Identifying Filled Pauses in Speech Across South and West Slavic Languages

Nikola Ljubešić[1,2,3], Ivan Porupski[1], Peter Rupnik[1],
[1]Department of Knowledge Technologies, Jožef Stefan Institute, Slovenia
[2]Faculty of Computer and Information Science, University of Ljubljana, Slovenia
[3]Institute of Contemporary History, Ljubljana, Slovenia

Abstract

Filled pauses are among the most common paralinguistic features of speech, yet they are mainly omitted from transcripts. We propose a transformer-based approach for detecting filled pauses directly from the speech signal, fine-tuned on Slovenian and evaluated across South and West Slavic languages. Our results show that speech transformers achieve excellent performance in detecting filled pauses when evaluated in the in-language scenario. We further evaluate cross-lingual capabilities of the model on two closely related South Slavic languages (Croatian and Serbian) and two less closely related West Slavic languages (Czech and Polish). Our results reveal strong cross-lingual generalization capabilities of the model, with only minor performance drops. Moreover, error analysis reveals that the model outperforms human annotators in recall and F1 score, while trailing slightly in precision. In addition to evaluating the capabilities of speech transformers for filled pause detection across Slavic languages, we release new multilingual test datasets and make our fine-tuned model publicly available to support further research and applications in spoken language processing.

1 Introduction

Most of the research in the discipline of computational linguistics was traditionally focused on the textual modality of language, while the spoken modality was only occasionally covered (Rohatgi et al., 2023). The main reason for this focus on text was the complexity of the speech signal compared to the textual modality.

With the advent of neural language representations (Goldberg, 2017), and especially pre-trained language models that allowed for embedding of speech in a manner comparable to text (Schneider et al., 2019), this trend started to change.

This paper is part of this change, investigating the possibility of identifying directly in the spoken

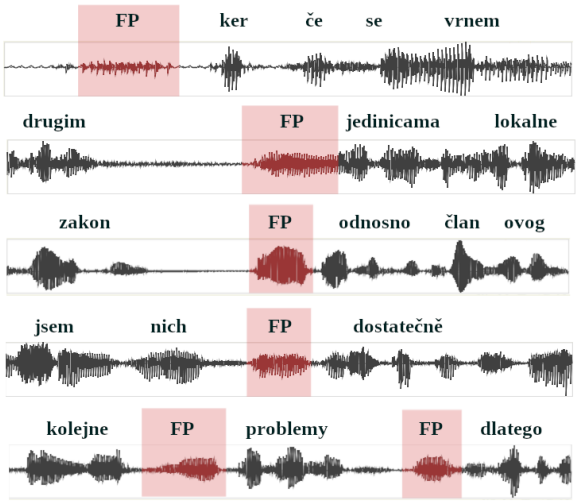

Figure 1: Predicting filled pauses (FP) from speech in Slovenian, Croatian, Serbian, Czech and Polish with a speech transformer fine-tuned on Slovenian data.

modality one of the most common paralinguistic features in speech – filled pauses (Lea et al., 2021; Bayerl et al., 2022a; Romana et al., 2024). The main motivation to focus on this feature is that it is most often not present in the transcript of the spoken signal (Romana et al., 2023), although it is a very frequent phenomenon that serves important communicative and cognitive functions, for instance, turn-taking management (Gósy, 2023).

This paper presents an automated approach to identify filled pauses directly from the speech signal by fine-tuning a transformer-based speech encoder (Barrault et al., 2023) to perform identification of the phenomenon on 20 ms audio frame level.

Our model fine-tuning data are in Slovenian, a South Slavic language. Besides investigating the capacity to perform filled pause identification inside the Slovenian language on dedicated test data, we investigate the capacity of this technology to perform the same task on two other South Slavic

Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025), pages 1–8

languages, Croatian and Serbian, as well as on two West Slavic languages, Czech and Polish (see Figure 1).

Our main findings are these: (1) speech transformers are highly capable of identifying filled pauses inside the language of the fine-tuning data (F1 above 0.95), (2) this technology is very much portable to related languages, with visible drops in performance, but an output that is still very useful (F1 between 0.87 and 0.94), (3) when comparing human and machine performance on the task, machines actually outperform humans in terms of recall as well as F1, but they fall slightly behind humans in terms of precision.

