Disfluency in Spontaneous Speech (DiSS) Workshop 2023
28-30 August 2023, Bielefeld, Germany

Variability in hesitations in Punjabi semi-spontaneous narrative speech: An automatic clustering based analysis

Farhat Jabeen, Petra Wagner

Bielefeld University, Germany
firstname.lastname@uni-bielefeld.de

Abstract

This research offers a first analysis of hesitations in Punjabi, an under-researched language, in conjunction with a cross-linguistic comparison. We show speaker related variation in the frequency of hesitations in Punjabi. Variability was also observed in the form of filled pauses which comprised vowels or vowel-consonant sequences with nasals or obstruents. The vowels in filled pauses differed based on their segmental context and individual speakers. Automatic clustering showed that (lexicalized) filled pauses were grouped by F0 register, instead of F0 contour. These results (1) have cross-linguistic significance and (2) provide insights for modeling hesitations in speech technological systems.

Index Terms: hesitation, filled pauses, lexicalized filled pauses, silent pauses, automatic clustering, semi-spontaneous, Punjabi

1. Introduction

In this article, we present a first analysis of hesitations in Punjabi semi-spontaneous narratives. [1] categorised hesitation as a type of disfluency and defined hesitation as a phenomenon "that temporally extends the delivery of the intended message for whatever reason." (p. 11). Among the hesitation types listed by [1], we focus on silent pauses (SP), filled pauses (FP), and lexicalized filled pauses (LFP). Far from being a barrier in communication, [2] argues that speech with hesitations is reflective of speakers' thought process during a communicative event. Their presence has also been associated with speech planning. [3] claims that speech genres such as narrative speech carry a large number of hesitations as recalling events from memory involves planning by speakers. Furthermore, hesitations are used when speakers are searching for a word [4] or aiming to focus some information [2]. Hesitations are also used by speakers to structure their discourse into different units [5, 6]. [7] reported that the presence and position of FPs in Dutch monologues was associated with major discourse boundaries.

The type and distribution of hesitations has been reported to be language and speaker specific. For example, the FPs used by Japanese speakers [8] differ from those used by speakers of Hungarian [9], English [10], German [11], Urdu/Hindi [12], and European Portuguese [13]. The position and duration of SPs has also been shown to be variable [14, 12]. Speaker based variation in the frequency and selection of hesitation types has been reported by [15, 16]. The frequency of hesitation clusters also varies within a language as shown by [17], who reported variation among individual speakers as well as hesitation types.

Hesitations vary in terms of their position in syntactic and prosodic units as well as in discourse structure. [18] showed that in Italian, silent and filled pauses were placed earlier in tonal units compared with lexicalized filled pauses and lengthenings. Moreover, LFPs were more often produced as tonal units on their own in comparison with filled and silent pauses. The occurrence of filled pauses also relates with the size of an utterance [6]. Accordingly, [19] found that the higher number of words in an utterance was related to the production of more hesitations at the beginning of an utterance.

Language specific variation is observed in the form of different hesitation types as well. For example, LFPs, being semantically bleached discourse markers [3], comprise the available discourse markers in a given language. As for FPs, the most frequently used filled pauses in English are vowel only or vowel-nasal sequences [10]. But FPs have been reported to consist of only a nasal consonant, a cluster of vowel-nasal-fricative, as well as a sequence of two vowels in Hungarian [9] and Urdu/Hindi [12]. Recently, [11] showed that FPs in German may consist of glottal stops as well as bisyllabic forms.

The vowels in FPs are generally transcribed as [ə] or /ɔ/ in English. However, the vowels used in FPs may differ in their quality. [20] showed that the FPs in English cannot be modelled after the centralized schwa only. [13] found similar results for European Portuguese and reported that FPs in that language are produced with either [ɨ:] or [ɐ:]. Similarly, [8] reported that the vowels in Japanese FPs differed from the vowels in other lexical items in terms of their duration, formants, intensity, and voice quality. This argues for a language specific analysis of the vowel quality of FPs that takes segmental context and speaker based differences into account.

