Title	Optimizing Speech Translation for Low-Resource Languages With SAMU Pretraining, Self-Distillation And Online Clustering
Author(s)	阮, 国強
Citation	
Issue Date	2025-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/19780
Rights	
Description	Supervisor: NGUYEN, Minh Le, 先端科学技術研究科, 修士 (情報科学)

Master's Thesis/ Master's Research Project Report

Optimizing Speech Translation for Low-Resource Languages With SAMU Pretraining, Self-Distillation And Online Clustering

NGUYEN, Quoc Cuong

NGUYEN, Minh Le

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

March, 2025

Abstract

Speech translation is a vital technology that bridges language barriers, enabling people from different linguistic backgrounds to communicate effectively. However, developing speech translation systems for low-resource languages poses significant challenges. Low-resource languages are those with limited available data for training models, making it difficult to create accurate and reliable translation systems.

The primary objective of this research is to enhance the performance of a direct speech translation model tailored for low-resource languages. A direct speech translation model translates spoken language directly from the source language to the target language without relying on intermediate text representations. The model architecture consists of three main components: an encoder, a dimensionality reduction module, and a transformer decoder layer.

A key contribution of this work is the exploration and implementation of two novel pretraining techniques. These techniques are derived from recent advancements in speech representation learning and are designed to improve the quality of the encoder's understanding of the source language audio. The first technique focuses on creating semantically aligned, multimodal, cross-lingual speech representations that enhance the model's ability to understand and translate spoken language across different languages. The second pretraining technique employs self-distillation and online clustering to learn robust and meaningful speech representations without requiring extensive labeled data. In this study, both pretraining techniques are applied to the encoder using audio data from a low-resource language. Specifically, the audio data of the Tamasheq language, a Niger-Congo language spoken in parts of Mali, Algeria, and Niger, is used. The entire direct speech translation architecture is fine-tuned after pretraining the encoder with the chosen techniques.

The research utilizes data from the IWSLT2024 competition, specifically focusing on the low-resource speech translation task involving the Tamasheq-French language pair. The IWSLT (International Workshop on Spoken Language Translation) competition provides a standardized benchmark for evaluating speech translation systems, allowing for consistent and objective comparisons of different models and techniques.

To evaluate the effectiveness of the two pretraining techniques, BLEU score metric is used. The results of this research are expected to demonstrate that both pretraining techniques significantly improve the performance of the direct speech translation model for the low-resource Tamasheq-French language pair. Moreover, the research highlights the potential of applying these pretraining techniques to other low-resource language pairs, contributing to the larger goal of making speech translation technologies more accessible.

Acknowledgements

I would first like to express my sincere gratitude to my principal mentor and supervisor, Professor Nguyen Le Minh of the Graduate School of Advanced Science and Technology at the Japan Advanced Institute of Science and Technology (JAIST), for his ongoing careful support and insightful guidance throughout my master's program.

I am also deeply grateful to Professor Sakriani Sakti of the Nara Institute of Science and Technology for her invaluable support and enthusiastic feedback, which were instrumental in the successful completion of my entrance process and this thesis.

I would like to express my appreciation to Dr. Racharak Teeradaj and Professor Ikeda Kokolo, Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, for sharing with me insightful perspectives, so that I can form up the idea for my research.

I am thankful for the kindness and support extended to me by all members of Professor Nguyen's laboratory and my friends at JAIST during my time here.

My studies at JAIST were conducted under the Collaborative Education Program between JAIST and the University of Information Technology, Ho Chi Minh City National University. I wish to acknowledge the excellent instruction provided by all the lecturers involved in this program.

Finally, I express my deepest gratitude to my family for their constant support throughout my academic journey.

Contents

This thesis was prepared according to the curriculum for the Collaborative Education Program organized by Japan Advanced Institute of Science and Technology and University of Information Technology, Ho Chi Minh City National University

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In our increasingly interconnected world, the ability to communicate between people from different languages has become more crucial than ever. The motivation behind this research comes from the desire to make communication more inclusive and accessible.

Speech translation, the task of converting spoken language from one language to another, holds great potential for breaking down these language barriers. While machine translation technology has made huge progress in recent years, particularly for widely-spoken languages, many of the world's approximately 7,000 languages remain underserved by these technological advances. Low-resource languages are spoken by communities that may lack access to the technological advancements enjoyed by speakers of high-resource languages like English, Spanish, or Mandarin. By developing speech translation systems for these languages, we can empower speakers to access information, participate in global conversations, and preserve their linguistic heritage.

1.2 Problem Statement

Speech translation, which enables translation of spoken content from one language to another, represents a particularly challenging frontier in this domain. This technology holds great potential for preserving cultural heritage, facilitating education, and enabling access to information in communities where written literacy might be limited. However, developing effective speech translation systems for low-resource languages presents unique challenges, primarily due to the insufficiency of training data and the limited availability of parallel speech-text corpora. These languages often lack the extensive datasets and linguistic resources that support the development of robust machine learning models. This scarcity comes from various factors, including limited digital presence, lack of standardized orthography, and the sheer diversity of languages spoken across the globe.

Traditional approaches to speech translation often rely on a cascaded system, where speech recognition and machine translation models are used together. While effective for high-resource languages, this approach suffers from error propagation in low-resource scenarios, where errors in the speech recognition stage can cascade into the machine translation stage, leading to inaccurate translations. Direct speech translation models, which directly map speech input to translated text, offer a promising alternative. These models eliminate the need for intermediate transcription, potentially reducing error accumulation and improving efficiency. Moreover, direct speech-to-speech translation (S2ST) models have the potential to preserve paralinguistic and non-linguistic features, such as tone and emotion, which can be crucial for conveying meaning in certain languages [1].

1.3 Research Objectives and Contributions

This thesis aims to address the challenges of developing speech translation systems for low-resource languages by focusing on the Tamasheq-French language pair. Tamasheq is a Niger-Congo language spoken in parts of Mali, Algeria, and Niger, and it exemplifies the characteristics of a low-resource language with limited digital resources. French, as a widely spoken language, serves as an appropriate target language for translation.

A key contribution of this work is the exploration and implementation of two novel pretraining techniques applied to the problem of speech translation for low-resource languages. These techniques are derived from recent advancements in speech representation learning and are designed to improve the quality of the encoder's understanding of the source language audio.

- Samu-xlsr Pretraining Technique: Based on the method described in the paper "Samu-xlsr: Semantically-aligned multimodal utterance-level cross-lingual speech representation," this technique focuses on creating multimodal, semantically aligned, cross-lingual speech representations. By aligning speech data across different languages at the utterance level, the model gains a better understanding of the semantic content, which enhances its ability to translate spoken language accurately.

- DinoSR Pretraining Technique: Inspired by the "Dinosr: Self-distillation and online clustering for self-supervised speech representation learning" paper, this method employs self-distillation and online clustering to learn robust and meaningful speech representations without requiring extensive labeled data. Self-supervised learning allows the model to capture the nuances of the source language's audio, which is particularly beneficial for low-resource settings where labeled data is scarce.

1.4 Thesis Outline

The structure of this thesis is organized as follows: In Chapter 1, the motivation, problem statement and objectives of this study were discussed. Chapter

2 covers related studies in this field. Chapter 3 presents the 2 audio encoder pretraining techniques used in this study. Chapter 4 describes our experimental setup and details. Chapter 5 is a place for error analysis and discussion. Chapter 6 concludes this entire study, as well as plans for future works.

Chapter 2

Related Works

2.1 Low-Resource Speech Translation

In this section we discussed about related works on low-resource speech translation. It includes 2 parts: cascaded speech translation and direct speech translation.

2.1.1 Low-Resource Cascaded Speech Translation

Early research in ST primarily focused on cascaded approaches, where speech recognition (ASR) and machine translation (MT) were treated as separate tasks. In this approach, the audio input in low-resource language could be first processed and converted into the transcription in that same language. Then a machine translation model takes place to translate it into the targeted language.

Automatic Speech Recognition (ASR) There have been many works that tried to develop new model architecture or training methods for better performance on limited data. Singh et al [5] introduced a modified Model-Agnostic Meta-Learning (MAML) approach, specifically through the implementation of a Multi-Step Loss (MSL) that enhances the training stability and convergence speed for low-resource speech recognition. This method significantly reduces character error rates when applied to various low resource languages, demonstrating its effectiveness over standard MAML. It is evalu-

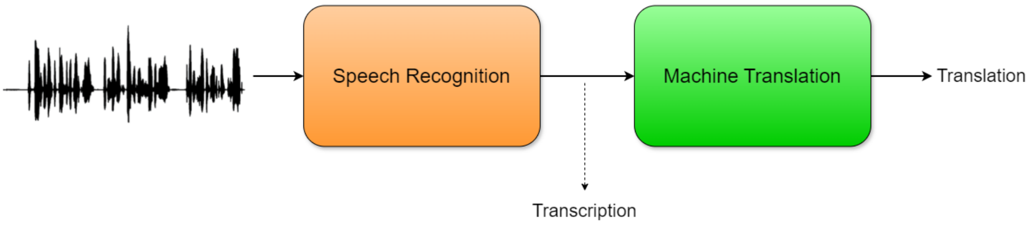


Figure 2.1: Cascaded Speech Translation Approach

ated on Common Voice v7.0 with 10 low-resource languages. Yi et al [6] built an innovative end-to-end automatic speech recognition (ASR) model that effectively integrates pretrained acoustic and linguistic encoders, specifically wav2vec2.0 [40] and BERT[8]. This model addresses the challenges of low-resource ASR by minimizing the need for labeled data during fine-tuning, while also employing a monotonic attention mechanism that aligns speech and language modalities efficiently. Zhao et al [9] explored the enhancement of Automatic Speech Recognition (ASR) systems for low-resource languages using self-supervised learning frameworks. Their findings indicate that multilingual pre-training enhances performance, while phoneme recognition tasks improve ASR outcomes. The study demonstrates significant improvements in ASR performance through the application of wav2vec2.0 pre-trained models, surpassing traditional hybrid systems. This research achieved notable success in the OpenASR21 Challenge[10], demonstrating the effectiveness of self-supervised learning in low-resource settings. Hamed et al [15] introduced a novel Automatic Speech Recognition (ASR) system specifically designed for Egyptian Arabic–English code-switching, addressing the challenges posed by low-resource dialectal languages. The researchers collected a new speech corpus named ArzEn, which includes a diverse range of code-switched sentences, and employed both DNN-based hybrid and Transformer-based end-to-end models to leverage their complementary strengths.

