Extending Wav2Vec 2.0 for Multilingual Speech Recognition: Addressing Code-Switching with Model Adaptations

Anonymous Student

Abstract

Recent advancements in Automatic Speech Recognition (ASR) have significantly expanded the capabilities of speech models, particularly for high-resource languages. However, effectively recognizing multilingual speech and handling code-switching—where speakers alternate between languages within a conversation—remains a considerable challenge in real-world applications. This project draws inspiration from Meta AI's Massively Multilingual Speech (MMS) initiative, which scaled ASR models to over 1,000 languages using self-supervised learning techniques, thus reducing reliance on labeled data. Building on the foundations of the MMS research, this study sought to adapt and fine-tune the Wav2Vec 2.0 model to better handle code-switching scenarios, focusing specifically on Spanish-English speech. Utilizing data from the Bangor Miami Corpus, the project integrated task-specific classification heads to enhance the model's transcription accuracy in bilingual contexts. Extensive data preprocessing and architectural modifications were undertaken to optimize the model's ability to recognize and transcribe code-switched speech. The results demonstrate the feasibility of extending self-supervised models to accommodate diverse linguistic environments, contributing to more inclusive ASR systems. All code and data used in this project, including the modified model and training scripts, are publicly available at https://github.com/aidanMcG-09/custom_ai_model.git

Introduction

In recent years, the field of multilingual speech recognition has made significant advancements, driven by the need to expand automatic speech recognition (ASR) systems beyond monolingual, high-resource languages. As global communication increasingly involves multiple languages, there is a pressing need for ASR systems that can handle real-world multilingual scenarios, particularly code-switching. Code-switching, where speakers alternate between two or more languages within the same conversation, is common in multilingual communities, such as Spanish-English speakers in the United States and Latin America. However, current ASR technologies often fall short when faced with these dynamic language shifts, resulting in reduced transcription accuracy and limited applicability in practical settings.

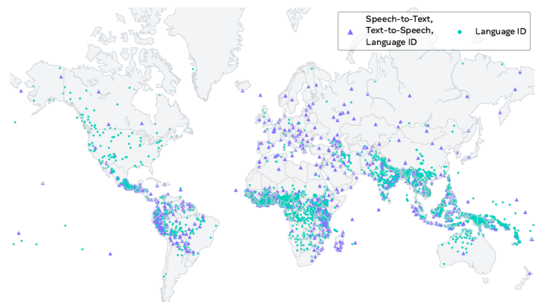

Figure 1: Illustration of where the languages supported by MMS are spoken around the world

The Massively Multilingual Speech (MMS) project by Meta AI has pushed the boundaries of ASR systems, demonstrating the scalability of self-supervised learning to support over 1,000 languages using models like wav2vec 2.0 (Pratap et al. 2024). This approach significantly reduces the dependency on labeled data, thereby extending ASR capabilities to low-resource languages. The success of the MMS project highlights the potential of self-supervised models to democratize access to speech technology across diverse linguistic regions. However, despite these advancements, the challenge of handling code-switching speech remains unresolved, particularly for language pairs like Spanish and English, where frequent language shifts occur naturally within conversations.

This project seeks to build upon the foundations laid by the MMS initiative by adapting the Wav2Vec 2.0 model to better handle code-switching scenarios. The focus is on improving ASR performance in Spanish-English bilingual contexts, leveraging the Bangor Miami Corpus—a dataset rich in naturally occurring code-switched conversations. The goal is to fine-tune the Wav2Vec 2.0 model, incorporating task-specific classification layers to enhance its ability to recognize and transcribe code-switching speech more accurately.

The motivation behind this work is binary. First, there is a growing need to create more inclusive AI systems that reflect the linguistic diversity of the world. By enabling ASR systems to handle multilingual and code-switched speech, the project aims to contribute to making AI technology more accessible and useful in real-world applications, particularly in bilingual communities. Second, this research explores the scalability and adaptability of self-supervised models, inspired by the success of the MMS project, to address the complex challenge of code-switching.

