

Article

Distribution and Acoustic Characteristics of Filled Pauses in Spontaneous Urdu Speech

Saira Zahid ^{1,2}, Ho-Young Lee ^{3,*} and Muhammad Asim Mahmood ¹ 

¹ Department of Applied Linguistics, Government College University Faisalabad, Faisalabad 38000, Pakistan; sara.linguistics@gmail.com (S.Z.); masimrai@gmail.com (M.A.M.)

² Basic Sciences and Humanities, University of Engineering and Technology Lahore, Faisalabad Campus, Lahore 54890, Pakistan

³ Department of Linguistics, Seoul National University, Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea

* Correspondence: hylee@snu.ac.kr

Abstract

This study examines the distribution and acoustic characteristics of filled pauses (FPs) in Urdu, a language underrepresented in disfluency research. Drawing on a spontaneous speech dataset from 18 female speakers, the analysis considers the types of FPs, their immediate segmental context, and their utterance position. The analysis also evaluates the effects of segmental context and utterance position on acoustic measures of FPs. Results show a dominant use of vocalic FPs. Moreover, FPs observe systematic contextual patterns and cluster in specific utterance positions. Acoustically, vowel-only and vowel–nasal FPs differ in duration and vowel height (F1). For vowel-only FPs, utterance position significantly conditions duration and prosodic properties (F0, intensity), whereas segmental context does not show any effects. Taken together, the findings demonstrate a language-specific organization of FPs in Urdu. This study offers a detailed phonetic account of Urdu FPs to date and highlights the importance of language-sensitive disfluency modeling in speech technology applications.

Keywords: disfluencies; filled pauses; spontaneous speech; acoustic analysis; Urdu

1. Introduction

Spontaneous speech involves real-time planning for conceptualization, formulation, and articulation (Levelt, 1993). During these processes, speakers often need to revise earlier segments, or pause to plan upcoming content, resulting in phenomena collectively termed disfluencies. Filled pauses (defined as gaps in speech filled with non-lexical vocalizations that do not add propositional meaning) is one type of such phenomenon.

Although early studies regarded them as indicators of hesitation/disfluency, more recent work has identified filled pauses (FPs) as poly-functional devices that contribute to discourse structuring and self-repair (Clark & Fox Tree, 2002; Götz, 2013). Within this literature, FPs have been discussed from multiple perspectives, including their relation to cognitive processing demands and their potential role in managing interaction. However, the extent to which these functions generalize across languages remains an open empirical question.



Academic Editor: Donna Erickson

Received: 14 August 2025

Revised: 9 February 2026

Accepted: 18 February 2026

Published: 25 February 2026

Copyright: © 2026 by the authors.

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.

Most existing research on FPs has focused widely on European languages, particularly English, Dutch, French, and German (see, e.g., [Betz, 2020](#); [Clark & Fox Tree, 2002](#); [de Boer & Heeren, 2020](#); [De Leeuw, 2007](#); [Kirjavainen et al., 2022](#); [Kosmala & Crible, 2022](#); [E. E. Shriberg, 1994](#); [Swerts, 1998](#)). However, research on FPs in South Asian languages remains extremely limited; to our knowledge, [Jabeen and Betz \(2022\)](#) and [Jabeen and Wagner \(2023\)](#) constitute the only published studies on the topic. This lack of empirical data restricts our understanding of how language-specific structures and prosody norms shape disfluency phenomena in diverse linguistic contexts. Given the growing importance of research on disfluencies in both spoken interaction and computational speech technologies, the present research is timely and necessary to address this gap in South Asian languages.

The present study addresses this gap by providing an empirical investigation of FPs in spontaneous Urdu speech produced by female speakers. Focusing on distributional patterns, segmental context, utterance position, and acoustic realization, we examine how FPs are organized in Urdu and which structural factors condition their phonetic properties. By doing so, the study aims to document language-specific regularities that can serve as a foundation for future theoretical and applied work on disfluency.

2. Review of Relevant Literature

2.1. The Function of Filled Pauses

Filled pauses (FPs) are often linked to hesitation, typically emerging when speakers experience uncertainty ([Smith & Clark, 1993](#)) or face decision-making moments requiring selection among available options ([Finlayson & Corley, 2012](#)).

FPs are seen as indicators of production difficulty, reflecting the speaker's cognitive load ([Levelt, 1983](#)). They are thought to reflect planning challenges at various linguistic levels like content, syntax and word choice ([Goldman-Eisler, 1968](#)) and serve as tools for getting additional time to manage verbal planning, much like silent pauses ([MacKay & Osgood, 1959](#)). This aligns with the broader understanding that hesitation arising from speech planning is a primary function of FPs, as they typically emerge during self-monitoring when additional time is needed for conceptualizing, selecting information, or encoding the intended message into appropriate linguistic structures, particularly when retrieving a specific word ([Levelt, 1993](#); [Tottie, 2016, 2020](#)). So, FPs increase during decision-making or while considering multiple options ([Christenfeld, 1994](#)), describing ambiguous situations ([Schnadt & Corley, 2006](#)), and with conceptually demanding or lengthy utterances ([Oviatt, 1995](#); [Watanabe et al., 2008](#)). However, more recent studies suggest that utterance length or conceptual complexity alone does not reliably predict filled pause frequency, indicating that the relationship between planning difficulty and filled pause use is more nuanced than previously assumed ([Paschen, 2023](#)).

Research highlights that FPs go beyond marking hesitation while retrieving words or planning speech; they also play a role in signaling discourse structure and managing conversational turns. From the "FPs-as-signal" perspective, they are seen as communicative tools that convey information about interactional dynamics. [Swerts \(1998\)](#) suggests that FPs can indicate shifts in discourse structure, while others emphasize their role in turn-taking ([Beňuš, 2009](#); [Kjellmer, 2003](#)). For instance, FPs may help a speaker maintain control of the conversational floor ([MacKay & Osgood, 1959](#)) or, in contrast, signal an intention to yield their turn ([Clark & Fox Tree, 2002](#)).

