Interaction Design and Architecture(s) Journal - IxD&A, N.62, 2024, pp. 141 - 152
DOI: 10.55612/s-5002-062-009

Echo: A Crowd-sourced Romanian Speech Dataset

Remus-Dan Ungureanu[1] and Mihai Dascalu[1,2]

[1] National University of Science and Technology POLITEHNICA Bucharest,
313 Splaiul Independentei, Sector 6, Bucharest, Romania
remus_dan.ungureanu@upb.ro, mihai.dascalu@upb.ro
[2] Academy of Romanian Scientists, Str. Ilfov, Nr. 3, 050044, Bucharest, Romania

Abstract. Romanian is the seventh most popular European language, with around 30 million speakers worldwide. Despite its popularity, the available speech resources are limited. As a result, there are few models that transcribe Romanian well, most of them being multilingual models that also cover less popular languages. Echo is a crowd-sourcing platform that has collected more than 300 hours of speech from various contributors. In this study, we document how a large speech dataset enables researchers to train automatic speech recognition, speaker verification, and diarization models to automatically process students' notes. We publicly release both the dataset and the Whisper-based baseline model as open-source.

Keywords: speech dataset, Romanian language, crowd-sourcing.

1 Introduction

In recent years, neural networks have rapidly advanced alongside faster hardware, enabling the use of more complex networks. This progress has also impacted speech processing, with traditional methods being replaced by end-to-end solutions powered by deep neural networks. However, the current challenge is the large amount of data required for those networks to learn accurately. Even more recently, having more data has become crucial in improving the accuracy of speech recognition systems [16].

A high-quality speech dataset is difficult to compile because various aspects must be covered [4]. First, the data must be collected from real-life environments and have a moderate amount of background noise. Additionally, the recordings should not be scripted but rather spontaneous, adding a degree of authenticity to the dataset. Finally, the diversity of recorded data is ideal, encompassing a range of speakers of various ages, genders, accents, and dictionaries.

These characteristics ensure the robustness and genericity of the model across different applications and ways of speaking. However, for very specific applications, the data may have a bias towards specific characteristics. When taking into consideration all the previous aspects, it is understandable why the cost of acquiring data is high. Furthermore, assessing the quality of speech data must consider all these criteria and is, in general, subjective.

The Romanian language remains significantly underrepresented when it comes to speech data despite being the seventh most popular language in the European Union, with over 30 million speakers worldwide. Even though Slavic-speaking countries surround Romania, the Romanian language is primarily a Romance language with influences and words coming from Bulgarian, German, Hungarian, and Turkish languages. One of the advantages of the language, which makes it easier to learn, is that it is a phonetic language, and words are pronounced exactly as they are spelled. However, there is also a characteristic that makes it more difficult to learn: the presence of 5 special characters called ”diacritics”: ˘a a mid-central vowel, pronounced like the ”i” in ”bird”; s, - voiceless postalveolar fricative, pronounced like the ”sh” in ”shoe”; t, - voiceless alveolar affricate, pronounced like ”ts” in ”pets”; ˆa andˆı - close central unrounded vowel, pronounced similar with the vowel heard in the second syllable of ”roses”.

The availability of local datasets is sparse, with only a few resources dedicated to building Romanian speech data. Some of the more recent multilingual models include Romanian among the languages they have been trained on; however, the representation of the language remains limited due to low representation in the global datasets: CommonVoice [2], FLEURS [5], and VoxPopuli [15].

To be more precise, the total cumulative hours of speech data available for Romanian amount to around 300 hours, which represents just a small fraction of the vast volume of English data used to train state-of-the-art models. As another example, Dutch is a language with similar number of speakers (˜30 million speakers), but has datasets [6] of over 2,000 hours of labeled dialogues. This lack of resources for the Romanian language hinders the development and advancement of speech recognition systems while pointing to the importance of expanding and diversifying the available datasets for this language.

