# A Survey on Multilingual End-to-End Automatic Continuous Speech Recognition Using Deep Neural Network

**[1]Jose Stephen, [2]Dr.Jeyakumar MK**

[1]Research Scholar, [2]Professor
Computer Science and Engineering,
Noorul Islam Centre for Higher Education
Tamil Nadu.

***Abstract*- Automatic Speech Recognition (ASR) systems have witnessed significant advancements in recent years, fueled by the development of deep neural networks and large-scale datasets. Multilingual end-to-end ASR, which aims to recognize speech in multiple languages using a unified framework, has gained substantial attention due to its potential applications in various domains. This article thoroughly analyses the design of deep neural frameworks for end-to-end, multilingual automatic continuous voice recognition. It offers an understanding of the latest techniques and possible directions for future research by examining the key components, challenges, and recent advancements within this domain through the lens of Wav2Vec.Wav2vec-U increases unsupervised recognition outcomes across a variety of languages, according to experiments, despite having a simpler conceptual framework.**

***Keywords*- Speech Recognition, Multilingual, Wav2Vec, Neural Networks**

## I. INTRODUCTION

The rapid development of deep neural networks and the accessibility of big voice datasets have been key drivers of Automatic Voice Recognition's (ASR) recent substantial improvements. Voice assistants, language translation, transcription services, and accessibility tools for those with speech impairments are just a few of the many fields where ASR technology has found extensive use. Traditionally, ASR systems were developed using modular approaches, involving separate components for acoustic modelling, pronunciation modelling, and language modelling [1]. However, these approaches often suffer from performance limitations and lack the ability to handle multiple languages seamlessly.The general block diagram for automatic voice recognition is shown in Figure 1.

Multilingual end-to-end ASR, on the other hand, aims to overcome the limitations of traditional ASR systems by recognizing speech in multiple languages using a unified framework [2]. This approach leverages the power of deep neural networks to directly map acoustic features to text without relying on intermediate representations. By formulating a deep neural framework for multilingual end-to-end ASR, it becomes possible to build systems that are capable of understanding and transcribing speech in multiple languages, catering to the demands of a multilingual and globally connected world.

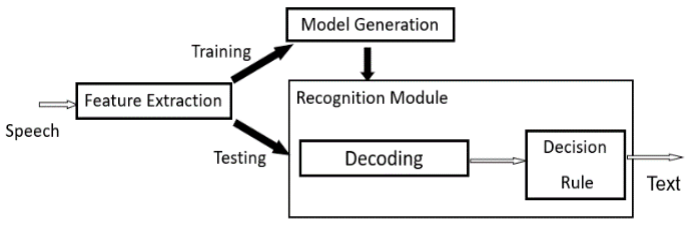


FIGURE1.Block Diagram for ASR

The research thoroughly analyses the design of deep neural frameworks for multilingual end-to-end continuous automatic speech recognition.To provide a glimpse into advanced techniques and emphasize potential directions for future research using the novel approach, itexplores the fundamental components, challenges, and contemporary advancements within this field.

Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformer-based models are just a few of the deep neural network subsystems that are employed in multilingual ASR. It will explore techniques for creating models that can handle multiple languages in audio processing, focusing especially on transfer learning, language-independent representations, and approaches suited for languages with limited resources. Furthermore, it will delve into the realms of multilingual language modeling and acoustic-phonetic modeling, exploring ways to effectively deal with various phonetic elements and harnessing knowledge from different languages.Furthermore, this review paper will address crucial training and data considerations, including multilingual datasets, data augmentation techniques, and strategies to tackle language imbalance. It will also discuss evaluation metrics and benchmarking initiatives specific to multilingual ASR to facilitate fair comparison and progress measurement in the field.
To provide a comprehensive overview, it will highlight recent advancements and state-of-the-art models in multilingual end-to-end ASR. This includes exploring self-supervised and unsupervised learning approaches, multitask learning, and joint optimization techniques. It will also discuss various applications of multilingual ASR, such as low-resource language recognition, code-switching, automatic translation, and voice assistant systems.
The remainder of Section 2 will describe Multilingual Speech Recognition in detail, as will Section 3's discussion of the various methods that have been proposed by researchers, Section 4's methodology with a novel approach, Section 5's performance analysis, and Section 6's conclusion.

## II. MULTILINGUAL SPEECH RECOGNITION

Multilingual Speech Recognition (MSR) is a rapidly evolving field within Automatic Speech Recognition (ASR) that focuses on the development of systems capable of understanding and transcribing speech in multiple languages. With the increasing demand for multilingual applications and services in our globally interconnected world, MSR plays a crucial role in breaking language barriers and enabling effective communication across diverse linguistic communities [3].

Traditional ASR systems typically rely on language-specific models and pipelines, making it challenging to handle multiple languages seamlessly. However, MSR aims to address this limitation by developing unified frameworks that can process and recognize speech in multiple languages using a single system. This not only simplifies the development process but also enhances the flexibility and scalability of ASR technology.

