The ParlaSpeech-HR benchmark for speaker profiling in Croatian

Nikola Ljubešić,*† Peter Rupnik*

*Department of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
nikola.ljubesic@ijs.si
peter.rupnik@ijs.si

†Faculty of Computer and Information Science
University of Ljubljana
Večna pot 113, SI-1000 Ljubljana

Abstract

Recent advances in speech processing have made speech technologies significantly more accessible to the research community. Beyond the most-popular task of automatic speech recognition, classifying speech acts by various criteria has also recently caught interest. In this paper we propose a benchmark constructed from a dataset of speeches given in the Croatian parliament, aimed at predicting the following speaker profile features: speaker identity, gender, age, and power position (whether the speaker is in the ruling coalition or opposition). We evaluate various pre-trained transformer models on our variables of interest, showing that speaker identification and power position prediction seem to rely mostly on language-specific features, while gender and age prediction rely more on generic speech features, available also in models not pre-trained on the target language. We release the benchmark to serve in measuring the strength of upcoming speech models on a lower-resourced language such as Croatian.

1. Introduction

Speech technologies have recently experienced a quantum leap in their development due to the successful application of the self-supervised pre-training of transformer models on speech data (Schneider et al., 2019). Due to this significant simplification of the development of speech technologies, their uptake has increased significantly (Fan et al., 2020; Pepino et al., 2021; Bartelds et al., 2022), which resulted also in the development of the first open dataset for training automatic speech recognition in Croatian (Ljubešić et al., 2022), based on data from the Croatian parliament. The parliamentary data are especially suited for speech experiments, not only because they are in the public domain, but also because they are rich in speaker metadata (Ljubešić et al., 2022).

In this work we are presenting a rather opportunistic benchmark for speaker profiling in Croatian, based on the ParlaSpeech-HR dataset and the available information on the speakers in that dataset. We define four tasks. In the first task, speaker identification, the task is to predict who of the possible 50 speakers is the speaker of a speech act. For the second task, male and female speakers are to be discriminated between. The third task is focused on discriminating between younger and older speakers, 49 years of age being the division point between the two age groups. In the fourth task we aim at discriminating the speech acts depending on whether they were given by MPs from the ruling coalition, or from the opposition.

We compare models pre-trained on the target language (Croatian) and models that were not pre-trained on this language, obtaining insights not only how well transformer models perform on these tasks, but also how much language-dependent these tasks are. While there have been many approaches to speaker profiling developed before the era of transformers, in this work, we limit ourselves on evaluating transformer models only, primarily due to their reported superior performance (Yang et al., 2021).

2. Related work

Various speech benchmarks, including speaker profiling tasks, exist, but mostly for the English language. The SUPERB benchmark (Yang et al., 2021) consists of tasks of speaker identification – identifying the speaker from a closed set of speakers, speaker verification – binary tasks of whether two utterances are spoken by the same speaker, and speaker diarization – predicting who is speaking when for each timestamp, where multiple speakers can also speak simultaneously.

Another recent benchmark, XTREME-S (Conneau et al., 2022), is focused on evaluating universal cross-lingual speech representations in many languages, the tasks based on speech classification covering spoken language identification among 104 languages, and intent classification from the e-banking domain.

A dataset used for speaker identification benchmarking is VoxCeleb (Nagrani et al., 2017), consisting of over 1,000 celebrities voice samples, obtained by applying facial recognition over YouTube videos.

A well known dataset used for benchmarking automatic speech recognition systems, but also used for speaker profiling is the TIMIT dataset (Garofolo et al., 1993), consisting of 630 speakers of 8 dialects of American English. It consists of speaker information on gender, age and height (Kalluri et al., 2020).

