applied sciences

MDPI

Article

Under-Represented Speech Dataset from Open Data: Case Study on the Romanian Language

Vasile Păiș *, Verginica Barbu Mititelu, Elena Irimia, Radu Ion and Dan Tufiș

Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, 050711 Bucharest, Romania; vergi@racai.ro (V.B.M.); elena@racai.ro (E.I.); radu@racai.ro (R.I.); tufis@racai.ro (D.T.)
* Correspondence: vasile@racai.ro

Abstract: This paper introduces the USPDATRO dataset. This is a speech dataset, in the Romanian language, constructed from open data, focusing on under-represented voice types (children, young and old people, and female voices). The paper covers the methodology behind the dataset construction, specific details regarding the dataset, and evaluation of existing Romanian Automatic Speech Recognition (ASR) systems, with different architectures. Results indicate that more under-represented speech content is needed in the training of ASR systems. Our approach can be extended to other low-resourced languages, as long as open data are available.

Keywords: speech dataset; under-represented voices; speech recognition; Romanian language

Citation: Păiș, V.; Barbu Mititelu, V.; Irimia, E.; Ion, R.; Tufiș, D. Under-Represented Speech Dataset from Open Data: Case Study on the Romanian Language. *Appl. Sci.* **2024**, *14*, 9043. https://doi.org/10.3390/app14199043

Academic Editor: Douglas O'Shaughnessy

Received: 21 August 2024
Revised: 20 September 2024
Accepted: 2 October 2024
Published: 7 October 2024

1. Introduction

State-of-the-art Automatic Speech Recognition (ASR) systems rely on large volumes of recorded speech data for each language. With projects like Common Voice (https://commonvoice.mozilla.org/en, accessed on 3 October 2024) [1], as well as other national and international projects, the volume of speech data available for research purposes has increased over the years. However, for many languages, only certain types of voices are recorded. For example, the Romanian segment of Common Voice contains mostly male speakers in their twenties or thirties. Other voice types are either scarcely represented (for example, out of the 17,738 validated samples in the Romanian Common Voice, 17% are feminine voices and only 0.6% are feminine voices in their thirties) or not represented (there are no samples for women in their fifties or above).

In the context of the project "Underrepresented speech dataset from open data: case study on the Romanian language" (USPDATRO) (https://www.racai.ro/p/uspdatro/, accessed on 3 October 2024), we created a new speech dataset for the Romanian language, specifically aimed at under-represented voices. For this purpose, we investigate openly available data (under a Creative Commons license) from online multimedia platforms. We show that platforms such as YouTube and Vimeo contain open data that can be exploited for the purposes of building a speech dataset useful for research purposes. Although showcased only for Romanian, our methodology can be extended to other under-resourced languages or under-represented voice types for which openly licensed content is available.

The rest of this paper is organized as follows: Section 2 presents related work, Section 3 briefly describes the methodology used, Section 4 reports the challenges in the USPDATRO project, Section 5 presents the dataset with associated statistics, and Section 6 gives the evaluation results of existing ASR systems on the new dataset. Finally, we conclude in Section 7. The dataset is publicly released and is freely available for research as mentioned in Section 5.

2. Related Work

According to the European Language Equality (https://european-language-equality.eu/, accessed on 3 October 2024) report on the Romanian language resources and technologies

(https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_29__Language_Report_Romanian_.pdf, accessed on 3 October 2024) [2], speech processing is a domain with fragmentary support, just like corpora development. Although speech corpora is not a separate category in this report, the multimodal corpora (text and speech) are relevant for this type of resource, and this is also evaluated as having fragmentary support.

Georgescu et al. [3] provide an inventory of Romanian speech corpora, which shows that they vary in size (from less than an hour to 280 h), they are mostly non-public, and they contain read or/and spontaneous speech, from various sources. However, the voices represented by these corpora are not described in terms of speaker's gender, age, country, or country region where they live. In what follows, we analyze several Romanian speech corpora, primarily those for which information about the speaker's gender and age is available. Details about their size is given in Table 1. One of the largest speech datasets, RSC [3], also openly available, has 164 voices, with the average age of 24 years (81% of them belonging to the 21–25 age group), male voices being twice as prevalent as female ones. Another corpus for which information about the speakers is available is ROBIN Technical Acquisition Speech Corpus RTASC [4]: it is balanced with respect to gender representation (three male and three female voices, each with the same amount of speech) but imbalanced with respect to age groups: 50% contains voices of speakers in their forties (two female and one male voice), and 33% contains voices of speakers in their thirties (with equal distribution of male and female voices), while the rest contain the voice of a single male speaker in his twenties. The voices in the RoDigits corpus [5] belong to people aged between 20 and 45, with 23 being the average age, while male voices are about 30% more prevalent than female ones. The speakers who volunteered for recording the SWARA corpus [6] are eight males and nine females, with ages between 20 and 35. The RSS corpus [7] contains only one female voice of a woman in her twenties. What we notice is that most of the voices represented are male, while the average age of speakers is in the range 20–35.

Table 1. Public Romanian speech corpora statistics.