The formulation of MSR systems involves tackling several unique challenges. Firstly, languages vary significantly in terms of phonetic and phonological characteristics, vocabulary sizes, and grammatical structures. These variations necessitate the design of models and algorithms that can handle diverse linguistic inputs effectively. Secondly, MSR systems need to address language-specific acoustic variations, including accents, dialects, and speaking styles, to ensure accurate and robust recognition across languages.

Deep Neural Networks (DNNs) have emerged as the dominant modelling approach in MSR due to their ability to capture complex patterns in speech data. DNNs can be leveraged for various components of MSR systems, including acoustic modelling, pronunciation modelling, and language modelling. By formulating deep neural frameworks specifically tailored for MSR, researchers can exploit the power of neural networks to build highly accurate and versatile multilingual speech recognition systems [4].

The objectives of this paper are to provide a comprehensive overview of the formulation of MSR systems and to highlight the key components, challenges, recent advancements, and future directions in the field.It will explore the following aspects of MSR:

### *A. Multilingual Acoustic Modeling:*

- Techniques for handling language-specific acoustic variations.
- Transfer learning and adaptation methods to leverage knowledge from resource-rich languages to low-resource languages.
- Language-agnostic feature representations to capture universal acoustic characteristics.

### *B. Multilingual Language Modeling:*

- Shared and separate language models to handle multilingual input.
- Cross-lingual transfer learning and adaptation approaches.
- Efficient handling of large vocabulary sizes and language-specific grammatical structures.

### *C. Multilingual Pronunciation Modeling:*

- Techniques for handling diverse phonetic units across languages.
- Cross-lingual phonetic embeddings for capturing language similarities and differences.

*D.* ***Training and Data Considerations:***

- Multilingual training datasets and their characteristics.
- Data augmentation and synthesis techniques to enhance training data diversity.
- Strategies for addressing language imbalance and low-resource scenarios.

*E.* ***Evaluation Metrics and Benchmarking:***

- Evaluation measures that are specific to MSR systems.
- Benchmark datasets and evaluation campaigns to facilitate fair comparison.

*F.* ***Recent Advancements and Applications****:*

- State-of-the-art MSR models and techniques.
- Multilingual ASR for low-resource languages, code-switching scenarios, and translation applications.
- Multilingual ASR in real-world applications such as voice assistants, transcription services, and language translation platforms.

By exploring these topics, this paper aims to provide researchers and practitioners in the field of MSR with a comprehensive understanding of the current state-of-the-art, challenges, and potential avenues for future research. MSR technology holds tremendous potential for enabling seamless communication and information access across languages, and this review paper serves as a valuable resource to advance the development of robust and effective multilingual speech recognition systems. Fig 2 shows the general block diagram of multilingual speech recognition.

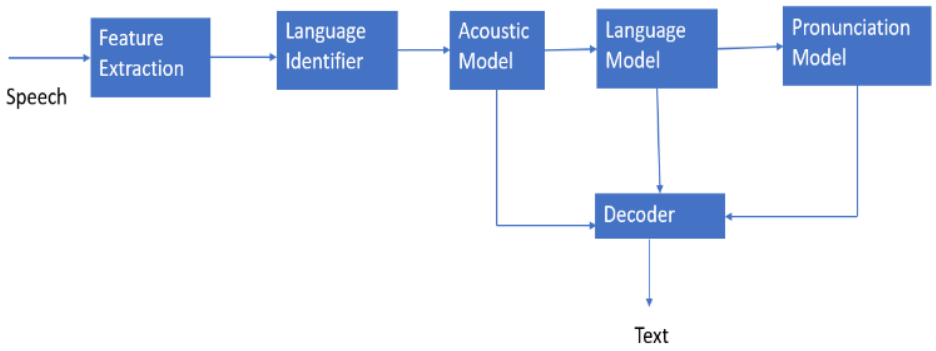


FIGURE 2. Block Diagram for Multilingual Speech Recognition

Deep Neural Networks (DNNs) are widely used in multilingual Automatic Speech Recognition (ASR) systems. These networks can effectively model the complex relationships between acoustic features and linguistic units such as phonemes or words. Here are the various components of DNNs employed in multilingual ASR, including Convolutional Neural Networks (CNNs) [5], Recurrent Neural Networks (RNNs) [6], and Transformer-based models [7].

***1. Convolutional Neural Networks (CNNs):***

- CNNs are commonly used for front-end acoustic modelling in ASR systems.
- They are effective in capturing local patterns in time-frequency representations of speech signals.
- CNN layers can learn hierarchical representations of acoustic features, extracting low-level features and progressively combining them into higher-level representations.
- CNNs are often used as feature extractors in conjunction with other neural network architectures.

***2. Recurrent Neural Networks (RNNs):***

- RNNs are frequently used in ASR for modeling the temporal dependencies in speech signals.
- Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are popular RNN variants.
- RNNs process sequential data by maintaining an internal state that allows them to remember past information.
- RNNs are well-suited for capturing long-term dependencies in speech, making them useful for language modeling and acoustic modeling tasks in ASR.