In addition to their production-related functions, FPs also affect listener processing. [Swerts \(1998\)](#) found that in Dutch, FPs at the beginning of major discourse segments show distinct prosodic patterns which may help listeners anticipate a shift in topic or structure. [Clark and Fox Tree \(2002\)](#) support this listener-oriented perspective by showing that, in American English, FPs can facilitate faster recognition of upcoming words, indicating their

role in previewing upcoming speech content. Furthermore, the acoustic form of FPs may correspond to the scale of the delay they signal: vocalic forms (e.g., uh) are often associated with minor delays¹, while vocalic-nasal forms (e.g., um) tend to mark more major delays (Clark & Fox Tree, 2002; E. E. Shriberg, 1994; Swerts, 1998).

The signal perspective has been widely questioned, particularly regarding the claim that FPs reliably signal upcoming delays. O’Connell and Kowal (2005) argued that their interpretation is context-dependent and not based on any inherent communicative function. Others have similarly contended that these are unintended by-products of production difficulty, rather than deliberate signals. Lickley (2015) explicitly rejects Clark and Fox Tree’s interpretation of his findings, emphasizing that FPs do not carry prosodic features indicative of intentional signaling.

The notion that FPs serve to hold the conversational floor, as proposed by Maclay and Osgood (1959), has become complicated by later evidence. Eklund (2004) reports that FPs are more frequent in human–machine interactions, where there is no risk of losing the floor, than in human–human conversations. Rather than serving interactional functions, this pattern is more consistent with increased cognitive load and task complexity, supporting the view that FPs reflect planning demands rather than intentional signaling behavior.

Taken together, FPs are multifunctional and potentially ambiguous, reflecting a complex interaction between cognitive processes, discourse structure, and language-specific conventions. Given the multifunctional and context-dependent nature of FPs, understanding their role in speech requires careful examination of their characteristics. Functional interpretations cannot be separated from how FPs are realized acoustically, how they are structured segmentally, and where they occur within an utterance. Consequently, research has increasingly focused on the phonetic characteristics, structural types, and positional distribution of FPs across languages, which is the focus of the next section.

2.2. Characteristics of Filled Pauses

Although FPs commonly involve a central or centralized vowel with an optional nasal segment across languages (Clark & Fox Tree, 2002; De Leeuw, 2007; Lickley, 2015; E. Shriberg, 2001), the specific quality of this central vowel differs by language (Candea et al., 2005; de Boer & Heeren, 2020). Candea et al. (2005), in their corpus-based study of eight languages (i.e., Arabic, Mandarin Chinese, French, German, Italian, European Portuguese, American English, and Latin American Spanish), reported language-specific normalized formant values, reflecting the influence of each language’s vowel system. This suggests that the vowels in FPs are not merely products of a generalized principle of the least effort, often referred to as the “articulatory rest position” (Candea et al., 2005, p. 51), but are instead shaped by language-specific phonological structures.

At the segmental level, FPs are often categorized into vocalic (V), vocalic-nasal (VN) and nasal consonant² (N), a distinction observed across multiple languages, for instance, English (Clark & Fox Tree, 2002; Kjellmer, 2003; E. Shriberg, 2001), and German (Fischer et al., 2017; Niebuhr & Fischer, 2019). Clark and Fox Tree (2002) also provide an overview of language-specific variants. Preferences for vocalic or vocalic-nasal forms appear to be language-dependent. German and English, for example, show a tendency toward a higher ratio of vocalic-nasal to vocalic forms (De Leeuw, 2007; Wieling et al., 2016), whereas French (Torreira et al., 2010), Dutch (De Leeuw, 2007) and Urdu (Jabeen & Betz, 2022) exhibit a greater preference for vocalic-only forms. These forms differ systematically in their distribution, duration, and discourse function, and have been linked to different types of planning delays (Clark & Fox Tree, 2002). For this reason, the present study focuses on FP types (vocalic, vocalic-nasal, nasal) when examining the acoustic and positional properties of FPs in Urdu.

FPs also differ in their phonetic realization across languages and speakers, with notable variation in duration, intensity, and fundamental frequency (F0). In terms of duration, FPs are typically longer in cognitively demanding contexts. Extended durations have been observed at phrase-initial positions in Dutch monologs (Swerts, 1998), during the introduction of new segments in Hungarian speech (Horváth, 2010), and when speakers experience lexical retrieval difficulty in Italian (Cataldo et al., 2019). Cross-linguistic evidence consistently shows that FPs tend to last longer than lexical vowels (Hughes et al., 2016; E. Shriberg, 2001), and their duration increases with anticipated delays (Clark & Fox Tree, 2002).

Maekawa and Mori (2017), in their acoustic analysis of Japanese, found that both prosodic (F0, intensity) and voice quality features distinguish FPs from ordinary lexical items. Using a random forest classifier, they reported that duration was the most informative feature and intensity showed the next highest importance. In terms of F0, FPs generally have lower F0 than surrounding speech (O'Shaughnessy, 1992). Their F0 is strongly influenced by utterance position: Swerts (1998) found that FPs occurring at the beginning of an utterance tend to have a higher F0 than those in medial positions. Furthermore, when FPs occur mid-clause, they tend to align prosodically with adjacent speech, indicating smooth integration into the utterance structure (E. E. Shriberg & Lickley, 1993). Overall, these features vary depending on the filled pause's position and function within the discourse. Hence, utterance position must be taken into account when analyzing FPs, as it has been shown to influence their phonetic realization in both theoretical modeling and forensic applications (de Boer & Heeren, 2020). In line with this, our study focuses on the position of FPs within the utterance and also the immediate context in which they occur.

2.3. Filled Pauses in Urdu

Research on FPs in Urdu remains limited, with Jabeen and Betz (2022) providing the only available study to date. Their analysis was based on approximately 25 min of speech from 14 speakers, collected in a semi-spontaneous, interview-style setting, where participants responded to predefined prompts. The study examined the frequency of hesitation markers, the properties of FPs, and the formant structure of vocalic FPs to assess vowel quality. Their findings indicate that vocalic (V) and vocalic-nasal (VN) forms are the most frequent FPs. While um-type FPs appear more often in turn-medial positions, uh-type FPs occur with similar frequency in turn-initial and turn-medial positions. Analysis of vowel quality showed that both forms share a close-mid central vowel. While their study provided important initial descriptive insights into hesitation patterns in Urdu, its scope was limited to semi-spontaneous speech and primarily focused on distributional frequency and vowel quality of vocalic FPs; the present study goes beyond their work in several key aspects. First, it examines FPs in spontaneous interaction, defined here as unscripted, non-read interaction elicited with minimal constraints. The speakers were given broad topics (e.g., future plans, daily routines, etc.) and then allowed to talk freely with each other without interruption from the researcher, selecting their own content and discourse structure. Second, the present analysis incorporates a broader set of parameters, including (i) filled pause type (vocalic, vocalic-nasal, nasal-only); (ii) segmental context: silence-FP-silence (henceforth SS), word-FP-word (WW), silence-FP-word (SW), word-FP-silence (WS); (iii) utterance position (initial, medial, final, single); and (iv) multiple acoustic measures (duration, F0, intensity, F1, and F2), allowing for a comprehensive examination of both structural and phonetic variation.

