RoDia: A New Dataset for Romanian Dialect Identification from Speech

Codruţ Rotaru[1] and **Nicolae Cătălin Ristea**[1,2] and **Radu Tudor Ionescu**[1]
[1]University of Bucharest, Romania
[2]National University of Science and Technology Politehnica Bucharest, Romania
Corresponding author: raducu.ionescu@gmail.com

Abstract

We introduce RoDia, the first dataset for Romanian dialect identification from speech. The RoDia dataset includes a varied compilation of speech samples from five distinct regions of Romania, covering both urban and rural environments, totaling 2 hours of manually annotated speech data. Along with our dataset, we introduce a set of competitive models to be used as baselines for future research. The top scoring model achieves a macro F_1 score of 59.83% and a micro F_1 score of 62.08%, indicating that the task is challenging. We thus believe that RoDia is a valuable resource that will stimulate research aiming to address the challenges of Romanian dialect identification. We release our dataset at https://github.com/codrut2/RoDia.

1 Introduction

Spoken dialect identification emerged as a challenging task aiming to achieve a fine-grained distinction between varieties of a certain language, having similar implications to spoken language identification (Barnard et al., 2014; Kimanuka et al., 2024; Ma et al., 2007). Despite being a more delicate task, spoken dialect identification received comparatively lower attention, most of it being devoted to dialect identification for widely spoken languages, such as English (Weinberger and Kunath, 2011), Chinese (Zhang et al., 2022), and Arabic (Ali et al., 2017, 2019; Shon et al., 2020). Spoken dialect identification for low-resource languages, such as Swiss German (Dogan-Schönberger et al., 2021; Plüss et al., 2023) and Finnish (Hämäläinen et al., 2021), has remained relatively underexplored (Ranathunga et al., 2023; Barnard et al., 2014; Hämäläinen et al., 2021). Different from prior studies, we focus on spoken language identification in Romanian, a low-resource language characterized by its intricate dialectal variations within the country of Romania. Romanian, a Romance language with Latin roots, boasts a rich linguistic landscape shaped by historical, geographical, and sociocultural factors (Barbu-Mititelu et al., 2018). However, despite its linguistic complexity, Romanian remains a low-resource language, with limited studies dedicated to understanding its regional linguistic diversity. This scarcity of resources is not unique to Romanian, numerous other languages around the world having similar challenges due to their lower visibility on the global linguistic stage (Ranathunga et al., 2023; Barnard et al., 2014; Hämäläinen et al., 2021). Notably, the VarDial workshop is one of the main drivers for growing the interest around language variety and dialect identification, through the organization of multiple shared tasks each year (Aepli et al., 2022; Chakravarthi et al., 2021; Gaman et al., 2020; Zampieri et al., 2019).

Due to the success of deep learning frameworks in speech processing (Mehrish et al., 2023), researchers started to employ such methods in the area of low-resource languages (Chan and Lane, 2015; Al-Ghezi et al., 2023). This has led to a growing need for resources on low-resource languages. Considering dialect identification datasets across different languages, we can distinguish between two types of resources: text-based datasets (Bouamor et al., 2018; Butnaru and Ionescu, 2019; Francom et al., 2014; Găman et al., 2023; Găman and Ionescu, 2022) and speech-based datasets (Ali et al., 2017, 2019; Shon et al., 2020; Dogan-Schönberger et al., 2021; Plüss et al., 2023; Hämäläinen et al., 2021). While various languages have benefited from text-based resources that leverage written materials capturing linguistic variations, the auditory dimension of dialects adds an intricate layer of complexity. Text data, although valuable, might not fully encapsulate the nuanced phonetic and prosodic characteristics that are pivotal in dialect differentiation. In contrast to text datasets, audio datasets (Ali et al., 2017, 2019; Shon et al.,

Findings of the Association for Computational Linguistics: NAACL 2024, pages 279–286
June 16-21, 2024

Figure 1: The administrative regions of Romania and the dominant dialect spoken within each region. RoDia is the first benchmark to contain samples representing these five Romanian dialects.

2020; Dogan-Schönberger et al., 2021; Plüss et al., 2023) offer a more holistic representation, capturing not only the lexical disparities, but also the subtle intonations and accents inherent in speech.