Corpus	# Hours	# Utterances	# Speakers
RSC	100	136.1 k	164
RoDigits	37.5	15.4 k	154
SWARA	21	19 k	17
RO-GRID	6.6	4.8 k	12
RSS	5.5	5.7 k	3
RASC	4.8	3 k	-
RTASC	6.5	3.8 k	6
CV	9	8k	130
VoxPopuli	83	27 k	164
MaSS	23	8.1 k	1
FLEURS	12	-	-

The RO-GRID [8] dataset contains readings of sequences of six words chosen from a list of alternatives. The first three words were designated as "keywords" and the speaker had to utter all combinations (400 in total). The last three words were designated as "fillers" and were randomly chosen while creating the sentence. The final corpus contained 6.6 h of audio from 12 speakers. The Romanian Anonymous Speech Corpus (RASC) [9] is a crowd-sourced dataset, gathered through an open interactive platform. The corpus currently contains 4.8 h of transcribed audio.

The Common Voice (CV) [1] corpus is a massively multilingual dataset of transcribed speech that continues to grow as more volunteers record their voice. The Romanian version contains 9 h of transcribed audio (6 h validated) recorded by 130 speakers, using sentences from the Romanian Wikipedia. VoxPopuli [10] is another multilingual corpus that contains 100,000 h of raw audios in 23 languages and 1800 h of transcribed speech in 16 languages. The Romanian language subset contains 4500 h of unlabeled speech and 83 h of transcribed

audio. The Multilingual corpus of Sentence-aligned Spoken utterances (MaSS) [11] is a speech dataset based on readings of the Bible. The dataset contains 8130 of parallel spoken utterances in eight languages, with 23 h of Romanian language. The Few-shot Learning Evaluation of Universal Representations of Speech (FLEURS) [12] is a parallel speech dataset in 102 languages, with approximately 12 h of speech supervision per language. Detailed statistics regarding the Romanian language are not available in the official dataset paper. The Representative Corpus of Contemporary Romanian Language (CoRoLa) [13] contains speech data, aligned with written text. This is created by integrating curated versions of different corpora already covered. No information is available on the age or gender of the speakers, nor on the number of voices in the oral component of the corpus. For other national corpora, different approaches were followed with respect to their spoken/oral component, when it existed. For example, for the British National Corpus the speakers were selected so as to ensure a balanced distribution of men and women of each age group and social grouping, as well as including diverse social contexts [14]. The oral component of the Czech National Corpus was also designed with concern for an equal distribution of major sociolinguistic categories, such as gender, age group, education, and region of origin [15].

In the case of other languages, Garnerin et al. [16] investigate gender representation in French broadcast corpora, on data recorded between 1998 and 2013, and its impact on ASR performance. The authors observe a lower WER score for women compared to men and they attribute this to the difference of women voices presence in training speech corpora. Tatman and Kasten [17] investigate the accuracy of ASR systems across gender, race and four dialects of American English. Their findings show that the best WER scores are achieved for general American talkers, and among ethnicities, for white talkers. Ngueajio and Washington [18] further acknowledge that ASR systems do not work equally well for everyone and actually hinder the productivity of some users. The authors also mention the increased diversification of a training set as a method for reducing ASR bias. Nevertheless, high-quality data collection is costly. When new data are added to a corpus, the data should be collected from the categories of speakers that speech recognition will benefit [19]. A Digital Language Equality Metric [20] is proposed to account for different types of resources and tools (including speech and speech processing) available for different languages.

Meyer et al. [21] introduce the Artie Bias corpus, an English curated subset of the Mozilla Common Voice corpus, with demographic indication for age, gender, and accent. It is intended to be an evaluation dataset for detecting demographic bias in ASR systems. Navarro et al. [22] propose a data augmentation technique via pairwise mixup across subgroups by adding new samples of under-represented groups to improve group fairness when training ASR systems.

3. Methodology

The dataset collection process involved three main steps: (a) identification of multimedia platforms that offer open content in the Romanian language; (b) collecting samples of such content from the identified platforms; and (c) manual annotation and transcription of the retrieved samples. First, well-known platforms like YouTube (https://www.youtube.com/), Vimeo (https://vimeo.com/), TikTok (https://www.tiktok.com), SoundCloud (https://soundcloud.com), and LinguaLibre (https://lingualibre.org/wiki/LinguaLibre:Main_Page) are checked for a number of features that are essential for our purposes: the availability of open licenses; search features (based on language, selected license, and features allowing the identification of under-represented speech types); and the actual availability of content that is of interest to our project in terms of language, voice type, recording length, recording quality, etc.

The very popular multimedia sharing platform YouTube uses, among others, the Creative Commons license as a standard way for content creators to grant permission to use their work. The user can select this license explicitly; otherwise, the default YouTube license is in force. The Creative Commons license is available as a filter of the keyword search. The platform explicitly mentions that creators should upload only video content

they made or they are allowed to use. This makes the users responsible for the content provided. The search interface does not allow language-based search or filtering by speaker characteristics (like gender or age, which must be inferred from the description, profile or video content) but provides a duration filter with three options: "Under 4 min", "4–20 min", and "Over 20 min". The feature is useful since we are interested in collecting recordings primarily in the range 4–20 min, to ensure voice diversity in the dataset. Otherwise, this cannot be ensured through long material covering a few voices.

Vimeo has similar features, offering Creative Commons licenses and filtering of the content (for license and duration) but still not filtering for the speaker's characteristics. Compared to YouTube, Vimeo has much less Romanian content, a small number of results with permissive licenses, many results with an unspecified license, and a majority of videos of long duration, which would unbalance the data if used. TikTok's maximum limit duration of 10 min initially made the platform a promising source of content, but further investigation led to discouragement due to the unclear terms of service (under a unique in-house license) and limited filtering options.

