employed, and subsequently confirmed the uniformity of the results obtained. In consideration of the issues highlighted by Fukushima [23] with regard to the review of words by participants outside the parameters of the study, explicit instructions were issued to the participants not to engage in any rehearsal of the words during the week, and strict compliance with this directive was subsequently ascertained in order to maintain the integrity of the data.

The word test was administered in a multiple-choice format, wherein participants were presented with an English word and were asked to select the corresponding Japanese meaning from a group of options. The order of the English words presented to the test participants differed from the order employed in the memorization application; this was done in the hope that the test subjects would not use the order of the words in their memory when answering the questions.

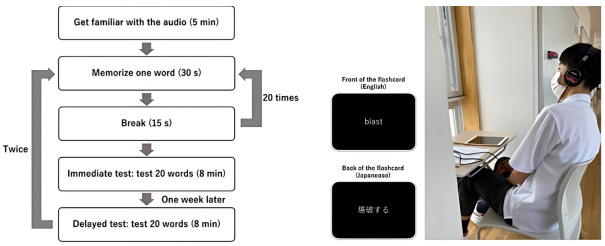


Fig. 2 Left: Experimental procedure, Right: Appearance of a participant

4. Result

A two-way repeated-measures Analysis of Variance (ANOVA) was conducted to compare the effect of audio condition (binaural vs. monaural) and time (immediate vs. one-week delayed test) on vocabulary retention. The results showed a significant main effect of audio condition ($F(1, 19) = 8.40, p < .01$) and time ($F(1, 19) = 102.82, p < .01$), but no significant interaction ($F(1, 19) = 3.14, p < .10$). These findings indicate that binaural audio yielded higher overall performance and that scores changed over time, yet these two factors operated independently.

Observationally, participants frequently used gestures and repeated pronunciations under the binau-

Table 2 Mean scores and standard deviations (SD) for each condition and time point (N=20).

Condition	Immediate Test M (SD)	Delayed Test M (SD)
Binaural Audio	17.5 (2.51)	12.4 (2.83)
Monaural Audio	15.6 (2.65)	10.5 (2.98)

ral condition. Words with clear auditory associations (e.g., “aerate,” “blast,” “whirl”) tended to yield higher correct response rates; by contrast, words with more obscure meanings (e.g., “suckle”) or indirect expressions (e.g., “maraud”) garnered lower correct response rates.

5. Discussion

This study compared the effects of binaural audio (hereafter, the “binaural condition”) and monaural audio (hereafter, the “monaural condition”) on English vocabulary retention. The results indicated that words learned under the binaural condition were recalled more accurately in both the short-term and the one-week delayed tests, showing a statistically significant advantage over those learned under the monaural condition. In the following sections, we interpret and contextualize these findings within existing theories and prior work, integrate preliminary EEG observations, and propose directions for future research.

5.1 Behavioral Findings and Theoretical Perspectives

A central theoretical framework relevant to these results is Dual Coding Theory [15], which posits that verbal information and nonverbal imagery rely on partially separate cognitive subsystems. Providing multiple forms of representation (textual plus sensory) can strengthen recall through the formation of multiple retrieval cues. While most dual-coding applications in second language learning emphasize visual aids [19], [60], the present study explored “imagery through sound” via binaural recordings that produce three-dimensional auditory effects. The heightened realism may have enabled participants to form stronger mental representations of word meaning. Indeed, vocabulary items associated with especially salient or dynamic auditory cues (e.g., “blast,” “whirl”) were recalled more reliably, corroborating the notion that *imageability* facilitates memory [60].

Additionally, participants in the binaural condition often exhibited spontaneous gestures and vocal repetitions more frequently than those in the monaural condition. According to embodied cognition theories, coupling motor actions or gestures with linguistic inputs can reinforce memory traces [50], [61]. A plausible interpretation is that the three-dimensional spatial cues elevated learners’ sense of presence, prompting them to engage physically with the learning material and further enhance retention. Similar findings have been reported in studies linking gestures to improved second language vocabulary recall [51].

At the same time, binaural audio alone may not suffice for representing more abstract or metaphorical words (e.g., “maraud,” “suckle”). These were remembered less effectively, mirroring earlier research on how the concreteness of a word affects recall [60]. Future approaches could integrate visual or textual enhancements to convey nuanced meanings. Furthermore, it remains an open question which technical attributes of three-dimensional audio—directional localization, distance cues, or reverberation—might most effectively bolster vocabulary retention.

