



# Identifying Primary Stress Across Related Languages and Dialects with Transformer-based Speech Encoder Models

Nikola Ljubešić<sup>1,2,3</sup>, Ivan Porupski<sup>1</sup>, Peter Rupnik<sup>1</sup>

<sup>1</sup>Department of Knowledge Technologies, Jožef Stefan Institute, Slovenia

<sup>2</sup>Faculty of Computer and Information Science, University of Ljubljana, Slovenia

<sup>3</sup>Institute of Contemporary History, Ljubljana, Slovenia

nikola.ljubesic@ijs.si, ivan.porupski@ijs.si, peter.rupnik@ijs.si

## Abstract

Automating primary stress identification has been an active research field due to the role of stress in encoding meaning and aiding speech comprehension. Previous studies relied mainly on traditional acoustic features and English datasets. In this paper we investigate the approach of fine-tuning a pre-trained transformer model with an audio frame classification head. Our experiments use a new Croatian training dataset, with test sets in Croatian, Serbian, the Chakavian dialect, and Slovenian.

By comparing an SVM classifier using traditional acoustic features with the fine-tuned speech transformer, we demonstrate the transformer's superiority across the board, achieving near-perfect results for Croatian and Serbian, with a 10-point performance drop for more distant Chakavian and Slovenian. Finally, we show that only a few hundred multi-syllabic training words suffice for strong performance. We release our datasets and model under permissive licenses.

**Index Terms:** primary stress detection, pre-trained encoder models, South Slavic languages

## 1. Introduction

Primary stress is a feature of each multi-syllabic word, where one syllable is perceived to stand out from its environment [1]. It has an important and varying function in different languages, including distinguishing word meaning and function, aiding speech comprehension, as well as communicating various sociolinguistic cues [2, 3, 4].

Automating the identification of the position of primary stress has attracted significant amount of previous work. Traditionally, the task was performed via supervised learning over acoustic features, while recently transformers pre-trained on speech started to be used, but only for probing experiments [5] and feature extraction [6]. Most of the intended use cases and datasets are related to computer-assisted language learning [1, 7], speech synthesis [8], and some applications in research of children's speech [9, 10]. The vast majority of work has been performed on English, with very infrequent exceptions such as German [11] and Arabic [12].

In this work, we investigate the capacity of pre-trained speech transformer models to predict the position of the primary stress in a multi-syllabic words. We achieve that by fine-tuning the transformer network on that task. Our imminent goal is to apply the resulting technology to describe the variation in spoken language by annotating large spoken corpora, but we also plan more specific downstream use cases, including child language and atypical speech processing, as well as language learning. We break from the tradition of English-centric research by building one training and four test sets in various South Slavic languages and dialects. With this setup, we also in-

vestigate the transferability of the proposed technology to similar languages and dialects.

Our training and one of the test languages is Croatian, particularly interesting for the task due to its high dialectal variability, which results in a varying position of the stress even in official communication. Croatian spreads across the Shtokavian, the Kajkavian and the Chakavian dialectal group. Serbian, our test language, mutually intelligible with standard Croatian, has less variability due to its dominant dialect, Shtokavian, which was used in the standardization of both Croatian and Serbian. Slovenian, another of our test languages, is standardized somewhat closely to the Croatian Kajkavian dialect. Finally, our test dialect is the Chakavian dialect, not present in the Croatian standard, nor as close to the Slovenian standard as Kajkavian [3, 4, 13].

Our primary contributions are the following: (1) we show excellent performance of the pre-trained speech transformers fine-tuned to the task of primary stress identification, (2) we investigate the limitations of the technology when applied to related languages and dialects, (3) we show that supervision of pre-trained models on a few hundred words already yields comparable results to those obtained after fine-tuning the model on ten thousand words, (4) we release a new training dataset and four test sets for related South Slavic languages and dialects, as well as a strong model for Croatian and Serbian<sup>1</sup>.

## 2. Related work

The traditional way of performing primary stress identification has been to use prosodic features, such as nucleus duration, intensity and pitch (F0) [1, 7, 12], as well as the sonority-based TCSSBC feature contour [14, 15].

These features are mostly exploited in a supervised machine learning setup, with infrequent takes on unsupervised approaches [16]. Most approaches use pre-neural classifiers such as Gaussian Mixture Models [1, 8, 7], Hidden Markov Models [17] and Support Vector Machines [8, 15], with some neural exceptions [9, 10, 12]. The performance reported from these experiments on various datasets ranges from 72 to 93% word-level accuracy.

