

Article

Modern Speech Recognition for Romanian Language

Remus-Dan Ungureanu ¹ and Mihai Dascalu ^{1,2,*}

¹ Computer Science & Engineering Department, National University of Science and Technology POLITEHNICA Bucharest, 060042 Bucharest, Romania; remus_dan.ungureanu@upb.ro

² Science and Information Technology Section, Academy of Romanian Scientists, Ilfov 3, 050044 Bucharest, Romania

* Correspondence: mihai.dascalu@upb.ro

Abstract

Despite having approximately 24 million native speakers, Romanian remains a low-resource language for automatic speech recognition (ASR), with few accurate and publicly available systems. To address this gap, this study explores the challenges of adapting modern speech recognition models, such as wav2vec 2.0 and Conformer, to Romanian. Our investigation is a comprehensive analysis of the two models, their capabilities to adapt to Romanian data, and the performance of the trained models. The research also focuses on unique attributes of the Romanian language, data collection techniques, including weakly supervised learning, and processing methodologies. Building on the previously introduced Echo dataset of 378 h, we release CRoWL (Crawled Romanian Weakly Labeled), a weakly supervised dataset of 9000 h created via automatic transcription. We obtain strong results that, to the best of our knowledge, are competitive with or exceed publicly reported results for Romanian under comparable open evaluation settings, with Conformer attaining 3.01% WER on Echo + CRoWL and wav2vec 2.0 reaching 4.04% (Echo) and 4.17% (Echo + CRoWL). In addition to the datasets, we also release our most capable models as open source, along with their training plans, thereby providing a solid foundation for researchers interested in languages with limited representation.

Keywords: Romanian automatic speech recognition; low-resource language; wav2vec 2.0; Conformer; weakly supervised learning

1. Introduction

Automatic Speech Recognition (ASR) technology has seen notable advances over the past several years, enabling effortless human-machine interaction across a wide range of applications. The Romanian language, despite having an estimated 24 million native speakers worldwide [1,2], is often categorized as a low-resource language within the speech domain due to the limited availability of extensive, high-quality speech corpora and specialized models. The development of robust ASR systems for Romanian would unlock a wide array of applications, ranging from automated transcription to personal assistants, to critical services or systems such as enhanced emergency services dispatch [3] and innovative educational tools [4].

There are several challenges to achieving high-accuracy automatic speech recognition for Romanian. The major challenge is the language's status low resource, characterized not just by the quantity of available data but also by its quality, diversity, and accessibility. This scarcity of comprehensive speech datasets hinders the training of sophisticated acoustic models. Language models suffer from the same issue, but to a lesser extent. Echo [5] is a



Academic Editor: Gloria Corpas Pastor

Received: 8 January 2026

Revised: 8 February 2026

Accepted: 12 February 2026

Published: 14 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

benchmark dataset that serves as a more difficult, realistic evaluation corpus, addressing this gap through a community-driven crowdsourcing effort. This dataset is critical for the present study.

Another challenge resides in the complex phonetic and linguistic features of Romanian [6,7]. The phonetic inventory includes specific central vowels such as 'ă' (schwa) and 'ɨ' (close central unrounded vowel), which distinguish it from other Romance languages [6,8]. Furthermore, the language exhibits a high degree of inflection and morpho-phonological alternations [9,10]. In addition, the use of diacritics (e.g., ă, â, î, ș, ț) is integral to Romanian orthography and meaning [11]; their omission or incorrect handling by ASR systems can lead to ambiguity and errors. Finally, regional variation further contributes to acoustic/prosodic diversity [12].

Nevertheless, Romanian has a relatively phonemic orthography and shares similarities with other Romance languages (e.g., French, Italian, Spanish, and Portuguese), while also incorporating lexical influences from Slavic languages. Consequently, research on these related languages can be applied to a certain extent, and transfer learning may be applicable.

This paper investigates the adaptability and performance of two state-of-the-art modern speech processing architectures, wav2vec 2.0 [13] and Conformer [14], when applied to Romanian. The selection of these models is intentional; wav2vec 2.0, particularly its cross-lingual variant XLS-R [15], represents the forefront of self-supervised learning and is highly beneficial in low-resource settings by reducing reliance on labeled data. Conformer (and FastConformer [16]) architectures, with their hybrid approach combining convolutional and attention mechanisms, offer a powerful and parameter-efficient way to model both local acoustic details and global contextual dependencies. These architectural choices are well-suited to address the primary challenge of data scarcity while achieving high accuracy.

Our results highlight the strong impact of the amount and quality of training data on Romanian WER. At the same time, the two architectures behave differently across regimes: XLS-R is more robust in low-data settings, given that it is also pre-trained on Romanian, whereas FastConformer benefits more from additional weakly supervised data.

The main contributions of this study are as follows:

- Public release of two Romanian resources: the 378 h Echo benchmark (<https://huggingface.co/datasets/upb-nlp/echo>; accessed on 30 December 2025) and the 9000 h CRoWL corpus (<https://huggingface.co/datasets/upb-nlp/crowl-speech>; Accessed on 30 December 2025).
- Competitive Romanian baselines with wav2vec 2.0 and Conformer, reaching down to 3.01% WER (<https://huggingface.co/upb-nlp/ro-wav2vec2> and <https://huggingface.co/upb-nlp/ro-fast-conformer>; accessed on 30 December 2025).
- colorblueOpen-source plans for dataset creation and model fine-tuning (<https://github.com/upb-nlp/speech-ai>; accessed on 8 February 2026), including the CRoWL crawling pipeline (<https://github.com/upb-nlp/crowl-speech>; accessed on 8 February 2026).

This paper is organized as follows: Section 2 reviews the related work on English and Romanian datasets and models for ASR. Section 3 describes the methodology, including the datasets (i.e., Echo and CRoWL data collection and processing), training and experimental setup for the two ASR models, followed by details on the evaluation metric (i.e., WER). Section 4 presents the results, followed by further discussion and analysis of each model in Section 5. Finally, Section 6 concludes the paper, summarizing the findings and outlining directions for future research.