On this basis, the present study addresses the following research questions:

- (1) How are different types of FPs distributed across segmental contexts and utterance positions in spontaneous Urdu speech?
- (2) Do filled pause types differ systematically in their acoustic realization?
- (3) How do utterance position and segmental context condition the distribution and acoustic realization of FPs in Urdu?

By addressing these questions, the present study provides a detailed description of FPs in Urdu and contributes new empirical evidence to cross-linguistic research on the structural and phonetic patterning of hesitation phenomena.

3. Corpus and Methods

The FPs of eighteen female native Urdu speakers were analyzed to investigate variation in their usage and phonetic realization. Participants were undergraduate students at the Government College University Faisalabad, aged 18–25 years ($M = 20$), with no reported language disorders. Male speakers were excluded from this study to avoid potential gender-related differences in filled pause distribution reported in previous research (Wieling et al., 2016), as examining such effects was beyond the scope of the present analysis.

Recordings were conducted in a quiet indoor space using a H4N recorder, positioned approximately 30 cm from the speaker. Each of the nine spontaneous conversations lasted 12 min, yielding a total of 108 min of recorded speech. The speakers discussed topics related to their daily routines, hobbies, and future plans. All sessions were facilitated by the same researcher, who introduced the discussion topics at the beginning and did not intervene during the conversations. All speakers were recorded in comparable conditions and did not know that FPs were the focus of analysis.

FPs and silences were identified across the recordings and analyzed for each speaker based on five parameters: (a) frequency of FPs; (b) three types of FPs (i.e., vocalic, vocalic-nasal, and nasal); (c) the placement of FPs in segmental context coded as SS, WW, SW, and WS; (d) their position within the utterance (start, mid, end, single); and (e) phonetic features (F_0 , F_1 , F_2 , duration, intensity), to examine phonetic differences in the realization of FPs.

3.1. Segmentation, Annotation, and Frequency Determination of FPs

To determine the frequency of FPs, all speech data were manually segmented and annotated using PRAAT (Boersma & Weenink, 2015). The most common filled pause types, i.e., vocalic, vocalic-nasal, and nasal were included. Less frequent but related phenomena were excluded, including vowel lengthening and lexical FPs. Each FP was identified and segmented by marking the onset and offset of each token based on auditory and visual inspection of the waveform and spectrogram. Annotation included labeling each token with its type (vocalic, nasal or vowel–nasal) and its context (e.g., surrounded by silence or speech). To distinguish FPs from segmental lengthening, we annotated only stand-alone hesitation tokens i.e., intervals showing a clear token boundary and not forming part of a neighboring lexical segment. The lower silent pause threshold was set at 150 ms. This detailed labeling allowed for accurate frequency counts per speaker and per category and enabled subsequent acoustic analysis. FP frequency was operationalized as FPs per 100 words of speech (FPs/100 words), calculated as $(\text{total FPs} \div \text{total words}) \times 100$. In addition, the presence or absence of creaky voice on FPs was manually coded.

To ensure consistency and reliability, a second trained annotator independently annotated the data. Inter-annotator agreement was calculated using Cohen's Kappa, which yielded a value of 0.89, indicating high reliability. Discrepancies were resolved through joint review and consensus, and annotation guidelines were refined where needed. This process ensured the accuracy of FPs identification and the validity of frequency counts used in the analysis. Only non-lexical, independently produced FPs were included in the

dataset to maintain a clear boundary between hesitation markers and lexical items. The proportion of each type of FP was calculated as a percentage of the total number of FPs produced by each participant. This allowed for an analysis of individual variations in the distribution of vowel, vowel–nasal, and nasal FPs.

All the FPs, vowel-only, vowel–nasal, and nasal, were analyzed across four immediate contextual environments (SS, WW, SW, and WS). SS positioning refers to FPs occurring between two silent pauses, where they are entirely enclosed by silence. WW positioning describes FPs that appear between two words, seamlessly integrated into speech flow. SW positioning includes FPs that follow a silent pause and precede a word, marking a transition from a pause to spoken content. Lastly, WS positioning refers to FPs that occur after a word and before a silent pause. These categorizations closely follow those proposed by O’Connell and Kowal (2005) and De Leeuw (2007).

In addition to immediate segmental context coding, each filled pause was also manually coded for its position within the utterance based on grammatical phrase boundaries (where “grammatical phrase” refers to a clause-level syntactic unit), adopting the four-position classification (start, mid, end, and single) from de Boer and Heeren (2020). FPs occurring between two grammatical phrases, with silent pauses of at least 150 ms on both sides, were classified as single. If a filled pause was adjacent to a silent pause on only one side and a grammatical phrase on the other, it was coded as either start or end, depending on its placement. FPs that interrupted a grammatical phrase were coded as mid-utterance, regardless of the length of surrounding silences, provided that the utterance remained grammatically fluent without the pause. Additionally, restarts, repairs, and repetitions were treated as new utterances.

3.2. Acoustic Measurements

Acoustic analysis involved measuring the duration of each filled pause and, for the vocalic element, fundamental frequency (F0), intensity, and first and second formants (F1, F2), using the PRAAT script³. Formant values (F1, F2) were extracted over the mid 50% of each vowel. Since all participants were female, the maximum frequency for formant extraction was set to 5500 Hz. The resulting values were then Lobanov-normalized. The duration of vocalic proportions (V) and nasal proportions (N) were extracted. The mean values for F0 and intensity were calculated over the full duration of each vowel.

A total of 21 tokens (11%) were excluded due to the presence of creaky voice, which can interfere with reliable estimation of acoustic measures, particularly F0 and formant frequencies. In addition, five FP tokens for which PRAAT could not extract reliable formant values were excluded. After these exclusions, a total of 193 tokens were retained for statistical analysis.