The contributions of this paper are the following. We (1) investigate the capacity of speech transformers to perform filled pause detection in Slavic languages, (2) prepare new test data in two South Slavic and two West Slavic languages, (3) investigate the capacity of the model to perform the task across more and less related Slavic languages, and (4) release the final model for downstream applications on spoken corpora.

The remainder of this paper is structured as follows. In the next section, we summarise the related work. In Section 3 we introduce fine-tuning data, as well as the five test datasets. Section 4 describes our fine-tuning and evaluation setup, while Section 5 presents quantitative results and its error analysis, followed by a qualitative and acoustic analysis. We wrap up with a conclusion, covering also data and model availability, as well as the path forward.

2 Related Work

Early work on filled pause detection focuses on acoustic features for classification such as the fundamental frequency (F0, pitch) and spectral stability (Goto et al., 1999), frame-level MFCCs (Mel-frequency Cepstral Coefficients) (Stouten and Martens, 2003), or vocal tract stability (formants) (Audhkhasi et al., 2009) with performance ranging between 0.3 and 0.7 in precision and recall. Medeiros et al. (2013) investigate the application of prosodic features to detect filled pauses in spontaneous speech, achieving an F1 score of about 61%. Reichel et al. (2019) improve the previous approach, focusing on prosodic discontinuity features, reaching an F1 score of 83%. More recent studies have shifted toward transformer-based models, demonstrating further advancements in detection accuracy (Romana et al., 2023; Mohapatra et al., 2022; Bayerl et al., 2022b).

Recent experiments predominantly focus on atypical speech, mainly stuttering, with only one study addressing typical speech. Specifically, Romana et al. (2023) investigate wav2vec2, HuBERT and WavLM transformer models for frame-level automatic disfluency detection and categorization on the Switchboard corpus (Godfrey et al., 1992), reaching a frame-level F1 score between 0.86 and 0.88, depending on the model.

Other recent experiments were performed on datasets of atypical speech: the English SEP-28k (Stuttering Events Podcasts) corpus (Lea et al., 2021), the German disfluency stuttering corpus KSoF (Kassel State of Fluency) (Bayerl et al., 2022a) and finally, the FluencyBank Timestamped corpus (Romana et al., 2024), which includes typically developing monolingual and bilingual children, children and adults who stutter or who clutter, as well as second language learners. Mohapatra et al. (2022) propose a model based on wav2vec2 contextual embeddings followed by 2D convolution feedforward layers, which scores an F1 score of 0.88 for filled pauses in the SEP-28k dataset. Bayerl et al. (2022b) fine-tune their wav2vec2 model on SEP-28k corpus and a portion of the FluencyBank corpus, before showing good transferability to the German KSoF corpus. Single-task learning on filled pauses returned an F1 score – for FluencyBank and KSoF respectively – of 0.83 and 0.71, while a multi-task learning model resulted in F1 scores of 0.84 and 0.74. Important to stress is that experiments on atypical speech are of limited use for typical speech processing due to the more complex nature of atypical speech, including more disfluencies per word than typical speech (Lea et al., 2021; Liu et al., 2023; Romana et al., 2024).

All of the mentioned approaches use the speech modality for identifying filled pauses. In addition to the speech modality, some approaches also use automatic speech recognition (ASR) systems to generate the transcripts and then exploit the text modality (Chatziagapi et al., 2022; Romana et al., 2023). However, the transcripts show to be useful more for detecting repairs and repetitions rather than filled pauses, for which transcripts show, as expected, to be much less informative than the speech signal (Romana et al., 2023).

The work presented in this paper builds on the set of experiments performed on typical English speech (Romana et al., 2023), investigating the applicability of the straightforward approach of

fine-tuning a speech transformer model to a Slavic language, namely Slovenian, with a shift from the technical frame-level evaluation to the more application-oriented event-level evaluation. In addition, it investigates the applicability of this model to other Slavic languages, along the lines of an experiment carried out on atypical speech (Bayerl et al., 2022b), covering in this case four different Slavic languages of different level of relatedness to the Slovenian language.

3 Data

This section describes the data used in our experiments. We first describe our Slovenian fine-tuning and in-language test data, moving forward to describe the construction process of our four additional cross-lingual test datasets in Croatian, Serbian, Czech, and Polish. A quantitative overview of the fine-tuning and evaluation data is given in Table 1.

