


Article

Segmenting Speech: The Role of Resyllabification in Spanish Phonology

Iván Andreu Rascón 

Department of Spanish and Portuguese, Rutgers University, New Brunswick, NJ 08904, USA;
ivan.andreu@rutgers.edu

Abstract: Humans segment speech naturally based on the transitional probabilities between linguistic elements. For bilingual speakers navigating between a first (L1) and a second language (L2), L1 knowledge can influence their perception, leading to transfer effects based on phonological similarities or differences. Specifically, in Spanish, resyllabification occurs when consonants at the end of a syllable or word boundary are repositioned as the onset of the subsequent syllable. While the process can lead to ambiguities in perception, current academic discussions debate the duration of canonical and resyllabified productions. However, the role of bilingualism in the visual perception of syllable and word segmentation remains unknown to date. The present study explores the use of bilingual skills in the perception of articulatory movements and visual cues in speech perception, addressing the gap in the literature regarding the visibility of syllable pauses in lipreading. The participants in this study, 80 native Spanish speakers and 195 L2 learners, were subjected to audio, visual-only, and audiovisual conditions to assess their segmentation accuracy. The results indicated that both groups could segment speech effectively, with audiovisual cues providing the most significant benefit. Native speakers performed more consistently, while proficiency influenced L2 learners' accuracy. The results show that aural syllabic segmentation is acquired at early stages of proficiency, while visual syllabic segmentation is acquired at higher levels of proficiency.

Keywords: speech perception; speech segmentation; second language acquisition; bilingualism; Spanish phonology; visual perception; auditory–visual integration



Citation: Andreu Rascón, Iván. 2024. Segmenting Speech: The Role of Resyllabification in Spanish Phonology. *Languages* 9: 346. <https://doi.org/10.3390/languages9110346>

Academic Editor: John Lipski

Received: 12 June 2024

Revised: 2 October 2024

Accepted: 1 November 2024

Published: 7 November 2024



Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Segmenting aural speech can present a significant challenge, particularly when there are no clear pauses between words. Understanding how speakers segment continuous speech into meaningful units is fundamental for communication and language learning. The segmentation process heavily relies on calculating transitional probabilities between segments and syllables. High transitional probabilities suggest that elements belong to the same word, while low probabilities indicate word boundaries (Saffran et al. 1996). Understanding how this mechanism functions in both a first language (L1) and second language (L2) is essential for understanding the cognitive processes underlying language comprehension and production.

When learning an L2, individuals often draw upon their foundational L1 knowledge. The influence of their L1 might facilitate or impede L2 acquisition, depending on the linguistic similarities or differences between the two languages. Similar language structures can accelerate L2 acquisition, while divergent features can lead to negative transfers and learning challenges (Best and Tyler 2007; Pajak and Levy 2012). Bilinguals' segmentation strategies are influenced by the phonological attributes of their dominant language, with the degree of phonological alignment between the two languages determining the extent of these transfer effects and their impact on comprehension (Alammar 2015; Carroll 2004; Katayama 2015; Sanders and Neville 2003).

The present study tackles the role of L2 English learners' perceptual resyllabification in both aural and audiovisual contexts. Resyllabification is present in both English and Spanish, but it functions differently in each language. In Spanish, resyllabification generally avoids vowel-initial syllables by moving a consonant to the onset of the following syllable, making glottalization rare. For example, the phrase 'las obras' ('the works') is resyllabified as [la.so.bras], where the final /s/ of 'las' shifts to the onset of the following syllable, making it sound identical to 'las sobras' ('the leftovers'). Conversely, English speakers often insert a glottal stop [ʔ] or glottalize the word-initial vowel, with glottalization rates varying significantly among speakers (Dilley and Shattuck-Hufnagel 1995; Ladd and Schepman 2003; Scobbie and Pouplier 2010; Umeda 1978). This tendency makes English speakers more likely to emphasize consonant sounds in V_CV sequences, such as 'an ice' [ən.ʔaɪs] versus 'a nice' [ə.naɪs], due to their inclination to glottalize vowel-initial words. Because of this, English speakers are more likely to activate /s/-initial words in V_CV sequences due to their tendency to glottalize vowel-initial words. In contrast, Spanish speakers must account for both V_CV and VC_V segmentations, such as distinguishing between 'las obras' and 'las sobras'. The difference demands that English speakers learning Spanish control their natural tendency to glottalize and their bias toward consonants. Additionally, current research on Spanish resyllabification, especially with /s/ and /n/, has shown that these consonants have longer durations in canonical V_CV utterances (Lahoz-Bengoechea and Jiménez-Bravo 2021; Scarpace 2017; Strycharczuk and Kohlberger 2016). However, the perception of resyllabified utterances in Peninsular Spanish by L2 learners remains underexplored.

In addition, the important role of the articulatory movements of the lips, tongue, and jaw are well known to play a crucial role in speech perception (e.g., Bicevskis et al. 2016; Chandrasekaran et al. 2009; Hartcher-O'Brien et al. 2017; Macaluso et al. 2016; Peelle and Sommers 2015; Rosenblum and Dorsi 2021; Schwartz et al. 2012; Turk 2014). According to Liberman's Motor Theory of Speech Perception, listeners discern speech by linking acoustic signals to articulatory motor commands, which aid in organizing speech sounds into phonetic categories despite variations among speakers (A. M. Liberman and Mattingly 1985). The McGurk effect further proves the multisensory nature of speech perception, where visual cues influence auditory perception (McGurk and MacDonald 1976).

To date, it remains unknown whether the visibility of syllable pauses in lipreading, as well as the timing of the articulatory sequence, can benefit L2 learners in their speech perception. This question has not been directly addressed in the literature, leaving it an open area of investigation.

1.1. Speech Segmentation

Speakers segment speech in spoken word recognition by calculating the transitional probabilities between units like segments and syllables. If one element strongly predicts the next, it is seen as part of the same word. Conversely, a low probability indicates a word boundary (Saffran et al. 1996).

When acquiring an L1, infants learn to recognize and differentiate between the distinct acoustic patterns of their native language, progressively enhancing their ability to understand speech (Maye et al. 2002; Werker et al. 2007).

The overall body of research argues that when learning an L2, individuals often draw upon their foundational L1 knowledge. The revised Speech Learning Model (Flege and Bohn 2021) also emphasizes the role of language experience. This model suggests that L2 learners map L2 sounds onto their existing L1 categories, and the ease or difficulty of acquiring new L2 sounds depends on the similarity between the L1 and L2's phonetic categories. Additionally, the SLMr posits that language learning is a dynamic process where learners continuously adjust their phonetic categories based on L2 experience, leading to gradual improvements in perceiving and producing L2 sounds.

The reliance on L1 knowledge extends beyond phonetic categorization. Studies using constructed and artificial languages have shown that listeners rely on the segmentation strategies of their L1 to segment and cue-weight an unfamiliar language into individual

words (Finn and Hudson Kam 2008; Pajak and Levy 2012; Tremblay et al. 2016, 2018). Previous research has identified an effect of prosodic learning interference on segmentation (Tremblay et al. 2016), as well as the functional weight of prosodic cues (Tremblay et al. 2018), with these findings also extending into L3 acquisition (Tremblay et al. 2021). The degree of phonological alignment between two languages determines the extent of these transfer effects and their impact on comprehension (Alammar 2015; Carroll 2004; Cutler 2012; Katayama 2015; Sanders and Neville 2003).

Ortega-Llebaria et al. (2019) studied the cue-driven window length, which refers to the time frame during which listeners can process prosodic cues like pitch, stress, and duration. They compared English and Spanish speakers. Sentences were manipulated so that target words contained only duration cues or only vowel quality cues. The findings showed that English speakers demonstrated greater flexibility in using prosodic expectations compared to Spanish speakers. They were able to adjust their processing strategies depending on whether the cue was based on vowel quality or duration.

Taking the case of English and Spanish resyllabification, the strategies across these languages differ significantly. In Spanish, resyllabification often occurs to avoid vowel-initial syllables by moving a consonant to the onset of the following syllable, as seen in *'busca socio'* ('seeks a partner') and *'buscas ocio'* ('you seek leisure'), both of which can be pronounced [bus.ka.so.sjo] after resyllabification. In contrast, English speakers often insert a glottal stop or glottalize the word-initial vowel, with studies showing glottalization in the VC_V context ranging from 25% to 100% depending on the speaker (Umeda 1978).

Additionally, Spanish is generally considered a syllable-timed language (Pike 1945), possessing a preponderance of open syllables and lacking vowel reduction. When emphasizing strong word-initial syllables, English relies on both suprasegmental cues like duration, pitch, and intensity (I. Y. Liberman 1973; Nakatani and Aston 1978) and segmental markers such as vowel quality (Campbell and Beckman 1997). In contrast, Spanish primarily uses suprasegmental cues (Hualde 2009).

Lexical access and speech segmentation are impacted by these variations. When processing VCV sequences, English speakers are more likely to activate /s/ initial words than, for example, /a/ initial words because of their glottalization tendency with vowel-initial words. In contrast, Spanish speakers must take into account both V_CV and VC_V segmentations. English speakers therefore need to control their natural tendency to glottalize and their bias toward consonants when learning Spanish.

Current academic discussions on the role of Spanish resyllabification with /s/ and /n/ point towards a longer duration of the consonant in canonical utterances V_CV (Lahoz-Bengochea and Jiménez-Bravo 2021; Scarpace 2017; Strycharczuk and Kohlberger 2016). The role of resyllabified utterances in the perception of Peninsular Spanish by L2 learners is understudied, as is the role of visual perception and whether visual access to mouth articulators can help speakers establish word boundaries.

1.2. Spanish Resyllabification

The phenomenon of resyllabification refers to the phonological process wherein consonants at the end of one syllable or word boundary can be repositioned to the onset of the subsequent syllable, especially in fluent speech or when a word ending in a consonant is followed by a word beginning with a vowel. This phenomenon is prominently observed across word and morpheme boundaries in Spanish (James Wesley Harris 1983; Hualde 2005).

