State-of-the-Art Review on Recent Trends in Automatic Speech Recognition

Abdou Karim Kandji[1(✉)], Cheikh Ba[1], and Samba Ndiaye[2]

[1] University of Gaston Berger, Saint-Louis, Senegal
{kandji.abdou-karim1,cheikh2.ba}@ugb.edu.sn
[2] Cheikh Anta Diop University, Dakar, Senegal
samba.ndiaye@ucad.edu.sn

Abstract. In the ever-changing technological landscape, speech recognition stands out as a growing discipline within the field of natural language processing (NLP). This major breakthrough in human-machine interfaces has dramatically reshaped the way we interact with the digital systems and intelligent environments around us. Speech recognition, as a cornerstone of this revolution, aims to accurately and quickly translate the complex modulations of the human voice into text, thus opening up a multitude of applications ranging from virtual assistants and voice control systems to digital devices, communication aid and automated transcription. It will also facilitate illiterate people's access to various digital services and improve financial inclusion. This article dives deep into the state of the art in speech recognition, exploring technological advances, cutting-edge algorithmic models, deep learning methodologies, and persistent challenges driving research such as low-resource languages, multilingual models and innovation in this constantly evolving field. By taking a close look at the progress made, current gaps and future prospects, this review aims to offer a comprehensive overview of the most recent and relevant developments in speech recognition.

Keywords: Automatic Speech Recognition · End-to-End models · Multilingual speech recognition · Low-resource languages · African Languages

1 Introduction

Automatic speech recognition (ASR) is the act of transcribing the human voice represented in a waveform into text. The text produced should respect grammatical rules of the language.

Here are the main tasks needed to perform speech-to-text transcription:

(1) Capturing the audio signal: It is necessary to have a microphone or an audio recording device in order to capture the audio of the speech;
(2) Signal pre-processing: A pre-processing phase of the audio signal is necessary in order to reduce or even eliminate the noise contained in the signal (background noise, environmental noise, etc.);

M. Masinde et al. (Eds.): AFRICATEK 2023, LNICST 520, pp. 185–203, 2024.
https://doi.org/10.1007/978-3-031-63999-9_11

(3) Segmenting speech: In this stage, it is necessary to segment the speech audio signal into short elements such that each element represents a phoneme which is defined as a sound unit of spoken speech. Each segment is then associated with one or more phonetic labels which represent labels associated with human speech sounds;
(4) Recognizing speech: In this task, the goal is to identify the spoken word(s) based on the phonetic labels associated with the acoustic representation of speech;
(5) Correction or post-processing: Last step is dedicated for correction of any error or inconsistency introduced in the previous step. This may include correcting mistranscribed word errors. A linguistic model is responsible for doing this processing.

It is important to note that the quality of the transcription depends on several factors, including the quality of the input audio, the performance of the speech recognition engine used and the possible post-processing steps with the language model. The traditional approach to speech recognition was essentially based on Bayesian statistical models during the 1970s. This approach consisted of three modules: *acoustics*, *phonetics* and *linguistics.*

The first two tasks listed earlier require using a recording microphone and then applying the Fourier transform for signal pre-processing. The acoustic module is responsible for processing task (3), then the phonetics module will in turn solve task (4). Finally, the last task (5) is processed by the linguistics module. After defining the different tasks necessary for voice recognition, then showing the solutions proposed for solving these tasks, we will present the evolution of the technology from the traditional approach to the present days.

From the 90s to the 2000s, a so-called 'hybrid' model combining a neural network model (for text recognition) and an HMM-based model (for sequential alignment) emerged. One of the disadvantages of traditional models is that they require the implementation and training of separate components, making the development of new speech recognition technologies difficult and expensive. In conclusion, this model requires a lot of expertise on each module as well as expert linguistic knowledge to make it work. It also requires separate training and design process. To overcome these drawbacks, the last decade has seen the appearance of '*End-to-End'* models which have the ability to take an audio sequence as input and produce a sequence of characters as output, thanks to a single neural network which offers a simpler and efficient approach compared to traditional models. This eliminates the need to build intermediate components.

The 'End-to-End' models provided solutions for the tasks listed earlier. We can note as a solution to the sequence alignment task, the Connectionist Temporal Classification (CTC) [1] loss function which will learn to align the audio to the transcription. A special symbol is introduced to handle the size difference between the input sequence and the output sequence using following equation:

$$p(l|x) = \sum_{\pi \in \mathcal{B}^{-1}(l)} p(\pi|x) \tag{1}$$

It is also possible to join the acoustic and linguistic information, thanks to the RNN-T architecture [2].

The seq2seq architecture is also an approach for solving voice recognition tasks using an '*Encoder'* capable of encoding the acoustic signal and then translating the

information produced by the Encoder into an output sequence using the '*Decoder'* via the attention mechanism [3].

However, we can note some limitations of the 'End-To-End' models such as:

- The CTC loss function assumes the independence of each word at the time of prediction, which can lead to inconsistent sentences;
- The attention mechanism is an auto-regressive model, and therefore it is able to produce only one token per time step, which results in a slow inference time;
- 'End-To-End' models require a large volume of data for its learning process.

Limits to CTCs and attention mechanism have been resolved with transducer models [2, 4].

To solve the limitations imposing a large volume of data, the following approaches have been proposed:

- Data augmentation techniques [5–7];
- Pre-training techniques, transfer learning [8, 9];
- Multitasking or multilingual learning techniques [10, 11];
- Unsupervised or self-supervised learning techniques [12, 13].

Another limit lies in low-resource African languages that have often been left behind in terms of research work. Resources in these languages are not only rare but they are even more so when it comes to specialized fields such as agriculture, finance or health. Current work [14–17] are mainly focused on language recognition in general and not in specific areas as mentioned.

In the next sections, we will start with the history of voice recognition going from the 1940s to the present days. Then we will detail the recent advances starting with monolingual then multilingual models, then we will end with models adapted to low-resource languages.

2 First Solutions Proposed in the Literature

2.1 Solutions Proposed in the 40s and 50s

The first automatic speech recognition solutions were developed in the late 1940s and early 1950s. A notable system, created by Bell Labs, was able to recognize any of 10 digits (0–9) spoken by a single speaker [18] with the system named Audrey. The system achieved an impressive 97 to 99% accuracy by selecting the model with the highest correlation coefficient with the input. The "Pattern matching" approach was used to compare each spoken word to a reference base and the choice was made according to the distance of the nearest word.

