INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2026, VOL. 72, NO. 1, PP. 1–7
Manuscript received July 4, 2025; revised February, 2026. doi: 10.24425/ijet.2026.157866

Analysis and categorization of the Rusyn language using the whisper model: demographic influences on linguistic convergence

Paweł Małecki

***Abstract*—The article presents a detailed linguistic analysis of the Rusyn language, focusing on its complex and evolving features, such as pronunciation, as well as individual, regional, and historical variabilities. The investigation employed an artificial neural network based on the OpenAI Whisper model to perform analysis and categorization. Although the Whisper model was trained on data from the majority of state official languages, it was not specifically trained with samples of the Rusyn language due to its niche and minority/ethnic status. Consequently, speech samples in Rusyn were classified according to the most closely related available labels, allowing for the assessment of linguistic similarity between Rusyn and other (mostly) Slavic languages. The study incorporated a diverse user base segmented by gender, age, and geographic location (Poland, Ukraine, Slovakia, Serbia), revealing significant resemblances to the dominant languages within these countries and demonstrating correlations between the computed linguistic similarity and the speakers' age.**

***Keywords*—Artificial Neural Network ANN; Automatic Speech Recognition ASR; Rusyn; Minority Languages; Automatic Classification**

I. Introduction

THE Rusyn language is an Eastern Slavic language employed by the Rusyn ethnic group, primarily within the Carpathian region of Central Europe—encompassing parts of Ukraine, Slovakia, Poland, Hungary, and Romania. It is characterized by the presence of multiple dialects and variants that mirror the diverse historical and cultural influences inherent to these regions. Classifying Rusyn within the broader framework of Eastern Slavic and other Slavic languages remains a complex and contentious issue, largely due to its unique geographical positioning and the resultant historical influences. Rusyn exhibits properties in common with Eastern Slavic languages (such as Russian, Ukrainian, and Belarusian) while also incorporating elements characteristic of West Slavic (e.g., Polish and Slovak) as well as South Slavic (e.g., Serbian and Croatian) languages. This multifaceted linguistic synthesis reflects its location at the crossroads of these language groups [1].

The Rusyn language has historically been influenced by various dominant powers, including the Austro-Hungarian Empire and the Soviet Union. The linguistic question is intrinsically linked to national identity, and while debates persist regarding whether Rusyn constitutes a distinct language or a dialect of Ukrainian, proponents of the latter position are predominantly individuals who identify as Ukrainian. In Ukraine, Rusyns are not officially recognized as a distinct ethnic group, which impedes language preservation efforts [2].

The codification of Rusyn occurred in the late 20th century, reflecting a renaissance of national and linguistic identity. These codification efforts resulted in the establishment of several regional literary standards, based on local dialects in Slovakia, Poland, Ukraine, and Serbia [3]. While Rusyn is predominantly classified as an East Slavic language due to its historical and linguistic origins, its classification is complicated by substantial influences from West and South Slavic languages [4].

Language variation, encompassing diachronic, diastratic, diaphasic, and diatopic aspects, presents significant challenges for Natural Language Processing (NLP) [5]. As noted by [6], dialectal differences and variations between national varieties of the same language affect the performance of applications such as machine translation and speech recognition. Consequently, there is growing interest in research on processing related languages, varieties, and dialects, as evidenced by numerous publications and scientific events, such as the VarDial workshops [6].

For low-resource languages like Rusyn, acquiring appropriate text corpora is particularly challenging [7]. Common solutions include speech transcription, as in the case of the ArchiMob corpus for German dialects, or the use of translations, as exemplified by the MADAR corpus for Arabic dialects [8]. [9] describe the challenges associated with creating NLP resources for the Rusyn language, proposing morphosyntactic lexicon induction using Slavic language resources.

Therefore, while specialized programs or applications exist for studying language characteristics, the application of a complex and sensitive model such as OpenAI Whisper represents an innovative approach. The integration of demographic factors with advanced language modeling techniques offers new perspectives on language change and maintenance in minority language communities, particularly valuable for understanding the evolution and preservation of the Rusyn language.

This work was supported by The Faculty of Mechanical Engineering and Robotics, AGH University of Krakow, Research Subsidy No. 16.16.130.942.

P. Małecki is with Faculty of Electrical Engineering and Communication, AGH University of Krakow, Krakow, Poland (e-mail: pawel.malecki@agh.edu.pl).