In this paper, we present Echo, a crowd-sourced Romanian speech dataset and platform used to train a state-of-the-art model for automatic speech recognition (ASR) that can transcribe student notes recorded in the Romanian language. We are releasing the Echo dataset and a fine-tuned Whisper model which can provide accurate Romanian transcriptions.

The second section of the article describes the most important speech datasets at a local and global level. The next section describes the crowd-sourcing speech platform we have developed, followed by another section that describes the particularities of the collected data. The final two sections describe the results obtained for the pre-trained and fine-tuned Whisper models and our conclusions.

2 State of the Art

The speech resources for Romanian are sparse and the few models that have been trained were tested on various datasets, usually non-public, which makes comparisons between models even more difficult.

In our previous experiments [12], the most accurate deep neural network had a word-error-rate (WER) of 2.73% on the ”clean” dataset and 3.77% on the ”complete” dataset. The ”complete” dataset has been defined as the sum of all available public or permission-based access, summing up to 104 hours of recordings. The ”clean” dataset has been limited to only the datasets that were known to be of higher quality (both better recordings and more accurate transcripts).

In another more recent experiment [13], we tested the same neural network, but with a previous version of the dataset presented in this paper, and our conclusion was that the models trained on Echo were able to generalize better as the results were consistent when tested with the test portions of other datasets.

Whisper [10], the renowned project developed by OpenAI, is one of the few available models for the Romanian language which has been benchmarked against popular multilingual datasets such as Common Voice [2], FLEURS [5], and VoxPopuli [15]. Details about these datasets and their corresponding results are described in Table 1.

Table 1. Characteristics of popular multilingual datasets that include Romanian speech.

Name	Language count	Duration (hours)	Romanian duration (hours)	Romanian speakers
Common Voice	60	7,335	47	428
FLEURS	102	1,400	12	19
VoxPopuli	16	1,791	89	18

The absence of a unified dataset akin to LibriSpeech poses significant obstacles to the development of robust ASR systems and innovation. Table 2 presents the characteristics of the most popular and largest local Romanian datasets and their characteristics.

Table 2. The most popular speech resources for the Romanian language.

Name	Recordings	Duration (hours)	Speakers	Notes
SWARA [11]	19,279	˜21	17	Read speech, High quality
RoDigits [7]	15,389	˜38	11	Read speech, High quality, Limited vocabulary
RSC [8]	136,120	˜100	164	Read speech, Many recordings of single words

These datasets typically exhibit limitations in terms of size and speaker diversity. Read corpora are often small-scale, containing at most a few tens of hours of speech data and a limited number of speakers. On the other hand, spontaneous speech datasets offer larger volumes of data, potentially ranging from tens to hundreds of hours, but are usually automatically labeled, which results in decreased accuracy.

We strongly believe that more data can increase the development rate of models for various tasks. Previous work showed how thousands of hours of transcribed speech are necessary to reach acceptable performance ([3, 9]). Better models can help the development of practical speech-processing applications, including those for domains such as smart learning.

3 The Echo Platform – Crowd-sourcing Speech Data

Echo is a crowd-sourcing platform created to address the problem of scarcity of speech datasets for the Romanian language. As opposed to other crowd-sourcing initiatives, the complete process is implemented within the platform: data acquisition/collection, data filtering, statistics, and other management functionalities. The same platform is used by both contributors and administrators to record their voices or manage the recordings.

The collected dataset has evolved from a generic speech dataset to a multidomain dataset. Transcripts have been sourced from multiple types of texts, ranging from news and legal documents to fiction and poetry, in order to achieve diversity of speech data. This approach not only enriches the dataset with a broad spectrum of linguistic contexts but also enhances its applicability across diverse domains of automatic speech processing.