Despite their being highly frequent and fulfilling a wide set of functions in discourse, hesitations and related phenomena are rarely taken into account in synthetic speech or dialogue system modeling. This is somewhat surprising given their manifest benefit for improving task success and the perception of system reactivity in HCI [21]. However, it is also explicable by the dilemma faced by human and synthetic speech alike. Despite its benefits for conversation, listeners tend to perceive hesitations in human and synthetic speech as less fluent [22, 23]. We argue that a thorough understanding of the form, function, distribution, and perception of these phenomena across languages is needed to successfully exploit the conversational benefits of hesitations in existing speech technology [24, 25] without compromising its quality. Likewise, conversational speech technology should be able to further inform phonetic models [26].

To this day, most of the existing research on hesitations is concerned with the European languages, and South Asian languages have been largely neglected. To our knowledge, [12] is the only systematic investigation of hesitations in Urdu/Hindi, a South Asian language, while there is no analysis of hesitations in Punjabi so far. In the current study, we aim to fill this gap and analyze the distribution and properties of hesitations in Punjabi semi-spontaneous narrative speech. Our analysis is concerned with silent pauses, filled pauses, and lexicalized filled pauses as we address the following research questions:

1. Are there speaker based differences in the form and distribution of hesitations in semi-spontaneous narrative

10.21437/DiSS.2023-15

speech?

2. What is the frequency of different types of hesitations?
3. Do hesitation types differ in terms of their position?
4. If found in semi-spontaneous narratives,
   (a) what types of FPs and LFPs are used in Punjabi?
   (b) what is the F0 contour of FPs and LFPs in Punjabi?
   (c) what is the quality of vowels in FPs compared with the quality of vowels in other lexical items?

To investigate these, we used a corpus of semi-spontaneous Punjabi narrative speech. The details are as follows.

2. Methodology

2.1. Corpus & annotation

We recorded semi-spontaneous narrative speech (42 minutes) using Zoom[1]. The narratives were produced by sixteen (nine male) speakers of Punjabi. The participants were shown a Punjabi animation story on YouTube (The Hungry Rat, 4:28 minutes). At the end of the video, they were asked to retell the story in their own words providing as much detail as they could remember. The first author was always present in the Zoom session to give the impression of an audience.

For the analysis, we divided each narrative into Inter-Pausal Units (IPUs) separated by a silent pause of at least 150ms. Each IPU was manually annotated for hesitation type, position of hesitations (IPU initial, medial, final, across IPU boundaries, FPs and LFPs produced as IPUs), and the segmental content of filled pauses (vowel, nasal, obstruent, irregular phonation), as well as LFPs. Apart from the manual annotation of hesitations' position in IPUs, we also measured their position over the course of a narrative. As the narratives by different speakers varied in their duration, we normalized it using the following: (start of hesitation - start of narrative) / narrative duration

2.2. Acoustic & statistical analysis

The filled pauses consisting of only a vowel were analyzed for their quality. The first two formants were extracted in the middle of each vowel with a Praat [27] script. To compare the quality of vowels in FPs with that of other vowels in Punjabi, we extracted F1 and F2 for the vowels produced in lexical items by each speaker in our data (n=6092). The resulting formant values were Lobanov normalized (grouped by vowels & speakers) using the 'PhonR' package [28] in R.

To analyze the F0 contour of LFPs and vowel-only FPs, each hesitation type was divided into five equidistant points using [29]'s script. To achieve its dynamic transition, F0 (semitones re 1Hz) was measured at each of those time points.

For statistical analysis , we ran Linear Mixed Effects Regression [30] in R [31] to analyze the difference in the log duration[2] of different FPs. We used filler type as a fixed factor and participants as random effect. P-values for multiple comparisons were adjusted using the Tukey method. A similar analysis was carried out for the position of hesitations over the time course of narratives. In the next section, our results are discussed with reference to existing analyses in other languages.

[1] https://www.zoom.us/

[2] Log duration was used to account for tempo differences.

3. Results & discussion

3.1. Hesitation frequency

The frequency of hesitations produced by different speakers in our data is given in Table 1. It illustrates that the participants varied in their production of absolute number of hesitations as well as in their relative frequency. For example, speakers 12 and 13 used a high number of hesitations per minute compared with speakers 11 and 14. Individual variation in the frequency of hesitations has also been reported in other languages such as Italian [15], English [16], German [11], and French [14].

