ParlaSpeech 3.0: Richly Annotated Spoken Parliamentary Corpora of Croatian, Czech, Polish, and Serbian

Nikola Ljubešić[†‡], **Peter Rupnik, Ivan Porupski, Taja Kuzman Pungeršek**
Jožef Stefan Institute, Ljubljana, Slovenia
† Faculty of Computer and Information Science, University of Ljubljana, Slovenia
‡ Institute of Contemporary History, Ljubljana, Slovenia
{nikola.ljubesic, peter.rupnik, ivan.porupski, taja.kuzman}@ijs.si

Abstract

ParlaSpeech is a collection of spoken parliamentary corpora currently spanning four Slavic languages – Croatian, Czech, Polish and Serbian – all together 6 thousand hours in size. The corpora were built in an automatic fashion from the ParlaMint transcripts and their corresponding metadata, which were aligned to the speech recordings of each corresponding parliament. In this release of the dataset, each of the corpora is significantly enriched with various automatic annotation layers. The textual modality of all four corpora has been enriched with linguistic annotations and sentiment predictions. Similar to that, their spoken modality has been automatically enriched with occurrences of filled pauses, the most frequent disfluency in typical speech. Two out of the four languages have been additionally enriched with detailed word- and grapheme-level alignments, and the automatic annotation of the position of primary stress in multisyllabic words. With these enrichments, the usefulness of the underlying corpora has been drastically increased for downstream research across multiple disciplines, which we showcase through an analysis of acoustic correlates of sentiment. All the corpora are made available for download in JSONL and TextGrid formats, as well as for search through a concordancer.

Keywords: spoken corpora, parliamentary corpora, Slavic languages, sentiment, filled pauses, speech alignment, primary stress

1. Introduction

Spoken corpora remain scarce compared to their written counterparts, largely due to the technical, logistical and legal challenges of their creation. Collecting natural speech often requires fieldwork, securing participant consent, and encouraging authentic conversation, while transcription and additional enrichment demands significant time, specialized skills, and deep knowledge of the communicative context. These difficulties are compounded when documenting unwritten languages or non-standard varieties (such as dialects, heritage languages, or L2 speech) which exist only in oral form. Even when available, spoken corpora, especially Slavic ones among European languages, are often small, limited in representativeness, and inconvenient to use, making them less effective for linguistic or phonetic analysis despite their importance for both research and speech technology development (Dobrushina and Sokur, 2022).

The CLARIN Resource Family (Fišer et al., 2018) for spoken corpora[1] reveals a highly skewed distribution of resources across European languages. While some, like Finnish and German, are well-resourced with thousands of hours of data, many others (including Croatian, Polish and Serbian) have only a small single corpus recorded. Most languages lack a corpus entirely. However, emerging corpus platforms, such as those hosted on CLARIN or TalkBank, have contributed significantly to the accessibility of spoken corpora across multiple languages concurrently.

Parliamentary proceedings offer an exceptional opportunity for building spoken corpora of various official languages due to the open status of the data, and their public status regarding privacy concerns. While there have been parliamentary corpora constructed in the past, they mostly focused only on the basic alignment of transcripts and recordings, having primarily automatic speech recognition (ASR) use cases in mind.

In this paper, we present the ParlaSpeech collection of spoken corpora, currently consisting of recordings from four parliaments, with speeches given in four languages. Different to all other parliamentary spoken corpora, the corpora have been enriched with various linguistic and paralinguistic layers, making the corpora highly useful for downstream research in various disciplines. The comparable nature of the corpora enables insights across languages, which significantly improves the strength of the resulting findings.

2. Related Work

There is a number of languages already represented via parliamentary spoken corpora, e.g.,

[1] https://www.clarin.eu/resource-families/spoken-corpora

Virkkunen et al. (2023) list corpora of 10 different languages. However, most of these corpora are primarily aimed at supporting development and improvement of automatic speech recognition (ASR) systems. A slight deviation from simplistic datasets, primarily aimed at ASR, is the Norwegian Parliamentary Speech Corpus (Solberg and Ortiz, 2022) which has its speeches automatically transcribed and manually checked, with non-standardized words explicitly marked and annotated with standardized equivalents. Another example of a parliamentary spoken corpus being more than just an ASR training dataset is the FT Speech Danish parliamentary corpus (Kirkedal et al., 2020), which includes normalizations of the official parliamentary transcripts, the lexic being aligned with the Språkbanken lexicon, and a grapheme-to-phoneme conversion. However, none of the available corpora (1) cover multiple languages, enabling cross-lingual studies, and, what is even more, (2) enrich the data on as many layers as is the case with the ParlaSpeech corpora.

Outside the realm of parliamentary corpora, there have been activities in automatic enrichment of spoken corpora, enabling researchers to analyze not just the linguistic content, but also the accompanying prosodic and paralinguistic signals such as disfluencies, prosodic variation, speaker-specific characteristics, laughter or gestures (Christodoulides et al., 2014; Nasr et al., 2014; Dobrushina and Sokur, 2022; Wu et al., 2022; Lemmenmeier, 2023). Just recently multi-layer automation approaches with high-quality results started to emerge (Liao et al., 2025), while there is a consensus in the community that automatic multilayer annotations of spoken corpora are very much needed (Wieczorkowska, 2025).

We organize the remainder of this section by types of enrichment layers applied inside this work.