The primary objective of this study is to assess the recognition and classification efficacy of the Rusyn language by the OpenAI Whisper automatic speech recognition system and to evaluate the impact of dominant languages in the regions inhabited by Rusyns on the classification results. The scope of the research includes an analysis of audio data collected from radio broadcasts in Rusyn across four countries—Poland, Ukraine, Slovakia, and Serbia—with additional segmentation based on speaker age groups.

The paper provides an overview of the Whisper model, detailing its architecture and multilingual transcription capabilities, followed by a comprehensive description of the research methodology. This methodology encompasses the discussion of the audio database, the segmentation and acoustic analysis algorithms employed, and the parameters utilized in the model. In the results section, the classification outcomes of the Rusyn language are presented with respect to the speakers' countries of origin and age-related differences. The findings are discussed in the context of the influence of dominant languages on Rusyn speech and the assimilation processes underway, while also addressing issues related to the limitations of the model and the nuanced challenges of language classification. The study concludes with a summary that highlights key insights and outlines potential directions for future research in automatic speech recognition for minority languages.

The complex sociolinguistic status of Rusyn presents unique challenges for computational analysis. The absence of standardized corpora and the existence of multiple regional variants necessitate careful consideration in the application of machine learning models.

II. The OpenAI Whisper Language Model

The OpenAI Whisper model is an advanced Automatic Speech Recognition (ASR) system [10]. The model and its source code are available under an open-source license, enabling dynamic development of applications and related research in advanced speech processing. The model was trained on a dataset comprising over 680,000 hours of multilingual recordings with transcriptions (supervised data obtained from online repositories). This extensive and diverse dataset has allowed the system to achieve high robustness against varied accents, background noise, and specialized terminology, while supporting transcription in multiple languages.

The Whisper system architecture exemplifies an ASR implementation using a Transformer-based neural network in an encoder-decoder configuration. The audio data input to the model is divided into 30-second segments, then transformed into log-Mel spectrograms and subsequently processed by the encoder. The decoder predicts the corresponding text content, introducing additional informational tokens such as speaker language identification and phrase-level timestamps.

In training Whisper, a very large and diverse dataset was used, as illustrated in Figure 1, which shows the relationship between correctly recognized words and the amount of training data. The dataset contains approximately one-third of recordings in languages other than English. Despite significantly fewer test recordings in other languages, the recognition effectiveness based on a dataset of over 100 hours is at 80% or higher.

The Whisper model recognizes the language of the input speech through a multi-stage process integrated into its encoder-decoder architecture. Logarithmic Mel spectrograms are passed to the encoder, which extracts complex audio features. The encoder identifies patterns in the acoustic signal that characterize different languages, including phonetic and phonological features. During training, the model is exposed to audio data in many languages, along with corresponding text labels that include language identification tokens. During decoding, the decoder uses these learned tokens to determine the input language. Thanks to this extensive training set, the model analyzes signals globally across different languages and accents, improving its ability to accurately recognize the language. Whisper utilizes zero-shot learning capabilities, meaning it can generalize based on training data to recognize and transcribe languages it was not explicitly trained on. It achieves this by leveraging the multilingual nature of the training data and the advanced feature extraction capabilities of the encoder.

Although the Whisper model was not specifically trained on Rusyn language data, its application for the classification of Rusyn can be justified on several grounds. Firstly, the Whisper model was trained on a large, multilingual dataset that includes various Slavic languages (such as Polish, Ukrainian, Slovak, Czech, Serbian, Croatian, and Russian), which share similarities with Rusyn in terms of grammatical structure, phonetics, and lexicon. This training enables the model to recognize patterns and features characteristic of the Slavic language family, thereby allowing for an approximate classification of the Rusyn language despite the absence of direct training data.

Moreover, Whisper utilizes a zero-shot learning mechanism, meaning it can identify and classify languages on which it was not explicitly trained by generalizing and analyzing similarities between language patterns. In the case of the relatively niche Rusyn language, employing a model trained on a broad spectrum of languages facilitates its classification based on common features with other languages. Additionally, Whisper is built on the Transformer architecture, which is highly effective for processing and analyzing diverse acoustic patterns. This capability enables the detection of distinct phonetic and lexical characteristics inherent in the Rusyn language, even without direct training on it. Consequently, the model can classify Rusyn speech as most similar to languages that exhibit comparable phonetic traits.