In the context of Spanish phonology, it has been conventionally suggested that consonants in the coda position at the end of a word undergo resyllabification when followed by a word beginning with a vowel; a process known as 'coda capture'. For instance, the phonemic representation /los#otros/ undergoes such a transformation to yield the phonetic realization [lo.so.tros] (Colina 1995, 1997, 2006; Crowhurst 1992; Robinson 2012; Torreira and Ernestus 2012).

In Spanish phonology, resyllabification is not universal and can vary both dialectally and phonologically. For instance, [Robinson \(2012\)](#) asserts that in Highland Ecuadorian Spanish, word-final [z] and [ɲ] avoid prevocalic resyllabification, aligning with [Lipski's \(1989\)](#) contention that resyllabification is not universally equal and that differences exist. James W. [Harris and Kaisse \(1999\)](#) discuss how Buenos Aires Spanish restricts [h]-realizations of /s/ to canonical codas, such as [dih.ko] ('disco'). On the one hand, resyllabification can prevent aspiration: /dos#alas/ becomes [do.sa.las]. On the other hand, some Caribbean dialects allow aspiration in phrasal contexts, resulting in [do.ha.las]. Another pattern in these dialects shows aspiration both within words and in phrasal contexts, such as [de.hi.ɣʌl]. [Bermúdez-Otero and Payne \(2011\)](#) address /s/-voicing in Highland Ecuadorian Spanish, where word-final /s/ voices to [z] before vowel-initial words, unlike word-medial /s/, which remains unvoiced: [ga.za.kre] ('acid gas') versus [ga.sa] ('gauze'). The presented variabilities are examples that illustrate the complex interplay of prosodic structure and phonological processes across Spanish dialects.

However, for the focus of this study, in Castilian Spanish phonology, the traditional perspective posits that derived onsets (resyllabified codas) are phonetically indistinguishable from canonical onsets, as described by [Hualde \(2005\)](#). More recent studies have argued against this view. [Hualde and Prieto \(2015\)](#) investigated the intervocalic alveolar fricatives /s̺/ and /z/ in Catalan and Spanish, examining how derived onsets differ from canonical onsets in relation to lenition. Their study finds that in Spanish, derived onsets are more lenited compared to canonical onsets. Specifically, intervocalic /s/ in derived onset positions (resyllabified from a coda position) exhibits more voicing and a shorter duration. Derived onsets at word boundaries (e.g., /VC_V/ sequences) are particularly susceptible to lenition, possibly due to their resyllabified nature, which means that they might not have the same robust gestural coordination as canonical onsets, leading to greater lenition and therefore a shorter duration.

[Strycharczuk and Kohlberger \(2016\)](#) further validated the acoustic properties of word-final /s/ in Peninsular Spanish. Acoustic data were collected from 11 speakers of Peninsular Spanish, showing that word-final pre-vocalic /s/ has a longer duration than coda /s/ but is shorter than word-initial or word-medial onset /s/. The study also suggests that partial resyllabification might occur, where word-final pre-vocalic /s/ is neither a true onset nor a true coda.

Expanding upon the findings of preceding studies, [Lahoz-Bengoechea and Jiménez-Bravo \(2021\)](#) studied how consonant duration can serve as a perceptual cue to identify the original lexical affiliation of resyllabified segments. The study employed 12 minimal pairs with their durations manipulated in 5 equidistant steps. A total of 65 native Spanish speakers were asked to recognize and perceive word boundaries in continuous speech while only relying on phonological properties. The results showed that /s/ and /n/ with a shorter duration biased participants towards perceiving the consonant as part of the preceding word. These results argue that consonant duration plays a major role in recognizing resyllabification contexts.

In understanding how Spanish L2 speakers' segment, Spanish resyllabification utterances have only been investigated by [Scarpace \(2017\)](#). The study employed speakers from various non-/s/-aspirating dialects of Spanish within Colombian and Mexican Spanish. In the perception tasks, target words were embedded in phrases like "Escribe(n)(s) X" to create minimal pairs differentiated by the word-affiliation of the target consonant.

Surprisingly, beginning L2 learners had a higher rate of perceiving consonants as word-initial (nearly 65% of the time) compared to advanced learners and native speakers, even without the presence of a glottal stop. There was a negative correlation between proficiency and the tendency to perceive consonants as word-initial (Pearson's $r = -0.524$, $p = 0.02$). Native speakers identified words as consonant-initial 55% of the time, while advanced learners performed similarly to native speakers, identifying words as consonant-initial 53% of the time. There was an asymmetry between /n/ and /s/: native speakers identified /n/ as word-initial 48% of the time and /s/ as word-initial 63% of the time.

Beginning L2 learners showed similar rates of perceiving both /n/ and /s/ as word-initial, whereas advanced learners and native speakers had a higher accuracy for /s/ than /n/. Native speakers used durational cues effectively to identify /n/ as word-initial but less so for /s/. Advanced learners approximated the native-like use of durational cues, especially for /n/, while beginning learners showed a medium correlation between duration and the perception of /s/ as word-initial. Overall, this suggests that speakers could use durational cues to determine the lexical affiliation of the pivotal consonant.

The present study builds upon Scarpace's research on resyllabification, focusing on sentences that feature resyllabification in Central Peninsular Spanish. Additionally, it investigates whether both L2 learners and native speakers can determine lexical segmentation duration when provided with additional information, such as visual access to mouth articulators, for both canonical and resyllabified utterances.

1.3. Visual Perception and Word Recognition in L2

The articulatory saliency of a speaker, such as the motion of their lips, tongue, and jaw, plays a crucial role in speech perception. A. M. Liberman and Mattingly's (1985) Motor Theory of Speech Perception suggests that listeners discern speech through acoustic signals by linking them to articulatory motor commands, which aid in organizing speech sounds into phonetic categories despite the 'lack of invariance' problem caused by speaker variations and coarticulation. The McGurk effect (McGurk and MacDonald 1976) further illustrates the multisensory nature of speech perception, where visual cues influence auditory perception, highlighting the interconnectedness of auditory and visual signals in speech.

The overall body of research on the subject underscores the significant prevalence of lipreading among listeners, highlighting its role in enhancing speech comprehension, particularly in acoustically challenging or ambiguous environments (e.g., Ross et al. 2007). Lewkowicz and Hansen-Tift (2012) further indicate that during demanding speech tasks, adults predominantly focus on a speaker's mouth rather than solely relying on auditory input. Complementing this, Peelle and Sommers's (2015) study argues that visual articulatory cues not only augment speech comprehension but also are able to reduce the cognitive load on listeners.

Multisensory processing integrates or segregates sensory stimuli based on their attributes and cognitive resources, driven by bottom-up sensory mechanisms and top-down cognitive functions. It is known that this dual process aids in the detection, discrimination, and recognition of speech (Hartcher-O'Brien et al. 2017; Macaluso et al. 2016). Speech comprehension involves merging congruent information across multiple sensory channels, resulting in the "fusion" of these data streams (Bicevskis et al. 2016; Chandrasekaran et al. 2009; Rosenblum and Dorsi 2021; Turk 2014).

Schwartz et al. (2012) proposed the Perception for Action Control Theory, which states that speech perception operates as a multisensory process. According to this theory, listening to language involves perceiving distinct gestures known as perceptuo-motor units, combining the cohesive nature of gestural actions with the perceptual significance of auditory and visual patterns. Schwartz and colleagues demonstrated that synchronizing silent articulation with matching ambiguous auditory and visual speech cues aids in identifying these cues.

Recent research into bilingual language experiences and audiovisual processing has highlighted the complexity and flexibility of language processing systems. Desroches et al. (2022) used a picture/spoken-word matching paradigm to investigate how bilinguals and monolinguals process language. Their study found that bilinguals can activate lexical options from both languages when identifying pictures, even when they expect input only in their dominant language. The authors suggest that bilinguals have cross-language connections between lexical representations and phonology, resulting in a more complex and flexible language processing system compared to monolinguals. Additionally, L2 learners undergo developmental stages similar to L1 listeners in integrating lipreading

with auditory speech. Proficiency levels influence this integration, with higher proficiency leading to better audiovisual integration and improved comprehension (Brancazio and Miller 2005; Hazan et al. 2006). However, L2 learners face challenges in perceiving L2 sounds due to differences in their L1 and L2 segmental inventories, which can affect their pronunciation and comprehension (Flege and Bohn 2021; Levy and Strange 2008; Polka 1995).

Previous research indicates that their L1 phoneme inventory impacts how L2 learners process visual input. For instance, Pegg et al. (1992) found that French listeners often confused the English interdental fricative /ð/ with /d/ or /t/, suggesting that L2 listeners adjust their perceptions based on their L1 phonology. Similarly, Hazan et al. (2002) demonstrated that Spanish L2 learners of English perceived visually salient sound contrasts more accurately when visual information from lipreading was available. The facilitation of L2 sound perception through AV input shows that some sound contrasts are acquired faster than others, depending on their frequency and visual salience (Best and Tyler 2007; Leonte et al. 2018).

Deepening our knowledge of duration and visual cues, while it is still unknown whether duration differences in Spanish resyllabification can be accessed through visual articulators, some studies have investigated the visual access to durational cues in both vowels (Lidestam 2009) and consonants (Scott and Idrissi 2018).

Lidestam (2009) investigated the phonetic perception of vowel duration in visual-only, auditory, and audiovisual modalities in Swedish, using 24 durational steps. The results indicated that visual-only presentations yielded the highest error rates. Auditory and audiovisual modalities demonstrated comparable performance levels, indicating an auditory dominance in bimodal perception. Nonetheless, participants were able to discern even the smallest duration of all the steps provided by having access to visual moving facial articulators. Scott and Idrissi (2018) studied whether participants could distinguish between short and long consonants, and between plain and pharyngealized consonants, using visual cues alone. Both consonant length and pharyngealization were shown to be influenced by visual information in an audiovisual context, though the auditory modality remains dominant for timing information. Visual information about pharyngealization did influence perception, but the effect was modest. Participants were more likely to perceive a consonant as pharyngealized if the video showed such cues.