2. Related Work

Automatic Speech Recognition has evolved significantly over the past few years, yet results are typically reported only for high-resource languages. The few multilingual models that report for Romanian usually have unsatisfactory word error rates. In terms of absolute values, the accuracies for the English language are still superior to those for the Romanian language, indicating that the datasets for English are better, larger, and/or of higher quality, models have not been adapted for Romanian language particularities, or that the Romanian language is more difficult overall. For example, English speakers can enjoy models that reach as low as 1.9% word error rates on LibriSpeech test-clean (Conformer with a language model [14]). This surpasses the capabilities of even native English speakers, who achieved 5.83% WER on the LibriSpeech test-clean set [17].

2.1. English ASR Models

English ASR serves as a reference point, as the language benefits from large-scale corpora such as LibriSpeech and long-established evaluation protocols. Modern systems span (i) self-supervised, encoder-only CTC models (wav2vec 2.0 [13]/XLS-R [15]), (ii) convolution–attention hybrids (Conformer [14]/FastConformer [16]), and (iii) large weakly supervised sequence-to-sequence models (Whisper [18]). Table 1 summarizes representative best-reported WERs on LibriSpeech; these values are taken from the cited sources and may differ in decoding (e.g., external language models), normalization, and test-time settings, so they should be read as indicative rather than strictly comparable. Rather than model size alone, reported WER depends strongly on training data, decoding, and evaluation choices; this observation motivates our focus on controlled, open evaluation for Romanian.

Table 1. Speech-to-text models for English, number of parameters, and representative best-reported WER (%) on LibriSpeech. Some entries use external language models and/or different decoding settings as described in the cited sources.

Model Name and Variant	Parameters	Test-Clean	Test-Other
Conformer (small) [14]	10 M	2.1%	5.0%
Conformer (medium) [14]	30 M	2.0%	4.3%
Conformer (large) [14]	118 M	1.9%	3.9%
FastConformer (large) [16] (available at https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_fastconformer_transducer_large ; accessed on 8 February 2026)	114 M	1.8%	3.8%
FastConformer (xlarge) [16] (available at https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_fastconformer_transducer_xlarge ; accessed on 8 February 2026)	600 M	1.6%	3.0%
FastConformer (xxlarge) [16] (available at https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_fastconformer_transducer_xxlarge ; accessed on 8 February 2026)	1.1 B	1.5%	2.7%
wav2vec 2.0 (base) [13]	95 M	2.1%	4.8%
wav2vec 2.0 (large) [13]	317 M	1.8%	3.3%
Whisper (tiny) [18]	39 M	5.6%	14.6%
Whisper (base) [18]	74 M	4.2%	10.2%
Whisper (small) [18]	244 M	3.1%	7.4%
Whisper (medium) [18]	769 M	3.1%	6.3%
Whisper (large) [18]	1.55 B	2.7%	5.6%
Whisper (large-v2) (available at https://huggingface.co/openai/whisper-large-v2 ; accessed on 8 February 2026)	1.55 B	2.7%	5.2%
Whisper (large-v3) (available at https://huggingface.co/openai/whisper-large-v3 ; accessed on 8 February 2026)	1.55 B	2.3%	4.6%
Humans [17]	-	5.83%	12.69%

2.2. Romanian Datasets and ASR Models

Research on Romanian language models is notably limited compared to English, whether for speech or language models. Many published speech-to-text solutions for Romanian are, in fact, large multilingual models that are not specifically optimized or trained for Romanian's nuances [18].

Historically, early research efforts in Romanian ASR relied on HMM-based systems, as did those in English. These systems were typically constrained to tasks like isolated word recognition or applications with small vocabularies, primarily due to the scarcity of adequate speech resources for acoustic modeling and text resources for language modeling [19].

To address this resource scarcity, various corpus development initiatives have been undertaken, from hiring voice actors to record readings to crowd-sourcing and crawling the web for audio recordings and automatically transcribing them. Among these datasets, the most notable resources are RSC [20], SWARA [21], CommonVoice [22], and Echo [5] (see Table 2). Echo [5] is central to this study, and further details are provided in Section 3.1.1. Overall, the total available Romanian speech data from various sources amounts to roughly 500 h today. To compare Romanian efforts with those of other languages, similar amounts of data are required. A good starting point is to catch up with LibriSpeech and collect around 1000 h of audio recordings and their associated transcripts.

Table 2. Overview of commonly used Romanian speech corpora in the literature (representative figures from the cited sources).

Corpus	Hours	Speakers	Speech Type/Domain	Supervision
SWARA [21]	>21	17	read (studio-quality)	manual transcripts
RSC [20]	100	164	read (varied microphones)	manual transcripts
Common Voice (RO) [22]	varies	varies	read (crowd-sourced)	manual transcripts
Echo [5]	378	343	multi-domain; read + spontaneous	manual transcripts
CRoWL (this work)	9000	N/A	parliamentary speech	weak labels

For ASR models, we summarize representative baselines along with the evaluation set on which they are reported, since different papers report results on different test sets and use different decoding/normalization conventions. Georgescu et al. reported a WER of 3.27% on RSC alone [20] and 2.79% when combining RSC with the SSC corpus using Kaldi-based DNNs [23]. Over the past few years, a couple of models have been released that are getting close to English results, even though efforts for Romanian have not been as steady. Table 3 shows that current multilingual models achieve low accuracy, likely due to the lack of Romanian data in well-known international datasets.

Table 3. Representative Romanian ASR baselines reported in prior work and/or evaluated in this study.

Model Name	Parameters	Echo-Test
Whisper (small) [18]	244 M	35.0%
Whisper (large-v3) [18]	1.55 B	7.6%
Whisper-RO (small) [5] (available at https://huggingface.co/readerbench/whisper-ro ; accessed on 8 February 2026)	244 M	7.3%

3. Method

To conduct a fair analysis of speech recognition models for Romanian, multiple factors have been considered, including model selection criteria, data collection, preprocessing and normalization techniques, training and fine-tuning strategies, and evaluation metrics. This section presents two considered datasets: Echo, first introduced by Ungureanu and

Dascalu [5] and pivotal for this work, and the newly introduced resource, CRoWL. It also details the training of two open-source ASR models for Romanian and their corresponding experimental setup. This section concludes with details on the evaluation metric used, namely WER.