In this work we are not trying to build on top of the existing benchmarks due to two reasons. The main reason is our interest in less-resourced languages, primarily South Slavic languages, for which there is little to no data available. The very recently released ParlaSpeech-HR dataset, on which this benchmark is based on, is the first openly available speech dataset for Croatian (Ljubešić et al., 2022). The second reason is the disruptive effect the speech transformers had on the field, drastically lowering the previous level of error (Yang et al., 2021), with significant improvements expected in the near future as well. This is why we opt for a new, very opportunistic benchmark on speakers from the Croatian parliament. Besides documenting the highly important data selection decisions, we are reporting first results on the current state-of-the-art technology. Given the current high pace of innovation in speech technologies, that is surely not to slow down soon, this benchmark will be highly useful in assessing what new technologies are and will be able to offer to a less-resourced language such as Croatian.

3. Benchmark construction

In this section we present the dataset our benchmark is constructed on, and the data selection protocols given the four variables of interest.

3.1. The dataset

The dataset this benchmark is based on is the ParlaSpeech-HR dataset (Ljubešić et al., 2022), aimed primarily at developing automatic speech recognition systems for Croatian. It consists of 1816 hours of speech obtained from 309 speakers. For each speaker metadata on age, gender, party affiliation, role in the parliament, and power status (opposition vs. coalition) is available. More details on the content and construction procedure of the ParlaSpeech-HR dataset can be found in the description paper (Ljubešić et al., 2022).

3.2. Data selection

For each of the four tasks a separate data selection procedure was set-up, given the limited data available, but also the different nature of the tasks. While most tasks are binary (gender, age, power status), the task of speaker identification is a 50-class task. Furthermore, while for the three binary tasks the training, development and testing subsets have to consist of different speakers, on the speaker identification task, in all three subsets the same speakers have to be present. Finally, in the tasks of age and power status prediction we decided to sample only from male speakers as there are too few female speakers in the dataset for a reasonable sampling that would not include any unwanted bias.

Additionally, in each of the four tasks we only selected instances that were at least 8 seconds in duration. While most of the ParlaSpeech-HR dataset consists of such instances (voice activity detection was set-up in such a fashion), there is a small number of instances, mostly coming from endings of audio files, that are shorter than 8 seconds.

We also discarded speakers producing more than 3,000 instances or less than 200 instances. While the speakers with a small production might complicate the data selection procedure as we want each selected speaker to be equally represented in a sample, the most prolific speakers were left out of the sampling procedures due to their very specific roles in the parliament, which quite likely carries different unwanted biases in their speech production.

In the four following subsections we describe the specific sampling criteria applied for each of our four tasks.

3.2.1. Speaker identification

For the task of speaker identification, 25 speakers per binary gender were sampled. Per speaker, 100 instances were included in the training subset, 10 in the development subset, and 10 in the test subset. Checks were performed to assure that for no speaker instances from the same video appear in more than one subset. With this sampling procedure, each of the three subsets consist of the same 50 speakers, the training subset having 5,000 instances, while the development and testing subsets of 500 instances each.

3.2.2. Gender prediction

For each of the two binary genders, male and female, 25 speakers were selected for the training subset, every speaker being represented with 20 instances. For each of the two genders, 5 speakers (that were not already in the training subset) were taken for the development split, and 5 speakers for the test split. Every speaker in the development and testing subset was represented with 200 instances. With this we assured three subsets of distinct speakers, the training subset consisting of 1,000 instances, and the development and testing subset of 2,000 instances.

3.2.3. Age prediction

Given that there are very few distinct female speakers in the ParlaSpeech-HR dataset, and that controlling for gender while performing any data split is necessary due to the likely strong signal coming from the gender of the speaker as a potential confounder, after some metadata analyses, we decided to setup the age prediction task on male speakers only.