Recently, pre-trained speech transformer models have been applied on the task as well, but only either for probing these networks for primary stress signal [5], or for feature extraction for various traditional and neural classifiers [6].

The probing experiments in [5] report promising results for the availability of the primary stress signal in neural representations, showing that CNN feature extractors already contain a level of relevant signal similar to traditional acoustic features,

<sup>1</sup>Data and model are available at  
<https://doi.org/10.57967/hf/5658>.

with higher transformer layers being significantly more predictive of the phenomenon.

The classification experiments in [6] compare traditional acoustic features with syllable-averaged neural representations from pre-trained speech transformers, with neural representations being more informative for the task. These experiments also compare traditional and neural classifiers, showing that the latter are more potent on the task. They, however, miss on the opportunity to fine-tune the transformer model to the task directly, which also introduces information loss while averaging the neural representations, available otherwise for each 20 ms frame, over the span of each syllable.

### 3. Data

#### 3.1. Sources

We construct new training and test datasets by exploiting recently released open datasets in three South Slavic languages and one dialect.

The Croatian `ParlaStress-HR` training and test datasets are constructed from a sample of the `ParlaSpeech-HR` open dataset of sentence-aligned parliamentary recordings and transcripts of the Croatian parliament [18]. Transcription sentences are sampled to assure the diversity of speakers and gender balance. We split the dataset into a training portion and a test portion, ensuring no speaker overlap while maintaining gender balance.

The Serbian `ParlaStress-SR` test set is built from the `ParlaSpeech-RS` dataset, another member of the `ParlaSpeech` collection of speech and text datasets [18], sampling transcript sentences to ensure maximum diversity of speakers, while ensuring gender balance.

The Chakavian `MiçiPrinc-CKM` test dataset is a sample of two chapters from the printed and audio book of the translation of *Le Petit Prince* into the Chakavian dialect. This multi-modal book has recently been released as an open dataset with word-level-aligned text and audio [19].

The Slovenian `Artur-SL` test set is sampled from the `ARTUR` dataset [20], part of the recently updated GOS corpus of spoken Slovenian [21]. Three speakers are sampled, one from a public, another from a private setting, and a third one from the parliamentary setting.

#### 3.2. Data pre-processing

To enable the manual annotation and subsequent modeling on the level of syllable nuclei, we perform a grapheme-level alignment on all three data sources except `Artur-SL`, which already had grapheme-level alignment present [22]. Phonemic transcription is not needed, except for three simple replacement rules that cover digraphs, due to the usage of phonemic orthography in all the languages and dialects addressed. We align the three datasets by using the forced alignment model from the Kaldi toolkit [23] that has previously been released as part of the initial `ParlaSpeech` dataset construction efforts [24].

#### 3.3. Manual data annotation

Each dataset is annotated by a native speaker using Praat [25] TextGrids, with syllable nuclei of multi-syllabic words pre-selected as candidates for annotation. The annotators are instructed to select one of the syllable nuclei in each multi-syllabic word as the primary stress. In rare cases of deviating transcripts or alignment errors, annotators are instructed to label

them with dedicated symbols, and they are not included in the final dataset.

To measure the subjectivity of the task at hand and perform quality control over the obtained manual annotations, we double annotate the whole `MiçiPrinc-CKM` dialectal test set, given the general agreement among the annotators and authors that primary stress is the hardest to determine in that dataset. We obtain a high observed word-level agreement of 96.2% and Krippendorff  $\alpha$  [26] of 0.92, which proves the quality of our annotations, but also the straightforwardness of the task for humans. Such high levels of inter-annotator agreement on language data are otherwise very rarely observed [27].

The final size of each of the datasets in terms of the number of syllables, multi-syllabic words and speakers is given in Table 1. The size of the training dataset follows similar English datasets, such as ISLE [28], while our test sets are large enough for a reasonable performance estimate of various models, as proven by confidence intervals reported in Section 5.

Table 1: Overview of the size of the train and the test datasets used. Dataset suffixes encode language or dialect (HR Croatian, SR Serbian, CKM Chakavian, SL Slovenian).

Dataset	Syllables	Words	Speakers
Training dataset			
ParlaStress-HR	30561	10443	46
Test datasets			
ParlaStress-HR	3843	1291	8
ParlaStress-SR	1766	580	40
MiçiPrinc-CKM	760	324	4
Artur-SL	382	136	3

#### 3.4. Data analysis

Before moving on to the use of these datasets in machine learning experiments, we performed two short analyses to gain a better understanding of the newly developed datasets.