On the other hand, some works proposed to focus on the quantity and quality of data. Reitmaier et al [2] developed an ASR (Automatic Speech Recognition) system tailored for low-resource languages, specifically isiXhosa and Marathi, through community partnerships. It emphasizes the importance of ethical data collection, community engagement, and the need for multidisciplinary collaboration to enhance data quality and address linguistic inequalities. The study advocates for the creation of mobile-friendly transcription tools that accommodate code-switching, empowering marginalized communities to generate high-quality transcripts and retain data sovereignty. Meng et al[11] introduced a data augmentation technique for automatic speech recognition (ASR) that leverages mixup methods to enhance model training. By combining two speech features, such as mel-spectrograms, it effectively generates new training samples, leading to improved recognition accuracy. Their experiments were conducted on datasets like TIMIT[12], WSJ[13], and HKUST[14]. Building upon the concept of data augmentation commonly used in the image domain, Google introduced SpecAugment [16], a technique aimed at enriching speech recognition datasets by applying augmentation directly to the log-Mel spectrogram. SpecAugment employs methods like time-warping and adding masking blocks to both the frequency and temporal dimensions, effectively reducing the risk of overfitting in Automatic Speech Recognition (ASR) models. While this approach shifts the challenge from overfitting to underfitting, it also demonstrates enhanced performance when paired with larger model architectures and extended training durations. Beside data augmentation, there has also been research about speech synthesis to enrich the data. Notably, Kaneko et al [17] introduced CycleGAN-VC, a voice conversion (VC) method designed to work without requiring parallel data. This approach utilized a Cycle-consistent Generative Adversarial Network (CycleGAN) equipped with gated convolutional neural networks (CNNs) and an identity-mapping loss function. By combining adversarial loss and cycle-consistent loss, CycleGAN-VC effectively learned bidirectional mappings between two domains, enabling it to create pseudo pairs from unpaired data. Building on this, Kameoka et al [19] proposed StarGAN-VC, an extension of CycleGAN-VC that supports non-parallel many-to-many VC. This model incorporated a domain classifier to identify the target speaker's class, significantly broadening its application. StarGAN-VC required only a few minutes of non-parallel, unannotated speech per speaker. Hsu et al [18] developed another non-parallel VC framework called VAW-GAN, which combined a Variational Autoencoder (VAE) and a Wasserstein Generative Adversarial Network (W-GAN). The VAE modeled speaker-specific speech characteristics, while the W-GAN generated speech for various speakers.

Machine Translation(MT) The initial focus is on model-centric approaches, where researchers aim to improve methods for modeling, training, and inference. Despite the scarcity of labeled data in the form of sentence pairs, abundant monolingual textual data offers opportunities for innovation. Many studies [20, 22, 21] have embraced unsupervised learning strategies to address this limitation. First, these methods align the latent spaces of source and target languages, reducing the distance between their respective representations. This alignment is often followed by a back-translation process, where a denoising autoencoder plays a critical role. The autoencoder integrates forward translation (source to target) with backward translation (target back to source), leveraging the concept of a shared latent space between languages. The model learns to reconstruct the original source sentence from its noisy translation, capitalizing on the premise that languages share common semantic structures. Additionally, Generative Adversarial Networks (GANs) are employed to strengthen the mapping between source and target languages. The GAN architecture includes a generator that produces translations and a discriminator that differentiates authentic target language data from translations. The training process incorporates two key loss functions: a reconstruction loss, which encourages accurate bidirectional reconstruction of noisy translations, and a discrimination loss, derived from the classifier's ability to distinguish translated text from original text. Together, these techniques facilitate robust unsupervised translation performance by exploiting both linguistic similarities and adversarial learning principles.

As in speech recognition, the performance of low-resource machine translation can also be improved with data augmentation. Nag et al [23] involved extracting a portion of sentences from an existing parallel or monolingual dataset and creating new synthetic sentences by modifying words or phrases within the selected set. A common strategy is leveraging a bilingual dictionary to substitute either all words or specific uncommon terms in the chosen sentences from a monolingual corpus with their equivalents in the target

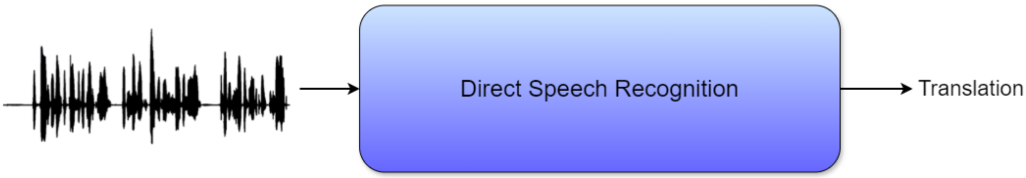


Figure 2.2: Direct Speech Translation Approach

language. This process effectively generates translated versions of the original sentences for enhancing linguistic diversity in the dataset. Sennrich et al [24] utilized back-translation technique, where a monolingual dataset in the target language is translated back into the source language using a pre-trained machine translation model. The resulting synthetic source-language sentences are paired with their corresponding original target-language sentences to create an artificial parallel corpus to use along with available data. Certain approaches leverage data mining techniques to identify sentence pairs within comparable corpora. These corpora consist of texts focused on similar topics but are not exact translations of each other. Instead, they may include segments that serve as translation equivalents, offering valuable opportunities for alignment and parallel data extraction. Mandy et al. [25] developed an approach leveraging bilingual dual encoder architectures to produce cross-lingual sentence embeddings. This method facilitates efficient sentence alignment across languages. During the sentence ranking phase, for each input sentence in one language, the model identifies a set of the most closely related sentences from another language as potential parallel pairs, based on the calculated similarity of their embeddings.

2.1.2 Low-Resource Direct Speech Translation

To solve the problem of overfitting in low-resource context, Hou et al [3, 21] proposed a module called SimAdapter for adapter-based cross-lingual speech translation. This module utilizes the attention mechanism to learn the similarity between the source and target language during fine-tuning using the adapters to boost the translation performance. They improve their performance on five low-resource languages from the Common Voice dataset[4]. Di et al [26] was one of the first to propose Speech Transformer which adapts the Transformer architecture[27] by first extracting audio feature (e.g. Fbank) with a convolutional layer before feeding to Transformer model. In Gu et al [28], convolutional neural networks is integrated with self-attention mechanisms. This design allows the model to capture both local correlations through convolutions and global interactions via self-attention. Another highlight method is non-autoregressive which predicts the whole sequence in parallel. Inaguma et al [29] featured a dual-decoder approach, combining a non-autoregressive (NAR) decoder for parallel token generation with an autoregressive (AR) decoder for candidate selection.

Other works revolved around utilizing related auxiliary tasks to enhance translation task, which is called multi-task learning. Sperber et al [30] built the model upon a two-stage architecture, where the first stage focuses on automatic speech recognition (ASR) and the second stage handles translation. Unlike traditional models that pass decoder states from the ASR to the translation component, the attention-passing model transmits context vectors, which helps alleviate the issues of error propagation commonly encountered in such systems. Instead of 2 sequential decoders, Liu et al [31] generated transcription and translation using two decoders synchronously since they considered generation processes of transcription and translation can help to improve each other. Instead of the decoupled decoder, other works [32, 33] also studied the use of dual-encoder architecture. The first encoder learns audio shallow features and the second one further learns the deep semantic representation used for translation decoding. Both encoders can be guided by information from the transcription, like phonetic details and text content.

Pino et al [34] expanded additional target language translation by using a high-quality off-the-shelf MT system on a large amount of ASR data. Inaguma et al [35] presented bidirectional sequence-level knowledge distillation (SeqKD). This method integrates both forward and backward SeqKD, where the forward SeqKD utilizes a text-based target-to-source NMT model to generate distilled translations, while the backward SeqKD leverages paraphrased source transcriptions generated from a backward NMT model.