SoundCloud is a platform that hosts high-quality material, mostly free of noise, with the exception of radio shows which may contain a faint musical background. The Romanian content is mostly recorded by either young or middle-aged people, with no obvious gender predominance. The platform offers Creative Common license material but no filtering by usage license. The tracks are usually long (over an hour, with the exception of short stories), making it very difficult to find short samples of spontaneous speech.

LinguaLibre is a project dedicated to building a multilingual multimedia corpus under free license, offering to the user the possibility to record their own voice and store it together with detailed metadata like the language or gender of the speaker. The interface then provides filters based on all metadata. Unfortunately, the Romanian language is poorly represented at the moment, but we intend to monitor the evolution of the content on the platform since it is, by design, suitable for our purposes.

The next step is to use the acquired knowledge to harvest the platforms and download multimedia content. We focus only on the three platforms that offer the most in terms of the discussed criteria—YouTube, Vimeo and SoundClound—and most of the corpus is actually gathered from YouTube. Under-represented Romanian language speakers (old or very young, and female) are targeted by means of search expressions since the platforms do not provide filtering options on these criteria. See examples of search words and phrases in Table 2. In this case, the category "young people" is aimed at age groups 14–19 and 19–29, while "older people" is aimed at age groups 50–70 and over 70. The speech content and style highly vary on these platforms, with many recordings being spontaneous and informal. We do not perform a pre-selection of the content based on style characteristics when searching for content. However, the collected samples are later classified by spontaneity and quality, as indicated in Section 5 and Table 3. In terms of domain distribution, the data are found to be heterogeneous, with many different domain categories such as motivational (for children and adolescents), childhood and personal memories, education, pedagogy, technology, activism, feminism, psychology, mathematics, literature, poetry reading, storytelling, anthropology, medicine, relationships, shopping, and news. The number of views of the videos on their respective platforms also varies substantially, from 6 to over 150,000 views. No specific filtering is considered based on domain or number of views.

The content is downloaded as .mp4 video files with a low video resolution (to reduce space requirements), using an online downloader application (https://en.savefrom.net/1-youtube-video-downloader-463/, accessed on 3 October 2024). For each downloaded material, information about the source platform, URL, license, and date of download is recorded, and a screenshot of the content as appearing in the platform at the download date is kept as proof for future potential claims denying the open license availability of the content. Metadata associated with the multimedia files are recorded in the downloading phase, to allow checking the conformity of the collected samples to our project goals.

The specific fields include information about the annotator, platform, URL, duration, license, speech type (read or spontaneous), quality, and speaker-related information (gender and age).

Table 2. Proposed search words and phrases.

Keywords	Target Group
elevii te învață ("the students teach you") probleme adolescenți ("problems teenagers")	Young people
sfaturi duhovnicești ("spiritual advice") viața la pensie ("life when retired")	Older people
emisiune pentru femei ("women show") feminism și literatură ("feminism and literature")	Women
editura ("publishing house") antropologie ("anthropology")	Generic

Annotations and, afterwards, transcriptions are performed by 4 native Romanian language speakers, expert annotators, who worked previously on the creation of other speech corpora. An annotation guide is created (https://www.racai.ro/p/uspdatro/resources/USPDATRO_Annotation_Guidelines_v1.0.pdf, accessed on 3 October 2024) and annotators are given specific instructions on how to complete the different annotations. Information about the age of the speaker at the moment of the recording is collected in different manners: from explicit mentions in the recording; from other sources on the web like social media profiles or personal websites; and deduced from the recording, e.g., a child in a kindergarten is obviously in the "under 14" category. The purpose is not to encode the exact age of the speaker (which we consider personal information) but to map the speakers as closely as possible to the appropriate age category. We consider 6 age categories (under 14, 14–19, 19–29, 30–50, 50–70, and over 70). As, most of the time, the exact age at the time of recording cannot be identified, the annotators are instructed to include the person in the most appropriate age group (for example, 14–19 is typically high school, while 19–29 corresponds to faculty students, PhD students, or young working individuals). The age and gender annotations are given for each speaking individual. At this stage, as the annotators process the files, recordings containing large amounts of non-Romanian speech are rejected, and suitable alternatives are downloaded. If a recording includes small amounts of non-Romanian speech, only information about Romanian speakers is extracted. Later at the transcription stage, non-Romanian speech is ignored.

The information about the speech being read or spontaneous is either inferred from the video (the speaker is seen reading from some written material or is clearly being interviewed on the street or in some other place, thus being spontaneous speech) or by common sense (very young kindergarten children cannot read; TV news presenters usually read from a teleprompter device). The type of speech annotation is available at the level of the entire video file. Annotators are instructed to avoid clips with mixed speech types.

For assessing the quality of speech, the annotators use the Mean Opinion Score (MOS) method [23]. Speech is rated on a scale from 1 to 5, where 1 is considered bad speech, with a high level of distortion, and 5 is excellent speech, with an imperceptible level of distortion. Annotators are instructed to discard any recordings that they would rate below a MOS value of 3. This is considered fair speech quality but with a level of distortion that is perceptible and slightly annoying. Thus, recordings with levels 1 or 2 are not present in the dataset. Finding spontaneous MOS 5 recordings is difficult due to the video blogging trend of adding soundtrack to the recording.

The actual transcription of the audio tracks is performed manually by the human annotators using the Subtitle Edit application, and transcriptions are saved in CSV format files with the same ID as the transcribed video file. In order to ensure the high quality of the transcripts, no automated technologies are used at any point of the work. In the transcription process, manual subtitle segmentation is carried out at the sentence level,

using the rendering of the waveform that Subtitle Edit tool makes available as an indication for the segmentation locations (while the annotators are free to segment at any point they considered relevant). From the resulting segments, 5% of them are transcribed by a second annotator and the corresponding video files are re-annotated. No discrepancies are found in the annotations. We consider this to be due to the clear annotation instructions as well as the experience of the annotators who previously worked on creating other audio corpora.

