Proceedings of ICEIPI 2025 Symposium: Understanding Religious Identity in Educational Contexts
DOI: 10.54254/2753-7048/2025.ND26257

Enhancing Listening Comprehension: A Multimodal Approach Grounded in Cognitive Load and Coding Theories

Yuqing Zhou

Faculty of Humanities and Social Science, Beijing Normal-Hong Kong Baptist University, Zhuhai, China
s230025117@mail.uic.edu.cn

Abstract. Listening, as a foundational skill in language acquisition, is crucial for effective communication and cultural understanding. Traditional teaching methods tend to overburden the phonological working memory while failing to fully engage the brain's visuospatial systems. These difficulties highlight the need for learner-centered techniques. Since 2010, the field has witnessed a significant shift: advances in multimedia technology, virtual reality (VR), and mobile-assisted language learning (MALL) have enabled the integration of multimodal input—audio, visual, and kinesthetic— into second language (L2) listening instruction. This marks a clear inflection point in pedagogical research, where cognitive theories have begun to intersect meaningfully with digital innovation. This paper reviewes multimodal L2 listening within the framework of cognitive load theory. The findings reveal that multimodal approaches significantly enhance L2 listening comprehension by reducing cognitive overload and increasing learner engagement, thereby supporting the development of more efficient listening strategies in language learners. This paper offers a novel pathway to enhance comprehension and engagement in language learners.

Keywords: Listening Comprehension, Multimodal Approach, Dual Coding Theory, Cognitive Load Theory

1. Introduction

Listening is the complex, dynamic process which uses neurological, linguistic, semantic and pragmatic systems to turn auditory input into understanding [1]. As the main channel for second language (L2) learners to receive comprehensible input—"i+1" in Krashen's terms—listening is crucial to interlinguistic development. Such a multifaceted model highlights the mental process of listening and its situational flexibility. Conventional teaching excessively burdens the phonological working memory, and does not exploit sufficiently the visuospatial systems of the brain in drawing on contextual inferences [1]. Most English as a Foreign Language (EFL) classrooms, teachers rely heavily on scripted materials, minimizing contextual relevance and communicative value. This lecture-centered model limits student interaction and reflective learning. Under this circumstance, learners passively receive information, focusing on test scores over skill development. These challenges highlight the urgent need for interactive, context-rich, and learner-centered approaches.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7048/2025.ND26257

This article explores multimodal L2 listening through the lens of cognitive load theory. It contributes theoretically by bridging cognitive theory and language pedagogy, and practically by offering evidence-based guidance for designing effective, engaging, and neurologically aligned listening instruction in the L2 classroom.

2. Theoretical framework

This study proposes a Multimodal Integrated Listening Comprehension (MILC) framework that synthesizes three complementary cognitive theories to explain how multimodal input enhances listening comprehension in L2 contexts.

Dual Coding Theory (DCT) suggests that information is processed through two semi-independent channels: verbal and imagery [2]. DCT assumes that data can be coded and stored in either one or both systems with some major implications that are applicable to Multimodal Information Learning (MIL). Firstly, there is additive effect where joint presentation of information in both verbally and non-verbally possible increases chances of recall compared to presentation of information in only one of the encoding system, forming numerous retrieval pathway. Secondly, there is also concreteness advantage according to which is easier to process and remember concrete concepts comparing with abstract ones, they easily touch verbal and non-verbal codes. Visual aids can hence, be used to make abstract information more concrete facilitating memory and understanding. Also, DCT stresses the role of the representational associations, including the links between non-verbal and verbal representations, like the association of the word “dog” with an image, which facilitates the understanding and memory. These associations result in a better put-together mental model of the information. DCT will therefore tell why a combination of listening in connection with visual or textual data input in MIL can be able to improve listening comprehension, and learning results by developing breadth and depth to mind representations [3].

Cognitive Load Theory (CLT) complements DCT by emphasizing the limited capacity of working memory. CLT distinguishes three types of cognitive load: intrinsic (task complexity), extraneous (inefficient presentation), and germane (schema-building effort) [4]. CLT focuses on reducing such a burden with evidence-based design principles in mind. Germane load is positive cognitive work aimed at achieving schema acquisition, automation, and assembling of powerful pattern of knowledge in the long-term memory. The crux of the teaching goal of CLT is to regulate the overall mental taxation in the limits of working memory regulated through a reduction in extraneous load and supporting germane load. In listening comprehension field, CLT offers an analytic tool applying to the analysis of relieving or increasing cognitive loads by using MIL interventions. For instance, a schematic diagram that is accompanied with an auditory description can lower extraneous load by freeing spatial processing ability up to the auditory working memory, but irrelevant animations could, on the other hand, increase it, as Mayer and Moreno proved [5].