The first analysis is directed at measuring the variation in the position of the stress in identically spelled words inside Croatian training data, which is one of the main motivations for this work. Already in our training dataset of slightly more than 10 thousand words, if we consider words that occur at least five times, around 6% of words have a varying position of the primary stress. By inspecting these words manually, we confirmed that the variation is not due to a different part-of-speech category or homography, but rather due to the expected variation in pronunciation of the same words.

The second analysis aims to measure the similarity of the three cross-lingual test sets to the Croatian training dataset in terms of the position of the primary stress in identically spelled words. The `ParlaStress-SR` dataset has an expected and significant lexical overlap of 305 words (53%), only 6 (2%) of them having a stress position not observed in the training data. The `MiçiPrinc-CKM` dataset has 63 identical words, 9 (14%) have a stress position not covered in the training data, showing a more significant deviation of the stress position than the Serbian dataset. Finally, the `Artur-SL` dataset has only 20 words covered by the training data, but 13 (65%) of these have a different position of the primary stress, accentuating the large differences in the stress positions between Croatian and Slovenian.

## 4. Methods

### 4.1. Pre-trained transformer model

Our solution for the problem at hand is the fine-tuning of the w2v-bert-2.0 model<sup>2</sup> with an audio frame classification head on top of the transformer model [29], allowing for every 20 ms frame to be classified into a specific category. The raw transformer model is a 580-million-parameters conformer model which was pre-trained on 4.5 Mh of unlabeled audio data covering more than 143 languages, and has shown state-of-the-art results in speech translation and transcription tasks, especially on less-resourced languages [30].

We transform each multi-syllabic word into a sequence of 20 ms frames, every frame being labeled as 0, except during the manifestation of the nucleus of the stressed syllable, where the label is 1. With this, we set our problem as a binary classification task on each audio frame.

Optimal hyperparameters are identified based on the hyperparameter search on our training data. We have investigated learning rates of  $\{8 \times 10^{-6}, 1 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}\}$ , number of epochs ranging from 1 to 20, and the impact of 1 or 4 gradient accumulation steps. These preliminary experiments showed that the learning rate of  $1 \times 10^{-5}$ , no gradient accumulation, and 20 epochs, our batch size being 32, deliver highly stable results on various splits of our training set. We fine-tune our model on an NVIDIA A100 with 40 GB of memory. Fine-tuning for one epoch takes 4.5 min.

### 4.2. Support Vector Machine model

To compare our approach with traditional methods, we train an SVM model with an RBF kernel and  $C = 10$ , using prosodic features. We consider each syllable nucleus as an instance, for which we calculate the prominence of intensity, pitch, and sonority by dividing the nucleus area under the curve (AUC), mean, and peak values by the corresponding word-level mean. This results in a total of nine input features, plus syllable nucleus duration. Binary classification is performed on each syllable nucleus.

For this classifier we also performed a number of hyperparameter search and feature selection experiments across the training dataset, the hyperparameters ranging between an RBF and linear kernel, and  $C$  selected from  $\{0.1, 1, 10, 100\}$ .

### 4.3. Evaluation

We evaluate each model on word-level accuracy. The predictions of each model are post-processed to ensure that only one syllable nucleus per word is selected as the most likely position of the primary stress.

In case of the transformer classifier, the syllable nucleus closest to the longest range of a span of positive predictions is selected as the final prediction. In very infrequent cases, multiple spans within a word are predicted to be primary stress positions.

For support vector machines, the syllable with the highest positive-class probability on the word level is selected as the final prediction.

## 5. Results

### 5.1. Traditional vs. deep features

In the first experiment, we compare the performance of the SVM classifier, trained on traditional acoustic features, with that of the transformer model. Each classifier was trained on the whole training dataset. The results of each classifier are given in Table 2.

The results show significant dominance of the transformer models over the SVM models, with word accuracy differences between 11 and 25 percentage points. However, a significant robustness of the traditional features can be observed as well, with a significantly smaller difference between the results on the various test sets, regardless of the distance to the training data, as discussed in Section 3.4.

Both the Croatian and Serbian test sets show to be relatively simple for both methods, but with a drastic difference in performance of less than one percent of error for the transformer and 20% or more of error for the SVM.