To the best of our knowledge, RoDia is the first dataset to tackle spoken dialect identification in the Romanian landscape in accordance with historical, geographical, and sociocultural factors, encouraging the research in this low-resource language. Although there are two text datasets addressing Romanian dialect identification, MOROCO (Butnaru and Ionescu, 2019) and MOROCO-Tweets (Găman and Ionescu, 2022), these cover only two dialects: Romanian (equivalent to the *Muntenesc* dialect) and Moldavian (*Moldovenesc*). In contrast, our dataset is focused on speech and covers five Romanian dialects, as shown in Figure 1. We underline that the extra dialects, namely *Ardelenesc*, *Bănățean*, and *Oltenesc*, are very well characterized by phonetic differences captured only in speech. This explains why MOROCO (Butnaru and Ionescu, 2019) and MOROCO-Tweets (Găman and Ionescu, 2022) only contain text samples from the other two dialects.

The number of audio datasets available in Romanian is rather low (Avram et al., 2022; Georgescu et al., 2020), confirming that Romanian is indeed a low-resource language. We underline that existing datasets comprising Romanian speech samples are mainly focused on automatic speech recognition, ignoring the diversity of dialects within the region. In contrast, our dataset is specifically designed to represent five distinct dialects spoken in Romania: *Muntenesc* (accepted as the official language of Romania), *Ardelenesc*, *Moldovenesc*, *Oltenesc*, and *Bănățean*[1].

2 Dataset

Data collection and annotation. We collected the vast majority of the audio samples by gathering interviews and shows from local TV channels from all five regions considered in the dataset (see Figure 1). To obtain a clean dataset, we employed a rigorous selection process. First, we manually cropped the gathered audio files with respect to each speaker, e.g. we split an interview into multiple samples, such that each sample contains a single speaker. Next, we discard samples with interfering speakers and with a low perceived intelligibility. To make sure the label assignment is robust, we submitted all samples gathered from the TV channels to local annotators to validate the assigned labels. The manual validation process eliminates samples with a questionable dialect. In addition, a small proportion of data was acquired by recording citizens native to the five regions, who were asked to read some random texts from the Romanian Wikipedia, in their own dialect. We cropped the recorded samples to minimize the amount of silence at the start and the end of each audio sample. Upon curating the gathered data, we are left with a clean dataset containing 2,768 audio samples, each having between 2.5 and 5.0 seconds of speech. The sample rate of all samples is 44.1 kHz.

We divide the dataset into 2,164 samples for training and 604 samples for testing, such that there is no overlap between speakers in training and test. Without separating the speakers between training and test, a model that overfits to certain speaker-specific features that are not related to dialect (e.g. pitch, loudness, rate, etc.) will reach high scores on the test set. However, these scores are unlikely to represent the actual performance of the model in a real-world scenario, where the audio samples come from unknown speakers. We thus consider that a more realistic evaluation is to separate the speakers. In the proposed setting, models that learn patterns related to speakers will not be able to capitalize on features unrelated to dialect identification.

There are five local annotators (one per region), who annotated all samples. The inter-rater Quadratic Weighted Kappa score is 0.83, indicating that the collected labels exhibit a substantial

[1]We refer to the original (untranslated) dialect names, since most of them have no translation in English.

Class	#speakers	Train			Test		
		#samples	SNR	SRR	#samples	SNR	SRR
Ardelenesc	47	427	28.8	36.4	119	30.5	38.5
Bănăţean	67	424	23.1	34.6	99	25.0	37.7
Moldovenesc	47	384	25.6	32.4	206	25.7	27.8
Muntenesc	64	603	29.0	35.3	106	26.7	37.8
Oltenesc	31	326	26.6	31.2	74	26.5	33.7
Overall	256	2164	26.9	34.2	604	26.8	34.0

Table 1: The number of training and test samples for each class in our dataset. For reference, we include the average SNR and SRR values for each category. The number of speakers per dialect is also provided.

agreement among human evaluators. The average accuracy of our raters is 86%, indicating that the task is fairly easy for humans. Note that all raters are native Romanian speakers who speak the literary language, as well as at least one of the five dialects.

The collected samples comprise interviews and read speech found on YouTube. The read speech is actually gathered from videos where various speakers read from different books (without any influence or preparation from our side). The percentage of read speech is 21%.