3.1. Datasets

3.1.1. Echo Dataset

Echo Ungureanu and Dascalu [5] also includes a crowdsourcing platform for data collection and verification, making for a frictionless experience for recording via a web interface and for comparing the decoded text from uploaded recordings with the original. The primary driver for creating the Echo dataset was the shortage of large, publicly accessible, and well-annotated speech corpora for Romanian. This initiative directly confronts the main obstacle to advancing Romanian ASR. The development and release of the Echo dataset enabled the current study, as Echo served as a strong foundation for training models that were later used to bootstrap the CRoWL dataset described in the next section via weakly supervised learning.

The Echo dataset stands out as a more challenging and realistic dataset compared to other existing Romanian speech corpora. It surpasses other Romanian datasets not only by the sheer volume of data (300+ h), but also by its diverse range of voices (300+ speakers), accents, speaking styles, and recording environments. All recordings consist of crowd-sourced read speech and were captured in non-professional (amateur) conditions. The challenging characteristics of Echo—including natural speech variability, ambient noise, and diverse acoustic conditions—make it a more representative test bed for evaluating ASR systems in real-world deployment scenarios, setting a higher bar for Romanian speech recognition performance.

Echo is an important part of the present research. An overview of the datasets used is provided in Table 4. To ensure robust evaluation and prevent spurious correlations, the Echo dataset was partitioned with careful attention to avoid data leakage. The primary constraint was that no transcript text appears in multiple splits. Given that some speakers recorded the entirety of certain transcript datasets (e.g., reading complete books or document collections), strict speaker separation would inevitably lead to text overlap across splits. Therefore, text-based partitioning took priority, with speaker overlap minimized where possible, ensuring the model cannot memorize texts while accepting some speaker presence across splits.

Table 4. Sub-datasets of the Echo dataset (with additional notes on speech style and recording conditions).

Domain	Recordings	Duration (h)	Speakers	Vocabulary	Details
Literature	34,896	69	207	10,661	high variability from different subtypes
- Drama	9077	13	198	2581	
- Epic	23,852	48	204	7643	
- Poems	1967	7	168	1182	
News	65,216	156	200	38,120	clean, up-to-date language
Emergency	8560	11	314	768	read with accent; disfluencies in speech
Legal	8832	28	194	2903	longer sentences and formal register
Wikipedia	45,193	111	329	7249	mixed topics and mixed conditions
Total I	162,697	378	343	49,664	

3.1.2. CRoWL: A Weakly Supervised Dataset

With the latest data additions, such as Echo, the total number of available hours remained unsatisfactory, at barely half of LibriSpeech. To supplement existing resources, we introduce CRoWL (Crawled Romanian Weakly Labeled), a dataset compiled by crawling

publicly available Romanian content and employing a weakly supervised learning approach for automatic alignment and transcription.

The CRoWL corpus was built exclusively from publicly available plenary sessions of the Romanian Chamber of Deputies (<https://www.cdep.ro/pls/steno/steno2015.home> accessed on 8 February 2026). We restrict the crawl to this source in order to ensure legal clarity and stable access conditions. The automatic transcription and alignment process leverages models initially trained on the Echo dataset, representing a weakly supervised learning paradigm where the model-generated transcriptions serve as noisy labels for further training.

CRoWL Processing Pipeline

Table 5 summarizes the end-to-end pipeline used to construct CRoWL: crawling plenary sessions, extracting audio, diarizing and segmenting speech turns, generating weak transcripts with an Echo-trained model, and applying automatic consistency filters (including characters-per-second). The released metadata includes per-utterance duration and trimming information to support reproducibility.

Table 5. CRoWL construction pipeline (high-level).

Step	Input	Output	Tooling/Notes
1. Crawling	session pages	media URLs + metadata	Official Parliamentary archive; session-level IDs
2. Audio extraction	video/stream	16 kHz mono WAV	ffmpeg; loudness normalization
3. Diarization + VAD	long-form audio	speech turns	PyAnnote-based diarization; remove non-speech
4. Segmentation	speech turns	short utterances	Split on pauses; keep within duration bounds
5. Weak transcription	segments	ASR pseudo-transcripts	Echo-trained ASR model; no manual correction
6. Filtering	audio + text	cleaned pairs	CPS + duration + trimming filters
7. Split + remove duplicates	cleaned pairs	train/val/test	transcript string matching; minimize speaker overlap

All collected audio data was standardized by converting it into a mono-channel, 16 kHz WAV format, processed and normalized as described in the following subsection.

Data Processing and Normalization

To ensure data quality and comparability across corpora, we apply the same audio and text normalization pipeline both during training and evaluation:

- Audio standardization: All audio is converted to mono 16 kHz WAV.
- Text normalization:
 - Lowercasing;
 - Whitespace normalization and removal of non-linguistic symbols frequent in crawled data;
 - Romanian diacritics are preserved; legacy forms (*ș*, *ț*) are mapped to Unicode-compliant (*ș*, *ț*);
 - Punctuation is removed for WER computation to avoid penalizing formatting;
 - Digits are expanded into their full word equivalents in Romanian (e.g., 10 → zece).
- Audio–text consistency filters:
 - Characters-per-second (CPS): We compute CPS as $|\text{text}|/\text{duration}$ and retain utterances with CPS in [1.2, 35.5]. Values outside this range usually indicate misalignment, non-speech regions, or truncated transcripts;
 - Duration: We kept segments in the [1 s, 80 s] range;
 - Trimming: We removed leading/trailing non-speech using an energy/VAD trimming pass and discarded clips with excessive remaining silence.

For reproducibility, the released dataset metadata includes per-utterance duration, trimming timestamps, and the normalized text used for WER scoring.