Table 1: *Individual variation in frequency of hesitations.*

Speaker	Speech time (m)	N. Hes	N. Hes/m
1	3.28	93	28
2	1.8	42	23
3	3.0	36	12
4	2.5	52	21
5	4.1	60	15
6	3.7	100	27
7	2.7	42	16
8	2.1	46	22
9	2.9	64	22
10	2.4	43	18
11	2.1	18	9
12	2.4	79	33
13	2.3	78	34
14	1.4	6	4
15	3.5	88	25
16	1.6	32	20

3.2. Hesitation types

Analysis showed that in our data, silent pauses were the most frequent type (59%), followed by filled pauses (24%), and lexicalized filled pauses (17%). Figure 1 illustrates that while individual speakers varied in their production of FPs and LFPs, they produced silent pauses as the most frequent hesitation type. Only speakers 9 and 11 deviated from this as they produced these hesitation types with almost similar frequency.

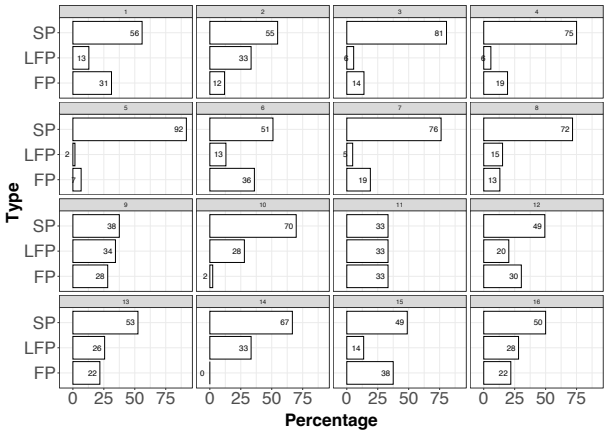

Figure 1: *Percentage of different hesitation types.*

3.3. Position of hesitations in IPUs

Table 2 offers the frequency of different hesitations at various positions in IPUs. As the IPUs were defined on the basis of pause duration, SPs was the most frequently found phenomenon across IPU boundaries. They are excluded from this Table as that data was not informative. Table 2 illustrates that Punjabi speakers have clear preferences for placing different hesitations at specific positions in IPUs. While LFPs were mostly placed at the initial position, the FPs were preferred at the medial position. Although almost a third of them were used at the IPU initial position as well. Moreover, LFPs were produced more frequently as IPUs compared with the filled pauses. [18] had reported a similar pattern for LFPs in Italian.

Table 2: *Occurrence (%) of hesitation types in IPUs.*

Hesitation	Position			
	Initial	Medial	Final	Unit
FPs	30	**59**	6	5
LFP	**69**	6	10	15

Figure 2 presents the production of different hesitations over the course of narratives. It shows that more hesitations were produced at the beginning of the narrative and speakers became more fluent later in their discourse. The production of hesitations towards the end of the narratives indicates that they were also used to perform other functions such as structuring discourse and prosodic units. No difference was found between hesitation types produced over a narrative.