Alignment. Transcript-aligned corpora are the most common type of corpora. All the parliamentary spoken corpora presented in Virkkunen et al. (2023) have been aligned on the level of instances of medium length, mostly up to a duration of 30 seconds, to streamline development of ASR systems. Most parliamentary corpora are aligned via time-stamped output of ASR, which is a practical solution given that ASR output on the spoken modality is needed to align the spoken modality with the official parliamentary transcripts. However, by now it is almost common knowledge that forced aligners such as the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) significantly outperform ASR-based time alignments (Rousso et al., 2024). Moreover, grapheme- or phoneme-level aligned corpora are rare, nne such corpus being the British National Corpus (Love et al., 2017).

In this paper we apply a MFA-based word-level and grapheme-level alignment model (Ljubešić et al., 2022) on two out of our four languages, showing significant improvements to the baseline ASR-based word alignment.

Primary word stress. Corpora containing stress labels for vowels or syllables are exceedingly rare, especially automatically annotated. One such corpus is the ISLE Corpus (Atwell et al., 2003), consisting of Italian and German learners' spoken English (18 hr), that contains word, phone and stressed syllable alignments done manually by experts. Another such corpus is the Russian National Corpus, specifically in the Accentological corpus (135 Mw). However, audio files are not available and their approach is based on software that accents words using a built-in lexicon, again followed by human expert review and correction (Savchuk, 2009).

Recently, a spoken dataset of English disyllabic words was automatically constructed from three datasets by using CNN-based models of accuracy around 90% (Allouche et al., 2025).

On our datasets, we apply a recently developed model based on a transformer speech encoder, which shows excellent performance in the two languages processed with word-level accuracy of 99% (Ljubešić et al., 2025b).

Filled pauses. Corpora with labeled disfluencies tend to be more available than primary stress corpora. However, these corpora often do not comprise of typical speakers. For example, The FluencyBank Timestamped (FBTS) (Romana et al., 2024) corpus contains 5.3 hours of annotated German speech from people who stutter, and the English SEP-28k (Lea et al., 2021) corpus consists of stuttering speakers with segment-level annotations of filled pauses.

Regarding corpora of typical speakers, the Switchboard corpus (Godfrey et al., 1992) has manually added disfluency markers, including filled pauses. Zayats et al. (2019) expand on the corpus, by using an automated approach to identify and add missed disfluency annotations.

In this work, we apply, similar as with the primary word stress layer, a recently developed model based on speech transformers, which achieves performance comparable to that of human annotators in all languages covered in this dataset (Ljubešić et al., 2025a).

Sentiment. Sentiment analysis determines the expressed stance towards a target, and has been a hugely popular task on text (Mao et al., 2024). Recently multimodal approaches have gained significant traction, which also includes speech (Das and Singh, 2023).

Regardless of a significant number of training datasets and benchmarks for sentiment identification on speech and the corresponding transcripts (Kaushik et al., 2015; Shon et al., 2022),

large spoken corpora annotated with sentiment information are almost non-existent. One example of such a corpus is the extension of Switchboard with manual sentiment labels (Chen et al., 2020), covering 140 hours of audio.

On the datasets presented in this paper, we exploit a recently developed multilingual text transformer model specifically developed for the parliamentary domain, showing very strong results on two out of our four languages, with comparable results on languages not seen during fine-tuning (Mochtak et al., 2024).

3. Dataset

The ParlaSpeech dataset is a derivative of the ParlaMint (Erjavec et al., 2025) corpora, the result of the CLARIN ERIC flagship project, which provided comparable encodings of transcripts of 29 European national and regional parliaments, together with rich speaker metadata. ParlaSpeech extends ParlaMint with recordings from parliaments aligned on the sentence level via an alignment procedure described in Ljubešić et al. (2022, 2024). The added value of the third version of the corpus is its speech and text enrichments on various levels, described in the next section.

ParlaSpeech 3.0 covers four parliaments: Croatian (HR), Czech (CZ), Polish (PL) and Serbian (RS). The size of each individual corpus and its composition is shown in Table 1. The dataset consists overall of around 6 thousand hours of speech, a large number of different speakers with good coverage of both genders. In terms of enrichments, all four corpora have rich speaker metadata from ParlaMint (Erjavec et al., 2023), sentiment predictions, linguistic labels and filled pause predictions. At this point, only the Croatian and the Serbian parliament have forced alignment and primary stress predictions.

The content distribution regarding speakers' gender, year of birth and age at the time of the recording are presented in Figure 1. The plots in the first row show the distribution of speakers regarding their year of birth and gender, while the second row shows the distribution of the content (in terms of number of words pronounced) given the gender and age at the time of recording. Each distribution is normalized on the parliament level, with that showing imbalances on the gender and age level inside and across the corpora. We can observe from the plots that there is an expected gender imbalance, which seems to be a little more emphasized on the content level than the speaker level. Regarding the age distribution, it seems to be reasonable and expected, with a median age of 50, and infrequent younger speakers.

4. Dataset Enrichments

ParlaSpeech 3.0 extends the base ParlaSpeech dataset (Ljubešić et al., 2022, 2024) with five annotation layers: linguistic annotations (ParlaSpeech-Ling following Universal Dependencies), sentiment (ParlaSpeech-Senti), filled pause detection (ParlaSpeech-Pause), precise word-level and grapheme-level alignments (ParlaSpeech-Align), and primary word stress markers (ParlaSpeech-Stress). Currently, Croatian and Serbian corpora contain all annotation layers, while Czech and Polish include only filled pause, sentiment and linguistic annotations. All the structured enrichments facilitate advanced research in interaction of linguistics, prosody, speech disfluencies, political science, and multimodal parliamentary analysis.