Authors of [19, 20] were the first to implement a phoneme recognition system capable of recognizing four vowels and five consonants.

2.2 Solutions Proposed in the 60s and 70s

Significant changes occurred during the late 1960s and early 1970s. The Shoebox project (1961) powered by IBM made it possible to perform arithmetic calculations with the

recognition of sixteen words which are the digits 0 to 9 as well as arithmetic operations such as “plus”, “minus”, or “total”.

We can cite the appearance of many feature extraction algorithms such as the Fast Fourier Transform (FFT) [21], the application of cepstral processing [22] or LPC for speech coding [23].

Another innovation during the 70s is based on models called Hidden Markov Models (HMM). [24, 25] have applied HMM models to various problems.

2.3 Solutions Proposed in the 80s and 90s

The association of HMM models with models named Gaussian Mixture Models (GMM) gained popularity in the 90s. During the same period, experiments were carried out with neural networks capable of predicting phonemes as well as performing other tasks related to speech processing. This constituted an alternative to HMM/GMM. Architectures have been created from neural networks such as the Time-Delay Neural Network (TDNN), models based on convolution networks [26, 27], and Recurrent neural network (RNN) [28].

There has also been the emergence of hybrid HMM/MLP models where the multilayer perceptron network is trained to recognize phonemes and extract a vector representation from them and then link it to the input of the HMM/GMM model [29].

The hybrid models showed performances close to the HMM/GMM models, however, their problem was related to their training speed which was very slow on CPU.

Performance at the time was limited to a layer containing 4000 units which was insufficient for consistent speech recognition.

2.4 Solutions Proposed in the 2000s and 2010s

The decades that followed saw the emergence of graphics cards, which made it possible to develop multi-layered neural networks. In 2009, the performances approached the HMM/GMM models on simple voice recognition tasks [30]. Then, in 2012, the performance of the hybrid systems ended up surpassing the traditional HMM/GMM systems [31, 32].

In 2006, the first works using the CTC loss function were carried out [1] as well as the RNN-Transducer in 2012 [2, 4].

Then appeared the End-To-End approach by rescoring [33] and by recognition [34] with advancements such as beam search [35]. The architecture of encoder-decoders based on the attention mechanism appeared in 2014 [36, 37]. The popular system called Listen, Attend and Spell (LAS) [3] appeared in the same period.

An alternative to RNNs named Transformer [38] started to appear in the speech field in 2018.

3 Recent Solutions Proposed in the Literature

3.1 Monolingual Approach

The last decade has seen the rise of models based on neural networks called 'End-To-End' [3, 39, 11]. Compared to traditional models which were based on HMM/GMM [30] and which required to train several components such as an acoustic model, a lexicon model as well as a language model. These 'End-To-End' models have also taken over the hybrid models [31, 32]. The notable difference of 'End-To-End' models is that they have the ability to take an audio sequence as input and produce a sequence of characters as output thanks to a single neural network.

One of the disadvantages of traditional models is that they require the implementation and training of separate components, making the development of new speech recognition technologies difficult and expensive. Furthermore, expertise is needed for every module that constitutes this model.

Hence the need to train a neural network called "End-To-End" [39] i.e., a single system capable of managing all the features listed earlier. It must also be able to learn a new language from scratch without making major changes to the system.

In 2016, End-To-End models being essentially based on attention mechanisms using RNN networks [39], as well as LSTM networks [3, 11] or both combined [3].

The model named "Listen, Attend and Spell" [3] consists of two sub-modules which are the listener and the speller. The listener encodes the acoustic information in order to be able to listen to the signal. The speller is an attention-based decoder. The Listen function of the listener takes as input a signal x and returns a representation noted h = $h_1, \ldots, h_U$ while the AttendAndSpell function receives data h to output a probability of character sequences noted P(y/x). The Listen function is a bidirectional LSTM or Bidirectional Long Short-Term Memory RNN (BLSTM) having a pyramidal structure. The disadvantage of using a single BLSTM resides in the difficulty of convergence of the model even after a month of training.

To overcome this limitation, the authors of [3] developed a pyramidal BLSTM denoted pBLSTM. In addition to reducing information, the model's neural network will be able to capture non-linear representations of the data. The pyramid structure also participates in reducing the computational complexity and therefore the time required for learning. To train the model, the functions Listen and AttendAndSpell are jointly trained by maximizing the probability of predicting the next character based on the previous one. The decoding phase of an acoustic representation is performed by the beam search algorithm.

Nevertheless, the LAS model produced less efficient results compared to the CLDNN-HMM [40] reference model. The results also showed a performance degradation on short utterances (2 words max) and also on longer utterances.

Another model called Deep Speech 2 capable of voice recognition of English and Mandarin was also presented during the same period by [39]. The architecture of Deep Speech 2 (DS2) is an RNN trained to consume spectrograms representing the audio signal in order to generate transcriptions in textual form. The model is responsible for generating graphemes for each language. The proposed model is an RNN composed

of several hidden layers. The architecture is structured with one or more convolutional layers followed by one or more recurrent layers, followed by one or more dense layers.

The model is trained using the CTC loss function. The benchmark of the model is performed on two datasets which are the Wall Street Journal (WSJ) and LibriSpeech corpus [41].

On WSJ 92, the DS2 model exceeds human performance with a Word Error Rate (WER) of 3.60% compared to 5.03% for humans. On WSJ 93, DS2 is still better with a score of 4.98% compared to 8.08% for the human.

DS2 outperforms human on LibriSpeech test-clean [41] with a score of 5.33% but on the other hand gives lower performance on LibriSpeech test-other [41] with a score of 13.25% compared to human which is at 12.69%.

In 2019, a new End-To-End model named Jasper published by [42] achieved state-of-the-art (SOTA) results on the popular LibriSpeech dataset [41]. Jasper consists of an input convolution layer followed by blocks and then three consecutive convolutional layers. At the output of the model, we find a CTC layer.

A variant of Jasper called Jasper Dense Residual (DR) has also been proposed and which has the particularity of adding the output of each convolution block to the input of all the following blocks.