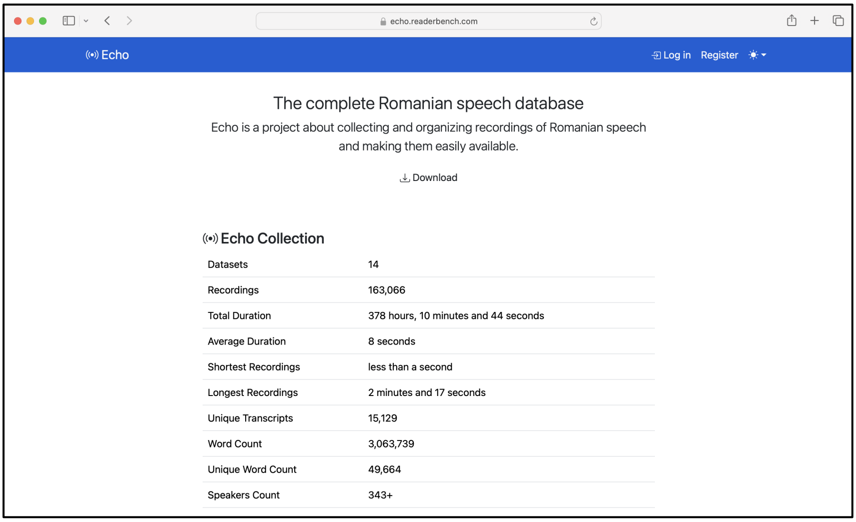
Echo is an open, accessible platform that welcomes contributions from individuals across all demographics. However, it has been particularly popular among students, resulting in a dataset skewed towards young contributors, predominantly aged between 19 and 25 years old.

Echo has also served as a collaborative platform for data collection efforts in Romanian. Notably, students from both the Faculty of Computer Science and Engineering and the Faculty of Philology and Literature have actively contributed to the dataset. This collaboration underscores the platform's potential to bridge disciplinary boundaries and foster interdisciplinary research.

Contributor Perspective The process of collecting speech data can be inherently cumbersome, which made us recognize the importance of a user-friendly interface. We have put in considerable effort to ensure that contributors encounter minimal barriers as they contribute to collecting Romanian speech data. To further enhance the contributor experience and incentivize active participation, Echo incorporates gamification elements into its platform design. Contributors can earn points based on their contributions, and their efforts are recognized through a global leaderboard.

Administrator Perspective From the administrator's perspective, Echo incorporates tools to facilitate the moderation and querying of the speech dataset, ensuring its quality and integrity. These tools include a mix of review processes, comprehensive statistics, and querying functionalities that empower administrators to perform detailed analyses and identify potential anomalies or inconsistencies within the dataset.

In addition to manual review capabilities, Echo leverages advanced algorithms and automated processes to support quality assurance efforts. By implementing various methods and algorithms for automatic reviewing, administrators can leverage previously trained ASR models to detect and flag potential errors or discrepancies within the dataset automatically.



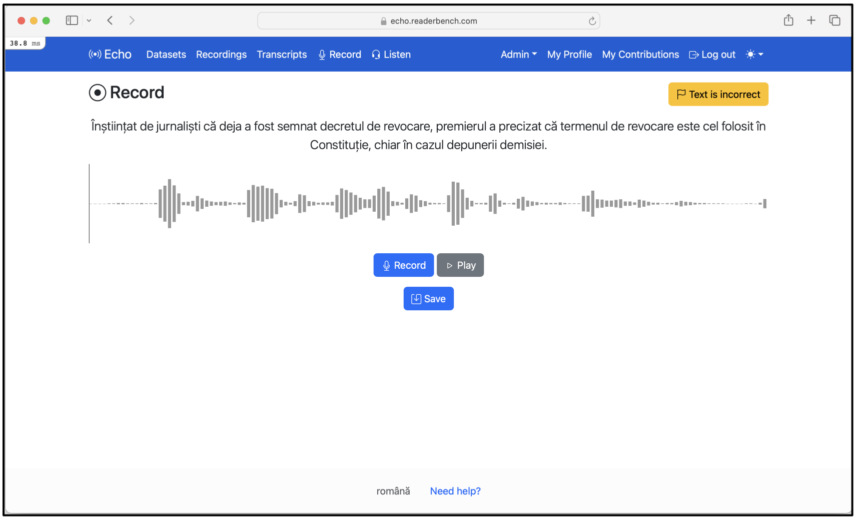
The complete Romanian speech database

Echo is a project about collecting and organizing recordings of Romanian speech and making them easily available.