4.1. ParlaSpeech-Ling

For linguistic annotation of all four corpora state-of-the art tools were used. For Croatian and Serbian the CLASSLA-Stanza tool was applied (Terčon and Ljubešić, 2023), while for Czech and Polish the Stanza tool was used (Qi et al., 2020). Both tools report very high annotation quality for standard text.

CLASSLA-Stanza performs for Croatian on level of morphosyntax (XPOS) accuracy of around 94%, while parsing (LAS) accuracy is around 87%. For Serbian text, it achieves an XPOS accuracy of about 96% and LAS accuracy of 90 % (Terčon and Ljubešić, 2023). Stanza reports for Czech an XPOS accuracy of around 95% and LAS accuracy of around 89%, while for Polish XPOS accuracy is around 95% and LAS accuracy around 92%[2]. Given the varying complexity of the used test data, we can overall expect a morphosyntactic performance of around 95% accuracy, with a syntactic parsing accuracy of around 90%.

4.2. ParlaSpeech-Senti

For annotating the sentiment layer, the domain-specific ParlaSent model (Rupnik et al., 2023) has been applied on the textual modality. The model is based on an additionally pre-trained XLM-R-large model on 1.7 billion words of parliamentary proceedings (Ljubešić et al., 2023), which was then fine-tuned on 13 thousand instances from English, Bosnian, Croatian, Czech, Serbian and Slovak manually annotated sentences (Mochtak et al., 2023), discriminating between six levels of sentiment. Important is that ablation experiments showed the quality of the automatic annotations to be comparable regardless of whether the test language was present in the fine-tuning data

[2] https://stanfordnlp.github.io/stanza/performance.html

Table 1: Detailed overview of the ParlaSpeech 3.0 corpora. The table presents the size (in hours and words), speaker distribution, and key annotations (filled pauses, grapheme alignment, and stress labels) for the Croatian, Czech, Polish, and Serbian datasets, covering the specified time ranges.

Size in	**Croatian**	**Czech**	**Polish**	**Serbian**	**Total**
Hours	3 110	1 218	1 009	896	6 233
Words	24,755,742	9,114,266	7,515,333	7,024,294	48,409,635
Sentences	922,679	717,682	535,465	290,778	2,466,604
Speakers (M/F)	348M / 145F	431M / 124F	610M / 226F	398M / 230F	1787M / 725F
Political parties	32	15	10	50	107
Filled pauses	323,514	205,528	196,671	73,861	799,574
Grapheme-level aligned words	17,765,585	–	–	3,138,747	20,904,332
Stress labels	11,341,813	–	–	1,998,020	13,339,833
Time range	2015-2022	2013-2023	2017-2022	2013-2022	2013-2023

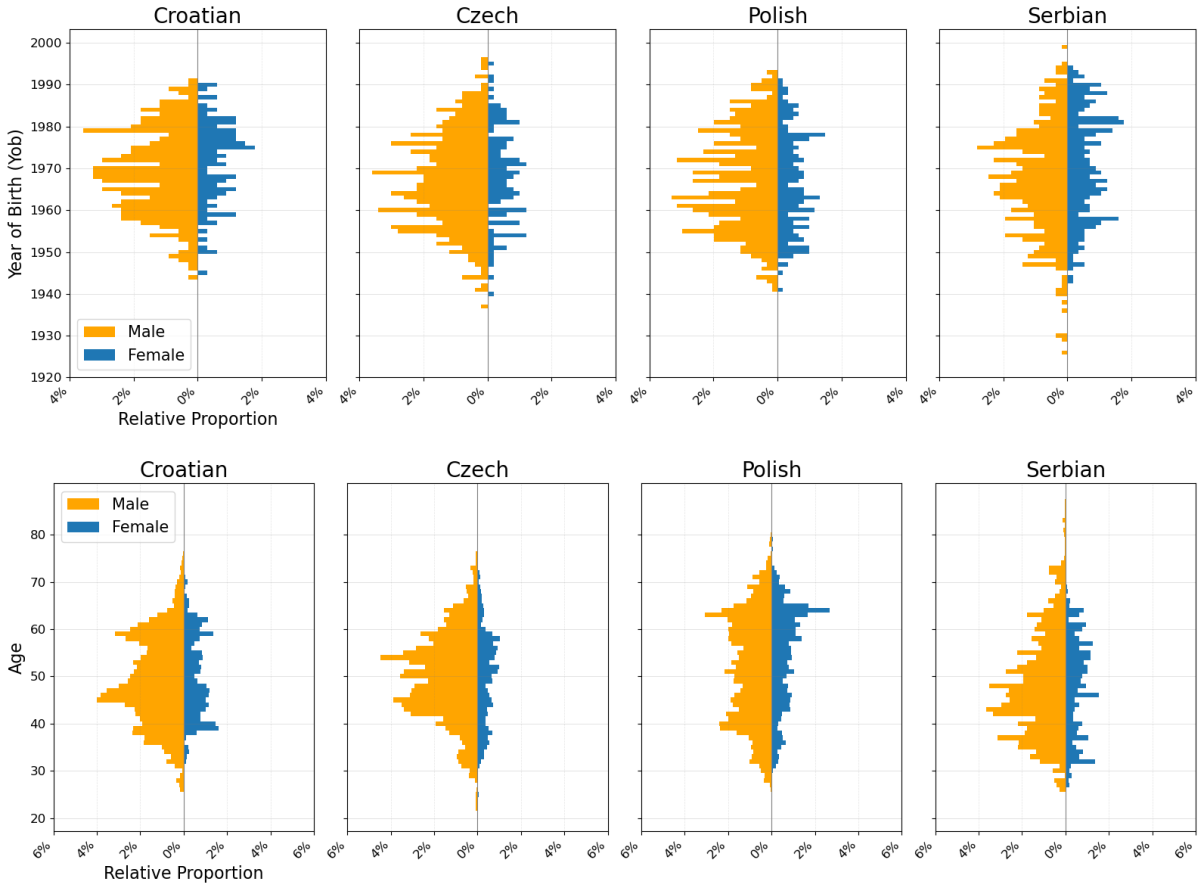

Figure 1: Relative distribution of the number of speakers by year-of-birth and gender across all four parliaments (top), and relative distribution of the number of words spoken by speakers across age, gender, and parliaments (bottom).