On the LibriSpeech dataset [41], the Jasper DR 10x5 model was trained on 400 epochs and achieved a SOTA performance on the test-clean subset with a WER score of 2.95%. A newer Jasper-based model named Jasper DR 10x5 + Time/Freq Masks was able to further improve performance with a WER 2.84%.

Authors of [43] focuses on sequence-to-sequence type models called Transformer [38] in order to carry out a comparison study with RNN models.

The experiments resulted in a significant performance improvement to the benefit of the Transformer model on several tasks compared to the RNN. The proposed solution is based on the sequence-to-sequence (S2S) method consisting of an encoder and a decoder. The S2S method is a neural network capable of learning to transform a source sequence denoted X into a target sequence denoted Y.

In the speech translation task, the Transformer improves the baseline RNN BLEU score with a score of 17.2%.

However, the Transformer model [38] shows some limitations in the decoding of filterbank features, namely a longer processing time compared to RNN. The comparative study showed better performance for the Transformer compared to the RNN, especially in the ASR task.

In 2020, [44] trained a model based on the LibriSpeech dataset [41] in conjunction with a model capable of generating pseudo-labeling from unlabeled audio datasets. The unlabeled dataset that was used is named LibriVox resulting in over 53,000 h of audio data. For the labeled dataset, LibriSpeech [41] was used. The Transformer model with GCNN + Transf LM produced as WER 2.09% on test-clean and 4.11% on test-other.

Also in 2020, [12] showed for the first time the principle of learning representations based solely on speech and then a fine-tining phase based on the transcription of speech. This methodology made it possible to surpass the performances obtained through semi-supervised models. The model in question named wav2vec 2.0 is based on the masking

of speech portions through the latent space in order to find the hidden portions by solving a contrastive task. This task is obtained from the quantification of latent representations.

The experiments carried out on the Librispeech annotated dataset [41] made it possible to obtain 1.8%/3.3% as the WER score on the clean/other samples. It only takes 100 times less annotated data for wav2vec 2.0 to surpass the state of the art. With ten minutes of labeled data for a pre-trained model on a 53k hour (unlabeled) dataset to score 4.8%/8.2% WER.

During this same period, [45] published a paper presenting the Conformer model which surpassed the state of the art. The authors demonstrated that the combination between a convolution model (CNN) and a Transformer-based model [38] made it possible to respectively capture local and global dependencies from an audio sequence with efficient parameters.

The release of the Conformer model has significantly improved state-of-the-art performance. The limitation of transformers [38] identified by the authors is the low capacity for extracting local information from the audio signal of speech. While CNNs have an inability to capture global signal information.

The proposed solution to address the limitations identified by the authors is to combine Transformers and CNNs in order to learn information in a local as well as a global context.

With a language model, Conformer in its large version with 118M parameters outperforms the state of the art with 1.9% WER on testclean and 3.9% WER on testother of LibriSpeech [41].

ContextNet presented by [46] is a model that aims to improve the performance of CNN models in the speech recognition task. The main flaw of CNNs is that it is limited in learning the overall context of a given audio signal because it is only able to capture a small window in the time domain. The authors identified this limitation as the reason why CNN-based models perform worse compared to RNN or Transformer-based models.

To enable the CNN model to capture the global context of the signal, the authors introduced a new concept called squeeze-and-excitation (SE) which is introduced in a CNN and constitutes the ContextNet model. By introducing the SE in a convolutional layer, we allow the model to have access to the global information context.

Three different versions of ContextNet have been evaluated on LibriSpeech. The small (alpha $= 0.5$), medium (alpha $= 1$) and large (alpha $= 2$) versions. The large ContextNet(L) model with a Language Model (LM) outperformed the state of the art with 1.9% WER on testclean and 4.1% WER on testother from LibriSpeech.

In [13] the authors explore the field of unsupervised pre-training by combining speech and text within a single model. The proposed solution is called Speech and Language Model (SLAM) which is a model based on Conformer [45] and trained by coupling text and speech with cost functions such as SpanBERT [47] for the textual part and w2v-BERT [48] for the speech part.

For the speech recognition task, the model pre-trained with w2v-bert XL produces results comparable to the state of the art with 1.6% on the test set and 2.9% on test-other of LibriSpeech.

There are currently two major approaches to improve the automatic speech recognition task based on a large number of unlabeled speech. The first strategy is called self-training, also known as pseudo-labeling, which consists in initially training a *Teacher* model from labeled data. Then, the *Teacher* model is used to generate labels on unlabeled data. The combination of labeled and pseudo-labeled data is then used to train a *Student* model. The pseudo-labeling process can be reused several times in order to improve the quality of the *Teacher* model.

The second strategy is to use the method called self-supervised pre-training. This involves pre-training a model from unlabeled data in order to initialize it on a good basis, then fine-tuning it on a labeled dataset.

In the work presented, the authors [48] have implemented a model named w2v-BERT which combines two approaches exploiting self-supervised pre-training as presented in the literature as wav2vec2.0 [12] and BERT [49]. w2v-BERT uses the method called contrastive task taken from wav2vec2.0 in order to obtain a finite number of discriminative voice units. Then, it uses the result obtained in a second method called Masked Language Model (MLM) proposed by BERT for learning contextualized speech representation. w2v-BERT is composed of an encoder called Convolution Subsampling responsible for encoding speech features. A module called contrastive module responsible for discretizing the features encoded in a finite number of discriminative vocal units. And finally, a module called masked prediction module which aims to extract contextualized speech representations.

w2v-BERT XXL outperforms the Conformer [45], HuBERT [50], w2v-Conformer and wav2vec2.0 [12] models by displaying as scores on the test set 1.4% of WER and 2.5% of WER on the test-other of the LibriSpeech dataset. These results make w2v-BERT the actual state-of-the-art model as shown on Table 1. The authors mentioned as a perspective, the evaluation of w2v-BERT in a low-resource environment.

Implementing applications covering spoken languages with few or no resources is a real challenge. [50] propose as a solution a model called Hidden unit BERT (HuBERT) capable of using the clustering approach to generate labels and then train the model in the style of BERT [49]. The objective is to predict hidden units of the acoustic signal. Note that these units are classes generated after clustering a k-means model.

Pre-trained models are based on the wav2vec2.0 architecture.

