Note: Paper summary is in the end. Please read it before reading the whole paper 🍻🍻

mSLAM: Massively multilingual joint pre-training for speech and text

Ankur Bapna [*1] **Colin Cherry** [*1] **Yu Zhang** [*1] **Ye Jia** [1] **Melvin Johnson** [1] **Yong Cheng** [1] **Simran Khanuja** [1]
Jason Riesa [1] **Alexis Conneau** [1]

Objective

Abstract

We present mSLAM, a multilingual Speech and LAnguage Model that learns cross-lingual cross-modal representations of speech and text by pre-training jointly on large amounts of unlabeled speech and text in multiple languages. mSLAM combines w2v-BERT pre-training on speech with SpanBERT pre-training on character-level text, along with Connectionist Temporal Classification (CTC) losses on paired speech and transcript data, to learn a single model capable of learning from and representing both speech and text signals in a shared representation space. We evaluate mSLAM on several downstream speech understanding tasks and find that joint pre-training with text improves quality on speech translation, speech intent classification and speech language-ID while being competitive on multilingual ASR, when compared against speech-only pre-training. Our speech translation model demonstrates zero-shot text translation without seeing any text translation data, providing evidence for cross-modal alignment of representations. mSLAM also benefits from multi-modal fine-tuning, further improving the quality of speech translation by directly leveraging text translation data during the fine-tuning process. Our empirical analysis highlights several opportunities and challenges arising from large-scale multimodal pre-training, suggesting directions for future research.

1. Introduction

Multilingual pre-trained models have demonstrated large quality gains on a variety of multilingual Natural Language Processing (NLP) tasks (Hu et al., 2020; Ruder et al., 2021). With the emergence of multilingual pre-trained models of speech like XLSR (Conneau et al., 2020; Babu et al., 2021), similar improvements have also been observed on speech understanding tasks. One key advantage of multilingual pre-trained models is the ability to overcome data skew across languages to improve quality on low resource languages (Arivazhagan et al., 2019). By training a shared set of (usually attention-based) parameters on many languages, these models can learn crosslingually aligned representations of text or speech in a shared representation space (Kudugunta et al., 2019; Wu et al., 2019). These shared multilingual representations allow multilingual pre-trained models to use supervised data in one language to benefit lower-resource languages (Conneau & Lample, 2019). An extreme scenario of cross-lingual transfer learning is zero-shot transfer, where using supervised data to fine-tune a pre-trained model on a source language exhibits non-zero performance on a target language; without utilizing any supervision for the target language (Johnson et al., 2017; Conneau et al., 2018).

Advantage of multilingual training

Given the convergence of architectures (Vaswani et al., 2017) and objectives (Devlin et al., 2019; Baevski et al., 2020; Chung et al., 2021) across the speech and text modalities, building a single model that could learn cross-lingual cross-modal representations of speech and text from hundreds of languages is the next natural step. Such a model can enable transfer learning across the two modalities, directly benefiting languages (and domains) with limited amounts of speech or text data. In addition, joint models of speech and text can likely enable end-to-end speech understanding tasks directly from the speech signal, including tasks like speech translation, speaker intent classification and speech language-identification, bypassing errors introduced by an intermediate automatic speech recognition (ASR) system.

Why joint training makes sense?

While there are several potential advantages from multilingual pre-trained models of speech and text, these models suffer from interference and capacity dilution (Bapna et al., 2021). This effect has also been documented in multilingual pre-trained models of text. While lower resource languages benefit from transfer learning, with increasing multilinguality, high resource languages lose quality (Caruana, 1997; Arivazhagan et al., 2019; Conneau et al., 2019). This deterioration is typically addressed by either increasing model capacity (Arivazhagan et al., 2019; Babu et al., 2021) or incorporating approaches from multi-task learning to reduce

- Problem with single models trained on joint data
- What happens when you build a single model with low leve resources?
- Approaches to address the above problems

[*] Equal contribution [1] Google, USA. Correspondence to: Ankur Bapna <ankurbpn@google.com>, Colin Cherry <colincherry@google.com>, Yu Zhang <ngyuzh@google.com>.

interference by leveraging architectural or optimization improvements (Wang et al., 2020; Raffel et al., 2019).