Aside from dialect labels, our annotators also labeled each audio sample with the gender and age of each speaker. More precisely, the age and gender of each speaker is estimated by two annotators who had to analyze both video and audio modalities. The age annotation consists in classifying each speaker into a 10-year age group, after watching the video available for the respective speaker. In summary, our audio samples come with dialect, age and gender labels, enabling the study of additional tasks such as gender prediction or age estimation from speech.

Dataset statistics. For a more comprehensive view, we present both demographic information and audio quality statistics for our new dataset. In Table 1, we report the number of samples for each dialect in both train and test splits, as well as the signal-to-noise ratio (SNR) and signal-to-reverberation ratio (SRR). Regarding data quality, we note that the SNR and SRR values are consistently higher than 23 dB, highlighting that the audio samples have relatively low noise and reverberation. The *Muntenesc* dialect has the largest number of audio samples. This dialect was easy to collect, since it represents the literary language, which is often borrowed by speakers native to other regions. On the opposite side, the *Oltenesc* dialect is least represented, having only 400 audio samples. However, the distribution gap between the five classes is not high enough to pose significant challenges to machine learning models.

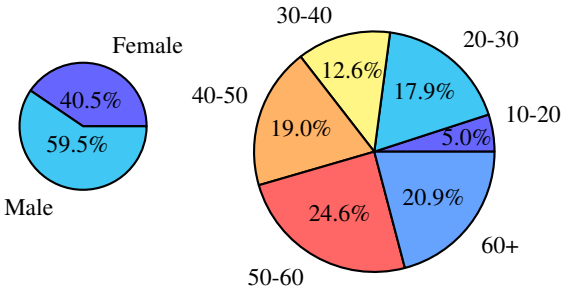

Figure 2: Age and gender statistics for the RoDia dataset.

We present demographic information in Figure 2. In terms of demographic insights, RoDia exhibits a relatively balanced gender distribution, having 59.5% male and 40.5% female speakers. Aside from separating the speakers between training and test, we also made sure to have similar demographics for the train and test splits, reducing unnecessary distribution gaps. In summary, we consider that RoDia is a suitable resource for spoken dialect identification.

3 Experiments

Baseline methods. We compile a lineup of four state-of-the-art neural architectures for speech processing to form a set of competitive dialect identification baselines for our novel dataset. We consider both convolutional (He et al., 2016) and transformer-based neural networks (Gong et al., 2021; Ristea et al., 2022), as well as a hybrid architecture (Baevski et al., 2020). We employ the ResNet-18 (He et al., 2016) convolutional network, as it was previously used for audio classification tasks (Ristea and Ionescu, 2020). Additionally, we explore two transformer-based architectures, namely the Audio Spectrogram Transformer (AST) (Gong et al., 2021) and the Separable Transformer (SepTr) (Ristea et al., 2022), due to their high performance in audio classification. We also employ the wav2vec 2.0 model (Baevski et al., 2020),

Model	Ardelenesc			Bănăţean			Moldovenesc			Muntenesc			Oltenesc			Overall F_1	
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	Micro	Macro
ResNet-18	59.45	73.94	65.91	47.54	58.58	52.48	73.98	62.13	67.54	44.11	56.60	49.58	88.00	29.72	44.44	58.94	55.99
AST	66.92	73.10	69.87	43.62	65.65	52.41	77.92	58.25	66.66	48.96	66.98	56.57	92.30	32.43	48.00	60.76	58.70
SepTr	67.71	72.26	69.91	41.31	69.69	51.87	81.75	58.73	68.36	53.33	67.92	59.75	85.18	31.08	45.54	61.42	59.08
wav2vec 2.0	68.00	71.42	69.67	43.13	66.66	52.38	77.63	60.67	68.10	54.34	70.75	61.47	88.88	32.24	47.52	62.08	59.83

Table 2: Spoken dialect identification results of ResNet-18 (He et al., 2016), AST (Gong et al., 2021), SepTr (Ristea et al., 2022) and wav2vec 2.0 (Baevski et al., 2020) on the RoDia test set. For a comprehensive evaluation, we report both per dialect and overall results. The best score on each column is highlighted in blue.

which uses a hybrid architecture combining the advantages of both convolutional and transformer blocks. All models are trained in the multi-class setting, since the ground-truth labels are constructed in a similar manner: one audio sample belongs to only one dialect.