4. Challenges

For certain videos, the transcription process poses several challenges. English words are sometimes used in conversation, either as short code-switching occurrences or as technical terms commonly used untranslated in conversations. These are transcribed phonetically. Romanian spelling is mostly phonetic (words are pronounced as they are spelled); thus, our approach is in line with this rule. For example, *Aș fi făcut ceva , **like** să merg unde trebuia* (I would have done something, **like** to go where I was supposed to). In this case, the English word *like* is transcribed as *laic*, corresponding to its pronunciation.

Spontaneous speech recordings sometimes contain overlapping voices. Whenever possible, the issue is solved by performing in-depth segmentation to separate the voices. This results in a number of 100 segments with a duration of less than 1 s. When separation of the voices is impossible, if the overlapping portion concerns a short fragment of recording, that segment is left untranscribed, while longer fragments of recording that represent clusters of overlapping regions are ignored, even if they contain non-overlapping parts. Furthermore, still in the case of spontaneous speech, for certain words, it is difficult to discern exactly what sounds the speaker pronounced. In many such situations, reducing the play rate to 40–50% is necessary to identify the uttered phonemes in words that are, most of the time, recognizable even with the missing sounds. This is particularly relevant for old people (over 70) and for recordings with MOS score 3.

For some speakers, the segmentation at sentence boundary is very difficult because of their tendency to systematically both pause in the middle of the sentence and not pause at the end of the sentence; in these cases, the segmentation is performed according to the speaker's pausing pattern. Furthermore, a flattening of the visualized waveform appears in situations where loud sounds (music and falling objects) are present somewhere in the recording because of the relative rendering of the sound intensity for the spoken parts; this phenomenon makes the waveform less useful as a cue for segmentation.

5. Dataset

The overall duration of the dataset is 4 h 18 m 55 s. This represents the usable speech for training or evaluating ASR systems. We analyze a total of 5 h 23 m 21 s of audio; however, many files also contain segments that are not useful (begin or end music, various sounds, silence, etc.). Thus, our findings suggest that only 80% of the available multimedia content may be useful for speech applications (with this percent varying with each multimedia file). The audio files are stored in WAV format, using 16-bit Signed Integer PCM encoding, 16-bit precision, single channel, with a sample rate of 16 KHz. Each audio file contains a voice segment, extracted from a larger media file, and has an associated TXT file with the corresponding transcription. Conversion from the original MP4 format to WAV is realized using the FFMPEG (https://www.ffmpeg.org/) application. The corpus metadata (in CSV format) contain detailed information about the source of each segment, including the platform, license, overall duration, start and end times of the segment, and speaker characteristics. The original source URL is included in order to allow users to download the original media and convert it into other formats if needed.

Detailed statistics are given in Table 3, with a graphical representation in Figure 1, and a speaker breakdown considering both gender and age is given in Table 4, with a percentage overview in Figure 2. As described in Section 3, the platform that is the richest in the type of multimedia material we are interested in is YouTube, and this is reflected in the amount of data distribution according to platforms: 83% of the content comes from YouTube, 12%

from SoundCloud, and 5% from Vimeo. The dataset covers primarily under-represented speech groups (outside the 19–29 male category), while some media files contain small portions of the common group represented by male voices with an age between 19 and 29 years (in these cases, the person has, most of the time, a supportive role in the recording and has generally few and short interventions). High-quality speech (MOS 5 or 4), easily usable for speech recognition systems, represents 99% of the data, with MOS 5 accounting for 55% of the entire corpus. The corpus contains primarily content licensed under the Creative Commons Attribution license, usable for both commercial and non-commercial applications. However, given our interest in research applications, we also include content available under non-commercial open licenses (CC Attribution Non-Commercial and CC Attribution Non-Commercial Share Alike). The license is indicated alongside each file in the corpus metadata. The majority of the content is spontaneous speech (as we find it more useful for improving ASR tools), while 33 min is read speech.

Table 3. Dataset statistics.

Indicator	Category	Duration	# Segments	Avg. Seg. Duration (s)
Gender	F	2 h 8 m 42 s	1506	5.13
	M	2 h 10 m 13 s	1131	6.91
Age	<14	10 m 44 s	175	3.68
	14–19	15 m 20 s	168	5.48
	19–29	1 h 5 m 34 s	676	5.82
	30–50	45 m 41 s	457	6.00
	50–70	1 h 1 m 22 s	674	5.46
	>70	1 h 0 m 14 s	487	7.42
MOS	5	2 h 20 m 34 s	1187	7.11
	4	1 h 56 m 44s	1435	4.88
	3	1 m 37 s	15	6.47
Platform	YouTube	3 h 32 m 48 s	2014	6.34
	Vimeo	13 m 53 s	194	4.29
	SoundCloud	32 m 14 s	429	4.51
License	CC BY	3 h 44 m 9 s	2169	6.2
	CC BY NC	2 m 32 s	39	3.9
	CC BY NC SA	32 m 14 s	429	4.51
Type	Read	33 m 32 s	467	4.31
	Spontaneous	3 h 45 m 23 s	2170	6.23

Table 4. Speaker breakdown by gender and age.