In this work we present mSLAM, a multilingual pre-trained model of speech and text that has been pre-trained with speech from 51 languages and text from 101 languages. mSLAM is a multilingual extension of SLAM (Bapna et al., 2021), with the addition of a Connectionist Temporal Classification (CTC) loss (Graves et al., 2006) on the paired speech-text data, to reduce interference and encourage stronger alignment across the two modalities.

On several downstream speech understanding tasks, including CoVoST-2 21→En speech translation (Wang et al., 2021b), Fleurs speech language identification (Section 4.2) and Minds-14 speech intent classification (Gerz et al., 2021), mSLAM demonstrates significant quality improvements over equivalent models trained only on speech. On multilingual ASR tasks, including MLS-10Hr (Pratap et al., 2020), VoxPopuli (Wang et al., 2021a) and Babel (Gales et al., 2014), mSLAM matches the performance of the speech-only baseline. We also evaluate mSLAM on XNLI (Conneau et al., 2018), to understand its strengths and limitations on text tasks. We find that the addition of the CTC loss significantly improves quality on several speech and text understanding tasks, highlighting the importance of alleviating interference in multi-modal pre-trained models.

CTC has its cons but it still is one of the best objectives out there for these tasks

We also conduct analyses to understand the extent of multimodal representation alignment in mSLAM. When fine-tuned with only speech translation data, mSLAM is capable of zero-shot text translation in several languages, suggesting that the model is capable of learning from data in one modality to improve quality in the other. mSLAM also benefits from multi-modal supervised data. On CoVoST-2, we jointly fine-tune mSLAM on multilingual speech translation and text translation, further improving speech translation quality by 2 BLEU; improving over a significantly larger XLS-R (2B) model (Babu et al., 2021) and establishing a new state of the art on this dataset. Increasing mSLAM model capacity to $2B$ parameters results in further quality improvements on most downstream tasks.

2. Background

Multimodal pre-training: SLAM (Bapna et al., 2021) is a multimodal speech and text pretraining method, which trains a single Conformer (Gulati et al., 2020) with SpanBERT (Joshi et al., 2020) and w2v-BERT (Chung et al., 2021) self-supervised losses that leverage unlabeled monomodal data, as well as a TLM loss (Conneau & Lample, 2019; Zheng et al., 2021) and a speech-text matching loss (Li et al., 2021) that both use supervised speech recognition data. Pre-trained speech representations have been shown to be close to text (Baevski et al., 2021) and SLAM leverages this similarity for cross-modal transfer. Compared to mono-modal pre-trained models, SLAM shows improvements on speech translation, similar performance on speech recognition but degradation on text downstream tasks, exposing a transfer-interference trade-off that has been previously studied in multilingual models (Arivazhagan et al., 2019). Because SLAM focuses on English it is harder to notice cross-modal transfer, as both modalities have a large amount of unlabeled data. In many languages, speech data is scarcer than text, or vice-versa. In this scenario cross-modal transfer is more likely, similar to how high-resource languages transfer to low-resource languages in multilingual pre-training. mSLAM exploits both cross-lingual and cross-modal transfer by simultaneously training on both modalities in a large number of languages.

SLAM pros and cons

Multilingual pre-training: In multilingual understanding literature, models like mBERT (Devlin et al., 2019), XLM-R (Conneau & Lample, 2019) or mT5 (Xue et al., 2021b) have shown the benefit of cross-lingual transfer for improving representations of low-resource languages: on these languages, multilingual models strongly outperform monolingual pre-trained models on public benchmarks (Conneau et al., 2018; 2019; Lewis et al., 2019; Hu et al., 2020; Ruder et al., 2021). Past work has also leveraged parallel data to improve multilingual text representations, e.g. with TLM (Conneau & Lample, 2019), explicit alignment (Hu et al., 2021) or nmT5 (Kale et al., 2021). Similarly, in speech understanding, multilingual pre-trained models (Kawakami et al., 2020; Conneau et al., 2020; Babu et al., 2021) based on self-supervised losses (Oord et al., 2018; Baevski et al., 2020) improve representations of low-resource languages at the cost of reduced performance on high-resource languages. In particular, multilingual pre-trained models like XLS-R expanded the few-shot learning capability of wav2vec 2.0 (Xu et al., 2021) to many other languages, both for speech recognition and speech translation (Wang et al., 2021b). Interestingly, for speech, no lexical overlap across languages is leveraged during training, but multilingual representations still emerge from parameter sharing of the Transformer network (Wu et al., 2019). Leveraging text can potentially create connections between speech representations across languages through shared text anchor embeddings of identical character strings. Past work also leverages supervised ASR data to build multilingual representations of speech (Kannan et al., 2019; Bai et al., 2021), similar to how multilingual machine translation in NLP is used to build multilingual representations (Eriguchi et al., 2018; Siddhant et al., 2020).