***3. Transformer-based Models:***

- Transformer-based models have gained significant attention in ASR due to their effectiveness in modeling long-range dependencies and parallelization.
- To identify global dependencies in the input sequence, transformers employ self-attention processes.
- By allowing each location in the sequence to pay attention to every other position, the self-attention mechanism enables the model to learn contextual representations.

- Transformers are made up of an encoder-decoder architecture, in which the encoder processes the speech features that are supplied and the decoder produces the transcriptions.
- Transformer models, such as the popular Transformer and its variants like the Conformer, have achieved state-of-the-art performance in multilingual ASR.

***4. Language Modeling:***

- Language modeling is an essential component of ASR systems, and it can be implemented using various neural network architectures.
- Neural language models, such as RNN-based models (LSTM, GRU) and Transformer-based models, are trained on large text corpora to learn the statistical properties of language.
- These language models help in predicting the next word or phoneme given the previous context, aiding in the transcription process and improving the overall ASR performance.

Overall, a combination of CNNs, RNNs, and Transformer-based models, along with language modeling techniques, is commonly employed in multilingual ASR systems to handle the complexity of speech signals and capture the linguistic properties across different languages.

## III. LITERATURE REVIEW

One sequence-to-sequence ASR model with extremely little script overlap was proven by Shubham Toshniwal et al.[8] using training data from nine different Indian languages.A grapheme-based sequence-to-sequence model was simultaneously trained on data from all languages using a union of language-specific grapheme sets. This model beats comparable sequence-to-sequence models trained on each language separately by 21% relative, despite not being explicitly given any knowledge about language identity, according to the researchers. They increased performance by an additional 7% relative, and they removed linguistic ambiguity, by changing the model to take language identification as an additional input feature.

Mari Ganesh Kumar and others [9] suggested two end-to-end (E2E) ASR systems. In the first system, the native language script was recovered using a brand-new data-driven back-end, and the E2E model was trained on the CLS representation.They proposed changing the E2E model for the second system to allow for concurrent training of the CLS representation and native language characters. The results of the multilingual and code-switching challenges for the Indic ASR Challenge 2021. The top results for the multilingual and code-switching tasks yield 6% and 5% (approximately) improvements in word error rate over the baseline system, respectively, according to the challenge development data.

This document [10] describes the CSTR submission for the Interspeech 2021 Multilingual and Code-Switching ASR Challenges. They developed a multilingual CNN-TDNN acoustic model for Gujarati, Hindi, Marathi, Odia, Tamil, and Telugu for the challenge's multilingual track. The model was then improved using solely monolingual training data. For decoding, a linguistic model created from training and CommonCrawl data was used. Furthermore, they showed how integrating semi-supervised training with data from YouTube crawls can successfully enhance the performance of the acoustic model. These models with confidence-based language identification produced an average WER of 18.1%, which is a 41% improvement over the provided multilingual baseline model. For the code-switching track of the challenge, their utilize a language model that has been trained on a combination of CommonCrawl data in Bengali and Hindi, along with in-domain English data. In contrast, their approach involves training a multilingual model specifically on technical lectures.To develop pronunciations for English terms, they also employed a cutting-edge transliteration technique. Comparing the final model to our multilingual baseline, improvements of 18% and 34% are seen. Both of their systems placed highly among the challenge entries.

A single multilingual speech recognition system that can recognize any of the training languages was created in this study by SuyounKimet al. [11] using current advancements in end-to-end voice recognition. To accomplish this, they suggested utilizing a universal character set that is shared by all languages. Additionally, they developed a language-specific gating system that can regulate the internal network representations in accordance with language. The Microsoft Cortana task was used to test the suggested approach, and the findings show that their system outperforms both standalone monolingual systems and systems developed utilizing a multi-task learning strategy.The authors also demonstrated how this model may be used to generate a bilingual model for usage in code-switching scenarios as well as to initialise a monolingual speech recognizer.

The CLSRIL-23 audio pre-trained model by Anirudh Gupta et al. learns cross-lingual speech representations from raw audio in 23 Indic languages using self-supervised learning. It is based on wav2vec 2.0 and simultaneously picks up the quantization of latents that all languages share. The problem is tackled by training a contrastive task across latent voice representations that have been masked. To examine the impact of monolingual and multilingual pretraining, they

compared the language-wise loss during pretraining. Using speech representations that encode phonetic similarity between languages, their research demonstrates that multilingual pretraining outperform monolingual training in downstream task performance as well as in the acquisition of such representations. Performance on various downstream speech recognition fine-tuning tasks was also compared. When a multilingual pre-trained model is employed for finetuning in Hindi, a 5% drop in WER and a 9.5% drop in CER were seen. Additionally, all of the code models have open sourcing. To aid in the study of speech recognition for Indic languages, the model CLSRIL-23 was trained on 23 languages and around 10,000 hours of audio data. The authors anticipate that, especially for Indic languages with limited resources, cutting-edge systems will be developed utilising the self-supervised technique.