After data processing and normalization, 9000 h of audio were obtained from over 2000 recordings; the total number of unique words is approximately 170,000. As speaker identity is not consistently available in the source material and diarization is imperfect, we do not report a reliable count of unique speakers for CRoWL.

The CRoWL dataset was partitioned into training, validation, and test sets using an 80%-10%-10% split, respectively. To avoid spurious correlations and reduce data leakage, the splitting strategy ensured that no text/transcript appeared in multiple sets—each transcript was assigned exclusively to the training, validation, or test set. The vocabulary is naturally shared across splits as it represents the language itself.

3.1.3. Consolidated Echo + CRoWL Test Set

For convenience and reproducibility, we provide a consolidated evaluation target, referred to as Echo + CRoWL. This consolidated test set is *nota* new dataset; it is the concatenation of the official test splits of Echo and CRoWL, evaluated with the same normalization and scoring protocol. The goal is to provide a single public test set that spans both crowd-sourced multi-domain speech (Echo) and formal parliamentary speech (CRoWL), enabling consistent comparisons across future Romanian ASR work; as such, the combination of CRoWL and Echo offers a valuable mix of speech data, blending weakly supervised and fully supervised learning paradigms. Training models on Echo + CRoWL is expected to yield more robust ASR systems capable of handling both controlled and challenging real-world scenarios.

Table 6 reports the observed statistics for the final consolidated Echo + CRoWL test: utterance and speaker counts, total duration, and vocabulary size. All entries underwent the same quality checks described earlier, and deduplication ensured that no audio–text pair appeared in both the training and test splits. An attempt was made to minimize overlap in speakers and vocabulary, but the priority was to keep the texts unique to each split.

Table 6. Composition of the consolidated Echo + CRoWL test set.

Source	Utterances	Duration (h)	Speakers	Vocabulary
Echo (test split)	53,818	72.26	337	41,979
CRoWL (test split)	16,305	33.69	N/A	2990
Total	70,123	105.95	N/A	42,957

Speaker counts are available for Echo (337 unique speakers). For CRoWL, speaker metadata is incomplete, so the total speaker figure is omitted.

To prevent data leakage, any transcript present in the consolidated Echo + CRoWL test set is removed from the corresponding training splits, and a deduplication pass via string matching to ensure that no audio clip appears in both the train and test splits. Speaker overlap with the training data is minimized and occurs only when transcript uniqueness would otherwise be violated, mirroring the priority order used when splitting Echo and CRoWL individually. This makes the test set a difficult but fair target: models cannot rely on memorizing text, and broad coverage of accents and acoustic conditions is preserved. We release the consolidated Echo + CRoWL test set alongside the trained models to facilitate continued research.

Introducing this consolidated Echo + CRoWL test set addresses a recurring problem in Romanian ASR research: Model comparisons are performed on small, domain-specific test sets that are not mutually compatible. By evaluating both models on exactly the same multi-

domain Romanian test set, we enable reproducible progress tracking and allow future work to compare against a single WER figure rather than a collection of incomparable results.

3.2. Training ASR Models for Romanian

3.2.1. Model Selection

The choice of ASR models for this investigation was guided by their alignment with current state-of-the-art approaches and their potential suitability for the Romanian language, particularly given its low-resource status. wav2vec 2.0 [13] (specifically, its cross-lingual variant XLS-R) and Conformer [14] (specifically, its efficient variant, FastConformer [16]) were selected. wav2vec 2.0 excels in self-supervised pre-training, learning robust representations from vast amounts of unlabeled data, which is crucial when transcribed data is scarce. Conformer models are known for their powerful hybrid architecture, which effectively blends convolutional layers for local feature extraction with attention mechanisms for global context modeling, resulting in high accuracy.

This selection was made after considering several alternatives, as detailed in Table 7. DeepSpeech, while historically significant, is no longer at the cutting edge of ASR accuracy. Kaldi remains a powerful and versatile toolkit, especially for research; however, its pace of integration with the latest neural architectures is slower, and its tooling is often more complex for end-to-end production-like systems. OpenAI's Whisper model, despite its strong multilingual capabilities and end-to-end nature, presents challenges due to its larger size, making fine-tuning computationally intensive and deployment slow; moreover, it suffers from a known issue of "hallucinations," where the model outputs a transcript entirely unrelated to the actual audio [18]. For a foundational study aiming to establish reliable models for Romanian, wav2vec 2.0 and Conformer deliver a better balance of performance and avoidance of known issues, making them more suitable for practical use cases in our low-resource context.

Table 7. Comparison of ASR model candidates and rationale for selection (selected models are marked with bold).

Model Name	Architecture	Strengths	Weaknesses
DeepSpeech	RNN	Well-known, historically important	No longer SOTA
Kaldi	HMM-DNN Hybrid	Powerful toolkit, flexible Ideal for research	Complex to develop new models
Whisper	Transformer	Multilingual, multitask End-to-end with LM	Large models, slow fine-tuning/ inference Hallucinations
wav2vec 2.0 (XLS-R)	Transformer	SOTA self-supervised learning Excellent for low-resource Robust pre-training Multilingual	Requires fine-tuning CTC output may need LM
Conformer	Transformer + CNN	SOTA hybrid Efficient architecture	Needs substantial data for best accuracy

3.2.2. Training and Fine-Tuning

The core experimental approach involves fine-tuning pre-trained wav2vec 2.0 (specifically, XLS-R, given its multilingual pre-training that includes Romanian data) and Conformer models using the collected and processed Romanian speech datasets. General best practices for fine-tuning large pre-trained models in low-resource language scenarios were considered:

- **Transfer learning:** The fundamental strategy is to leverage the rich representations learned by models pre-trained on vast datasets. XLS-R, for instance, has already been exposed to thousands of hours of Romanian speech during its pre-training phase, providing a strong starting point.
- **Layer freezing:** In some low-data regimes, it can be beneficial to freeze the initial layers of the pre-trained model (e.g., the feature encoder) and only fine-tune the upper layers (e.g., the Transformer blocks and the final classification head). This helps preserve the general acoustic representations learned during pre-training while adapting the task-specific layers. The wav2vec 2.0 paper notes that the feature encoder is not trained during fine-tuning.
- **Data augmentation:** Techniques like SpecAugment [24], which involve masking frequency bands and time steps in the spectrogram, are often applied during fine-tuning to improve model robustness and prevent overfitting, especially with limited data.
- **Tokenizer:** For character-based models like wav2vec 2.0-CTC (as used for Librispeech), the output vocabulary consists of characters. For Conformer models that might use sub-word units, a Romanian-specific tokenizer or adaptation of a multilingual tokenizer would be necessary.

