THE GOVORI.SI SPEECH TRANSCRIPTION PLATFORM

Klara ŽNIDERŠIČ,[1] Vid KLOPČIČ,[1] Matevž PESEK,[1] Matija MAROLT[1]
[1]University of Ljubljana, Faculty of Computer and Information Science

Despite considerable progress in automatic speech recognition, there are very few open speech transcription solutions with user-friendly graphical interfaces, especially for low-resource languages. To fill this gap, we have developed and implemented *govori.si*, a speech transcription platform specifically tailored to the Slovenian language. We trained an automatic speech recognition model that is among the best current Slovenian automatic speech recognition models and used state-of-the-art methods to tackle other transcription challenges: diarization, capitalization, punctuation, custom substitution dictionaries and numerical notation. The platform is available free of charge for research and non-commercial purposes and has been well received by users, making it a valuable tool for various applications, including legislative processes, journalism, and research.

Keywords: Slovenian language, speech recognition, natural language processing

1 INTRODUCTION

The field of automatic speech recognition has seen remarkable advancements over the years and automatic speech transcription services have become indispensable tools in numerous domains, streamlining time-consuming tasks. However, despite the ubiquity of these services, there are still challenges in extending their capabilities to non-English languages. Recent studies have mainly focused on English language processing in neutral environments (Al-Fraihat et al., 2024), but language-specific special cases, particularly for languages with limited training data and complex linguistic variations, need to be additionally considered to broaden the scope of speech recognition.

In this paper, we present a state-of-the-art speech transcription platform *govori.si*, which is specifically tailored to the Slovenian language and enables seamless speech-to-text conversion with high accuracy and efficiency. The main advantages of the proposed platform lie in the full implementation of the speech transcription pipeline and the adaptability to the different needs of end users in a wide range of applications. Whether it is the transcription of lectures, speeches, interviews, meetings or other lengthy voice recordings, *govori.si* can increase productivity by speeding up the transcription process and saving valuable time and resources.

The article is structured as follows: Section 2 presents the relevant ASR tools and an overview of related interfaces for speech transcription. Section 3 describes the methods used in each phase of the transcription process, Section 4 discusses the technologies used for the development of our platform, and Section 5 describes the user interface of *govori.si*. The results of an informal user study are presented in Section 6, and the article concludes with Section 7.

2 LITERATURE REVIEW

Automatic speech recognition (ASR) is the process of transcribing speech into text. Since the diversity of speech signals poses challenges to automatic methods, researchers have pursued different strategies to overcome them. In recent years, deep learning models trained on large corpora are the predominant choice, especially the models based on transformer blocks such as OpenAI's Whisper (Radford et al., 2023) or on convolution-augmented transformer blocks such as Google's USM (Zhang et al., 2023).

Although the Slovenian language is included in many multilingual ASR models (Li et. al., 2022; Pratap et al., 2020; Radford et al., 2023; Zhang et al., 2023), it may be underrepresented in the training data, which has a negative impact on the accuracy of these models. Therefore, efforts have been made in recent years to develop ASR datasets and models specifically for the Slovenian language. The ARTUR dataset (Verdonik et al., 2023) contains 884 hours of transcribed speech and was specifically designed to train ASR models for the Slovenian language. SloBENCH (Žitnik & Dragar, 2021) serves as a central evaluation plat-form for benchmarking the progress of Slovenian natural language processing technologies, including machine translation between Slovenian and English, named entity recognition, universal dependency parsing, and speech recogni-tion. Several ASR models have also been developed in recent years. The ASR system of (Gril & Dobrišek, 2022) uses a hybrid acoustic modeling approach that combines hidden Markov models (HMMs) with deep neural networks (DNNs).

Two models based on convolution-augmented transformer blocks (conformers) were also introduced (Lebar Bajec et al., 2022; *True-bar 23.02 ASR model*, 2023).

Integrated solutions for transcribing and editing speech are often found in commercial software packages. Unfortunately, the high associated costs are often an obstacle for researchers or projects with limited budgets. Manual transcription, while accurate, is time-consuming and labor-intensive. Therefore, open source solutions are the only viable option for many individuals or organizations looking to streamline the analysis of spoken text, interviews, and other audio content.