The results obtained on low resource configurations reveal better scores for the HuBERT X-LARGE model compared to other models in the literature such as wav2vec2.0 and DiscreteBERT [51]. The results of HuBERT are close but less efficient than those of wav2vec2.0 + self-training on LibriSpeech 960h with a WER score of 1.8% on test-clean and 2.9% on test-other.

The approach proposed by [52] is the combination of two learning methods, which are iterative self-training and pre-training. Different versions of models are pre-trained and then trained afterwards in self-training mode. The unlabeled dataset will be used at each stage of the process i.e., used for pre-training and then used again for generating pseudo-labels by the *Teacher* model for training the *Student* model.

The ASR system is based on the Conformer model without the relative positional embedding layer in order to speed up the training process. Four versions of Conformer

were used namely, Conformer L (100M), XL (600M), XXL (1B) and XXL +. Four generations of models have been trained (Gen0 to Gen3).

The best results were obtained with the 4th generation Gen3 Conformer XXL + model which combines pre-training and self-training with 1.4% WER on the test set and 2.6% WER on the test-other from LibriSpeech.

Table 1. Comparison of the Librispeech ASR benchmark

Paper	Year	Base model	Test-clean / Test-other
Deep speech 2 [39]	2016	Bi-RNN + CTC	4.3 / 13.2
ESPnet Transformer [43]	2019	Transformer	2.6 / 5.7
Jasper [42]	2019	CNN + CTC	2.8 / 7.8
SpecAugment [7]	2019	LAS with LM	2.2 / 5.2
Semi-supervised Transformer [44]	2020	Transformer + LM	2.1 / 4.1
wav2vec 2.0 [12]	2020	Transformer + CTC	1.8 / 3.3
Conformer [45]	2020	Conformer	1.9 / 3.9
ContextNet [46]	2020	CNN-RNN-Transducer	1.9 / 4.1
Conformer + NST + SpecAugment [52]	2020	Conformer	1.4 / 2.6
SLAM [13]	2021	Conformer	1.6 / 2.9
HuBERT [50]	2021	Transformer + BERT	1.8 / 2.9
W2v-BERT [48]	**2021**	**wav2vec 2.0 + BERT**	**1.4 / 2.5**

After having compared state of the art on monolingual models, we will present in the next section, models capable of recognizing and transcribing several languages.

3.2 Multilingual Approach

A multilingual version of wave2vec2.0 named XLSR was presented by [53]. The results showed improvements from the multilingual model compared to the monolingual models, i.e., models trained on a single language.

The solution is based on the wave2vec2.0 architecture [12] with audio data from several languages in order to take advantage of the sharing of representations learned through these various languages. The large model built by the authors named XLSR-53 was trained on 53 languages by combining all the BABEL [54], CommonVoice [55] and Multilingual LibriSpeech datasets [56]. The CTC loss function was used in the fine-tuning phase.

Analysis of latent speech representations revealed that the multilingual model has the ability to share common knowledge especially from close languages. The authors [10] highlighted the limitations of unsupervised or self-supervised pre-training methods

that require little or no labeling such as the Wav2Vec2.0 model [12]. Indeed, this model needs to be fine-tuned to be able to adapt to a specific task (e.g., voice recognition). On the other hand, the task of fine-tuning requires very technical skills because of its complexity.

The authors [10] further emphasize the goal of a speech recognition system which must be able to generalize to any domain without requiring fine-tuning tasks on each domain-specific dataset. To address the problems listed above, [10] proposed a model named *Whisper* which was trained by supervised learning on 680,000 h of labeled audio data. The authors were able to demonstrate that models trained on such a large amount of data manage to generalize by zero-shot on any dataset, that is to say without having to resort to fine-tuning on a specific dataset. Note that the dataset includes 97 different languages. The Whisper model is mainly based on the Transformer encoder-decoder architecture [38].

This resulted in a dataset diversified by its data whose distributions are varied as well as environments, languages and actors. To evaluate Whisper, zero-shot technique was used to measure the generalization quality of the model on datasets on which it has never been trained.

- ASR task: Across 14 different datasets, Whisper Large outperforms wave2vec2.0 averaging 12.9% WER versus 29.5% WER. Note that on LibriSpeech test-clean, Whisper displays the same performance as wave2vec2.0, i.e., 2.7% of WER;
- Multilingual speech recognition task: Zero-Shot Whisper outperforms state-of-the-art models on the MLS corpus with 8.1% WER, but does much less than XLSR and mSLAM-CTC on VoxPopuli [57];

To build a model capable of learning multilingual and multimodal representations from hundreds of different languages, [58] proposed an improvement of the SLAM model [13]. The proposed model named mSLAM has the particularity of learning several languages compared to its predecessor SLAM. To avoid interference between modalities, the mSLAM model was trained with the CTC loss function on the speech-text pair of a dataset. The model is based on the SLAM architecture [13] by combining, together, pre-training based on unlabeled speech dataset with w2v-BERT [48] and unlabeled text with spanBERT[47]. The Translation Language Modeling (TLM) method was used on the labeled dataset coupling speech and transcription. The TLM method introduced by [59] aims to improve cross-language pre-training by extending to the Masked Language Modeling (MLM) method, which is more suitable for monolingual texts. TLM will concatenate the sentences of a source language and a target language in parallel and then randomly hide words in each language. To predict the hidden word in a specific language, the model will be able to rely on the representation from the parallel language.

One of the key challenges for Google is to be able to extend voice technologies to several languages. This results in enough data to be able to train high quality models. The task of manually labeling data from low-resource languages is a big challenge for supervised learning both in terms of time and cost. The objective of this study [60] is to manage to produce in the long term, a universal model of voice recognition capable of covering all the languages spoken in the world.

The approach used by the authors [60] is mainly based on the construction of a model called Universal Speech Models (USMs) trained on large datasets of three types

which are unlabeled audio data, unlabeled textual data and data pairs (audio-text). The construction of the model is based on three steps such as unsupervised pre-training, a supervised pre-training with multiple objectives and supervised training on specific tasks such as ASR and AST. The solution is based on the large model named Conformer [45] with 2 billion parameters. The USM model and its variants outperform Whisper [10] on all multilingual datasets. USM also outperforms Whisper on all low-resource languages in the FLEURS dataset [61].