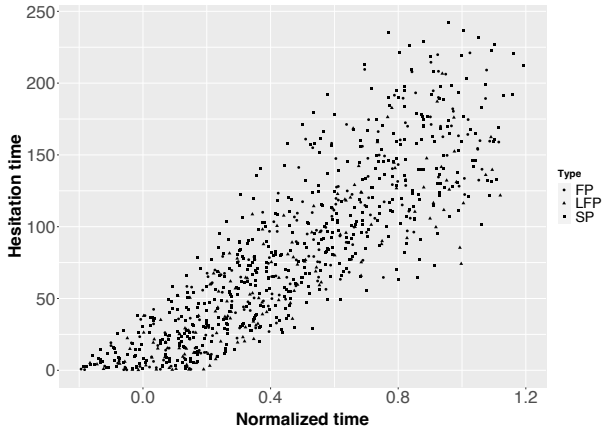

Figure 2: *Position of hesitations over the course of narratives.*

3.4. Form of pauses

3.4.1. Lexicalized filled pauses

We found that 'ʈe̪' (then) was the most frequently used LFP in our data. As the narratives involved a temporal organization of events, the use of this LFP is understandable. However, 'ʈe̪' was not used to indicate the temporal organization of events only. This claim is supported by the fact that 'ʈe̪' was frequently preceded or followed by a silent pause (15%). Moreover, some instances of 'ʈe̪' were combined with a temporal adverbial 'o.ʈo bɑ:ɖ̪' (after that). This shows that 'ʈe̪' was stripped of its lexical meaning and used as an LFP.

The second LFP used in our data was 'ət∫.tʃʰɑ' (ok), invariably produced at the beginning of a narrative. The use of 'ok' as an LFP at the beginning of narrative speech has been reported by [2] for English.

Figure 3 illustrates that most of the LFPs were produced with a level F0 contour at varying F0 registers. However, speakers 4 and 15 produced a few dynamic contours with a rising or a falling F0. We ran an automatic clustering algorithm [29] to analyze the F0 of LFPs and found that an analysis with three clusters had the lowest information cost. Cluster 1 consisted of a low falling F0 contour (n=99). Cluster 2 comprised a similar F0 contour but produced at a lower F0 register (n=25). The third cluster contained a high falling contour (n=7). This indicates that the clustering of LFPs was based on F0 register, instead of their F0 contour.

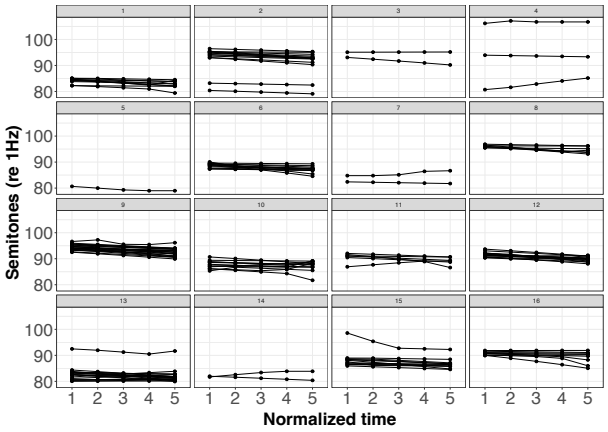

Figure 3: *Speaker & item based variation in the F0 of LFPs.*

3.4.2. Filled pauses

The analysis of filled pauses showed that they mostly consisted of a single vowel 'V' (n=113, 56%). However, we found a few instances of FPs with nasals (12%) that consisted of either an elongated nasal [m:] (n=11), a nasal-vowel sequence [m:V, mV] (n=5), or vowel-nasal combination [Vm] (n=8). These were collectively labeled as (V)N filler type. We also found a few vowel-obstruent sequences [Vh, kV, bV, V.hV, V.kV, V.vV, V.zV] (n=7, 3%). Furthermore, many filled pauses consisted of only irregular phonation (n=58, 29%). This inventory corresponds to [11]'s proposal of vocalic and glottal fillers covering a continuum of single to polysyllabic forms. The regression analysis showed that the FPs with irregular phonation differed significantly in their log duration from other filled pauses. FPs with irregular phonation were significantly shorter than the (V)N filled pauses (β = -0.63, SE = 0.1, t = -4.4, p = 0.0001) as well as the FPs with single vowels (β = -0.40, SE = 0.09, t = -4.2, p = 0.0002).

The low frequency of vowel-nasal fillers in Punjabi is surprising as [əm] was the second most frequently used FP in Urdu/Hindi [12], a closely related South Asian language. Further investigation of vowel-nasal FPs in our data showed that 'Vm' was mostly followed by a lexical word beginning with a bilabial nasal as shown in (1). This leads us to claim that the instances of VN filled pauses in Punjabi resulted from their segmental context and the FPs with a single vowel or irregular phonation are the mostly frequently used ones in this language. Future research should analyze the functions of these FPs to further investigate this.