On the Chakavian and Slovenian test sets, the transformer model achieves ten to twelve points lower performance than on the two other datasets. This is in line with the findings in Section 3.4, which show a high level of similarity between Croatian and Serbian, and a decreasing similarity of Chakavian, followed by Slovenian. However, regardless of this drop in performance, and the robustness of SVMs across the test sets, transformers still outperform SVMs with more than 10 accuracy points difference.

Table 2: Comparison of the word-level accuracies of the SVM and the transformer model with 95% confidence intervals.

Dataset	SVM		w2v-bert-2.0	
	acc	95% CI	acc	95% CI
ParlaStress-HR	74.0	[71.6,76.3]	99.1	[98.6,99.6]
ParlaStress-SR	80.2	[77.1,83.3]	99.3	[98.6,99.8]
MiGiPrinc-CKM	78.7	[73.8,83.0]	88.9	[85.2,92.3]
Artur-SLO	72.1	[64.0,79.4]	89.0	[83.1,94.1]

### 5.2. Stress position

In this set of experiments we investigate whether the decreasing performance of transformer models on the Chakavian and Slovenian test sets are due to the model’s bias towards a specific stress position introduced by fine-tuning on the Croatian dataset. We calculate confusion matrices between the true and the predicted syllable position from our transformer evaluation results on each of the four test sets. The results are depicted in Figure 1.

In the Croatian and Serbian test set there is an obvious preference for the first syllable of 78% cases, while that preference drops to 66% for Chakavian and 43% on Slovenian. The wrong predictions on both Chakavian and Slovenian are mostly due to the first-syllable stress being preferred.

Additionally, we performed a short manual qualitative analysis of the erroneously classified words in Chakavian and Slovenian. In the Chakavian dataset the most frequent reason for misclassification was a less clear position of the primary stress, while in the Slovenian dataset the most frequent reason was the

<sup>2</sup><https://huggingface.co/facebook/w2v-bert-2.0>

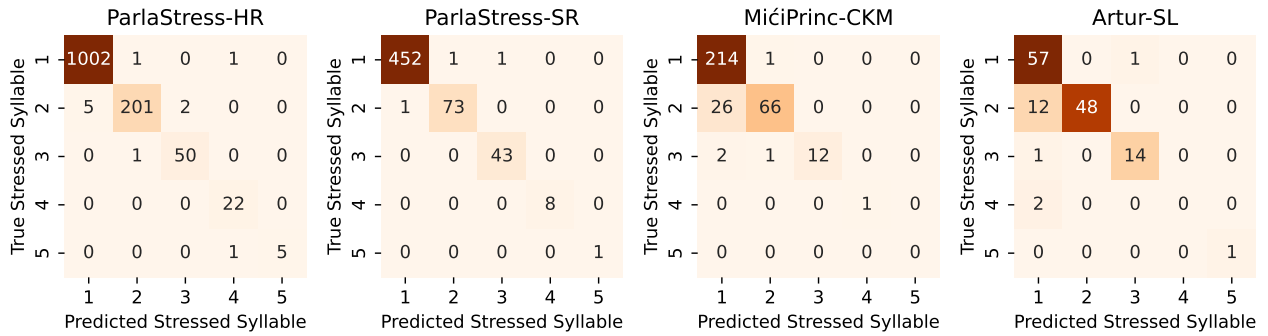


Figure 1: Confusion matrices of stress positions on the four test sets.

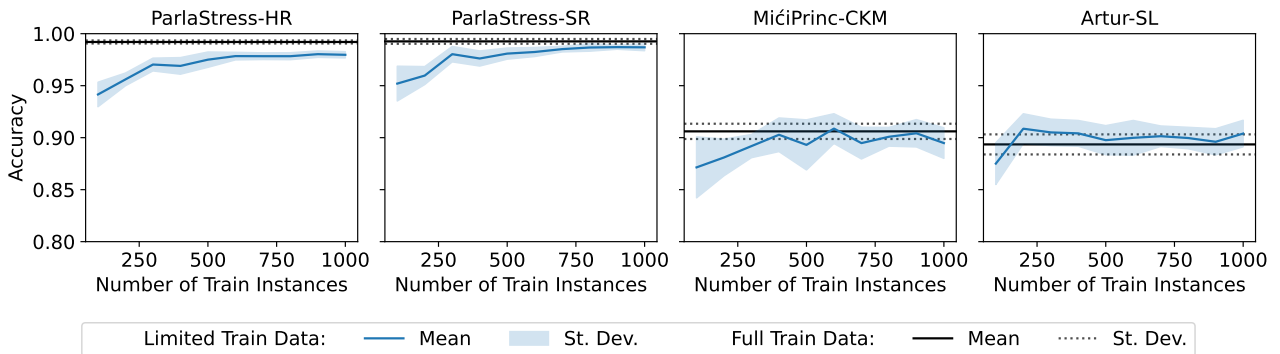


Figure 2: Learning curves as the number of training instances increases, compared to the performance of the final transformer model trained on all available instances.

