Audiovisual Input and L2 Pronunciation Development: A Review of Caption Design, Attention, and Exposure Conditions

Jiaying Fu *

University of Nottingham, School of English, Nottingham, United Kingdom

* Corresponding Author Email: hvyjf1@nottingham.ac.uk

Abstract. This review analyzes research on how audiovisual input supports second language pronunciation development by examining perceptual mechanisms, caption design, and exposure conditions. Studies on multi-modal alignment show that captions can guide attention to phonetic cues when they are selectively displayed or temporally synchronized with the speech signal. However, in the condition of full captions and form-focused tasks often cause excessive cognitive load. Research on caption types demonstrates that partial captions, keyword captions, and audio-synchronized highlighting promote more efficient decoding than dense or static text. Although proficiency consistently predicts learning outcomes more strongly than caption mode, long-term evidence indicates that extensive viewing still has benefit on lexical consolidation, listening confidence, and gradual reduction of caption dependence. Overall, audiovisual input is most effective when textual support is minimal, well-timed, and embedded within sustained and increasingly authentic listening. The review identifies the need for longitudinal and dose–response designs to clarify how input amount shapes measurable phonetic gains.

Keywords: Audiovisual input, caption design, phonetic perception, cognitive load, extensive viewing.

1. Introduction

Pronunciation remains one of the most challenging in second language acquisition (SLA). A large body of research has supported that pronunciation is the domain that learners are least likely to reach the native-like proficiency [1, 2]. Over the past two decades, research on audiovisual input, including film, series, and online video platform is increasing, working on providing L2 learners with multi-modal access to audio speech and further exploring the new domain of exposure material. This shift has prompted a surge of research exploring whether and how audiovisual input, particularly when supported by captions, can enhance pronunciation and other fields of speech capacity.

While earlier studies focused primarily on listening comprehension and vocabulary gains, more recent research has investigated how audiovisual input affects phonetic perception, speech segmentation, and pronunciation accuracy [3-6]. However, as research focuses differently on task design, target sounds, and learning conditions, findings still remain fragmented.

Some studies demonstrate that there are obvious perceptual benefits of captioned input, whereas others notice limited or even negative effects with high cognitive load of different types of caption modes [5, 7]. Moreover, although few studies have explored the long-term exposure influence on vocabulary acquisition, while other studies integrate language perception and production, the relationship of time and pronunciation learning remains further explored [5, 8]. Meanwhile, the studies on phonetic production lack extension from small-scale samples to wide range of different phonetic categories.

Building on this foundation, it is crucial to synthesize existing evidence to understand how and under what conditions audiovisual input supports pronunciation development, considering exposure time and speech production. The present review aims to critically examine empirical studies from three complementary dimensions: (1) the perceptual mechanisms underlying multi-modal input, (2) the design and cognitive demands of different captioning modes, and (3) the effect of exposure variability and transfer. This paper aims to clarify the benefit of audiovisual input on facilitating L2 phonetic learning and to identify directions for future work.

2. Theory Framework

The influence of audiovisual input on pronunciation development can be supported by an integrated framework combining multi-modal learning, cognitive load, and phonetic learning theories. Rather than treating these models as isolated constructs, this review combines them as complementary lenses for explaining how visual and textual information interact with auditory input to promote L2 phonetic learning.

From a multi-modal perspective, learners process information through multiple sensory channels. Paivio's Dual Coding Theory and Mayer's Multimedia Learning Theor both propose that combining verbal and non-verbal input enhances memory and comprehension [9]. Paivio demonstrates that human cognition processes information through two channels simultaneously, combining sensory information of sound and vision [9]. Similarly, Mayer indicates that learning is more effective when information is presented through both words and pictures, building connections between them. These theories provide solitary foundation for audiovisual input in the context of pronunciation. Captions and visual cues can reinforce auditory signals, supporting the mapping of sounds to words. Studies such as Galimberti et al. and Montero-Pérez & Pattemore exemplify this principle, indicating that audio-synchronized textual enhancement and multi-modal video exposure led to faster lexical recognition and improved speech segmentation [6, 10].