(1) V**m: m**eri salgira vaa
'<FP> it's my birthday.'

3.4.3. Vowel quality in FPs

Figure 4 shows the F1 and F2 of vowel-only FPs in comparison with the formants of other vowels produced by the same speakers in our data. A first glance shows that the mean formant values of vowels in FPs are clustered around schwa. However, there is variation in the quality of these vowels as they are also produced as fronted or back. Although not to the same extent as observed here, [13] had also reported variation in the vowel quality of FPs in European Portuguese. Unlike [13]'s data, our analysis is based on FPs consisting of vowels only. Therefore, this variation in their vowel quality cannot be explained on the basis of filler type. Further analysis revealed two sources of this variability: Speaker related variation and coarticulation. In the latter context, the quality of vowels in FPs was assimilated with that of the vowels in the preceding or following lexical words.

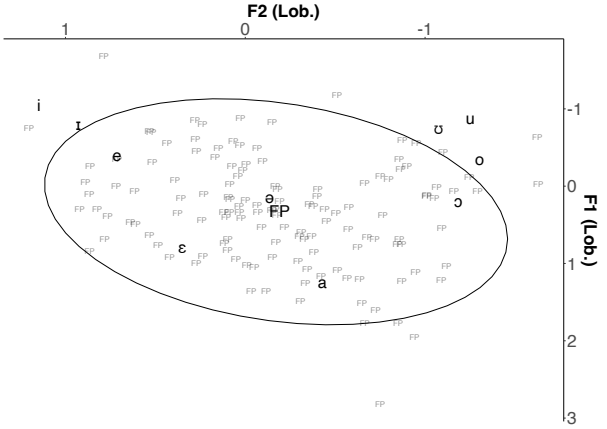

Figure 4: *'FP' in bold face shows mean formant values of vowels in FPs. Mean formants of vowels produced in lexical items are indicated with corresponding IPA labels.*

Speaker based variability in the vowel quality of FPs is illustrated in Figure 5. It shows that speaker 12 produced most of the FPs with a central vowel. Comparatively, the FPs produced by speaker 5 and 10 are highly variable. This data illustrates the challenges involved in the synthesis of natural sounding FPs.

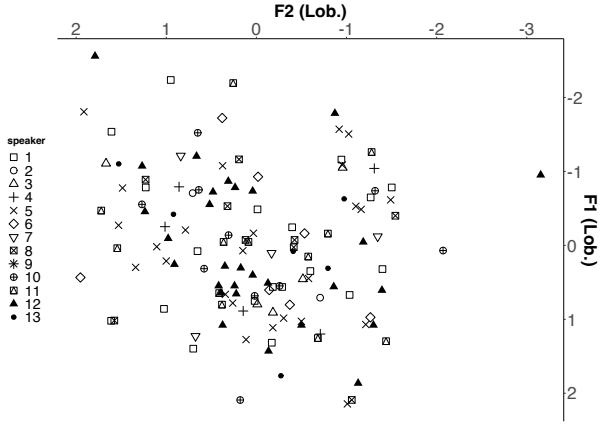

Figure 5: *Individual variation in vowel quality of FPs.*

3.4.4. F0 contour of FPs

Unlike their vowel quality, the F0 contour of FPs follows a regular pattern. Figure 6 illustrates that most of the FPs were produced with a level F0 contour. However, speakers 7 and 13 produced FPs with a rising or a falling F0. Variable F0 contour for FPs in German has been reported as a turn management cue [32]. As our data consists of narratives, the variation in Punjabi FPs may not be explained as a turn management device.