The age distribution of male speakers showed a rather narrow and normal distribution around the median of 49 years of age. The age distribution is far from a uniform and wide distribution, that would allow for a diverse age prediction task, being set-up as a regression task, or a classification task with many categories. This is why we decided to define this as a binary task, predicting whether a speaker is below or above the median age. For the training portion of the task, 60 speakers were selected, with 20 instances per speaker. For the development and test set, 20 speakers were selected for each subset, each speaker being represented by 50 instances. While performing the split, additional checks were put in place to ensure that the age distribution in each of the subsets is as-close-as-possible to the distribution in the full dataset. Additional checkups were also performed to ensure that no speaker leakage existed between the three subsets. With this data selection, the training subset consists of 1,200 instances, while the development and testing subsets consist of 1,000 instances each. Given that the median was chosen as the classification boundary, the final dataset is balanced regarding the two levels of the age variable.

3.2.4. Power status prediction

We decided to wrap up the benchmark with a quite likely less acoustic task, and a more semantic task. Given that we are currently proposing a shared task on predicting whether a transcript of a speech was given by the ruling coalition or opposition, we decided to add that task in this benchmark as well, but performed on speech and not on text transcripts. The ParlaSpeech-HR data come from a single term of the Croatian parliament, which means that the ruling coalition members are mostly from the right political spectrum, while the opposition members are mostly from the left side of the political spectrum. Disentangling the party affiliation or political orientation, and the power status was rather impossible here, which has to be taken into account while analysing the results.

Similar to the task of age prediction, we, again, sampled only among male speakers as the number of female speakers was too low for well-stratified samples. Similar as with age, given the high predictability of gender, we did not want to allow for gender to become a confounder of our primary prediction task, which is power status in this case. We sampled 25 speakers per each power status for train, each speaker being represented by 50 instances. For the development and test sets we selected 9 speakers for each subset, again, representing each speaker with 50 instances. Additional checks were performed that there is no speaker leakage between the three subsets. With this, the size of the training subset is 2,500 instances, while the development and test subsets consist of 900 instances each. For simplicity of evaluation, the division of instances regarding the power status variable is balanced, with 50% instances coming from each side of the political power spectrum.

The benchmark is made available for reproducibility and further benchmarking to the public via the GitHub repo `https://github.com/clarinsi/parlaspeech-hr-benchmark/`.

4. Experimental setup

In this section we give a short description of the setup of the experiments performed on the newly constructed benchmark.

We perform all our experiments with transformer models (Vaswani et al., 2017) that were pre-trained on spoken data. We use the Transformers library (Wolf et al., 2019) and retrieve pre-trained models from the Huggingface model repository.

We use the model pre-trained on Croatian that has proven to perform best on the task of automatic speech recognition (ASR) (Ljubešić et al., 2022), namely the `Slavic` model.[1] We compare the performance of the pre-trained-only model to the model that was additionaly fine-tuned on the ASR task (`Slavic-asr`[2]) to investigate whether fine-tuning the model on the same data, but another task, improves performance.

We compare the performance of the model pre-trained on Croatian to the model that was pre-trained on an unrelated language, in our case English (`English-asr`[3]). We have decided to use the English model fine-tuned for ASR as the non-finetuned model[4] was giving random results after fine-tuning to any of our four tasks. This suspiciously bad result is probably to be followed back to a technical issue in the model, rather than the fact that the model was not fine-tuned on ASR before, as will be seen in the comparison between the performance of the `Slavic` and the `Slavic-asr` model.

The overview of models used in our experiments, together with a short description on the type and amount of data the models were pre-trained and fine-tuned on, is given in Table 1. The non-finetuned Croatian model was pre-trained on around 99 thousand hours of raw recordings of speeches in various Slavic languages that were given in the European parliament. The fine-tuned Croatian model was additionally fine-tuned on the ASR task on around 300 hours of the ParlaSpeech-HR dataset. The English model was pre-trained on 53 thousand hours of raw speech material obtained from audio books and was fine-tuned for the ASR task on 960 hours of similar material.

Regarding hyperparameter optimization, we investigate only the number of epochs required for performance improvements to stall, which is performed by training on the training portion and evaluating on the development portion. For the first two tasks of speaker identification and gender prediction, two epochs were shown to be enough, while for the tasks of age prediction and power status prediction, 15 epochs over the training subset were chosen as optimal.