However, multi-modal input is not always beneficial regardless of conditions. Sweller's Cognitive Load Theory (CLT) points out that working memory capacity limits how much information can be processed simultaneously in human's brain [11]. The efficiency of processing information may decline when learners need to comprehend the caption and pay attention to the audio at the same time, especially when the caption and audio are in different language. Wisniewska and Mora further supported this theory, indicating that meaning-focused viewing tasks with captions improved pronunciation, while form-focused captioned tasks overloaded learners and reduced gains [5]. Similarly, Pannatier and Bétrancourt observed that low-proficiency learners benefit from captions under appropriate cognitive load, while high-proficiency learners have reducing outcomes [7]. These results highlight why balancing visual support and cognitive efficiency is important for effective audiovisual design.

Finally, Flege's Speech Learning Model (SLM) provides an explanation for how multi-modal exposure influences category formation [1]. As repeated audiovisual cues can provide rich context and exposure to authentic L2 language environment, learners are able to rearrange L2 phonetic categories and less likely to be impacted by L1 knowledge. Hutchinson and Dmitrieva's further provided evidence through conducting a study on film's influence for vowel [12]. SLM interprets learners' reconstruction of perception, while multi-modal learning accounts for how learners integrate auditory, visual and textual information during L2 speech learning.

In summary, this framework demonstrates three dimensions to examine audiovisual pronunciation research, including mechanism, design, and exposure. It acknowledges that multi-modal facilitates perceptual encoding, cognitive load shapes how effectively learners integrate multiple inputs, and phonetic learning theories explain how these processes result in durable category restructuring. Together, these perspectives provide the conceptual foundation for analyzing how audiovisual input, when properly designed, can foster L2 pronunciation development.

3. Research Analysis

3.1. Audiovisual Input and Phonetic Perception

Research on audiovisual input has already provided evidence that the integration of auditory, visual, and textual cues can facilitate phonetic processing in second-language (L2) perception and production to different extent. Mitterer and McQueen provided one of the earliest demonstrations that subtitle has same language with the auditory can improve listeners' ability to encode their second language phonetic categories, helping them distinguish segmental contrasts more accurately [4]. Their

results indicated that, when watching videos especially with English captions, subtitles serve as lexical support which reflects auditory with text cue. Such lexically guided perceptual learning process help listener to locate target words accurately, improving their sentence comprehension capacity and phonetic perception. Subsequent studies extended the research to speech segmentation, exploring beyond the process of shape pronunciation categories.

Montero-Pérez and Pattemore further emphasized the influence of multi-modal alignment on speech segmentation [6]. They found that a single 72-minute short exposure to a captioned Spanish television episode improved learners' ability to identify word boundaries, regardless of the condition whether there is caption. As a result, they suggested that audiovisual input alone can foster segmentation. However, captions appeared to accelerate recognition for words already represented in the mental lexicon and the vocabulary size can also the result. This implies that captions assist the mapping of known lexical items rather than introducing new ones. However, Galimberti et al. examined the micro level influence of audio-synchronized textual enhancement on phonology and lexical processing [10]. They highlighted each target words shortly before it was spoken in the audio and learners in these conditions subsequently rejected mispronounced tokens more quickly, indicating sensitivity to both phonological and lexical accuracy. These findings collectively reveal that synchronization and temporal alignment are crucial for prompting learners to pay attention to phonetic form while processing meaning.

Hutchinson and Dmitrieva further illustrated perception development and production through naturalistic film exposure [12]. After viewing a French-language film, English-speaking learners improved discrimination and imitation of the French /y/ vowel, demonstrating that even brief audiovisual contact can influence articulatory mapping when sufficient multi-modal cues are available.

Together, these studies suggest that audiovisual materials promote perceptual restructure by reinforcing auditory representations with visual and audio material. However, as most experiments used short exposure and immediate post-tests, the durability of such perceptual gains remains uncertain. There are also other limitations. The scope of the target sounds also narrow, while many studies focus on single vowels or a few segmental contrasts. This led to further studies on broader phonetic categories. Meanwhile, the link between perception and production also remains insufficiently explored, while most studies focus on perceptual improvement, neglecting the articulacy improvement. In summary, existing evidence supports the notion that multi-modal alignment enhances learners' ability to decode and segment the speech, thereby forming a perceptual foundation upon which later pronunciation learning may build.