Beside the studies above, another promising method for low-resource speech translation is pretraining. Pretraining has proven to be a highly effective strategy for enhancing model performance in low-resource scenarios. It typically utilizes readily available data sources, such as vast collections of raw text or speech data. It also often involves foundational tasks such as reconstruction, masked prediction, and contrastive learning. These methods enable the model to capture deeper contextual representations of the data. Chen et al [36] developed Masked Acoustic Modeling (MAM), a self-supervised technique that improves speech representation by randomly masking segments of the speech spectrogram and training the encoder to reconstruct them. Building on MAM, Zheng et al [37] proposed the FAT framework, which effectively integrates speech and text representations through a unified masked language modeling process. In another approach, Wang et al [38] introduced a curriculum learning strategy for encoder pretraining, designed to systematically enhance both syntactic and semantic comprehension capabilities. **This thesis places a strong emphasis on exploring the pretraining process, driven by a clear motivation that will be elaborated upon in detail in the upcoming ”Proposed Model” section.**

2.2 Audio Pretrained Models

Here exists some multi-lingual pretrained models that can be utilized for this task of speech translation. For instance, SeamlessM4T [39] (Massively Multilingual & Multimodal Machine Translation) is a multilingual model designed for seamless text and speech translation across multiple languages. The method integrates advanced self-supervised learning for speech and text. SeamlessM4T uses a combination of massive multimodal datasets, including CommonVoice, VoxPopuli, and multilingual corpora from publicly available sources. Its capacity spans dozens of languages, providing high-quality translation for both high-resource and low-resource languages. wav2vec 2.0 [40] is another model for speech representation. It learns directly from raw audio waveforms by first applying a convolutional encoder to extract low-level features and then masking portions of the latent representations. The model predicts the masked portions using a contrastive loss. Trained on large-scale speech datasets like LibriSpeech and LibriLight, wav2vec 2.0 achieves superior results in Automatic Speech Recognition (ASR) tasks with minimal labeled data. HuBERT [41] is another self-supervised speech representation model that builds upon the strengths of wav2vec. It introduces a hidden-unit prediction approach, where pseudo-labels for speech frames are iteratively refined using k-means clustering on acoustic features. Like wav2vec, HuBERT utilizes a masked prediction task, encouraging the model to learn contextualized speech representations. It is trained on large audio datasets, such as LibriSpeech and LibriLight

2.3 IWSLT competition

Our experiments follow the data and settings defined by the Low-Resource Speech Translation track of the IWSLT competition. This section details the approaches used by other teams in past IWSLT competitions for Tamasheq-French language pair.

Back to 2022, ON-TRAC [44] submitted both primary and contrastive end-to-end speech translation (ST) systems. Their primary system utilized a wav2vec 2.0 base model that was trained on 234 hours of Tamasheq audio in order to generate intermediate representations. The end-to-end ST system consisted of a partial wav2vec 2.0 module, a linear layer, and a Transformer decoder. The contrastive model used mel filterbank features as input, and it used approximate transcriptions in Tamasheq which were produced by a French phonemic ASR model. This model used a conformer architecture and jointly optimized ASR, MT, and ST losses. The use of ASR transcriptions as additional supervision was an effective strategy for low-resource settings. GMU's model [45] used the fairseq S2T extension, employing a Transformer architecture. They fine-tuned a pre-trained XLS-R 300M encoder on French and Arabic ASR, and then trained the entire model on the speech translation task using all of the data provided. However, this model, despite multilingual fine-tuning, failed to produce meaningful outputs for this particular task. TalTech [45] also submitted a system that used XLS-R and mBART-50, but this submission did not perform well either. This suggests that applying off-the-shelf pre-trained multilingual models can be challenging for low-resource tasks.

In 2023, ALEXA AI [46] submitted one primary and three contrastive systems for Tamasheq-French, all in unconstrained condition. Their systems reused the end-to-end Automatic Speech Translation (AST) model proposed by the ON-TRAC Consortium [44] in the previous IWSLT edition. This model uses a speech encoder initialized with the wav2vec 2.0 base model pre-trained on 243 hours of Tamasheq audio. The decoder is a shallow stack of two transformer layers with four attention heads. A feed-forward layer is placed between the encoder and decoder. In their work, they focused on leveraging different data augmentation techniques such as audio stretching, back translation, paraphrasing, and weighted loss. They also experimented with post-processing approaches using Large Language Models (LLMs), such as re-ranking, token masking, and sentence correction, and they also ensembled AST models trained with different seeds and data augmentation methods. NAVER [47] submitted one primary and two contrastive systems, concentrating on parameter-efficient training methods. They initialized their models with a pre-trained multilingual machine translation (MT) model, either mBART[49] or NLLB [48] and then fine-tuned this model on the ST task by inputting features extracted with a frozen pre-trained speech representation model, either wav2vec 2.0 or HuBERT. The encoder of their translation model is modified by stacking several modality-specific layers at the bottom, with adapter layers inserted between layers of the pre-trained MT model. This method allows the same model to perform both speech-to-text and text-to-text translation, maximizing knowledge transfer for improved low-resource performance. Other teams also submitted systems to this track, and it was noted that, in general, the 2023 submissions achieved better results compared to the previous year's best system. It was also observed that cascaded systems (separated speech recognition and machine translation) were not favored in this track, as none of the submitted systems were of this type.

Chapter 3

Proposed Model

3.1 System Architecture

This section details the architecture of the direct speech-to-text translation system we are using, focusing on the encoder, dimensionality reduction, and decoder components. It also details how the SAMU-XLSR[43] and DinoSR[42] models are integrated into the encoder and the overall training and fine-tuning process.

Overall System Architecture

Our system adopts a direct speech-to-text translation approach, which means that the system directly translates the input speech into a text sequence in the target language, without explicit intermediate representations such as text from the source language. The overall architecture is composed of three main components:

- Encoder: This component processes the input speech and converts it into a high-dimensional representation. The encoder is where the SAMU-XLSR and DinoSR models are integrated in our experiments.

- Dimensionality Reduction: This intermediate component reduces the high-dimensional representation from the encoder into a lower-dimensional representation suitable for the decoder. This helps to reduce computational cost and make it easier to learn the mapping between speech and text.

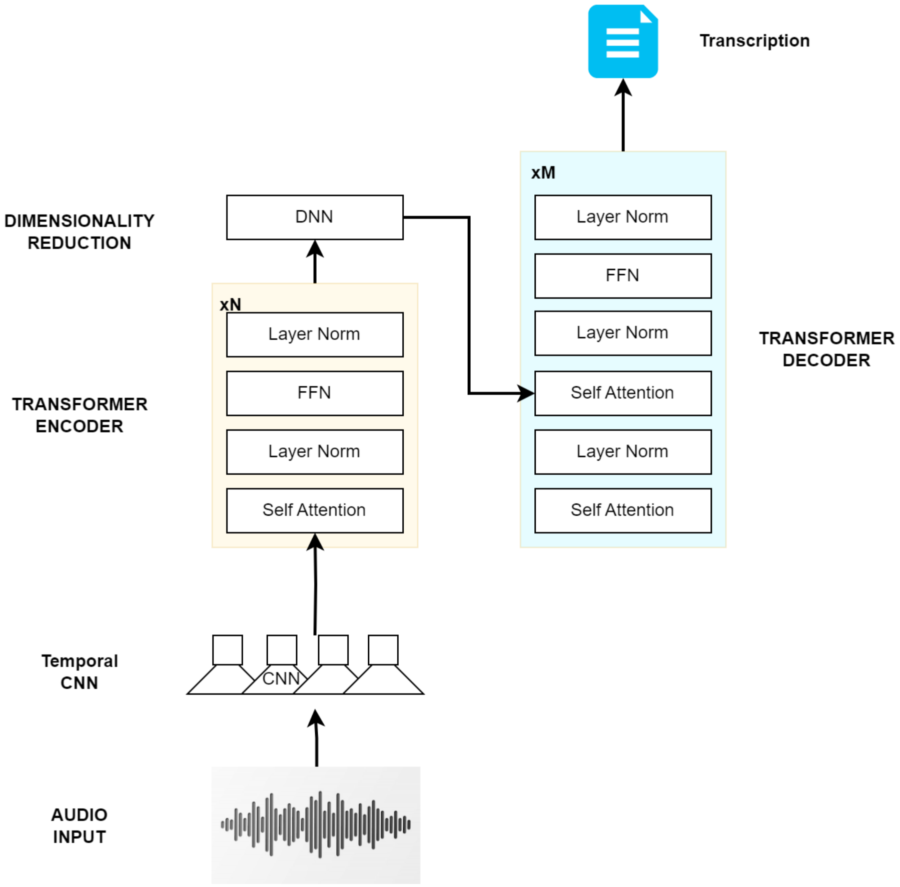


Figure 3.1: The overall architecture of our direct speech to text translation. It comprises of 3 main components: encoder (along with its temporal CNN to extract feature from raw audio), a dimensionality reduction module and a decoder

- Decoder: This component takes the reduced representation and generates the corresponding text sequence in the target language.

Encoder Details

The encoder is the core component where we incorporate different self-supervised learning models, including SAMU-XLSR and DinoSR. This component transforms the input speech waveform into a sequence of contextualized feature vectors. Here's a breakdown:

- Input: The input to the encoder is a raw speech waveform denoted as x, which is a 1D time series representing sample values of the speech signal.