Echo Collection

Datasets	14
Recordings	163,066
Total Duration	378 hours, 10 minutes and 44 seconds
Average Duration	8 seconds
Shortest Recordings	less than a second
Longest Recordings	2 minutes and 17 seconds
Unique Transcripts	15,129
Word Count	3,063,739
Unique Word Count	49,664
Speakers Count	343+

Fig.1. The crowd-sourcing platform and its welcome page



Record

Înștiințat de jurnaliști că deja a fost semnat decretul de revocare, premierul a precizat că termenul de revocare este cel folosit în Constituție, chiar în cazul depunerii demisiei.

Fig.2. Users can read transcripts and record audio files directly through the browser. Then, they can check the recording using the ”Play” button, press ”Record” again to overwrite the recording, or ”Submit” if everything is fine

3.1 Manual and Automatic Verifications

We have found that irrespective of the contributors' backgrounds and training, not all transcripts are of high quality due to various reasons, such as technical glitches, a noisy environment, transcript misreading, contributor stuttering, or losing focus.

Initially, manual verification was employed for each submitted recording. However, the volume of contributed data quickly exceeded the capacity for manual inspection. A system was implemented where three contributors manually ranked each recording as ”good” or ”bad”. Agreement among two or three contributors has been achieved only for 83% of the 17,503 verified recordings. While this approach increased the verified volume, the relatively low agreement rate required additional verification measures.

Attempts were made to utilize different statistics, such as the average duration per transcript, average character per second, and average words per second. Those statistics were computed per transcript, and outliers were flagged for further manual review. While these methods helped discover inaccurate recordings, they could only discover obvious mistakes.

Attempts were made to utilize automatic speech recognition (ASR) to filter out poor recordings. However, the ASR systems used for evaluation were trained on data that could have been faulty. Consequently, the automatic detection of subpar recordings was flawed due to the systems learning from low-quality data.

In reality, some errors are going to happen during the verification process because there is no known automatic speech recognition system with perfect accuracy and even humans have a non-zero error rate [1]. Despite the differences in methodology among these methods, all models share a common characteristic: there is a trade-off between accuracy and throughout, and as accuracy improves, throughput diminishes. Manual methods had a higher accuracy but a low throughput; automatic methods had an accuracy similar to the quality of the data but a high throughput.

In light of these considerations, we developed a comprehensive solution (see Figure 3) that leverages both human and automated methods. Initially, all recordings were transcribed using ASR systems trained on all available datasets. Recordings with a word error rate below 60% were filtered out, followed by sorting and selection based on WER scores for manual review.

To minimize errors, the ASR systems used language models and were continuously retrained.

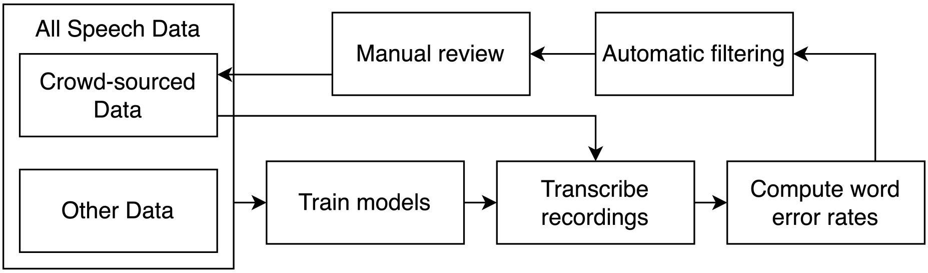


Fig.3. The main steps from the recordings review flow.

At the same time, during the manual review, we observed that recordings for certain transcripts consistently exhibited low quality. Further analysis revealed pervasive errors, including misspellings, missing words, and inadequate punctuation, leading to ambiguous intonation.