The specific hyperparameters and configurations used for fine-tuning wav2vec 2.0 and Conformer in this study are detailed in the Experimental Setup section.

3.2.3. Experimental Setup

This subsection describes the experimental setup for fine-tuning and evaluating the wav2vec 2.0 and Conformer models on Romanian-language datasets. The fine-tuning process was conducted under two primary data conditions:

1. **Echo only:** The models were fine-tuned solely on Echo, a fully supervised dataset.
2. **Echo + CRoWL:** The models were fine-tuned on Echo and CRoWL, incorporating both fully supervised and weakly supervised learning approaches.

The experimental design, particularly the comparison between Echo only and Echo + CRoWL training conditions for both models, is structured to directly investigate the impact of dataset size and diversity on ASR performance for Romanian. This is a critical aspect of research for low-resource languages, where understanding the trade-offs between data quantity, quality, and diversity is essential for developing effective systems.

wav2vec 2.0

We use the XLS-R checkpoint `facebook/wav2vec2-xls-r-300m` (~317 M parameters, available at <https://huggingface.co/facebook/wav2vec2-xls-r-300m>; accessed on 8 February 2026). This checkpoint is pre-trained on large-scale multilingual data and includes substantial Romanian coverage (i.e., 17,515 h of Romanian speech data [15]), which is an important confound when comparing to models without Romanian in pre-training. The model was initially trained on 436,000 h, representing 128 languages.

Architecturally, XLS-R follows the wav2vec 2.0 design [13,15]: a convolutional feature encoder maps raw audio to latent representations (at roughly 20 ms frame stride, with a receptive field of about 25 ms), which are then processed by a Transformer context network. During self-supervised pre-training, spans of these latent features are masked, and the model is trained with a contrastive objective: A quantization module discretizes the unmasked features using product quantization with Gumbel-Softmax codebook selection, and the contextualized representation at each masked time step must identify the true quantized target among $K = 100$ distractors; an auxiliary diversity term encourages uniform codebook usage [15].

The specific XLS-R checkpoint used here corresponds to a wav2vec 2.0 model with 24 Transformer blocks, model dimension $H_m = 1024$, feed-forward inner dimension $H_{ff} = 4096$, and 16 attention heads (about 317 M parameters in total) [15]. For ASR fine-tuning, a linear CTC projection layer is added on top of the Transformer outputs, while the quantization module is only used for pre-training [13].

Conformer

We initialize from a pre-trained multilingual FastConformer Hybrid Large model with punctuation/capitalization (stt_multilingual_fastconformer_hybrid_large_pc) ~114 M parameters; available at https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_multilingual_fastconformer_hybrid_large_pc; accessed on 8 February 2026). This checkpoint was trained on approximately 20,000 h across 10 European languages (Belarusian, German, English, Spanish, French, Croatian, Italian, Polish, Russian, and Ukrainian) and does not include Romanian. We fine-tune the model on Romanian using the same two training conditions (Echo; Echo + CRoWL) and evaluate with greedy CTC decoding (no external language model) under the same normalization as XLS-R.

The architecture of FastConformer builds upon the Conformer encoder [14]: An initial convolutional subsampling front-end reduces the input sequence length (e.g., from 10 ms to 40 ms frame rate), followed by a stack of Conformer blocks in a Feed-Forward–Multi-Headed Self-Attention–Convolution–Feed-Forward arrangement with half-step residual connections. The self-attention module uses relative sinusoidal positional encodings (Transformer-XL style) and pre-norm residual connections, while the convolution module combines pointwise convolution + GLU with a depthwise 1-D convolution, batch normalization, and Swish activations [14]. FastConformer speeds up this design by increasing the effective downsampling from $4\times$ to $8\times$ using a 2/4/8 subsampling schedule implemented with depthwise-separable convolutions and smaller convolution kernels (e.g., $K = 9$ vs. $K = 31$ reported for the original Conformer subsampling) [16]; for CTC variants, the FastConformer paper notes that $8\times$ subsampling can require switching from character to subword tokenization [16].

Training Configurations

To improve reproducibility, we report the main training configuration and evaluation settings here. Exact configuration files are provided in the accompanying open-source recipes (see Table 8).

Table 8. Key training configuration (shared across experiments unless noted).

Item	Setting
Audio sampling rate	16 kHz; mono
Optimizer	AdamW
Mixed precision	FP16
Epochs	20
Decoding for WER	Greedy CTC decoding without an external language model; same text normalization for all evaluations
Seed	Fixed for data shuffling and initialization
Hardware	Single-node GPU (NVIDIA A100 with 80 GB)
Total training time	69 h for wav2vec 2.0 when training on Echo + CRoWL and 42 h when training only on Echo 24 h for Conformer when training on Echo + CRoWL and 11 h when training only on Echo

3.3. Evaluation Metric

The primary metric employed to assess the performance of the ASR models in this study is the Word Error Rate (WER). WER is a standard ASR metric that measures the dissimilarity between the reference transcription and the ASR model's hypothesis. It is calculated as the minimum number of substitutions (S), deletions (D), and insertions (I) required to transform the hypothesis into the reference, normalized by the total number of words (N) in the reference text:

$$\text{WER} = \frac{S + D + I}{N}$$

A lower WER indicates better ASR performance, signifying fewer errors in the transcribed text. It is typically expressed as a percentage.