We evaluate each model on our test subset by reporting both the accuracy and macro F1 metric. Given that all our tasks consist of datasets with a balanced distribution of the response variable, our random baseline lies at 0.5 in case of the binary classification schema, and 0.02 in case of the 50-class speaker identification schema.

For the less challenging tasks of speaker identification and gender prediction, we perform two types of evaluation, on full instances, and on the first 2 seconds of each instance only.

5. Results

5.1. Speaker identification

The results on the speaker identification task are presented in Table 2. The results show for task to be quite easy for the `Slavic` and `Slavic-asr` models applied on full instances. The model fine-tuned on ASR seems to perform slightly better in the full-data scenario, keeping an even score on instances clipped to two seconds, while in that case the non-finetuned model experiences a significant drop of 20 points. This result seems to show how important it is for the model to experience the exact speakers it is supposed to differentiate between, even on another task such as ASR. We do not believe that transfer has occurred between the ASR task and the speaker identification task directly (the model exploiting what people are saying while deciding on the speaker identity) but rather that its parameters

[1] `https://huggingface.co/facebook/wav2vec2-large-slavic-voxpopuli-v2`

[2] `https://huggingface.co/classla/wav2vec2-large-slavic-parlaspeech-hr`

[3] `https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self`

[4] `https://huggingface.co/facebook/wav2vec2-large`

model name	short name	pre-training	ASR fine-tuning
facebook/wav2vec2-large-slavic-voxpopuli-v2	Slavic	Slavic (99k hours)	-
classla/wav2vec2-large-slavic-parlaspeech-hr	Slavic-asr	Slavic (99k hours)	Croatian (300 hours)
facebook/wav2vec2-large-960h-lv60-self	English-asr	English (53k hours)	English (960 hours)

Table 1: List of models used in our experiments, with amount and type of pre-training and fine-tuning data.

model	clipped	accuracy	macro F1
Slavic	no	0.998	0.998
	2 sec	0.806	0.784
Slavic-asr	no	1.000	1.000
	2 sec	1.000	1.000
English-asr	no	0.334	0.275
	2 sec	0.106	0.048

Table 2: Speaker identification results.

were previously adapted to focus better at the peculiarities of the 50 speakers in question.

For the English model, it shows interestingly to perform rather badly, with predictions over the full length of each instance (between 8 and 20 seconds) being correct only in 33% of cases. This is still quite far apart from the random baseline of 2%, but also very far from the stellar performance of the models pre-trained on Croatian. Predicting only on 2 seconds of speech further deteriorates the results to an accuracy of 10%. For the speaker identification task the pre-training language seems to be very important, as the model quite likely models phonetic peculiarities of each speaker, rather than only acoustic features for which any speech transformer should be useful.

To investigate which speakers get confused between by the `Slavic` model, when only two seconds are available for prediction, we present the confusion matrix in Figure1. The matrix shows that speakers of the same gender are being confused between each other, e.g. Arsen Bauk, Davor Bernardić and Božo Petrov being confused for Žarko Katić, or Sunčana Glavak and Ljubica Lukačić being misclassified as Ivana Ninčević-Lesandrić.

5.2. Gender prediction

The results on task of gender prediction are presented in Table 3. On this task all three models, regardless of the language they are pre-trained on, achieve very good performance, the lowest result being accuracy of 98.5%, and the difference in the length of test instances not having a strong impact. Interestingly, the `Slavic-asr` model that performed perfectly on the speaker identification task is the one that performs the worse on the gender prediction task.

Investigating what type of confusion occurs on this task we analyse the output of the `Slavic` model on 2-second instances. We represent the results via a confusion matrix in Figure 2, showing that male instances are sometimes confused for female instances, but not vice versa. Investigating further what speakers are being confused most of the time, it shows that it is a limited number of speakers whose voice has, at least in some occasions, a higher pitch.