3.3. Statistical Analysis

We employed linear mixed-effects models to examine how FP type (vowel, nasal, vowel–nasal), contextual environment, and positional placement influence the acoustic characteristics of FPs. The acoustic features included log-transformed duration of FPs, F0, F1, F2, and intensity. For each acoustic feature, models were fitted treating it as the dependent variable. In the initial model, FP type was entered as the fixed effect.

In the second model, context and position were included simultaneously as fixed effects. Owing to the comparatively low frequency of vocalic–nasal and nasal FPs, inclusion of FP type together with segmental context and utterance position resulted in sparse cells and unstable estimates. The multivariate model was therefore restricted to vocalic (V) tokens, which constituted most of the data. All models included random intercepts for Speaker to account for repeated observations within speakers.

Segmental context captured the environment in which the FPs occurred and was classified as silence–silence (SS), silence–word (SW), word–silence (WS), and word–word (WW, used as the reference category). Utterance position was included as a second fixed effect (medial, final, single; initial as the reference category).

4. Results

4.1. Distribution of FPs

The results of this study reveal considerable variation in the use of FPs across speakers (see Figure 1). FP rates range from 0.00 FPs/100 words (e.g., Speakers 4 and 17) to values exceeding 2.0 FPs/100 words (speakers 9 and 15). Most speakers cluster below 1.5 FPs/100 words, indicating relatively low overall FP frequency, while a smaller subset of speakers exhibits markedly higher rates. This dispersion suggests that FP usage is strongly speaker-dependent.

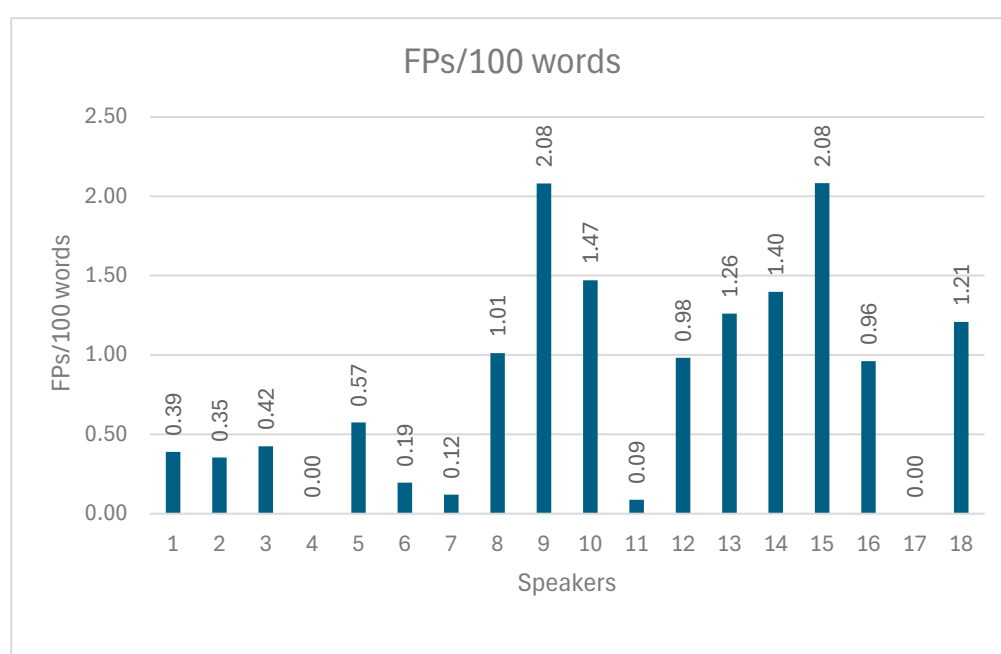


Figure 1. Number of FPs/100 words per speaker.

4.2. Distribution and Acoustic Characteristics of FPs Types

The analysis of FP types reveals a clear pattern, with vocalic forms occurring most frequently across speakers, accounting for 92.7% of all observed FPs. In contrast, vowel–nasal FPs occur far less frequently, representing 5.8%, while nasal FPs are rare, comprising only 1.6% of the total (see Figure 2). Even among speakers who used multiple types, vocalic forms remained dominant, making up more than 85–95% of their total FPs. This distribution reinforces that the dominant type in Urdu is the vocalic FPs.

4.3. Contextual Position of FPs

Descriptive analysis is reported for each FP form, indicating how tokens of a given form are distributed across the four positional contexts (SS, SW, WS, WW). For vocalic FPs (V), the SW (silence–FP–word) context contains the largest share of tokens (44.5%), followed by WW (word–FP–word) (33.5%). The remaining vocalic tokens are evenly split between SS (silence–FP–silence) and WS (word–FP–silence) (each 11.0%). For vowel–nasal FPs (VN), tokens cluster in SS (45.5%), with additional occurrences in WS (27.3%) and WW (18.2%), while SW is comparatively rare (9.1%). Nasal FPs (N) occur exclusively in the WW context, with 100% of nasal tokens found in word–FP–word environments (see Figure 3).

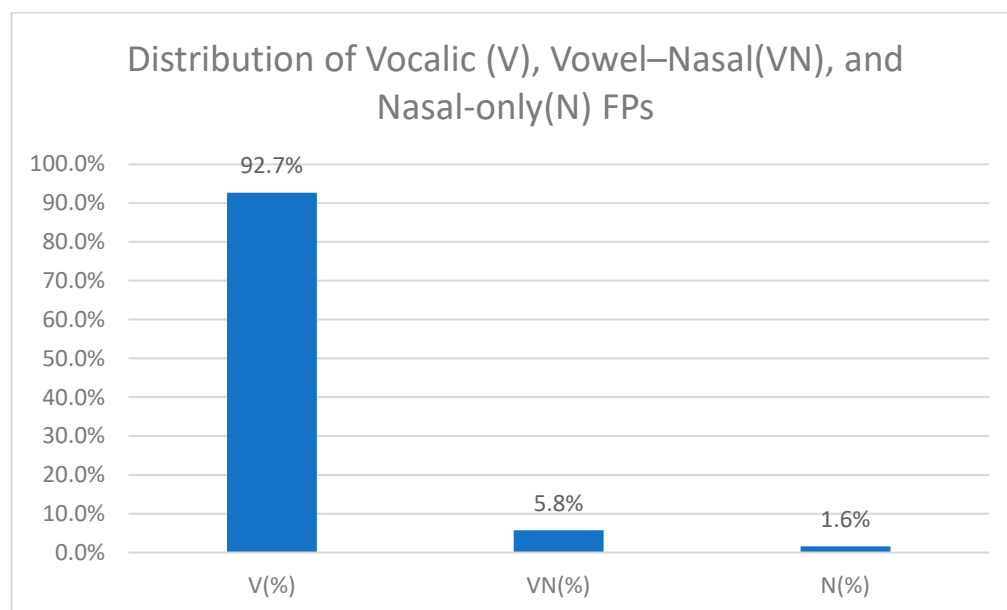


Figure 2. Distribution of Vocalic (V), Vocalic-Nasal (VN), and Nasal (N) FPs.