Finally, the visibility of syllable pauses in lipreading as well as the timing of the articulatory sequence has not been directly addressed in the literature, and it remains unknown whether L2 learners could benefit from these cues in speech perception.

1.4. The Present Study

The present study aims to further understand word boundary segmentation abilities and category formation during the acquisition of Spanish by native English speakers. Specifically, it focuses on the case of Castilian Spanish resyllabification and its processing in relation to language proficiency across audio-only, visual-only, and audiovisual modalities. To achieve this, this study will first confirm that the speakers selected for the study can produce the expected durational differences in coda and onset positions for /n/ and /s/, as described in previous studies (Lahoz-Bengoechea and Jiménez-Bravo 2021; Scarpace 2017; Strycharczuk and Kohlberger 2016). Following this, the study will test and compare the perception and segmentation decisions of native speakers with those of L2 learners, examining how language proficiency influences their processing abilities. The present study is driven by the following research questions:

RQ1 If there are any durational differences, can L2 learners accurately detect temporal distinctions in V_CV and VC_V resyllabification utterances? Do L2 learners tend to perceive resyllabified consonants as word-initial due to the influence of English when learning Spanish?

While uncertainty remains about whether resyllabification is absolute in Castilian Spanish (Hualde 2005), recent research indicates that durational differences are anticipated.

Specifically, a pre-vocalic /s/ is expected to have a longer duration than a coda /s/ (Lahoz-Bengoechea and Jiménez-Bravo 2021; Scarpace 2017; Strycharczuk and Kohlberger 2016).

Moreover, bilinguals' segmentation strategies are expected to be influenced by the phonological attributes of their dominant language (Cutler 2012) and therefore they may experience a transfer of segmentation abilities from their L1 to their L2 (Finn and Hudson Kam 2008; Pajak and Levy 2012; Tremblay et al. 2016, 2018).

English speakers typically introduce a glottal stop in word-initial vowels, leading to the expectation that they might apply this pattern when perceiving Spanish. However, Scarpace (2017) found that beginning L2 learners identified words as consonant-initial more often than chance would predict. It remains unknown whether this phenomenon applies to Castilian Spanish utterances and how visual cues affect a broader sample of participants.

RQ2 How does a learner's proficiency in their L2 interact with their perception of resyllabification utterances?

Research on bilingual segmentation strategies in the context of resyllabification is limited. Bilinguals' segmentation strategies are anticipated to be influenced by the phonological attributes of their dominant language. Lower-proficiency learners with less exposure to their L2 are expected to be more influenced by English segmentation rules (Cutler 2012; Pajak and Levy 2012).

Beginning learners, as native English speakers, are unfamiliar with the resyllabification processes in Spanish and are therefore likely to parse words as vowel initial. Contrary to this expectation, Scarpace (2017) found that beginning learners showed some ability to use durational cues to identify word boundaries even in the absence of glottalization. Specifically, there was a negative correlation between proficiency and the tendency to perceive consonants as word-initial.

It remains under discussion whether the durational differences in CV_C and C_VC sequences apply to Castilian Spanish speech, and whether low-proficiency learners develop specific parsing abilities that differentiate them from higher-proficiency learners and native speakers. Additionally, a broader and larger pool of learners must be examined to confirm these findings.

RQ3 If durational differences are present, how do visual articulatory cues influence the ability and proficiency of L2 learners in distinguishing between canonical and resyllabified utterances?

Multisensory processing, which uses both bottom-up sensory mechanisms and top-down cognitive functions, has been shown to facilitate speech detection, discrimination, and recognition (Hartcher-O'Brien et al. 2017; Macaluso et al. 2016). Speech comprehension can merge congruent information across sensory channels, resulting in the fusion of these data streams (Bicevskis et al. 2016; Chandrasekaran et al. 2009; Rosenblum and Dorsi 2021; Turk 2014). Therefore, higher-proficiency learners, speakers with more language experience, are expected to benefit more from visual cues due to their stronger phonetic and phonological categories, allowing the integration of auditory and visual information (Brancazio and Miller 2005; Hazan et al. 2006).

While it remains unknown whether duration differences in Spanish resyllabification can be visually perceived, previous studies have confirmed that speakers can visually perceive and access durational cues in both vowels (Lidestam 2009) and consonants (Scott and Idrissi 2018). Lidestam (2009) found that visual-only presentations had high error rates, but auditory and audiovisual modalities performed similarly.

Based on previous research, speakers are expected to benefit from having access to visual cues. However, it is also possible that the visual category may interfere with possible early segmentation strategies between both languages, potentially causing a confounding effect. In the auditory-only condition, English speakers might again be biased by the absence of the glottal stop before a word-initial vowel.

2. Materials and Methods

2.1. Participants

Participants in this study were divided into two groups: a control group of native Spanish speakers (N = 80) and a group of L2 learners (N = 165) ranging from very early learners to high-proficiency learners. Proficiency in this study is considered a continuous variable. Participants were recruited through the Qualtrics platform from multiple universities in the United States. The control group consisted of native speakers who identified themselves as not comfortable with English at all, had less than six years of English instruction (mostly in high school), had not spent any time in an English-speaking country, had no family members who spoke English, and reported 0% usage of English in their daily activities. These native speakers did not speak any other languages fluently. For the English speakers, participants who were proficient in other languages, heritage speakers of any language, or those who had spent time with family members who spoke another language were excluded from the study.

2.2. Tasks and Procedures

The study was conducted entirely on Qualtrics. After consenting to participate, the participants filled out the Bilingual Language Profile (Birdsong et al. 2012) specific to their Spanish or English usage. The questionnaire inquired about every language the participants knew, their age when they learned each language, and how frequently they used these languages in different settings.

After the questionnaire, L2 learners were administered the Lexical Test for Advanced Learners of Spanish (LexTALE-ESP) (Izura et al. 2014), a common lexical decision task, to assess their Spanish proficiency. In this task, participants identified whether displayed words were real or fake by pressing designated keys ('1' for real, '0' for non-real). LexTALE scores span from 0 to 100. Monolingual Spanish speakers often score above 80. For this study, English-speaking participants with little or no Spanish proficiency typically had scores below 50. The study dataset, shown in Table 1, displays the LexTALE scores of the participants. The study encompassed a total of 195 participants. 165 successfully completed all designated tasks, with LexTALE proficiency scores that spanned a range of 38.22 to 96.67.

Table 1. LexTALE descriptives.

N	Avg.	SD	Range
165	61.74	19.06	[38.22–96.67]

Participants were then exposed to the three aforementioned conditions (audio, audiovisual, and audio-only) in a two-alternative forced-choice (2AFC) task. Their task was to choose the input they thought they were being exposed to for both V_CV and VC_V sequences involving instances of /n/ and /s/. The stimuli recordings were conducted in a soundproof booth, with no background noise. The employed microphone captured frequencies from 20 Hz to 20 kHz and maintained a 24/48 bit depth/sample rate for high-quality audio recordings with detailed sound and low noise. Additionally, the study's camera featured a 50-megapixel sensor, measuring 1/1.56 inches, with 1.0 micrometer pixels and an f/1.8 lens. The camera was positioned roughly forty-five centimeters from the subjects.

The video images were edited to consistently display an area from the septal cartilage to the mentalis muscle of the mandible. To compensate for involuntary participant movements, edits ensured the visibility of at least 5 cm to the left and right of the risorius muscle. The goal of this editing was twofold: to provide participants with a clear view of key visual articulators and to maintain uniformity across the video dataset while preventing participants from focusing on the speakers' eyes, physical characteristics, or other unrelated visual cues. The only input they had access to was the mouth's articulators.

The stimuli items included inputs from native Spanish speakers within carrier sentences. Four native speakers from Madrid, who did not aspirate /s/ or elide final /n/ and who were unaware of the study’s goals, were asked to naturally read 26 sentences containing canonical and resyllabified utterances with /n/ and /s/ (see Supplementary Materials). For example, participants read sentences such as *ve naves espaciales* (“sees spaceships”) versus *ven aves migratorias* (“they see migratory birds”). Other pairs included phrases like *la salas* (“you salt it”) and *las alas* (“the wings”), as in *La salas demasiado y la comida queda muy salada* (“You salt it too much, and the food becomes too salty”) versus *Las alas del avión se extendían majestuosamente en el cielo* (“The wings of the airplane spread majestically in the sky”). The input was later trimmed to *ven aves* and *ve naves* or *las alas* and *la salas*, creating three conditions: audio-only, audiovisual, and visual-only.

The target items were embedded at the beginning of sentences to create natural reading inputs. The sentences were presented in a random order to prevent readers from noticing the minimal pairs and the triggering of potential disambiguation through pauses or stress on either the vowel or the consonant that could undergo resyllabification. Special attention was given to identifying and avoiding unnatural pauses, ensuring consistent lexical stress in both V_CV and VC_V conditions while maintaining similar prosodic structures. The recordings were then normalized for intensity to ensure uniform loudness across samples. Before presenting the stimuli to L2 speakers, measurements of duration were taken using Praat (Boersma and Weenink 2018).

Differences in duration were found, as displayed in Figure 1, which thus argues against the possibility of finding ambisyllabicity and aligning with previous research on the segment duration of resyllabification (Lahoz-Bengochea and Jiménez-Bravo 2021; Scarpace 2017; Strycharczuk and Kohlberger 2016).