or not (Mochtak et al., 2024). In addition, when comparing this model to a state-of-the-art proprietary LLM, both show comparable performance, with the Spearman correlation coefficient twice as high as for a dictionary approach. The ParlaSent model achieved an average correlation of 0.786, the GPT-4o model of 0.807, while the dictionary approach (Young and Soroka, 2012) achieved an average correlation of 0.408. For more details see Mochtak et al. (2025).

4.3. ParlaSpeech-Pause

Filled pauses were annotated with the wav2vecbert2 model[3] (Rupnik et al., 2024), with a 20 ms audio frame classification head fine-tuned on the the ROG dataset for training speech processing technologies (Verdonik et al., 2024). Evaluation of the model on the four languages covered here shows very high performance, with F1 for Croatian of 0.913, Czech of 0.874, Polish of 0.924, and Serbian of 0.940. What is even more interesting, these F1 metrics touch on human performance measured through inter-annotator agreement. Once the disagreement between human and automatic annotation is compared, the automatic approach actually shows to perform better, with a significantly lower number of false negatives, but with a minor increase in false positives (Ljubešić et al., 2025a).

4.4. ParlaSpeech-Align

The ParlaSpeech-Align layer is added via a Montreal Forced Aligner (MFA) model trained on an early version of the Croatian ParlaSpeech corpus (Ljubešić et al., 2022). To measure the quality of this alignment, it was manually evaluated by an expert phonetician on a random sample of 324 multisyllabic words. For each word, the alignments produced by MFA and ASR (available in all four languages as a byproduct of the speech and text alignment procedure described in Ljubešić et al. (2024)) were compared with the underlying audio. Deviations of aligned boundaries relative to the actual spoken word were classified as Perfect ($<$30 ms), Close (30–100 ms), Bad ($>$100 ms) or Fail (does not match audio). The thresholds reflect practical usability: shifts $<$30 ms are acceptable for phonetic research, while errors $>$100 ms are considered unreliable for that use. However, even the "Bad" alignment should be very much sufficient for retrieval purposes. The results of the comparison of the two word-level alignment approaches are shown in Table 2.

Table 2: Evaluation of word-level alignments by Montreal Forced Aligner (MFA) and Automatic Speech Recognition (ASR).

	Perfect $<$30 ms	Close	Bad $>$100 ms	Fail
MFA	85.8%	8.3%	5.3%	0.6%
ASR	39.5%	28.7%	30.6%	1.2%

We can very clearly observe that ASR word alignment is decent, with 99% of words at least roughly corresponding to the recording, but still around 30% of the alignments not being suitable for phonetic research. On the other hand, the MFA output has only 6% of words that are not suitable for phonetic research, with around 85% of alignments considered perfect. Most ASR errors were caused by silent pauses, speaker noise, or difficulty locating boundaries in fricatives and nasals. MFA performed more consistently, but struggled with disfluencies such as filled pauses, repetitions, and repairs.

4.5. ParlaSpeech-Stress

The primary word stress layer has been applied over Croatian and Serbian data, as the underlying enrichment technology (Ljubešić et al., 2025b) currently supports only these two languages. The enrichments have been added on top of the ParlaSpeech-Align layer, i.e., grapheme-level alignments. The model used to identify primary stress is a wav2vecbert2 model (Barrault et al., 2023) fine-tuned on manually annotated 10 thousand words from the Croatian ParlaSpeech corpus (Ljubešić et al., 2025). The model was evaluated on both Croatian and Serbian parliamentary data, in both cases showing word-level accuracy above 99% (Ljubešić et al., 2025b). The primary stress predictions were encoded as raw predictions on 20 ms audio frames, as well as a syllable-level annotation, with the longest raw span prediction assigned to the closest syllable nucleus.

5. Dataset Encoding

To facilitate downstream research, we have encoded the dataset into three different formats.

JSONL. The primary format of the corpora is the JSONL (JavaScript Object Notation Lines) file format that contains all the layer data, including speaker metadata. This format is uniquely suited for computational processing because each line is a self-contained, valid JSON object representing a sentence. This line-by-line structure allows for highly efficient data streaming and parallel processing of the entire dataset without the need to load massive files into memory. Furthermore, the hierarchical nature of JSON logically organizes the complex layer annotations alongside essential speaker metadata. This structure provides both standardization for cross-lingual research and the flexibility to accommodate future expansions or additional data layers. Documentation on the full dataset schema (and more) is available in the online documentation [4].

TextGrid. The JSONL entries have been converted into individual Praat TextGrid files (Boersma

[3] https://huggingface.co/classla/wav2vecbert2-filledPause

[4] https://clarinsi.github.io/parlaspeech/

and Weenink, 2025), a format particularly suited for phonetic analysis. These files map phonetic information onto a set of time-aligned, named tiers. Specifically, the TextGrids include the ParlaSpeech-Align tiers , the ParlaSpeech-Stress tier , and the ParlaSpeech-Pause tier . This format includes the Croatian and the Serbian language for which the enrichments are available. This format makes the dataset immediately convenient and user-friendly for phoneticians and linguists who rely on the Praat software. Importantly, TextGrids are easily editable and extensible, enabling researchers to add new tiers and annotations for their own downstream analyses. We hope that researchers will use this opportunity and provide the community with additional data for analysis and modeling purposes.