3.1 In-language data

For fine-tuning the transformer model to the task of filled pause identification, we exploited the ROG dataset (Verdonik et al., 2024). The dataset contains recordings of Slovenian speech and manual annotations on multiple layers, including that of disfluencies, which also covers filled pauses. To exploit the ROG training data to their maximum, the recordings were split into 30 s chunks with 50 % overlap. As presented in Table 1, the fine-tuning data contain 1314 filled pauses, while the evaluation dataset contains 558 filled pauses.

3.2 Cross-lingual data

To test cross-lingual performance of our model, we constructed test datasets in four languages present in the ParlaSpeech collection[1] of spoken parliamentary corpora (Ljubešić et al., 2024). For each of the available languages, namely Croatian, Serbian, Czech, and Polish, we sampled 400 instances (transcript sentences and the speech recordings) with speech lengths between 6 and 20 s. While sampling, two additional criteria were taken into account. The first criterion was to ensure a 50-50 gender balance. The second criterion required pre-annotation of the data with the Slovenian fine-tuned model, to sample half instances with automatically identified filled pauses, and the other half of instances without automatically identified filled pauses. With this final sampling criterion, we ensured a reasonable number of positive instances in our test data regardless of the data coming from parliamentary proceedings, while sampling randomly would require the test dataset to be very large to include enough examples of filled pauses.

For the manual annotation campaign, we prepared audio recordings and ELAN files with an empty tier to be used by the annotators. The annotation guidelines were kept as short as possible. Annotators were asked to mark the "schwa"-like filled pauses wherever they noticed them. However, the annotators were made aware that beginning and endings of instances from the ParlaSpeech collection might include incomplete words due to the instance separation based on ASR-based automatic word alignment, and that incomplete words should not be confused for filled pauses. With this manual annotation process, between 288 and 394 filled pauses were annotated inside the 400 test sentences. Detailed statistics can be inspected in Table 1.

In cases of Croatian and Serbian test sets, we introduced a second annotator who annotated 10 % of the data, which allowed us to estimate agreement between annotators. The results of the inter-annotator agreement are presented in Table 2 in terms of observed F1 and Krippendorff α (Castro, 2017). While performing these calculations, we followed the overall evaluation protocol of our experiments, focusing on event-level evaluation, as described in detail in Section 4.2. Important at this point are two things: (1) the overall agreement is rather high, with observed agreement around 0.9 and a very good Krippendorff score of around 0.8 (2) besides proving that we have high-quality annotations in our cross-lingual test datasets, we also want to emphasize that the observed agreement can be considered a ceiling of what can be measured in quantitative analyses described in Section 5.1.

4 Experiments

In this section, we present the approach to fine-tuning the transformer model, as well as the data representation and evaluation setup that we follow throughout our experiments.

4.1 Fine-tuning protocol

We fine-tune a Wav2Vec2Bert model (Barrault et al., 2023) in its 'Audio Frame Classification' mode, which means the labels at input and output

[1] https://huggingface.co/collections/classla/parlaspeech-670923f23ab185f413d40795

lang	split	words	filled pauses
SL	train+dev	38 881	1 314
SL	test	9 440	558
HR	test	10 525	289
SR	test	10 762	288
CZ	test	8 368	318
PL	test	8 928	394

Table 1: The Slovenian (SL) fine-tuning and five evaluation datasets in Slovenian, Croatian (HR), Serbian (SR), Czech (CZ), and Polish (PL), presented through number of words and number of filled pauses in each dataset.

lang	F1	Krippendorff α
HR	0.932	0.791
SR	0.889	0.814

Table 2: Inter-annotator agreement in terms of F1 score and Krippendorff α for test instances sampled in Croatian (HR) and Serbian (SR), annotated by two annotators.

stages are binary vectors, with each element corresponding to a $20\,\mathrm{ms}$ frame and describing whether a filled pause occurs in that frame. After preparing all the training data labels into appropriate binary vectors, we fine-tuned a Wav2Vec2Bert model with some initial experiments on provisional training data splits to determine the optimal hyperparameters. We investigated learning rates of 3×10^{-5}, 1×10^{-6}, and 8×10^{-6}, training duration of 10 and 20 epochs, and gradient accumulation steps of 1 and 4.