Recent scientific studies have focused on expanding the language coverage of speech processing technologies. However, current technologies are limited to the recognition of a hundred languages which is very low compared to more than 7000 languages spoken in the world. To address this problem, the authors of the article [62] constructed a new dataset comprising 1107 labeled languages and another dataset comprising 3809 languages in unlabeled audio format. A multilingual model called Massively Multilingual Speech (MMS) was created for the occasion from the new datasets to cover the recognition and speech synthesis of 1107 languages, as well as the identification of 4017 languages. The architecture of the MMS model is based on wav2vec2.0 [12].

The experiments are based on the pre-trained wav2vec2.0 model on several languages implemented with the fairseq tool [63]. The MMS model was finetuned on 61 languages in the FLEURS [61] dataset and was compared to the XLS-R model [53]. MMS outperforms XLS-R [53] on most languages, especially low-resource languages. On the ASR task, MMS was finetuned on the MMS-lab labeled dataset using the CTC criterion. The model was first compared to Whisper [10] and the results showed better performance for MMS in 31 out of 54 languages (FLEUR-54 dataset [61]). Regarding Google-USM [60], MMS slightly outperforms USM. Table 2 compares state-of-the-art multilingual models.

Table 3 compares number of supported languages by each model.

Table 2. Comparison on Multilingual LibriSpeech, VoxPopuli and FLEURS benchmark. All results are reported with the WER score except for the FLEURS dataset which is in Character Error Rate (CER).

Model	Year	MLS	VoxPopuli	FLEURS
MMS [62]	2023	8.7	10.3	6.2
USM [60]	2023	-	-	6.5
Whisper [10]	2022	8.1	15.2	-
XLSR [53]	2020	10.9	10.6	-
mSLAM [58]	2022	9.7	9.1	-

3.3 Automatic Speech Recognition for Low-Resource Languages

This article [15] presents the steps that were necessary to collect data from four sub-Saharan African languages such as Swahili, Hausa, Amharic and Wolof. The authors

Table 3. Number of languages supported

Model	Supported languages
MMS [62]	1162
USM [60]	102
Whisper [10]	99
XLSR [53]	53
mSLAM [58]	51

specifically focused on Wolof by setting up the very first voice recognition system for this low-resource West African language. Access to technologies is mainly done through mobile phones or smartphones. Speech technology is an essential mean to reduce the gap for illiterate people, both in the field of health, food and in social networks. It is in this sense that the project named ALFFA was initiated in order to develop ASR and Text-To-Speech (TTS) technologies for African languages. Note that the ALFFA project brought together profiles in various fields such as experts in technology and also in linguistics. Wolof is spoken in Senegal, Gambia and Mauritania. This language is spoken by more than 10 million people. Kaldi software was used to build the speech recognition system. The ASR for Swahili was trained over ten hours of audio and evaluated over 1.8 h. The language model contains approximately 28 million words. For Hausa, the model was trained on seven hours of data and tested on one hour. The language model has a size of 41000 words. The Amharic meanwhile was trained on twenty hours of data and tested on two hours. The language model is created using 3-g and the text has been segmented into morphemes. The studies in this article are mainly focused on the Wolof spoken in Dakar the capital of Senegal.

The construction of the audio corpus in Wolof was carried out by selecting 6000 utterances at random. Eighteen people (ten men and eight women) were selected from different socio-professional categories. These eighteen people were responsible for making voice recordings of 1000 utterances extracted from the 6000 mentioned earlier. The age of the people varies from 24 to 48 years old and the microphone used for the recording is a Samson G-track in an environment containing no noise.

A total of 18,000 utterances were recorded, representing 21 h and 22 min of audio signal. Two language models were set up and the tool named Phonetisaurus was used to transform the vocabulary into phonemes. Kaldi was used to build the ASR model. Three acoustic models have been constructed, namely HMM/GMM which is a hidden Markovian model coupled with a Gaussian model. Another model named SGMM or subspace Gaussian mixture model and a last one based on DNN deep neural networks. For the HMM/GMM model, it receives as input 133 MFCCs and 16.8 h of data. The DNN model was trained with the same MFCCs used for the GMM models. The DNN was fine-tuned with the Stochastic Gradient Descent (SGD).

The performances of the different models were around 30% WER. The best WER score of 27.21% was obtained on the LM2 language model with the DNN model. The limits are essentially on so-called diatric data which refers to a word which can take

several forms. For example, the word “jél” which can also be written “jël” and which means “to take”. One track mentioned in the paper would be to standardize the text by choosing only one form for all the words referring to several representations.

This article [64] studies the technique of learning transfer to a low-resource language such as Amharic. The transfer of learning took place through two high-resource languages (English and Mandarin) towards the target, which is Amharic. The results showed a strong reduction in the WER error rate. The best result was obtained from the English language.

A TDNN (Time Delay Deep Neural Network) neural network is proposed as a baseline. The alignment of text and voice data is performed by a GMM-HMM model before being injected into a neural network.

Amharic corpus contains 20 h of training data and 2 h of test data from 100 native speakers. English corpus containing 100 h for English-1 set and 460 h for English-2 set. 5.4 h are dedicated to the test set They are sourced from OpenSLR[1]. Mandarin corpus containing 178 h including 85% dedicated to the train set and the rest for the test set. It is sourced from Beijing Shell Technology.

The best results were obtained with English-2 as the source language with a WER score of 25.5% compared to the baseline which is 38.72%.

The authors [65] carried out experiments on the contribution of multilingual models on low-resource languages. Four low-resource Ethiopian languages as target were considered such as Amharic, Oromo, Tigrana and Wolaytta. What motivated the development of multilingual systems is justified by the number of high-resource languages which are not widespread and which require colossal means for their development. The idea is to take advantage of these languages in order to develop voice recognition models requiring few resources. The objective of this approach is to enable illiterate populations to benefit from being able to access digital services, mainly in rural areas, through their phones. The proposed solution is the implementation of multilingual models based on a corpus of more than 20 languages in order to build an ASR system for four Ethiopian languages.