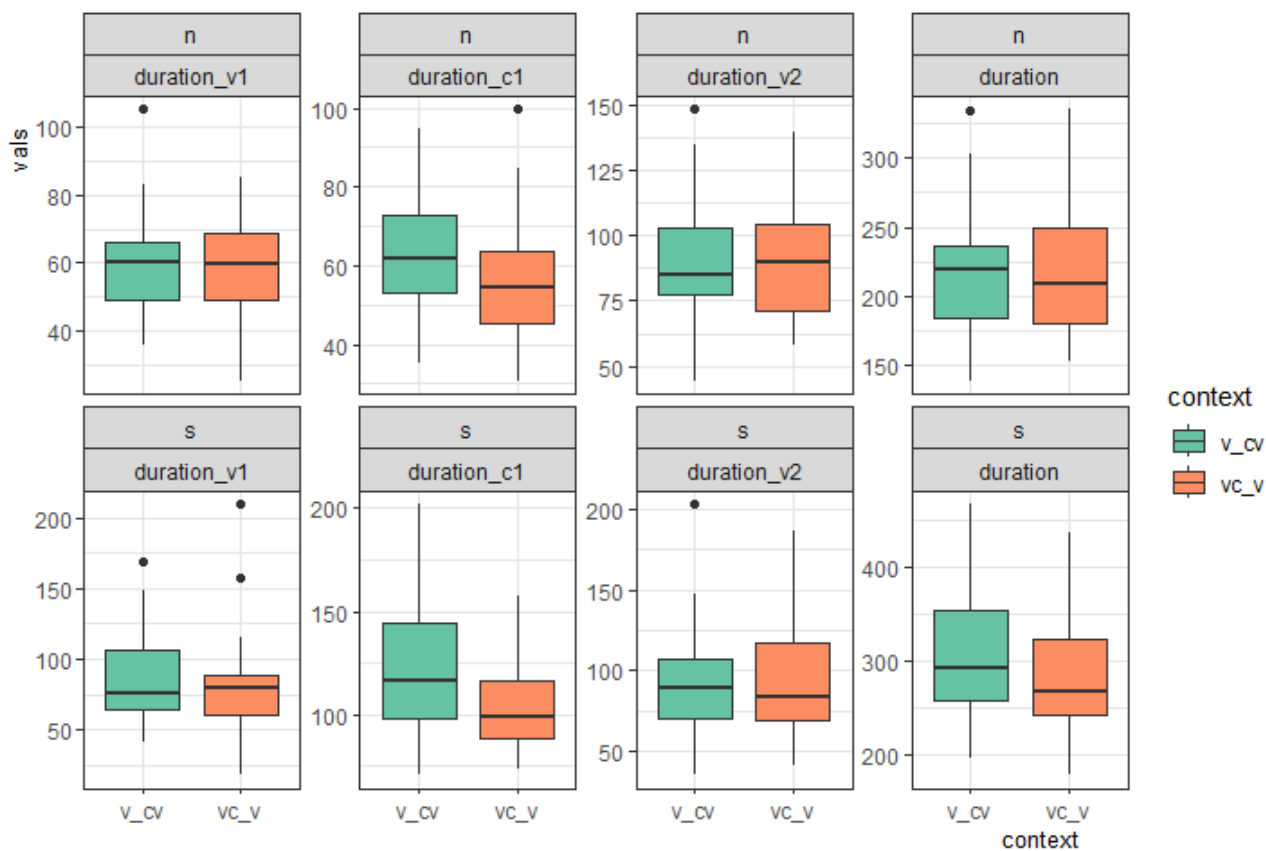


Figure 1. Durational differences (in ms) of initial vowel, target consonant, subsequent vowel, and total V_CV and VC_V sequences.

2.3. Data Analysis

The present study utilized a binomial logistic regression to evaluate the likelihood of participants accurately identifying the correct syllable combination and its association with individual language proficiency. The model incorporated parameters including consonant status (canonical or resyllabified), consonant type (/n/ or /s/), LexTALE score, and their interactions. The same model was employed for the control group of native speakers, with the proficiency LexTALE variable removed from the model formula. These parameters were assessed across three conditions: audio, visual, and audiovisual (A/V). The likelihood of the model followed a Bernoulli distribution with a logit link function. The criterion response was coded as “1” for correct responses and “0” for incorrect responses. Intercept provides insight into the log-odds of accurately distinguishing the consonant pairing, standardized around the mean values of the predictor variables. The analysis employed a categorical likelihood with a logit link function, accounting for individual variability by introducing varying intercepts for participants and items. Additionally, the model permitted varying slopes for the interplay between consonant status and testing conditions and integrated regularizing, informative priors.

Metrics for interpreting the model’s parameters included the Region of Practical Equivalence (ROPE), the Maximum Posterior Estimate (MPE), Rhat, and the Effective Sample Size (ESS). When 95% of the Highest Density Interval (HDI) for a parameter, β , lies outside the ROPE and the MPE approaches 1, it is interpreted as substantial evidence of a specific effect. All analyses were performed using R, fitting the models with the probabilistic programming language Stan through the R package brms (Bürkner 2017).

3. Results

An initial statistical analysis was performed for the mean (M) and standard deviation (SD) for each experimental condition, as summarized Table 2. Both L2 speakers and native speakers demonstrated their ability to parse and segment both canonical and resyllabified boundaries, performing very similarly across the three conditions. Native speakers performed slightly better in the audio and audiovisual conditions, while L2 learners performed better when only visual access to mouth articulators was provided. The audiovisual condition was where all participants performed best, whereas the visual-only condition proved to be the most challenging. Specifically, in the audiovisual condition, for L2 speakers, the mean probability of correct responses was 0.611 (SD = 0.488), while native speakers had a mean of 0.626 (SD = 0.270). In the audio-only condition, L2 speakers had a mean probability of giving the correct response of 0.601 (SD = 0.490), whereas native speakers had a mean of 0.615 (SD = 0.270). Finally, in the visual-only condition, L2 speakers had a mean probability of giving the correct response of 0.545 (SD = 0.498), and native speakers had a mean of 0.513 (SD = 0.291). Both groups exhibited lower averages and higher variability in this condition. Native speakers consistently performed better and more uniformly, with less spread.

Table 2. Accuracy in perception of Spanish resyllabification of native English and Spanish speakers: summary of multi-response variables.

Condition	L2 Group (M ± SD)	NS Group (M ± SD)
Audio-Only	0.601 ± 0.490	0.615 ± 0.270
Audio-visual	0.611 ± 0.488	0.626 ± 0.270
Visual-Only	0.545 ± 0.498	0.513 ± 0.291

The results of the logistic regression model are presented in Table A1, including the ROPE, MPE, Rhat, ESS, HDI, and the intercept. Figure 2 shows the predicted probability distribution of a correct response for both canonical and resyllabified utterances. The leftmost box displays the overall probability, the middle one compares canonical and resyllabified utterances, and the rightmost one shows performance relative to the type of

consonant (/n/ or /s/). Horizontal lines with circles indicate the mean and 95% credible intervals of the predicted values for L2 learners.

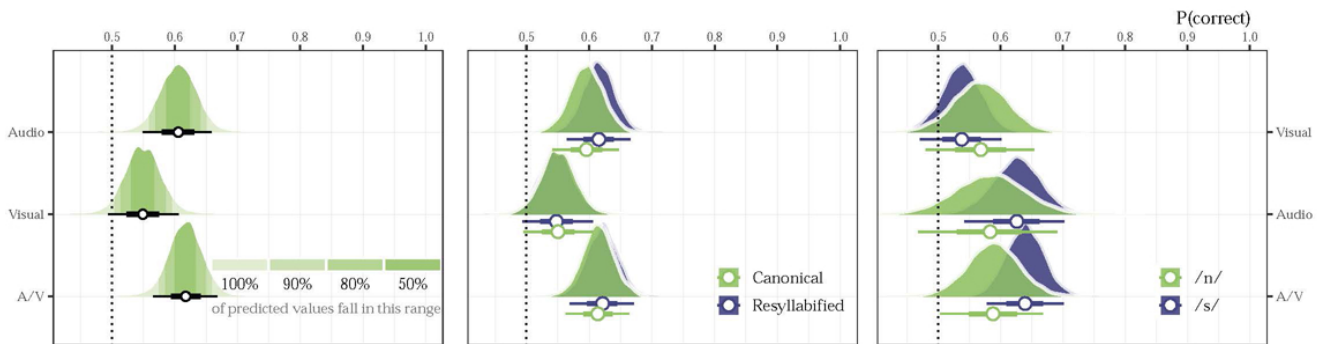


Figure 2. Probability of a correct response for canonical vs. resyllabified utterances by Spanish L2 learners. Values are standardized and points represent posterior medians along with the 95% HDI.

Similarly, Figure 3 shows the same distributions for native speakers. The probability of a correct response for canonical versus resyllabified utterances by native Spanish speakers is displayed in the left-most box. Values are standardized, and the points represent posterior medians along with the 95% HDI. Both groups exhibit very similar performance and response patterns, with almost no difference between the types of syllable utterances and a tendency to be more accurate with /s/ items than with /n/. Additionally, the 95% credible intervals of the predicted values are shorter for native speakers, indicating that this group is more consistent in their responses compared to L2 learners.

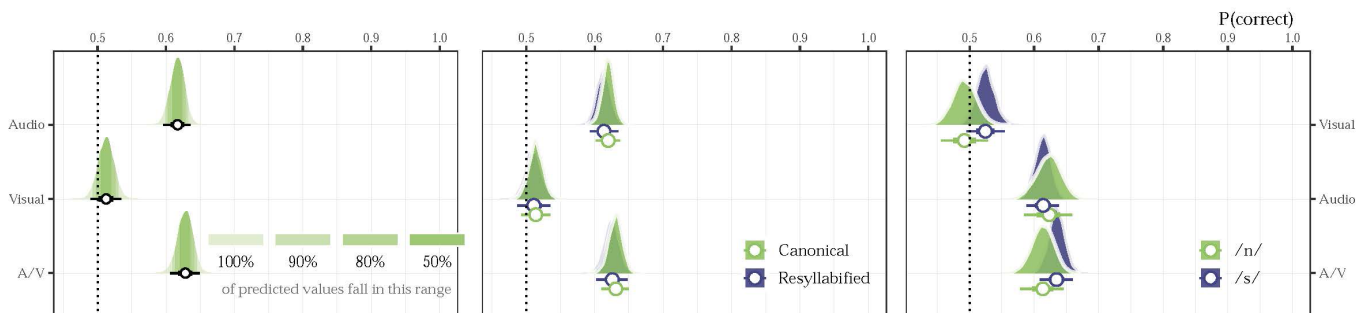


Figure 3. Probability of a correct response for canonical versus resyllabified utterances by native Spanish speakers. Values are standardized and points represent posterior medians along with the 95% HDI.