Exploratory EEG Findings: Although the primary focus of our study was on behavioral outcomes, we also conducted a preliminary EEG(Electroencephalography) investigation on a separate group of participants who listened to the same Binaural/Monaural stimuli. This group consisted of 7 right-handed male participants (M_{age} = 22.1, SD = 0.9 years). Because these EEG participants did not overlap with the main behavioral group and the number of valid EEG trials was limited by strict artifact rejection, these neural observations must be treated as exploratory. However, they offer potential insights into the neurocognitive mechanisms that underlie immersive-audio learning.

Overview of ERSP Analysis: An Event-Related Spectral Perturbation (ERSP) measures the *time-varying* changes in EEG power within specific frequency bands. It compares the spectral power following stimulus onset (e.g., auditory presentation) to a baseline period preceding the stimulus:

$$\mathrm{ERSP}(f,t) = 10 \times \log_{10}\left(\frac{\mathrm{Power}_{\mathrm{stimulation}}(f,t)}{\mathrm{Power}_{\mathrm{baseline}}(f)}\right),$$

where $\text{Power}_{\text{stimulation}}(f, t)$ is the signal power at frequency f in time window t post-stimulus, and $\text{Power}_{\text{baseline}}(f)$ is the average power for the same frequency band during a baseline interval $[-200, 0]$ ms.

For our analysis, the post-stimulus 500 ms window was segmented into five 100 ms sub-intervals, and we examined four frequency bands: **theta** (4–8 Hz), **lower alpha** (8–11 Hz), **higher alpha** (11–13 Hz), and **beta** (13–40 Hz). EEG signals were recorded using a Polymate Pocket MP208 system (Miyuki Giken) from midline electrodes Fz, Cz, and Pz (international 10–20 system), as shown in Figure 3. This is a common practice for memory and attention studies [62], [63]. To minimize physical interference between the EEG sensors and headphones, each participant's head was stabilized using a chin rest, and an expert visually monitored the recordings for noise. Trials with large artifacts (e.g., exceeding $\pm 100\,\mu$V) or minimal amplitude differences (indicating possible amplifier saturation) were excluded.

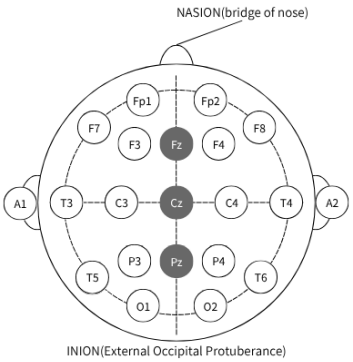


Fig. 3 The position of sensor

Theta-Band Activity and Memory Trends: Preliminary correlations between ERSP and vocabulary recall (the latter obtained from a separate behavioral test) suggested that *theta-band* (4–8 Hz) activity at the Pz site, especially within the $[0, 100)$ ms and $[300, 400)$ ms intervals, was positively related to recall accuracy. These correlations did not remain statistically significant after robust corrections for multiple comparisons; nevertheless, they align with a body of literature that implicates midline and parietal theta oscillations in episodic memory encoding and retrieval [62]~[64]. The ERSP analysis results are visualized in Figure 4.