4. Results

Table 9 reports word error rates (WER, %) for each trained checkpoint across five evaluation sets. All WER values are computed with the same normalization and greedy CTC decoding without an external language model (see Table 8).

Table 9. Experimental Results (WER %) for trained models.

Model	Training Dataset	Evaluation Dataset				
		Common Voice	Echo	SWARA	RSC	Echo + CRoWL
wav2vec 2.0	Echo	9.21	4.04	7.51	6.63	6.39
	Echo + CRoWL	4.58	4.51	2.98	3.04	4.17
Conformer	Echo	9.47	8.43	8.98	7.91	12.16
	Echo + CRoWL	2.81	4.23	2.80	2.75	3.01

5. Discussion

The wav2vec 2.0 and Conformer models evaluated in this study outperform the Whisper-small model fine-tuned on Echo (i.e., Whisper-RO small) reported by Ungureanu and Dascalu [5]. In their evaluation, Whisper-RO (small) achieved WERs of 12.2% (Common Voice), 10.9% (FLEURS), 9.4% (VoxPopuli), 7.3% (Echo), and 5.4% (RSC), while Whisper large-v3 attained 10.8%, 8.2%, 13.8%, 27.2%, and 24.9% on the same datasets, respectively [5]. Figure 1 argues for the better scalability of the Conformer architecture.

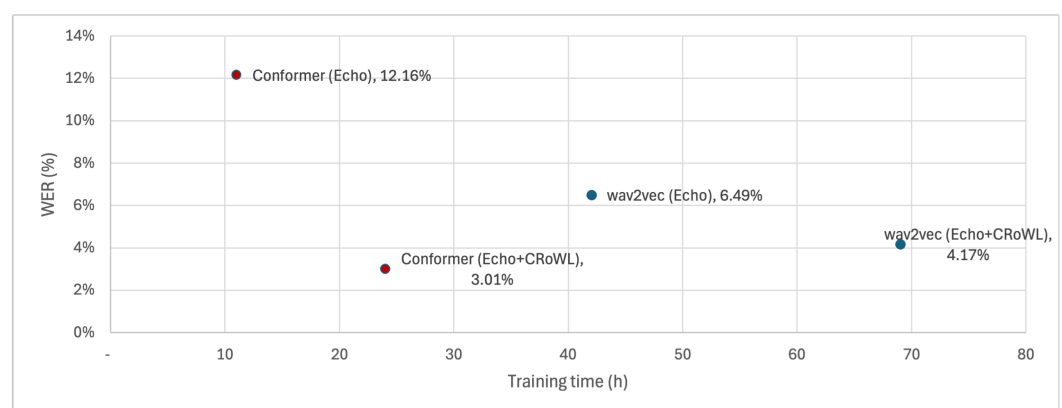


Figure 1. WER as a function of training time on Echo + CRoWL.

To the best of our knowledge, there are no other publicly released Romanian ASR models with reported WER results that are directly comparable on our consolidated test set under the same normalization. Commercial ASR APIs (e.g., Google, Azure) support

Romanian but generally do not publish WER on open Romanian benchmarks, and systematic comparison to proprietary systems is outside the scope of this work. Instead, we release the consolidated test set, normalization scripts, and model checkpoints to enable transparent and independent comparisons.

Based on manual review of a sample of decoded utterances, the most frequent qualitative error sources were (a) proper nouns and foreign words; (b) numeric expressions and abbreviations; (c) diacritics in rare words and names; (d) disfluencies and partial words in spontaneous speech (Echo Emergency); and (e) domain-specific terminology (i.e., legal and parliamentary registers). These observations are intended as a qualitative complement to WER, highlighting practical failures that guide future improvements.

5.1. Analysis of wav2vec 2.0 Performance

The wav2vec 2.0 model, an XLS-R variant, achieved robust performance across the tested conditions. When fine-tuned on Echo + CRoWL and evaluated on the same comprehensive test set, it achieved a WER of 4.17%. This result underscores the model's capability to effectively leverage larger, more diverse datasets, including weakly supervised data.

An interesting observation arises from the cross-evaluation scenarios. When fine-tuned on Echo only, the model achieved a WER of 4.04% on the Echo test set. However, this same model, when evaluated on the Echo + CRoWL test set, yielded a worse WER of 6.39%. Conversely, the model fine-tuned on Echo + CRoWL also achieved a worse WER of 4.51% when evaluated specifically on the Echo test set.

This pattern suggests that Echo may exhibit a cleaner, more homogeneous distribution of speech than CRoWL. Training exclusively on Echo seems to produce a model that generalizes reasonably well to the broader, potentially more varied, Echo + CRoWL. However, training on the more diverse Echo + CRoWL results in a model that, while superior on the Echo + CRoWL test set, is slightly less specialized for Echo. This could imply that CRoWL, with its weakly supervised nature, introduces acoustic or linguistic variability (e.g., noise, different speaking styles, domain-specific parliamentary vocabulary, imperfect alignments) that, while beneficial for general robustness, slightly degrades performance on the more focused Echo if the model has adapted to these broader conditions. This highlights the importance of test set composition and the potential benefits of domain adaptation when optimal performance on a specific subset, such as Echo, is desired after broad training.

5.2. Analysis of Conformer Performance

The Conformer model achieved the lowest overall WER of 3.01% when fine-tuned and evaluated on Echo + CRoWL. This result is superior to the best wav2vec 2.0 performance (4.17% WER on the same Echo + CRoWL), indicating Conformer's higher potential when provided with sufficient training data.

However, Conformer's performance is more sensitive to the amount of training data. When fine-tuned on Echo only, it yielded a WER of 12.16%, which is considerably higher than wav2vec 2.0's 6.39% WER on the same limited dataset; however, the model had not been previously exposed to any Romanian. This supports the characteristic of Conformer models, which require substantial data to effectively train their numerous parameters and complex interactions between convolutional and attention layers. The performance progression from 12.16% (limited training data, difficult test data), 8.43% (same limited training data, but similar test data), 4.23% (increased training data), and 3.01% (increased training data and similar test data) clearly illustrates its strong scalability and benefit from increased training data volumes for Romanian.