3.1. Audio Condition

In the audio-only context, both L2 speakers and native speakers showed positive intercept estimates. For L2 speakers, the intercept has a point estimate of 0.42 with a 95% HDI of [0.22, 0.63], indicating a positive baseline log-odds of correctly identifying syllable combinations ($\beta = 0.42$, HDI = [0.22, 0.63], ROPE = 0, MPE = 1). This suggests that L2 learners have a strong likelihood of correct responses in this context. Native speakers also show a positive intercept ($\beta = 0.11$, HDI = [0.05, 0.17], ROPE = 10.21, MPE = 0.53), although their baseline performance is slightly lower than that of the L2 group.

The consonant status, whether canonical or resyllabified, did not present a significant effect for L2 learners. However, the overall impact of consonants on these L2 speakers showed a slightly negative effect ($\beta = -0.09$, HDI = [-0.28, 0.11], ROPE = 53.03, MPE = 0.82). Conversely, both consonant status (V_CV vs. VC_V) and consonant type (/n/ vs. /s/) had slightly positive effects on native speakers. Specifically, for consonant status, the estimate was ($\beta = 0.10$, HDI = [0.04, 0.16], ROPE = 9.13, MPE = 0.53), and for the type of consonant, the estimate was ($\beta = 0.09$, HDI = [0.03, 0.15], ROPE = 9.71, MPE = 0.51). Interestingly, proficiency did have a small negative effect ($\beta = -0.05$, HDI = [-0.14, 0.05], ROPE = 88.47,

MPE = 0.82). As depicted in Figure 4, which shows the probability of a correct response as a function of increasing proficiency, there was a slight reduction in consonant accuracy. In both V_CV and VC_V patterns, lower-proficiency participants showed greater accuracy with /s/ than with /n/. However, as proficiency increased, their accuracy in selecting /s/ marginally decreased.

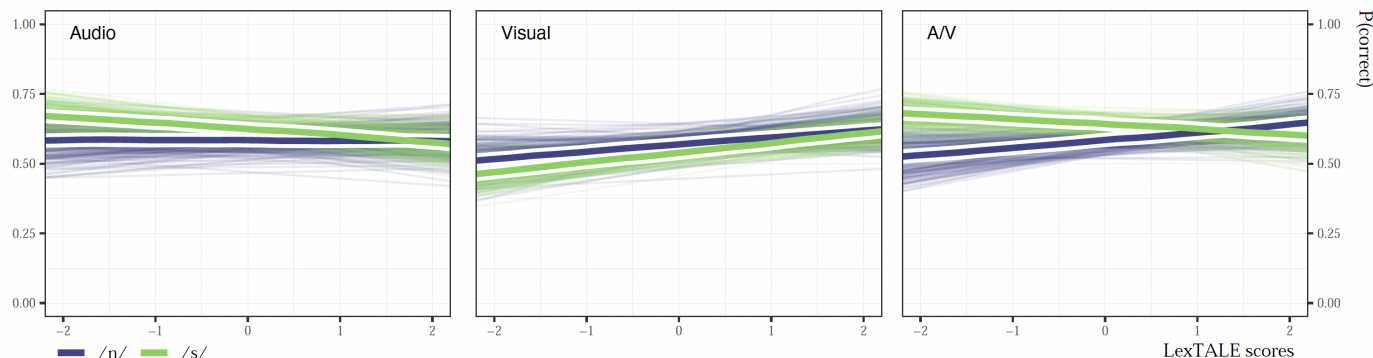


Figure 4. The probability of a correct response as a function of LexTALE. Thick solid lines indicate the median lines of best fit from the posterior. Thinner, transparent lines represent 300 draws from the posterior distribution.

3.2. Audiovisual

The results from the audiovisual condition followed a very similar pattern as the audio-only condition; speakers showed a slight increase in the probability of selecting the correct response. For the L2 learners, the intercept estimate was ($\beta = 0.46$, HDI = [0.3, 0.63], ROPE = 0, MPE = 1.00) and for the native speakers it was ($\beta = 0.12$, HDI = [0.06, 0.18], ROPE = 8.03, MPE = 0.53).

Similarly, the L2 group was not affected by consonant status or the specific consonant. In contrast, just like in the audio condition, native speakers exhibited a slightly positive effect for consonant status, with an estimate of ($\beta = 0.14$, HDI = [0.08, 0.20], ROPE = 8.71, MPE = 0.54), while for the consonant itself this was ($\beta = 0.08$, HDI = [0.02, 0.14], ROPE = 9.29, MPE = 0.52).

As opposed to the audio condition, the role of proficiency slightly helped in audio-visual processing ($\beta = 0.02$, HDI = [-0.08, 0.12], ROPE = 95.89, MPE = 0.64). As seen in Figure 4, this is true for /n/ but not for /s/, where an increase in proficiency actually decreased the likelihood of selecting the correct consonant. Overall, language proficiency does not show an effect on the processing of canonical versus resyllabified utterances for the L2 learners. Figures A1 and A2 in Appendix A provide a broader overview, including the proficiency distribution across all conditions.

3.3. Visual Condition

The visual scenario was the only condition where learners had a positive intercept estimate ($\beta = 0.21$, HDI = [0.05, 0.37], ROPE = 0, MPE = 1) while native speakers had a negative one ($\beta = -0.02$, HDI = [-0.08, 0.04], ROPE = 8.18, MPE = 0.50).

In the other two conditions, L2 learners were not substantially affected by consonant status ($\beta = 0.01$, HDI = [-0.13, 0.15], ROPE = 89.63, MPE = 0.55) or consonant ($\beta = 0.06$, HDI = [-0.09, 0.22], ROPE = 70.58, MPE = 0.80). On the other hand, native speakers had a negative estimate for both consonants. Their estimate for consonant status was ($\beta = -0.01$, HDI = [-0.07, 0.05], ROPE = 8.82, MPE = 0.51), and for the consonant itself, their estimate was ($\beta = -0.01$, HDI = [-0.07, 0.05], ROPE = 8.08, MPE = 0.50). Additionally, the role of language experience in the visual context was notably influential, with an estimate of 0.13 ($\beta = 0.13$, HDI = [0.04, 0.22], ROPE = 0, MPE = 1).

The results showed that participants with a higher language proficiency were more likely to select the correct consonant. As seen in Figure 2, as proficiency increased, the likelihood of selecting the correct word sequence also increased.

4. Discussion and Conclusions

The present study examined L2 learners' abilities to segment and perceive durational cues in Spanish resyllabification across audio-only, visual-only, and audiovisual conditions. A binomial logistic regression was used to analyze participants' accuracy in identifying syllable combinations and how this correlated with their language proficiency, in terms of both L2 speakers and native speakers. The model included parameters for consonant status, type (/n/ or /s/), LexTALE score, and their interactions across audio-only, visual-only, and audiovisual conditions. For native Spanish speakers, only consonant status, type, and their interactions were included in the model.

The study first confirmed previous research arguing against the perspective of ambisyllabicity and then examined the perceptual role of visual cues in comparison to audio. Consistent with previous research (Lahoz-Bengoechea and Jiménez-Bravo 2021; Scarpace 2017; Strycharczuk and Kohlberger 2016), it was found that prevocalic consonants had a longer duration than coda consonants in the case of /s/ and /n/.

After confirming durational differences in V_CV and VC_V, the first research question asked whether L2 learners were able to detect temporal differences and, therefore, select the item they were being exposed to correctly. The findings reveal that both native Spanish speakers and L2 learners can segment speech accurately in both canonical and resyllabified contexts. L2 learners showed variability in their performance across different conditions, with an average probability of giving the correct response of 0.601 (SD = 0.490) in the audio-only condition, compared to 0.615 (SD = 0.270) for native speakers. The results indicate that L2 learners can detect temporal differences, although with a wider credible interval that is proficiency-dependent. The present study did not observe a reportable distinction in performance between canonical and resyllabified utterances.

However, a perceptual bias was observed for utterances containing /s/ compared to those with /n/. As proficiency increased, the likelihood of selecting /s/ slightly decreased, while the likelihood of selecting /n/ increased. A possible rationale that might explain these results could lie in cross-linguistic interference, which is often articulated as the competition hypothesis. According to this hypothesis, bilinguals tend to exhibit a slower pace in accessing sounds and words, both during perception and production. The presented inaccuracy is attributed to the need to mediate the competition arising from the simultaneous activation of both languages, even when only one language is in use (Kroll et al. 2015; Marian and Spivey 2003). For example, on the input item "la salas" (you salt it) vs. "las alas" (the wings), participants could also activate other options such as "las salas" (the rooms) as well as "salads" and "solace" in English.

There is also evidence supporting the idea of cross-linguistic competition being due to the parallel activation stemming from a variety of linguistic tasks, encompassing phoneme, word, and sentence processing (Blumenfeld and Marian 2013; Li and Gollan 2018; Libben and Titone 2009; Sullivan et al. 2018). This may explain why, when exposed to two very similar utterances in both audio and visual forms, the overload of both sources of information, coupled with the dual system of language structures, might affect how efficiently bilingual individuals perform when tasked with deciding between canonical or resyllabified utterances.

The second research question focused on L2 learners' language proficiency and their ability to adequately segment speech. Contrary to expectations based on dominant-language influences on L2 lexical segmentation (Finn and Hudson Kam 2008; Pajak and Levy 2012; Tremblay et al. 2016, 2018), English speakers with low proficiency in Spanish were not influenced by their L1 in their natural tendency to insert a glottal stop. Contrary to this prediction, proficiency was not a significant factor in the audio and audiovisual sections. The results for the audio condition showed that participants scored above chance, indicating that even at a low proficiency level, students were able to correctly identify the shorter duration of the resyllabified utterances and differentiate them from the longer duration of the canonical forms of both consonants, /n/ and /s/.