3.2. Caption Modes and Pronunciation Learning

Beyond perceptual mechanisms, researchers have examined how specific caption designs affect learners' attention, processing load, and eventually pronunciation outcomes. Wisniewska and Mora investigated different task focuses during an eight-week captioned-video program by manipulating instructions given to learners [5]. Learners who watched videos with captions while focusing on meaning showed significant gains in speech processing and pronunciation accuracy. However, learners did not show much improvement with the instruction that focused on phonetic form directly. The authors attributed the result to increased cognitive load, indicating that multi-tasking on both auditory and textual information while consciously monitoring pronunciation presumably overwhelmed working memory. This finding aligns with Sweller's Cognitive Load Theory, highlighting the importance of balancing form-oriented and meaning-oriented attention [11].

Some studies further addressed this problem of cognition load, exploring the influence of different caption types. Through comparing full caption and partial caption, Mohsen and Mahdi draw similar conclusions as Wisniewska and Mora [5,13]. Partial captions, which display only key lexical items, led to better pronunciation scores than full captions, suggesting that reduced visual text encourages greater reliance on auditory decoding. Galimberti et al. supported this idea by manipulating synchrony [10]. The synchronized highlight links the written form to the unfolding speech signal, guiding

learners' attention to the phonetic cues. Their audio-synchronized highlighting not only captured attention but also avoided the overload often caused by static full captions. Collectively, these studies indicate that caption benefits are conditional on how the text is presented in time and space relative to the audio signal.

Additional work has considered learners' proficiency as an important factor. Panna tier and Bétrancourt reported that foreign language proficiency can predict comprehension and transfer [7]. High-proficiency learners performed better than low-proficiency learners on all learning measures and experienced low cognitive load. Their results showed that subtitles regardless of whether it was native language or same language, had no significant effect on learning performance and did not compensate for low foreign language proficiency. Morton et al. examined another design parameter, examining subtitle duration in a computer-assisted learning environment [14]. They compared subtitles that disappear after characters' words with subtitles that last for longer time after characters' turn. Finally, they found that learners displayed no progress on both length and complexity of the spoken response, even though many learners thought they performed better with long duration caption.

Overall, caption design research suggests that the key to effective pronunciation learning lies in optimizing the interplay between information richness and cognitive economy. Captions that are temporally synchronized, selectively displayed, or progressively reduced appear to enhance focus on the auditory channel without overwhelming working memory. In contrast, static full captions or L1 subtitles may foster dependency and superficial comprehension. Therefore, future pedagogical design should emphasize adaptive caption modes that vary density and synchrony in response to learner proficiency, ensuring that attention remains directed toward the evolving phonetic form.

3.3. Extensive Viewing and Long-term Exposure

A growing body of work has investigated how the amount, variability, and continuity of audiovisual exposure influence the durability and transferability of pronunciation learning. Almusharraf et al. combined High Variability Phonetic Training (HVPT) with captioned video on the YouGlish platform [15]. Almusharraf et al. combined High-Variability Phonetic Training (HVPT) with captioned video using the YouGlish platform [15]. In their study, fifty-six Saudi learners are trained with English diphthongs under high variability and low variability, caption and non-captioned input. All groups showed improvement on pronunciation, but the low-variability no-caption condition facilitated production accuracy most, while high variety group got the lowest score. This implies that excessive variability and redundant visual input may hinder focused input and raise cognitive load for learners. The study confirmed that multi-modal platforms can sustain learners' motivation and provide effective exposure to authentic speech.

Similarly, Hutchinson and Dmitrieva demonstrated that even limited film exposure can affect production as well as perception, although effects were modest and short-term [12]. These findings raise the question that how long and how frequently learners should engage with audiovisual materials to achieve durable articulatory change.

Additionally, few studies focus on longitudinal experiment. Gesa and Miralpeix proposed the concept of extensive viewing [8]. They applied this concept through exposing learners with additional L2 input for several months, providing learners with captioned videos regularly. Although their year-long study focused vocabulary, its implications for pronunciation are clear. They pointed out that repeated audiovisual exposure can promote phonological and lexical knowledge, as well as facilitating unconscious learning. Meanwhile, through longitudinal research on extensive viewing, Pattemore et al. revealed that learners' perceived learning benefits and viewing preferences change with different language proficiency [16]. They suggest that students with high level language proficiency tend to shift from L1 subtitled to L2 captioned or uncaptioned modes more easily. This indicated that learners can shift to a natural progression toward viewing videos automatically and reducing their reliance on text. Cárdenas-Claros and Ramírez-Orellana put this shift into practice through the process of slowly removing the caption so that learners have to rely more on listening

[17]. The study found that progressive-reduction approach of video exposure benefit learners on not relying on subtitles and being more confident on their listening capacity.