The data-dependent performance disparity between wav2vec 2.0 (XLS-R) and Conformer (FastConformer) is a noteworthy observation. XLS-R, benefiting from its extensive

cross-lingual pre-training that included a considerable amount of Romanian speech, exhibits a stronger performance baseline when fine-tuned on smaller target-language datasets like Echo. It essentially has a "head start" for Romanian. Conformer, while potentially more powerful at modeling speech, requires a larger volume of in-domain Romanian data to learn effectively and surpass the performance of the heavily pre-trained XLS-R. This suggests that in extremely low-resource scenarios for Romanian ASR, a model like XLS-R might be a more pragmatic initial choice. However, as more Romanian-specific data becomes available (as demonstrated by Echo + CRoWL), Conformer architectures are likely to yield superior accuracy. This trade-off between pre-training benefits and fine-tuning data requirements provides valuable guidance for future model selection strategies, given the availability of Romanian speech resources.

5.3. Limitations

While this study provides important results, certain limitations should be acknowledged. The experiments did not exhaustively explore the full hyperparameter space for model fine-tuning and did not consider all possible data augmentation or advanced language modeling techniques that could further enhance performance. The CRoWL dataset, which is weakly supervised and uses automatically generated transcriptions, introduces uncontrolled variability and label noise that could influence model training. The evaluation primarily focused on WER, with less-detailed qualitative error analysis.

Several additional limitations should be acknowledged. First, we did not conduct an exhaustive exploration of the hyperparameter space for fine-tuning (e.g., learning-rate schedules, layer-wise freezing, regularization, batch size, decoding settings), nor did we systematically evaluate the effect of alternative training recipes such as SpecAugment variants, noise/reverberation augmentation, curriculum strategies, or more advanced decoding with external language models.

Second, the CRoWL dataset is weakly supervised and uses automatically generated transcriptions. As a result, the dataset likely contains transcription errors, segmentation mismatches, and stylistic variation across sources, which introduce uncontrolled variability and label noise that may affect both convergence and the robustness of the learned representations.

Third, while we report results across multiple test sets, we did not perform statistical significance testing (e.g., bootstrap resampling) or stratified analyses by speaker attributes, recording conditions, or dialectal region.

Fourth, our evaluation emphasizes WER and provides only a limited qualitative analysis of error types; we do not quantify error categories (e.g., numbers, named entities, or diacritics) or report complementary metrics such as CER and diacritic-sensitive scoring that could better capture orthographic fidelity.

Finally, we restrict comparisons to publicly released open-source systems; we do not benchmark against commercial ASR APIs, and results may change under different normalization conventions or domain shifts beyond the datasets considered here.

6. Conclusions and Future Work

This study investigated the difficulties encountered in adapting and evaluating two ASR architectures for Romanian, wav2vec 2.0 (and its cross-lingual variant, XLS-R) and Conformer (and its more efficient variant, FastConformer), and quantified the impact of adding weakly supervised speech from publicly available parliamentary recordings. Beyond the models, we release CRoWL (~9000 h) and open-source checkpoints and recipes to facilitate reproducible Romanian ASR research.

The main findings reveal distinct performance characteristics for the two models. Conformer, when trained on Echo + CRoWL (blending fully supervised and weakly supervised learning), achieved the best overall Word Error Rate of 3.01%. This highlights its capacity to effectively leverage large datasets, including noisy weakly supervised data, to achieve high accuracy. In contrast, wav2vec 2.0 (XLS-R) performed better in low-data regimes; when fine-tuned on Echo only, it achieved a WER of 6.39%, outperforming Conformer's 12.16% on the same data. This advantage is attributable to XLS-R's extensive cross-lingual pre-training, which included a substantial volume of unlabeled Romanian speech, providing it with a robust initial understanding of Romanian acoustic properties. The best WER achieved by wav2vec 2.0 was 4.04% when trained and evaluated on Echo.

This research establishes an updated performance baseline for modern open-source ASR models in Romanian. Given the limited prior work focusing specifically on applying these architectures to Romanian, this study may serve as a guide for future research and development in Romanian speech technology. The CRoWL dataset illustrates a practical application of weakly supervised learning for low-resource ASR, showing how models trained on clean data (Echo) can be leveraged to automatically annotate and learn from noisier web-crawled data. The study showcases a practical workflow for low-resource ASR development: identifying data scarcity, creating and collecting a curated dataset (Echo), scaling training data through weakly supervised learning (CRoWL), adapting strong pre-trained models, and establishing reproducible baselines. Together, these steps provide a clear starting point and concrete directions for future improvements.

A primary direction for future research is to integrate external language models trained on large Romanian text corpora to improve the fluency and accuracy of both wav2vec 2.0 and Conformer outputs. Exploring advanced data augmentation techniques tailored to Romanian speech characteristics and the specifics of the available datasets could further enhance model robustness. Deeper investigations into modeling unique Romanian phonetic features, the consistent handling of diacritics, and dialectal variation are also warranted. Continued expansion of Romanian corpora and further refinement of weakly supervised learning methods to better leverage noisy web-crawled Romanian data (expanding CRoWL) are required. Evaluating different model variants (e.g., various sizes of XLS-R, other FastConformer configurations) and optimizing models for computational efficiency for practical deployment are also important next steps. Finally, a more detailed qualitative and quantitative error analysis will be essential for understanding the specific challenges these models face in Romanian speech and guiding targeted improvements in future iterations.

In conclusion, while the core model architectures themselves are not novel, we provide an open evaluation in the underrepresented Romanian setting and release of our most capable open-source models and their training recipes, which together constitute a substantial contribution. This work extends the empirical understanding of these models and provides the research community with valuable data and effective weakly supervised learning resources for low-resource ASR (CRoWL), all of which contribute to continued progress in making speech recognition technology more accessible and effective for Romanian.