The other various methods with their observations and limitations are listed inTable 1.

TABLE 1 Observations and DrawbacksofExisting Methods

| SL. No. | Title | Observations/ Drawbacks |
|---|---|---|
| 1 | Multilingual Speech Recognition With a Single End-to-end Model [8] | When comparing model performances across languages, Malayalam and Kannada show the worst results. The agglutinative nature of such languages, causes the average word in those languages to be lengthier than in languages like Hindi or Gujarati, is the explanation, according to authors.<br>Not able to handle code-switching problems.<br><br>The authors conclude that they would have improved recognition accuracy if they had combined the conditional variations of the model with distinct language-specific language models. |
| 2 | Dual Script E2E Framework for Multilingual and Code-Switching ASR [9] | Corpus Limitation:- Marathi & Odiya training corpus contains many English contents with only a few unique data.<br><br>Code Switching was not very accurate and it is very noisy. |
| 3 | The CSTR System For Multilingual And Code-switching ASR Challenges For Low Resource [10] | The authors conclude that they would have improved recognition accuracy even more if they had combined the conditional variations of the model with distinct language-specific language models. |
| 4 | Towards Language-universal End-to-end Speech Recognition [11] | It is necessary to investigate the best way to include a new language into a multilingual system that has previously been trained and assess the performance gains that can be made by utilizing an external language model. |
| 5 | CLSRIL-23: Cross-Lingual Speech Representations For Indic Language [12] | The findings show that multilingual pretraining performs better than monolingual pretraining both when learning speech representations during pre-training and when refining performance in a subsequent speech recognition exam. It demonstrates that the model can encode phonetic informationsimilarity in speech representations. |
| 6 | A Survey of Multilingual Models for Automatic Speech Recognition<br>[13] | More research is needed if these techniques can be applied to languages with extremely minimal resources using only a few hours of data.<br>The majority of the languages employed in the research on multilingual ASR are from the Indo-European language family. To determine whether results are consistent across languages, more research is needed.<br>It would be wise to keep comparing performance with simpler monolingual models as models get bigger in terms of |

| | | |
|---|---|---|
| | | parameters, training time, and training data needed, especially for low-resource languages.<br>An intriguing area of research that has already made some headway is the transfer of representations learned by ASR models trained with SSL. |
| 7 | Adaptive Multilingual Speech Recognition With Pretrained Models [14] | Utilizing a combination of three techniques—encoder pre-training, decoder pre-training, and adaptive weights—great progress in multilingual ASR was made in this study. The experimental findings show that each technique differs in its effects on various languages with various linguistic characteristics and data sizes. The model still has trouble getting to a reasonable error rate at the very low condition. Future work that could benefit from multimodal pretraining to address the cross-attention layers will be included in the analysis utilizing adaptive weights. |
| 8 | Spoken Language Identification Using Deep Learning [15] | The overall accuracy was 95%. The performance of the feature vector log-Mel spectrogram can be further enhanced by removing audio noise. The performance can be improved by using a variety of techniques, including pitch shifting, cropping, rotating, flipping, adding random noise, adjusting audio speed, and other approaches. It is frequently observed or discussed in more depth how different feature extraction techniques, such as Constant-Q transform and Fast Fourier Transform, affect language recognition. These are known to have an advantage over convolutional neural networks in terms of performance. |
| 9 | End-to-end Oriental Language Speech Recognition with Integrated Language Identification [16] | It can be observed that the overall language speech recognition effect has been improved, but the low recognition accuracy of Hokkien, Japanese, and Mandarin existing in the baseline system still exists in the multi-task system.<br>Since the potential relationship between the auxiliary task language identification and the main task speech recognition is unknown, and the proportion of the loss function of the language identification task to the total loss function is uncertain, this paper sets the proportion of the language identification task to be 0.1, 0.2, 0.3, 0.4, 0.5 through experiments for comparison. Need to do further research on the task relationship of the multi-task model. |
| 10 | Joint ASR And Language Identification Using Rnn-t: An Efficient Approach To Dynamic Language Switching [17] | Studies using spoken Hindi and Indian English demonstrate that (a) code-switched utterances are inherently difficult to identify and categorize, (b) coupled training offers better LID accuracy than multi-task learning, and (c) multi-task models with a dedicated LID feedforward network offer the best performance overall. The suggested joint ASR-LID architectures are multilingual and, in theory, can scale to support more than two languages. A possible future option for further reducing the memory and computation footprint is investigating systems with parameter sharing between the RNN-T encoder and acoustic LID classifier. |

Based on the information presented in the aforementioned Table 1, it is evident that there remain numerous areas within research and development that require further exploration. Several of these gaps include:

1. Less accuracy in the handling of multiple languages in speech (code-switching problem).
2. Performance degradation of ASR due to the presence of contradictory features in the languages,
3. Performance degradation on adding of new language to an already-trained multilingual system
4. Performance degradation of ASR performance due to the unbalanced training data.