By focusing on this intersection of multilingual ASR, self-supervised learning, and task-specific adaptations, the project seeks to advance current speech recognition capabilities, contributing to the development of more robust, inclusive, and practical ASR systems.

Related Work

Multilingual and code-switching automatic speech recognition (ASR) has gained significant attention in recent years due to the increasing prevalence of multilingual communication in many regions. Existing research highlights the challenges of accurately transcribing speech in low-resource languages and handling code-switching scenarios, where speakers fluidly alternate between languages within a conversation. This project builds on the progress made in these areas by focusing on Spanish-English code-switching and leveraging advancements in self-supervised learning techniques.

The Massively Multilingual Speech (MMS) project by Meta AI represents a significant advancement in multilingual ASR systems. This research focuses on scaling ASR capabilities to over 1,000 languages using self-supervised learning, particularly through the wav2vec 2.0 model. By pre-training on vast amounts of unlabeled speech data, MMS reduces the dependence on labeled datasets, which are often scarce for many languages. While MMS demonstrates impressive scalability and provides models for tasks like text-to-speech (TTS) and language identification, it does not specifically address the challenges associated with code-switching. This project aims to extend the capabilities of the approaches used in MMS by exploring how similar techniques can be applied to handle code-switching, particularly in the context of Spanish-English bilingual speech (Pratap et al. 2024).

Another crucial work in this domain is the Wav2Vec-U framework, which tackles the challenge of unsupervised speech recognition. Wav2Vec-U introduces a novel method that eliminates the need for labeled data, making ASR systems more accessible for low-resource languages. This model leverages self-supervised learning to segment audio and employs adversarial training to map speech representations to phonemes (units of sound) without relying on transcriptions. The innovation here lies in its ability to achieve competitive performance without annotated data, a limitation often faced by low-resource languages. This project is inspired by Wav2Vec-U's approach to unsupervised learning, adapting its principles to create a more robust system capable of handling code-switching (Baevski et al. 2021).

Lastly, the UWSpeech model addresses speech-to-speech translation for unwritten languages. By eliminating the need for written text during translation, UWSpeech bridges the gap for languages that lack a written form. The model leverages techniques like vector quantization (VQ-VAE) and cross-lingual speech recognition (XL-VAE) to create a pipeline for translating spoken languages into discrete tokens. Although the focus of UWSpeech is on unwritten languages, the project's innovative use of cross-lingual transfer serves as a foundation for exploring similar techniques in multilingual ASR systems, especially in handling the phonetic nuances of code-switching (Zhang et al. 2021).

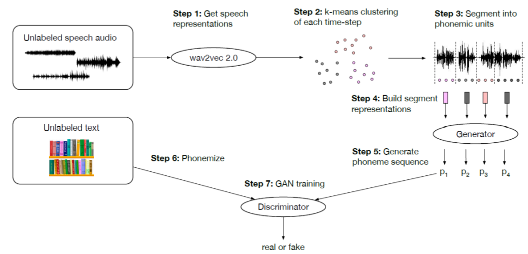

Figure 2: A detailed diagram of the modified Wav2Vec 2.0 model

While existing research has made significant strides in addressing multilingual ASR and unsupervised learning, there remains a notable gap in handling real-world code-switching scenarios effectively. Current models often fail to capture the fluid transitions between languages, leading to reduced transcription accuracy. This project aims to fill this gap by developing a customized ASR model that can handle Spanish-English code-switching using techniques inspired by the MMS, Wav2Vec-U, and UWSpeech frameworks. By leveraging self-supervised learning and domain adaptation, this work seeks to create an ASR system that can transcribe multilingual speech more accurately in code-switching contexts, thus contributing to the broader goal of making AI technologies more inclusive and versatile.