Concordancer. The corpora are also searchable through the online NoSke concordancer, hosted on the CLARIN.SI infrastructure[5]. The main purpose of the concordancers is to allow researchers to easily search and explore the corpus data and metadata using the intuitive NoSke interface. This approach allows users to search for specific words, lemmas, and part-of-speech tags, or use Corpus Query Language (CQL) for more advanced pattern matching. Crucially, it is also possible to filter results using text types, allowing researchers to restrict searches by attributes such as gender, age, and sentiment. Each concordance can be also played back, either as a recording of the whole sentence, or as a recording of a chunk up to 6 seconds in length. For ease of use, we provide a tutorial on how to leverage the full functionality of the concordancers[6].

6. Use Case

In this section, we discuss one of the many possible use cases of the ParlaSpeech 3.0 dataset – investigating the differences in acoustic features between speech with positive and speech with negative sentiment. Does a speaker's intonation (pitch), loudness (intensity) or speech rate vary in different sentiments?

This use case cross-references the sentiment layer with the acoustic layer, and there is a significant number of similar layer interactions that can answer relevant research questions. One example are the filled pause layer and the syntactic parsing layer, which enables inspecting the syntactic role of filled pauses. Another one is the primary word stress layer in Croatian and that of the speaker metadata, investigating the consistency of speakers in using one of the two dominant word stress systems in that language.

6.1. Data Selection

To follow upon the question tackled in this section, namely the interaction of sentiment and acoustic features, we first selected only the instances with clear positive or negative predicted sentiment, omitting instances with mixed sentiment. We next filtered the remaining data to include only utterances that were 10 to 40 words long. This length constraint was applied to ensure both a good prediction of sentiment and a consistent sentiment across the instance.

Finally, we retained only instances of those speakers who had at least 50 positively and 50 negatively labelled instances to allow for good per-speaker estimates.

6.2. Feature Extraction

With the instance and speaker selection completed, we moved on to feature selection. To quantify intonation, we extracted the pitch (F0, in Hz) envelope (time series of values). Jitter, the measure of vocal (F0) perturbation, was also selected as an accompanying feature to intonation. To quantify loudness, we extracted the intensity (in dB) envelope. We also extracted shimmer, a measure of variability in the amplitude, representing variation in the intensity signal.

The selected four acoustic features were extracted using Praat and two OpenSMILE extractors (GeMAPS (Eyben et al., 2015) and ComParE_2016 (Schuller et al., 2016)). The two OpenSMILE extractors served as a control for Praat, showing highly comparable measurements. In the remainder of the section, only results for Praat extracts are shown. We also investigated one timing feature, namely speech rate. Speech rate (in syllables/s) was calculated as the number of syllable nuclei (vowels) in a word, divided by the word duration, excluding silent pauses.

Feature extraction was first performed over the entire instance, creating a feature envelope. From the entire envelope, only intervals of actual words were retained, skipping over silence and possible noise. These word feature envelopes were then averaged using the median to further counter noise in specific values of the word-level feature envelope. Finally, all word medians, within a single instance, were mean-averaged into one feature value to describe that particular instance.

6.3. Statistical Analysis

When comparing the feature values calculated on instances with positive and negative sentiment, we always rely on measurements from the same

[5] https://www.clarin.si/ske/#open

[6] https://clarinsi.github.io/parlaspeech/concordancer/concordancer-guide.html

Table 3: Statistical analysis results on the dependence of acoustic and temporal features on the sentiment of the sentence. Speaker average and instance-level measurements are given.

SPEAKER AVERAGE				INSTANCE-LEVEL			
Feature	**p-value**	**RBC**	**P(Neg>Pos)**	**Feature**	**p-value**	**RBC**	**P(Neg>Pos)**
Croatian				**Croatian**			
F0	**0.0000**	0.9940	0.9663	F0	**0.0000**	0.3590	0.6292
Intensity	**0.0000**	0.5700	0.7260	Intensity	**0.0000**	0.0890	0.5320
Jitter	**0.0227**	0.1120	0.5673	Jitter	0.0509	0.0190	0.5134
Shimmer	**0.0000**	-0.3040	0.3798	Shimmer	0.1090	-0.0250	0.4921
Speech rate	**0.0000**	0.4020	0.6683	Speech rate	**0.0001**	0.0560	0.5140
Czech				**Czech**			
F0	**0.0000**	0.8780	0.8678	F0	**0.0000**	0.1790	0.5600
Intensity	**0.0002**	0.2330	0.5702	Intensity	**0.0162**	0.0520	0.5198
Jitter	0.5672	0.1610	0.5702	Jitter	0.5305	0.0160	0.5067
Shimmer	**0.0000**	0.3030	0.6116	Shimmer	**0.0226**	0.0440	0.5140
Speech rate	**0.0000**	0.0970	0.5372	Speech rate	0.1806	-0.0200	0.4934
Polish				**Polish**			
F0	**0.0000**	0.9700	0.9427	F0	**0.0000**	0.3330	0.6163
Intensity	**0.0000**	0.7540	0.7834	Intensity	**0.0000**	0.1780	0.5703
Jitter	0.2458	0.1650	0.5987	Jitter	0.3924	0.0130	0.5085
Shimmer	0.6462	0.2240	0.5924	Shimmer	0.2052	0.0210	0.5056
Speech rate	**0.0000**	0.3800	0.6369	Speech rate	0.3845	0.0090	0.5012
Serbian				**Serbian**			
F0	**0.0000**	0.9560	0.9167	F0	**0.0000**	0.2840	0.5938
Intensity	**0.0000**	0.7090	0.8125	Intensity	**0.0000**	0.1540	0.5563
Jitter	0.3903	0.0530	0.5208	Jitter	0.5050	0.0290	0.5202
Shimmer	0.3953	-0.2230	0.3958	Shimmer	0.8632	0.0020	0.4994
Speech rate	**0.0000**	0.5030	0.6875	Speech rate	**0.0083**	0.0700	0.5185