III. Methodology

A. Database

For the linguistic analysis, archival recordings from the Lemko radio station lem.fm were utilized [11]. The audio dataset comprises a highly diverse collection of spoken Rusyn language samples, enabling a comprehensive investigation from a broad perspective. In this dataset, all segments containing music, jingles, and commercial advertisements were removed; however, for interviews and dialogues, only the voice of a single speaker was retained per instance. The curated

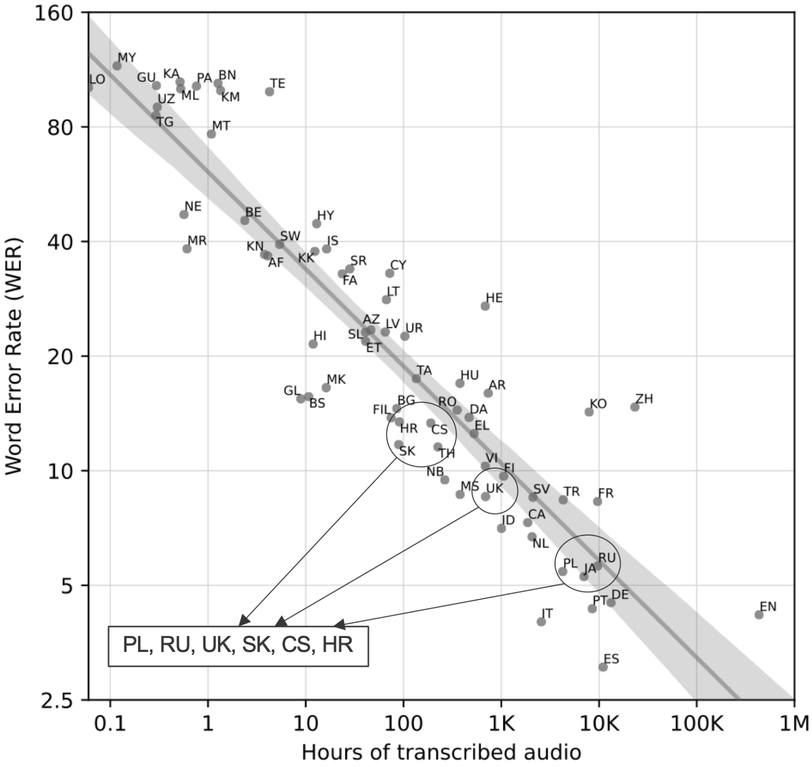


Fig. 1: Comparison of Word Error Rate (WER) across various languages using the Common Voice 15 dataset and OpenAI Whisper ASR [10]. The graph highlights selected Slavic languages (Polish, Ukrainian, Slovak, and Russian).

TABLE I: Characteristics of audio database used in language analysis

Abbr.	Country	Programs quantity	Total duration [h]	Speaker quantity
PL	Poland	209	45.4	over 100
UK	Ukraine	99	37.5	over 40
SK	Slovakia	131	23.2	over 50
CS	Serbia	31	11.5	10

dataset consists of 470 distinct broadcasts with a cumulative duration of approximately 121.8 hours. The distribution of speakers' origins—along with the number of broadcasts and the respective durations of the analyzed samples—is summarized in Table I.

Additional metadata assigned to each processed broadcast includes:

- **Age Group:** Two categories of speakers, delineated by whether they are above or below 70 years of age.
- **Speaker Identification:** Names or pseudonyms of the speaker.
- **Source File Name:** Names of the log files generated during the analysis process.
- **Total Recording Duration (in seconds):** The duration of each audio sample.

B. Speaker Language Recognition Using the Whisper Model

To facilitate data analysis and categorization, a Python script was developed to automate the selection and processing of audio files, perform linguistic analysis, and log the results. The script systematically ingests audio data from the curated dataset, applies necessary pre-processing routines, and utilizes the Whisper model for speaker language identification. This automated workflow not only streamlines file handling and analysis but also ensures that the outcomes—comprising language classification results and associated metadata—are accurately recorded for subsequent evaluation.