By operationalizing this framework, research can systematically compare modes (audio-only, audio+captions, audio+visuals, full MILC) and assess not only performance but cognitive load, using validated measures like mental-effort scales and eye-tracking [6]. Such studies should also account for learner differences—novice learners may require more structured visual aids, whereas experts may benefit from reduced support (expertise-reversal effect) [7].

DOI: 10.54254/2753-7048/2025.ND26257

3. Current applications of multimodal integrated learning in listening comprehension

3.1. Modality integration effecton listening comprehension

Extensive research into the field shows that audiovisual multimodal content always results in more gains in listening comprehension provided it is carefully combined. The initial experiments by Suvorov demonstrated that the visual context (in the form of photographs or video) was helpful to learners than the audio-only context that enhanced topical knowledge and interest [8]. A recent review showed that audio-visual input provides more rich contextual clues and results in much better comprehension outcomes in comparison to unimodal input [9]. These improvements align with CLT, as visuals help offload auditory processing and manage intrinsic cognitive demand [3,4].

Audio-Visual integration pairs spoken language with visual elements, encompassing dynamic formats like videos and animations, and static formats like photographs or illustrations. This common form of MIL makes direct use of Dual Coding Theory allowing the encoding of information to take place simultaneously of into verbal and a non-verbal image system. Simultaneously, it conforms to the theory of Cognitive Load because the goal is the diminution of extraneous cognitive load, the carefully selected images can explain abstract ideas, display space correlation, and divide tedious stories, thus lessening the strains on the auditory working memory. An iconic example of its use is in TED-Ed lessons, where professional talks would be used in combination with animated explainer videos that visually illustrate exemplifying notions, processes or allegories. The studies validate the fact that these A-V materials lead to a huge jump in the comprehension and retention levels when compared to audio only presentations.

Audio-Textual integration supplements auditory input with written text, including captions, subtitles, transcripts, keyword displays, or summaries. The main peculiarity of this combination of modalities is the use of the potential of adding to the verbal system of DCT making the connections between the auditory-verbal and visual-verbal (orthographic) code. As a CLT perspectives states, its overall objective is to strategically manage intrinsic and extraneous load and it tends to be a scaffold. This can be seen in language learning websites such as Duolingo and FluentU, which provide videos or dialogues, with optional and frequently interactive, subtitles and transcriptions. Managing cognitive load is adaptively done by learners by being able to toggle text support, clicking to see definitions, and replaying parts. When proficiency is enhanced, learners are able to minimize the need to make use of text therefore maximizing germane load to construct schema to greater depths.

Audio-Kinesthiastic/Tactile fusion is a combination of audible perception with physical actions like movement, handling of objects or moving through space, more commonly facilitated by means of immersive experiences, like VR and AR. The form reflects the DCT in its connection between auditory-verbal codes and motoric and spatial ones, possibly building up stronger, embodied memory traces based on the concept of embodied cognition. Its main CLT objective is to improve the germane load through more experiential learning. For instance VR language simulations, where the learner follows instructions or conversations and manages virtual objects or tours an environment (e.g., a virtual café scenario). While promising for creating rich contextual learning, significant CLT considerations involve managing potential extraneous load induced by the inherent complexity of the virtual environment itself [10].

Captioning (both L1 and L2 subtitles) is one of the most widely investigated multimodal strategies. Scholars found that captioned videos significantly improve listening comprehension and memory, particularly for vocabulary. However, Diao et al. caution that captions may induce extraneous load if learners focus on reading rather than listening [11]. Zhang et al. examined how

DOI: 10.54254/2753-7048/2025.ND26257

AI-generated subtitles affected Thai EFL learners and found that auto-subtitles can enhance comprehension and satisfaction, although edited subtitles were slightly superior in cognitive load efficiency [12]. Yang extended this by showing that dual-code display models supported all proficiency levels but learners with low visual aptitude reported higher loads, suggesting that caption benefits may vary across individual differences [13].

These types of visual support strongly mediate efficacy: static images such as diagrams or maps generally support comprehension without increasing extraneous load, especially when placed near corresponding audio cues to avoid split-attention effects [14]. In contrast, full-motion video yields mixed results: relevant, content-rich video often supports inference and contextualization but irrelevant animation or multitasking visuals can overload working memory [15]. Learners exposed to busy or decorative visual distractors performed worse than those receiving minimal, aligned visuals (Reddit commentary; seductive-details research). Suvorov and Wagner both report that content-aligned visuals improve performance, whereas extraneous imagery splits attention and diminishes focus.