Revolving around the role of language proficiency and applying the present study and its results to the revised Speech Learning Model (Flege and Bohn 2021), the results also challenge the perspective of a dynamic learning of phonetic categories that develop over time with language exposure. Only the case of the nasal /n/, which is naturally more acoustically difficult to differentiate from the fricative /s/, fits within the framework of the SLMr and the progressive development of phonetic categories. Yet, as language proficiency increased, participants had more difficulty acoustically identifying the resyllabified /s/, suggesting that greater language experience can negatively affect specific resyllabified distinctions.

The present results align with those of Scarpace (2017), which reported a negative correlation between proficiency and accuracy. In Scarpace's study, beginner learners performed similarly to native speakers in some tasks. Likewise, in this study, increased proficiency was associated with a slight decrease in the likelihood of selecting the correct combination. At the early stages of L2 acquisition, learners appear to develop a parsing strategy that allows them to effectively extract duration cues. The key difference between native speakers and L2 learners lies in the credible interval: native speakers, who rely more on suprasegmental cues like duration, demonstrated a more uniform performance, while L2 speakers exhibited a wider interval in their processing, indicating more variability in how they segment speech. The results can also be framed within the cue-driven window length (Ortega-Llebaria et al. 2019), which suggests that listeners can dynamically adjust the length of their processing window based on the type of acoustic cues available in the speech signal, such as duration. Under the specific conditions of temporal resyllabification in Spanish, some listeners demonstrated the ability to adapt the length of their processing window according to the acoustic cues present in the speech signal, effectively accounting for durational cues more accurately than initially anticipated.

In the audiovisual condition, participants' likelihood of selecting a correct response was slightly higher than in the audio-only condition, indicating that visual articulation aided participants. These results align with previous research on bimodal perception, such as that by Darcy and Thomas (2019), where the presence of visual cues, like lip movements and facial expressions, reduced the reliance on proficiency for correct syllable identification. This, in turn, could decrease the perceptual distance between speech differences, thereby facilitating word recognition. Additionally, the interplay between top-down (visual) and bottom-up (audio) processes (Hartcher-O'Brien et al. 2017; Macaluso et al. 2016) seemed to contribute to the "fusion" of these data streams, as depicted by (Bicevskis et al. 2016; Chandrasekaran et al. 2009; Rosenblum and Dorsi 2021; Turk 2014), resulting in slightly better perception.

The last research question tackled the role of the visual articulatory cues' impact on L2 learners. There was no surprise in the decline in overall accuracy due to the absence of auditory cues, marking this condition as the one with the poorest accuracy rate. On the other hand, with increasing proficiency, the likelihood of correctly identifying both consonants across canonical and resyllabified utterances increased, contrary to the previous conditions. This condition was the only one where high-proficiency speakers had a greater capability to disambiguate durational cues using only visual cues compared to low-proficiency speakers. These results build on previous studies where participants successfully identified visual cues related to duration (e.g., Lidestam 2009; Scott and Idrissi 2018). These findings extend this body of research to Spanish, showing that L2 learners of Spanish can also perceive visual articulatory durations. The spread in the credible interval was larger for L2 learners compared to native speakers, a variation which highlights the influence of language experience. As learners become more proficient, they improve their ability to detect and interpret visual pauses, which aids in both speech perception and syllable segmentation.

These results pave the way for future research, suggesting that visual category formation and syllable segmentation may evolve with language experience. This parallels the Speech Learning Model (Flege and Bohn 2021), which posits that as L2 learners gain proficiency, they refine their phonetic categories for L2 sounds. Similarly, my findings

suggest that with increased proficiency, learners may also develop distinct visual categories that enhance their ability to segment speech more effectively, contributing to better lexical segmentation in their L2.

The present study contributes to research in this area by exploring speech segmentation strategies in native Spanish speakers and English-speaking L2 learners of Spanish in the context of Spanish resyllabification. This study provided evidence that both native Spanish speakers and L2 learners can accurately segment speech in both canonical and resyllabified contexts, even at low proficiency levels, suggesting that L2 learners adopt lexical duration segmentation strategies early in their language acquisition process. When comparing the results with those of native speakers, it is observed that the credible interval for native speakers is shorter, while the interval for L2 learners is wider. Although both groups exhibited a similar pattern—performing best in the audiovisual condition, followed by the audio condition, and worst in the visual-only condition—native speakers showed more consistent responses, with a narrower spread of accuracy compared to L2 learners, indicating proficiency-dependent results. The visual-only condition posed the greatest challenge, but higher-proficiency participants benefited the most from visual cues. Overall, these findings suggest that aural syllabic segmentation develops early in L2 proficiency, while visual syllabic segmentation becomes more refined as proficiency increases.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/languages9110346/s1>, and includes the sentences utilized in the study.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Rutgers (protocol code eIRB Pro2022000980, approved 1 July 2022).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The original contributions presented in this study are included in the article/Supplementary Materials; further inquiries can be directed to the corresponding author.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A

Table A1. Summary of multi-response variables.

Condition	Speaker	Parameter	Estimate	% in ROPE	MPE	Rhat	ESS
Audio	L2	Intercept	0.42 [0.22, 0.63]	0.00	1.00	1	1838.73
	L2	Consonant status	0.05 [−0.13, 0.24]	67.21	0.72	1	1623.46
	L2	Consonant	−0.09 [−0.28, 0.11]	53.03	0.82	1	2242.49
	L2	LexTALE	−0.05 [−0.14, 0.05]	88.47	0.82	1	3443.41
	L2	Consonant status × consonant	0.03 [−0.15, 0.23]	71.74	0.65	1	1702.05
	L2	Consonant status × LexTALE	0 [−0.07, 0.07]	100.00	0.54	1	6708.80
	L2	Consonant × LexTALE	0.05 [−0.04, 0.13]	90.58	0.86	1	4617.13
	L2	Consonant status × consonant × LexTALE	0.1 [0.03, 0.17]	54.55	1.00	1	6531.75
Visual	L2	Intercept	0.21 [0.05, 0.37]	6.13	0.99	1	1736.45
	L2	Consonant status	0.01 [−0.13, 0.15]	89.63	0.55	1	1766.08
	L2	Consonant	0.06 [−0.09, 0.22]	70.58	0.80	1	1984.80
	L2	LexTALE	0.13 [0.04, 0.22]	26.92	1.00	1	2737.85
	L2	Consonant status × consonant	0.05 [−0.1, 0.2]	77.47	0.77	1	1828.02
	L2	Consonant status × LexTALE	0.03 [−0.04, 0.1]	99.00	0.82	1	5675.13
	L2	Consonant × LexTALE	−0.02 [−0.1, 0.06]	100.00	0.69	1	3797.62
	L2	Consonant status × consonant × LexTALE	0.1 [0.03, 0.17]	53.61	1.00	1	5270.62
A/V	L2	Intercept	0.46 [0.3, 0.63]	0.00	1.00	1	1997.48

Table A1. Cont.

Condition	Speaker	Parameter	Estimate	% in ROPE	MPE	Rhat	ESS
Audio	L2	Consonant status	0 [−0.14, 0.15]	88.11	0.52	1	2013.43
	L2	Consonant	−0.11 [−0.26, 0.05]	44.16	0.93	1	2385.81
	L2	LexTALE	0.02 [−0.08, 0.12]	95.89	0.64	1	2931.19
	L2	Consonant status × consonant	−0.04 [−0.18, 0.1]	83.08	0.70	1	2110.23
	L2	Consonant status × LexTALE	0.02 [−0.05, 0.09]	100.00	0.68	1	4792.99
	L2	Consonant × LexTALE	0.09 [0, 0.19]	55.08	0.98	1	3902.98
	L2	Consonant status × consonant × LexTALE	−0.02 [−0.09, 0.05]	100.00	0.70	1	4449.51
Visual	Native	Intercept	0.11 [0.05, 0.17]	10.21	0.53	1	3415.77
	Native	Consonant status	0.10 [0.04, 0.16]	9.13	0.53	1	3234.50
	Native	Consonant	0.09 [0.03, 0.15]	9.71	0.51	1	3042.35
	Native	Consonant status × consonant	0.11 [0.05, 0.17]	7.24	0.52	1	3368.30
A/V	Native	Intercept	−0.02 [−0.08, 0.04]	8.18	0.50	1	3408.58
	Native	Consonant status	−0.01 [−0.07, 0.05]	8.82	0.51	1	4031.86
	Native	Consonant	−0.01 [−0.07, 0.05]	8.08	0.50	1	4280.15
	Native	Consonant status × consonant	0.02 [−0.04, 0.08]	8.95	0.51	1	3742.78
A/V	Native	Intercept	0.12 [0.06, 0.18]	8.03	0.53	1	3779.17
	Native	Consonant status	0.14 [0.08, 0.20]	8.71	0.54	1	4092.63
	Native	Consonant	0.08 [0.02, 0.14]	9.29	0.52	1	4468.32
	Native	Consonant status × consonant	0.12 [0.06, 0.18]	8.84	0.52	1	4211.01