Figure 2 schematically illustrates the operational flow of the script. The individual stages are as follows:

- **Preprocessing phase:** A graphical user interface (GUI) is employed to select the directory containing the audio files. All fragments containing music or additional voices (i.e., voices other than that of the primary speaker) were previously removed to isolate the target speech. This ensures that the subsequent transcription and language recognition processes operate exclusively on the desired spoken content.
- **Segmentation:** Each audio file is divided into 30-second fragments.
- **Language recognition:** The Whisper model (large variant) is used for speech transcription and language detection. The model architecture comprises 32 layers with a width of 1280 neurons per layer and 20 attention heads, totaling 1.55 billion parameters. This configuration represents the largest model in the Whisper family, offering high transcription accuracy

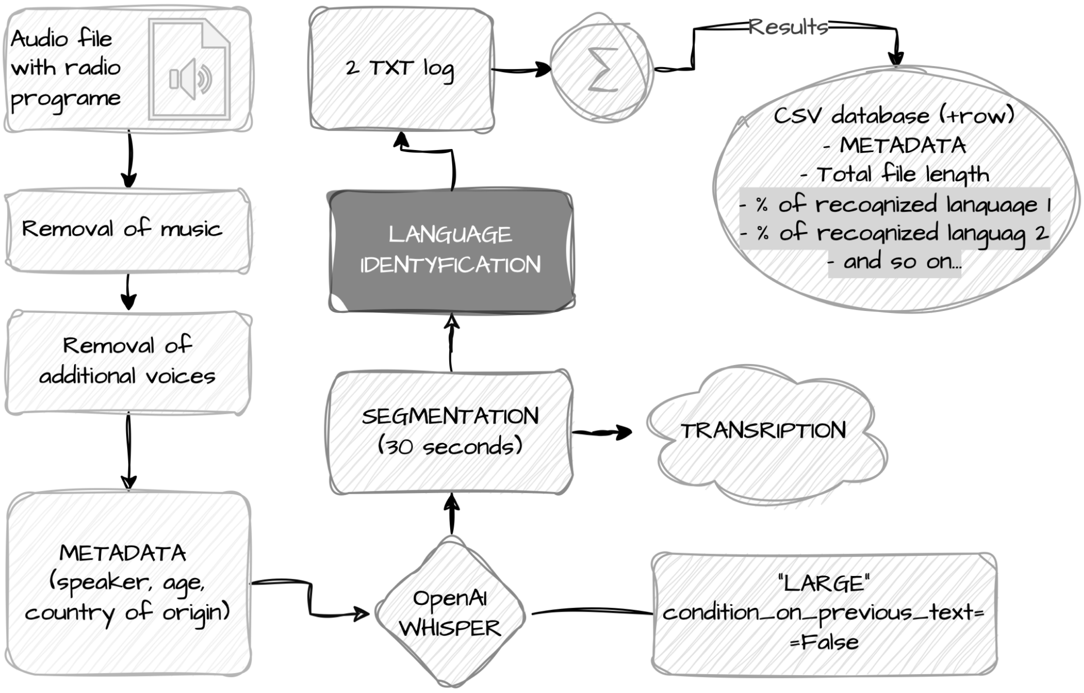

Audio file with radio programe
Removal of music
Removal of additional voices
METADATA (speaker, age, country of origin)
OpenAI WHISPER
"LARGE" condition_on_previous_text= =False
SEGMENTATION (30 seconds)
TRANSRIPTION
LANGUAGE IDENTYFICATION
2 TXT log
Σ
Results
CSV database (+row)
- METADATA
- Total file length
- % of recognized language 1
- % of recognized languag 2
- and so on...

Fig. 2: A schematic of the process for segmenting radio broadcasts, recognizing language, and logging results.

and robust language detection thanks to its complex structure. The model transcribes speech for each audio segment, and its output includes additional information such as the detected language. The parameters are configured with `no_speech_threshold = 0.8` and `condition_on_previous_text = False`, ensuring that each 30-second sample is processed independently. The network width defines the number of neurons per layer (i.e., the size of the representation vector), while the number of attention heads refers to the distinct "attention heads" in each attention mechanism layer, allowing the model to focus on different parts of the input sequence and capture complex dependencies within linguistic patterns.

- **Output logging:** The transcription results and detected language for each segment are recorded into individual text files.
- **Data aggregation:** The final results are aggregated into a CSV file, which includes the original broadcast metadata along with information on recognized languages and their percentage share in the total recording duration.

The script leverages the `librosa` library for audio file loading and `pydub` for segmentation.