Preprocessing. For models operating in the time-frequency domain (He et al., 2016; Gong et al., 2021; Ristea et al., 2022), we apply the Short-Time Fourier Transform with a window size of 512 and a hop size of 256. Then, we compute the square root of the magnitude, obtaining the spectrogram map. The other steps and parameters are exactly as described in the original papers. For wav2vec 2.0 (Baevski et al., 2020), we apply the preprocessing steps described by the authors, which are mainly used for normalization. In all our experiments, we use the following data augmentation methods: noise perturbation, time shifting, speed perturbation, mix-up and SpecAugment (Park et al., 2019).

Evaluation metrics. We report the precision (P), recall (R), and F_1 scores computed for each dialect. These metrics provide insights into the ability of models to correctly classify instances within each class. To quantify the overall performance, we aggregate the individual scores via the micro and macro F_1 measures. The micro F_1 score combines the performance metrics across all examples, while the macro F_1 score offers a balanced average of the F_1 scores across all classes.

Training environment. All models are optimized on an Nvidia GeForce GTX 3090 GPU with 24 GB of VRAM.

Hyperparameter tuning. For each model, we employed grid search to find the optimal learning rate (between 10^{-2} and 10^{-6}) and the optimal batch size (between 8 and 128 samples). We take the wav2vec 2.0 (Baevski et al., 2020) model, which is pretrained on English data, and fine-tune it for 10 epochs on RoDia using a learning rate of 10^{-5} and mini-batches of 16 samples. We train ResNet-18,

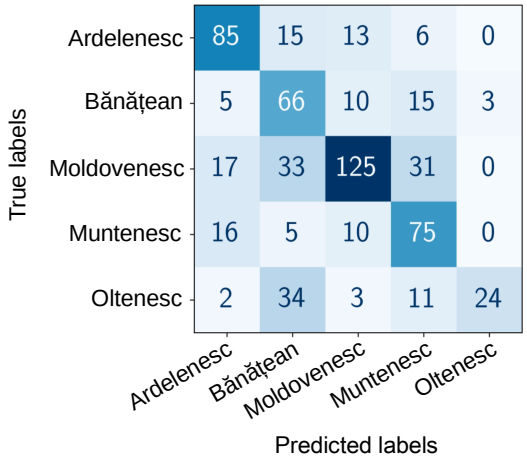

Figure 3: Confusion matrix on the test set for the wav2vec 2.0 (Baevski et al., 2020) model.

AST (Gong et al., 2021) and SepTr (Ristea et al., 2022) from scratch. The models are trained for 50 epochs with early stopping, using a learning rate of 10^{-4} and mini-batches of 32 samples. All models are optimized with the Adam optimizer (Kingma and Ba, 2015).

Dialect identification results. In Table 2, we present the spoken dialect identification results of the baseline models on the RoDia test set. The convolutional network, ResNet-18, obtains the lowest overall performance. Still, ResNet-18 reaches competitive F_1 scores for the *Bănăţean* and *Moldovenesc* dialects. Unlike the other baselines, the ResNet-18 model struggles with the *Muntenesc* and *Ardelenesc* dialects, which explains its low overall performance. The transformer-based models, AST (Gong et al., 2021) and SepTr (Ristea et al., 2022), yield superior results, with a slight upper hand from the SepTr model. In terms of the overall F_1 scores, the best model appears to be wav2vec 2.0 (Baevski et al., 2020). However, the F_1 scores per dialect seem to tell a slightly different story, since the wav2vec 2.0 is outperformed by at least one of the other models on four dialects: *Ardelenesc*, *Bănăţean*, *Moldovenesc* and *Oltenesc*. The competitive edge of wav2vec 2.0 lies in its ability to better identify the *Muntenesc* dialect. We underline

that the audio samples were recorded in uncontrolled scenarios, so the reported results directly reflect the capability of systems in the real-world case.

