12th edition of the Disfluency in Spontaneous Speech Workshop (DiSS 2025)
4-5 September 2025, Lisbon, Portugal

The Effects of Spontaneous Speech on Disfluencies Assessment of Spanish Speakers with Down Syndrome

David Fernández-García[1], *César González-Ferreras*[1], *Valentín Cardeñoso-Payo*[1,2], *David Escudero-Mancebo*[1,2], *Mario Corrales-Astorgano*[1]

[1]ECA-SIMM Research Group, Universidad de Valladolid, Spain
[2]AI Center, Universidad de Valladolid, Spain

david.fernandez@uva.es,cesargf@uva.es,valentin.cardenoso@uva.es,
escuderomancebo.david@uva.es,mario.corrales@uva.es

Abstract

The aim of this study is to investigate the phonetic and fluency characteristics of spontaneous speech produced by Spanish speakers with Down syndrome (DS) compared to nonspontaneous speech modes (read, elicited and imitation) and assess the impact of these differences both on expert speech quality assessment and on automatic speech recognition (ASR) performance. The PRAUTOCAL corpus includes four different speech generation modes of utterances spoken by people with DS. The results show that there are minor differences in some features between spontaneous speech and other modes, but specific types of disfluencies and phonetic errors are more prevalent in spontaneous speech. The Whisper model showed improved performance on spontaneous speech, achieving a significantly lower Word Error Rate (WER) and fewer substitution errors. The Wav2Vec phoneme recognition model performed significantly worse, showing higher phoneme error rate (PER), more substitutions, and greater total errors, no matter the automatic segmentation tool used (MFA or WebMAUS).

Index Terms: spontaneous speech, disfluencies, Down syndrome

1. Introduction

Individuals with Down syndrome (DS) typically present speech and language disorders that involve both articulation impairments and prosodic deficits [1, 2]. Their atypical prosodic production, affecting rhythm, intonation, and stress, often leads to their speech being perceived as different by typically developing listeners [3]. Specifically, they show difficulties in producing prosodic functions related to turn-end, chunking, and focus, and the acoustic features they use for these functions are often less informative than those used by their typically developing peers [4]. Furthermore, individuals with DS may have trouble with articulatory control, such as articulating declination in declarative sentences and using appropriate pausing for chunking [5].

Additionally, speech disfluencies, including stuttering-like disfluencies, are reported to be more common in children with DS than in typically developing children [6]. This study indicates that approximately 30 percent of children with DS between 3 and 13 years of age may stutter, which is notably higher than the prevalence in typically developing children, and the speech of children with DS shows a different distribution of types of disfluencies compared to typically developing children [6]. The types of disfluencies observed can differ, with a higher occurrence of blocks and interjections reported in some studies [6, 7].

Regarding spontaneous speech, a wide range of expressions, including varied grammatical structures, linguistic variations, and notably, disfluencies such as pauses, repetitions, and sound prolongations, are common due to the unplanned nature of the discourse. While much research on speech, especially prosody, often controls the linguistic content with pre-established sentences, studies focusing on the analysis of prosodic characteristics in spontaneous speech are less common [8, 9]. One of these studies analyzes that, during spontaneous communication, pre-school children show problems to express prosodic contours in interrogative sentences [8].

Speech disfluencies significantly hinder the accuracy of deep learning-based transcription models. As noted in this study [10], while the intended speech word error rate (isWER) for Whisper was comparable between typical speech and speech from people who stutter, Whisper transcribed filled pauses and partial words at higher rates in the latter, and the isWER increased with stuttering severity. Furthermore, both text-based (BERT) and audio-based (Whisper encoder) disfluency detection models showed decreased performance when evaluated on speech with disfluencies from people who stutter compared to typical speech [10]. As highlighted previously, individuals with DS also exhibit more frequent and potentially different types of disfluencies in their speech, which would similarly pose challenges for these models attempting to produce accurate transcriptions [11].

Therefore, this study aims to investigate the phonetic and fluency characteristics of spontaneous speech produced by Spanish speakers with DS, comparing it directly with nonspontaneous speech from the same individuals. Specifically, we seek to answer the following research questions:

- Is the spontaneous speech of Spanish speakers with Down syndrome significantly different, with respect to phonetics and fluency, compared to their nonspontaneous speech?
- Do these potential differences significantly affect the performance of deep learning-based ASR models?