Vink (Tolle et al., 2024) is an open-source automatic transcription tool that simplifies the use of OpenAI's Whisper ASR model for non-programmers engaged in qualitative research. The tool is available as a standalone application for the Windows operating system. Users reported that transcription accuracy varied widely across the 14 languages tested, with the highest error rates for the languages with the smallest training datasets. A similar result was observed when evaluating the tool aTrain (Haberl et al., 2024), which also uses the Whisper transcription model and is implemented as an offline application for Windows. As these tools are not dependent on an internet connection and do not require the data to be uploaded to online servers, they meet the criteria of data protection, ethical guidelines, and legal compliance. While both Vink and aTrain offer a graphical user interface, they lack additional tools for text editing. Their functionality is limited to basic functions such as copying to the clipboard or exporting transcripts in various formats used for synchronizing audio and transcripts. In addition, the parameter settings for transcription are severely limited in both applications.

Another open source transcription tool, SpeechToText (Negrão & Domingues, 2021), was originally developed as a software module for the forensic software Autopsy, with the primary goal of integrating speech content into the standard workflow of digital forensic investigations. Powered by Mozilla's DeepSpeech for speech transcription, SpeechToText was tested with English recordings of non-native speakers obtained from various Android applications that can generate audio files with speech. These tests aimed to replicate scenarios that occur during forensic investigations. However, SpeechToText can also be used independently and is available under an open source license. Compared to Vink and aTrain, its advantage lies in the graphical user interface, as it also includes a transcription viewer and a keyword browser. The disadvantage, however, is the limited language support of DeepSpeech, which is no longer being actively developed.

An application closely related to *govori.si* was developed in Estonia by Olev and Alumäe (2022). The Estonian speech recognition and transcription editing system uses the Wav2Vec2 model, and its evaluation shows that even smaller languages can benefit from such a pre-trained model. It is provided as a publicly available service and offers a graphical user interface for transcription editing, interactive listening to recordings, and speaker annotation features.

3 THE GOVORI.SI SPEECH PROCESSING PIPELINE

The *govori.si* platform processes the input speech in several phases, starting with diarization, automatic speech recognition, capitalization, punctuation, replacement of regular expressions and ending with the processing of numbers. We briefly present the phases in the following subsections.

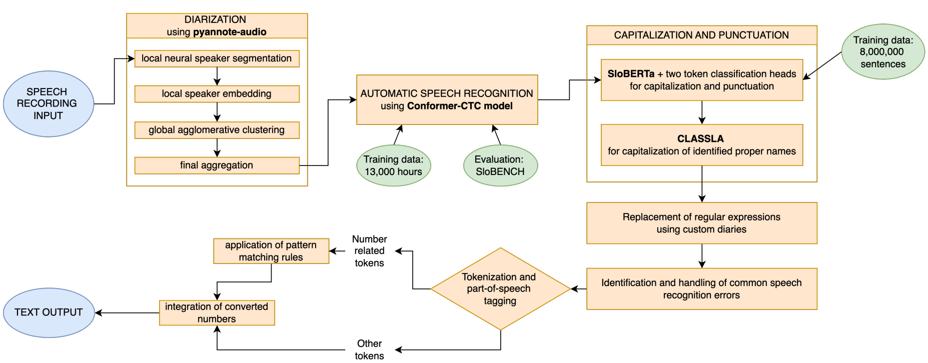

Figure 1: Graphic block scheme.

3.1 Diarization

For speech diarization, we use pyannote-audio (Bredin, 2023; Plaquet & Bredin, 2023)[1], an open-source Python toolkit known for its good performance in

[1] https://github.com/pyannote/pyannote-audio

speaker diarization. The diarization model consists of a speaker segmentation model applied to short sliding windows, neural speaker embedding of each (local) speaker, and (global) agglomerative clustering. We chose pyannote because it is seamlessly implemented in the Python language and has a modular architecture that facilitates the integration of new modules and provides the flexibility to customise the parameters of existing components to our specific needs.

3.2 Speech recognition

We have trained a new speech recognition model for the Slovenian language. The model is based on convolution-augmented transformer (conformer) blocks, which combine convolutional neural networks and transformer models to utilise the best of both approaches. While transformer models are well suited to capture content-based global interactions, CNNs effectively utilise local features (Gulati et al., 2020). The Conformer-CTC model is specifically designed for ASR tasks and uses connectionist temporal classification (CTC) loss and decoding, making it a non-autoregressive model. The NeMo framework from Nvidia was used to train and evaluate the model.