IV. Results

For each broadcast, the recognized languages were ordered based on frequency, starting with the most frequently detected language (#1), along with the corresponding country identifier. The percentage share for each recognized language (e.g., primary language #1, secondary language #2, etc.) was then computed. These results are graphically represented in Figure 3, which displays the most frequently recognized languages in broadcasts conducted by Rusyn speakers residing in Poland, Slovakia, Ukraine, and Serbia.

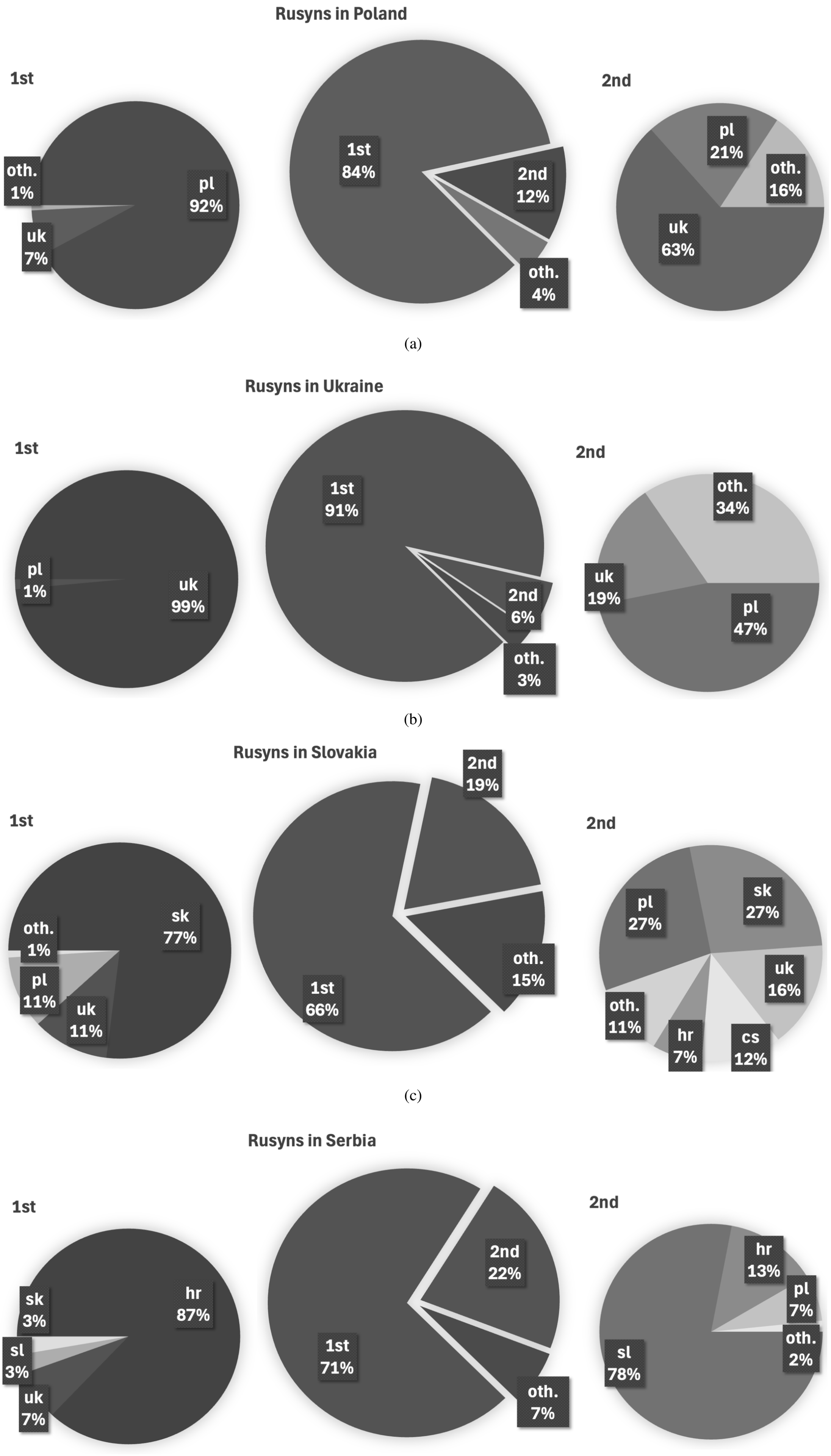
In broadcasts from Poland (Figure 3a), the primary recognized language (#1) accounted for 84% of the total recording duration. Within this primary language segment, 92% of the cases were identified as Polish and 7% as Ukrainian. The secondary recognized language (#2) comprised 12% of the audio, with 63% attributed to Ukrainian and 21% to Polish.