	F			M		
Age	Duration	# Segments	Avg. Durat.	Duration	# Segments	Avg. Duration
<14	1 m 38 s	27	3.63	9 m 06 s	148	3.69
14–19	-	-	-	15 m 20 s	168	5.44
19–29	54 m	557	5.82	11 m 34 s	119	5.83
30–50	33 m 51 s	352	5.77	11 m 51 s	105	6.77
50–70	26 m 26 s	431	3.68	34 m 56 s	243	8.63
>70	12 m 48 s	139	5.53	47 m 26 s	348	8.18

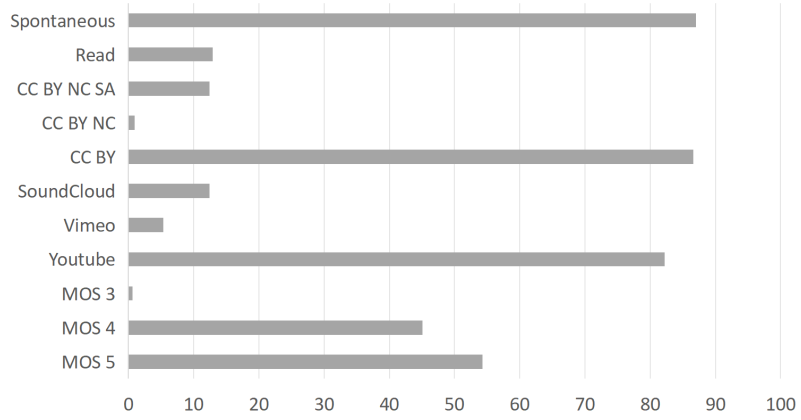

Figure 1. Dataset statistics(as duration percentage of the entire dataset).

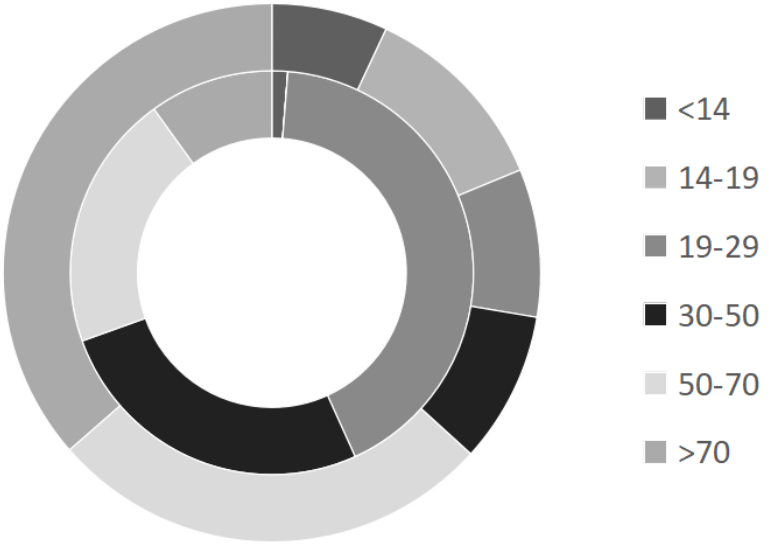

Figure 2. Percentage of ages represented in the corpus, by gender M (outer doughnut), and F (inner doughnut).

The text files holding the transcriptions are UTF-8 encoded, with the appropriate Romanian characters. Punctuation is available as entered by the dataset authors in the transcription process. The files are further annotated for lemma, part-of-speech, and dependency parsing, using UDPipe 1.3 [24], integrated in the RELATE platform (https://relate.racai.ro) [25–27], with a recent model [28]. We use the RELATE-integrated UDPipe text processing engine with the latest model because it strikes a good balance between the annotation accuracy and processing speed. As we show in [28], UDPipe is competitive (see Table 3 of the cited paper) with other text processing engines that are trained for Romanian (such as NLP-Cube [29], RNNTagger [30], and Stanza [31]), specifically in terms of accuracy when processing out-of-domain corpora, as our dataset does, when compared to the training data of UDPipe. The annotations are available in separate CoNLL-U Plus (https://universaldependencies.org/ext-format.html, accessed on 3 October 2024) files, associated with the raw text files. Statistics on the text part are computed in the RELATE platform and are given in Table 5. Transcriptions are segmented in locations natural from the speech point of view. This results in text files containing multiple sentences (an average of 2.5 sentences in each file, with an average sentence length of 7.3 tokens). A large number of words appear only once in the corpus (hapax legomena), indicating the need for an externally trained language model if the ASR system makes use of this feature. With regard to part of speech tags, common nouns are the most represented, followed by verbs.

A number of proper nouns are also included, potentially making the recognition more challenging in the case of systems employing pre-trained language models.

Table 5. Characteristics of the text part of the dataset.

Indicator	Value	Indicator	Value
Text files	2637	UPOS Noun	8471
Sentences	6652	UPOS Verb	5793
Tokens	48,530	UPOS Adp	4009
Unique tokens	8221	UPOS Adv	3717
Unique lemmas	5509	UPOS Adj	1952
Hapax legomena	5055	UPOS Num	615
Avg. Sentence Length	7.30	UPOS PropN	851

The dataset is publicly released on the Zenodo platform (https://doi.org/10.5281/zenodo.7898232). It is available under an open license, Creative Commons Attribution Non Commercial Share Alike (CC BY NC SA). The original content is available under its own license, as indicated in the metadata. The dataset is further indexed in the European Language Grid [32] catalogue (https://live.european-language-grid.eu/catalogue/corpus/21567, accessed on 3 October 2024).