We trained the model with a dataset containing 13,000 hours of speech and the corresponding transcriptions. The dataset includes existing public collections: Gos (Zwitter Vitez et al., 2021), Gos VideoLectures (VideoLectures.NET, 2019), CommonVoice, SiTEDx (Žgank et al., 2016), Sofes (Dobrišek et al., 2017) and ARTUR (Verdonik et al., 2023). We augmented the datasets with publicly available online sources, including dialectal speech from *narecja.si*, radio and television broadcasts, and parliamentary debates, as well as a smaller proportion of datasets from other Slavic languages (Croatian, Serbian, Czech, and Russian).

We evaluated the model (labelled CON-ASR-1.1) as part of the SloBENCH evaluation. The SloBENCH evaluation data for ASR consists of 15 recordings with a total duration of almost 3.5 hours, including public and private speech from southwestern and northeastern Slovenia, represented by male and female speakers. The character error rate of our model was 0.019 and the word error rate was 0.050, which is comparable to the best evaluated model (True-bar

23.02) and significantly better than the OpenAI Whisper multilingual model. For more details on the model's performance see CJVT SloBench Leaderboard.[2]

3.3 Capitalization and punctuation

To capitalize and punctuate the output of the speech recognition model, we use a two-step process. In the first step, we trained two token-level classifiers on top of the monolingual Slovenian SloBERTa model (Ulčar & Robnik-Šikonja, 2020). SloBERTa is a pre-trained BERT-like model trained on corpora with over 3 billion tokens and a subword vocabulary of 32,000 tokens. We extended its embeddings with two different token classification heads—one for punctuation and one for capitalization—and jointly trained them on a set of 8,000,000 sentences from corpus Gigafida (Krek et al., 2020).

In the second step, each sentence is processed with the CLASSLA pipeline. CLASSLA (Ljubešić & Dobrovoljc, 2019; Terčon & Ljubešić, 2023)[3] is the official fork of Stanza (the official Python NLP library of the Stanford NLP Group) for processing Slovenian, Croatian, Serbian, Macedonian, and Bulgarian. Although CLASSLA supports extensive linguistic processing (including tokenization, sentence splitting, lemmatization, part-of-speech tagging, dependency parsing, and named-entity recognition) for both standard and non-standard Slovenian, our main focus was on using CLASSLA to capitalize identified proper names. For this purpose, we used the morphosyntactic XPOS tags, which are more detailed and specific than standard POS tags and contain information that determines whether a word should be capitalized, e.g. whether the word is a noun and a proper name.

This two-step approach ensures that the final text maintains consistent and accurate punctuation and capitalization, which improves readability and the overall quality of the text.

3.4 Regular Expressions Replacement

Search and replace with regular expressions enables automatic correction of ASR output using custom dictionaries, allowing for customization of text output

[2] https://slobench.cjvt.si/leaderboard/view/10
[3] https://github.com/clarinsi/classla?tab=readme-ov-file

and increasing the flexibility and effectiveness of the platform. For example, colloquial terms can be replaced with their formal counterparts (e.g. *pršu* can be replaced with *prišel* etc.) In addition, at this stage we integrate the identification and handling of common speech recognition errors.

3.5 Numeric notation

The final step in the recognition pipeline is the conversion of text representations of numbers into their corresponding digit forms. While there are some online tools for conversion in widely used languages such as English,[4] there is a lack of comprehensive solutions for less common languages such as Slovenian. In our solution, we have implemented a rule-based approach for converting Slovenian text numbers into their digit counterparts. The conversion tool uses a set of regular expressions and linguistic patterns to recognise and convert various number representations, including cardinals, ordinals, fractions, and decimals. It also takes into account common abbreviations, units of measurement, and contextual clues specific to the Slovenian language. First, the input text is subjected to tokenization and part-of-speech tagging to identify relevant number-related tokens. A series of pattern matching rules is then applied to extract and convert the numbers in the text. These rules cover a wide range of number formats and take into account the grammatical subtleties of the Slovenian language. Some examples are:

- Special handling of ”pol” (half) to correctly convert phrases like ”pol milijona” (half a million).
- Recognition of unconventional formats, such as ”dvajset dvajset” (2020) or ”devetnajststo dvajset” (1920).
- Correct conversion of decimal separators and fractions like ”dve celi pet” (2,5) or ”tri četrtine” (3/4).
- Contextual disambiguation that distinguishes between the use of ”sto” (hundred) as an independent number and as part of a larger number such as ”dvesto” (two hundred) or ”petsto” (five hundred).
- Conversion of number ranges, such as ”od pet do deset” (from five to ten).

[4] https://pypi.org/project/word2number/
https://www.atatus.com/tools/word-to-number
https://codebeautify.org/word-to-number-converter

- Handling abbreviations like ”mio” for million and ”mrd” for billion.

4 IMPLEMENTATION

The *govori.si* platform is implemented as a web application. The technology stack consists of several components:

1. Backend: We use the Django web framework together with Django Channels and Redis for handling real-time web functionality, and PostgreSQL for database management.
2. Frontend: Our frontend is developed with React and MobX libraries, providing a dynamic and responsive user interface.
3. Pipeline worker: Our worker module is developed with Python and includes the PyTorch machine learning library and models from Section 3 to support the functionalities of our application.
4. Containerization: We also use Docker for application containerization and Nginx for web server functionalities, resulting in its seamless deployment and efficient web traffic management.

To ensure the accuracy and coverage of the solution, a comprehensive set of test cases was developed. These test cases include a variety of real-world examples taken from various Slovenian texts. The solution has been extensively validated against these test cases, proving its robustness and reliability.

5 USER INTERFACE

The user interface of *govori.si* allows the user to upload and transcribe speech recordings, customising various parameters to their preferences, as shown in Figure 2. The user can specify punctuation and capitalization preferences, and toggle the options for transcribing numbers, segmenting speakers, and using dictionaries. The solution also offers three preset settings: automatic transcription, dictation with punctuation, and raw transcription.

After uploading, the recordings are processed according to the selected settings and the user can edit the transcribed text manually, as shown in Figure 3. The

Figure 2: User interface for uploading recordings and selecting parameters.

top frame provides an overview of the audio recording and allows the user to navigate through the recording and select the playback speed. The transcribed text is displayed segmented by speaker, with each speaker identified by a colored marker that can be renamed if required. The user has the option to replace all instances of a particular word and adjust the font size. Once editing of an individual speaker turn is complete, the user can lock it by clicking the check mark icon on the side. After editing, the text can be exported in various formats, including as an interview in .docx format, as plain text in .txt format, or as a subtitle-compatible .vtt file.

Interaction with the tool is simplified by a series of keyboard shortcuts for text editing and navigation through the audio recording. In addition to individual projects, users can also collaborate on joint projects.

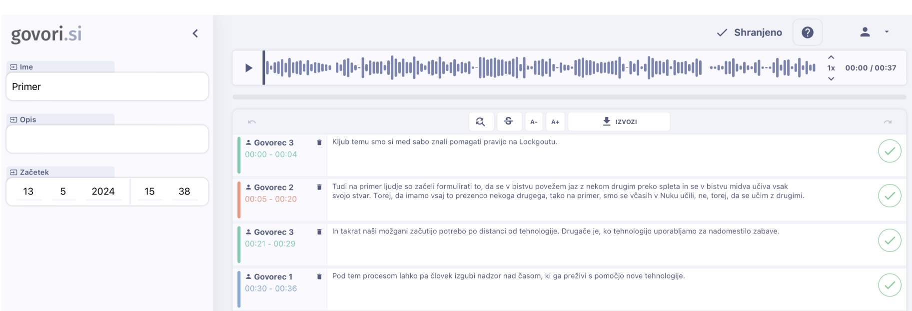

Figure 3: User interface for transcription editing.

6 EVALUATION

Our goal in developing the *govori.si* platform was to provide an easy-to-use speech recognition tool for a variety of applications, including transcription of field recordings, interviews, manual dictation in professional environments, subtitle generation, etc. We have made the tool available for research purposes and it is currently used by over 60 users from different backgrounds who have transcribed over 100 hours of recordings in the past year. It has been systematically used as part of the History of Journalism course at the Faculty of Social Sciences, where students used it to transcribe their interviews with media audiences in socialist Yugoslavia. We gathered informal feedback from 25 participants.