Statistical Notes: Both analyses use the Paired Wilcoxon signed-rank test. The statistical value W has been left out for brevity.
P = Concordance Probability; $P = 0.5$ means no difference between sentiments, $P > 0.5$ means negatives tend to be higher than positives.
RBC = Rank-Biserial Correlation. A positive RBC value corresponds with an effect towards the Negative sentiment. p-values < 0.05 are shown in **bold**.

speaker. There are two reasons for this. The first reason is the high inter-speaker variation of the features in question, esp. the pitch between genders. The second reason is that with this approach, we obtain a paired dataset, which has higher statistical strength than unpaired datasets.

We compare the paired features on two levels: the speaker average level and the instance level. Speaker level analysis averages all positive instances to one value and all negative instances to another, thereby forming one single paired sample per speaker. Instance-level analysis pairs positive and negative utterances within speakers in a randomized fashion, collecting 50 paired samples per speaker.

For testing statistical significance of the two types of paired samples, the Wilcoxon signed-rank paired test (two-sided) was applied. For calculating the effect size, we used rank-biserial correlation (RBC) and concordance probability ($\mathrm{P}\left(\mathrm{Neg} > \mathrm{Pos}\right)$). RBC behaves like any correlation coefficient, ranging from -1 (strong negative correlation), via 0 (no correlation) to 1 (strong positive correlation). Regarding the concordance probability, for the speaker level, the probability states how likely is that the average feature value of a speaker in negative sentiment is higher than the average feature value of the same speaker, but in positive sentiment. For the instance-level, the probability states how likely is that a random negative-sentiment instance will have a higher feature value than a random positive-sentiment instance of the same speaker.

6.4. Results

Statistical analysis revealed consistent patterns in the acoustic correlates of sentiment across all four parliaments, as presented in Table 3. At

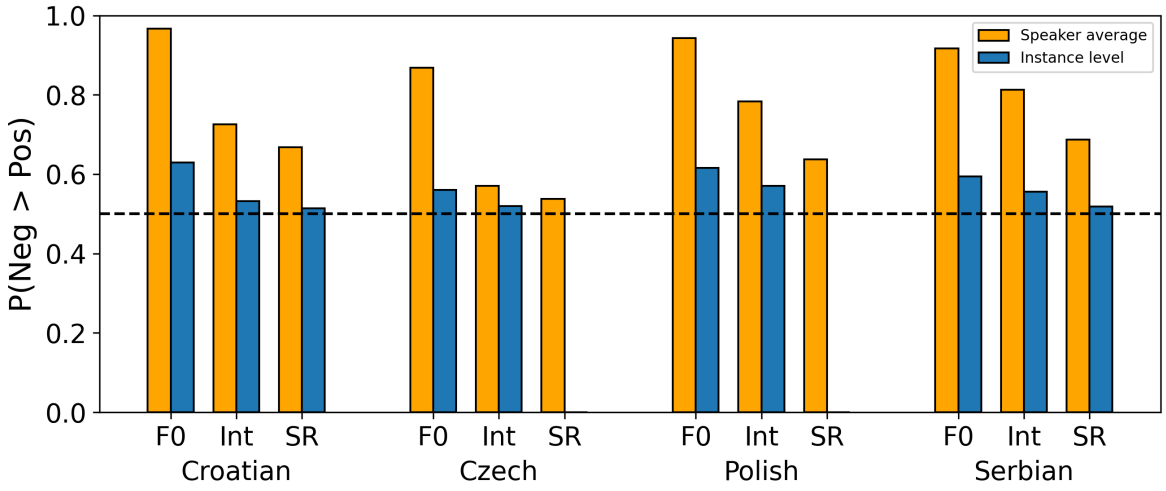

Figure 2: Visualization of the $\mathrm{P}\,(\mathrm{Neg} > \mathrm{Pos})$ effect size on three strong sentiment predictors – pitch (F0), intensity (Int) and speech rate (SR) – across the four languages. Speaker average and instance results are shown. Statistically non-significant results (instance-level speech rate in Czech and Polish) are omitted from the plot.

the speaker-average level, pitch (F0) and intensity emerged as the most robust indicators of sentiment, showing statistically significant differences ($p < 0.0001$) across all parliaments with mostly large effect sizes (correlation coefficients ranging from 0.878 to 0.994 for F0 and from 0.233 to 0.709 for intensity). At the instance level, both features remained statistically significant across all parliaments, though with low-to-moderate effect size (F0: RBC = 0.179—0.359; intensity: RBC = 0.052—0.178). Speech rate also demonstrated significant effects at the speaker level across all parliaments ($p < 0.0001$), with small-to-moderate effect sizes (RBC = 0.097—0.503) favoring faster articulation in negative utterances. However, at the instance level, only Croatian ($p = 0.0001$, RBC = 0.056) and Serbian ($p = 0.0083$, RBC = 0.070) maintained significance, while Czech ($p = 0.181$) and Polish ($p = 0.385$) did not show significant effects. In contrast, voice quality measures showed much less signal: jitter reached significance only in Croatian at the speaker level ($p = 0.023$, RBC = 0.112), while shimmer showed inconsistent effects, even producing negative correlations in some cases (Croatian: RBC = -0.304; Serbian: RBC = -0.223). While some consistency in jitter and shimmer values can be observed at the speaker-average level, any consistency disappears at the instance level, pointing to an overall weaker relationship between these two features and sentiment when compared to pitch, intensity, and speech rate.