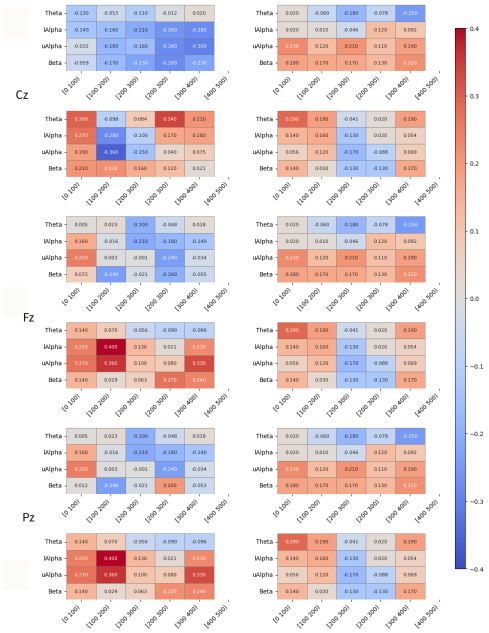


Fig. 4 The result of analysis

A plausible interpretation is that the immersive quality of binaural audio may heighten attention or cognitive engagement, supported by two key points. First, previous research indicates that parietal theta power is related to attention and cognitive engagement. Second, our ERSP analysis found that memory success or failure was tied to changes in theta power. Together, these findings suggest that improvements in word memory could hinge on enhanced attention or cognitive engagement. Meanwhile, our behavioral data showed that binaural stimuli improved word-memory performance. Hence, it is reasonable to hypothesize that the immersive nature of binaural audio boosts attention or cognitive engagement, ultimately leading to better word-memory outcomes, as evidenced by elevated theta power in parietal regions. This finding resonates with our behavioral observations that binaural stimuli induced more active, embodied responses. Although these EEG results cannot be generalized due to separate participants and a small trial count, they highlight the potential for immersive audio to influence both behavioral and neural correlates of memory.

5. 2 Limitations and Future Directions

This work has several limitations that frame our findings and guide future research.

First, a key limitation is the lack of quantitative data on participants' physical behaviors. While we anecdotally observed that participants in the binau-

ral condition appeared to use more gestures and vocal repetitions, this was not systematically measured or statistically compared. Therefore, we cannot rule out the possibility that these behaviors, rather than the binaural audio itself, contributed to the observed memory improvement. We hypothesize that the immersive quality of binaural audio may naturally increase learners' engagement, with these behaviors being a byproduct. Disentangling the direct effect of immersive audio from the secondary effect of physical engagement is a critical task for future research.

Second, our observation that words with clear auditory associations (e.g., "blast") yielded higher accuracy lacks formal statistical support. A post-hoc statistical comparison between word categories was not performed. This suggests a valuable direction for future work: studies should be designed with word characteristics as a primary factor to confirm this effect.

Third, as our study design suggests, the technique is most effective for words with concrete, spatial meanings, and its utility for abstract vocabulary is restricted. This limited application method is a key consideration for practical application.

Finally, our reliance on multiple-choice items captures only one dimension of vocabulary knowledge (receptive recognition). The EEG data were also exploratory and drawn from a separate sample.

However, these findings underscore the promise of binaural audio to improve the retention of the L2 vocabulary. The preliminary trends in EEG suggest a neural basis for such benefits, possibly linked to increased attentional or memory-related processing. Future investigations may:

- **Collect simultaneous EEG and behavioral data**: Matching neural responses and recall outcomes within the same participants could elucidate individual differences and more precise correlations.
- **Explore diverse audio attributes**: Systematically alter depth cues, reverberation, or directional movement to isolate which binaural features most effectively enhance learning.
- **Adopt varied assessment forms**: Incorporate free-recall, oral production, or other tasks that capture deeper vocabulary mastery and practical language use.

Building on these directions, it is possible to further clarify both the pedagogical value and the neurocognitive underpinnings of immersive auditory techniques in the acquisition of second language.

6. Future work

Future research may explore the integration of binaural audio with multimodal learning environments, such as virtual or augmented reality, in order to examine whether the simultaneous presence of three-dimensional visual and auditory stimuli yields additional improvements in vocabulary retention. Another line of inquiry may involve exploring how binaural audio interacts with task-induced cognitive processing, for instance by manipulating the degree of “search” and “evaluation” as specified in the Involvement Load Hypothesis [14], to determine whether immersive audio can amplify the benefits of deeper cognitive engagement.

In conclusion, this research demonstrates the potential of binaural audio for facilitating English vocabulary learning by providing immersive and emotionally engaging experiences that appear to support stronger memory traces. The findings align with Dual Coding Theory and suggest that multisensory encoding—particularly when enhanced by three-dimensional sound—can increase both short-term and longer-term word recall. At the same time, the effects are likely moderated by factors such as word concreteness, individual learner differences, and the need for complementary modalities. By investigating these nuances and building on emerging technologies, educators and designers of language-learning materials can harness the advantages of binaural audio to create more dynamic and effective instructional experiences in the field of second language acquisition.

Disclosure of Conflicts of Interest

There are no conflicts of interest to report for this article.