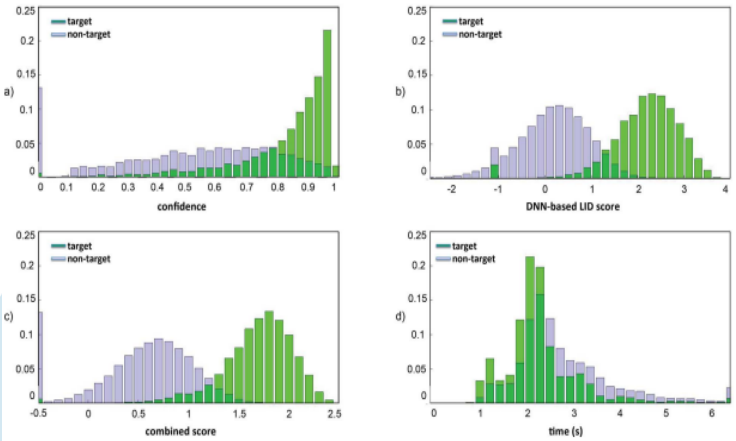


FIGURE. 3. Language identification over the Google Multilang Corpus. Distribution of target/non-target scores by using a) ASR confidences, b) the DNN-based LIDsystem, c) linear weighting of confidences and DNN-based LID system; and d) as a function of the response time of the speech recognizers.

Hence, it can characterize the issue as the creation of a system capable of precisely identifying and transcribing speech in numerous languages. This endeavor entails crafting a resilient algorithm capable of

- Handle variations in pronunciation, dialects, and accents across different languages
- Easily and efficiently adapt to the specific characteristics of each language.
- Recognize and differentiate between languages spoken within the same sentence and accurately transcribe the words spoken in each language.

## IV. METHODOLOGY

To initiate the process, our initial focus should be on examining the efficacy of transfer learning in the context of multilingual speech recognition. This entails a series of phases illustrated in Figure 3.

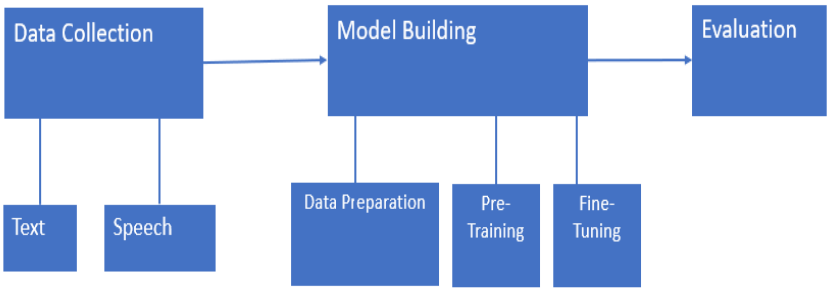


FIGURE 4. Transfer learning for multilingual speech recognition

In this context, they offer a variety of pre-trained models that come in both monolingual and multilingual options.Pre-trained models can be used for both acoustic and language models. Next is the collection of data for fine-tuning. Data can be in monolingual format or in multilingual format. Data consists of both speech and text. The third phase is the model building which is supervised learningand involves fine-tuning pre-trained models with collected data. Finally, evaluation of the model with the standard performance evaluation metrics.

The task completed are:

- Selected Pre-trained Model (English + Hindi)

~4200Hours of Data.
- Collected multilingual data for fine-tuning
  ~ 2000 sentences.
- Build a multilingual model that recognises both Hindi & English but the output is in Devanagari script.
- Evaluate the model & compared it with an open-source model.

And the task planned are:
- Select a pre-trained model that contains multiple Indian languages (more than two)
- Collect data for fine-tuning two Indian languages – Hindi & Malayalam.
- Fine-tune and built a model that can handle multiple Indian languages,
- Evaluation of the system.

The next step is to optimize language detection/ identification to handle code-switching, the process is shown in Figure 4.

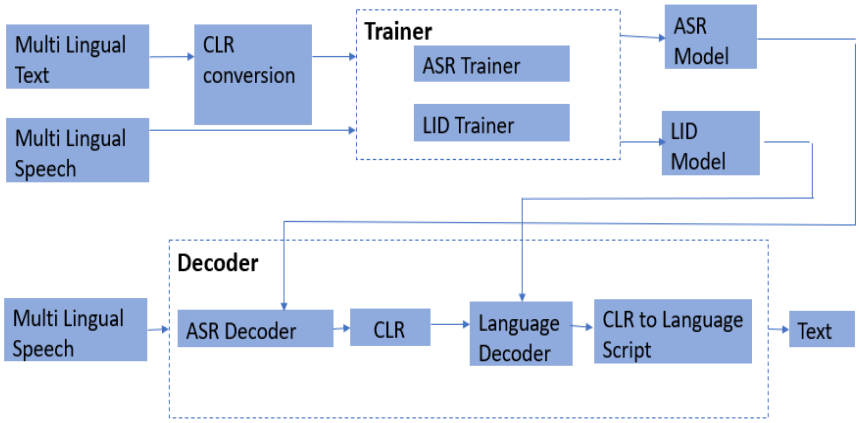


FIGURE 5. General Block Diagram of Multilingual Speech Recognizer

The basic flow diagram for language identification from speech has been depicted in Figure 7.