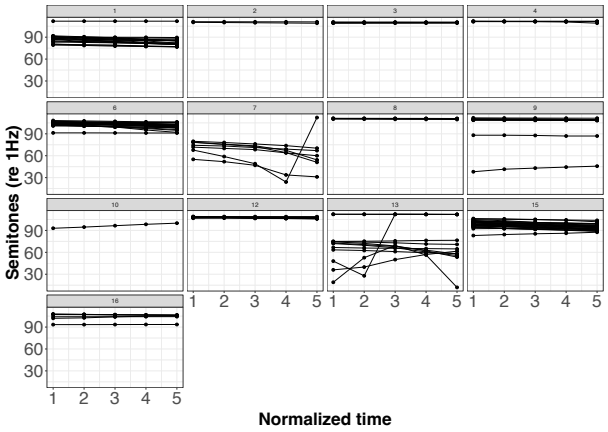

Figure 6: *Speaker & item based variation in the F0 of FPs. Three speakers did not produce any FPs with vowels.*

The clustering analysis showed that the F0 contour of FPs can be divided into three clusters with lowest information cost: Falling with low F0 register (n=20), falling with a mid level register (n=74), and a falling contour with high F0 register (n=25). Overall, the clusters for both FPs and LFPs were distinguished on the basis of their F0 register. As the F0 has been converted into semitones, this difference may not be attributed to speakers' gender. Considering that Punjabi is a tonal language [33], the restricted use of F0 in hesitations is understandable.

4. Summary & conclusion

The current study presents evidence for speaker based variation in the form and distribution of hesitations in Punjabi semi-spontaneous narratives. The variation in the position of hesitations shown in our data has also been reported for Italian [15], German, and Dutch [14]. Punjabi speakers' use of variable vowel quality in FPs is similar to that reported for European Portuguese [13]. However, Portuguese speakers only switched between two central vowels. This shows that although the form of Punjabi FPs follows the patterns reported for other languages, the extent of variation in the acoustic realization of filled pauses exhibited in our data has not been reported in detail before.

Our findings make a twofold contribution. They add to the ecological diversity of the field by offering an insight into the use of hesitations in an under-researched language. Our results also contribute to understanding the sources of variation in the form and structure of hesitations reported in existing literature [34, 16, 15]. These findings may be used to improve the modelling of hesitations in speech technological systems.

Importantly, our analysis is based on narrative speech in Punjabi. Given that their use is context specific, it is expected that the distribution and form of hesitations in Punjabi would differ on the basis of genre. In future, we aim to investigate hesitations in interview style speech as well.

5. References

[1] S. Betz, "Hesitations in Spoken Dialogue Systems," Ph.D. dissertation, 2020. [Online]. Available: https://nbn-resolving.org/urn:nbn:de:0070-pub-29422545, https://pub.uni-bielefeld.de/record/2942254

[2] W. Chafe, "Some reasons for hesitating," *Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler*, pp. 169–180, 1980.

[3] G. Tottie, "Planning what to say: Uh and um among the pragmatic markers," in *Outside the Clause*. John Benjamins, 2016, pp. 97–122. [Online]. Available: https://www.jbe-platform.com/content/books/9789027266552-slcs.178.04tot

[4] ——, "Word-search as word-formation? The case of uh and um," in *Crossing linguistic boundaries: Systemic, synchronic and diachronic variation in English*. Bloomsbury Publishing, 2020, pp. 29–42.

[5] J. Ginzburg, R. M. Fernández, and S. David, "Disfluencies as intra-utterance dialogue moves," *Semantics and Pragmatics*, vol. 7, no. 9, p. 64, 2014.

[6] M. Watanabe and Y. Korematsu, "Factors affecting clause-initial filler probability in an English monologue corpus," *Journal of the Phonetic Society of Japan*, vol. 21, no. 3, pp. 24–32, 2017.

[7] M. Swerts, "Filled pauses as markers of discourse structure," *Journal of Pragmatics*, vol. 30, no. 4, pp. 485–496, 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378216698000149

[8] X. Li, C. T. Ishi, C. Fu, and R. Hayashi, "Prosodic and voice quality analyses of filled pauses in Japanese spontaneous conversation by Chinese learners and Japanese native speakers," in *Proceedings of Speech Prosody 2022*, 2022, pp. 550–554.

[9] V. Horváth, "Filled pauses in Hungarian: Their phonetic form and function," *Acta Linguistica Hungarica*, vol. 57, no. 2–3, pp. 288–306, 2010.

[10] H. H. Clark and J. E. Fox Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[11] M. Belz, "Defining filler particles: A phonetic account of the terminology, form, and grammatical classification of "Filled Pauses"," *Languages*, vol. 8, no. 1, p. 57, 2023.