- Feature Extraction:
  - SAMU-XLSR: When using SAMU-XLSR, the input speech x is first processed through a convolutional neural network (CNN) feature extractor, which is part of the pre-trained XLS-R model. This CNN maps the 1D input to a 2D sequence of feature vectors $H \in \mathbb{R}^{T \times 512}$. Each feature vector h_t in H represents a 20ms segment of speech, similar to an acoustic frame. These are considered frame-level representations.
  - DinoSR: When using DinoSR, the input waveform is first downsampled to 50Hz using a convolutional feature encoder.
- Contextualization:
  - SAMU-XLSR: The frame-level representations H are then fed into a deep transformer encoder. This transformer encoder has 24 Multi-Headed Self-Attention (MHSA) blocks. The transformer encoder converts H into contextual representations $C \in \mathbb{R}^{T \times 1024}$. Each vector in C has a dimension of 1024. These are also considered frame-level representations.
  - DinoSR: In DinoSR, a K-layer transformer encoder is used. This encoder is identical in both the student and teacher network. The input speech is partially masked for the student model to generate a masked representation $z_t^K \in \mathbb{R}^D$, where $t = 1, ..., T$ is the sequence length and D is the embedding dimension. The teacher network takes unmasked input and produces the output representation denoted as $\tilde{z}_t^K$. Both the student and teacher transformer encoders have the same architecture.
- Utterance-Level Embedding (SAMU-XLSR): To get an utterance-level embedding, a pooling mechanism is applied to C. SAMU-XLSR uses a self-attention pooling mechanism followed by a non-linear projection layer. This produces a single embedding vector $z_s \in \mathbb{R}^d$ that

represents the entire utterance. The formula can be represented as $z_s = NonLinearProjection(Pooling(C))$, where $Pooling$ represents the self-attention pooling mechanism and $NonLinearProjection$ is a trainable non-linear layer.

- Discrete Units (DinoSR): DinoSR leverages online clustering to derive discrete units from the teacher network's output which are used to guide the student network. This is done by using a codebook (set of centroids) $E^k = \{e_1^k, ..., e_V^k\}$, where k is the layer and V represents the number of codewords. A weighted sum of embeddings is used to update each codeword.

$$\begin{aligned}
\tilde{Z}_v^k &= \left\{ \tilde{z}_t^k \;\middle|\; v = \underset{i \in V}{\operatorname{argmin}} \left\| \tilde{z}_t^k - e_i^k \right\|_2 \right\}, \\
s_v^k &\longleftarrow \tau \; s_v^k + (1-\tau) \; \sum \tilde{Z}_v^k, \\
n_v^k &\longleftarrow \tau \; n_v^k + (1-\tau) \; \left| \tilde{Z}_v^k \right|, \\
e_v^k &\longleftarrow \frac{\mathbf{s}_v^k}{n_v^k}.
\end{aligned}$$

  - Where $\tilde{Z}_v^k$ is the set of teacher output frames closest to the current representation of v according to the codebook, s_v^k is the sum of neighboring teacher representations, n_v^k is the count of the neighbors, and e_v^k is the moving average of its neighbor set, and τ is the decay rate.

Dimensionality Reduction

The dimensionality reduction component takes the high-dimensional output of the encoder (either the utterance-level embedding z_s from SAMU-XLSR or the contextualized representations from DinoSR), and reduces it to a lower dimension. This is done to reduce the computational complexity and ease the decoder's training. The dimensionality reduction can be done using a simple linear layer, a non-linear layer, or a more complex technique like an attention mechanism. The reduced representation is fed to the decoder.

$z_{reduced}$ = f(z_s or $\tilde{z}_t^K$), where f is a dimensionality reduction function.

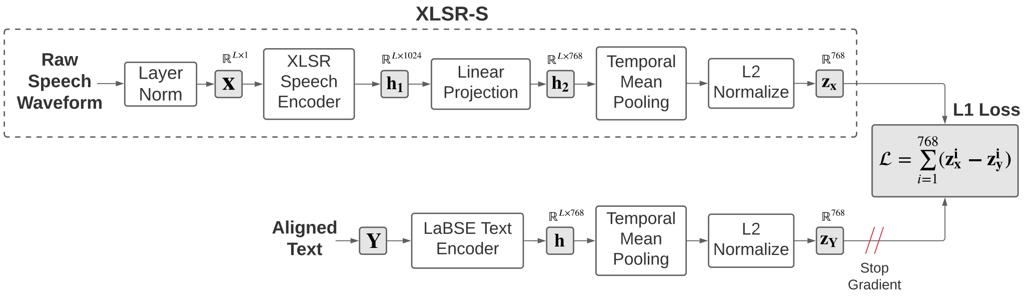


Figure 3.2: The model detail of SAMU-XLSR

For a linear layer, f can be expressed as $z_{reduced} = Wz_s + b$, where W is the weight matrix and b is the bias vector.

Decoder Details

The decoder is responsible for generating the target text sequence based on the reduced representation. It is an autoregressive model that predicts the next token given the previously predicted tokens and the encoder's output. In our design, decoder is 8-layer Transformer-based decoder. A transformer decoder is a sequence-to-sequence model which predicts the next token based on the output of the encoder, and a given sequence of tokens. The output of the decoder is a probability distribution over the vocabulary, which is used to predict the next token using the beam search technique.

Integration of SAMU-XLSR and DinoSR into the Encoder

Our research involves using both SAMU-XLSR and DinoSR for pre-training the encoder. After pre-training, we fine-tune the entire model on our low-resource speech-to-text translation task. This involves updating all parameters of the model, including encoder, dimensionality reduction, and decoder components using the target language training dataset.

In figure 3.4, we show the original architecture of SAMU-XLSR. Instead of the XLSR-S encoder, in our research, we replace it with different speech encoder, specificially HuBert, Wav2vec 2.0, DinoSR

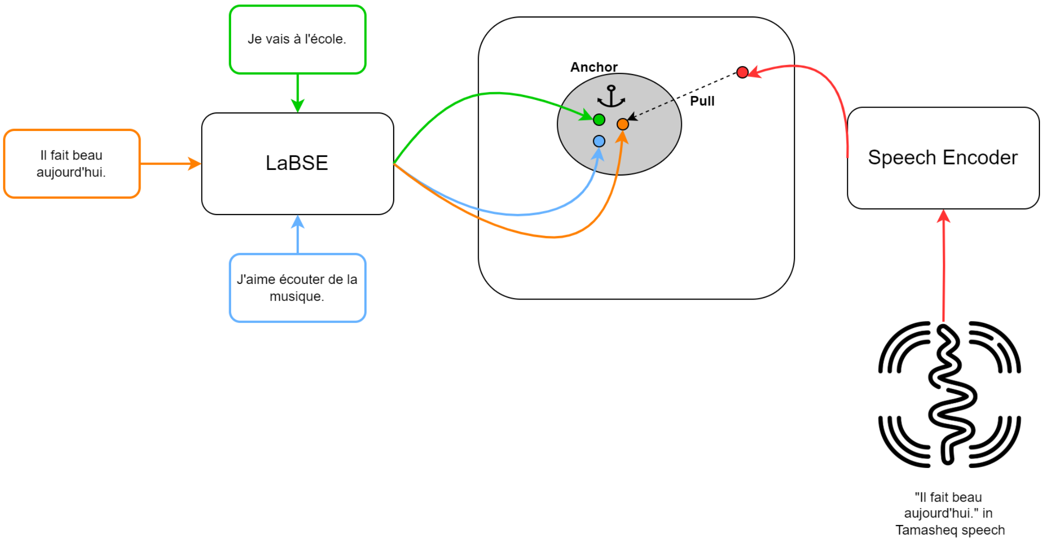

Je vais à l'école.
Il fait beau aujourd'hui.
LaBSE
J'aime écouter de la musique.
Anchor
Pull
Speech Encoder
"Il fait beau aujourd'hui." in Tamasheq speech

Figure 3.3: An explanation of how utilizing transcribed speech data with LaBSE as a guiding model can foster connections between spoken and written language across different tongues. In this example, we train with Tamasheq audio paired with its French transcription.

3.2 SAMU-XLSR

SAMU-XLSR is a framework designed to learn semantically-aligned, multimodal, utterance-level speech representations that are shared across multiple languages. Unlike previous models that focus on acoustic frame-level embeddings, SAMU-XLSR aims to create sentence-level embeddings where the spoken utterance is clustered together with its speech and text translations in various languages. This is achieved by leveraging a pre-trained multilingual speech encoder, XLS-R, and a language-agnostic text encoder, LaBSE, through a knowledge distillation process.