To further assess the behavior of the best baseline model, namely wav2vec 2.0, we consider its confusion matrix illustrated in Figure 3. The confusion matrix reveals some interesting patterns. While the model tends to mislabel samples from the *Oltenesc* dialect, we observe that most of these mistakes are caused by a high confusion with the *Bănăţean* dialect. Since Banat and Oltenia are neighboring regions, there are several similarities between these two dialects. For instance, both dialects are characterized by the frequent use of the perfect simple tense, which is hardly encountered in the other Romanian dialects. Another noticeable problem is with the *Moldovenesc* dialect, which is often wrongly identified as *Bănăţean* and *Muntenesc*. The confusion between the *Moldovenesc* and *Bănăţean* dialects is caused by the fact these two dialects kept the form of words such as *câne* (dog), *pâne* (bread) and *mâne* (tomorrow), from the old Romanian. In the literary language, as well as the other dialects, these words are pronounced with an 'i' before the consonant 'n', as follows: *câine*, *pâine* and *mâine*. The confusion with the *Muntenesc* dialect can be attributed to the fact that some residents of the southern part of Moldavia lost some of the dialectal features, e.g. they prefer to use the word *pantofi* to refer to shoes, and the word *papuci* to refer to slippers, just as the residents of Muntenia. In contrast, the residents of the northern side of Moldavia regularly use *papuci* to refer to shoes, and *şlapi* to refer to slippers. In summary, the confusion matrix shows that Romanian dialect identification is not an easy task, requiring researchers to address specific issues in order to come up with more accurate models in the future.

The confusion matrix illustrated in Figure 3 also shows that the training data distribution does not affect wav2vec 2.0. For example, the Moldovenesc dialect is the second-least popular dialect in our dataset, but wav2vec places many of the test samples into the Moldovenesc class. Overall, the confusion matrix of wav2vec reflects the test data distribution, although the model was trained on a slightly different class distribution. This confirms that the imbalance is not high enough to bias models.

Speech recognition results. We manually transcribed our data samples to test the performance of a state-of-the-art automatic speech recognition (ASR) system on RoDia. Then, we applied the open source Whisper-Large model (Radford et al., 2023) on our test set and obtained the word error rates reported in Table 3. The *Muntenesc* dialect is almost identical to the literary language, explaining why it exhibits the lowest WER. The WER obtained by the Whisper-Large model for the Romanian language on the Common Voice dataset is 19.8%. The difference between the WER for the Muntenesc dialect and the WER reported on Common Voice can be attributed to the distribution gap between the RoDia and the Common Voice datasets. Considering the generally higher error rates for the other Romanian dialects shown in Table 3, we conclude that ASR for dialectal speech is more difficult. This justifies the utility of our novel dataset for ASR.

Dialect	WER
Ardelenesc	31.5%
Bănăţean	30.2%
Moldovenesc	32.8%
Muntenesc	24.1%
Oltenesc	29.7%

Table 3: Word error rates (WER) of the Whisper-Large model (Radford et al., 2023) on the five dialects from the RoDia dataset. The Whisper-Large model is trained on the literary Romanian language. The ASR transcripts are compared with manual transcripts to establish the performance of the Whisper-Large model.

4 Conclusion

In this paper, we introduced RoDia, the first dataset for Romanian dialect identification from speech. Our dataset contains 2,768 speech samples representing five Romanian dialects. The audio samples were manually annotated with dialect, age and gender labels, enabling the study of spoken dialect identification in a realistic scenario, where the speakers in the training and test splits are disjoint. We conducted experiments with four state-of-the-art speech processing models, establishing a range of baseline performance levels for future research.

5 Limitations

This work is focused on spoken Romanian dialect identification, but the performance levels of the considered approaches might be different on other languages. Due to our specific focus on the Romanian language, we did not evaluate the performance of the considered models across other languages.

However, we consider that the evaluation on other languages is beyond the scope of the current study.

Another limitation of our work is the slightly limited number of samples with manual labels included in our corpus. This limitation is caused by scarcity of resources available online. Most of the Romanian video or audio samples available online use the literary language. Local news and content creators commonly use the literary language taught in schools. Hence, dialects are mostly used by rural residents, which often have no Internet access. This situation significantly limits the dialectal resources that are publicly available.

6 Ethics Statement

The manual labeling was carried out by volunteers who agreed to annotate the audio samples for free. Prior to the annotation, they also agreed to let us publish their labels along with the dataset. Our data is collected from YouTube, which resides in the public web domain. We note that the European regulations[2] allow researchers to use data in the public web domain for non-commercial research purposes. Thus, we release our data and code under the CC BY-NC-SA 4.0 license[3].

During data collection, we made sure the audio samples do not contain information that names or uniquely identifies individual people.

References

Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. 2022. Findings of the VarDial evaluation campaign 2022. In *Proceedings of VarDial*, pages 1–13.

Ragheb Al-Ghezi, Yaroslav Getman, Ekaterina Voskoboinik, Mittul Singh, and Mikko Kurimo. 2023. Automatic rating of spontaneous speech for low-resource languages. In *Proceedings of SLT*, pages 339–345.