More than half of them emphasized the time-saving benefits of the transcription tool. Some also pointed out the ability to translate informal language into formal language, although the handling of certain dialects still needs improvement. In addition, users appreciated the availability of keyboard shortcuts and the helpful documentation, as well as the very effective processing of background noise. The export function was also praised for its user-friendliness. Overall, the tool was rated as faster, simpler, and more efficient than manual transcription methods.

However, challenges were also identified, such as difficulties with speaker segmentation, which were mentioned by more than three quarters of users. Users reported cases where the tool recognized three or more speakers when only two were involved in the conversation, as well as incorrectly tagged speakers. Some respondents also reported problems with excessive capitalization and punctuation. There were also some problems with copying and pasting text.

Suggestions for improvement included adding features such as text underlining and bolding in the editing interface, providing information on the duration of pauses in speech, and improving search functions within audio recordings. One suggested solution to the problem of speaker segmentation was the ability to specify the number of speakers in advance.

7 CONCLUSION

In this paper, we present an overview of our solution for Slovenian speech-to-text transcription. We have used state-of-the-art methods to overcome various transcription challenges. We trained our own deep speech recognition model, which is among the best current Slovenian ASR models, and integrated diarization, capitalization, punctuation, custom substitution dictionaries, and numerical notation parsing into a common platform that is available for free for non-commercial use. However, access is granted via registration credentials issued by the authors upon request. Users rated the platform very positively, although its accuracy can of course always be improved. We consider it a useful tool for a variety of use cases, including legislative processes, journalism, and research.

In the future, we aim to further improve the performance and usability of the tool. This includes improving the speaker segmentation model, integrating large language models for text summarization and correction, and continuously developing the user-friendly interface for different use cases. With all these improvements, we hope to make the *govori.si* platform a useful tool for Slovenian language processing and also contribute to advances in transcription accuracy, efficiency, and usability.

8 ACKNOWLEDGMENTS

This study was supported by the Slovenian Research Agency (research core funding No. P6-0436, Digital humanities: resources, tools and methods).

REFERENCES

Al-Fraihat, D., Sharrab, Y., Alzyoud, F., Qahmash, A., Tarawneh, M., & Maaita, A. (2024). Speech Recognition Utilizing Deep Learning: A Systematic Review of the Latest Developments. *Human-centric Computing and Information Sciences*, *14*, 41–76. https://doi.org/10.22967/HCIS.2024.14.015

Bredin, H. (2023). pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Interspeech 2023* (pp. 1983–1987). https://doi.org/10.21437/Interspeech.2023-105

Dobrišek, S., Žganec Gros, J., Žibert, J., Mihelič, F., & Pavešić, N. (2017). *Speech Database of Spoken Flight Information Enquiries SOFES 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1125

Gril, L., & Dobrišek, S. (2022). Akustično modeliranje z različnimi osnovnimi enotami za avtomatsko razpoznavanje slovenskega govora. In *Proceedings of Conference on Language Technologies and Digital Humanities, Ljubljana, 2022.*

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented Transformer for Speech Recognition. In *Interspeech 2020* (pp. 5036–5040). https://doi.org/10.21437/Interspeech.2020-3015

Haberl, A., Fleiß, J., Kowald, D., & Thalmann, S. (2024). Take the aTrain. Introducing an interface for the Accessible Transcription of Interviews. *Journal of Behavioral and Experimental Finance, 41*. https://doi.org/10.1016/j.jbef.2024.100891

Krek, S., Holdt, Š. A., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I., & Dobrovoljc, K. (2020). Gigafida 2.0: the reference corpus of written standard slovene. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 3340–3345).