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

Funding: This research was supported by the project “Romanian Hub for Artificial Intelligence—HRIA”, Smart Growth, Digitization and Financial Instruments Program, 2021–2027, MySMIS No. 351416.

Data Availability Statement: Echo is available at <https://huggingface.co/datasets/upb-nlp/echo> (accessed on 30 December 2025), and the CRoWL corpus is available at <https://huggingface.co/datasets/upb-nlp/crowl-speech> (accessed on 30 December 2025).

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

Abbreviations

The following abbreviations are used in this manuscript:

ASR	Automatic Speech Recognition
CNN	Convolutional Neural Network
CTC	Connectionist Temporal Classification
DNN	Deep Neural Network
HMM	Hidden Markov Model
LM	Language Model
RNN	Recurrent Neural Network
WER	Word Error Rate

References

1. Eberhard, D.M.; Simons, G.F.; Fennig, C.D. *Ethnologue: Languages of the World*, 26th ed.; SIL International (SIL Global Publishing): Dallas, TX, USA, 2023.
2. Posner, R. Romanian Language. Encyclopaedia Britannica. 2026. Available online: <https://www.britannica.com/topic/Romanian-language> (accessed on 7 February 2026).
3. Ungureanu, D.; Toma, S.A.; Filip, I.D.; Mocanu, B.C.; Aciobăniței, I.; Marghescu, B.; Balan, T.; Dascalu, M.; Bica, I.; Pop, F. ODIN112–AI-Assisted Emergency Services in Romania. *Appl. Sci.* **2023**, *13*, 639. [\[CrossRef\]](#)
4. Ungureanu, D.; Ruseti, S.; Toma, I.; Dascalu, M. pRONounce: Automatic Pronunciation Assessment for Romanian. In *Conference on Smart Learning Ecosystems and Regional Development*; Springer: Singapore, 2022; pp. 103–114.
5. Ungureanu, D.; Dascalu, M. Echo: A Crowd-sourced Romanian Speech Dataset. *Interact. Des. Archit. J.—IxD&A* **2024**, *62*, 141–152. [\[CrossRef\]](#)
6. Chitoran, I. *The Phonology of Romanian: A Constraint-Based Approach*; Walter de Gruyter: Berlin, Germany, 2013; Volume 56.
7. Pană Dindelegan, G. (Ed.) *The Grammar of Romanian*; Oxford University Press: Oxford, UK, 2013.
8. Renwick, M.E. Vowels of Romanian: Historical, Phonological and Phonetic Studies. Ph.D. Thesis, Cornell University, Ithaca, NY, USA, 2012.
9. Stan, C.; Moldoveanu Pologea, M. Inflectional and Derivational Morphophonological Alternations. In *The Grammar of Romanian*; Pană Dindelegan, G., Ed.; Oxford University Press: Oxford, UK, 2013; pp. 607–611.
10. Șulea, O.M. Semi-supervised Approach to Romanian Noun Declension. *Procedia Comput. Sci.* **2016**, *96*, 664–671. [\[CrossRef\]](#)
11. Tufis, D.; Ceausu, A. Diacritics Restoration in Romanian Texts. In *Proceedings of the a Common Natural Language Processing Paradigm for Balkan Languages—RANLP 2007 Workshop Proceedings, Borovets, Bulgaria, 27–29 September 2007*; Paskaleva, E.; Slavcheva, M., Eds.; INCOMA Ltd.: Shoumen, Bulgaria, 2007; pp. 49–56.
12. Roseano, P.; Turculeț, A.; Bibiri, A.D.; Cerdà Massó, R.; Fernández Planas, A.M.; Elvira-García, W. A dialectometric approach to Romanian intonation. *Onomázein* **2022**, 105–139. [\[CrossRef\]](#)
13. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12449–12460.
14. Gulati, A.; Qin, J.; Chiu, C.C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv* **2020**, arXiv:2005.08100. [\[CrossRef\]](#)
15. Babu, A.; Wang, C.; Tjandra, A.; Lakhotia, K.; Xu, Q.; Goyal, N.; Singh, K.; Von Platen, P.; Saraf, Y.; Pino, J.; et al. XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv* **2021**, arXiv:2111.09296.
16. Rekesch, D.; Koluguri, N.R.; Krizan, S.; Majumdar, S.; Noroozi, V.; Huang, H.; Hrinchuk, O.; Puvvada, K.; Kumar, A.; Balam, J.; et al. Fast conformer with linearly scalable attention for efficient speech recognition. In *Proceedings of the 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Taipei, Taiwan, 16–20 December 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–8.
17. 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, New York, NY, USA, 19–24 June 2016*; International Machine Learning Society (IMLS): San Diego, CA, USA, 2016; pp. 173–182.

18. 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 International Conference on Machine Learning (PMLR), Honolulu, HI, USA, 23–29 July 2023; pp. 28492–28518.
19. Caranica, A.; Burileanu, C. An automatic speech recognition system with speaker-independent identification support. In *Proceedings of the Advanced Topics in Optoelectronics, Microelectronics, and Nanotechnologies VII, Constanta, Romania, 21–24 August 2014*; SPIE: Bellingham, WA USA, 2015; Volume 9258, pp. 769–775.
20. 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.
21. Stan, A.; Dinescu, F.; Țiple, C.; Meza, Ș.; Orza, B.; Chirilă, M.; Giurgiu, M. The SWARA speech corpus: A large parallel Romanian read speech dataset. In *Proceedings of the 2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, 6–9 July 2017*; IEEE: Piscataway, NJ, USA, 2017; pp. 1–6.
22. Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tysers, F.M.; Weber, G. Common voice: A massively-multilingual speech corpus. *arXiv* **2019**, arXiv:1912.06670.
23. Georgescu, A.L.; Cucu, H.; Burileanu, C. Kaldi-based DNN architectures for speech recognition in Romanian. In *Proceedings of the 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Timisoara, Romania, 10–12 October 2019*; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6.
24. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Interspeech 2019*; ISCA: Grenoble, France, 2019. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.