The optimal hyperparameters used in the final fine-tuning were learning rate 3×10^{-5}, training duration 20 epochs, and gradient accumulation steps set to 4.

4.2 Evaluation protocol

The output of our fine-tuned model is a series of $20\,\mathrm{ms}$ frame-level predictions encoding whether there is a filled pause present in each frame or not. Given that it is not easy to state where exactly a filled pause has started and ended, the human annotators often selecting additional silence around a filled pause, for evaluation purposes, we transformed our data representation from a binary frame-based representation to a span-based representation, each filled pause being represented by its start and end time. This allows us to evaluate the output of the model in terms of true positives, i.e., when the true and predicted filled pause overlap, false positives, i.e., when there is a predicted filled pause in an interval with no true filled pause, and false negatives, i.e., when there is a true filled pause annotated, but none predicted.

By having the output of a machine compared to the human annotations in terms of true positives, false positives and false negatives, we can report precision, recall, and F1 for our quantitative evaluation.

This event-level evaluation, measuring the percentage of filled pauses we managed to correctly identify, and the percentage of those we missed, is much more useful for informing downstream applications than the $20\,\mathrm{ms}$ frame-level overlap between human and machine output, the evaluation followed in most related work, including the only experiment on applying transformers to typical speech data (Romana et al., 2023).

While quantitatively evaluating the model, we also investigate a post-processing technique, especially aimed at the cross-lingual ParlaSpeech-based test sets. These test sets consist of instances that were segmented via imperfect ASR-based automatic word alignment, each instance covering one transcript sentence. Because of these segmentation imperfections, the post-processing rule discards predicted filled pauses at the beginning and ending of an instance, as it is rather possible for incomplete words to be mistaken for filled pauses. The post-processing technique also discards very short predictions (less than $80\,\mathrm{ms}$ long), as such brief instances are unlikely to be reliably perceived by humans.

5 Results

This section presents a quantitative analysis of the results, followed by an error analysis to further clarify the results of the quantitative analysis. These are followed by a qualitative and an acoustic interpretation of the output of our model.

5.1 Quantitative analysis

In Table 3 we report recall, precision, and F1 scores of our model on each of the five test sets, both with and without post-processing applied (column 'post-proc').

The post-processing overall lowers recall while improving precision, which is to be expected given that it only discards specific filled pause predictions. Also, as anticipated, post-processing improves results on the ParlaSpeech-based cross-lingual test sets because of the imperfect segmen-

lang	post-proc	recall	precision	F1
SL	no	0.973	0.914	0.943
	yes	0.959	0.922	0.940
HR	no	0.940	0.872	0.905
	yes	0.940	0.887	0.913
SR	no	0.974	0.900	0.936
	yes	0.966	0.915	0.940
CZ	no	0.905	0.814	0.857
	yes	0.889	0.859	0.874
PL	no	0.910	0.924	0.917
	yes	0.903	0.947	0.924

Table 3: Recall, precision, and F1 score of the positive class on the test datasets, calculated on raw and post-processed outputs.

tation of the original ParlaSpeech data. On the Slovenian test data that were manually segmented, post-processing does not have a global positive impact.

The in-language results on Slovenian show to be very strong, with recall, precision and F1 being above 0.9. If we compare these results to the English-based Switchboard experiments achieving frame-level F1 of 0.86 to 0.88 (Romana et al., 2023), our Slovenian results show to be roughly comparable.

The cross-lingual evaluation of our models shows a visible, but acceptable drop, with performance of the post-processed output ranging from 0.87 to 0.94 in F1.

The reported results sometimes go even above the observed agreement between two human annotators achieved on Croatian and Serbian data, as discussed in Section 3.2. For this reason, an error analysis comparing human and machine output is necessary, which we perform in the following section.

5.2 Error analysis

The automatic evaluation results presented in the previous section reach or even surpass the level of inter-annotator agreement we have measured on the Croatian and Serbian double-annotated data, therefore a natural question arises – given that we have in some cases surpassed the limits in measuring the quality of the automatic responses via human-annotated data, we wonder who is actually better at this task, human or machine? We hypothesize that, based on the numbers we observed, where automatic evaluation was sometimes higher than inter-human agreement, machines might actually perform better than humans.