The structure of the paper is as follows. Section 2 details the methodology used, Section 3 presents the results of our analyzes, Section 4 discusses the implications of these findings, and Section 5 concludes the paper.

2. Methodology

2.1. Data

This study is based on the PRAUTOCAL corpus [12], a Spanish corpus of Down syndrome speech comprising 120 minutes of recordings. The corpus includes four production modes: *read* (R), *elicited* (E), *imitation* (I), and *spontaneous* (S). This *production modes* are not equally distributed. In fact, the vast majority of audios belonged to the *Read* class, and a small minority belonged to the *Spontaneous* class (see Table 1). In our

10.21437/DiSS.2025-15

analyses, we consider only a binary grouping: nonspontaneous (read, elicited, and imitated) versus spontaneous. This choice is motivated by our goal of testing whether any differences exist between spontaneous and nonspontaneous speech.

The corpus also includes disfluency transcriptions, which annotate the presence of fillers, interruption points, and editing terms throughout the utterances. These annotations serve as a key input for our experiments.

In addition, we incorporate a categorical, multi-aspect phonetic and fluency analysis, conducted by a professional linguist. The linguist listened to all the recordings and annotated them using a predefined rubric. For full details of the annotation guidelines and individual descriptions of each metric, see [13].

Phoneme segmentation was obtained using two automatic tools, Montreal Forced Aligner (MFA) [14] and WebMAUS [15]. As these segmentations were not manually corrected, using both systems helps to mitigate potential tool-specific segmentation errors.

2.2. Speech Annotation and Statistical Analysis

To address our research questions, we conducted statistical analyses comparing the two-group configurations across a variety of variables. We applied either parametric (*Student's t-test*) or nonparametric (*Mann-Whitney*) significance tests, depending on the normality of the data, in order to find if any significant difference really exists. We used a significance threshold of $\alpha = 0.05$.

- **Disfluency annotations**: We computed the mean number of disfluency errors (filler and editing terms) per production mode. For this work, we discarded the *interruption points* (that can be seen in Table 10 of [12]) because they were obtained by comparing the disfluency transcription with the ground truth transcription, which obviously does not exist for spontaneous speech. We saw that spontaneous utterances tend to be longer than the others. To solve that, we applied a normalization at utterance level, dividing each individual utterance value (e.g. number of fillers of one utterance) by the length (number of words) of the utterance itself.
- **Professional linguist analysis**: As explained before, a professional linguist has listened and analyzed all the audios of the speakers with DS. Here we can distinguish three different analysis:
  - **General assessment**: Each utterance was assigned a general score (1 to 3) based on phonetic and fluency criteria.
  - **Fine-Grained Fluency analysis**: The linguist also performed a fine-grained fluency analysis on different fluency aspects. For each aspect, she annotates if the error occurs zero, one or more than once, as a categorical annotation. We distinguish the following errors: blocks, prolongations, sound repetition, word repetition, and interjections.
  - **Phonetic word analysis**: Finally, the linguist also analyzed each word of each utterance individually and annotated errors at phoneme level. These errors are substitutions, omissions, distortions, and additions of phonemes in an individual word. Similar to what we have seen above, the linguist annotates, if the error occurs or not, in a binary way, for each different variable mentioned before. We also calculate a *total errors* variable, by adding all the error variables of the word.
- **Automatic phonetic segmentation**: We used the segmentations obtained by Montreal Forced Aligner and WebMAUS in two different ways. First of all, we use them to calculate the mean duration by phoneme, for each production mode. Then, we used the phonetic transcription as ground truth for one of the deep learning models we evaluated.
- **Deep learning-based transcriptors results**: We evaluated two different types of deep learning transcriptors. First, we measure how *Whisper* [16] performance differs along the different production modes. For that experiment, we used *whisper-large-v3* and disfluency transcriptions as ground truth. In order to make a fair comparison with the *Whisper* output (which does not contain disfluency annotations), we apply the same preprocessing pipeline as the one seen in [17].