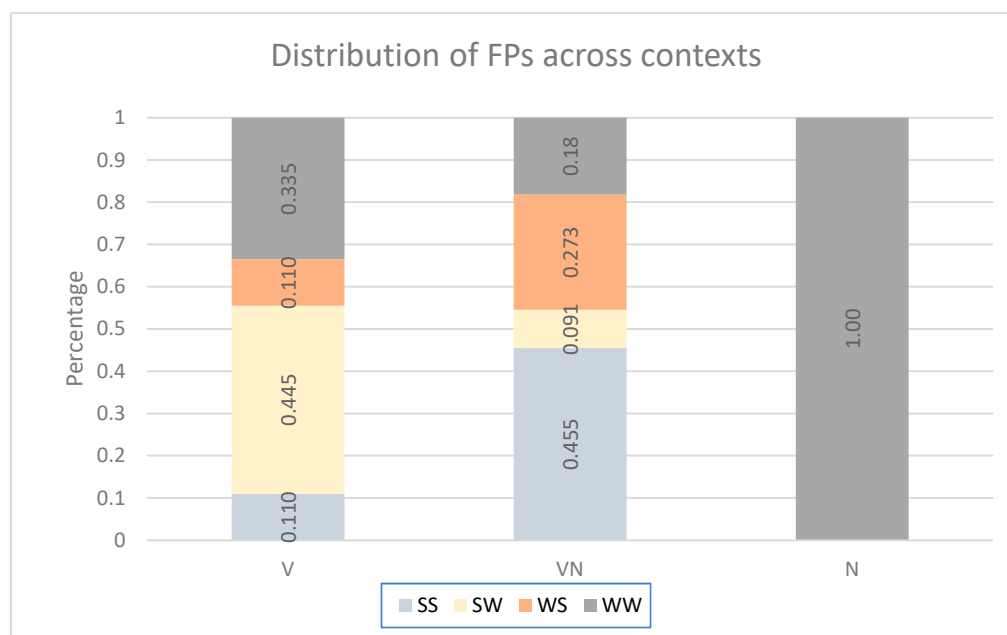


Figure 3. Distribution of FPs in four contexts (SS, WW, WS, SW).

4.4. FPs' Position in the Utterance

Figure 4 shows that vowel-only FPs were distributed across all four utterance positions, with medial position accounting for about 70%, initial for approximately 17%, final for around 8%, and single position for about 5% (see Figure 4). Vowel–nasal FPs also occurred predominantly in the medial position (approximately 91%), with only around 9% produced in the initial position, and none in final or single positions. Nasal FPs were used exclusively in the medial position. Overall, these findings indicate a strong preference for medial positioning across all FP types.

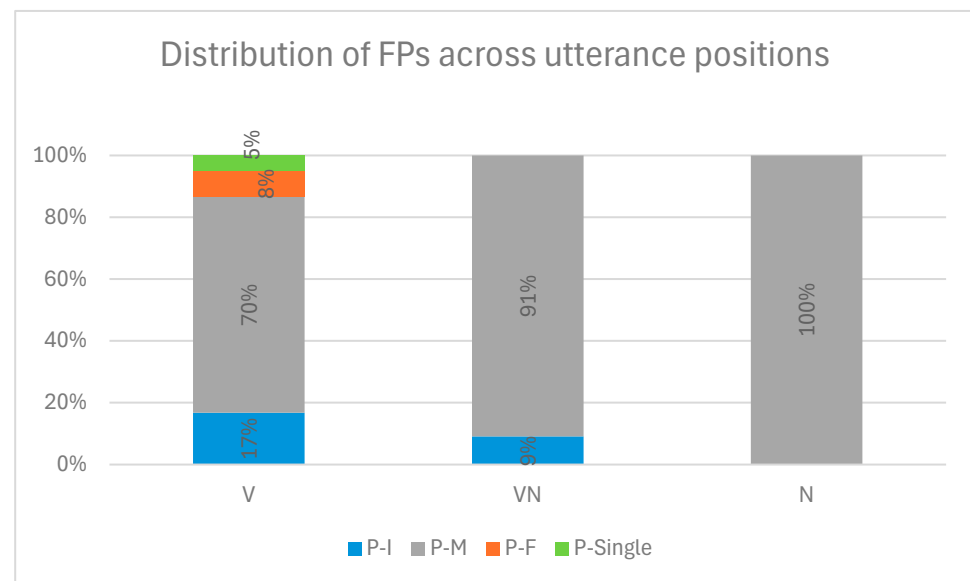


Figure 4. Distribution of FPs in utterance positions.

4.5. Acoustics of Filled Pauses in Urdu

Table 1 presents the effects of FP type (vowel-only vs. vowel–nasal) on the acoustic measures of FPs, with random intercepts for Speaker. Vowel-only filled pauses show a strong, significant positive effect on duration (logfp; $\beta = 0.569$, $p < 0.001$), indicating longer durations than the vowel portion of vowel–nasal tokens. FP type also shows a significant negative effect on F1 ($\beta = -0.645$, $p = 0.038$), indicating lower F1 values for vowel-only filled pauses, consistent with relatively higher vowel articulation than for vowel in vowel–nasal tokens.

Table 1. Linear mixed-effects model: FP type effects on acoustic measures.