The results on gender prediction show that transformer

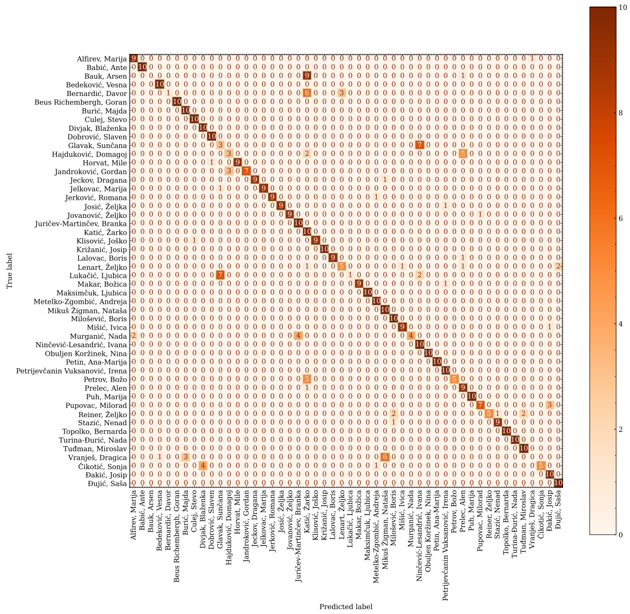

Figure 1: Confusion matrix for speaker identification with the `Slavic` model on instances clipped to two seconds.

eval split	clipped	accuracy	macro F1
Slavic	no	0.997	0.997
	2 sec	0.989	0.989
Slavic-asr	no	0.985	0.985
	2 sec	0.985	0.985
English-asr	no	0.999	0.999
	2 sec	0.994	0.994

Table 3: Gender prediction results.

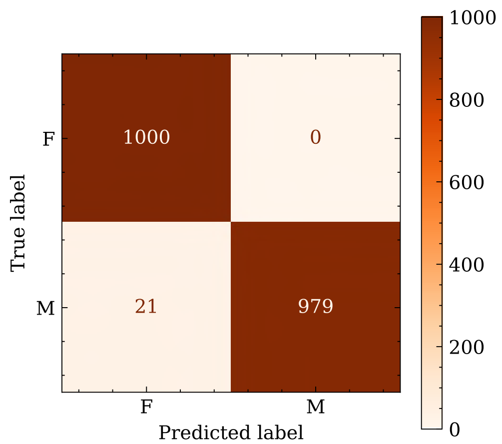

Figure 2: Confusion matrix for speaker gender prediction of the `Slavic` model on 2-second test instances.

model	clipped	accuracy	macro F1
Slavic	no	0.694	0.690
Slavic-asr	no	0.722	0.722
English-asr	no	0.678	0.672

Table 4: Age prediction results.

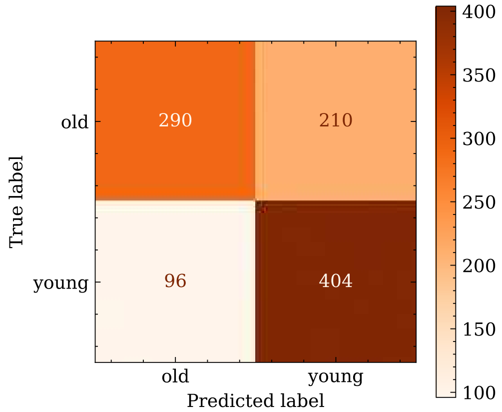

Figure 3: Confusion matrix for speaker age classification by the `Slavic` model.

models do not rely on language-specific features, but quite likely on the pitch of a speaker's voice, with best results being reported by the English model, with almost perfect results even on 2-second test instances.

5.3. Age prediction

The results on age prediction, guessing whether a speaker is younger or older than 49 years, which is the median speaker age in the dataset, are given in Table 4. Here we do not perform experiments on speech samples clipped to two seconds as the task is already demanding enough on full-length instances. The `Slavic-asr` model seems to perform best, with accuracy of 72%, 50% being a random result. The `Slavic` and `English-asr` model seem to be suspiciously close in performance, only with a point and a half difference, which shows that the age prediction task does not rely on language-specific features, but rather general acoustic features.