For broadcasts from Ukraine (Figure 3b), the primary language (#1) comprised 91% of the overall audio data, with 99% of these segments assigned to Ukrainian.

In broadcasts of Rusyn speakers in Slovakia (Figure 3c), the primary language (#1) represented 66% of the recording time, of which 77% was identified as Slovak. The secondary language (#2) accounted for 19% of the audio, with 27% each corresponding to Polish and Slovak, and 16% to Ukrainian.

For broadcasts from Serbia (Figure 3d), the primary recognized language (#1) made up 71% of the recordings, with 87% of these segments assigned to Croatian. The secondary language (#2) constituted 22% of the total duration, wherein 78% was identified as Slovenian.

Additional analyses were conducted for the Polish Rusyn community. Figure 4 presents the language recognition analysis in broadcasts by Polish Rusyn speakers, categorized into two age groups: individuals aged over 70 ("70+") and those aged below 70 ("70–"). Notably, the recordings from the "70+" group comprise over 60% of the Polish speaker dataset. The decision to restrict the age group analysis to Polish speakers was driven by two factors. Firstly, the sample sizes for speakers from other regions were insufficient to obtain statistically significant results. Secondly, determining the ages of speakers from other countries proved challenging, as several



(a)

(b)

(c)

(d)

Fig. 3: Percentage breakdown for Rusyn language classification by OpenAI Whisper for different countries of speaker origin.

broadcast editors responsible for the recordings are no longer active at the radio station.

In the "70+" group (Figure 4a), Polish was the most frequently recognized primary language (#1), accounting for 89% of broadcasts and 77% of the total analyzed duration. For the secondary recognized language (#2), Ukrainian was predominantly detected (75% within this category), representing 17% of all analyzed broadcasts. Additionally, a minor presence of other languages was observed, constituting 6% of the recognitions.

In contrast, within the younger "70–" group (Figure 4b), Polish was overwhelmingly the primary recognized language (#1), covering over 99% of the analyzed duration and accounting for 98% of the total broadcasts. Other languages in this group appeared only marginally.

V. Discussion

Based on the obtained results, it is evident that the dominant official languages of a given country significantly influence how the AI language model classifies the Rusyn language. With the exception of Rusyn speakers from Vojvodina (Serbia), the speech of Rusyn speakers from Ukraine, Poland, and Slovakia was classified by the algorithm as most similar to the dominant language in their respective countries. Although the flexional, lexical, and syntactic differences among variants of Rusyn are considered by both its speakers and linguists to be smaller than those between Rusyn and the dominant languages, the algorithm—designed to analyze global language similarities—tends to align more closely with the respective dominant languages.

In particular, Ukrainian dominates among Rusyn speakers from Poland and Slovakia, whereas the case in Serbia shows a different pattern. For Rusyn speakers from Vojvodina, the highest similarity was observed with Croatian and Slovenian, with Serbian contributing only marginally; however, the sample size from Serbia was notably smaller due to limited availability of material.

The exact mechanisms and criteria by which the model decides on language classification are not fully understood. The operation of a Transformer network is an inherently complex statistical process that identifies patterns in language data, enabling the detection of similarities without any genuine semantic comprehension.

Traditional ASR models—based on Hidden Markov Models (HMM)—exhibited a certain degree of determinism, which facilitated the analysis of how acoustic features influenced transcription and language classification. In contrast, modern ASR models that employ deep neural networks, including recurrent neural networks (RNN), convolutional neural networks (CNN), and Transformer architectures, are inherently implicit and non-deterministic. This means that identical speech inputs can yield slightly different transcriptions across separate instances. The non-deterministic nature arises from the vast number of model parameters and the complex, nonlinear relationships between them, complicating any direct analysis or extraction of the specific acoustic signal components that determine the transcription outcome. In other words, pinpointing exactly which portions of the speech signal influenced the classification of individual phonemes or words is extremely challenging.