To answer the above question, we perform a manual analysis of 20 test instances per language where human and machine disagree. This comparison was performed by a trained phonetician, who has a good understanding of three out of five languages, using transcripts for easier decision-making on the remaining two. The phonetician-annotator discriminated between false positives, i.e., situations where human or machine would claim there was a non-existing filled pause, and false negatives, i.e., situations where human or machine would miss an existing filled pause.

The results of the disagreement analyses show that human error is the more frequent reason for a disagreement between human and machine, proving our assumption that machines overall perform better on the task. However, while humans mostly miss existing filled pauses, resulting in more false negatives in comparison to machines, machines generate more false positives than humans do. Given that humans generate twice as many false negatives as machines do, while machines generate around $40\,\%$ more false positives than humans, machines generating a similar amount of false negatives and positives, we can conclude that machines generate stable, high-quality output.

It is important to stress that both human and machine perform very well on the task. In all the test data, humans and machines agreed in $95\,\%$ of their predictions, showing that both manual or automatic annotations of filled pauses can be safely used in downstream data analyses.

5.3 Qualitative analysis

The following qualitative analysis is performed by a phonetician, documenting the sources of confusion and discrepancies observed in the model's results in all evaluation languages.

Prolonged vowel sounds (e.g. the conjunction /a/), prolonged nasals (e.g. /m, n/) and noise (e.g. a cough or other speaker in the background) can all have a negative impact on the model's performance. On occasion, a repetition or repair might be wrongly flagged as filled pause. Generally speaking, short filled pauses are reliably detected by the model, even when they are subtle or barely perceptible to human listeners. The model demonstrates particular strength in capturing voiced, but weakly articulated sounds, often unnoticed by annotators.

The most frequent model errors in terms of false positives for each language are as follows: In both

Slovenian and Croatian, nasals caused the most false positives. In Serbian and in Czech, the source of false positives were vowel sounds, e.g., the conjunction /a/, especially if prolonged. In Polish, nasals and background noise have caused the most false positives.

5.4 Acoustic analysis

To gain a deeper understanding of the problem at hand, we analyse the acoustic features of filled pauses in the five Slavic languages and examine how their formant overlap relates to the model's cross-lingual performance. We perform this analysis on around one thousand gender-balanced predicted filled pauses per language.

Formants are the resonant frequencies of the vocal tract that shape vowel sounds in speech. We analyse the first and second formants (F1 and F2), which are commonly used to represent differences in vowel quality, including those found in vowel-like filled pauses. These values are visualized in a vowel diagram (see Figure 2). By comparing formants of filled pauses, it is possible to conclude how similarly they are articulated between languages.

Formant measurements for both vowels and filled pauses were extracted at the phoneme level using Praat (Boersma and Weenink, 2001). Median values were selected in lieu of means to mitigate the impact of erroneous formant readings occasionally produced by Praat. To provide a reference framework within the vowel space, we included the five Croatian vowels /i, e, a, o, u/, the language with the strongest support in the original dataset. This contextualization enables a clearer interpretation of the relative positioning of the filled pauses.

To visualize distributional tendencies, kernel density estimation (KDE) was used, with a single isoline drawn at 80 % of the peak density, resulting in smoother and more interpretable contour representations.

Figure 2 shows a clear similarity between filled pauses in Slovenian (SL_FP), Croatian (HR_FP), Serbian (SR_FP) and Czech (CZ_FP). All four languages seem to share universal filled pauses, characterised in the /ə/ region of the vowel plot, having a median F1 between 500-600 Hz and median F2 between 1300-1500 Hz.

The obvious outlier, Polish (PL_FP), stands out substantially from the other languages, having filled pauses localized more around the vowel /e/, instead of /ə/, suggesting that not all languages