Acknowledgements

This work was supported in part by JSPS KAKENHI No. JP24834281, UTokyoGSC and JST-Mirai Program Grant Number JPMJMI21D3, Japan

Reference

[1] Alderson, J. C.: Reading in a Foreign Language: A Reading Problem or Language Problem?, *Read-*

ing in a Foreign Language (Alderson, J. C. and Urquhart, A. H., eds.), Longman (1984).

[2] Deville, G., Vandecasteele, M., Ostyn, P. and Kelly, P.: Measuring a FL learner's lexical needs, *Proceedings of the 5th European LSP Symposium*, Leuven, Belgium (1985).

[3] Nation, P. and Coady, J.: Vocabulary and reading, *Vocabulary and Language Teaching* (Carter, R. and McCarthy, M., eds.), Longman, London, pp. 97–110 (1987).

[4] Laufer, B.: What percentage of lexis is essential for comprehension, *From Humans Thinking to Thinking Machines* (Lauren, C. and Nordman, M., eds.), Multilingual Matters, Clevedon, pp. 316–323 (1989).

[5] Laufer, B.: Vocabulary Acquisition in a Second Language: Do Learners Really Acquire Most Vocabulary by Reading? Some Empirical Evidence, *The Canadian Modern Language Review*, Vol. 59, pp. 567–587 (2003).

[6] Laufer, B.: How Much Lexis is Necessary for Reading Comprehension?, *Vocabulary and Applied Linguistics* (Arnaud, P. J. L. and Béjoint, H., eds.), Palgrave Macmillan, London, pp. 126–132 (online), 10.1007/978-1-349-12396-4_12 (1992).

[7] Zeeland, H. V. and Schmitt, N.: Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension?, *Applied Linguistics*, Vol. 34, No. 4, pp. 457–479 (2013).

[8] Laufer, B. and Ravenhorst-Kalovski, G. C.: Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension, *Reading in a Foreign Language*, Vol. 22, No. 1, pp. 15–30 (2010).

[9] Qian, D. D.: Investigating the Relationship Between Vocabulary Knowledge and Academic Reading Performance: An Assessment Perspective, *Language Learning*, Vol. 52, No. 3, pp. 513–536 (online), 10.1111/1467-9922.00193 (2002).

[10] Lewis, M.: *The Lexical Approach: The State of ELT and a Way Forward*, LTP teacher training, Language Teaching Publications (1993).

[11] Lewis, M.: Implementing the Lexical Approach: Putting Theory into Practice, (online), available from (https://api.semanticscholar.org/CorpusID:60992313) (1997).

[12] Krashen, S.: Some issues relating to the monitor model, *Teaching and Learning English as a Second Language: Trends in Research and Practice* (Brown, H., Yorio, C. and Crymes, R., eds.), Teachers of English to Speakers of Other Languages, Washington, DC, pp. 144–158 (1977). Selected Papers from the Eleventh Annual Convention of Teachers of English to Speakers of Other Languages, Miami, Florida, April 26 – May 1, 1977.

[13] Krashen, S.: *Explorations in Language Acquisition and Use* (2003).

[14] Hulstijn, J. H. and Laufer, B.: Some empirical evidence for the involvement load hypothesis in vocabulary acquisition, *Language Learning*, Vol. 51, No. 3, pp. 539–558 (2001).

[15] Paivio, A.: Dual coding theory: Retrospect and current status, *Canadian Journal of Psychology/Revue canadienne de psychologie*, Vol. 45, No. 3, pp. 255–287 (1991).

[16] Clark, J. M. and Paivio, A.: A dual coding perspective on encoding processes, *Imagery and Related Mnemonic Processes: Theories, Individual Differences, and Applications*, Springer New York, New York, NY, pp. 5–33 (1987).

[17] Çakici, D.: The use of ICT in teaching English as a foreign language, *Participatory Educational Research*, Vol. 4, No. 2, pp. 73–77 (2016).

[18] Miura, T.: Report on a ' Not-So-Expensive ICT Seminar for English Language Teachers '., *Annual Bulletin of the Centre for Culture and Language Education, Tohoku University*, Vol. 9 (2023).