Nonetheless, advanced methods exist for examining and interpreting ASR models [12], including:

- **Analysis of attention layers**, which allows for visualizing the dependencies between segments of the speech signal and the elements of the transcription, thereby highlighting the portions of the audio that the model "attends to" during recognition.
- **Gradient-based methods**, which enable the identification of input signal segments that most significantly impact the activation of particular neurons and, consequently, the final transcription.
- **Input perturbation techniques**, which involve introducing minor modifications to the speech signal and observing the resulting changes in transcription, thereby identifying key acoustic features.

The results also indicate an almost complete absence of Russian language recognition—a finding that is surprising given its significant presence and influence in Ukraine, especially in the eastern regions, from where some of the speakers originated.

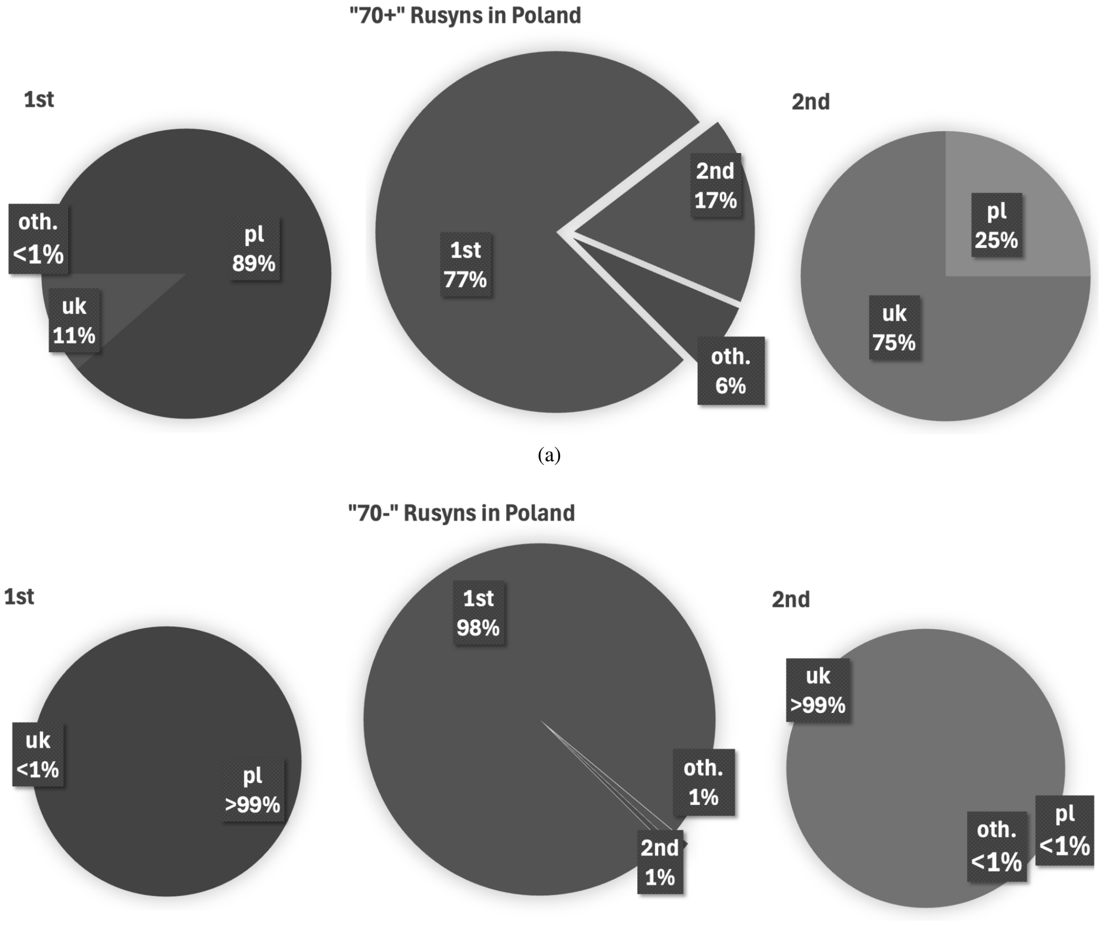
Furthermore, the analysis of the Lemko language (Rusyn variant from Poland) revealed significant variations in classification results based on the age of the speakers. As a result, an additional analysis was conducted by splitting the samples into two age groups: those below 70 years and those 70 years and older. For the 70+ group, the similarity between the Lemko language and Polish was considerably lower than for the younger group, suggesting a progressive process of linguistic assimilation, particularly in the realms of phonetics and phonology. In contrast, with respect to lexical items, older speakers exhibited a comparable—or even higher—frequency of Polish loanwords, which may indicate that different mechanisms of linguistic assimilation are at play across generations.