FIGURE 6. Language Identification from Speech

Here, Wav2vec 2.0 is used for speech recognition which is a model developed by Facebook AI Research. It utilizes a self-supervised learning approach to learn representations from raw audio data. The model is trained to distinguish between original speech examples and modified versions by performing a predictive coding task.

In the training process, Wav2vec 2.0 breaks down the audio data into short overlapping segments, typically spanning a few milliseconds. It then aims to predict the original audio sample corresponding to each segment but with a twist. Instead of predicting the current audio sample, the model is trained to predict future audio samples, typically milliseconds ahead.

By training the model to predict the correct audio samples in the future, Wav2vec 2.0 forces itself to capture meaningful representations of the input audio. This approach is based on the assumption that in order to predict future audio samples accurately, the model needs to understand the underlying structure and content of the input audio, including phonetic and linguistic information.

The training process of Wav2vec 2.0 repeats this predictive coding task multiple times for each second of audio. This repetition helps the model to learn progressively better representations of the audio data. The model learns to encode relevant features and discard irrelevant noise or variations introduced by the modified versions of the audio.

Overall, the training procedure of Wav2vec 2.0 involves predicting future audio samples from short overlapping segments of the input audio, repeating this process multiple times for each second of audio. This allows the model to learn robust representations and capture meaningful information from raw audio data, enabling it to perform tasks such as speech recognition and transcription more accurately.

Wav2Vec 2.0 is one of the most sophisticated models for automatic speech recognition currently available with a self-supervised training process, which is a relatively novel concept in this field. They can utilize this training strategy to pre-train a model using unlabeled data, which is consistently easier to acquire.The model can then be modified for a given objective on a particular dataset. This method of training is quite effective, as evidenced by earlier studies [18].

Figure 6 illustrates the two processes needed to train the model. The best voice representation is what is intended for the first phase, which is carried out in a self-supervised mode utilizing unlabeled data that can be approached similarly to how word embeddings are approached. Word embeddings also aim to represent real language as accurately as possible. The main difference is that Wav2Vec 2.0 processes audio instead of text. In the second stage of training, supervised fine-tuning, the model is instructed to predict certain words or phonemes using labelled data.
The primary benefit of this technique is the first training period. On a small amount of labelled data, learning an excellent speech representation permits obtaining cutting-edge outcomes. For instance, the model was pre-trained by the paper's authors using a sizable LibriVox dataset. When they used the entire Libri Speech dataset for fine-tuning, the test-clean subset had a Word Error Rate (WER) of 1.8%, whereas the test-other had a WER of 3.3%. Allowing for 2.0% WER on the test-clean and 4.0% on the test-other with over ten times less data. Only 10 minutes of labelled training data were used, which is practically no data, and the test-clean and test-other subsets of Libri Speech yielded WERs of 4.8% and 8.2%, respectively.

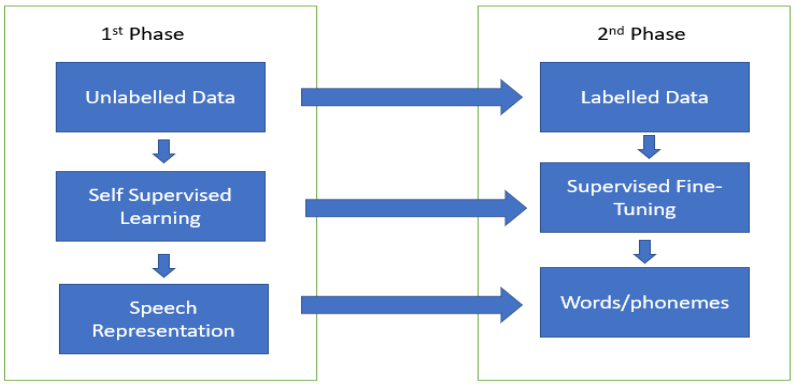


FIGURE 7. Phases of Wav2vec

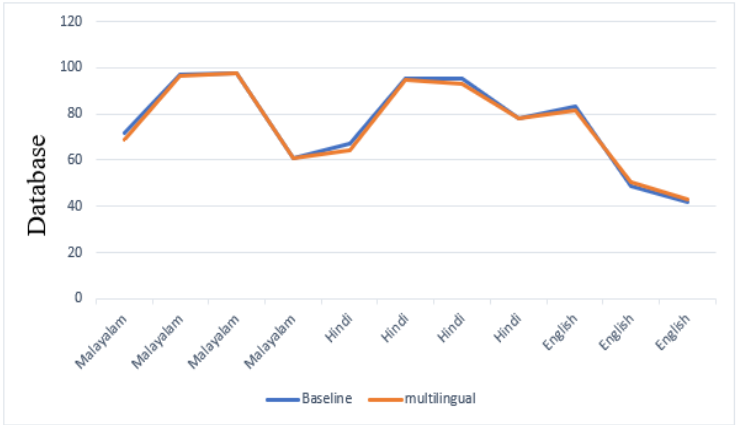


FIGURE 8.Comparison of (%) Word Accuracy resultsusing the baseline and the multilingual network tested on different languages and tasks.