The neural network is based on the Factored Time Delay Neural Networks with additional Convolutional layers (CNN-TDNNf) model. To train this model, the data is first aligned (voice and text) using an HMM-GMM model. The source dataset used is the GlobalPhone (GP) [66] which is a multilingual corpus covering 20 languages. For the Amharic language we have AMH2005 containing 20 h of training with 11k utterances. AMH2020 is used as the second corpus. Note that Amharic data was collected in Ethiopia as well as for Oromo, Tigrana and Wolaytta. All models were built using the Kaldi ASR toolkit. Tri-gram language models have been developed for each target language. The 1st experiment carried out led to the training of a multilingual ML22 model on the GP corpus. The 2nd named ML23 (22 languages + 1 Ethiopian) was trained by adding an Ethiopian language in the train data. The 3rd named ML25 (22 languages + 3 Ethiopian) was trained by adding three Ethiopian languages. The last experiment named ML26 (22 languages + 4 Ethiopian) was trained by adding four Ethiopian languages. Two models from transfer learning were built and named ML22_Ada and ML26_Ada.

[1] http://www.openslr.org/.

The best multilingual model on the Amharic language was obtained with ML22_Ada with 8.21% of WER. On Tigrina we have the MonoLing model which scores 16.82% WER. On Oromo, ML22_Ada wins with 32% WER. Finally, on Wolaytta the MonoLing model produces better results with 23.23% WER. We can conclude that training a model whose source and target language are close enough produces good results.

In this article the author [17] presents the creation of two datasets, one of which is labeled for ten African languages and the other for four African languages. The contribution of this paper is intended to reduce the digital divide within illiterate populations located in Sub-Saharan Africa. The work presents the exploitation of resources from radio archives for low-resource languages and the use of self-supervised methods based on wav2vec [67] to train models capable of solving speech recognition tasks intended for these languages. The Architecture of the West African wav2vec (WAwav2vec) model is based on wav2vec which is used for the feature extraction part.

On the Multilingual Speech Recognition task which consists in identifying 105 classes of utterances on four languages, WAwav2vec and wav2vec give similar results. The virtual assistant only understands contact management vocabulary, it would be possible to extend it in the field of micro-finance, agriculture and education. It would be useful to take advantage of the abundance of radio data to train the encoder to identify other nearby languages.

Much effort has gone into developing ASR models for so-called "high-resource" languages such as English, Mandarin and Japanese. However, the so-called "low-resource" models are much less robust due to an insufficient number of training data. The other point raised by the authors [9] is related to the multilingual models which have been tested in the literature and which produce limited results due to the non-proximity of the languages chosen for training.

To provide a solution to the points mentioned above, the authors have set up a multilingual model dedicated to low-resource Turkish languages. Ten Turkish languages were considered in the study such as: Azerbaijani, Bashkir, Chuvash, Kazakh, Kyrgyz, Sakha, Tatar, Turkish, Uyghur, and Uzbek. The models are based on a dataset created by the authors containing 218 h of speech transcribed into the Turkish language. The models are based on the Conformer model architecture.

The dataset created by the authors is called Turkish Speech Corpus (TSC) which is composed of 218 h of speech in Turkish. CVC includes ten different Turkish languages and one language in English. Russian language comes from OpenSTT. Data augmentation techniques were used such as speed perturbation and spectral augmentation. For the experiments, 22 models were trained in total (13 monolingual and 9 multilingual). All models were trained in Pytorch with the ESPnet framework. The monolingual models were trained with 1 NVIDIA DGX A100 GPU card (40 GB), the multilingual ones required four GPUs of the same type. The language model used is based on Transformer.

We can observe as a result that the multilingual models outperform the monolingual ones. The model trained only on Turkish languages (all_turkic) also outperforms those trained on non-Turkish languages (English, Russian). This proves that the proximity of the language has a positive impact on the model.

The authors [14] made a contribution to the modeling of the ASR task for low-resource Congolese languages. Would it be possible to surpass the state of the art with

the wav2vec2.0 model on the Lingala? Would pre-training on large datasets be a solution to reduce the amount of data to label on a low-resource language? Would it be possible to benefit from the proximity of languages in order to develop efficient multilingual models?

To answer these questions, the authors proposed various solutions such as the creation of two datasets named *Congolese Speech Radio Corpus* and *Linguala Read Speech Corpus*, establishment of a baseline from the collected data and creation of a multilingual model from four languages spoken mainly in Congo.

The developed solution is based on wav2vec2.0 [12].

Two baselines were considered for the experiments such as GMM-HMM (T-DNN) and DeepSpeech 2 [39].Whisper medium [10] model was also used. For self-supervised experimentation, the BASE model wav2vec2.0 was used. The 1st experiment consisted in pre-training wav2vec2.0 on the Congolese Speech Radio dataset. For the second experiment wav2vec2.0 was finetuned in two scenarios: fine-tuning of the pre-trained model on four Congolese languages on Linguala Read Speech Corpus and fine-tuning of XLSR-53 and XLS-R multilingual models respectively on Linguala Read Speech Corpus. The tools used are HuggingFace and Fairseq. The best model is the multilingual version CdWav2Vec pre-trained on Congolese Speech Radio corpus and fine-tuned on Linguala Read Speech Corpus which obtains a WER score of 21.4%. The results prove that the pre-trained multilingual models show better performance than the monolingual ones. This is due to the richness of the representation of close languages which is transferable to target languages close to the source languages.

4 Conclusion

We reviewed the history of automatic speech recognition starting from traditional systems requiring the implementation of several models such as the acoustic model, then the phonetic model and finally the linguistic model. We have seen that these systems have limitations that so-called "*End-to-End"* systems try to simplify thanks to the development of neural networks. Nowadays, state-of-the-art models tend to offer multilingual approaches in order to model the largest possible number of spoken languages which is around 7000 [68] and, to allow illiterate people to be able to interact with all kinds of application from their native languages. Our future work will mainly focus on African languages, considered as low-resources due to the scarcity of their digitized data in order to allow people speaking African languages to be able to interact with digital services specifically related to financial sector.

Acknowledgements. This work is part of the ongoing PhD training supported by the Partnership for skills in Applied Sciences, Engineering and Technology (PASET) - Regional Scholarship and Innovation Fund (RSIF).