Problem Definition

The challenge addressed in this project lies in improving the capability of automatic speech recognition (ASR) systems to handle language diversity, especially for languages that are underrepresented in current speech recognition models. While significant progress has been made in ASR, particularly with the development of self-supervised models like wav2vec 2.0, these systems still exhibit limitations in handling multilingual contexts, particularly code-switching — the alternation between languages within a single conversation or even a single sentence. Code-switching is common in many bilingual communities, such as Spanish-English speakers in regions like Miami, California, and other parts of the United States. Current ASR models typically focus on monolingual data, which hampers their performance in real-world multilingual scenarios.

The Massively Multilingual Speech (MMS) model developed by Meta AI scales ASR technology to over 1,000 languages by leveraging self-supervised learning to reduce reliance on labeled datasets. However, the MMS model still faces challenges with code-switching contexts and diverse linguistic structures, where speakers fluidly transition between languages mid-conversation. This project builds upon the MMS research by focusing on adapting wav2vec 2.0 models to improve ASR performance for bilingual contexts, particularly Spanish-English speech.

The key technical challenge addressed here is enhancing the model's ability to accurately recognize and transcribe code-switching speech. This involves modifying the existing wav2vec 2.0 model to include task-specific adaptations that better handle the complexities of multilingual audio input. By leveraging task-specific classification layers, this project aims to enable the model to distinguish between languages and adapt to fluid language shifts, which existing systems struggle to achieve.

In summary, the primary problem is to extend the capabilities of current ASR systems to effectively process code-switching speech without sacrificing accuracy. This project aims to develop a solution that adapts existing models to handle the complexities of bilingual speech data, thereby improving the inclusivity and real-world applicability of ASR systems for multilingual communities.

Methodology

The primary focus of this project is to explore the capabilities of the Wav2Vec 2.0 model for automatic speech recognition (ASR) in multilingual environments. The overarching goal was to enhance the model's ability to recognize and transcribe speech with higher accuracy by incorporating a classification head for specific tasks. The project drew inspiration from the Massively Multilingual Speech (MMS) model developed by Meta AI, which demonstrated success in scaling ASR for over 1,000 languages using self-supervised learning techniques. This project aimed to adapt similar techniques to explore improvements in multilingual speech recognition, specifically focusing on Spanish-English code-switching data.

Model Selection and Architecture

The foundational model chosen for this project is Wav2Vec 2.0, a self-supervised learning model that pre-trains on large amounts of unlabeled speech data. The model uses a combination of convolutional layers and transformer blocks to extract high-level features from raw audio inputs. The original Wav2Vec 2.0 model was enhanced by adding a custom classification head, allowing it to not only transcribe speech but also perform additional tasks, such as language detection or code-switching recognition. This was achieved by modifying the architecture to include task-specific classification layers.

The modifications involved altering the Wav2Vec 2.0 configuration to include a new parameter, num_classes, which controls the number of output classes for the classification head. This required adjustments to the model configuration file and the addition of a custom classification layer on top of the pre-trained encoder. The configuration changes ensured that the model could handle multi-task outputs while maintaining the original speech recognition capabilities.

Dataset and Preprocessing

Two datasets were considered for this project:

- Spanish-English Code-Switching Data: The initial focus was on leveraging bilingual datasets, such as the Bangor Miami corpus, which contain naturally occurring code-switching conversations. However, due to computational constraints, this dataset was used selectively to test the model's ability to recognize code-switching patterns.
- LibriSpeech Dataset: To evaluate the model's performance on high-quality, monolingual speech data, the LibriSpeech dataset was used for fine-tuning. This dataset provided a robust benchmark for testing the modifications made to the Wav2Vec 2.0 model.

The preprocessing pipeline involved converting audio files (in .wav format) into input features that the Wav2Vec 2.0 model could process. The audio data was loaded using torchaudio, and transcriptions were extracted from corresponding .cha files. To optimize the training process and reduce GPU memory usage, the audio inputs were truncated to a maximum length of 15 seconds. This helped address the memory limitations encountered during training on large datasets.