[19] Zhu, Y., Wang, Y., Yu, C., Shi, S., Zhang, Y., He, S., Zhao, P., Ma, X. and Shi, Y.: ViVo: Video-augmented dictionary for vocabulary learning, *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 5568–5579 (online), 10.1145/3025453.3025779 (2017).

[20] : *Moetan, DS*, FACTORY Co.: IDEA, Ltd. (2008).

[21] O., H.: *Moesta Moerutodaieigojyuku*, NOISE FACTORY Co., Ltd. (2014).

[22] Ogura, H.: *Maruoboe Eitango 2600*, KADOKAWA CORPORATION (2014).

[23] Fukushima, S.: EmoTan: enhanced flashcards for second language vocabulary learning with emotional binaural narration, *Research and Practice in Technology Enhanced Learning*, Vol. 14, No. 1 (online), 10.1186/s41039-019-0109-0 (2019).

[24] Shimizu, K., Fukushima, S. and Naemura, T.: Effects of Binaural Audio on English Vocabulary Learning, *Proceedings of the 30th International Conference on Computers in Education*, Vol. 2, pp. 665–667 (2022).

[25] Weblio: Online dictionary (2005). Available at `http://www.weblio.jp/`.

[26] NTT Resonant Incorporated: Goo Dictionary, `https://dictionary.goo.ne.jp/` (1999). Accessed: 11 September 2024.

[27] Houghton Mifflin: American Heritage Dictionary, `https://ahdictionary.com/` (1969). Accessed: 11 September 2024.

[28] Oxford University Press: Oxford Learner's Dictionaries, `http://www.oxfordlearnersdictionaries.com/` (1948). Accessed: 11 September 2024.

[29] mikan: English learning application (2014). Available at `http://mikan.link/`.

[30] Smart Language Apps Limited: Learn English (US) Flashcards, `https://itunes.apple.com/us/app/learn-english-us-flashcards/id970002864?mt=8` (2015). Accessed: 11 September 2024.

[31] Hwang, G.-J. and Wu, P.-H.: Applications, impacts and trends of mobile technology-enhanced learning: a review of 2008-2012 publications in selected SSCI journals, *International Journal of Mobile Learning and Organisation*, Vol. 8, No. 2, pp. 83–95 (online), 10.1504/IJMLO.2014.062346 (2014).

[32] Hwang, G.-J. and Fu, Q.-K.: Trends in the research design and application of mobile language learning: a review of 2007–2016 publications in selected SSCI journals, *Interactive Learning Environments*, Vol. 27, No. 4, pp. 567–581 (online), 10.1080/10494820.2018.1486861 (2019).

[33] Wright, S., Fugett, A. and Caputa, F.: Using e-readers and internet resources to support comprehension, *Educational Technology & Society*, Vol. 16, No. 1, pp. 367–379 (2013).

[34] Hong, R., Wang, M., Xu, M., Yan, S. and Chua, T.: Dynamic captioning: Video accessibility enhancement for hearing impairment, *Proceedings of the 18th ACM International Conference on Multimedia*, Firenze, Italy, pp. 421–430 (2010).

[35] Hu, Y., Kautz, J., Yu, Y. and Wang, W.: Speaker-following video subtitles, *ACM Transactions on Multimedia Computing, Communications, and Applications*, Vol. 11, No. 4, pp. 1–17 (online), 10.1145/2795100 (2015).

[36] Brown, A., Jones, R., Crabb, M., Sandford, J., Brooks, M., Armstrong, M. and Jay, C.: Dynamic subtitles: The user experience, *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, Brussels, Belgium, pp. 103–112 (2015).

[37] Nishida, Y., Kusunoki, F., Hiramoto, M. and Mizoguchi, H.: Learning by Doing: Space-Associate Language Learning Using a Sensorized Environment, *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Edmonton, AB, Canada, pp. 3636–3641 (2005).

[38] Hautasaari, A., Hamada, T., Ishiyama, K. and Fukushima, S.: VocaBura: A Method for Supporting Second Language Vocabulary Learning While Walking, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 3, No. 4, pp. 1–23 (online), 10.1145/3369822 (2020).

[39] Dearman, D. and Truong, K.: Evaluating the implicit acquisition of second language vocabulary using a live wallpaper, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Austin, TX, USA, pp. 1391–1400 (online), 10.1145/2207676.2208598 (2012).