References

1. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: ACM International Conference Proceeding Series (2006)

2. Graves, A.: Sequence Transduction with Recurrent Neural Networks (2012)
3. Chan, W., Jaitly, N., Le, Q.V., Vinyals, O.: Listen, Attend and Spell (2015). https://doi.org/10.48550/arxiv.1508.01211
4. Graves, A., Mohamed, A.R., Hinton, G.: Speech recognition with deep recurrent neural networks. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings (2013)
5. Sriram, A., Auli, M., Baevski, A.: Wav2Vec-Aug: Improved self-supervised training with limited data (2022). https://doi.org/10.48550/arxiv.2206.13654
6. Park, D.S., Chan, W., Zhang, Y., et al.: SpecAugment: a Simple Data Augmentation Method for Automatic Speech Recognition (2019). https://doi.org/10.21437/interspeech.2019-2680
7. Park, D.S., Zhang, Y., Chiu, C.-C., et al.: SpecAugment on Large Scale Datasets (2019). https://doi.org/10.48550/arxiv.1912.05533
8. Yi, C., Wang, J., Cheng, N., et al.: Applying Wav2vec2.0 to Speech Recognition in Various Low-resource Languages (2020). https://doi.org/10.48550/arxiv.2012.12121
9. Mussakhojayeva, S., Dauletbek, K., Yeshpanov, R., Varol, H.A.: Multilingual speech recognition for Turkic languages. Information **14**, 74 (2023). https://doi.org/10.3390/info14020074
10. Radford, A., Wook Kim, J., Xu, T., et al.: Robust Speech Recognition via Large-Scale Weak Supervision (2022). https://cdn.openai.com/papers/whisper.pdf. Accessed 23 Sep 2022
11. Kim, S., Hori, T., Watanabe, S.: Joint CTC-Attention based End-to-End Speech Recognition using Multi-task Learning (2016). https://doi.org/10.48550/arxiv.1609.06773
12. Baevski, A., Zhou, H., Mohamed, A., Auli, M.: wav2vec 2.0: a Framework for Self-Supervised Learning of Speech Representations (2020). https://doi.org/10.48550/arxiv.2006.11477
13. Bapna, A., Chung, Y., Wu, N., et al.: SLAM: a Unified Encoder for Speech and Language Modeling via Speech-Text Joint Pre-Training (2021)
14. Kimanuka, U., wa Maina, C., Büyük, O.: Speech recognition datasets for low-resource congolese languages. In: 4th Workshop on African Natural Language Processing (2023)
15. Gauthier, E., Besacier, L., Voisin, S., et al.: Collecting resources in Sub-Saharan African languages for automatic speech recognition: a case study of Wolof. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Portorož, Slovenia, pp. 3863–3867 (2016)
16. Gauthier, E., Séga Wade, P., Moudenc, T., et al.: Preuve de concept d'un bot vocal dialoguant en wolof (Proof-of-Concept of a Voicebot Speaking Wolof). In: Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles, Volume 1 : conférence principale. ATALA, Avignon, France, pp. 403–412 (2022)
17. Doumbouya, M., Einstein, L., Piech, C.: Using Radio Archives for Low-Resource Speech Recognition: Towards an Intelligent Virtual Assistant for Illiterate Users (2021). https://doi.org/10.48550/arxiv.2104.13083
18. Davis, K.H., Biddulph, R., Balashek, S.: Automatic recognition of spoken digits. J. Acoust. Soc. Am. **24** (1952). https://doi.org/10.1121/1.1906946
19. Fry, D.B.: Theoretical aspects of mechanical speech recognition. J. Br. Inst. Radio Eng. **19** (1959). https://doi.org/10.1049/jbire.1959.0026
20. Denes, P.: The design and operation of the mechanical speech recognizer at University College London. J. Br. Inst. Radio Eng. **19** (1959). https://doi.org/10.1049/jbire.1959.0027
21. Cooley, J.W., Tukey, J.W.: An algorithm for the machine calculation of complex Fourier series. Math. Comput. **19** (1965). https://doi.org/10.1090/s0025-5718-1965-0178586-1
22. Oppenheim, A.V., Schafer, R.W., Stockham, T.G.: Nonlinear filtering of multiplied and convolved signals. Proc. IEEE **56** (1968). https://doi.org/10.1109/PROC.1968.6570
23. Atal, B.S., Hanauer, S.L.: Speech analysis and synthesis by linear prediction of the speech wave. J Acoust. Soc. Am. **50** (1971). https://doi.org/10.1121/1.1912679

24. Baum, L.E., Petrie, T.: Statistical inference for probabilistic functions of finite state Markov chains. Ann. Math. Stat. **37** (1966). https://doi.org/10.1214/aoms/1177699147
25. Baum, L.B., Eagon, J.A.: An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. Bull. Am. Math. Soc. **73** (1967). https://doi.org/10.1090/S0002-9904-1967-11751-8
26. Waibel, A., Hanazawa, T., Hinton, G., et al.: Phoneme recognition using time-delay neural networks. IEEE Trans. Acoust. **37** (1989). https://doi.org/10.1109/29.21701
27. Lang, K.J., Waibel, A.H., Hinton, G.E.: A time-delay neural network architecture for isolated word recognition. Neural Netw. **3** (1990). https://doi.org/10.1016/0893-6080(90)90044-L
28. Robinson, T., Fallside, F.: A recurrent error propagation network speech recognition system. Comput. Speech Lang. **5** (1991). https://doi.org/10.1016/0885-2308(91)90010-N
29. Morgan, N., Bourlard, H.: Continuous speech recognition using multilayer perceptrons with hidden Markov models. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings (1990)
30. Mohamed, A.-R., Dahl, G., Hinton, G.: Deep belief networks for phone recognition. Scholarpedia **4** (2009). https://doi.org/10.4249/scholarpedia.5947
31. Jaitly, N., Nguyen, P., Senior, A., Vanhoucke, V.: Application of pretrained deep neural networks to large vocabulary speech recognition. In: 13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012 (2012)
32. Dahl, G.E., Yu, D., Deng, L., Acero, A.: Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. **20** (2012). https://doi.org/10.1109/TASL.2011.2134090
33. Graves, A., Jaitly, N.: Towards end-to-end speech recognition with recurrent neural networks. In: 31st International Conference on Machine Learning, ICML 2014 (2014)
34. Maas, A.L., Xie, Z., Jurafsky, D., Ng, A.Y.: Lexicon-free conversational speech recognition with neural networks. In: NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference (2015)
35. Hannun, A.Y., Maas, A.L., Jurafsky, D., Ng, A.Y.: First-Pass Large Vocabulary Continuous Speech Recognition using Bi-Directional Recurrent DNNs (2014)
36. Chorowski, J., Bahdanau, D., Cho, K., Bengio, Y.: End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results (2014)
37. Bahdanau, D., Chorowski, J., Serdyuk, D., et al.: End-to-End Attention-based Large Vocabulary Speech Recognition. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings 2016 May, pp. 4945–4949 (2015). https://doi.org/10.1109/ICASSP.2016.7472618
38. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. 5999–6009, December 2017)
39. Amodei, D., Ananthanarayanan, S., Anubhai, R., et al.: Deep speech 2: end-to-end speech recognition in English and Mandarin. In: 33rd International Conference on Machine Learning, ICML 2016, pp. 312–321 (2016)
40. Sainath, T.N., Vinyals, O., Senior, A., Sak, H.: Convolutional, long short-term memory, fully connected deep neural networks. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings, pp. 4580–4584, August 2015. https://doi.org/10.1109/ICASSP.2015.7178838
41. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an ASR corpus based on public domain audio books. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings, pp. 5206–5210, August 2015. https://doi.org/10.1109/ICASSP.2015.7178964
42. Li, J., Lavrukhin, V., Ginsburg, B., et al.: Jasper: an End-to-End Convolutional Neural Acoustic Model (2019). https://doi.org/10.48550/arxiv.1904.03288