6. Evaluation

The recognition performance of several existing Romanian ASR systems is evaluated against the newly created dataset. RO-DS2 [33] and RO-DS2-ROBIN [34] are based on the DeepSpeech2 [35] architecture, with RO-DS2-ROBIN being trained with additional data and using an improved language model. RO-WAV2VEC2 [36] is based on the Wav2Vec2 architecture [37]. RO-Whisper [38] is based on the OpenAI Whisper [39] architecture. All the systems are fine-tuned on publicly available Romanian language speech data which do not cover the under-represented categories from the USPDATRO dataset, since such data were not available at the time of training of the models (for the purpose of this work, we do not retrain any models). Evaluating the performance of state-of-the-art pre-trained models on under-represented speech is relevant in order to understand the necessity of including more such data in the training of new models or in the fine-tuning process. If the model's performance on the USPDATRO dataset is significantly lower compared to regular data, this is a strong indication that a larger under-represented speech dataset is needed to be included in the model's training. Results are given in Table 6. The baseline values are the best scores reported in the corresponding system papers.

Table 6. Evaluation of several Romanian language ASR systems.

	Baseline		USPDATRO	
System	WER	CER	WER	CER
RO-DS2	0.0991	0.0280	0.5714	0.3638
RO-DS2-ROBIN	0.0991	-	0.6491	0.3381
RO-WAV2VEC2	0.1393	0.0983	0.9115	0.6675
RO-Whisper medium	0.1379	-	**0.2800**	**0.1319**
RO-Whisper large-v2	0.1261	-	0.4330	0.2875

The baseline results presented in Table 6 are computed on slightly different combinations of Romanian language corpora, which are described in Section 2. The Whisper baselines are an average of the results given on individual corpora by the authors. This accounts for the slight decrease in performance. Nevertheless, all models give significantly worse results on the USPDATRO dataset. Even for the best-performing system on USPDATRO (the RO-Whisper medium), the WER is double compared to the baseline. Unexpectedly, the medium variant of the Whisper model performs better compared to

the large-v2 version. This is likely due to the few-shot training scenario employed for the model, where the total number of training epochs is limited to 10. Further training is expected to improve over the large-v2 results.

We further expand on the work by [38] and evaluate Whisper models (without fine-tuning) on the test part of the Representative Corpus of Contemporary Romanian Language (CoRoLa) corpus and on the USPDATRO corpus. Results are given in Table 7. For each model, we report results with and without beam search. When the beam search algorithm is activated, a beam with size 32 is used. As expected, the error of the prediction decreases with the increase in model size. Thus, the best performing model on both CoRoLa and USPDATRO is the *large-v2* variant of the model. Furthermore, the use of the beam search algorithm has an impact on the predictions, improving performance. However, even the best results on the USPDATRO dataset are below the CoRoLa-based results by 8% WER. This indicates the need to include under-represented speech during model training or fine-tuning.

Table 7. Evaluation of Whisper models without Romanian language fine-tuning.

			CoRoLa		USPDATRO	
Model	**Param**	**Beam**	**WER**	**CER**	**WER**	**CER**
tiny	39M	N	1.2218	0.8135	1.1502	0.6169
tiny		Y	0.7903	0.3439	0.9115	0.4771
base	74M	N	0.6079	0.2534	0.9347	0.5086
base		Y	0.6433	0.2625	0.7275	0.3391
small	244M	N	0.5027	0.2144	0.6789	0.3169
small		Y	0.5005	0.2143	0.5794	0.2380
medium	769M	N	0.5562	0.2768	0.5566	0.2516
medium		Y	0.4347	0.1887	0.5051	0.2064
large-v2	1550M	N	0.4561	0.2142	0.5104	0.2189
large-v2		Y	**0.4052**	**0.1777**	**0.4874**	**0.1952**

7. Conclusions

The word error rate (WER) of current Romanian ASR systems is high on the USPDATRO dataset, which further means that such datasets need to be developed in order to bring the ASR WER down for under-represented categories of speakers. However, it is clear that for whatever language is in focus, under-represented categories of speech data need to be added to training datasets. With 4 h, 18 min and 55 s of speech in total but with an even split between men and women and with significantly more speech outside the 19–29 years of age category, the USPDATRO dataset is an example of a speech dataset specifically targeted at under-represented categories of speakers.

The precise composition of the USPDATRO dataset was created by manually assembling short samples of speech that were available through online audio/video platforms (such as YouTube or Vimeo) that offered filtering by license. We were thus able to create a freely accessible speech corpus, containing speech samples from men and women, either very young or middle-aged or older. The corpus creation technique described in this paper can be easily reproduced for any language of interest, owing to the fact that speech data with open access licenses is available online. Besides investigating a new source of speech data, the USPDATRO project aimed to contribute towards increasing the Digital Language Equality Metric with regard to the Romanian language, by providing a new resource as well as an opportunity for building even larger speech resources.

Collection of the USPDATRO dataset focused only on certain types of under-represented Romanian language speech. Future research may involve extending the dataset with content and enhancing it based on other characteristics of under-represented voices, such as dialects or non-Romanians speaking the Romanian language. Our approach can prove useful for creating speech datasets for other low-resourced languages. Furthermore, even though our work focused on the speech recognition task (ASR), gathering more under-represented speech may enhance the capabilities of text to speech (TTS) synthesis models

for such voices. While exploiting social media platforms, such as YouTube, other platform-specific criteria can be used for selecting or categorizing the content of future corpora, such as audience, engagement, and categories.