## V. PERFORMANCE ANALYSIS

Our intention is to offer a comprehensive summary of the research efforts that have been explored and showcased within this section.For evaluation speech recognition, two metrics are generally used: word error rate and character error rate.

- The Character Error Rate (CER) is the ratio of errors in characters to the total number of characters
- Word Error Rate (WER) is the ratio of errors in a transcript to the total number of words spoken.

The formula for calculating WER is as follows:

Word Error Rate = (Substitutions + Insertions + Deletions) of words / Number of Words Spoken (1)

Similarly,

Character Error Rate = (Substitutions + Insertions + Deletions) of characters /Total Number of Characters (2)

## VI. CONCLUSION

This paper delves into the realm of automatic speech recognition, examining a range of methodologies, strategies, and advancements utilized within this field. Through our investigation, it becomes apparent that while researchers have devised multiple techniques, there remains a shortfall in adequately fulfilling the necessary criteria.

The formulation of a deep neural framework for multilingual end-to-end automatic continuous speech recognition (ASR) using Wav2Vec 2.0 is a promising approach in the field of multilingual ASR. Wav2Vec 2.0, with its self-supervised pre-training approach, has shown remarkable performance in monolingual ASR tasks. Extending it to the multilingual setting opens possibilities for more efficient and robust speech recognition across different languages.By leveraging Wav2Vec 2.0's capabilities, the proposed framework eliminates the need for language-specific phonetic transcriptions or acoustic models. Instead, it learns language-agnostic representations directly from raw audio data, making it adaptable to various languages without extensive language-specific training data or resources.

The framework benefits from the advantages of deep neural networks, such as Convolutional Neural Networks (CNNs) and Transformer-based models, which effectively capture local and global dependencies in speech signals. The use of these components, along with Recurrent Neural Networks (RNNs) for temporal modeling, enhances the overall performance of the multilingual ASR system.

In the future,it can expand the domain by:

1. Enhanced Multilingual Training: While the proposed framework already demonstrates the ability to handle multiple languages, further improvements can be made by augmenting the multilingual training process. Techniques like curriculum learning, language-specific fine-tuning, or cross-lingual adaptation can be explored to optimize the model's performance for individual languages.
2. Low-Resource Language Support: One challenge in multilingual ASR is the scarcity of labelled data for low-resource languages. Future enhancements can focus on techniques such as unsupervised or semi-supervised learning to leverage the available unlabeled or limited labelled data for these languages, enabling better performance in low-resource scenarios.
3. Code-Switching and Language Mixing: Many multilingual environments involve code-switching or language mixing, where speakers switch between languages within a single utterance. Enhancements can be made to the framework to handle code-switching scenarios effectively, allowing for seamless recognition of mixed-language speech.
4. Speaker Adaptation: Personalization and adaptation to individual speakers can be explored to improve ASR performance. Speaker-specific fine-tuning or techniques like speaker embeddings can be incorporated to enhance the model's ability to recognize speech from specific individuals, leading to better accuracy and user experience.
5. Robustness to Noise and Variability: ASR systems must be robust to various environmental conditions, such as background noise, reverberation, or channel distortions. Future enhancements can focus on incorporating robustness mechanisms, such as data augmentation techniques, multi-microphone processing, or domain adaptation, to improve the model's performance in real-world and challenging conditions.

By addressing these future enhancements, the deep neural framework for multilingual end-to-end ASR using Wav2Vec 2.0 can further advance the field of multilingual speech recognition, enabling more accurate and versatile speech-to-text conversion across diverse languages and real-world scenarios.