The key concepts driving SAMU-XLSR's design are multimodality, cross-linguality, semantic alignment, and utterance-level representation. Multimodality signifies that the embedding space is shared between both speech and text modalities, enabling direct comparison and interaction between them. Cross-linguality ensures that this shared space is consistent across different languages, facilitating cross-lingual understanding. Crucially, SAMU-XLSR aims for semantic alignment, meaning that semantically similar utter-

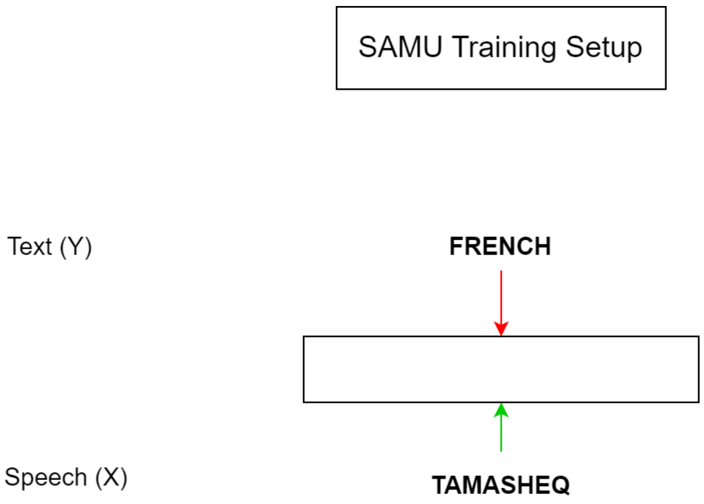


Figure 3.4: The training setup of SAMU-XLSR in our research. The model is trained to push the embedding of Tamasheq speech to be as close as possible with the text embedding in French

ances, regardless of their language or modality (speech or text), are clustered closely together in the embedding space. This is achieved through utterance-level processing, focusing on representations at the sentence level (approximately 5-10 seconds of speech) rather than the frame level (10-20 milliseconds). This broader context is essential for capturing the complete meaning of an utterance. Ultimately, SAMU-XLSR aims to enable zero-shot translation, where speech utterances and their translations are naturally clustered together in the embedding space without requiring explicit parallel training data.

The SAMU-XLSR model combines a pre-trained frame-level speech encoder (XLS-R) with a mechanism for pooling the frame-level representations into a single utterance-level embedding vector. This model is trained using transcribed speech data. The framework's architecture can be broken down into the following key parts:

- Speech Encoder: The input speech waveform is processed by a deep convolutional neural network (CNN), which maps it to a sequence of feature vectors. The feature sequence is then transformed into contextual representations by a deep transformer encoder.

- Pooling Mechanism: To transform the frame-level contextual representations into a single utterance-level embedding vector, a pooling mechanism is applied.

- LaBSE Text Encoder: The LaBSE text encoder, g_ϕ, is used to create sentence embeddings z_T from text transcripts. LaBSE has a transformer encoder with 12 attention layers that process tokens, and then a pooling step creates a compact sentence-level vector. The input text is tokenized into word pieces, and the CLS token embedding is used as the sentence embedding z_T. The input text is tokenized into word pieces, and the CLS token embedding is used as the sentence embedding z_T. This text encoder is language-agnostic and produces embeddings in a semantically aligned vector space shared across 109 languages.

Through training, cross-lingual speech-to-text and speech-to-speech associations emerge in the learned representation space. This happens without the model explicitly being trained on cross-lingual data during training. The model learns to cluster speech utterances with their corresponding text and speech translations in different languages, thereby creating a semantically-aligned, multimodal, cross-lingual embedding space.

In our research, we experiment with starting from some pretrained speech encoder, pretraining that speech encoder using SAMU. As we are working with Tamasheq-French translation, the pretraining setup runs on the dataset of Tamasheq speech and French text, aims to push these 2 multimodal multilingual data into the same feature space.

3.3 DinoSR

DinoSR is a self-supervised speech representation learning framework that combines masked language modeling, self-distillation, and online clustering. It aims to learn strong speech representations by leveraging the complementary strengths of these three techniques.

DinoSR leverages several key concepts and goals to achieve state-of-the-art performance in speech recognition, particularly in low-resource settings.

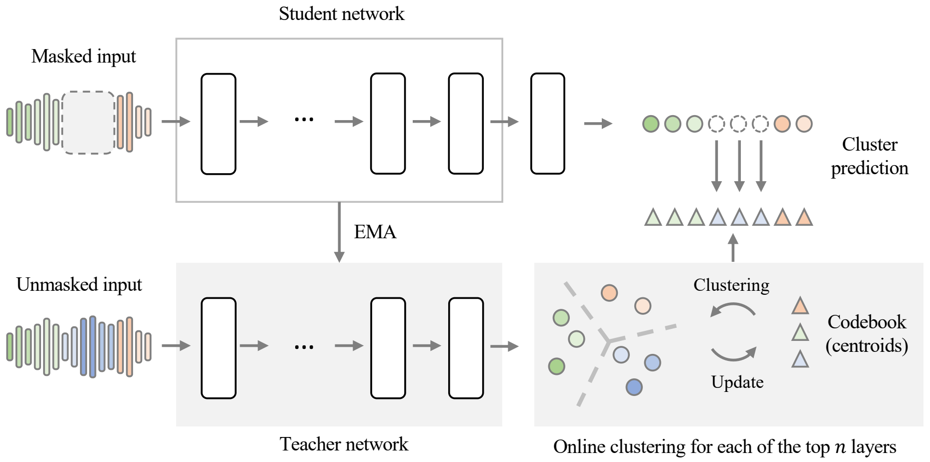


Figure 3.5: A summary of DinoSR: The teacher model is updated using an exponential moving average of the student model and processes unmasked speech to generate target features. Multiple layers of the teacher employ independent codebooks for online clustering. The student model learns by predicting the cluster assignments of masked input. Neither the teacher model nor the clustering mechanism (highlighted areas) require gradient computation. Source: [42]

At its core, DinoSR employs self-supervised learning, enabling it to learn directly from unlabeled data by exploiting the inherent structure of speech. This eliminates the dependency on large, costly, and often unavailable transcribed datasets. A crucial component of DinoSR is Masked Language Modeling (MLM), a technique borrowed from natural language processing. In this context, MLM involves masking portions of the speech signal and training the model to predict the masked segments based on the surrounding context. This process encourages the model to learn rich, contextualized representations of the speech signal. DinoSR further incorporates self-distillation, utilizing a teacher-student framework. The student network learns from a teacher network, which is an exponentially moving average of the student itself. This approach provides a more stable learning target, leading to improved training stability and performance. Another key innovation is the use of online clustering. DinoSR employs an online clustering system to dynamically discover a discrete inventory of acoustic units directly from the teacher's embeddings. This allows the model to effectively create its own "dictionary" of acoustic units without relying on any prior phonetic knowledge or human-defined transcriptions. This process of acoustic unit discovery results in units that closely align with human phonetic understanding, enhancing the model's interpretability. By effectively combining these techniques – self-supervised learning, masked language modeling, self-distillation, and online clustering for acoustic unit discovery – DinoSR achieves significant performance gains, surpassing previous state-of-the-art results in both limited-resource speech recognition and unsupervised acoustic unit discovery tasks.

The DinoSR architecture is built around a teacher-student framework, comprising two networks that share an identical K-layer transformer encoder structure. The teacher network ($\theta_{teacher}$) processes the input audio without any masking, generating contextualized embeddings that capture the full context of the audio sequence. Its parameters are updated using an exponential moving average (EMA) of the student's parameters, providing a more stable and generalized learning target. Crucially, the teacher network serves two primary functions: it provides the target features that the student network is trained to predict, and it is responsible for performing the online clustering that discovers the discrete acoustic units. The student network ($\theta_{student}$), on the other hand, processes a masked version of the same audio input. This masking forces the student to learn contextual representations by predicting the masked portions based on the surrounding unmasked context. The student network is trained to predict the clustered outputs produced by the teacher network, effectively learning to mimic the teacher's representations. While the student network's parameters are updated using standard gradient descent, the teacher's parameters are updated using the EMA of the student's parameters, creating a dynamic and mutually beneficial learning process.

DinoSR employs an online clustering mechanism to discover a discrete phone inventory directly from the teacher network's representations. This process operates on the top N layers of the teacher network. For each layer k within these top N layers, a codebook E^k is introduced, consisting of V codewords (e_i^k), where each codeword is a vector in a D-dimensional embedding space. The codebook update occurs in two distinct steps. First, during the assignment step, each teacher output frame ($\tilde{z}_t^k$) is assigned to the closest codeword in the codebook E^k. This is achieved by finding the codeword e_v^k that minimizes the Euclidean distance to the output frame, effectively finding the nearest neighbor:

$$v = argmin_{i \in V} ||\tilde{z}_t^k - e_i^k||^2$$

Following the assignment step, the update step refines the codewords. Each codeword e_v^k is updated using an exponential moving average (EMA) of all the teacher output frames that have been assigned to it. This update is formalized as:

$$\begin{aligned}
\tilde{Z}_v^k &= \left\{ \ \tilde{z}_t^k \ \middle| \ v = \underset{i \in V}{\operatorname{argmin}} \left\| \tilde{z}_t^k - e_i^k \right\|_2 \right\}, \\
s_v^k &\longleftarrow \tau \ s_v^k + (1-\tau) \ \sum \tilde{Z}_v^k, \\
n_v^k &\longleftarrow \tau \ n_v^k + (1-\tau) \ \left| \tilde{Z}_v^k \right|, \\
e_v^k &\longleftarrow \frac{\mathbf{s}_v^k}{n_v^k}.
\end{aligned}$$

where $\tilde{Z}_v^k$ represents the set of teacher output frames closest to the codeword e_v^k, and τ is a codebook decay rate that controls the influence of past updates on the current codeword value. This online clustering process dynamically adapts the codebooks to the evolving representations learned by the teacher network, effectively discovering a set of acoustic units that capture the underlying phonetic structure of the speech data.