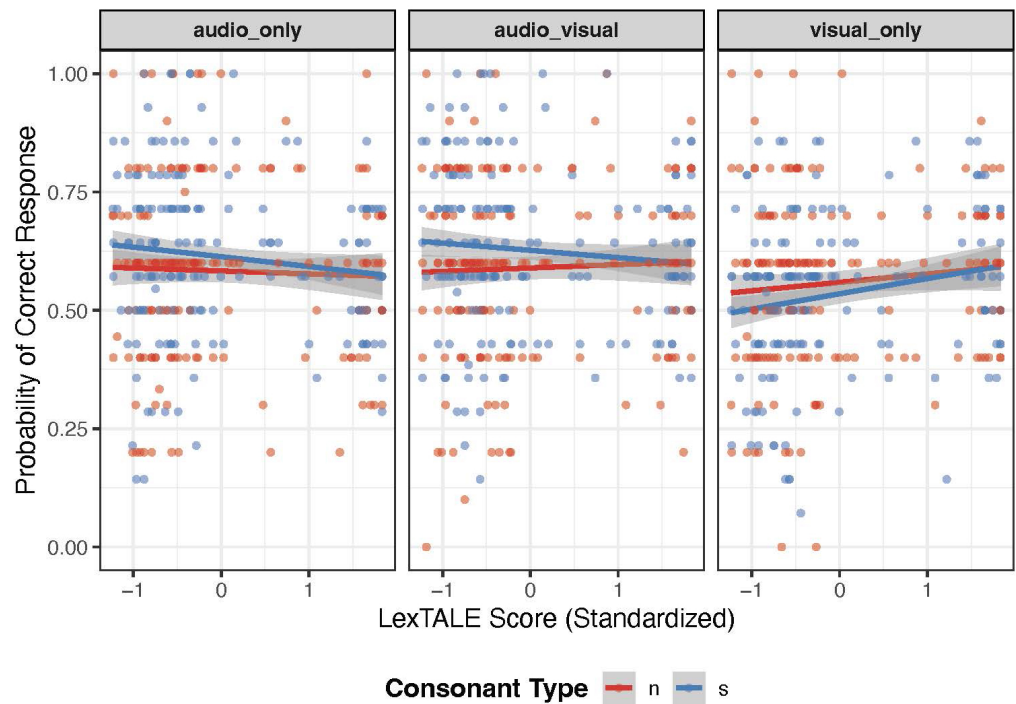


Figure A1. Probability of a correct response as a function of LexTALE scores (standardized) across three presentation modes: audio-only, audio-visual, and visual-only. Each point represents individual participant responses, with trend lines indicating the relationship between LexTALE scores and response accuracy for each consonant type. Gray shading represents the 95% confidence intervals.

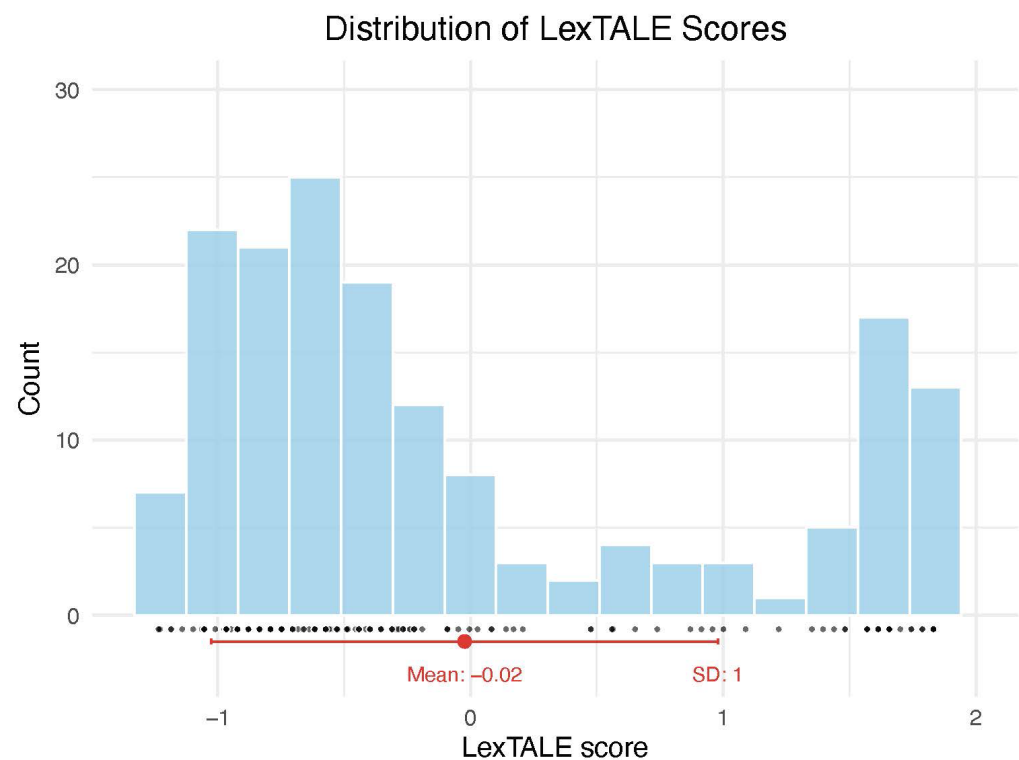


Figure A2. Distribution of LexTALE scores among participants. The histogram displays the frequency of LexTALE scores. Black dots along the x-axis represent individual scores.

References

- Alammar, Aseel. 2015. Second language perception of word segmentation. In *Proceedings of LSUGA's Second Interdisciplinary Conference in Linguistics*. Edited by Lisa Lipani, Michael Olsen and Douglas Merchant. The Linguistics Society at the University of Georgia. Athens: University of Georgia, pp. 1–13.
- Bermúdez-Otero, Ricardo, and John Payne. 2011. There are no special clitics. In *Morphology and Its Interfaces*. Amsterdam and Philadelphia: John Benjamins.
- Best, Catherine T., and Michael D. Tyler. 2007. Nonnative and second-language speech perception. In *Language Experience in Second Language Speech Learning*. Edited by Ocke-Schwen Bohn and Murray J. Munro. Amsterdam: John Benjamins Publishing Company, pp. 13–34. [CrossRef]
- Bicevskis, Katie, Donald Derrick, and Bryan Gick. 2016. Visual-tactile integration in speech perception: Evidence for modality neutral speech primitives. *The Journal of the Acoustical Society of America* 140: 3531–39. [CrossRef] [PubMed]
- Birdsong, David, Libby M. Gertken, and Mark Amengual. 2012. *Bilingual Language Profile: An Easy-to-Use Instrument to Assess Bilingualism*. Austin: COERLL, University of Texas at Austin.
- Blumenfeld, Henrike K., and Viorica Marian. 2013. Parallel language activation and cognitive control during spoken word recognition in bilinguals. *Journal of Cognitive Psychology* 25: 547–67. [CrossRef] [PubMed]
- Boersma, Paul, and David Weenink. 2018. *Praat: Doing Phonetics by Computer* [Computer Program]. Available online: <http://www.praat.org/> (accessed on 31 October 2024).
- Brancazio, Lawrence, and Joanne L. Miller. 2005. Use of visual information in speech perception: Evidence for a visual rate effect both with and without a McGurk effect. *Perception & Psychophysics* 67: 759–69.
- Bürkner, Paul-Christian. 2017. Brms: An r package for Bayesian multilevel models using stan. *Journal of Statistical Software* 80: 1–28. [CrossRef]
- Campbell, Nick, and Mary Beckman. 1997. Stress, prominence, and spectral tilt. Paper presented at ESCA Tutorial and Research Workshop on Intonation: Theory, Models and Applications, Athens, Greece, September 18–20.
- Carroll, Susanne E. 2004. Segmentation: Learning how to “hear words” in the L2 speech stream. *Transactions of the Philological Society* 102: 227–54. [CrossRef]
- Chandrasekaran, Chandramouli, Andrea Trubanova, Sébastien Stillitano, Alice Caplier, and Asif A. Ghazanfar. 2009. The natural statistics of audiovisual speech. *PLoS Computational Biology* 5: e1000436. [CrossRef]
- Colina, Sonia. 1995. *A Constraint-Based Analysis of Syllabification in Spanish, Catalan, and Galician*. Champaign: University of Illinois at Urbana-Champaign.
- Colina, Sonia. 1997. Identity constraints and Spanish resyllabification. *Lingua* 103: 1–23. [CrossRef]