[12] F. Jabeen and S. Betz, "Hesitations in Urdu/Hindi: Distribution and properties of fillers & silences," in *Proceedings of Interspeech 2022*, 2022, pp. 4491–4495.

[13] H. Moniz, A. I. Mata, and M. C. Viana, "On filled-pauses and prolongations in European Portuguese," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[14] S. Betz, N. Bryhadyr, L. Kosmala, and L. Schettino, "A crosslinguistic study on the interplay of fillers and silences," *Proceedings of DiSS 2021*, 2021.

[15] L. Schettino, "The role of disfluencies in Italian discourse. Modelling and speech synthesis applications," Ph.D. dissertation, Universitá Degli Studi di Salerno, Italy, 2021.

[16] K. McDougall and M. Duckworth, "Profiling fluency: An analysis of individual variation in disfluencies in adult males," *Speech Communication*, vol. 95, pp. 16–27, 2017.

[17] D. C. O'Connell and S. Kowal, "Uh and um revisited: Are they interjections for signaling delay?" *Journal of Psycholinguistic Research*, vol. 34, pp. 555–576, 2005.

[18] L. Schettino, S. Betz, and P. Wagner, "Hesitations distribution in Italian discourse," in *DiSS 2021*, 2021.

[19] E. Shriberg, "Preliminaries to a theory of speech disfluencies," *Ph D. thesis University of California*, 1994.

[20] R. Dall, M. Wester, and M. Corley, "The effect of filled pauses and speaking rate on speech comprehension in natural, vocoded and synthetic speech," in *Proc. Interspeech*, 2014, pp. 56–60.

[21] S. Betz, B. Carlmeyer, P. Wagner, and B. Wrede, "Interactive hesitation synthesis: modelling and evaluation," *Multimodal Technologies and Interaction*, vol. 2, no. 1, 2018.

[22] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "Spontaneous conversational speech synthesis from found data." in *Proceedings of Interspeech*, 2019, pp. 4435–4439.

[23] O. Niebuhr and K. Fischer, "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 *Proceedings of Interspeech*, 2019, pp. 544–548.

[24] K. Horii, M. Fukuda, K. Ohta, R. Nishimura, A. Ogawa, and N. Kitaoka, "End-to-end spontaneous speech recognition using disfluency labeling," in *Proceedings of Interspeech 2022*, 2022, pp. 4108–4112.

[25] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "How to train your fillers: uh and um in spontaneous speech synthesis," in *The 10th ISCA Speech Synthesis Workshop*, 2019.

[26] A. Kirkland, H. Lameris, E. Székely, and J. Gustafson, "Where's the uh, hesitation? the interplay between filled pause location, speech rate and fundamental frequency in perception of confidence," in *Proceedings of Interspeech*, 2022, pp. 18–22.

[27] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program, [v. 6.3]," 2022, available at http://www.praat.org/ [retrieved 15.11.2022].

[28] D. R. McCloy, *phonR: tools for phoneticians and phonologists*, 2016, r package version 1.0-7.

[29] C. Kaland, "Contour clustering: A field-data-driven approach for documenting and analysing prototypical f0 contours," *Journal of the International Phonetic Association*, pp. 1–30, 2021.

[30] H. Baayen, D. J. Davidson, and D. M. Bates, "Mixed-effects modeling with crossed random effects for subjects and items," *Journal of Memory and Language*, vol. 59, no. 4, pp. 390–412, 2008.

[31] R Core Team, *R: A language and environment for statistical computing[v. 3.6.1]*, R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org/

[32] M. Belz and U. D. Reichel, "Pitch characteristics of filled pauses in spontaneous speech," in *Proceedings of DiSS 2015*, 2015. [Online]. Available: http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-25259-7

[33] R. K. Dhillon, "Stress and Tone in Indo-Aryan languages," Doctoral dissertation, Yale University, 2010.

[34] R. J. Lickley, "Fluency and disfluency," in *The Handbook of Speech Production*, M. A. Redford, Ed. John Wiley & Sons, Inc, 2015, pp. 445–474.