[40] Al-Mekhlafi, K., Hu, X. and Zheng, Z.: An Approach to Context-Aware Mobile Chinese Language Learning for Foreign Students, *2009 Eighth International Conference on Mobile Business*, pp. 340–346 (online), 10.1109/ICMB.2009.65 (2009).

[41] Ogata, H. and Yano, Y.: Context-aware support for computer-supported ubiquitous learning, *The 2nd IEEE International Workshop on Wireless and Mobile Technologies in Education, 2004. Proceedings*, pp. 27–34 (online), 10.1109/WMTE.2004.1281330 (2004).

[42] Hautasaari, A., Hamada, T., Ishiyama, K. and Fukushima, S.: VocaBura: A Method for Supporting Second Language Vocabulary Learning While Walking, *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, Vol. 3, No. 4 (online), 10.1145/3369824 (2020).

[43] Mizuho, T., Narumi, T. and Kuzuoka, H.: Exploratory Study on the Reinstatement Effect Under 360-Degree Video-Based Virtual Environments, *29th ACM Symposium on Virtual Reality Software and Technology*, Christchurch, New Zealand (2023).

[44] Ebert, D., Gupta, S. and Makedon, F.: Ogma: A virtual reality language acquisition system, *Proceedings of the 9th ACM International Conference on Pervasive Technologies Related to Assistive Environments*, pp. 1–5 (2016).

[45] Vázquez, C., Xia, L., Aikawa, T. and Maes, P.: Words in Motion: Kinesthetic Language Learning in Virtual Reality, *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, Mumbai, India, pp. 272–276 (online), 10.1109/ICALT.2018.00068 (2018).

[46] Ratcliffe, J., Ballou, N. and Tokarchuk, L.: Actions, not gestures: Contextualising embodied controller interactions in immersive virtual reality, *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*, Osaka, Japan, pp. 1–11 (2021).

[47] Brübach, L., Westermeier, F., Wienrich, C. and Latoschik, M. E.: Breaking Plausibility Without Breaking Presence—Evidence For The Multi-Layer Nature Of Plausibility, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 28, No. 5, pp. 2267–2276（オンライン）, 10.1109/TVCG.2022.3152275 (2022).

[48] Li, S., Gu, X., Yi, K., Yang, Y., Wang, G. and Manocha, D.: Self-Illusion: A Study on Cognition of Role-Playing in Immersive Virtual Environments, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 28, No. 5, pp. 3035–3049 (online), 10.1109/TVCG.2022.3151272 (2022).

[49] Ogawa, N., Narumi, T., Kuzuoka, H. and Hirose, M.: Do You Feel Like Passing Through Walls?: Effect of Self-Avatar Appearance on Facilitating Realistic Behavior in Virtual Environments, *2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, HI, USA, pp. 1–14 (online), 10.1145/3313831.3376163 (2020).

[50] Macedonia, M. and Knösche, T. R.: Body in Mind: How Gestures Empower Foreign Language Learning, *Mind, Brain, and Education*, Vol. 5, No. 4, pp. 196–211 (online), 10.1111/j.1751-228X.2011.01129.x (2011).

[51] Tellier, M.: The effect of gestures on second language memorisation by young children, *Gesture*, Vol. 8, No. 2, pp. 219–235 (online), 10.1075/gest.8.2.06tel (2008).

[52] Zhang, X. and Zuber, S.: The effects of language and semantic repetition on the enactment effect of action memory, *Frontiers in Psychology*, Vol. 11, p. 515 (online), 10.3389/fpsyg.2020.00515 (2020).

[53] Hung, H.-T., Yang, J.-C., Hwang, G.-J., Chu, H.-C. and Wang, C.-C.: A scoping review of research on digital game-based language learning, *Computers and Education*, Vol. 126, pp. 89–104 (online), 10.1016/j.compedu.2018.07.001 (2018).

[54] Kensinger, E. A. and Corkin, S.: Memory enhancement for emotional words: are emotional words more vividly remembered than neutral words?, *Memory & Cognition*, Vol. 31, No. 8, pp. 1169–1180 (online), 10.3758/BF03195800 (2003).

[55] Phelps, E. A., LaBar, K. S. and Spencer, D. D.: Memory for emotional words following unilateral temporal lobectomy, *Brain and Cognition*, Vol. 35, No. 1, pp. 85–109 (online), 10.1006/brcg.1997.0929 (1997).