The self-distillation training process in DinoSR is a crucial aspect of its learning mechanism. The process begins with input masking: the input audio is partially masked before being fed to the student model, while the teacher model receives the original, unmasked audio. This difference in input is key to the knowledge distillation process. The student network processes the masked input, producing a masked representation z_t^K, while the teacher model processes the unmasked input, generating the unmasked representation $\tilde{z}_t^K$. Following the forward pass, the teacher network's parameters are updated using an exponential moving average (EMA) of the student's parameters. This update is governed by the equation

$$\theta_{teacher} \longrightarrow \lambda\theta_{teacher} + (1 - \lambda)\theta_{student}$$

where λ represents the decay rate of the teacher model at each training step. This EMA update ensures that the teacher network maintains a stable and generalized view of the learned representations, acting as a consistent target for the student. The core training objective for the student network is to predict the codeword index of the corresponding frame from the teacher model's output. In other words, the student tries to match the discrete acoustic units discovered by the teacher. This is achieved by minimizing a loss function defined as

$$\sum_{t \in M} \sum_{k \in (K-N, K]} \log p_{\phi_k}(v | \mathbf{z}_t^K)$$

where M is the set of masked timesteps, ϕ_k is a linear projection followed by a softmax activation applied to the student's output, and v represents the index of the nearest codeword identified by the teacher's online clustering process. By minimizing this loss, the student learns to align its representations with the clustered representations of the teacher.

The pre-training phase of DinoSR utilizes a substantial dataset of 960 hours of speech drawn from the LibriSpeech corpus. The architecture employed for pre-training is a base-sized transformer with K = 12 layers and an embedding dimension of D = 768. The raw 16 kHz input waveform undergoes a downsampling process to 50Hz using a convolutional feature encoder, preparing it for input into the model. For the student model, a significant portion of the input features, specifically 80%, is masked. To ensure contextual learning, each masked span is constrained to be no shorter than 10 frames. The online clustering process, which discovers the discrete acoustic units, is performed on the top N = 8 layers of the teacher network. Each of these layers maintains a codebook consisting of V = 256 codewords. Optimization of the student model is carried out using the Adam optimizer with a learning rate schedule that ramps up to a peak of 0.0005 and then undergoes exponential decay. The computational demands of this pre-training phase are considerable, requiring approximately 180 hours of training time on a cluster of 16 NVIDIA V100 GPUs.

DinoSR presents several key differences compared to existing self-supervised pre-training methods. Unlike HuBERT, which uses a multi-stage, iterative training process, DinoSR employs a more streamlined, one-stage approach. Furthermore, DinoSR performs clustering online during training, rather than relying on offline clustering as in HuBERT. When compared to Wav2vec 2.0, another prominent method, DinoSR diverges in its training objective. While Wav2vec 2.0 utilizes contrastive learning coupled with vector quantization, DinoSR focuses on cluster prediction as its primary learning task.

3.4 Pretrained Models

HuBERT (Hidden Unit BERT) HuBERT is a self-supervised learning method that uses masked language modeling (MLM) to predict discrete units derived from the speech signal. Unlike models that predict raw audio features, HuBERT targets automatically discovered hidden units.

HuBERT starts by generating initial discrete units using a simple clustering algorithm (like K-means) on the raw audio features, or even random linear projections. During training, the input speech is masked, and the model is trained to predict the discrete units corresponding to the masked regions, based on the surrounding context. The key feature of HuBERT is its iterative refinement process. After the initial training, HuBERT uses the pre-trained model's output to generate new, improved discrete units by running offline K-means clustering. This process can be repeated multiple times to further enhance the quality of the discrete units and the learned representations. The model can be re-trained with the new targets. HuBERT relies on offline k-means clustering to generate the discrete targets, which can be computationally expensive and requires careful tuning of hyperparameters. HuBERT employs a transformer encoder as its core architecture to learn the contextualized representations.

wav2vec 2.0

wav2vec 2.0 is a self-supervised learning framework that learns speech representations through a contrastive learning task using vector quantization (VQ). It focuses on learning robust representations at the acoustic frame-level. The raw audio is passed through a convolutional neural network to extract a sequence of feature vectors. The extracted features are then quantized into discrete units using vector quantization. The model is trained to distinguish between the true quantized representation of a masked audio segment and a set of distractors. wav2vec 2.0 masks portions of the input feature sequence and then trains the model to predict the quantized representation of the masked segment using the unmasked context. Vector Quantization is also used to convert the continuous representations into discrete ones, which are then used in the contrastive task Similar to HuBERT, wav2vec 2.0 utilizes a transformer network to process the contextualized representations.

Compared to HuBERT, both wav2vec 2.0 and HuBERT utilize masking strategies during training; however, their objectives differ. wav2vec 2.0 focuses on a contrastive task with vector quantization, aiming to distinguish true latent representations from negative samples. In contrast, HuBERT employs a masked language modeling (MLM) approach, predicting discrete units derived from offline clustering of acoustic features. This distinction leads wav2vec 2.0 to learn frame-level representations, while HuBERT captures more contextualized representations that can be aggregated into utterance-level embeddings.

Compared to DinoSR, DinoSR introduces an online clustering mechanism, differing from wav2vec 2.0's contrastive objective with vector quantization. This online clustering allows DinoSR to adapt its cluster assignments dynamically during training, potentially leading to more efficient learning processes. In contrast, wav2vec 2.0 relies on a fixed vector quantization process to generate discrete units for its contrastive task.

In comparison to SAMU-XLSR, SAMU-XLSR is designed to learn semantically aligned multimodal utterance-level cross-lingual speech representations. Unlike wav2vec 2.0, which focuses on frame-level acoustic details, SAMU-XLSR emphasizes capturing semantic information across different languages at the utterance level. This approach facilitates tasks that require understanding the overall meaning of spoken sentences, whereas wav2vec 2.0 is more suited for tasks that benefit from detailed acoustic modeling.

3.5 Pre-training Strategies

Pre-training has become an essential step in the field of speech processing, allowing models to learn powerful representations from massive amounts of unlabeled data, which can then be fine-tuned for specific downstream tasks. This approach can drastically reduces the need for labeled data and improves model performance.

Using SAMU-XLSR to pre-train the encoder component of a speech translation system:

- SAMU-XLSR is trained using transcribed speech data. The model learns by minimizing the distance between the speech embedding produced by SAMU-XLSR and the text embedding of the corresponding transcript provided by LaBSE. Because LaBSE's embedding space is semantically aligned across languages, this process allows SAMU-XLSR to learn cross-lingual speech-text associations without explicitly seeing cross-lingual training pairs.

- The pre-trained SAMU-XLSR encoder can then be used as the encoder component in a speech translation system. The idea is that the encoder will produce semantically rich and cross-lingually aligned representations that improve translation performance.

- The pre-trained SAMU-XLSR encoder is then fine-tuned using a supervised speech translation dataset. This fine-tuning adjusts the encoder's parameters for the specific speech translation task, making use of the general cross-lingual speech and text representations acquired during pre-training.

Using DinoSR to pre-train the encoder component of a speech translation system:

- DinoSR first extracts contextualized embeddings from the input audio using the teacher network, then runs an online clustering system on these embeddings to create a machine-discovered phone inventory. The student network is then trained to predict these cluster assignments from masked input audio.

- The pre-trained student encoder from DinoSR can be used as the encoder component in a speech translation system.
- The pre-trained encoder is then fine-tuned on a task-specific dataset for speech translation. This stage adapts the encoder's learned features from a phonetic space to the specifics of the speech translation task, while retaining the discrete units that capture phonetic content.

Pre-training a model with DinoSR, then with SAMU-XLSR, and then fine-tuning it:

This multi-stage pre-training approach combines the strengths of both DinoSR and SAMU-XLSR by first pretraining the speech encoder with DinoSR and then continuing to pretrain with SAMU. Finally, that pretrained encoder is used to finetune on speech translation task

- First Stage (DinoSR): The model is first pre-trained using DinoSR, allowing the encoder to learn strong, discrete phonetic representations of speech using a clustering method.
- Second Stage (SAMU-XLSR): The DinoSR pre-trained encoder is then used as a basis for pre-training with SAMU-XLSR. In this stage, the encoder learns cross-lingual, semantically aligned representations by using text transcriptions as an anchor point using LaBSE.
- Fine-tuning: The resulting encoder is then fine-tuned on a speech translation dataset. This process combines the phonetic robustness from DinoSR with the cross-lingual and multimodal understanding from SAMU-XLSR to potentially improve overall translation performance.

Chapter 4

Evaluation

4.1 Data

The Tamasheq-French dataset was a key component of the low-resource speech translation shared task at the 19th International Conference on Spoken Language Translation (IWSLT) in 2022. This task focused on developing speech translation tools for languages with limited resources. The Tamasheq language, being primarily oral, falls under this category. The goal was to translate Tamasheq speech into French text.

Dataset Composition

- Tamasheq Speech: The dataset includes approximately 17 hours of Tamasheq speech. This speech data comprises 5,829 utterances that have been translated into French.

- French Translations: Each of the Tamasheq utterances has a corresponding French text translation. This forms the parallel data necessary for training speech translation models.

- Additional Audio Data: Besides the parallel data, additional audio data was made available, including:

  – 224 hours of Tamasheq audio.

  - 417 hours of audio in geographically close languages, such as French from Niger, Fulfulde, Hausa, and Zarma. It is important to note that this additional audio data does not have transcriptions.

- Speech Style: All of the speech data in the dataset is characterized by a radio broadcasting style.

Data Characteristics

The dataset is considered low-resource due to the limited amount of parallel speech-translation data, typical of many of the world's languages. This contrasts with high-resource language pairs, which often have large quantities of training data. Tamasheq is a predominantly oral language, which means there's a lack of written text, making it challenging to adapt techniques from text-based machine translation. The additional 224 hours of Tamasheq audio and 417 hours of audio in geographically close languages are provided without transcriptions.