	(1)	(2)	(3)	(4)	(5)
Variables	logfp	f0sem	LOB_F1	LOB_F2	LOB_Intensity
Vocalic	0.569 *** (0.000)	2.054 (0.432)	−0.645 * (0.038)	−0.412 (0.186)	−0.219 (0.488)
Constant	−1.704 *** (0.000)	16.236 *** (0.000)	0.922 ** (0.012)	0.543 (0.141)	−0.238 (0.525)
Observations	190	190	188	188	188
Number of groups	16	16	14	14	14

p-values in parentheses, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Note: Vocalic is a dummy variable, equals to 1 if FP is vocalic and 0 for non-vocalic FP. “n” is removed from the analysis.

Table 2 reports linear mixed-effects model estimates for vocalic-only FPs, testing the effects of segmental context (SW, WS, WW; with SS as the reference category) and utterance position (medial, final, single; with initial as the reference category) on each dependent variable (logfp, f0sem, LOB_F1, LOB_F2, and LOB_intensity). Random intercepts were included for speakers.

Across outcomes, segmental context effects are not statistically significant. In contrast, utterance position shows several significant associations: compared to initial position, medial tokens have shorter durations (logfp; $\beta = -0.209$, $p = 0.034$) and higher intensity ($\beta = 0.591$, $p = 0.048$); final tokens show lower F0 ($\beta = -7.405$, $p = 0.010$); and single tokens show shorter durations ($\beta = -0.274$, $p = 0.007$), lower F0 ($\beta = -5.639$, $p = 0.024$), and higher intensity ($\beta = 0.715$, $p = 0.020$). No significant effects are observed for LOB_F1 or LOB_F2.

Table 2. Linear mixed-effects model: context and position effects on acoustic measures (v tokens only).

	(1)	(2)	(3)	(4)	(5)
Variables	logfp	f0sem	LOB_F1	LOB_F2	LOB_Intensity
SW	0.099 (0.164)	−0.928 (0.596)	−0.287 (0.162)	−0.178 (0.385)	0.011 (0.958)
WS	0.039 (0.741)	2.320 (0.416)	0.090 (0.787)	0.176 (0.596)	0.213 (0.528)
WW	0.060 (0.688)	−5.495 (0.136)	−0.507 (0.260)	−0.355 (0.430)	0.613 (0.181)
Medial	−0.209 * (0.034)	−3.768 (0.120)	0.104 (0.725)	−0.021 (0.942)	0.591 * (0.048)
Final	0.091 (0.438)	−7.405 * (0.010)	−0.410 (0.237)	−0.605 (0.081)	0.519 (0.141)
Single	−0.274 ** (0.007)	−5.639 * (0.024)	−0.010 (0.972)	0.218 (0.471)	0.715 * (0.020)
Constant	−1.185 *** (0.000)	19.763 *** (0.000)	0.186 (0.580)	0.116 (0.730)	−0.618 (0.071)
Observations	179	179	177	177	177
Number of groups	16	16	14	14	14

p-values in parentheses, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

5. Discussion

This study examines the use and phonetic realization of FPs in spontaneous Urdu speech. The analysis was conducted at multiple levels: distributional, contextual, and acoustic, providing a detailed account of how FPs are patterned in the language. The results show a strong preference for vowel-only FPs over vowel–nasal and nasal forms. Previously, in semi-spontaneous Urdu dialogs, [Jabeen and Betz \(2022\)](#) reported uh-type FPs as the most frequent form (69%), with um-type FPs as the second most frequent (25%). Our spontaneous-speech data show an even stronger skew toward vowel-only FPs (92.7%), while vowel–nasal forms constitute only 5.8% of the dataset. This contrast may be attributed to differences between spontaneous and semi-spontaneous speech contexts. Similar patterns have been observed in other languages. For instance, Dutch shows a strong preference for vowel-only FPs (v), with um being used much less frequently ([De Leeuw, 2007](#); [Swerts, 1998](#)). This contrasts with languages such as English and German, which tend to favor vowel–nasal FPs ([De Leeuw, 2007](#)). These cross-linguistic differences reinforce the idea that FP realization is shaped by language-specific phonological and discourse conventions.

A distributional asymmetry emerges when FPs are examined with respect to their immediate segmental context. In Urdu, vowel-only FPs occur most frequently in SW and WW contexts, whereas vowel–nasal FPs, though rare overall, cluster primarily in SS and WS contexts, where the filler is followed by silence. This pattern indicates that vowel-only forms tend to precede lexical continuation, while vowel–nasal forms are more often associated with speech suspension. This asymmetry can be interpreted in articulatory terms. In contexts where speech is about to resume, vowel-only FPs may facilitate smoother transitions into the following word by avoiding velum lowering that could carry over into the upcoming segment. By contrast, in contexts where no immediate oral articulation is planned, the nasal component in vowel–nasal FPs may reflect reduced articulatory constraint during ongoing voicing. The association of vowel–nasal FPs with surrounding silence is consistent with previous findings in other languages ([Clark & Fox Tree, 2002](#); [Hughes et al., 2016](#)). Finally, nasal FPs appear exclusively in WW contexts and are restricted to a small number of speakers, (given their very low frequency in the present dataset, no firm conclusions can be drawn about the phonetic or functional status of nasal FPs; future work using larger and more diverse speech corpora will be necessary to assess their distribution and role). These findings indicate distinct contextual preferences across FP types.

Moreover, the results show that all types of FPs are most often placed at the medial position of utterance. Contrary to [de Jong's \(2016\)](#) claim that FPs are uncommon at mid-utterance in L1 speech, our data, consistent with findings from [de Boer and Heeren \(2020\)](#), show that medial positioning was the most frequent placement across all FP types. These positional patterns may suggest that, in Urdu, FPs are more closely tied to cognitive effort during lexical retrieval than to discourse structuring.

Furthermore, the results indicate that vowel-only FPs (V) are produced with longer durations and a relatively higher tongue position (lower F1) than the vowel portion of vowel–nasal FPs (VN). This duration asymmetry aligns with previous work showing that the vowel component in VN FPs tends to be shorter than in purely vocalic FPs ([Belz, 2021](#); [Hughes et al., 2016](#)). By contrast, pitch (F0), vowel backness (F2), and intensity show no reliable differences between the two types. Notably, our F1 result diverges from [Jabeen and Betz \(2022\)](#), who reported no F1 or F2 differences between the vowel components of V and VN FPs in their semi-spontaneous Urdu data. This discrepancy may reflect differences in speech style (spontaneous vs. semi-spontaneous), dataset size, or the distribution of VN tokens across speakers.