To investigate the confusion patterns between the two age groups, we plot a confusion matrix of the `Slavic` model in Figure3. The confusion matrix shows clearly that more frequently older speakers tend to be misclassified as younger speakers than vice versa.

Given that we have divided the speakers by age on the median point, and that the speaker age is rather normally distributed, we wanted to additionally check whether most of the prediction errors occur on users who are close to the class boundary. To investigate this, we plot an instance-level age histogram in Figure 4, encoding the correctly and incorrectly classified instances by the `Slavic` model with different colour. The histogram shows that most misclassifications happen, as expected, close to the median class boundary, with almost all instances of speakers of 50 and 51 years of age being misclassified as younger speakers. Classifications on the youngest (35 years) and oldest speakers (68 and 69 years) show to be perfectly performed by the model.

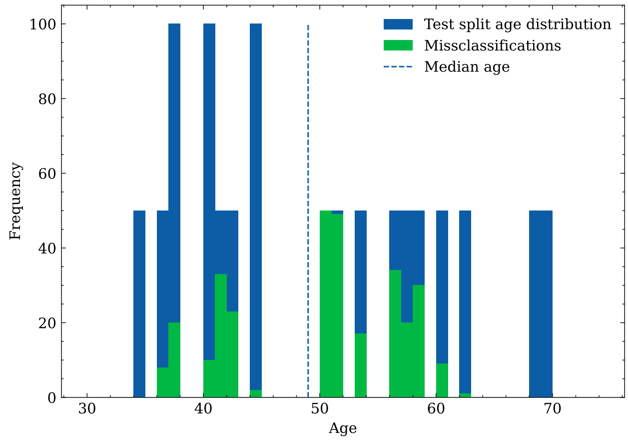

Figure 4: Distribution of age in our test subset, along with misclassifications by the `Slavic` model.

model	clipped	accuracy	macro F1
Slavic	no	0.590	0.587
Slavic-asr	no	0.627	0.626
English-asr	no	0.549	0.531

Table 5: Power status identification.

This insight might motivate us to organise the age prediction task in the future as a classification task into three categories, the middle category, around the median age, being considered hard, and discarded in the easier setup of the classification task.

5.4. Power status prediction

The results of our final task, power status prediction, are given in Table 5. The results show to be, as expected, the lowest of all four tasks defined in this benchmark. The `Slavic-asr` model performs best, with the difference to the non-finetuned model being 2.7 accuracy points. The model that was not pre-trained on Croatian achieves a significantly lower result, 5 points lower than any model pre-trained on Croatian, showing that for solving this task mostly language-specific features are used.

Which features exactly are actually used is hard to identify. The only attempt we perform in this direction is a per-speaker analysis of correct and incorrect classifications by the `Slavic-asr` model, which we present in Figure 5. The results show that people in power seem to be easier to identify than those who are in opposition, as the speakers having the lowest percentage of correctly classified instances are mostly from the opposition. The error also seems to be rather speaker-dependent, with eight of the speakers having accuracy above 80%, and the five worst-performing speakers having accuracy below 40%.

Analysing the five worst-performing speakers, a trend can be observed, with the two speakers in power being two of the most fine-mannered speakers, while two out of three speakers from the opposition are rather known for their harsh speech. This analysis has also shown that the signal the classifier has caught on is quite likely based on the polit-

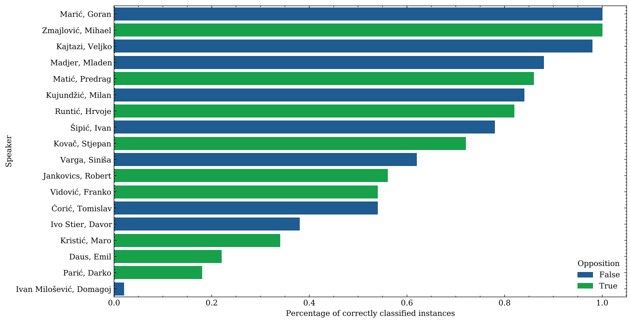

Figure 5: Per-speaker accuracy level with the `Slavic-asr` model on the power status prediction task.

ical orientation rather than power status itself. For performing modelling of the power status in speech, the training and evaluation data should consist of multiple-terms data, with the same political orientations having speeches given while in power and while in opposition.