These studies provide a map of audiovisual input study that learners start with controlled caption input and build their confidence. With the gradual reduce of textual support and the exposure becomes more authentic, learners become showing better comprehension, transfer, and finally improve their production, less being influenced by cognitive load. However, quantitative evidence on the ideal duration, frequency, or variability level remains scarce. Future research should therefore control how much audiovisual input learners receive and test whether large amounts of exposure lead to correspondingly greater improvements in measurable phonetic outcomes. Such work could inform adaptive learning systems that personalize video length, speaker diversity, and caption support to maximize long-term pronunciation improvement.

4. Discussion

The present review shows that audiovisual input supports L2 pronunciation development in specific ways with different and well-designed conditions. Therefore, the discussion must consider how the reviewed studies collectively refine our understanding of when and why such input is effective. Across the literature, one consistent finding is that captions matter less as a source of linguistic information and more as a tool for directing attention. For example, Mitterer and McQueen demonstrated that captions can guide listeners toward phonetic cues embedded in the speech stream, but the later studies in this review proposed more assumptions [4]. Wisniewska and Mora, for instance, showed that even when captions are present, learners only gain knowledge focusing on meaning rather than explicitly on phonetic form [5]. This suggested that learners should comprehend and gain knowledge naturally while watching video, rather than deliberately focus on phonetic segments. Similarly, Mohsen and Mahdi found that partial captions can better facilitate pronunciation outcomes than full captions, reinforcing the idea that reduced text led to less distraction and leaves more space for learners to recognize the sound [13].

Synchrony is also a key element that influence the learning efficiency. Galimberti et al. demonstrated that when text is highlighted with audio material spontaneously, learners can easily match what they see with what they hear, which makes them more sensitive to pronunciation details [10]. This reflects a conclusion that visual support is beneficial only when it combines attention and the unfolding signal, rather than distracting them from it. In contrast, study of Pannatier and Bétrancourt indicated that captions alone cannot support the insufficient language proficiency [7]. In their work, the transfer between comprehending and learning depended on learners' learning proficiency, and depended less on caption types. This means that audiovisual materials are most efficient when learners process basic language skills.

Caption duration and learner perception can also impact the effect of caption, presenting as the important role during the processing of viewing. Morton et al. found no overt effect of subtitle duration on spoken responses, even though learners themselves preferred longer on-screen text [14]. This difference between perceived and real benefits aligns with findings from Pattemore et al., which suggested that caption mode did not affect perceived learning but that learners' preferences extended exposure strongly let learners to shift away from L2 captions [16]. These patterns highlight the need to distinguish between cognitive effects and learner beliefs, as they do not always align.

Long-term exposure appears to be particularly important. Hutchinson and Dmitrieva demonstrated that even brief film exposure can facilitate production of difficult vowels, as well as Gesa and Miralpeix's longitudinal study further showed that months of captioned television support vocabulary development beyond classroom instruction [6, 8, 12]. Together, these findings suggest that audiovisual input functions can be benefit when learners encounter repeated and varied examples of the target language forms through period of time. Finally, recent work emphasizes progression rather than static design. Cárdenas-Claros and Ramírez-Orellana showed that gradually reducing caption density helps learners transition from reliance on text to reliance on listening [17]. It aligns with the

developmental shift observed in extensive viewing studies. There is a common point across the reviewed studies that audiovisual input supports pronunciation not through constant visual input but through designs that gradually shift learners' attention toward the acoustic signal. Overall, the review indicates that pronunciation benefits arise when captions are minimal, well-timed, and embedded within extended and increasingly authentic exposure. The challenge for future research is to specify the exact relationship between input quantity, attention, and measurable phonetic production