Unlike most prior studies, we examined the effects of utterance position and segmental context on the acoustic characteristics of FPs, since both factors could plausibly condition acoustic realization. These effects were examined for vocalic FPs only, as this is the predominant FP form in Urdu and the only type that occurs with sufficient frequency across multiple segmental contexts and utterance positions to permit reliable statistical modeling.

While descriptive analyses indicate that FPs may cluster in certain local environments, the mixed-effects models for vocalic FPs show that segmental context does not exert statistically significant effects on any acoustic parameter. This suggests that, for vocalic FPs in Urdu, the immediate phonetic environment (e.g., whether the filler is adjacent to silence or lexical material) does not independently shape duration or acoustic realization; instead, any contextual patterning is more plausibly distributional than phonetic conditioning.

In contrast to segmental context, utterance position emerges as a meaningful structural factor shaping the acoustic realization of vocalic FPs. With respect to duration, vocalic FPs occurring in medial position and in single position are significantly shorter than those produced in the utterance-initial position. This indicates temporal reduction when vocalic FPs are embedded within an ongoing utterance or produced in isolation, relative to utterance-initial FPs. Utterance position has also been shown to affect filler duration ([de Boer & Heeren, 2020](#)), but the single-position pattern differs: single FPs are longer in [de Boer and Heeren's](#) study, whereas single vocalic FPs are shorter than utterance-initial tokens in our data.

Utterance position also conditions fundamental frequency (F0). Vocalic FPs produced in final and single positions exhibit significantly lower F0 values than those in initial position. Lower pitch in these positions may reflect reduced prosodic prominence or diminished articulatory effort at points where speakers are closing an utterance or momentarily disengaging from lexical production. Finally, intensity shows positional sensitivity for vocalic FPs: tokens in medial and single positions display significantly higher intensity than those in the initial position. This increase may reflect tighter integration into the surrounding speech stream, particularly in the medial position where articulatory activity is continuous. No significant positional effects are observed for F1 or F2, indicating that vowel quality remains relatively stable across utterance positions.

We acknowledge that there are several limitations in the current study that should be addressed in future research. First, the speaker sample was limited to female participants, which may restrict the generalizability of the findings across gender. Second, the analysis focused on the most common filled pause types observed in the data (vocalic, vowel–nasal,

and nasal) and did not include other hesitation-related phenomena such as clicks, glottal FPs, or prolongations. Finally, the study examined speech production only. Future work will address these limitations by incorporating a more diverse speaker pool and by conducting listener-based perception experiments to directly examine how FPs are interpreted.

6. Conclusions

This study documents the distributional, contextual, and acoustic properties of FPs in spontaneous Urdu speech, providing a detailed, language-specific account of how these hesitation phenomena are patterned. The findings show a clear organization of FPs in Urdu, characterized by a strong dominance of vowel-only (vocalic) forms across speakers. Distributional analyses further suggest systematic contextual tendencies, with vowel-only FPs occurring most often before lexical continuation, whereas vowel–nasal FPs are more likely to be followed by silence. Across filled pause types, medial utterance position is strongly preferred, indicating that Urdu FPs typically occur within ongoing utterances rather than at utterance edges.

Acoustically, vowel-only and vowel–nasal FPs differ reliably in duration and vowel height (F1). For vocalic FPs, mixed-effects modeling shows that utterance position conditions their acoustic realization, whereas segmental context does not exert independent effects on duration or acoustic measures once the position is controlled. Overall, the results underline the value of documenting FPs within individual languages before drawing broader functional or theoretical conclusions. These findings pave the way for language-specific modeling of disfluency in Urdu speech technology, including recognition and synthesis.

Author Contributions: Conceptualization and data collection, S.Z. and M.A.M.; Formal analysis, Writing—original draft preparation, S.Z.; Writing—review and editing, H.-Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Review Committee of Government College University Faisalabad, Pakistan (Ref. No. GCUF/ERC/519; 18 December 2024).

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

Data Availability Statement: Dataset available on request from the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Notes

- ¹ In line with Swerts (1998) and Clark and Fox Tree (2002), FPs occurring with surrounding silence are interpreted as major delays (often associated with um), whereas FPs occurring between words are interpreted as minor delays (often associated with uh).
- ² Throughout the manuscript, the term nasal refers specifically to a nasal consonant (i.e., [m]/[n]) and not to nasalization of the vowel.
- ³ https://github.com/stylerw/styler_praat_scripts (accessed on 5 July 2025).

References

- Belz, M. (2021). *The phonetics of “äh” and “ähm”: Acoustic variation of filled particles in German..* Metzler. [CrossRef]
- Beňuš, Š. (2009, September 6–10). *Variability and stability in collaborative dialogues: Turn-taking and filled pauses.* INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association (pp. 796–799), Brighton, UK.
- Betz, S. (2020). *Hesitations in spoken dialogue systems* [Doctoral dissertation, Bielefeld University].
- Boersma, P., & Weenink, D. (2015). *Praat: Doing phonetics by computer* (Version 5.4) [Computer software]. University of Amsterdam. Available online: <http://www.praat.org/> (accessed on 20 January 2025).