The speech translation systems in this study are categorized based on the resources they are permitted to utilize during training. We follow the original setup of the competition to categorize the systems as follow:

- Constrained Systems: These systems operate under strict limitations regarding the training data. They are restricted to a medium-sized framework, primarily to manage training time and resource consumption. Crucially, constrained systems are not permitted to use any pre-trained language models.

- Constrained with Large Language Models (Constrained+LLM) Systems: This category builds upon the constrained setting by allowing the inclusion of a select group of large language models. While still subject to limitations on the overall training data (similar to the constrained systems), these systems can leverage the knowledge encoded within pre-trained LLMs.

- Unconstrained Systems: In contrast to the constrained approaches, unconstrained systems face virtually no restrictions on the resources they can employ. They are free to utilize any available data, including large language models, pre-trained models of all kinds, and any other relevant resource (excluding the evaluation datasets themselves).

4.2 Evaluation Metric

The Bilingual Evaluation Understudy (BLEU) score is a metric for automatically evaluating machine-translated text. It's one of the most widely used metrics in natural language processing (NLP), particularly in machine translation. BLEU works by comparing the machine-generated translation against one or more human-produced reference translations. The core idea is that a good translation should have a high degree of overlap with the reference translations. This overlap is measured using n-gram precision, combined with a brevity penalty to account for overly short translations.

The final BLEU score is calculated as follows:

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n log(p_n))$$

Where:

BP is the brevity penalty. N here represents the maximum n-gram order (typically 4). p_n is the modified n-gram precision for n-grams of order n. w_n are the weights for each n-gram order. Usually, uniform weights are used (e.g., $w_n = \frac{1}{N}$ for all n). In simpler terms, the BLEU score is the brevity penalty multiplied by the geometric mean of the modified n-gram precisions.

For example, let's consider a simple example with one reference translation and one candidate translation:

Reference: "the cat sat on the mat"

Candidate: "the cat sat on mat"

Let's calculate the BLEU score with N=2 (bigrams):

Unigram precision: 5/5 = 1.0 (after clipping)

Bigram precision: 4/4 = 1.0 (after clipping)

Candidate length (c): 5

Reference length (r): 6

Brevity penalty (BP): $e^{1-6/5} \approx 0.82$

BLEU score = 0.82 * exp((1/2 * log(1.0)) + (1/2 * log(1.0))) = 0.82

4.3 Result

This section details the performance of our proposed direct speech translation model for the Tamasheq-French language pair. We conducted experiments under constrained and constrained with large language models (constrained+LLM) settings, comparing our approach against the current state-of-the-art (SOTA) system developed by NAVER, which operates under unconstrained conditions.

Table 4.1: Results of pretraining by SAMU on wav2vec 2.0 and HUBERT for Tamasheq-French dataset

	System constraint	Valid BLEU	Test BLEU
NAVER	Unconstrained	N/A	23.59
wav2vec2-base-960h	Constrained with LLM	2.32	1.72
wav2vec2-base-960h+SAMU	Constrained with LLM	2.57	2.06
HUBERT	Constrained with LLM	2.16	1.67
HUBERT+SAMU	Constrained with LLM	8.02	6.23
DINOSR	Constrained with LLM	4.12	3.84
DINOSR+SAMU	Constrained with LLM	8.34	6.89

Table 4.1 presents the BLEU scores obtained on the validation and test sets for our models and the NAVER system. The NAVER system, utilizing a substantial amount of diverse audio data including 17 hours of Tamasheq speech with French transcriptions, 111 hours of French audio, 109 hours of Fulfulde audio, 100 hours of Hausa audio, and 95 hours of Zarma audio, achieves a test BLEU score of 23.59.

Our experiments focus on leveraging pre-trained models and a novel pre-training approach using SAMU. Specifically, we fine-tuned the wav2vec 2.0 base model pre-trained for 960 hours and the HUBERT model, both under constrained+LLM conditions. Direct fine-tuning of wav2vec2-base-960h yields BLEU scores of 2.32 on the validation set and 1.72 on the test set. Applying our proposed SAMU pre-training to wav2vec2-base-960h before fine-tuning results in a modest improvement, achieving 2.57 on the validation set and 2.06 on the test set. Similarly, fine-tuning the HUBERT model achieves BLEU scores of 2.16 and 1.67 on the validation and test sets, respectively. However, pre-training HUBERT with SAMU significantly boosts performance, leading to BLEU scores of 8.02 on the validation set and 6.23 on the test set.

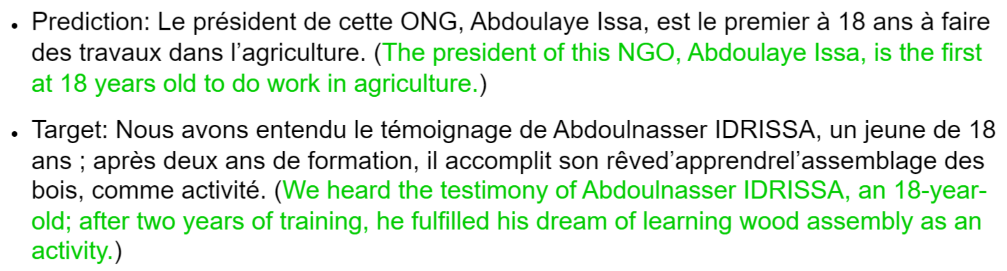
Furthermore, we explored the impact of DINOSR pre-training. Fine-tuning the DINOSR model results in BLEU scores of 4.12 and 3.84 on the validation and test sets. Applying SAMU pre-training on top of DINOSR before fine-tuning further improves the results, yielding BLEU scores of 8.34 on the validation set and 6.89 on the test set.

Table 4.2: The data used by the State-of-the-art method and ours

Method	Data
NAVER	17 hours of Tamasheq speech - French transcription 111 hours of French audio 109 hours of Fulfulde audio 100 hours of Hausa audio 95 hours of Zarma
Our methods	17 hours of Tamasheq speech - French transcription

Table 4.2 details the data used by the NAVER system and our proposed methods. While NAVER leverages a large and diverse multilingual audio dataset, our models are trained using only 17 hours of Tamasheq speech with French transcriptions, adhering to the constrained+LLM setting. This highlights the potential of our proposed SAMU pre-training method to effectively leverage limited resources and significantly improve the performance of direct speech translation systems, especially when combined with strong acoustic models like HUBERT and DINOSR. Despite the significant data disparity, our best system achieves a respectable performance, demonstrating the effectiveness of our approach in low-resource scenarios.

4.4 Error Analysis And Discusion



- Prediction: Le président de cette ONG, Abdoulaye Issa, est le premier à 18 ans à faire des travaux dans l’agriculture. (The president of this NGO, Abdoulaye Issa, is the first at 18 years old to do work in agriculture.)
- Target: Nous avons entendu le témoignage de Abdoulnasser IDRISSA, un jeune de 18 ans ; après deux ans de formation, il accomplit son rêved’apprendrel’assemblage des bois, comme activité. (We heard the testimony of Abdoulnasser IDRISSA, an 18-year-old; after two years of training, he fulfilled his dream of learning wood assembly as an activity.)

Figure 4.1: A sample of error remained in our best combination of DinoSR and SAMU

Our error analysis reveals that the model’s mistakes are not primarily due to grammatical errors, incorrect word order, or repetitive phrases, which would suggest decoder issues. Instead, the errors point to weaknesses in the encoder and a mix of encoder and decoder problems. Specifically, as shown in 4.1 the model fails to grasp the meaning of words and phrases within the larger context of the sentence. In the example, the model translates ”Abdoulaye Issa” as the president of an NGO doing agricultural work, while the target refers to ”Abdoulnasser IDRISSA” fulfilling his dream of learning woodworking. The model missed the connection between the name and the described activity. Moreover, the model misinterprets words with multiple meanings. It chooses the wrong meaning for words in the source language, leading to incorrect translations. Finally, the model’s translation drifts away from the original meaning of the source text. It sometimes even ”hallucinates” content that is not present in the source at all. This suggests problems with both understanding the input (encoder) and generating the correct output (decoder). The example shows a clear hallucination, as the target text mentions woodworking, not agriculture. From these consideration, there may be more room for improvement on encoder than with the decoder.

Chapter 5

Conclusion

In this work, we investigated the impact of different pretraining techniques, specifically DINOSR and our proposed SAMU, on the performance of direct speech translation models for the low-resource Tamasheq-French language pair within the IWSLT competition framework. Our experiments showed that pretraining with DINOSR and SAMU helps. Using these methods made our model better at learning from the limited data. We also found that combining DINOSR and SAMU pretraining gave us the best results of all our experiments. This shows that our pretraining methods are promising for improving speech translation when data is scarce.

Even though our best results are not as good as the top system, that system used a lot more data than we did. We only used the data given to us for the IWSLT competition, while the top system used much more data from other sources.

Furthermore, our method can be easily combined with other improvements, like better decoders or different ways of training the model.