43. Karita, S., Chen, N., Hayashi, T., et al.: A Comparative Study on Transformer Vs RNN in Speech Applications
44. Synnaeve, G., Xu, Q., Kahn, J., et al.: End-to-end ASR: from supervised to semi-supervised learning with modern architectures a preprint (2020)
45. Gulati, A., Qin, J., Chiu, C.-C., et al.: Conformer: convolution-augmented transformer for speech recognition (2020). https://doi.org/10.48550/arxiv.2005.08100
46. Han, W., Zhang, Z., Zhang, Y., et al.: ContextNet: improving convolutional neural networks for automatic speech recognition with global context (2020). https://doi.org/10.48550/arxiv.2005.03191
47. Joshi, M., Chen, D., Liu, Y., et al.: SpanBERT: improving pre-training by representing and predicting spans. Trans. Assoc. Comput. Linguist. **8**, 64–77 (2019). https://doi.org/10.1162/tacl_a_00300
48. Chung, Y.A., Zhang, Y., Han, W., et al.: W2v-BERT: combining contrastive learning and masked language modeling for self-supervised speech pre-training. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021 – Proceedings, pp. 244–250 (2021). https://doi.org/10.1109/ASRU51503.2021.9688253
49. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, vol. 1, pp. 4171–4186 (2018)
50. Hsu, W.N., Bolte, B., Tsai, Y.H.H., et al.: HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 3451–3460 (2021). https://doi.org/10.1109/TASLP.2021.3122291
51. Baevski, A., Auli, M., Mohamed, A.: Effectiveness of self-supervised pre-training for speech recognition (2019)
52. Zhang, Y., Qin, J., Park, D.S., et al.: Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition (2020)
53. Conneau, A., Baevski, A., Collobert, R., et al.: Unsupervised cross-lingual representation learning for speech recognition (2020). https://doi.org/10.48550/arxiv.2006.13979
54. Gales, M., Knill, K., Ragni, A., Rath, S.: Speech recognition and keyword spotting for low-resource languages: Babel project research at CUED (2014)
55. Ardila, R., Branson, M., Davis, K., et al.: Common Voice: a massively-multilingual speech corpus. In: LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings, pp. 4218–4222 (2019)
56. Pratap, V., Xu, Q., Sriram, A., et al.: MLS: a large-scale multilingual dataset for speech research. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2020, pp. 2757–2761, October 2020 https://doi.org/10.21437/Interspeech.2020-2826
57. Wang, C., Rivière, M., Lee, A., et al.: VoxPopuli: a large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In: ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference, pp. 993–1003 (2021). https://doi.org/10.18653/v1/2021.acl-long.80
58. Bapna, A., Cherry, C., Zhang, Y., et al.: mSLAM: Massively multilingual joint pre-training for speech and text (2022)
59. Conneau, A., Lample, G.: Cross-lingual language model pretraining. Adv. Neural Inf. Process. Syst. **32** (2019)
60. Zhang, Y., Han, W., Qin, J., et al.: Google USM: scaling automatic speech recognition beyond 100 languages (2023)

61. Conneau, A., Ma, M., Khanuja, S., et al.: FLEURS: few-shot learning evaluation of universal representations of speech. In: 2022 IEEE Spoken Language Technology Workshop, SLT 2022 – Proceedings, pp. 798–805 (2022). https://doi.org/10.1109/SLT54892.2023.10023141
62. Pratap, V., Tjandra, A., Shi, B., et al.: Scaling speech technology to 1,000+ languages (2023)
63. Ott, M., Edunov, S., Baevski, A., et al.: fairseq: a fast, extensible toolkit for sequence modeling (2019). https://doi.org/10.48550/arxiv.1904.01038
64. Woldemariam, Y.: Transfer learning for less-resourced semitic languages speech recognition: the case of Amharic. In: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL). European Language Resources association, Marseille, France, pp. 61–69 (2020)
65. Tachbelie, M.Y., Abate, S.T., Schultz, T.: Development of multilingual ASR using GlobalPhone for less-resourced languages: the case of Ethiopian languages. In: INTERSPEECH (2020)
66. Schultz, T., Vu, N.T., Schlippe, T.: GlobalPhone: a multilingual text & speech database in 20 languages. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings, pp. 8126–8130 (2013). https://doi.org/10.1109/ICASSP.2013.6639248
67. Schneider, S., Baevski, A., Collobert, R., Auli, M.: wav2vec: unsupervised pre-training for speech recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2019, pp. 3465–3469, September 2019. https://doi.org/10.21437/Interspeech.2019-1873
68. Ethnologue | Languages of the world. https://www.ethnologue.com/. Accessed 20 Aug 2023