- Candea, M., Vasilescu, I., & Adda-Decker, M. (2005, September 10–12). *Inter- and intra-language acoustic analysis of autonomous fillers*. Disfluency in Spontaneous Speech (DiSS 2005) (pp. 47–51), Aix-en-Provence, France.
- Cataldo, V., Schettino, L., Savy, R., Poggi, I., Origlia, A., Ansani, A., Sessa, I., & Chiera, A. (2019). Phonetic and functional features of pauses, and concurrent gestures, in tourist guides' speech. *Audio Archives at the Crossroads of Speech Sciences, Digital Humanities and Digital Heritage*, 6, 205–231.
- Christenfeld, N. (1994). Options and ums. *Journal of Language and Social Psychology*, 13(2), 192–199. [\[CrossRef\]](#)
- Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84(1), 73–111. [\[CrossRef\]](#) [\[PubMed\]](#)
- de Boer, M. M., & Heeren, W. F. (2020). Cross-linguistic filled pause realization: The acoustics of uh and um in native Dutch and non-native English. *The Journal of the Acoustical Society of America*, 148(6), 3612–3622. [\[CrossRef\]](#) [\[PubMed\]](#)
- de Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching*, 54, 113–132. [\[CrossRef\]](#)
- De Leeuw, E. (2007). Hesitation markers in English, German, and Dutch. *Journal of Germanic Linguistics*, 19(2), 85–114. [\[CrossRef\]](#)
- Eklund, R. (2004). *Disfluency in Swedish human–human and human–machine travel booking dialogues* [Doctoral dissertation, Linköping University].
- Finlayson, I. R., & Corley, M. (2012). Disfluency in dialogue: An intentional signal from the speaker? *Psychonomic Bulletin & Review*, 19, 921–928. [\[CrossRef\]](#) [\[PubMed\]](#)
- Fischer, K., Niebuhr, O., Novák-Tót, E., & Jensen, L. C. (2017, March 6–9). *Strahlt die negative Reputation von Häsitationsmarkern auf ihre Sprecher aus?* 43rd Annual Meeting of the German Acoustical Society (DAGA 2017) (pp. 1450–1453), Kiel, Germany.
- Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. Academic Press.
- Götz, S. (2013). *Fluency in native and nonnative English speech*. John Benjamins Publishing Company.
- Horváth, V. (2010). Filled pauses in Hungarian: Their phonetic form and function. *Acta Linguistica Hungarica (Since 2017 Acta Linguistica Academica)*, 57(2–3), 288–306. [\[CrossRef\]](#)
- Hughes, V., Wood, S., & Foulkes, P. (2016). Strength of forensic voice comparison evidence from the acoustics of filled pauses. *The International Journal of Speech, Language and the Law*, 23(1), 99–132. [\[CrossRef\]](#)
- Jabeen, F., & Betz, S. (2022, September 18–22). *Hesitations in Urdu/Hindi: Distribution and properties of fillers and silences*. Interspeech 2022 (pp. 3113–3117), Incheon, Republic of Korea.
- Jabeen, F., & Wagner, P. (2023, August 28–30). *Variability in hesitations in Punjabi semi-spontaneous narrative speech: An automatic clustering based analysis*. Disfluency in Spontaneous Speech (DiSS) Workshop 2023, Bielefeld, Germany.
- Kirjavainen, M., Crible, L., & Beeching, K. (2022). Can filled pauses be represented as linguistic items? Investigating the effect of exposure on the perception and production of um. *Language and Speech*, 65(2), 263–289. [\[CrossRef\]](#) [\[PubMed\]](#)
- Kjellmer, G. (2003). Hesitation. In defence of er and erm. *English Studies*, 84(2), 170–198. [\[CrossRef\]](#)
- Kosmala, L., & Crible, L. (2022). The dual status of filled pauses: Evidence from genre, proficiency and co-occurrence. *Language and Speech*, 65(1), 216–239. [\[CrossRef\]](#) [\[PubMed\]](#)
- Levelt, W. J. (1983). Monitoring and self-repair in speech. *Cognition*, 14(1), 41–104. [\[CrossRef\]](#) [\[PubMed\]](#)
- Levelt, W. J. (1993). *Speaking: From intention to articulation*. MIT Press.
- Lickley, R. J. (2015). Fluency and disfluency. In M. A. Redford (Ed.), *The handbook of speech production* (pp. 445–474). John Wiley.
- Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous English speech. *Word*, 15(1), 19–44. [\[CrossRef\]](#)
- Maekawa, K., & Mori, H. (2017). Comparison of voice quality between the vowels in filled pauses and ordinary lexical items. *Journal of the Phonetic Society of Japan*, 21(3), 53–62.
- Niebuhr, O., & Fischer, K. (2019, September). Do not hesitate!—Unless you do it shortly or nasally: How the phonetics of filled pauses determine their subjective frequency and perceived speaker performance. In *Interspeech 2019* (pp. 544–548). International Speech Communication Association.
- O'Connell, D. C., & Kowal, S. (2005). Uh and um revisited: Are they interjections for signaling delay? *Journal of Psycholinguistic Research*, 34, 555–576. [\[CrossRef\]](#)
- O'Shaughnessy, D. (1992, March). Recognition of hesitations in spontaneous speech. In *IEEE international conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. 521–524). IEEE Computer Society.
- Oviatt, S. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language*, 9(1), 19–36. [\[CrossRef\]](#)
- Paschen, L. (2023, August 28–30). *Filled pauses and false starts do not reliably preface longer or more complex utterances across typologically diverse languages*. Proceedings of the Disfluency in Spontaneous Speech (DiSS) Workshop 2023 (pp. 13–17), Bielefeld, Germany.
- Schnadt, M. J., & Corley, M. (2006, July 26–29). *The influence of lexical, conceptual and planning based factors on disfluency production*. Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 28, No. 28), Vancouver, BC, Canada.
- Shriberg, E. (2001). To 'errrr' is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31(1), 153–169. [\[CrossRef\]](#)
- Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies* [Doctoral dissertation, University of California].

- Shriberg, E. E., & Lickley, R. J. (1993). Intonation of clause-internal filled pauses. *Phonetica*, 50(3), 172–179. [[CrossRef](#)] [[PubMed](#)]
- Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language*, 32(1), 25–38. [[CrossRef](#)]
- Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics*, 30(4), 485–496. [[CrossRef](#)]
- Torreira, F., Adda-Decker, M., & Ernestus, M. (2010). The Nijmegen corpus of casual French. *Speech Communication*, 52(3), 201–212. [[CrossRef](#)]
- Tottie, G. (2016). Planning what to say: Uh and um among the pragmatic markers. In *Outside the clause* (pp. 97–122). John Benjamins Publishing Company.
- Tottie, G. (2020). Word-search as word-formation?: The case of “Uh” and “Um”. In *Crossing linguistic boundaries: Systemic, synchronic and diachronic variation in English* (pp. 29–42). Bloomsbury Academic.
- Watanabe, M., Hirose, K., Den, Y., & Minematsu, N. (2008). Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech Communication*, 50(2), 81–94. [[CrossRef](#)]
- Wieling, M., Grieve, J., Bouma, G., Fruehwald, J., Coleman, J., & Liberman, M. (2016). Variation and change in the use of hesitation markers in Germanic languages. *Language Dynamics and Change*, 6(2), 199–234. [[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.