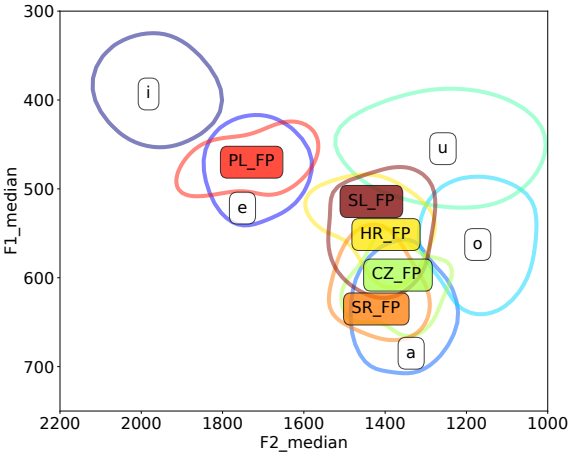

Figure 2: Filled pauses (FP) across languages presented in the vowel diagram, with Croatian vowels /i, e, a, o, u/ as reference. Each filled pause and vowel distribution is presented as a KDE plot at 80% peak density.

share a single filled pause form.

While filled pauses are frequently assumed to approximate a mid-central /ə/ in quality, Lickley (2015) cautions that this may constitute an overgeneralization.

Notably, although Slovenian and Croatian show the greatest acoustic overlap in filled pauses, the fine-tuned model performs better on Serbian than on Croatian test data. Even more striking, the model achieves the highest precision on Polish, despite its filled pauses being acoustically most distinct from Slovenian. It also outperforms Czech on all metrics, though Czech and Slovenian filled pauses are acoustically closer. These discrepancies indicate that the model does not rely solely on acoustic similarity and instead leverages contextual or language-general features, demonstrating strong generalization capability across typologically and acoustically varied languages.

6 Conclusion

This paper has investigated the capacity of pre-trained speech transformer models to identify filled pauses – one of the most frequent paralinguistic phenomena in speech. The fine-tuning language was Slovenian, while evaluation languages were Slovenian, Croatian, Serbian, Czech, and Polish, which allowed the model to be evaluated across the South and West Slavic languages.

The evaluation showed very strong results in Slovenian and an acceptable drop of around 5 F1 points in prediction quality in the remaining languages. What is more, the quantitative evaluation

revealed that the model's performance surpassed the observed inter-annotator agreement. Consequently, an error analysis was conducted, showing that the model's outputs were actually of higher quality than human annotations on all languages, with an important limitation – machines showed a slightly stronger tendency towards confusing linguistic elements such as unclearly pronounced words or lengthenings for filled pauses. At the same time, humans were twice as likely to miss filled pauses.

While the fine-tuning and test data for Slovenian were already available (Verdonik et al., 2024), as part of this work, we release four new datasets based on the ParlaSpeech collection, covering Croatian, Serbian, Czech, and Polish. These test datasets are available upon request, to prevent the integration of these data into future large language models. These test sets could soon be useful for evaluating speech-enabled large language models in a prompting scenario.

We also release our fine-tuned filled pause identifier via the HuggingFace repository[2]. Aside from that, we can report that the model has already been applied to the ParlaSpeech spoken corpus collection, spanning 5 thousand hours, 4 languages, and 800 thousand identified filled pauses, together with the linguistic annotation of the transcript, which allows for downstream research on linguistic contexts inside which filled pauses occur. The resulting datasets are available through a FAIR repository[3].

Our future plans include extending the source of speech data from parliamentary discussions to sources covering more variation. We also envision to expand the approach investigated here to other paralinguistic features. By combining the extended speech data sources with the possibility of automatic speech annotation, we hope to empower a new era of data-driven speech research.

Limitations

The main limitations of our work are the following: (1) the Slovenian test data come from the same source as the fine-tuning data, although the data source is rather diverse, (2) the cross-lingual test data come all from a single domain of parliamentary debates, (3) while we do test cross-lingual performance across a number of languages, we are positive that performance would further drop if the model was applied on phonologically more distant languages, (4) our model is more prone to false positives than humans, showing need for further performance improvements.

[2] https://huggingface.co/classla/Wav2Vec2BertPrimaryStressAudioFrameClassifier

[3] http://hdl.handle.net/11356/1833

Acknowledgements

This work was supported by the Projects "Spoken Language Resources and Speech Technologies for the Slovenian Language" (Grant J7-4642), "Large Language Models for Digital Humanities" (Grant GC-0002), the Research Programme "Language Resources and Technologies for Slovene" (Grant P6-0411), and the Research Infrastructure DARIAH-SI (I0-E007), all funded by the ARIS Slovenian Research and Innovation Agency.