To summarize the findings visually, we represent in Figure 2 the concordance probabilities $\mathrm{P}\,(\mathrm{Neg} > \mathrm{Pos})$ for the three features with consistently significant differences between the two sentiments, namely pitch (F0), intensity (Int), and speech rate (SR), across all four parliaments, at both speaker-average and instance levels. The speaker-average probabilities (orange bars) are consistently elevated above the 0.5 random baseline, with F0 showing the strongest effects (0.868 to 0.966), followed by intensity (0.570 to 0.813) and finally speech rate (0.537 to 0.688). The instance-level probabilities (blue bars) are notably lower across all features and parliaments, clustering closer to the 0.5 border of randomness.We omit instance-level speech rate results for Czech and Polish as they showed to be statistically insignificant ($p \geq 0.05$).

7. Conclusion

This paper presented the collection of parliamentary spoken corpora of four Slavic languages, spanning together more than six thousand hours in size. The textual modality of the corpora has been automatically enriched with linguistic and sentiment annotations. The spoken modality has been annotated with filled pauses in all four languages, while detailed word- and grapheme-level alignment and subsequent annotation of the primary stress in each multisyllabic word has been applied for two out of the four languages for which the technology is already available.

The corpora are available in various formats by following the FAIR paradigm[7] – JSONL for com-

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

pleteness, TextGrid for simple use by phoneticians and other speech scientists, and corpus concordancers for corpus linguists.

We wrapped up the paper with one use case among many potential ones for such a rich dataset – acoustic and temporal correlates of sentiment. The analysis shows consistent strong effects for higher pitch, strong to medium effects for increased intensity in negative content, and medium to low effects for increased speech rate in negative content.

In addition to a heavy use of the presented dataset in downstream research, we will continue developing this dataset in two directions: (1) adding additional languages and (2) adding additional annotation layers.

8. Ethical Considerations and Limitations

This paper presents a dataset comprising over six thousand hours of human speech in four less-resourced languages. Owing to the public nature of the recordings, we are able to freely enrich and share these data, thereby facilitating further enrichments and analyses by third parties. However, responsible use by downstream users remains essential to ensure the ethical and appropriate application of the dataset.

While the dataset provides an immense opportunity to study spoken communication in the four covered languages, it also entails certain limitations – including a narrow domain, limited spontaneity in speech production, and imbalances in gender and age representation.

9. Acknowledgments

This work was supported in part 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), all funded by the ARIS Slovenian Research and Innovation Agency, and the research project "Embeddings-based techniques for Media Monitoring Applications" (L2-50070), co-funded by the Kliping d.o.o. agency.

10. Bibliographical References

Itai Allouche, Itay Asael, Rotem Rousso, Vered Dassa, Ann Bradlow, Seung-Eun Kim, Matthew Goldrick, and Joseph Keshet. 2025. How Does a Deep Neural Network Look at Lexical Stress? *arXiv preprint arXiv:2508.07229*.

ES Atwell, PA Howarth, and DC Souter. 2003. The ISLE corpus: Italian and German spoken learner's English. *ICAME Journal: International Computer Archive of Modern and Medieval English Journal*, 27:5–18.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Paul Boersma and David Weenink. 2025. Praat: Doing phonetics by computer. [Computer program]. Retrieved 28 August 2025.

Eric Chen, Zhiyun Lu, Hao Xu, Liangliang Cao, Yu Zhang, and James Fan. 2020. A large scale speech sentiment corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6549–6555.

George Christodoulides, Mathieu Avanzi, and Jean-Philippe Goldman. 2014. DisMo: A morphosyntactic, disfluency and multi-word unit annotator. an evaluation on a corpus of French spontaneous and read speech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ringki Das and Thoudam Doren Singh. 2023. Multimodal sentiment analysis: a survey of methods, trends, and challenges. *ACM Computing Surveys*, 55(13s):1–38.

Nina Dobrushina and Elena Sokur. 2022. Spoken corpora of Slavic languages. *Russian Linguistics*, 46(2):77–93.

Tomaž Erjavec, Matyáš Kopp, Nikola Ljubešić, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, Çağrı Çöltekin, Danijel Koržinek, Katja Meden, et al. 2025. ParlaMint II: advancing comparable parliamentary corpora across Europe. *Language Resources and Evaluation*, 59(3):2071–2102.

Tomaž Erjavec, Matyáš Kopp, Maciej Ogrodniczuk, and Petya et al. Osenova. 2023. Multilingual comparable corpora of parliamentary debates ParlaMint 4.0. Slovenian language resource repository CLARIN.SI.

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202.

Darja Fišer, Jakob Lenardič, and Tomaž Erjavec. 2018. CLARIN's key resource families. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

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.

Lakshmish Kaushik, Abhijeet Sangwan, and John HL Hansen. 2015. Automatic audio sentiment extraction using keyword spotting. In *INTERSPEECH*, pages 2709–2713.

Andreas Kirkedal, Marija Stepanović, and Barbara Plank. 2020. FT Speech: Danish Parliament Speech Corpus. In *Interspeech 2020*, pages 442–446.

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.