classifier’s preference for an earlier stress position regardless of the stress being clearly pronounced later in the word.

### 5.3. Training data size

In this final set of experiments we investigate the capacity of the transformer model to perform well on smaller amounts of fine-tuning data. During our previous experiments we have observed very good performance already after a single epoch of fine-tuning, which signals that the more than 10 thousand multi-syllabic words we have at our disposal might not be necessary to obtain our final results. The following insights are especially relevant as they illustrate how one could bring the performance on Chakavian and Slovenian up to the levels of Croatian and Serbian via additional manual data annotation.

We perform experiments on varying the amount of training instances from 100 to 1000, with step of 100. The number of training steps is kept constant at 1200 to control for the amount of updates the transformer has received. On each amount of training data, 10 models are trained on a random selection of the training data. We compare the learning curves with the performance of 10 models trained on all data. The variability of model performance is quantified via standard deviation. The results, given in Figure 2, show that even with a few hundred words available for fine-tuning, performance becomes useful and improves significantly up to a training dataset size of only 500 words. At that point the final performance on the Chakavian and Slovenian test sets is obtained already, while for the Croatian and Serbian test set, more similar to the training data,

there is slow growth continuing even after the 1000 instances depicted here.

## 6. Conclusion

This paper has investigated the performance of pre-trained transformer speech encoders on the task of primary stress identification, comparing them to SVM classifiers trained on traditional acoustic features. The experiments were performed on a newly constructed training and four test datasets in various South-Slavic languages and dialects. Although SVM classifiers show to be more robust to language and dialect change, their performance in each setup is drastically lower to that of transformers. On Croatian and Serbian, transformer models achieve close-to-perfect results, with a 10-percent accuracy drop on more distant Chakavian and Slovenian.

Insights in the true and predicted position of the stress in transformer models show that the main reason for the drop in performance on Chakavian and Slovenian is the very strong preference of the first syllable in the Croatian training data. Experiments on the impact of training data size show that 500 annotated words for fine-tuning already generate peak performance in Chakavian and Slovenian, while for Croatian and Serbian, having multiple thousands of fine-tuning instances does improve the results further.

Future work will include developing techniques for model robustness to language and dialect change, as well as more in-depth analyses such as gender and word memorization effects.

## 7. Acknowledgements

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), and the Research Infrastructure DARIAH-SI (IO-E007), all funded by the ARIS Slovenian Research and Innovation Agency.

We are especially grateful to our two data annotators, Mirna Potočnjak and Nejc Robida.