  The second model we evaluated was *Wav2Vec* fine-tuned for phone recognition task [18]. Specifically, we used the *wav2vec2-lv-60-espeak-cv-ft*, which has been self-pretrained on *LibriSpeech* [19], and then fine-tuned with *Common Voice* [20]. In this case, as mentioned above, we use the phone transcriptions obtained by Montreal Forced Aligner and WebMAUS as ground truth. No adjustment in phoneme dictionaries has been made between *Wav2Vec Phoneme* and any of the segmentators.

  In the first case, we calculated word error rate (*WER*), and in the second case, we calculated phoneme error rate (*PER*). For both models, we also obtained the number of substitutions, deletions, and insertions for each utterance. As we did before, we built a new variable called *total errors* as the sum of all the errors mentioned before. As these values are sensible to the length of the utterance, we did length normalization at utterance level.

Finally, in order to do a fair comparison between speakers and production modes, we only take into account the audios of the speakers that had done, at least, one audio of each type of production mode.

3. Results

The values shown in Tables 2 and 3 represent group-wise means, regardless of whether the variable is continuous or categorical. We also show the *p-value* for two groups, which will be bold if there is significant difference between the spontaneous group and all the others together.

The first result we found during the analysis was that all the variables (except one) followed a nonnormal distribution, which means that we mainly applied only nonparametric tests.

3.1. Disfluency annotations

As shown in Table 2, there are no statistically significant differences in disfluency annotations between our two groups. Based on these results, we conclude that the disfluency patterns, measured by the selected annotation categories, did not differ significantly between the groups.

3.2. Professional linguist analysis

As previously outlined, we can distinguish three different types of analysis: **General assessment**, **Fine-Grained Fluency analysis** and **Phonetic word analysis**. Regarding the first, we can see in Table 2 that there is no significant difference between the groups.

Regarding the analysis of **fine-grained fluency** (see Table 2), we find that only **Interjections** and **Prolongations**, are significantly higher between spontaneous and nonspontaneous speech.

Production mode	#audios	#words	#phonemes MFA	#phonemes WebMAUS
Read	1110	7416	28687	28522
Elicited	434	1960	7976	7874
Imitation	140	776	3194	3147
Spontaneous	35	281	1174	1156

Table 1: *Number of audios, word and phonemes by Production Mode. In the case of phonemes, as we employ two different segmentation, we report the values for each tool.*

	Variable	4 groups				2 groups		
		R	**E**	**I**	**S**	**NS**	**S**	**p-value**
DISFLUENCIES	ET	0.07	0.07	0.10	0.08	0.07	0.08	0.4448
	ETw	0.08	0.12	0.14	0.15	0.10	0.15	0.3788
	Fil	0.02	0.02	0.04	0.04	0.02	0.04	0.0719
GENERAL ASSESSMENT	Phonetic	2.09	2.17	2.01	1.86	2.11	1.86	0.0771
	Fluency	2.65	2.71	2.60	2.46	2.66	2.46	0.2147
FLUENCY	Blocks	0.32	0.24	0.35	0.43	0.30	0.43	0.5518
	Prolongations	0.11	0.08	0.13	0.34	0.10	0.34	**0.0278**
	Repeated Sounds	0.19	0.12	0.19	0.17	0.17	0.17	0.4235
	Repeated Words	0.15	0.16	0.23	0.31	0.16	0.31	0.1190
	Interjections	0.04	0.04	0.06	0.20	0.04	0.20	**0.0176**
WORDS	Substitutions	0.03	0.05	0.05	0.10	0.04	0.10	**0.0000**
	Omisions	0.09	0.10	0.12	0.11	0.09	0.11	0.4657
	Distorsions	0.25	0.26	0.25	0.31	0.25	0.31	**0.0289**
	Additions	0.02	0.03	0.03	0.02	0.03	0.02	0.4141
	Total errors	0.39	0.44	0.45	0.53	0.41	0.53	**0.0023**

Table 2: *Results of the analysis of disfluency variables, general assessment variables, fine-grained fluency variables and phonetic word variables for 2 and 4 groups. ET: average number of editing terms per recording, ETw: average number of words in the editing terms per recording, Fil: average number of fillers per recording. Phonetic and Fluency: average ratings for each level. The features in the FLUENCY and WORDS groups represent the average number of occurrences of each aspect, as provided by the categorical expert evaluation. WORDS groups represent error at phoneme level found during the analysis of each word individually.*

The last type of analysis performed by the linguist was the **Phonetic word analysis** (see Table 2), which revealed further group differences. Considering our two groups, we will see that the significant evidence appears in, **Substitutions**, **Distorsions** and **Total errors** variables. These differences tell us that Spanish speakers with DS tend to make more mistakes, of those types, when they produced spontaneous speech.