Lebar Bajec, I., Bajec, M., Bajec, Ž., & Rizvič, M. (2022). *Slovene Conformer CTC BPE E2E Automated Speech Recognition model RSDO-DS2-ASR-E2E 2.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1737

Li, B., Pang, R., Zhang, Y., Sainath, T. N., Strohman, T., Haghani, P., Zhu, Y., Farris, B., Gaur, N., & Prasad, M. (2022). Massively Multilingual ASR: A Lifelong Learning Solution. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6397–6401). https://doi.org/10.1109/ICASSP43922.2022.9746594

Ljubešić, N., & Dobrovoljc, K. (2019). What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing* (pp. 29–34). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3704

Negrão, M., & Domingues, P. (2021). SpeechToText: An Open-source Software for Automatic Detection and Transcription of Voice Recordings in Digital Forensics. *Forensic Science International: Digital Investigation, 38*. https://doi.org/10.1016/j.fsidi.2021.301223

Olev, A., & Alumäe, T. (2022). Estonian Speech Recognition and Transcription Editing Service. *Baltic Journal of Modern Computing, 10*(3). https://doi.org/10.22364/bjmc.2022.10.3.14

Plaquet, A., & Bredin, H. (2023). Powerset multi-class cross entropy loss for neural speaker diarization. In *Interspeech 2023* (pp. 3222–3226). https://doi.org/10.21437/Interspeech.2023-205

Pratap, V., Sriram, A., Tomasello, P., Hannun, A., Liptchinsky, V., Synnaeve, G., & Collobert, R. (2020). Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters. In *Interspeech 2020* (pp. 4751–4755). https://doi.org/10.21437/Interspeech .2020-2831

Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 28492–28518). https://proceedings.mlr.press/v202/radford23a.html

Terčon, L., & Ljubešić, N. (2023). *CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages.* https://arxiv.org/abs/2308.04255

Tolle, H., Castro, M. d. M., Wachinger, J., Putri, A. Z., Kempf, D., Denkinger, C. M., & McMahon, S. A. (2024). From voice to ink (Vink): development and assessment of an automated, free-of-charge transcription tool. *BMC Research Notes*, *17*(1), 95. https://doi.org/10.1186/s13104-024-06749-0

True-bar 23.02 asr model. (2023). https://vitasis.si/products/truebar (Retrieved: 2024-05-10)

Ulčar, M., & Robnik-Šikonja, M. (2020). *Slovenian RoBERTa contextual embeddings model: SloBERTa 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1387

Verdonik, D., Bizjak, A., Sepesy Maučec, M., Gril, L., Dobrišek, S., KrižajЈ., ... Dretnik, N. (2023). *ASR database ARTUR 1.0 (transcriptions).* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1772

VideoLectures.NET. (2019). *Spoken corpus Gos VideoLectures 4.0 (audio).* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1222

Žgank, A., Maučec, M. S., & Verdonik, D. (2016). The SI TEDx-UM speech database: a new Slovenian Spoken Language Resource. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 4670–4673). https://aclanthology.org/L16-1740

Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., ... Wu, Y. (2023). Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages. *ArXiv*. https:// arxiv.org/abs/2303.01037

Žitnik, S., & Dragar, F. (2021). *SloBENCH evaluation framework.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1469

Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., & Erjavec, T. (2021). *Spoken corpus Gos 1.1.* Slovenian language resource repository CLARIN.SI. http://hdl .handle.net/11356/1438

PLATFORMA ZA TRANSKRIPCIJO GOVORA GOVORI.SI

Kljub velikemu napredku pri razvoju samodejne razpoznave govora je odprtokodnih integriranih rešitev z uporabniku prijaznimi grafičnimi vmesniki še vedno zelo malo. To vrzel naslavljamo z razvojem slovenskemu jeziku prilagojene platforme *govori.si* za transkripcijo govora. Razvili smo nov model za samodejno razpoznavo slovenskega govora, ki se trenutno uvršča med najboljše tovrstne modele in uporabili druge sodobne pristope k reševanju izzivov pri transkripciji. Platforma *govori.si* združuje prepoznavo govora s segmentacijo govorcev, samodejnim določanjem velikih začetnic in ločil, uporabniško definiranimi slovarji in logiko za zapisovanje števil. Za raziskovalne in nekomercialne namene je platforma prosto dostopna. Med uporabniki je bila pozitivno sprejeta in postaja dobrodošlo orodje za uporabo v zakonodajnih postopkih, novinarstvu in raziskavah.

Keywords: slovenski jezik, samodejna razpoznava govora, obdelava naravnega jezika