3. Pre-training Method

3.1. Architecture and Objectives

Our pre-training approach builds on SLAM (Bapna et al., 2021) and extends it to the massively multilingual setting.

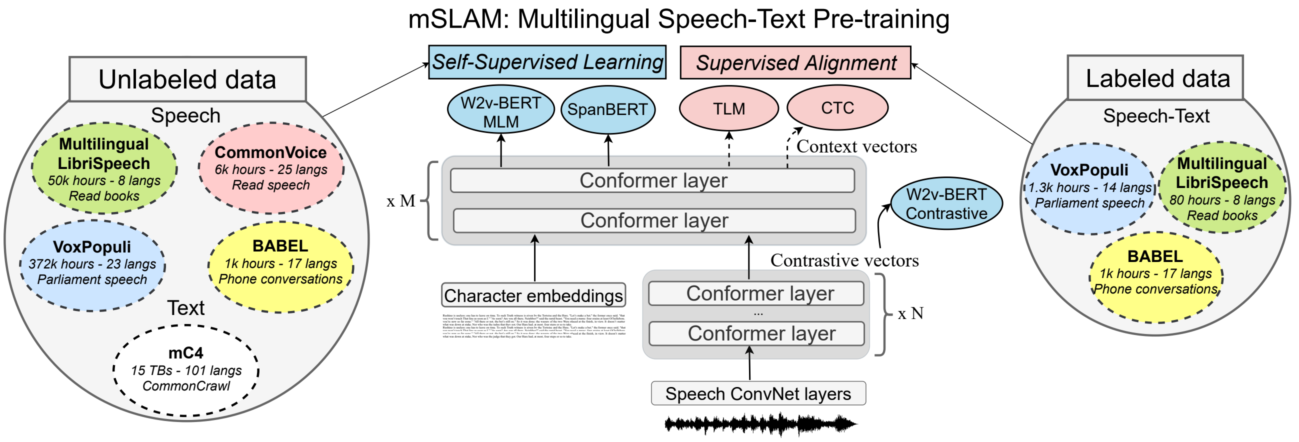

Figure 1: Multilingual Speech-Text Pretraining We pre-train a large multilingual speech-text Conformer on 429K hours of unannotated speech data in 51 languages, 15TBs of unannotated text data in 101 languages, as well as 2.3k hours of speech-text ASR data.

Specifically, we build on SLAM-TLM, that combines together pre-training on speech unlabeled data with w2v-BERT (Chung et al., 2021), text with spanBERT (Joshi et al., 2020) and on paired speech-transcript data with TLM (Conneau & Lample, 2019). We skip the Speech-Text-Matching (STM) task since preliminary experiments didn't reveal any advantages over SLAM-TLM.

How mSLAM is different from SLAM?

mSLAM pre-training differs from SLAM on a couple of points. First, instead of using $32k$ token sentence-piece tokenization (Kudo & Richardson, 2018), we use a character vocabulary, containing 4096 tokens spanning 101 languages. This results in longer sequence lengths, which we cap to 512 characters. We also increase the length of masked spans from 5 to 20 tokens for the spanBERT objective. Second, we apply a CTC loss (Graves et al., 2006; Graves & Jaitly, 2014) on the speech portion of the paired input, using the character-level transcript as the target. This CTC loss is applied in addition to TLM, so the input consists of a concatenated masked speech and masked text sequence, with the CTC loss applied to the speech portion of the output. We share the softmax vocabulary and parameters used for CTC with the softmax used for training the spanBERT objective with text input. We find that this CTC loss ensures stronger alignment between the speech and text representations learnt by the model, as further discussed in Sections 5 and 6.

In our $2B$ model, we increase the model dimension from 1024 to 1408, and the number of conformer layers from 24 to 40. We keep 8 layers in the contrastive block, and 32 in the MLM block. The peak learning rate is reduced from $6.0e-4$ to $3.6e-4$ for better training stability. Other hyper-parameters remain the same as the base $600M$ model. Note that our $2B$ model contains close $1.84B$ parameters.