## 8. References

- [1] J. Tepperman and S. Narayanan, “Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners,” in *Proceedings (ICASSP’05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1. IEEE, 2005, pp. 1–937.
- [2] P. Garde, *Naglasak*. Školska knjiga, Zagreb, 1993.
- [3] I. Škarić, *Fonetika hrvatskoga književnoga jezika*. Zagreb: Nakladni zavod Globus, 2007, pp. 16–157.
- [4] L. Subotić, D. Sredojević, and I. Bjelaković, “Fonetika i fonologija: ortoepska i ortografska norma standardnog srpskog jezika,” *Novi Sad: Filozofski fakultet*, 2012.
- [5] M. Bentum, L. ten Bosch, and T. Lentz, “The processing of stress in end-to-end automatic speech recognition models,” in *Proc. Interspeech 2024*, 2024, pp. 2350–2354.
- [6] J. Mallela, S. H. Aluru, and C. Yarra, “Exploring the use of self-supervised representations for automatic syllable stress detection,” in *2024 National Conference on Communications (NCC)*. IEEE, 2024, pp. 1–6.
- [7] K. Li, S. Zhang, M. Li, W.-K. Lo, and H. Meng, “Prominence model for prosodic features in automatic lexical stress and pitch accent detection,” in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.
- [8] A. Verma, K. Lal, Y. Y. Lo, and J. Basak, “Word independent model for syllable stress evaluation,” in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1. IEEE, 2006, pp. 1–1.
- [9] M. A. Shahin, B. Ahmed, and K. J. Ballard, “Classification of lexical stress patterns using deep neural network architecture,” in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 478–482.
- [10] J. McKechnie, M. Shahin, B. Ahmed, P. McCabe, J. Arciuli, and K. J. Ballard, “An automated lexical stress classification tool for assessing dysprosody in childhood apraxia of speech,” *Brain Sciences*, vol. 11, no. 11, p. 1408, 2021.
- [11] A. S. Vakil and J. Trouvain, “Automatic classification of lexical stress errors for German CAPT,” in *SLaTE*, 2015, pp. 47–52.
- [12] M. A. Shahin, J. Epps, and B. Ahmed, “Automatic classification of lexical stress in English and Arabic languages using deep learning,” in *Interspeech*, 2016, pp. 175–179.
- [13] M. L. Greenberg, *A short reference grammar of Standard Slovene*. SEELRC Reference Grammar Network, 2006.
- [14] S. Narayanan and D. Wang, “Speech rate estimation via temporal correlation and selected sub-band correlation,” in *Proceedings (ICASSP’05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1. IEEE, 2005, pp. 1–413.
- [15] C. Yarra, O. D. Deshmukh, and P. K. Ghosh, “Automatic detection of syllable stress using sonority based prominence features for pronunciation evaluation,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5845–5849.
- [16] M. K. Ramanathi, C. Yarra, and P. K. Ghosh, “ASR inspired syllable stress detection for pronunciation evaluation without using a supervised classifier and syllable level features,” in *INTER-SPEECH*, 2019, pp. 924–928.
- [17] M. Lai, Y. Chen, M. Chu, Y. Zhao, and F. Hu, “A hierarchical approach to automatic stress detection in English sentences,” in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1. IEEE, 2006, pp. 1–1.
- [18] N. Ljubešić, P. Rupnik, and D. Koržinek, “The ParlaSpeech collection of automatically generated speech and text datasets from parliamentary proceedings,” in *International Conference on Speech and Computer*. Springer, 2024, pp. 137–150.
- [19] N. Ljubešić, P. Rupnik, and T. Perinčič, “Mići Princ - A little boy teaching speech technologies the Chakavian dialect,” 2024. [Online]. Available: <https://doi.org/10.5281/zenodo.13936404>
- [20] D. Verdonik, A. Bizjak, A. Žgank, M. S. Maučec, M. Trojar, J. Ž. Gros, M. Bajec, I. L. Bajec, and S. Dobrišek, “Strategies for managing time and costs in speech corpus creation: insights from the Slovenian ARTUR corpus,” *Language Resources and Evaluation*, pp. 1–26, 2024.
- [21] D. Verdonik, K. Dobrovoljc, T. Erjavec, and N. Ljubešić, “Gos 2: A new reference corpus of spoken Slovenian,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 7825–7830.
- [22] J. Križaj, J. Žganec Gros, and S. Dobrišek, “Utilizing forced alignment for phonetic analysis of Slovene speech,” in *Proceedings of the Language Technologies and Digital Humanities Conference 2024*, 2024.
- [23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, “The Kaldi speech recognition toolkit,” in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.
- [24] N. Ljubešić, D. Koržinek, P. Rupnik, and I.-P. Jazbec, “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*, 2022, pp. 111–116.
- [25] P. Boersma and D. Weenink, “Praat: doing phonetics by computer,” <http://www.praat.org/>, 2001, version 6.4, accessed 2024-01-30.
- [26] K. Krippendorff, “Computing Krippendorff’s alpha-reliability,” 2011.
- [27] J.-C. Klie, R. E. d. Castilho, and I. Gurevych, “Analyzing dataset annotation quality management in the wild,” *Computational Linguistics*, vol. 50, no. 3, pp. 817–866, 2024.
- [28] W. Menzel, E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton, and C. Souter, “The ISLE corpus of non-native spoken English,” in *Proceedings of LREC 2000: Language Resources and Evaluation Conference*, vol. 2. European Language Resources Association, 2000, pp. 957–964.
- [29] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>
- [30] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, and *et al.*, “Seamless: Multilingual Expressive and Streaming Speech Translation,” 2023. [Online]. Available: <https://arxiv.org/abs/2312.05187>