Ahmed Ali, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James Glass, Steve Renals, and Khalid Choukri. 2019. The MGB-5 Challenge: Recognition and Dialect Identification of Dialectal Arabic Speech. In *Proceedings of ASRU*, pages 1026–1033.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic MGB-3. In *Proceedings of ASRU*, pages 316–322.

[2] https://eur-lex.europa.eu/eli/dir/2019/790/oj

[3] https://creativecommons.org/licenses/by-nc-sa/4.0/

Andrei-Marius Avram, Mihai-Virgil Nichita, Razvan-George Bartusica, and Mădălin-Virgil Mihai. 2022. RoSAC: A Speech Corpus for Transcribing Romanian Emergency Calls. In *Proceedings of COMM*, pages 1–5.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of NeurIPS*, volume 33, pages 12449–12460.

Verginica Barbu-Mititelu, Elena Irimia, and Dan Tufiş. 2018. The reference corpus of the contemporary Romanian language (CoRoLa). In *Proceedings of LREC*, pages 1178–1185.

Etienne Barnard, Marelie H. Davel, Charl van Heerden, Febe de Wet, and Jaco Badenhorst. 2014. The NCHLT speech corpus of the South African languages. In *Proceedings of SLTU*, pages 194–200.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic Dialect Corpus and Lexicon. In *Proceedings of LREC*, pages 3387–3396.

Andrei Butnaru and Radu Tudor Ionescu. 2019. MOROCO: The Moldavian and Romanian Dialectal Corpus. In *Proceedings of ACL*, pages 688–698.

Bharathi Raja Chakravarthi, Gaman Mihaela, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke, Eswari Rajagopal, Yves Scherrer, and Marcos Zampieri. 2021. Findings of the VarDial evaluation campaign 2021. In *Proceedings of VarDial*, pages 1–11.

William Chan and Ian Lane. 2015. Deep convolutional neural networks for acoustic modeling in low resource languages. In *Proceedings of ICASSP*, pages 2056–2060.

Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2021. SwissDial: Parallel Multidialectal Corpus of Spoken Swiss German. *arXiv preprint arXiv:2103.11401*.

Jerid Francom, Mans Hulden, and Adam Ussishkin. 2014. ACTIV-ES: A comparable, cross-dialect corpus of everyday Spanish from Argentina, Mexico, and Spain. In *Proceedings of LREC*, pages 1733–1737.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A report on the VarDial evaluation campaign 2020. In *Proceedings of VarDial*, pages 1–14.

Mihaela Găman and Radu Tudor Ionescu. 2022. The unreasonable effectiveness of machine learning in Moldavian versus Romanian dialect identification. *International Journal of Intelligent Systems*, 37(8):4928–4966.

Alexandru-Lucian Georgescu, Horia Cucu, Andi Buzo, and Corneliu Burileanu. 2020. RSC: A Romanian read speech corpus for automatic speech recognition. In *Proceedings of LREC*.

Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. In *Proceedings of INTERSPEECH*, pages 571–575.

Mihaela Găman, Adrian-Gabriel Chifu, William Domingues, and Radu Tudor Ionescu. 2023. FreCDo: A New Corpus for Large-Scale French Cross-Domain Dialect Identification. In *Proceedings of KES*, pages 366–373.

Mika Hämäläinen, Khalid Alnajjar, Niko Partanen, and Jack Rueter. 2021. Finnish Dialect Identification: The Effect of Audio and Text. In *Proceedings of EMNLP*, pages 8777–8783.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of CVPR*, pages 770–778.

Ussen Kimanuka, Ciira wa Maina, and Osman Büyük. 2024. Speech Recognition Datasets for Low-resource Congolese Languages. *Data in Brief*, 52:109796.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic gradient descent. In *Proceedings of ICLR*.

Bin Ma, Haizhou Li, and Rong Tong. 2007. Spoken Language Recognition Using Ensemble Classifiers. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2053–2062.

Ambuj Mehrish, Navonil Majumder, Rishabh Bharadwaj, Rada Mihalcea, and Soujanya Poria. 2023. A review of deep learning techniques for speech processing. *Information Fusion*, page 101869.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proceedings of INTERSPEECH*, pages 2613–2617.

Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, and Mark Cieliebak. 2023. STT4SG-350: A speech corpus for all Swiss German dialect regions. In *Proceedings of ACL*, pages 1763–1772.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of ICML*, pages 28492–28518.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Nicolae-Cătălin Ristea and Radu Tudor Ionescu. 2020. Are you Wearing a Mask? Improving Mask Detection from Speech Using Augmentation by Cycle-Consistent GANs. In *Proceedings of INTERSPEECH*, pages 2102–2106.

Nicolae-Catalin Ristea, Radu Tudor Ionescu, and Fahad Shahbaz Khan. 2022. SepTr: Separable Transformer for Audio Spectrogram Processing. In *Proceedings of INTERSPEECH*, pages 4103–4107.

Suwon Shon, Ahmed Ali, Younes Samih, Hamdy Mubarak, and James Glass. 2020. ADI17: A Fine-Grained Arabic Dialect Identification Dataset. In *Proceedings of ICASSP*, pages 8244–8248.

Steven H. Weinberger and Stephen A. Kunath. 2011. The Speech Accent Archive: towards a typology of English accents. In *Proceedings of Corpus-based studies in language use, language learning, and language documentation*, pages 265–281. Brill.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei M. Butnaru, and Tommi Jauhiainen. 2019. A report on the third VarDial evaluation campaign. In *Proceedings of VarDial*, pages 1–16.

Yongmao Zhang, Zhichao Wang, Peiji Yang, Hongshen Sun, Zhisheng Wang, and Lei Xie. 2022. AccentSpeech: Learning Accent from Crowd-sourced Data for Target Speaker TTS with Accents. In *Proceedings of ISCSLP*, pages 76–80.

A Appendix

Peculiarities of Romanian dialects. As per Wikipedia[4], the Romanian dialects are not easy to classify, and their classification is still highly debated by experts, who proposed various classifications, ranging from 2 to even 20 dialects. Since there is no standard classification, we underline that the dialects from RoDia are not unanimously considered as dialects by experts. In Romanian, these are called "grai", which is translated (perhaps abusively) as "idiom" or "dialect" in English. Aside from phonetic differences, we note that there are a few hundred words that are specific to each such "grai". For example, lists of such words, called regionalisms, are available online[5]. The dialects included in our dataset have lists comprising between 300 and 800 regionalisms. In summary, the Romanian dialects have several distinctive features, such as:

- Phonetic differences, e.g. the word "ce" (what) is pronounced "ci" in the Moldovenesc dialect and "ce" in other dialects.
- Regionalisms, e.g. the word "melon" is translated as "pepene" in the Muntenesc dialect, "harbuz" in the Moldovenesc dialect, "lubeniţă" in the Bănăţean and Oltenesc dialects, and "curcubete" or "lebeniţă" in the Ardelenesc dialect.
- The addition of unnecessary dialect-specific words, e.g. "What time is it?" is normally translated as "Cât este ceasul?", but in the Ardelenesc dialect, people commonly use "Oare cât este ceasul?" (which can be translated as "I wonder what time is it"). In general, it is common to use "Oare" when addressing a question to another person in the Ardelenesc dialect.
- Preference for using a different past tense in the Oltenesc and Bănăţean dialects than in other Romanian dialects, e.g. for the phrase "I was", speakers of the Oltenesc and Bănăţean dialects dialects say "fusei", but the speakers of other Romanian dialects use "am fost".

Criteria for choosing the five dialects. To establish the set of dialects for RoDia, we used two criteria. On the one hand, we aimed to include as many dialects as possible. On the other hand, we were limited by the low number of audio samples available for dialects spoken in small regions (by low populations). We thus selected the top five most popular dialects, which are representative for the regions depicted in Figure 1 from our paper. In Table 4, we provide the size of the population in each region of Romania corresponding to one of the top seven Romanian dialects. Dialects that are not included in RoDia correspond to smaller subregions, e.g. Crişana and Maramureş, for which it is even harder to collect sufficient audio samples.

Dialect	Population
Ardeal (without Banat)	5.5M
Moldova	4.2M
Muntenia	3.3M
Oltenia	2M
Banat	1.25M
Crişana	1.2M
Maramureş	0.46M

Table 4: Population size for seven of the largest regions in Romania. Our dataset includes representative dialects for the top five regions.

[4] https://en.wikipedia.org/wiki/Romanian_dialects

[5] http://regionalisme.ro