3.2. Cognitive load measurement and individual differences

Accurate measurement of cognitive load is critical for validating the theoretical assumptions of multimodal instruction. As cognitive load is not directly observable, scholars have developed a taxonomy of subjective, behavioral, physiological, and performance-based indicators. Subjective tools, such as NASA-TLX or Likert-scale mental effort ratings, are widely used for their simplicity but may lack precision. To overcome these limitations, objective and real-time methods have gained traction. For instance, eye-tracking metrics like fixation duration, saccadic movement, and percentage dwell time offer moment-to-moment indicators of attentional allocation and processing effort [16]. When learners are forced to divide attention across poorly aligned multimodal elements —such as separated subtitles and imagery—studies consistently show increased fixation durations and regression counts, indicating higher extraneous load [3, 17].

Electroencephalography (EEG) provides another layer of precision, capturing neural correlates of cognitive load. Research has shown that increased theta activity and suppressed alpha rhythms, especially in the frontal cortex, are positively correlated with working memory engagement during complex listening tasks [18,19]. These physiological findings corroborate CLT's assumptions regarding limited cognitive capacity and help triangulate data from eye-tracking and behavioral outcomes. Integrating these multimodal measures allows researchers to track not only the existence but also the temporal evolution of cognitive load during listening comprehension activities.

Beyond measurement, individual learner differences substantially mediate the effects of multimodal instruction. Factors such as proficiency level, working memory capacity (WMC), and cognitive style have all been identified as moderators. Sung and Mayer found that learners with high visual-spatial ability showed significant gains from image-enhanced materials, whereas verbal-dominant learners benefited less [20]. In early stages of language acquisition, learners may benefit from rich visual supports to reduce intrinsic load and scaffold meaning. However, as proficiency increases, the same supports may become redundant, or even disruptive, by imposing unnecessary processing demands—thereby increasing extraneous load.

Such results evidence the need of an adaptive multimodal design. Learner-responsive instructional systems should replace the usual forms of instruction, which are always static and provide visual and verbal stimuli, depending on the real-time learner profile and cognitive load indices. Potentially, emerging systems combining AI-driven personalization and neuroadaptive feedback can become solutions in this respect. Further investigations should also be done on how

DOI: 10.54254/2753-7048/2025.ND26257

load-sensitive multimodal systems may dynamically adjust the level of support to ensure optimum germane load with regards to proficiency levels.

3.3. Advantages of MIL based on CLT and DCT principles

Multimodal Integrated Learning, when strategically designed, offers significant advantages for listening comprehension. This approach is grounded in the robust theoretical foundations of Cognitive Load Theory (CLT) and Dual Coding Theory (DCT).

First, MIL demonstrably reduces extraneous cognitive load. Well-integrated non-auditory modalities serve to clarify ambiguities inherent in auditory input, illustrate spatial configurations, define unfamiliar vocabulary contextually (visually or textually), and segment intricate information streams. This directly alleviates the burden on limited auditory working memory resources. For example, a diagram matching a process that is explained verbally allows the students not to waste too much mental energy on creating flawed mental images that one might have, dedicating cognitive capacities to a better understanding instead..

Second, MIL significantly enhances encoding and retrieval processes through the mechanism of dual coding. By presenting information simultaneously through verbal channels (auditory speech, written text) and non-verbal channels (visuals, spatial arrangements, kinesthetic feedback), MIL facilitates the encoding of information into multiple, semi-independent representational systems within long-term memory. This creates richer, elaborate memory traces and establishes multiple potential retrieval paths, leading to measurably improved accuracy in comprehension and superior long-term retention [2]. The simultaneous presentation of an image alongside its auditory description epitomizes this creation of interconnected verbal and imaginal codes.

Third, MIL effectively supports schema construction and automation, thereby optimizing germane cognitive load. Complementary modalities provide crucial scaffolding for building accurate mental models of the listened content. Visuals can offer inherent organizational structure (e.g., timelines, flowcharts), while textual summaries can highlight critical relationships and hierarchies. MIL frees working memory thereby enabling the learner to spend more of their mental resources attending to the process of schema building and the automation of comprehension procedures which is the primary aim as defined by Sweller and others [4]. This germane processing is further enhanced by interactive MIL features, i.e., clicking on a visual item to elaborate it.

Fourth, MIL enhances learner engagement and motivation. Its multimodal presentations provide greater variety and dynamism than monomodal audio, reducing monotony and sustaining attention critical for maintaining the focused cognitive effort essential for successful listening comprehension. Furthermore, interactive MIL formats, particularly those utilizing VR, AR, or app-based interactivity, significantly heighten this motivational effect, as noted in the broader literature on multimedia learning by scholars like Plass et al. [6].

4. Theoretical implications

The accumulated findings confirm that integrating Communicative Language Teaching (CLT) and Dual Coding Theory (DCT) provides a robust explanatory framework. Multimodal listening comprehension is most effective when visual and verbal modalities are semantically aligned, extraneous information is minimized, and learners' cognitive profiles are taken into account. Mayer operationalizes this by advocating principles such as reducing unnecessary processing, ensuring coherence, maintaining contiguity, avoiding redundancy, and using signaling effectively. These

DOI: 10.54254/2753-7048/2025.ND26257

principles reinforce the demand for CLT to reduce unnecessary processing and leverage the dual-channel strengths theorized by DCT.