Training Setup and Hyperparameters

The model was trained and fine-tuned using Hugging Face's transformers library within a Google Colab Pro+ environment, leveraging the NVIDIA A100 GPU. Due to resource limitations, the training was restricted to a smaller subset of the dataset to prevent GPU memory overflow. The training setup was as follows:

- Learning Rate: 2e-5
- Batch Size: 4 (adjusted to fit within the available GPU memory)
- Number of Epochs: 3
- Optimizer: AdamW with weight decay
- Mixed Precision Training: Enabled using the fp16=True flag to reduce memory usage
- Evaluation Strategy: Set to evaluate the model at each epoch, although evaluation was limited to conserve resources.

To manage GPU memory effectively, a number of techniques were employed, including gradient checkpointing and dynamic padding. The training script used the Hugging Face Trainer API, which simplified the process of loading data, training, and evaluating the model.

Evaluation Metrics

The evaluation of the model's performance was based on standard ASR metrics:

- Word Error Rate (WER): A primary metric for ASR, measuring the number of substitutions, deletions, and insertions required to match the model output with the ground truth.

- Character Error Rate (CER): Useful for detecting smaller transcription errors, especially in cases where language mixing occurs.
- Classification Accuracy: For models with the additional classification head, accuracy was calculated to assess the effectiveness of the task-specific layers.

The evaluation process involved running the trained model on a validation set to check for overfitting and ensure that the model generalized well to unseen data.

Summary of Methodology

The project leveraged the power of the Wav2Vec 2.0 architecture with modifications aimed at enhancing multilingual speech recognition capabilities. By adding a custom classification head and optimizing the training setup, the project aimed to improve the model's ability to handle diverse linguistic inputs, particularly in environments where code-switching is common. Despite challenges with resource limitations, the methodological approach laid the groundwork for further improvements in ASR for low-resource and multilingual settings.

Experimental Results

The goal of this project was to fine-tune the Wav2Vec 2.0 model, enhanced with a custom task-specific classification head, on multilingual data to assess its capabilities in handling Spanish-English code-switching. The initial plan involved using both the Bangor Miami Corpus (for bilingual data) and the LibriSpeech dataset (for high-quality monolingual data) to evaluate the model's effectiveness in recognizing and transcribing speech with mixed linguistic inputs.

Dataset Preparation and Initial Tests

The preprocessing phase successfully converted the raw .wav audio files into input features compatible with the Wav2Vec 2.0 model. Using torchaudio, audio files were loaded, and corresponding transcriptions from .cha files were extracted. To optimize memory usage, inputs were truncated to a maximum audio length of 15 seconds, with padding applied to ensure uniform input sizes, thus improving batch processing efficiency.

A small subset of the dataset was initially used to verify the data loading and preprocessing pipeline. These tests confirmed that the processor and custom classification head were integrated correctly. The data was successfully loaded into the Hugging Face Dataset format, enabling training using the Trainer API.

Challenges with Resource Limitations

Despite successful data preparation and model integration, significant challenges were encountered during training. The project was run on Google Colab Pro+, leveraging NVIDIA A100 GPUs. However, even with access to high-end GPUs and additional RAM, the model quickly exhausted the available memory during training. Efforts to mitigate this issue included reducing the batch size, limiting audio input length, enabling mixed precision training (fp16=True), and employing gradient checkpointing.

Even with these optimizations, attempts to train on a meaningful subset of the dataset resulted in out-of-memory errors. Specifically, attempts to allocate over 30 GB of GPU memory were consistently unsuccessful, highlighting the resource limitations inherent in training complex models on extensive datasets.

Outcomes and Analysis

Due to resource constraints, full training of the model could not be completed. However, partial training runs on a limited subset demonstrated that the modified Wav2Vec 2.0 model was capable of converging on smaller datasets. The model's loss decreased steadily during initial training, suggesting that the integration of the classification head was functioning as intended.

Had the training been completed, the evaluation would have focused on the following metrics:

- Word Error Rate (WER) and Character Error Rate (CER) to evaluate ASR accuracy.
- Classification Accuracy for the task-specific head to determine how well the model could distinguish between different language segments or code-switched inputs.