3.3. Automatic phonetic segmentation

To investigate how production mode affects articulation timing, we analyzed the mean phoneme durations extracted from the automatic phoneme segmentation systems (see Table 3). We found significant difference that indicates, that phoneme duration is significantly shorter in spontaneous speech.

When individual phoneme durations extracted from forced alignment are analyzed, we found that silence, /e/ and /i/ phonemes have an average longer duration in spontaneous speech utterances, for both aligners. This might resemble a slower speech production and a higher number of fillers made up by prolongations of the phonemes /e/ and /i/.

3.4. Deep learning-based transcriptors results

Results from the **Whisper model** (Table 3) show us that this model mainly works better with spontaneous speech, rather than with the other types. Differences can be seen in **Substitutions** and **WER**, which are significantly lower.

Regarding the results from **Wav2Vec Phoneme model** (Table 3), we can see that the evidence is very similar between both segmetations. In both cases, **PER**, **Substitutions** and **Total errors** are significantly higher when we talk about our binary grouping.

4. Discussion

Our analyses revealed that while normalized counts of general disfluencies (fillers, editing terms) did not show significant differences between spontaneous and nonspontaneous speech, specific types of disfluencies and phonetic errors were more prevalent in spontaneous utterances. Professional linguistic analysis indicated significantly higher rates of interjections and prolongations, as well as increased phonetic substitutions and distortions in spontaneous speech. Furthermore, automatic analysis showed that overall phoneme duration was significantly shorter in spontaneous speech, although certain phonemes (/e/, /i/) and silences were longer. These findings suggest that while the overall quantity of manually annotated disfluencies might be similar after normalization, the nature of fluency breaks and phonetic execution differs significantly in spontaneous contexts for speakers with DS, leaning towards patterns potentially indicative of speech planning or motor control difficulties rather than just pragmatic pausing.

	Variable	4 groups				2 groups		
		R	E	I	S	NS	S	p-value
AVG. PHONEME DURATION (ms)	MFA	98.1	93.1	89.4	94.9	96.4	94.9	**0.0004**
	WebMAUS	86.8	85.0	83.4	85.5	86.2	85.5	**0.0070**
WHISPER ASR	WER	0.33	0.38	0.48	0.24	0.42	0.24	**0.0419**
	Substitutions	0.21	0.23	0.32	0.11	0.22	0.11	**0.0060**
	Deletions	0.14	0.12	0.23	0.13	0.13	0.13	0.4315
	Insertions	0.03	0.04	0.05	0.02	0.04	0.02	0.6354
	Total errors	0.39	0.39	0.60	0.26	0.39	0.26	0.0901
WAV2VEC PHONEME RECOGNIZER								
MFA [14]	PER	0.43	0.48	0.48	0.53	0.45	0.53	**0.0039**
	Substitutions	0.31	0.35	0.31	0.38	0.32	0.38	**0.0148**
	Deletions	0.16	0.18	0.20	0.18	0.17	0.18	0.5554
	Insertions	0.03	0.03	0.04	0.04	0.03	0.04	0.3086
	Total errors	0.50	0.56	0.55	0.61	0.52	0.61	**0.0106**
WebMAUS [15]	PER	0.47	0.51	0.49	0.59	0.48	0.59	**0.0000**
	Substitutions	0.34	0.38	0.34	0.45	0.35	0.45	**0.0003**
	Deletions	0.15	0.17	0.19	0.17	0.16	0.17	0.9330
	Insertions	0.03	0.04	0.04	0.05	0.04	0.05	0.2125
	Total errors	0.53	0.58	0.57	0.67	0.55	0.67	**0.0007**

Table 3: *Duration and Error Rate recognition results using forced aligners and ASR models (silence is not included as a phoneme). WER/PER is the average of the error rate metric by utterance. Subtitutions, Deletions, Insertions, and Total errors are the average of the number of errors of each type per utterance.*