- Colina, Sonia. 2006. Optimality-theoretic advances in our understanding of Spanish syllable structure. *Optimality-Theoretic Studies in Spanish Phonology* 99: 172–204.
- Crowhurst, Megan J. 1992. Diminutives and augmentatives in Mexican Spanish: A prosodic analysis. *Phonology* 9: 221–53. [\[CrossRef\]](#)
- Cutler, Anne. 2012. *Native Listening: Language Experience and the Recognition of Spoken Words*. Cambridge, MA: MIT Press.
- Darcy, Isabelle, and Trisha Thomas. 2019. When blue is a disyllabic word: Perceptual epenthesis in the mental lexicon of second language learners. *Bilingualism: Language and Cognition* 22: 1141–59. [\[CrossRef\]](#)
- Desroches, Amy S., Deanna C. Friesen, Matthew Teles, Chloe A. Korade, and Evan W. Forest. 2022. The dynamics of spoken word recognition in bilinguals. *Bilingualism: Language and Cognition* 25: 705–10. [\[CrossRef\]](#)
- Dilley, Laura, and Stefanie Shattuck-Hufnagel. 1995. Individual differences in the glottalization of vowel-initial syllables. *The Journal of the Acoustical Society of America* 97 S5: 3418–19. [\[CrossRef\]](#)
- Finn, Amy S., and Carla L. Hudson Kam. 2008. The curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition* 108: 477–99. [\[CrossRef\]](#) [\[PubMed\]](#)
- Flege, James E., and Ocke-Schwen Bohn. 2021. The revised speech learning model (SLM-r). In *Second Language Speech Learning: Theoretical and Empirical Progress*. Cambridge: Cambridge University Press, vol. 10. [\[CrossRef\]](#)
- Harris, James W., and Ellen M. Kaisse. 1999. Palatal vowels, glides and obstruent's in Argentinian Spanish. *Phonology* 16: 117–90. [\[CrossRef\]](#)
- Harris, James Wesley. 1983. Syllable structure and stress in Spanish. A nonlinear analysis. *Linguistic Inquiry Monographs Cambridge, Mass* 8: 1–158.
- Hartcher-O'Brien, Jess, Salvador Soto-Faraco, and Ruth Adam. 2017. A matter of bottom-up or top-down processes: The role of attention in multisensory integration. In *Frontiers in Integrative Neuroscience*. Lausanne: Frontiers Media SA, vol. 11, p. 5.
- Hazan, Valerie, Anke Sennema, and Andrew Faulkner. 2002. Audiovisual perception in L2 learners. Paper presented at Seventh International Conference on Spoken Language Processing, Denver, CO, USA, September 16–20.
- Hazan, Valerie, Anke Sennema, Andrew Faulkner, Marta Ortega-Llebaria, Midori Iba, and Hyunsong Chung. 2006. The use of visual cues in the perception of non-native consonant contrasts. *The Journal of the Acoustical Society of America* 119: 1740–51. [\[CrossRef\]](#)
- Hualde, José Ignacio. 2005. *The Sounds of Spanish with Audio CD*. Cambridge: Cambridge University Press.
- Hualde, José Ignacio. 2009. Unstressed words in Spanish. *Language Sciences* 31: 199–212. [\[CrossRef\]](#)
- Hualde, José Ignacio, and Pilar Prieto. 2015. Lenition of intervocalic alveolar fricatives in Catalan and Spanish. *Phonetica* 71: 109–27. [\[CrossRef\]](#)
- Izura, Cristina, Fernando Cuetos, and Marc Brysbaert. 2014. Lextale-esp: A test to rapidly and efficiently assess the Spanish vocabulary size. *Psicológica* 35: 49–66.
- Katayama, Tamami. 2015. Effect of phonotactic constraints on second language speech processing. *I-Perception* 6: 2041669515615714. [\[CrossRef\]](#)
- Kroll, Judith F., Paola E. Dussias, Kinsey Bice, and Lauren Perrotti. 2015. Bilingualism, mind, and brain. *Annual Review of Linguistics* 1: 377–94. [\[CrossRef\]](#) [\[PubMed\]](#)
- Ladd, Dwight Robert, and Astrid Schepman. 2003. "Sagging transitions" between high pitch accents in English: Experimental evidence. *Journal of Phonetics* 31: 81–112. [\[CrossRef\]](#)
- Lahoz-Bengochea, José María, and Miguel Jiménez-Bravo. 2021. *Spoken Word Boundary Detection in Ambiguous Resyllabification Contexts in Spanish*. Madrid: Complutense University of Madrid.
- Leonte, Anna, Lorenza S. Colzato, Laura Steenbergen, Bernhard Hommel, and Elkan G. Akyürek. 2018. Supplementation of gamma-aminobutyric acid (GABA) affects temporal, but not spatial visual attention. *Brain and Cognition* 120: 8–16. [\[CrossRef\]](#) [\[PubMed\]](#)
- Levy, Erika S., and Winifred Strange. 2008. Perception of French vowels by American English adults with and without French language experience. *Journal of Phonetics* 36: 141–57. [\[CrossRef\]](#)
- Lewkowicz, David J., and Amy M. Hansen-Tift. 2012. Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences of the United States of America* 109: 1431–36. [\[CrossRef\]](#)
- Li, Chuchu, and Tamar H. Gollan. 2018. Cognates facilitate switches and then confusion: Contrasting effects of cascade versus feedback on language selection. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 44: 974. [\[CrossRef\]](#)
- Libben, Maya R., and Debra A. Titone. 2009. Bilingual lexical access in context: Evidence from eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35: 381. [\[CrossRef\]](#)
- Liberman, Alvin M., and Ignatius G. Mattingly. 1985. The motor theory of speech perception revised. *Cognition* 21: 1–36. [\[CrossRef\]](#)
- Liberman, Isabelle Y. 1973. Segmentation of the spoken word and reading acquisition. *Bulletin of the Orton Society* 23: 65–77. [\[CrossRef\]](#)
- Lidestam, Björn. 2009. Visual discrimination of vowel duration. *Scandinavian Journal of Psychology* 50: 427–35. [\[CrossRef\]](#)
- Lipski, John. 1989. /s/-voicing in Ecuadorian Spanish: Patterns and principles of consonantal modification. *Lingua* 79: 49–71. [\[CrossRef\]](#)
- Macaluso, Emiliano, Uta Noppeney, Durk Talsma, Tiziana Vercillo, Jess Hartcher-O'Brien, and Ruth Adam. 2016. The curious incident of attention in multisensory integration: Bottom-up vs. Top-down. *Multisensory Research* 29: 557–83. [\[CrossRef\]](#)
- Marian, Viorica, and Michael Spivey. 2003. Competing activation in bilingual language processing: Within-and between-language competition. *Bilingualism: Language and Cognition* 6: 97–115. [\[CrossRef\]](#)
- Maye, Jessica, Janet F. Werker, and LouAnn Gerken. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82: B101–11. [\[CrossRef\]](#) [\[PubMed\]](#)

- McGurk, Harry, and John MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264: 746–48. [[CrossRef](#)] [[PubMed](#)]
- Nakatani, Lloyd H., and Carletta H. Aston. 1978. Perceiving the stress pattern of words in sentences. *The Journal of the Acoustical Society of America* 63 S1: S55. [[CrossRef](#)]
- Ortega-Llebaria, Marta, Daniel J. Olson, and Alba Tuninetti. 2019. Explaining cross-language asymmetries in prosodic processing: The cue-driven window length hypothesis. *Language and Speech* 62: 701–36. [[CrossRef](#)]
- Pajak, Bożena, and Roger Levy. 2012. Distributional learning of L2 phonological categories by listeners with different language backgrounds. In *Proceedings of the 36th Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press, vol. 2, pp. 400–13.
- Peelle, Jonathan E., and Mitchell S. Sommers. 2015. Prediction and constraint in audiovisual speech perception. *Cortex* 68: 169–81. [[CrossRef](#)]
- Pegg, Judith E., Janet F. Werker, and Peter J. McLeod. 1992. Preference for infant-directed over adult-directed speech: Evidence from 7-week-old infants. *Infant Behavior and Development* 15: 325–45. [[CrossRef](#)]
- Pike, Kenneth L. 1945. *The Intonation of American English*. Ann Arbor: University of Michigan Press.
- Polka, Linda. 1995. Linguistic influences in adult perception of non-native vowel contrasts. *The Journal of the Acoustical Society of America* 97: 1286–96. [[CrossRef](#)]
- Robinson, Kimball. 2012. The dialectology of syllabification: A review of variation in the Ecuadorian highlands. *Romance Philology* 66: 115–45. [[CrossRef](#)]
- Rosenblum, Lawrence D., and Josh Dorsi. 2021. Primacy of multimodal speech perception for the brain and science. In *The Handbook of Speech Perception*. Hoboken: John Wiley & Sons, Inc., pp. 28–57.
- Ross, Lars A., Dave Saint-Amour, Victoria M. Leavitt, Daniel C. Javitt, and John J. Foxe. 2007. Do you see what i am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex* 17: 1147–53. [[CrossRef](#)] [[PubMed](#)]
- Saffran, Jenny R., Elissa L. Newport, and Richard N. Aslin. 1996. Word segmentation: The role of distributional cues. *Journal of Memory and Language* 35: 606–21. [[CrossRef](#)]
- Sanders, Lisa D., and Helen J. Neville. 2003. An ERP study of continuous speech processing: I. Segmentation, semantics, and syntax in native speakers. *Cognitive Brain Research* 15: 228–40. [[CrossRef](#)] [[PubMed](#)]
- Scarpace, Daniel L. 2017. *The Acquisition of Resyllabification in Spanish by English Speakers*. Doctoral dissertation, University of Illinois at Urbana-Champaign, Champaign, IL, USA.
- Schwartz, Jean-Luc, Anahita Basirat, Lucie Ménard, and Marc Sato. 2012. The perception-for-action-control theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics* 25: 336–54. [[CrossRef](#)]
- Scobbie, James M., and Marianne Pouplier. 2010. The role of syllable structure in external sandhi: An EPG study of vocalisation and retraction in word-final English /l/. *Journal of Phonetics* 38: 240–59. [[CrossRef](#)]
- Scott, Mark, and Ali Idrissi. 2018. Audiovisual perception of gemination and pharyngealization in Arabic. *Speech Communication* 98: 17–27. [[CrossRef](#)]
- Strycharczuk, Patrycja, and Martin Kohlberger. 2016. Resyllabification reconsidered: On the durational properties of word-final /s/ in Spanish. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 7: 3.
- Sullivan, Margot D., Gregory J. Poarch, and Ellen Bialystok. 2018. Why is lexical retrieval slower for bilinguals? Evidence from picture naming. *Bilingualism: Language and Cognition* 21: 479–88. [[CrossRef](#)]
- Torreira, Francisco, and Mirjam Ernestus. 2012. Weakening of intervocalic /s/ in, the Nijmegen corpus of casual Spanish. *Phonetica* 69: 124–48. [[CrossRef](#)]
- Tremblay, Annie, Mirjam Broersma, and Caitlin E. Coughlin. 2018. The functional weight of a prosodic cue in the native language predicts the learning of speech segmentation in a second language. *Bilingualism: Language and Cognition* 21: 640–52. [[CrossRef](#)]
- Tremblay, Annie, Mirjam Broersma, Caitlin E. Coughlin, and Jiyoung Choi. 2016. Effects of the native language on the learning of fundamental frequency in second-language speech segmentation. *Frontiers in Psychology* 7: 985. [[CrossRef](#)] [[PubMed](#)]
- Tremblay, Annie, Sahyang Kim, Seulgi Shin, and Taehong Cho. 2021. Re-examining the effect of phonological similarity between the native- and second-language intonational systems in second-language speech segmentation. *Bilingualism: Language and Cognition* 24: 401–13. [[CrossRef](#)]
- Turk, Matthew. 2014. Multimodal interaction: A review. *Pattern Recognition Letters* 36: 189–95. [[CrossRef](#)]
- Umeda, Noriko. 1978. Occurrence of glottal stops in fluent speech. *The Journal of the Acoustical Society of America* 64: 88–94. [[CrossRef](#)]
- Werker, Janet F., Ferran Pons, Christiane Dietrich, Sachiyu Kajikawa, Laurel Fais, and Shigeaki Amano. 2007. Infant-directed speech supports phonetic category learning in English and Japanese. *Cognition* 103: 147–62. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.