4. The Echo Speech Dataset and Speech-to-Text Models

The Echo dataset presents several opportunities for the development of innovative applications, particularly in the realm of smart learning, leveraging its unique characteristics and rich linguistic content. The skewed distribution of speakers towards young individuals renders the dataset particularly well-suited for smart learning applications, as it aligns closely with the demographic profile of students, who often comprise a significant portion of the user base for educational tools.

For instance, Philology and Sociology students engage in transcription tasks as part of their academic pursuits yet lack accessible tools tailored to their needs. Smart learning applications powered by the Echo dataset could automate transcription processes, enabling students to convert their audio notes into text effortlessly. Moreover, these tools can facilitate further analysis by segmenting transcriptions and automatically identifying speakers, thereby streamlining the organization and review of study materials.

While accuracy remains an important consideration, particularly in educational settings, the dynamic nature of the Echo dataset allows for ongoing refinement and improvement. Students can actively contribute to enhancing transcription accuracy by updating and correcting errors as needed.

Overall, the Echo dataset holds great potential for the development and evaluation of speech-based smart learning applications that empower students with innovative tools for transcription and analysis of study materials.

4.1 Automatic Speech Recognition

Training an ASR model with the Echo dataset is facilitated by its high-quality and normalized transcripts, minimizing the need for extensive data preprocessing or augmentation while simplifying the overall model training processes. The absence of significant noise or distortions in the dataset contributes to improved model performance and accuracy, allowing developers to focus their efforts on optimizing model architecture and training parameters.

In addition to its high quality and diversity, the Echo dataset offers a normalized transcript, which includes translating numerical figures into their literal format, removing punctuation, and standardizing other linguistic elements. This normalization process enhances the consistency and coherence of the dataset, reducing variability and improving the overall quality of transcriptions, thereby facilitating more accurate and reliable automatic speech recognition.

4.2 Our Dataset

Currently, the Echo Dataset contains over 350 hours of speech data from 14 different text corpora, recorded by over 300 speakers. There are over 150,000 phrases, with a unique transcript count of about 15,000 and about 50,000 unique words. The characteristics of each subdomain are presented in Table 3.

Table 3. Component datasets of Echo

Domain	Recordings	Duration (hours)	Speakers	Unique texts	Unique words
Literature	34,896	69	207	1,945	10,661
- Drama	9,077	13	198	524	2,581
- Epic	23,852	48	204	1,309	7,643
- Poems	1,967	7	168	112	1,182
Journalism	65,216	156	200	11,516	38,120
Emergency Services	8,560	11	314	168	768
Legal	8,832	28	194	482	2,903
Wikipedia	45,193	111	329	1,000	7,249
Total	162,697	378	343	15,129	49,664

We are releasing version 1.0 of the dataset, which can be accessed from HuggingFace at https://huggingface.co/datasets/readerbench/echo. The packaged dataset contains all audio files, associated transcripts, and other metadata about the speaker such as the gender and age range.

4.3 Fine-tuning Whisper

Whisper [10] is a weakly supervised speech recognition system originally trained on around 680,000 hours of labeled audio data. It was a breakthrough in the field of speech processing because it not only scales weekly supervised training to a high volume of data but also extends the model to be multilingual and multitask. The authors acknowledge that more labeled data is the key to more accurate transcriptions and that unsupervised learning cannot be used further to train generic and reliable models. They focused their efforts on constructing a very diverse dataset with recordings that have a broad set of characteristics but are more conservative on the side of transcripts. They developed multiple heuristics to detect and remove transcripts that were automatically generated in different languages. However, they decided against normalizing the transcripts, maintaining whitespace, casing, punctuation, and other stylistic aspects.

The underlying model is an encoder-decoder Transformer [14], an architecture previously validated to work well for automatic speech recognition applications as it scales with high volumes of data and is robust for weakly supervised methods.

We decided to use Whisper as a starting point in our analysis for similar reasons. On top of that, Whisper fits well within our application, and it has already been trained on a large amount of data. The Echo dataset was used to continue the training and evaluation of the ”small” Whisper model of 12 layers with a width of 768 and 12 heads for a total of 244M parameters. In terms of data processing, we used the original method: recordings of 30 seconds, resampled to 16,000 Hz, split in windows of 25 milliseconds with a 10-millisecond stride. The only difference is that we decided on a partially normalized transcript to convert numbers into their literal form.