Our results show that, contrary to what is observed in typical speakers [21, 22], people with Down syndrome do not present a significant increase in disfluencies in spontaneous speech compared to nonspontaneous speech. This observation can be interpreted in light of Clark's proposal [23], who argues that disfluencies should not be understood as failures in speech production, but as solutions to problems of speech planning and execution. In this sense, disfluencies serve relevant communicative functions, such as facilitating synchronization between interlocutors, improving auditory comprehension or signaling the complexity of the upcoming message [23, 24, 25, 26, 27, 28]. The absence of modality-sensitive modulation of disfluency use suggests that these speakers may lack flexible control of disfluency mechanisms. These speakers may have reduced control over the pragmatic management of disfluencies, which could represent an additional constraint on their spontaneous communicative competence. This deficit would not only affect discourse fluency, but also the ability to use disfluencies as strategic tools to manage communicative interaction. Since disfluencies contribute to the regulation of interaction and the listener's processing of information, impaired use of them may indicate broader pragmatic difficulties, limiting the speaker's ability to adapt to the communicative context and effectively manage interlocutor expectations.

Regarding the impact on ASR, the results were model-dependent. The Whisper model showed improved performance on spontaneous speech, achieving a significantly lower Word Error Rate (WER) and fewer substitution errors compared to nonspontaneous modes. Conversely, the Wav2Vec phoneme recognition model performed significantly worse on spontaneous speech, exhibiting higher Phoneme Error Rate (PER), more substitutions, and greater total errors, irrespective of the automatic segmentation tool used (MFA or WebMAUS). This divergence highlights that the specific challenges posed by spontaneous DS speech (e.g., increased phonetic variability, altered timing, specific disfluency types) affect different ASR architectures and evaluation granularities (word vs. phoneme) in distinct ways. Whisper's robustness might stem from its large-scale, diverse training data, potentially making it less sensitive to certain variations, while the phoneme-level Wav2Vec model appears more susceptible to the increased phonetic errors observed in spontaneous speech.

5. Conclusions

This study investigated the characteristics of spontaneous speech in Spanish speakers with Down syndrome (DS) compared to nonspontaneous speech modes (read, elicited, imitation) and assessed the impact of these differences on automatic speech recognition (ASR) performance. Our objective was to determine if spontaneous speech presents significantly greater phonetic and fluency challenges and how these might affect deep learning-based transcription models.

In summary, our study confirms that the spontaneous speech of Spanish speakers with DS presents distinct phonetic and fluency characteristics compared to nonspontaneous speech, manifested in specific types of disfluencies and phonetic errors . Furthermore, it shows that these differences have variable implications for the performance of ASR models , improving for word-level models like Whisper but worsening for phoneme-level models like Wav2Vec.

6. Acknowledgements

This work was carried out in the Project PID2021-126315OB-I00 that was supported by MCIN / AEI / 10.13039/501100011033 / FEDER, EU.

7. References

[1] R. D. Kent and H. K. Vorperian, "Speech impairment in down syndrome: A review," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 1, pp. 178–210, 2013. [Online]. Available: https://pubs.asha.org/doi/abs/10.1044/1092-4388%282012/12-0148%29

[2] E. M. Wilson, L. Abbeduto, S. M. Camarata, and L. D. Shriberg, "Speech and motor speech disorders and intelligibility in adolescents with down syndrome," *Clinical linguistics & phonetics*, vol. 33, no. 8, pp. 790–814, 2019.

[3] M. Corrales-Astorgano, D. Escudero-Mancebo, C. Gonzalez-Ferreras, V. C. Payo, and P. Martinez-Castilla, "Analysis of atypical prosodic patterns in the speech of people with down syndrome," *Biomedical Signal Processing and Control*, vol. 69, p. 102913, 2021.

[4] M. Corrales-Astorgano, D. Escudero-Mancebo, and C. González-Ferreras, "Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with down syndrome," *Speech Communication*, vol. 99, pp. 90–100, 2018.

[5] E. López-Riobóo and P. Martínez-Castilla, "Prosodic skills in spanish-speaking adolescents and young adults with down syndrome," *International Journal of Language & Communication Disorders*, vol. 59, no. 4, pp. 1284–1295, 2024.

[6] K. Eggers and S. Van Eerdenbrugh, "Speech disfluencies in children with down syndrome," *Journal of Communication Disorders*, vol. 71, pp. 72–84, 2018.