CoRoLa is the open-access national reference corpus of contemporary Romanian and includes both textual and oral data. While the textual part is highly used in many national and international projects, the oral part is under-exploited, and our explanation is that unlike the textual part which is classified and documented along multiple parameters, the oral part is stored with limited metadata. The work reported in this paper aimed at remedying this deficiency. The current oral data and all its multivalued metadata will be added to CoRoLa. The collecting of diversified speech data will continue, as there is still a need for special groups of speakers with respect to age, speech disabilities, and regional dialects. We hope that, in this way, the oral part of CoRoLa will become as popular (if not more so) as the textual part.

Author Contributions: Conceptualization, V.P.; methodology, V.P., E.I., V.B.M., R.I. and D.T.; software, V.P.; validation, V.P., V.B.M. and E.I.; formal analysis, V.P., E.I., V.B.M., R.I. and D.T.; investigation, V.P., E.I., V.B.M., R.I. and D.T.; resources, V.P., V.B.M., R.I. and E.I.; data curation, V.P., V.B.M., E.I., R.I. and D.T.; writing—original draft preparation, V.P., V.B.M., E.I., R.I. and D.T.; writing—review and editing, V.P., V.B.M., E.I., R.I. and D.T.; visualization, V.P. and E.I.; supervision, V.P.; project administration, V.P.; funding acquisition, V.P. All authors have read and agreed to the published version of the manuscript.

Funding: The USPDATRO project was funded by the European Language Equality 2 project (ELE 2), which has received funding from the European Union under the grant agreement no. LC-01884166-101075356 (ELE 2).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in Zenodo at https://zenodo.org/doi/10.5281/zenodo.7898232, reference number 7898233.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASR	Automatic Speech Recognition
CoRoLa	The Representative Corpus of Contemporary Romanian Language
CC BY	Creative Commons Attribution
CC BY NC	Creative Commons Attribution Non-Commercial
CC BY NC SA	Creative Commons Attribution Non-Commercial Share Alike
CV	Common Voice
MaSS	Multilingual corpus of Sentence-aligned Spoken utterances
MOS	Mean Opinion Score
RASC	Romanian Anonymous Speech Corpus
RTASC	Romanian Technical Acquisition Speech Corpus

References

1. 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. In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), Marseille, France, 11–16 May 2020; pp. 4211–4215.
2. Păiș, V.; Tufiș, D., Language Report Romanian. In *European Language Equality: A Strategic Agenda for Digital Language Equality*; Springer International Publishing: Cham, Switzerland, 2023; pp. 199–202. [CrossRef]
3. 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, Marseille, France, 11–16 May 2020; pp. 6606–6612.
4. Păiş, V.; Ion, R.; Avram, A.M.; Irimia, E.; Mititelu, V.B.; Mitrofan, M. Human-Machine Interaction Speech Corpus from the ROBIN project. In Proceedings of the 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, 13–15 October 2021; pp. 91–96. [CrossRef]

5. Georgescu, A.; Caranica, A.; Cucu, H.; Burileanu, C. Rodigits—A Romanian Connected-Digits Speech Corpus For Automatic Speech And Speaker Recognition. *Univ. Politeh. Buchar. Sci. Bull. Ser. C* **2018**, *80*, 45–62.
6. Stan, A.; Dinescu, F.; Ţiple, C.; Meza, S.; Orza, B.; Chirilă, M.; Giurgiu, M. The SWARA speech corpus: A large parallel Romanian read speech dataset. In Proceedings of the 9th International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, 6–9 July 2017; pp. 1–6.
7. Stan, A.; Yamagishi, J.; King, S.; Aylett, M. The Romanian Speech Synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate. *Speech Commun.* **2011**, *53*, 442–450. [CrossRef]
8. Kabir, A.; Giurgiu, M. A Romanian corpus for speech perception and automatic speech recognition. In Proceedings of the 10th WSEAS International Conference on Electronics, Hardware, Wireless and Optical Communications, and 10th WSEAS International Conference on Signal Processing, Robotics and Automation, and 3rd WSEAS International Conference on Nanotechnology, and 2nd WSEAS International Conference on Plasma-Fusion-Nuclear Physics, Cambridge UK, 20–22 February 2011; pp. 323–326.
9. Dumitrescu, S.D.; Boroș, T.; Ion, R. Crowd-sourced, automatic speech-corpora collection—Building the Romanian Anonymous Speech Corpus. In Proceedings of the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL2014), Reykjavik, Iceland, 26 May 2014; pp. 90–94.
10. 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. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Virtual Conference, 1–6 August 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 993–1003. [CrossRef]
11. Zanon Boito, M.; Havard, W.; Garnerin, M.; Le Ferrand, É.; Besacier, L. MaSS: A Large and Clean Multilingual Corpus of Sentence-aligned Spoken Utterances Extracted from the Bible. In Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 6486–6493.
12. 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 Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, 9–12 January 2023; pp. 798–805. [CrossRef]
13. Tufiș, D.; Mititelu, V.B.; Irimia, E.; Păiș, V.; Ion, R.; Diewald, N.; Mitrofan, M.; Onofrei, M. Little strokes fell great oaks. Creating CoRoLa, the reference corpus of contemporary Romanian. *Rev. Roum. Linguist.* **2019**, *64*, 227 – 240.
14. Love, R.; McEnery, T. The Spoken British National Corpus 2014: Design, compilation and analysis. Ph.D. Thesis, Lancaster University, Lancaster, UK, 2018.
15. Waclawičová, M.; Křen, M.; Válková, L. Balanced corpus of informal spoken Czech: Compilation, design and findings. In Proceedings of the Interspeech 2009, Brighton, UK, 6–10 September 2009; pp. 1819–1822. [CrossRef]
16. Garnerin, M.; Rossato, S.; Besacier, L. Gender Representation in French Broadcast Corpora and Its Impact on ASR Performance. In Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery, AI4TV '19, Nice France , 21 October 2019; pp. 3–9. [CrossRef]
17. Tatman, R.; Kasten, C. Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 934–938. [CrossRef]
18. Ngueajio, M.K.; Washington, G. Hey ASR System! Why Aren't You More Inclusive? In Proceedings of the HCI International 2022—Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence, Virtual Conference, 26 June–1 July 2022; Springer: Cham, Switzerland, 2022; pp. 421–440.
19. Doğruöz, A.S.; Sitaram, S. Language Technologies for Low Resource Languages: Sociolinguistic and Multilingual Insights. In Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, Marseille, France, 24–25 June 2022; pp. 92–97.
20. Gaspari, F.; Gallagher, O.; Rehm, G.; Giagkou, M.; Piperidis, S.; Dunne, J.; Way, A. Introducing the Digital Language Equality Metric: Technological Factors. In Proceedings of the Workshop Towards Digital Language Equality (TDLE 2022; co-located with LREC 2022), Marseille, France, 20–25 June 2022; pp. 1–12.
21. Meyer, J.; Rauchenstein, L.; Eisenberg, J.D.; Howell, N. Artie Bias Corpus: An Open Dataset for Detecting Demographic Bias in Speech Applications. In Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 6462–6468.
22. Navarro, M.; Little, C.; Allen, G.I.; Segarra, S. Data Augmentation via Subgroup Mixup for Improving Fairness. In Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 7350–7354. [CrossRef]
23. Loizou, P.C., Speech Quality Assessment. In *Multimedia Analysis, Processing and Communications*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 623–654. [CrossRef]
24. Straka, M.; Straková, J. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, BC, Canada, 3–4 August 2017; pp. 88–99. [CrossRef]
25. Păiș, V.; Ion, R.; Tufiș, D. A Processing Platform Relating Data and Tools for Romanian Language. In Proceedings of the 1st International Workshop on Language Technology Platforms, Marseille, France, 11–16 May 2020; pp. 81–88.