**REFERENCES:**


1. S. Qin, L. Wang, S. Li,"Improving low-resource Tibetan end-to-end ASR by multilingual and multilevel unit modelling," J Audio Speech Music Proc. 2022, 2 (2022). https://doi.org/10.1186/s13636-021-00233-4
2. S. Liang, W. Q. Yan,"A hybrid CTC+Attention model based on end-to-end framework for multilingual speech recognition," Multimed Tools Appl 81, 2022, 41295–41308. https://doi.org/10.1007/s11042-022-12136-3
3. H. Inaguma, K. Duh, T. Kawahara and S. Watanabe, "Multilingual End-to-End Speech Translation," 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 2019, pp. 570-577, doi: 10.1109/ASRU46091.2019.9003832.
4. S.S. Aluru, B. Mathew, P. Saha andA. Mukherjee, Deep learning models for multilingual hate speech detection. arXiv preprint arXiv:2004.06465, 2020.
5. S. Venkatasubramanian, R. Mohankumar,"A Deep Convolutional Neural Network-Based Speech-to-Text Conversion for Multilingual Languages," In: Smys, S., Tavares, J.M.R.S., Balas, V.E. (eds) Computational Vision and Bio-Inspired Computing. Advances in Intelligent Systems and Computing, vol 1420. 2022, Springer, Singapore. https://doi.org/10.1007/978-981-16-9573-5_44
6. M. Corazza, S. Menin, E. Cabrio, S. Tonelli and S. Villata, "A multilingual evaluation for online hate speech detection," ACM Transactions on Internet Technology (TOIT), 20(2), 2020, pp.1-22.
7. J. Cho, M.K. Baskar, R. Li, M. Wiesner, S.H. Mallidi, N. Yalta, M. Karafiat, S. Watanabe, and T. Hori, "Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling," In 2018 IEEE Spoken Language Technology Workshop (SLT), 2018, December, pp. 521-527. IEEE.
8. S. Toshniwal, "Multilingual Speech Recognition with a Single End-to-End Model," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 4904-4908, doi: 10.1109/ICASSP.2018.8461972.
9. M.G. Kumar, J. Kuriakose, A. Thyagachandran, A, A.K., A. Seth, L.V.S.V.D. Prasad, S. Jaiswal, A. Prakash, H.A. Murthy, "Dual Script E2E Framework for Multilingual and Code-Switching ASR," Proc. Interspeech 2021, 2021, 2441-2445, doi: 10.21437/Interspeech.2021-978
10. O.Klejch, E. Wallington, &P. Bell,"The CSTR System for Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages," In Proceedings of Interspeech, 2021, pp. 2881-2885. International Speech Communication Association. https://doi.org/10.21437/Interspeech.2021-1035
11. S. Kim and M. L. Seltzer, "Towards Language-Universal End-to-End Speech Recognition," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 4914-4918, doi: 10.1109/ICASSP.2018.8462201.
12. Gupta, Anirudh & Chadha, Harveen & Shah, Priyanshi & Chimmwal, Neeraj & Dhuriya, Ankur & Gaur, Rishabh & Raghavan, Vivek. 2021. CLSRIL-23: Cross-Lingual Speech Representations for Indic Languages.
13. Hemant Yadav and Sunayana Sitaram. "A Survey of Multilingual Models for Automatic Speech Recognition,"In Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pages 5071–5079, Marseille, France. European Language Resources Association.
14. N,-Q. Pham, A. Waibel, and J. Niehues, "Adaptive multilingual speech recognition with pretrained models", arXiv e-prints, 2022. doi:10.48550/arXiv.2205.12304.
15. G. Singh, S. Sharma, V. Kumar, M. Kaur, M. Baz, M. Masud,"Spoken Language Identification Using Deep Learning," Comput Intell Neurosci. 2021 Sep 20;2021:5123671. doi: 10.1155/2021/5123671. PMID: 34594371; PMCID: PMC8478554.
16. A. Qi, X. Xie, Q. Zhan, C. Hu and X. Su, "End-to-end Oriental Language Speech Recognition with Integrated Language Identification," 2022 International Conference on Machine Learning, Control, and Robotics (MLCR), Suzhou, China, 2022, pp. 27-31, doi: 10.1109/MLCR57210.2022.00014.
17. S. Punjabi., "Joint ASR and Language Identification Using RNN-T: An Efficient Approach to Dynamic Language Switching," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 7218-7222, doi: 10.1109/ICASSP39728.2021.9413734.
18. J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019).
19. R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," arXiv preprint arXiv:1912.06670, 2019.
20. J. Iranzo-S´anchez, J. A. Silvestre-Cerd`a, J. Jorge, N. Rosell ´o, A. Gim´enez, A. Sanchis, J. Civera, and A. Juan, "Europarl-st: A multilingual corpus for speech translation of parliamentary debates," in ICASSP, 2020.
21. N.-Q. Pham, T.-L. Ha, T.-N. Nguyen, T.-S. Nguyen, E. Salesky, S. St ¨uker, J. Niehues, and A. Waibel, "Relative Positional Encoding for Speech Recognition and Direct Translation," in Proc. Interspeech 2020.
22. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, 2017.

23. A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, "Xls-r: Selfsupervised cross-lingual speech representation learning at scale," arXiv preprint arXiv:2111.09296, 2021.
24. V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," arXiv preprint arXiv:2012.03411, 2020.
25. Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual translation with extensible multilingual pretraining and finetuning," arXiv preprint arXiv:2008.00401, 2020.
26. N.-Q. Pham, T.-N. Nguyen, S. Stueker, and A. Waibel, "Efficient weight factorization for multilingual speech recognition," arXiv preprint arXiv:2105.03010, 2021.