5. Conclusion

This review synthesized research on how audiovisual input contributes to L2 pronunciation development. It shows that the effects of captions and related design features depend on proficiency, attention, and exposure conditions. Reduced or selective subtitles include partial captions, keyword captions, and synchronized highlighting. These different types of subtitles can help captioning, segmental learning by reducing cognitive load and directing attention toward the acoustic signal. In contrast, full captions rarely promote pronunciation improvement and may divert learners' attention across other parts of the input. According to multiple studies, proficiency consistently predicted comprehension and confidence more strongly than caption type, suggesting that audiovisual learning is based on foundational linguistic skills. In addition, long-term exposure is another crucial factor. While short-term input show limited gains, extended viewing promotes vocabulary development, increases listening confidence. As a result, it allows learners to shift away from caption support gradually, leading them less dependent on subtitles while watching video. Research on extensive learning also shows that learners' preferences for subtitles shift over time. These preference shifts of viewing mode suggest that audiovisual input not only improves language learning, but also shapes learners' beliefs about how difficult materials are and how confident they are on authentic input. This aligns with studies showing that staged reduction of captions can facilitate the transition toward autonomous listening. Future research could control different amount of input, therefore tracking the production accuracy. Additionally, more longitudinal production studies could shift focus on examining how audiovisual input contributes to sustainable articulatory change over time. In the future, researchers could record and analyze caption timing, visual load, variability of talkers and learners' individual difference more precisely, combining different element. In conclusion, audiovisual input represents a promising but nuanced resource for L2 pronunciation development. Its effectiveness lies not in the presence of captions, but in how textual support is calibrated, reduced, and integrated within authentic and sustained listening experiences.

References

[1] Flege, J. E. Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), Speech perception and linguistic experience: Issues in cross-language research. Timonium, MD: York Press. 1995, 233 - 277.

[2] Derwing, T. M., & Munro, M. J. Second language accent and pronunciation teaching: A research-based approach. TESOL Quarterly, 2005, 39 (3), 379 – 397.

[3] Norris, J. M., & Ortega, L. Does type of instruction make a difference? Substantive findings from a meta-analytic review. Language Learning, 2005, 55 (S1), 157 – 213.

[4] Mitterer, H., & McQueen, J. M. Foreign subtitles help but native-language subtitles harm foreign speech perception. PLoS ONE, 2009, 4 (11), e7785.

[5] Wisniewska, N., & Mora, J. C. Can captioned video benefit second language pronunciation? Studies in Second Language Acquisition, 2020, 42, 599 - 624.

[6] Montero Pérez, M., & Pattemore, A. Effects of captioned video on L2 speech segmentation in intermediate learners of Spanish. Language Learning & Technology, 2025, 29 (3), 205 – 225.

[7] Pannatier, F., & Bétrancourt, M. Learning from academic video with subtitles: When foreign language proficiency matters. Learning and Instruction, 2024, 90, 101863.

[8] Gesa, F., & Miralpeix, I. Extensive viewing as additional input for foreign language vocabulary learning: A longitudinal study. Language Learning and Technology, 2023, 27 (3), 99 – 121.

[9] Paivio, A. Mental Representations: A Dual Coding Approach. Oxford University Press. 1986.

[10] Galimberti, V., Mora, J. C., & Gilabert, R. Audio-synchronized textual enhancement in foreign language pronunciation learning from videos. System, 2023, 116, 103078.

[11] Sweller, J. Cognitive load during problem solving: Effects on learning. Cognitive Science, 1988, 12 (2), 257 – 285.

[12] Hutchinson, A., & Dmitrieva, O. Exposure to speech via foreign film and its effects on non-native vowel production and perception. Journal of Phonetics, 2022, 95, 1129 – 1156.

[13] Mohsen, M. & Mahdi, H. Partial versus full captioning mode to improve L2 vocabulary acquisition in a mobile-assisted language learning setting: words pronuncio] ation domain. Journal of Computing in Higher Education, 2021, 33, 524 - 543.

[14] Morton, H., Gunson, N., & Jack, M. Attitudes to Subtitle Duration and the Effect on User Responses in Speech Interactive Foreign Language Learning. Journal of Multimedia, 2011, 6, 5.

[15] Almusharraf, A., Al-Furais, M., & Al-Ghamdi, N. Exploring the effects of modality and variability on EFL learners' pronunciation of English diphthongs: A student perspective on HVPT implementation. Humanities and Social Science Communications, 2024, 11, 141.

[16] Pattemore, A., Suárez, M., & Muñoz, C. Perceptions of learning from audiovisual input and changes in L2 viewing preferences: The roles of on-screen text and proficiency. System, 2024, 121, 103662.

[17] Cárdenas-Claros, M. S., & Ramírez-Orellana, P. Progressive reduction of captions in language learning: An exploratory study on learners' perceptions. Journal of Information Technology Education: Innovations in Practice, 2024, 23, 1 – 22.