Future research directions could include:

- Combining the described pretraining techniques with other methods such as better decoders.
- Exploring different model architectures, including end-to-end and cascaded approaches, and how they interact with pre-training methods

Bibliography

[1] Gupta, Mahendra and Dutta, Maitreyee and Maurya, Chandresh Kumar. Direct Speech-to-Speech Neural Machine Translation: A Survey, *arXiv preprint arXiv:2411.14453* (2024)

[2] Reitmaier, Thomas, Electra Wallington, Dani Kalarikalayil Raju, Ondrej Klejch, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. Opportunities and challenges of automatic speech recognition systems for low-resource language speakers. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1-17 (2022)

[3] Hou, Wenxin, Han Zhu, Yidong Wang, Jindong Wang, Tao Qin, Renjun Xu, and Takahiro Shinozaki. Exploiting adapters for cross-lingual low-resource speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 30* pp. 317-329 (2021)

[4] Ardila, Rosana, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670* (2019).

[5] Singh, Satwinder, Ruili Wang, and Feng Hou. Improved meta learning for low resource speech recognition. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4798-4802. IEEE, 2022.

[6] Yi, Cheng, Shiyu Zhou, and Bo Xu. Efficiently fusing pretrained acoustic and linguistic encoders for low-resource speech recognition. *IEEE Signal Processing Letters 28* pp. 788-792 (2021)

[7] Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems 33*, pp.12449-12460 (2020)

[8] Kenton, Jacob Devlin Ming-Wei Chang, and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of naacL-HLT*, vol. 1, no. 2. (2019)

[9] Zhao, Jing, and Wei-Qiang Zhang. Improving automatic speech recognition performance for low-resource languages with self-supervised models. *IEEE Journal of Selected Topics in Signal Processing 16*, no. 6, pp.1227-1241 (2022)

[10] Peterson, Kay, Audrey Tong, and Yan Yu. OpenASR21: The Second Open Challenge for Automatic Speech Recognition of Low-Resource Languages. *INTERSPEECH*, pp. 4895-4899. (2022)

[11] Meng, Linghui, Jin Xu, Xu Tan, Jindong Wang, Tao Qin, and Bo Xu. Mixspeech: Data augmentation for low-resource automatic speech recognition. *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7008-7012. IEEE, (2021)

[12] Garofolo, John S., Lori F. Lamel, William M. Fisher, David S. Pallett, Nancy L. Dahlgren, Victor Zue, and Jonathan G. Fiscus. TIMIT acoustic-phonetic continuous speech corpus. (1993).

[13] Hershey, John R., Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 31-35. IEEE, (2016)

[14] Liu, Yi, Pascale Fung, Yongsheng Yang, Christopher Cieri, Shudong Huang, and David Graff. Hkust/mts: A very large scale mandarin telephone speech corpus. *Chinese Spoken Language Processing: 5th International Symposium, ISCSLP 2006*, pp. 724-735, (2006)

[15] Hamed, Injy, Pavel Denisov, Chia-Yu Li, Mohamed Elmahdy, Slim Abdennadher, and Ngoc Thang Vu. Investigations on speech recognition systems for low-resource dialectal Arabic–English code-switching speech. *Computer Speech Language 72*, pp.101278 (2022)

[16] Park, Daniel S., William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, SpecAugment: A simple data augmentation method for automatic speech recognition, *Proc. Interspeech* , pp. 2613–2617 (2019)

[17] Kaneko, Takuhiro, and Hirokazu Kameoka. Parallel-data-free voice conversion using cycle-consistent adversarial networks. *arXiv preprint* arXiv:1711.11293 (2017)

[18] Hsu, Chin-Cheng, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang. Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. *arXiv preprint* arXiv:1704.00849 (2017)

[19] Kameoka, Hirokazu, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 266-273. IEEE, (2018)

[20] Chronopoulou, Alexandra, Dario Stojanovski, and Alexander Fraser. Reusing a pretrained language model on languages with limited corpora for unsupervised nmt. *arXiv preprint arXiv:2009.07610* (2020)

[21] Edman, Lukas, Antonio Toral Ruiz, and Gertjan van Noord. Low-resource unsupervised NMT: Diagnosing the problem and providing a linguistically motivated solution. *22nd Annual Conference of the European Association for Machine Translation*, pp. 81-90. (2020)

[22] Graça, Yunsu Kim Miguel, and Hermann Ney. When and why is unsupervised neural machine translation useless. *22nd Annual Conference of the European Association for Machine Translation*, p. 35. (2020)

[23] Nag, Sreyashi, Mihir Kale, Varun Lakshminarasimhan, and Swapnil Singhavi. Incorporating bilingual dictionaries for low resource semi-supervised neural machine translation. *arXiv preprint*, arXiv:2004.02071 (2020)

[24] Sennrich, Rico. Improving neural machine translation models with monolingual data. *arXiv preprint*, arXiv:1511.06709 (2015)

[25] Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Effective parallel corpus mining using bilingual sentence embeddings. In Proceedings of the 3rd Conference on Machine Translation: Research Papers. 165–176, (2018)

[26] Di Gangi, Mattia A., Matteo Negri, and Marco Turchi. Adapting transformer to end-to-end spoken language translation. *Proceedings of INTERSPEECH 2019*, pp. 1133-1137, (2019)

[27] Waswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *NIPS*. (2017)

[28] Gulati, Anmol, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han et al. *Conformer: Convolution-augmented transformer for speech recognition.* arXiv preprint arXiv:2005.08100 (2020)

[29] Inaguma, Hirofumi, Yosuke Higuchi, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. Orthros: Non-autoregressive end-to-end speech translation with dual-decoder. *ICASSP 2021-2021 IEEE International Conference on Acoustics*, pp. 7503-7507. IEEE, (2021)

[30] Sperber, Matthias, Graham Neubig, Jan Niehues, and Alex Waibel. Attention-passing models for robust and data-efficient end-to-end speech

translation. *Transactions of the Association for Computational Linguistics 7*, pp. 313-325 (2019)

[31] Liu, Yuchen, Jiajun Zhang, Hao Xiong, Long Zhou, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. Synchronous speech recognition and speech-to-text translation with interactive decoding. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 8417-8424. (2020)

[32] Liu, Yuchen, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. Bridging the modality gap for speech-to-text translation. *arXiv preprint* arXiv:2010.14920 (2020)

[33] Dong, Qianqian, Rong Ye, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. Listen, understand and translate: Triple supervision decouples end-to-end speech-to-text translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 14, pp. 12749-12759. (2021)

[34] Pino, Juan, Qiantong Xu, Xutai Ma, Mohammad Javad Dousti, and Yun Tang. Self-training for end-to-end speech translation. *arXiv preprint* arXiv:2006.02490 (2020)

[35] Inaguma, Hirofumi, Tatsuya Kawahara, and Shinji Watanabe. Source and target bidirectional knowledge distillation for end-to-end speech translation. *arXiv preprint* arXiv:2104.06457 (2021)

[36] Chen, Junkun, Mingbo Ma, Renjie Zheng, and Liang Huang. Mam: Masked acoustic modeling for end-to-end speech-to-text translation. *arXiv preprint* arXiv:2010.11445 (2020)

[37] Zheng, Renjie, Junkun Chen, Mingbo Ma, and Liang Huang. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation. *International Conference on Machine Learning*, pp. 12736-12746. PMLR, (2021)

[38] Wang, Chengyi, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. Curriculum pre-training for end-to-end speech translation. *arXiv preprint* arXiv:2004.10093 (2020)

[39] Barrault, Loïc, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar et al. SeamlessM4T-Massively Multilingual & Multimodal Machine Translation. *arXiv preprint* arXiv:2308.11596 (2023)

[40] Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems 33*, pp. 12449-12460 (2020)

[41] Hsu, Wei-Ning, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing 29*, pp. 3451-3460, (2021)

[42] Liu, Alexander H., Heng-Jui Chang, Michael Auli, Wei-Ning Hsu, and Jim Glass. Dinosr: Self-distillation and online clustering for self-supervised speech representation learning. *Advances in Neural Information Processing Systems 36*, pp. 58346-58362 (2023)

[43] Khurana, Sameer, Antoine Laurent, and James Glass. Samu-xlsr: Semantically-aligned multimodal utterance-level cross-lingual speech representation. *IEEE Journal of Selected Topics in Signal Processing 16*, no. 6, pp.1493-1504 (2022)

[44] Boito, Marcely Zanon, John Ortega, Hugo Riguidel, Antoine Laurent, Loïc Barrault, Fethi Bougares, Firas Chaabani et al. ON-TRAC consortium systems for the IWSLT 2022 dialect and low-resource speech translation tasks. *arXiv preprint*, arXiv:2205.01987 (2022)

[45] Anastasopoulos, Antonios, Loc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey et al. "Findings of the IWSLT 2022 Evaluation Campaign." *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pp. 98-157. Association for Computational Linguistics, (2022)

[46] Akshaya Vishnu, Kudlu Shanbhogue, Ran Xue, Soumya Saha, Daniel Zhang, and Ashwinkumar Ganesan. Amazon Alexa AI's Low-Resource Speech Translation System for IWSLT2023. *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*, (2023)

[47] Edward Gow-Smith, Alexandre Berard, Marcely Zanon Boito, and Ioan Calapodescu. NAVER LABS Europe's Multilingual Speech Translation Systems for the IWSLT 2023 Low-Resource Track. *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT)*, 2023.

[48] Costa-jussà, Marta R., James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint*, arXiv:2207.04672 (2022)

[49] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, 8:726–742, (2020)