[56] Barratt, E. L. and Davis, N. J.: Autonomous Sensory Meridian Response (ASMR): a flow-like mental state, *PeerJ*, Vol. 3, p. e851 (2015).

[57] CA Sega Joypolis Ltd.: 3D Sound Attraction of Joypolis (2016).

[58] Mather, M. and Sutherland, M. R.: Arousal-Biased Competition in Perception and Memory, *Perspectives on Psychological Science*, Vol. 6, No. 2, pp. 114–133 (online), 10.1177/1745691611400234 (2011).

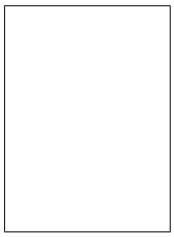
[59] LaBar, K. S. and Cabeza, R.: Cognitive Neu-

roscience of Emotional Memory, *Nature Reviews Neuroscience*, Vol. 7, No. 1, pp. 54–64 (2006).
[60] Clark, J. M. and Paivio, A.: Dual Coding Theory and Education, *Educational Psychology Review*, Vol. 3, No. 3, pp. 149–210 (1991).
[61] Macedonia, M., Müller, K. and Friederici, A. D.: The Impact of Iconic Gestures on Foreign Language Word Learning and Its Neural Substrate, *Human Brain Mapping*, Vol. 32, No. 6, pp. 982–998 (2011).
[62] Klimesch, W.: EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis, *Brain Research Reviews*, Vol. 29, No. 2–3, pp. 169–195 (1999).
[63] Hsieh, L.-T. and Ranganath, C.: Frontal midline theta oscillations during working memory maintenance and episodic encoding and retrieval, *NeuroImage*, Vol. 85, pp. 721–729 (2014).
[64] Kahana, M. J.: The cognitive correlates of human brain oscillations, *Journal of Neuroscience*, Vol. 26, No. 6, pp. 1669–1672 (2006).

(received May 7, 2025, revised Jul. 29)

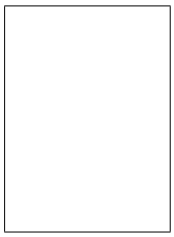
Biography

Kosuke Shimizu (Student Member)



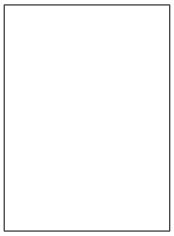
Kosuke Shimizu is currently an undergraduate student at the University of Tsukuba, School of Informatics. He is interested in the field of Virtual Reality and Human-Computer Interaction. He is a student member of the Human Interface Society.

Shogo Fukushima (Member)



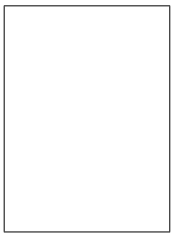
Shogo Fukushima received his Ph.D. in Engineering from the University of Electro-Communications in 2013. After serving as a JST PRESTO Researcher and Assistant Professor at the Graduate School of Interdisciplinary Information Studies, The University of Tokyo, he has been an Associate Professor at the Faculty of Information Science and Electrical Engineering, Kyushu University, since 2022. During his doctoral studies, he spent a research period at MIT. His research interests include virtual reality (VR), emotional intelligence augmentation, and technology-enhanced learning (TEL).

Hirokazu Doi



Hirokazu Doi holds a Ph.D. in Academics and is currently Professor in the Information and Management Systems Engineering course at Nagaoka University of Technology. His research interests include cognitive neuroscience, behavioral endocrinology, digital phenotyping, and internal state inference. He is a member of the Japanese Psychological Association, Cognitive Science Society of Japan, IEICE, Japanese Society for Baby Science, and Japanese Society for Social Welfare Management.

Takeshi Naemura



Takeshi Naemura received his Ph.D. in Engineering from The University of Tokyo in 1997. Since 2013, he has been Professor at the Graduate School of Interdisciplinary Information Studies, The University of Tokyo. He currently serves as Vice President of the Virtual Reality Society of Japan and chairs the Handbook/Knowledge-Base Committee of Institute of Electronics, Information and Communication Engineers. His research interests include augmented reality, interaction design, creativity support, and human–computer media interfaces.