Instructional materials based on a Multimodal Integrated Listening Comprehension (MILC) design should synchronize audio and visual elements, reinforce verbal content without mere repetition, and avoid irrelevant or overly complex images. They should also provide more visual or textual aids for low-proficiency learners while reducing such aids for expert or highly verbal learners.

Objective load measurements, such as eye-tracking (fixation count, dwell time), EEG rhythms (frontal theta/alpha activity), and real-time mental effort ratings, can dynamically validate the effectiveness of these designs. Moreover, employing these objective cognitive load measurement tools allows for continuous validation of multimodal design efficacy and learner fit. This triangulated assessment approach strengthens causal inference and supports iterative instructional optimization.

5. Conclusion

This review presents compelling evidence that multimodal listening instruction—when undergirded by Cognitive Load Theory and Dual Coding Theory—enhances comprehension, retention, and engagement in second-language contexts. The key to effectiveness lies not simply in adding modalities, but in designing coherent, aligned, and cognitively calibrated multimodal experiences.

Nevertheless, this review is not without limitations. First, non-English literature was underrepresented, potentially introducing language or publication bias. Second, learners with auditory impairments — who may be uniquely impacted by multimodal strategies—were notably absent from the reviewed studies.

Future research should pursue adaptive MILC implementations that respond to learner feedback in real time, leveraging eye-tracking or neural data to adjust modality presentation dynamically. Controlled experiments could further test the interplay between modality, proficiency, and working memory constraints. Ultimately, establishing empirically driven, theoretically anchored instructional design frameworks will enable educators to transform multimodal listening materials from a pedagogical novelty into a scientifically validated methodology.

References

[1] Baddeley, A. (2003). Working memory and language. Journal of Communication Disorders.
[2] Paivio, A. (2006). Mind and its evolution: A dual coding theoretical approach. Psychology Press.
[3] Mayer, R. E. (2009). Multimedia learning (2nd ed.). Cambridge University Press.
[4] Sweller, J. (2011). Cognitive load theory. Psychology of Learning.
[5] Montero Pérez, M., Van Den Noortgate, W., & Desmet, P. (2014). Captioned video for L2 listening comprehension: A meta-analysis. System, 42, 79–95.
[6] Plass, J.L. (2003). Cognitive load in reading. Learning and Instruction.
[7] Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. Educational Psychology Review, 19(4), 509–539.
[8] Suvorov, R. (2009). Context visuals in L2 listening tests: The effects of photographs and video vs. audio-only format. Developing and evaluating language learning materials, 53-68.
[9] Shaojie, T., Samad, A. A., & Ismail, L. (2022). Systematic literature review on audio-visual multimodal input in listening comprehension. Frontiers in Psychology, 13, 980133.
[10] Van Merrienboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. Educational psychology review, 17(2), 147-177.
[11] Diao, Y., Chandler, P., & Sweller, J. (2007). Written text, diagrams, and cognitive load in learning from multimedia materials. Educational Psychology, 27(2), 129–141.

DOI: 10.54254/2753-7048/2025.ND26257

[12] Zhang, H., Zou, D., & Wang, J. (2022). Cognitive load and listening comprehension in AR-based learning. British Journal of Educational Technology, 53(2), 423–441. https: //doi.org/10.1111/bjet.13143
[13] Yang, H. Y. (2014). Does multimedia support individual differences?–EFL learners' listening comprehension and cognitive load. Australasian Journal of Educational Technology, 30(6).
[14] Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. Cognition and Instruction, 8(4), 293–332.
[15] Pink, A., & Newton, P. M. (2020). Decorative animations impair recall and are a source of extraneous cognitive load. Advances in Physiology Education.
[16] Zu, B., Liu, J., & Wang, S. (2018). Eye-tracking analysis of split-attention in multimedia learning. Computers in Human Behavior, 89, 108–117. [https: //doi.org/10.1016/j.chb.2018.07.033
[17] Liu, Y., Liu, M., Wang, Y., & Cheng, X. (2020). Eye-tracking research on split-attention in multimedia learning: A meta-analysis. Educational Psychology Review, 32, 1083–1114. https: //doi.org/10.1007/s10648-020-09530-3
[18] Antonenko, P., Paas, F., Grabner, R., & van Gog, T. (2010). Using electroencephalography to measure cognitive load. Educational Psychology Review, 22(4), 425–438.
[19] Debue, N., & Van De Leemput, C. (2014). What does neurophysiological evidence tell us about cognitive load theory? Frontiers in Psychology, 5, 1072.
[20] Sung, E., & Mayer, R. E. (2012). Supporting L2 vocabulary learning through imagery. Computer Assisted Language Learning, 25(4), 329–349.