6. Conclusion

In this paper we have presented a benchmark for speaker profiling in Croatian, based on the recordings of the Croatian parliament. We have carefully selected the speakers and instances to be used in the benchmark, paying special attention to any type of bias or confounders that might be included in the tasks.

We have performed initial experiments with transformer models pre-trained on speech, obtaining interesting insights. The task of speaker identification seems to be rather language-dependent, and can be further improved if the model has seen speakers to be identified before the final fine-tuning process. Gender prediction seems to be the least language specific, obtaining very good results regardless of the model, and quite likely relying simply on the pitch of the speaker. Age prediction, in our case set up as a binary task, with the boundary being the age median, shows to be hard, but very feasible on instances that are further away from the classification boundary. The task shows to use language-specific features to a small amount, but the model that has experienced the same speakers before the final fine-tuning still performing visibly better than the model that has not. Power status prediction is the hardest of all four tasks, and shows to rely on language-specific features, again profiting additionally from experiencing the speakers prior to the final fine-tuning. Analysing the accuracy by speaker shows that the power status model seems to have caught on the political orientation rather than the language of power itself. For working on modelling that phenomenon, a dataset controlling for political orientation should be constructed, which requires a much wider data range than is currently available.

We are releasing the benchmark definitions, to be coupled with the full ParlaSpeech-HR dataset (Ljubešić et al., 2022) in a GitHub repository.[5]

[5] `https://github.com/clarinsi/parlaspeech-hr-benchmark/`

Acknowledgements

This work has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278341 (MaCoCu project). This communication reflects only the author's view. The Agency is not responsible for any use that may be made of the information it contains.

This work was also funded by the Slovenian Research Agency within the Slovenian-Flemish bilateral basic research project "Linguistic landscape of hate speech on social media" (N06-0099 and FWO-G070619N, 2019–2023) and the research programme "Language resources and technologies for Slovene" (P6-0411).

7. References

Martijn Bartelds, Wietse de Vries, Faraz Sanal, Caitlin Richter, Mark Liberman, and Martijn Wieling. 2022. Neural representations for modeling variation in speech. *Journal of Phonetics*, 92:101137.

Alexis Conneau, Ankur Bapna, Yu Zhang, Min Ma, Patrick von Platen, Anton Lozhkov, Colin Cherry, Ye Jia, Clara Rivera, Mihir Kale, et al. 2022. Xtreme-s: Evaluating cross-lingual speech representations. *arXiv preprint arXiv:2203.10752*.

Zhiyun Fan, Meng Li, Shiyu Zhou, and Bo Xu. 2020. Exploring wav2vec 2.0 on speaker verification and language identification. *arXiv preprint arXiv:2012.06185*.

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, and David S. Pallett. 1993. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93:27403.

Shareef Babu Kalluri, Deepu Vijayasenan, and Sriram Ganapathy. 2020. Automatic speaker profiling from short duration speech data. *Speech Communication*, 121:16–28.

Nikola Ljubešić, Danijel Koržinek, Peter Rupnik, Ivo-Pavao Jazbec, Vuk Batanović, Lenka Bajčetić, and Bojan Evkoski. 2022. ASR training dataset for Croatian ParlaSpeech-HR v1.0. Slovenian language resource repository CLARIN.SI, `http://hdl.handle.net/11356/1494`.

Nikola Ljubešić, Danijel Korzinek, 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 Third ParlaCLARIN Workshop*, Marseille, France.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng- I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2021. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.