[7] M. C. Coppens-Hofman, H. R. Terband, B. A. Maassen, H. M. van Schrojenstein Lantman-De, Y. van Zaalen-op't Hof, A. F. Snik *et al.*, "Dysfluencies in the speech of adults with intellectual disabilities and reported speech difficulties," *Journal of Communication Disorders*, vol. 46, no. 5-6, pp. 484–494, 2013.

[8] L. Zampini, M. Fasolo, M. Spinelli, P. Zanchi, C. Suttora, and N. Salerni, "Prosodic skills in children with down syndrome and in typically developing children," *International Journal of Language & Communication Disorders*, vol. 51, no. 1, pp. 74–83, 2016.

[9] D. O'Leary, A. Lee, C. O'Toole, and F. Gibbon, "Perceptual and acoustic evaluation of speech production in down syndrome: A case series," *Clinical linguistics & phonetics*, vol. 34, no. 1-2, pp. 72–91, 2020.

[10] A. Romana, M. Niu, M. Perez, and E. M. Provost, "Fluencybank timestamped: An updated data set for disfluency detection and automatic intended speech recognition," *Journal of Speech, Language, and Hearing Research*, vol. 67, no. 11, pp. 4203–4215, 2024.

[11] P. J. Lou, P. Anderson, and M. Johnson, "Disfluency detection using auto-correlational neural networks," *arXiv preprint arXiv:1808.09092*, 2018.

[12] D. Escudero-Mancebo, M. Corrales-Astorgano, V. Cardeñoso-Payo, L. Aguilar, C. González-Ferreras, P. Martínez-Castilla, and V. Flores-Lucas, "Prautocal corpus: a corpus for the study of down syndrome prosodic aspects," *Language Resources and Evaluation*, vol. 56, no. 1, pp. 191–224, 2022.

[13] M. Corrales-Astorgano, C. González-Ferreras, D. Escudero-Mancebo, L. Aguilar, V. Flores-Lucas, V. Cardenoso-Payo, and C. Vivaracho-Pascual, "Pronunciation assessment and automated analysis of speech in individuals with down syndrome: Phonetic and fluency dimensions," in *Proc. IberSPEECH 2024*, 2024, pp. 26–30.

[14] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.

[15] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, Sep. 2017.

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[17] D. Fernández-García, V. Cardeñoso-Payo, C. González-Ferreras, and D. Escudero-Mancebo, "Adaptación de asr al habla de personas con síndrome de down," *Procesamiento del lenguaje natural*, vol. 73, pp. 209–220, 2024.

[18] Q. Xu, A. Baevski, and M. Auli, "Simple and effective zero-shot cross-lingual phoneme recognition," *arXiv preprint arXiv:2109.11680*, 2021.

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[20] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222. [Online]. Available: https://aclanthology.org/2020.lrec-1.520/

[21] P. Wagner and A. Windmann, "Re-enacted and spontaneous conversational prosody—how different?" in *Proceedings of Speech Prosody*, 2016, pp. 518–522.

[22] E. Shriberg, "To 'errrr'is human: ecology and acoustics of speech disfluencies," *Journal of the international phonetic association*, vol. 31, no. 1, pp. 153–169, 2001.

[23] H. H. Clark, "Speaking in time," *Speech communication*, vol. 36, no. 1-2, pp. 5–13, 2002.

[24] R. L. Rose, "The communicative value of filled pauses in spontaneous speech," *MA Diss., Univ. of Birmingham*, 1998.

[25] J. E. Fox Tree, "Listeners' uses of um and uh in speech comprehension," *Memory & cognition*, vol. 29, pp. 320–326, 2001.

[26] M. Watanabe, K. Hirose, Y. Den, and N. Minematsu, "Filled pauses as cues to the complexity of following phrases." in *Interspeech*, 2005, pp. 37–40.

[27] M. Corley and O. W. Stewart, "Hesitation disfluencies in spontaneous speech: The meaning of um," *Language and Linguistics Compass*, vol. 2, no. 4, pp. 589–602, 2008.

[28] H. Moniz, I. Trancoso, and A. I. Mata, "Classification of disfluent phenomena as fluent communicative devices in specific prosodic contexts." in *INTERSPEECH*, 2009, pp. 1719–1722.