VI. Summary

In this study, the capabilities of the Whisper model for recognizing and classifying the Rusyn language were examined, even though the model was not directly trained on Rusyn materials. The investigation relied on radio broadcasts produced by Rusyn speakers residing in Poland, Ukraine, Slovakia, and Serbia, with additional consideration given to differences among age groups.

The results indicate that the Whisper model predominantly classifies Rusyn speech in a manner that mirrors the dominant official language of each respective country. This outcome suggests that the influence of dominant languages plays a critical role in shaping the classification of minority languages. Notably, there is also evidence of a higher degree of linguistic assimilation among younger speakers. This is particularly apparent in Poland, where younger speakers exhibit a considerably stronger alignment with Polish phonetics and phonology compared to the older generation, even though older speakers might display a similar or greater frequency of Polish loanwords. Such findings point to distinct assimilation mechanisms operating across generations.

Planned future research aims to modify the Whisper model by eliminating its bias toward the dominant language of each



(a)

(b)

Fig. 4: Percentage breakdown for Rusyn language classification by OpenAI Whisper for Polish-located speakers differentiated by age group.

country. This adjustment should allow for a more accurate evaluation of the similarity between the Rusyn language and other Slavic languages without interference from the official language influences. Additionally, further analyses are proposed to investigate the recognition of language in the context of singing—specifically, through the study of Rusyn and other Slavic songs—where the vocal emission mechanisms differ notably from those of spoken language.

Together, these efforts are expected to contribute to a more precise classification of minority languages, as well as to a deeper understanding of the processes underlying linguistic assimilation and globalization.

References

[1] N. Kushko, "Literary standards of the rusyn language: The historical context and contemporary situation," *The Slavic and East European Journal*, vol. 51, no. 1, pp. 111–132, 2007.

[2] A. G. Nikitin, I. T. Kochkin, C. M. June, C. M. Willis, I. Mcbain, and M. Y. Videiko, "Mitochondrial dna sequence variation in the boyko, hutsul, and lemko populations of the carpathian highlands," *Human Biology*, vol. 81, no. 1, pp. 43–58, 2009. [Online]. Available: https://doi.org/10.3378/027.081.0104

[3] A. Plišková, "Practical spheres of the rusyn language in slovakia," *Studia Slavica Academiae Scientiarum Hungaricae*, vol. 53, no. 1, pp. 95–115, 2008. [Online]. Available: https://doi.org/10.1556/SSlav.53.2008.1.6

[4] M. Moser, "Rusyn: A new-old language in-between nations and states," in *The Palgrave Handbook of Slavic Languages, Identities and Borders*, T. Kamusella, M. Nomachi, and C. Gibson, Eds. Palgrave Macmillan UK, 2016, pp. 124–139. [Online]. Available: https://doi.org/10.1007/978-1-137-34839-5_7

[5] S. U. Maheswari, A. Shahina, and Nayeemulla, "A study on the impact of lombard effect on recognition of hindi syllabic units using cnn based multimodal asr systems," *Archives of Acoustics*, 2020. [Online]. Available: https://doi.org/10.24425/aoa.2020.134058

[6] M. Zampieri, P. Nakov, and Y. Scherrer, "Natural language processing for similar languages, varieties, and dialects: A survey," *Natural Language Engineering*, vol. 26, no. 6, pp. 595–612, 2020. [Online]. Available: https://doi.org/10.1017/S1351324920000492

[7] Y. Scherrer and A. Rabus, "Neural morphosyntactic tagging for rusyn," *Natural Language Engineering*, vol. 25, no. 5, pp. 633–650, 2019.

[8] H. Bouamor, S. Hassan, and N. Habash, "The madar shared task on arabic fine-grained dialect identification," in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, 2019, pp. 199–207.

[9] A. Rabus and Y. Scherrer, "Lexicon induction for spoken rusyn–challenges and results," in *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, 2017, pp. 27–32.

[10] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," https://doi.org/10.48550/ARXIV.2212.04356, 2022.

[11] Trochanowski. (2022) Jezyk łemkowski. Lem.fm – . [Online]. Available: https://www.lem.fm/jezyk-lemkowski/

[12] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, "Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions," *Information Fusion*, vol. 81, pp. 203–239, 2022.