26. Păiș, V.; Tufiș, D.; Ion, R. Integration of Romanian NLP tools into the RELATE platform. In Proceedings of the International Conference on Linguistic Resources and Tools for Natural Language Processing, Cluj-Napoca, Romania, 18–20 November 2019; pp. 181–192.
27. Păiș, V. Multiple annotation pipelines inside the RELATE platform. In Proceedings of the 15th International Conference on Linguistic Resources and Tools for Natural Language Processing, Virtual Conference, 14–16 December 2020; pp. 65–75.
28. Păiş, V.; Ion, R.; Avram, A.M.; Mitrofan, M.; Tufiș, D. In-depth evaluation of Romanian natural language processing pipelines. *Rom. J. Inf. Sci. Technol. (ROMJIST)* **2021**, *24*, 384–401.
29. Boros, T.; Dumitrescu, S.D.; Burtica, R. NLP-Cube: End-to-End Raw Text Processing with Neural Networks. In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Brussels, Belgium, 31 October–1 November 2018; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 171–179. [CrossRef]
30. Schmid, H. Deep Learning-Based Morphological Taggers and Lemmatizers for Annotating Historical Texts. In Proceedings of the DATeCH. ACM, Brussels, Belgium, 8–10 May 2019; pp. 133–137.
31. Qi, P.; Zhang, Y.; Zhang, Y.; Bolton, J.; Manning, C.D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Online Conference, 5–10 July 2020; pp. 101–108. [CrossRef]
32. Rehm, G.; Berger, M.; Elsholz, E.; Hegele, S.; Kintzel, F.; Marheinecke, K.; Piperidis, S.; Deligiannis, M.; Galanis, D.; Gkirtzou, K.; et al. European Language Grid: An Overview. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020), Marseille, France, 11–16 May 2020; pp. 3359–3373.
33. Avram, A.M.; Păiș, V.; Tufis, D. Towards a Romanian end-to-end automatic speech recognition based on DeepSpeech2. *Proc. Rom. Acad. Ser. A* **2020**, *21*, 395–402.
34. Avram, A.M.; Păiș, V.; Tufiș, D. Romanian speech recognition experiments from the ROBIN project. In Proceedings of the 15th International Conference on Linguistic Resources and Tools for Natural Language Processing, Online Conference, 14–16 December 2020; pp. 103–114.
35. 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 Proceedings of the International Conference on Machine Learning, PMLR, New York City, NY, USA, 19–24 June 2016; pp. 173–182.
36. Avram, A.M.; Păiș, V.; Tufiș, D. Self-Supervised Pre-Training in Speech Recognition Systems. In *Speech Recognition Technology and Applications*; Păiș, V., Ed.; Nova Science Publishers: Hauppauge, NY, USA, 2022; pp. 27–56.
37. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20, Vancouver, BC, Canada, 6–12 December 2020 ; pp. 12449–12460.
38. Păiș, V.; Barbu Mititelu, V.; Ion, R.; Irimia, E. Evaluating a Fine-Tuned Whisper Model on Underrepresented Romanian Speech. In Proceedings of the 2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, 25–27 October 2023; pp. 141–145. [CrossRef]
39. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine Learning, JMLR.org, ICML'23, Honolulu, HI, USA, 23–29 July 2023.