References

Kartik Audhkhasi, Kundan Kandhway, Om D Deshmukh, and Ashish Verma. 2009. Formant-based technique for automatic filled-pause detection in spontaneous spoken English. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4857–4860. IEEE.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, and et al. 2023. Seamless: Multilingual Expressive and Streaming Speech Translation.

Sebastian P Bayerl, Alexander Wolff von Gudenberg, Florian Hönig, Elmar Nöth, and Korbinian Riedhammer. 2022a. KSoF: The Kassel state of fluency dataset–a therapy centered dataset of stuttering. *arXiv preprint arXiv:2203.05383*.

Sebastian P. Bayerl, Dominik Wagner, Elmar Noth, and Korbinian Riedhammer. 2022b. Detecting dysfluencies in stuttering therapy using wav2vec 2.0. In *Interspeech*.

Paul Boersma and David Weenink. 2001. Praat: doing phonetics by computer. http://www.praat.org/. Version 6.4, accessed 2024-01-30.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. https://github.com/pln-fing-udelar/fast-krippendorff.

Aggelina Chatziagapi, Dimitris Sgouropoulos, Constantinos Karouzos, Thomas Melistas, Theodoros Giannakopoulos, Athanasios Katsamanis, and Shrikanth Narayanan. 2022. Audio and ASR-based filled pause detection. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Yoav Goldberg. 2017. *Neural Network Methods for Natural Language Processing*, volume 10 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Mária Gósy. 2023. Occurrences and durations of filled pauses in relation to words and silent pauses in spontaneous speech. *Languages*, 8(1):79.

Masataka Goto, Katunobu Itou, and Satoru Hayamizu. 1999. A real-time filled pause detection system for spontaneous speech recognition. In *Sixth European Conference on Speech Communication and Technology*.

Colin Lea, Vikramjit Mitra, Aparna Joshi, Sachin Kajarekar, and Jeffrey P Bigham. 2021. Sep-28k: A dataset for stuttering event detection from podcasts with people who stutter. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6798–6802. IEEE.

Robin J Lickley. 2015. Fluency and disfluency. *The handbook of speech production*, pages 445–474.

Jiajun Liu, Aishan Wumaier, Dongping Wei, and Shen Guo. 2023. Automatic speech disfluency detection using wav2vec2. 0 for different languages with variable lengths. *Applied Sciences*, 13(13):7579.

Nikola Ljubešić, Peter Rupnik, and Danijel Koržinek. 2024. The ParlaSpeech Collection of Automatically Generated Speech and Text Datasets from Parliamentary Proceedings. In *International Conference on Speech and Computer*, pages 137–150. Springer.

Henrique Medeiros, Helena Moniz, Fernando Batista, Isabel Trancoso, Hugo Meinedo, et al. 2013. Experiments on automatic detection of filled pauses using prosodic features. *Actas de Inforum*, 2013:335–345.

Payal Mohapatra, Akash Pandey, Bashima Islam, and Qi Zhu. 2022. Speech disfluency detection with contextual representation and data distillation. In *Proceedings of the 1st ACM international workshop on intelligent acoustic systems and applications*, pages 19–24.

Uwe D Reichel, Benjamin Weiss, and Thilo Michael. 2019. Filled pause detection by prosodic discontinuity features. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, pages 272–279.

Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2023. The ACL OCL corpus: Advancing open science in computational linguistics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10348–10361, Singapore. Association for Computational Linguistics.

Amrit Romana, Kazuhito Koishida, and Emily Mower Provost. 2023. Automatic disfluency detection from untranscribed speech. *Preprint*, arXiv:2311.00867.

Amrit Romana, Minxue Niu, Matthew Perez, and Emily Mower Provost. 2024. Fluencybank timestamped: An updated data set for disfluency detection and automatic intended speech recognition. *Journal of Speech, Language, and Hearing Research*, 67(11):4203–4215.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Frederik Stouten and Jean-Pierre Martens. 2003. A feature-based filled pause detection system for Dutch. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*, pages 309–314. IEEE.

Darinka Verdonik, Kaja Dobrovoljc, Peter Rupnik, Nikola Ljubešić, Simona Majhenič, Jaka Čibej, and Thomas Schmidt. 2024. Training corpus of spoken Slovenian ROG 1.0. Slovenian language resource repository CLARIN.SI.