We are also releasing a finetuned version of the Whisper model using the data presented in the previous sections. The model can be accessed from HuggingFace at https://huggingface.co/readerbench/whisper-ro.

4.4 Performance Evaluation

The metric used for evaluating the accuracy of the trained model is word error rate, a common metric used to measure the performance of speech recognition or machine translation systems. It is defined as the ratio between the count of errors (substitutions, insertions, and deletions) and the total number of words.

$$WER = \frac{S + D + I}{N} \quad (1)$$

Even though the word-error-rate as a metric is strictly defined, the input strings that are compared are sometimes normalized, resulting in different error rates in the end. In our experiments, we have normalized the strings by ignoring capitalization and removing punctuation.

5 Results

Our model was evaluated on the original datasets and the test subset of Echo, comprising 20% of the original data. The data was split between train and test subsets so that a transcript does not exist simultaneously in both subtests. The data was divided so the speaker sets were as different as possible between train and test splits. However, that is not always achievable because some speakers have contributed recordings for a large fraction of all transcripts. The results are presented in Table 4.

Table 4. Word error rates for evaluated datasets for the original small and large models and for the fine-tuned dataset using the presented Echo dataset

Name	Small	Large V1	Large V2	Large V3	Fine-tuned small
Common Voice	33.2	19.8	15.8	10.8	12.2
FLEURS	29.8	15.4	14.4	8.2	10.9
VoxPopuli	28.6	17.0	14.4	13.8	9.4
Echo	>100	>100	>100	27.2	7.3
RSC	38.6	33.4	28.5	24.9	5.4

Comparing the ”small” but fine-tuned model with the original ”large” and the newer ”large-v2” models shows a 30% lower word error rate on average for all tested datasets while having six times fewer parameters. The newest ”largev3” shows mixed results as it performs better than the fine-tuned model on the Common Voice and FLEURS dataset but much worse on VoxPopuli, Echo, and RSC.

The difference between ”large-v2” and ”large-v3” models can be attributed to the amount of data used for training, the latter being trained on 1 million hours of weakly labeled audio and 4 million hours of pseudolabeled audio, around 7 times more data in total.

Finally, these results show that a higher volume of data (356 hours of data used during the initial training plus the 378 hours of recordings of the Echo dataset) is more important than a deeper network.

One of the problems that the Whisper-based models are exhibiting is the generation of ”hallucinations.” In this context, a ”hallucination” is an output transcript unrelated to the audio it transcribes. They are usually generated during a prolonged period of silence or around the end of the audio file. We have not noticed any hallucinations during the evaluation of the fine-tuned model, and we believe this can be attributed to the shorter length and higher quality of the audio files.

6 Conclusions

The Romanian language remains an under-represented language in terms of data, available models, and practical applications with automatic speech recognition functionality. The resources and initiatives dedicated specifically to Romanian language processing are limited compared to more resourced languages. This under-representation poses challenges for advancing the development of robust models and applications tailored to the unique linguistic characteristics of the Romanian language.

We anticipate that Echo, as a platform, will support collaboration and innovation within the field of automatic speech recognition, particularly within the Romanian context. By providing free access to a growing repository of speech data, Echo can help researchers, developers, and enthusiasts dedicated to advancing the state-of-the-art in Romanian language processing. While initiatives like CommonVoice offer similar methodologies, Echo’s localized focus and growing popularity within the Romanian-speaking community hopefully positions it as an important resource for addressing the specific linguistic nuances and challenges inherent to Romanian.

We expect that the Echo dataset will attract more researchers to engage with Romanian language processing and inspire the development or adaptation of models tailored specifically for the Romanian language. By offering a rich and diverse dataset encompassing various linguistic contexts and speaker demographics, Echo provides researchers with a valuable resource for training and evaluating ASR models.