Dolores Lemmenmeier. 2023. *Spoken language corpora: Approaches for facilitating linguistic research*. Ph.D. thesis, University of Zurich.

Huan Liao, Qinke Ni, Yuancheng Wang, Yiheng Lu, Haoyue Zhan, Pengyuan Xie, Qiang Zhang, and Zhizheng Wu. 2025. NVSpeech: An Integrated and Scalable Pipeline for Human-Like Speech Modeling with Paralinguistic Vocalizations.

Nikola Ljubešić, Danijel Koržinek, Peter Rupnik, and Ivo-Pavao Jazbec. 2022. ParlaSpeech-HR-a freely available ASR dataset for Croatian bootstrapped from the ParlaMint corpus. In *Proceedings of the workshop ParlaCLARIN III within the 13th language resources and evaluation Conference*, pages 111–116.

Nikola Ljubešić, Ivan Porupski, and Peter Rupnik. 2025a. Identifying Filled Pauses in Speech Across South and West Slavic Languages. In *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*, pages 1–8, Vienna, Austria. Association for Computational Linguistics.

Nikola Ljubešić, Ivan Porupski, and Peter Rupnik. 2025b. Identifying Primary Stress Across Related Languages and Dialects with Transformer-based Speech Encoder Models. In *Proceedings of Interspeech 2025*, pages 5768–5772.

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.

Nikola Ljubešić, Peter Rupnik, Ivan Porupski, Nejc Robida, and Mirna Potočnjak. 2025. Dataset for primary stress identification in Croatian and related languages and dialects. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Peter Rupnik, and Rik van Noord. 2023. XLM-R-Parla (Revision 8c7f9b2). Hugging Face.

Robbie Love, Claire Dembry, Andrew Hardie, Vaclav Brezina, and Tony McEnery. 2017. The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3):319–344.

Yanying Mao, Qun Liu, and Yu Zhang. 2024. Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University – Computer and Information Sciences*, 36(4):102048.

Michael McAuliffe, Michael Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Proceedings of Interspeech 2017*, pages 498–502.

Michal Mochtak, Peter Rupnik, Taja Kuzman Pungeršek, and Nikola Ljubešić. 2025.

ParlaSent: Mapping Sentiment in Political Discourse with Large Language Models. *Political Research Exchange*, 7(1).

Michal Mochtak, Peter Rupnik, and Nikola Ljubešić. 2024. The ParlaSent multilingual training dataset for sentiment identification in parliamentary proceedings. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16024–16036, Torino, Italia. ELRA and ICCL.

Michal Mochtak, Peter Rupnik, Katja Meden, and Nikola Ljubešić. 2023. The multilingual sentiment dataset of parliamentary debates ParlaSent 1.0. Slovenian language resource repository CLARIN.SI.

Alexis Nasr, Frederic Bechet, Benoit Favre, Thierry Bazillon, Jose Deulofeu, and Andre Valli. 2014. Automatically enriching spoken corpora with syntactic information for linguistic studies. In *LREC*, pages 854–858.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

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.

Rotem Rousso, Eyal Cohen, Joseph Keshet, and Eleanor Chodroff. 2024. Tradition or Innovation: A Comparison of Modern ASR Methods for Forced Alignment. In *Proceedings of Interspeech 2024*, pages 1525–1529.

Peter Rupnik, Nikola Ljubešić, and Michal Mochtak. 2023. XLM-R-ParlaSent (Revision a04de02). Hugging Face.

Peter Rupnik, Nikola Ljubešić, Ivan Porupski, and Darinka Verdonik. 2024. wav2vecbert2-filledpause (revision 5e75061). Hugging Face.

Svetlana Savchuk. 2009. Spoken texts representation in the Russian National Corpus: Spoken and Accentologic sub-corpora. In *NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference, Smolenice, Slovakia, 25-27 November 2009. Proceedings*, pages 310–320.

Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini. 2016. The INTERSPEECH 2016 computational paralinguistics challenge: deception, sincerity and native language. In *17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), Vols 1-5*, volume 8, pages 2001–2005. ISCA.

Suwon Shon, Ankita Pasad, Felix Wu, Pablo Brusco, Yoav Artzi, Karen Livescu, and Kyu J Han. 2022. Slue: New benchmark tasks for spoken language understanding evaluation on natural speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7927–7931. IEEE.

Per Erik Solberg and Pablo Ortiz. 2022. The Norwegian parliamentary speech corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1003–1008, Marseille, France. European Language Resources Association.

Luka Terčon and Nikola Ljubešić. 2023. CLASSLA-Stanza: The next step for linguistic processing of south slavic languages. *arXiv preprint*.

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.

Anja Virkkunen, Aku Rouhe, Nhan Phan, and Mikko Kurimo. 2023. Finnish parliament ASR corpus: Analysis, benchmarks and statistics. *Language Resources and Evaluation*, 57(4):1645–1670.

Alicja Wieczorkowska. 2025. Methodology for Obtaining High-Quality Speech Corpora. *Applied Sciences*, 15(4):1848.

Yaru Wu, Fabian Suchanek, Ioana Vasilescu, Lori Lamel, and Martine Adda-Decker. 2022. Using a knowledge base to automatically annotate speech corpora and to identify sociolinguistic variation. In *13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 1054–1360. European Language Resources Association (ELRA).

Lori Young and Stuart Soroka. 2012. Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication*, 29(2):205–231.

Vicky Zayats, Trang Tran, Richard Wright, Courtney Mansfield, and Mari Ostendorf. 2019. Disfluencies and human speech transcription errors. *arXiv preprint arXiv:1904.04398*.