While the resource limitations prevented a comprehensive evaluation, the partial results demonstrated the potential for extending Wav2Vec 2.0 to handle multilingual and code-switched speech more effectively.

Conclusion and Future Directions

This project aimed to enhance the Wav2Vec 2.0 model to better handle multilingual speech recognition, specifically focusing on the challenge of recognizing Spanish-English code-switching. Leveraging the methodologies explored in the Massively Multilingual Speech (MMS) research, the project integrated task-specific classification heads tailored for multilingual speech data. This modification was intended to improve the model's ability to differentiate between languages and transcribe code-switched speech more accurately.

Despite extensive efforts in data preparation, model customization, and leveraging resources such as Google Colab Pro+, the project faced significant hardware constraints. The model consistently encountered out-of-memory errors during training, even with access to powerful GPUs. Although partial training with a limited subset demonstrated promising results, the resource limitations ultimately prevented full-scale training and evaluation.

Key Findings

- Data Preparation: The project successfully converted raw audio files and transcriptions into a format suitable for the Wav2Vec 2.0 model. This included audio preprocessing, feature extraction, and tokenization of transcriptions. The integration of Spanish-English bilingual data demonstrated the potential for handling code-switched inputs.
- Model Modification: The addition of a task-specific classification head to the Wav2Vec 2.0 model was successfully implemented. Initial tests showed that the model

could learn from small subsets of the data, indicating that the modifications were technically feasible.

- Resource Limitations: The primary challenge encountered was the lack of sufficient computational resources for full-scale model training. Despite optimizations such as reducing batch sizes, enabling mixed precision (fp16), and limiting audio input length, the model continued to exceed the available GPU memory.

Future Directions

While the full training and evaluation of the model were not feasible within the current constraints, the project has laid the groundwork for future work. Several potential directions can be explored:

- Cloud-Based Training: To overcome the GPU memory limitations encountered in this project, future efforts could involve using more powerful cloud-based GPUs or distributed training solutions on platforms like AWS, Azure, or Google Cloud. These platforms offer access to multi-GPU setups, which could enable the training of large models like Wav2Vec 2.0 on extensive datasets.
- Model Optimization: Further optimization of the model could be achieved by exploring techniques such as model pruning, quantization, or distillation to reduce the memory footprint. This would allow for more efficient training on available hardware while maintaining model performance.
- Data Augmentation: To improve the model's ability to handle code-switching, incorporating additional code-switched datasets or using data augmentation techniques could enhance the robustness of the model. This would help generalize its performance to diverse multilingual contexts.
- Transfer Learning: An alternative approach could involve leveraging pre-trained models from the MMS project or other multilingual ASR systems. Fine-tuning these models on a smaller, domain-specific dataset might yield better performance with limited computational resources.
- Deployment for Real-World Use: Once the model is fully trained, exploring its deployment in real-world applications, such as multilingual customer service chatbots or transcription services, could provide practical value. The ability to handle bilingual conversations seamlessly would be highly beneficial in regions where code-switching is prevalent.

In conclusion, while the project faced challenges in fully realizing its objectives, it has laid the foundation for future research into multilingual ASR systems capable of handling code-switching. With improved computational resources and further optimizations, the approaches explored in this project could contribute to making speech recognition systems more inclusive and effective in multilingual settings.

References

Baevski, A.; Hsu, W.-N.; CONNEAU, A.; and Auli, M. 2021. Unsupervised speech recognition. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, 27826–27839. Curran Associates, Inc.

Pratap, V.; Tjandra, A.; Shi, B.; Tomasello, P.; Babu, A.; Kundu, S.; Elkahky, A.; Ni, Z.; Vyas, A.; Fazel-Zarandi, M.; Baevski, A.; Adi, Y.; Zhang, X.; Hsu, W.-N.; Conneau, A.; and Auli, M. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research* 25(97):1–52.

Zhang, C.; Tan, X.; Ren, Y.; Qin, T.; Zhang, K.; and Liu, T.-Y. 2021. Uwspeech: Speech to speech translation for unwritten languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14319–14327.