Acknowledgments. This research is partially supported by the project “Romanian Hub for Artificial Intelligence - HRIA”, Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 334906.

CRediT author statement. Remus-Dan Ungureanu: Conceptualization, Investigation, Methodology, Formal analysis, Software, Resources, Data curation, Writing – original draft preparation. **Mihai Dascalu**: Conceptualization, Methodology, Validation, Writing – review and editing, Supervision, Project administration, Funding acquisition.

References

1. Amodei D., Ananthanarayanan S., Anubhai R., Bai J., Battenberg E., Case C., Casper J., Catanzaro B., Cheng Q., Chen G., et al.: Deep speech 2: End-to-end speech recognition in english and mandarin. In: International conference on machine learning, pp. 173–182, PMLR (2016)
2. Ardila R., Branson M., Davis K., Henretty M., Kohler M., Meyer J., Morais R., Saunders L., Tyers F.M., Weber G.: Common voice: A massively-multilingual speech corpus. arXiv preprint arXiv:1912.06670 (2019)
3. Baevski A., Zhou Y., Mohamed A., Auli M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems 33, pp. 12449–12460 (2020)
4. Chen G., Chai S., Wang G., Du J., Zhang W.Q., Weng C., Su D., Povey D., Trmal J., Zhang J. et al.: Gigaspeech: An evolving, multidomain ASR corpus with 10,000 hours of transcribed audio. arXiv preprint arXiv:2106.06909 (2021)
5. Conneau A., Ma M., Khanuja S., Zhang Y., Axelrod V., Dalmia S., Riesa J., Rivera C., Bapna A.: Fleurs: Few-shot learning evaluation of universal representations of speech. In: 2022 IEEE Spoken Language Technology Workshop (SLT), pp. 798–805, IEEE (2023)
6. Defined.ai: Dutch spontaneous dialogue dataset (nd), URL https://defined. ai/datasets/dutch-spontaneous-dialogue, accessed on 14.06.2024
7. Georgescu A.L., Caranica A., Cucu H., Burileanu C.: Rodigits-a romanian connected-digits speech corpus for automatic speech and speaker recognition. University Politehnica of Bucharest Scientific Bulletin, Series C 80(3), pp. 45–62 (2018)
8. Georgescu A.L., Cucu H., Buzo A., Burileanu C.: Rsc: A romanian read speech corpus for automatic speech recognition. In: Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 6606–6612 (2020)
9. Hannun A., Case C., Casper J., Catanzaro B., Diamos G., Elsen E., Prenger R., Satheesh S., Sengupta S., Coates A., et al.: Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567 (2014)
10. Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning, pp. 28492–28518, PMLR (2023)
11. Stan A., Dinescu F., T¸iple C., Meza S¸., Orza B., Chiril˘a M., Giurgiu M.: The swara speech corpus: A large parallel romanian read speech dataset. In: 2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), pp. 1–6, IEEE (2017)
12. Ungureanu D., Badeanu M., Marica G.C., Dascalu M., Tufis D.I.: Establishing a baseline of romanian speech-to-text models. In: 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), pp. 132–138, IEEE (2021)
13. Ungureanu D., Toma S.A., Filip I.D., Mocanu B.C., Aciob˘anit,ei I., Marghescu B., Balan T., Dascalu M., Bica I., Pop F.: Odin112–aiassisted emergency services in romania. Applied Sciences 13(1), p. 639 (2023)
14. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser, Polosukhin I.: Attention is all you need. In: 31st Conference on Neural Information Processing Systems (NIPS 2017), pp. 5998–6008 (2017)

15. Wang C., Riviere M., Lee A., Wu A., Talnikar C., Haziza D., Williamson M., Pino J., Dupoux E.: Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. arXiv preprint arXiv:2101.00390 (2021)
16. Zhang Y., Park D.S., Han W., Qin J., Gulati A., Shor J., Jansen A., Xu Y., Huang Y., Wang S. et al.: Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. IEEE Journal of Selected Topics in Signal Processing 16(6), pp. 1519–1532 (2022)