3.2. Pre-training Data

We use three types of data for pre-training mSLAM; unlabeled speech drawn from multiple public datasets, unlabeled text from mC4 (Xue et al., 2021a) and paired speech and text transcript data from multiple sources.

3.2.1. Unlabeled Speech Data

We use approximately $429k$ hours of unlabeled speech data in 51 languages[1]. Our unlabeled speech data closely follows the pre-training data used for XLS-R (Babu et al., 2021) with one major difference: we do not use VoxLingua. As a consequence our model is pre-trained on speech from 51 languages as compared to 128 for XLS-R, and our pre-training set is smaller by $6.6k$ hours.

We train on $372k$ hours of speech data spanning 23 languages from VoxPopuli (Wang et al., 2021a), read speech data in 25 languages drawn from the v6.1 release of Common Voice (Ardila et al., 2019), $50k$ hours of read books data in eight European languages from Multilingual LibriSpeech (Pratap et al., 2020) and $1k$ hours of telephonic conversation data spanning 17 African and Asian languages from BABEL (Gales et al., 2014).

3.2.2. Unlabeled Text Data

For pre-training with unlabeled text, we use the mC4 dataset (Xue et al., 2021b) spanning 101 languages. We upsample lower resource languages using temperature-based sampling (Arivazhagan et al., 2019), with $T = 3.0$.

3.2.3. Paired Speech-Transcript Data

In addition to training with unlabeled speech and text, we also use approximately $2.4k$ hours of paired speech and

[1] Counting languages with more than 1 hour of speech data.

transcript data spanning 32 languages, for training with the CTC and TLM alignment losses. This data is drawn from the following sources:

VoxPopuli: Approximately $1.3k$ hours of speech and transcript data spanning 14 languages. We exclude languages with less than 1 hour of data following Wang et al. (2021a).

Multilingual LibriSpeech (MLS): We use the 10 hour training splits of the paired data for each of the 8 MLS languages. We exclude any paired data outside the 10 hour splits to align with our downstream evaluations.

Babel: $1k$ hours of speech and transcript data spanning 17 languages from the Babel ASR task.

3.3. Optimization and Hyperparameters

At each training step, we train mSLAM on all three types of data; each batch is composed of 2048 sequences of unlabeled speech, 8192 sequences of text and 256 sequences of paired data. Our speech-only baseline, w2v-bert-51 (0.6B), sees a batch composed of 4096 unlabeled speech sequences at every step. For our best run based on CoVoST dev performance, the speech loss has a coefficient of 1.0, the text loss has a coefficient of 0.3 and the paired CTC loss has a coefficient of 0.03 (to avoid over-fitting to the small paired data). We use the Adam optimizer (Kingma & Ba, 2014) with a Transformer learning rate schedule (Vaswani et al., 2017). We use $40k$ warmup steps, linearly increasing the learning rate to 6×10^{-4}, followed by inverse square root decay. We train all $600M$ models for $1.3m$ steps. The $2B$ model was pre-trained for $350k$ steps.

4. Downstream tasks

4.1. Multilingual Speech Translation

CoVoST 2 Speech Translation: CoVoST 2 (Wang et al., 2021b) is a multilingual speech translation (ST) dataset created by professional translation of the Common Voice speech corpus (Ardila et al., 2020). The audio consists of read speech crowd-sourced through the Mozilla Common Voice project. We evaluate on a multilingual XX-to-English task that covers translation from 21 source languages into English. The training data ranges in size from 264 hours speech for French to about 1 hour speech for Indonesian.

Multi-modal fine-tuning: Apart from fine-tuning with just ST data, we leverage the ability of mSLAM to learn from both speech and text modalities by using text translation data in addition to the CoVoST 2 ST data for multi-modal joint fine-tuning. For each CoVoSt 2 XX-to-English language pair, we use the text translation data from CoVoST 2 combined with all data from either WMT or TED Talks, as available. Specifically, we pair with WMT20 (Barrault et al., 2020) for ja, ta, WMT19 (Barrault et al., 2019) for de, ru, zh, WMT18 (Bojar et al., 2018) for et, tr, WMT17 (Bojar et al., 2017) for lv, WMT15 (Bojar et al., 2015) for fr, WMT13 (Bojar et al., 2013) and TED59 (Qi et al., 2018) for ar, fa, id, it, nl, pt, sl, sv, leaving ca and cy unpaired.

We attach a 6-layer, 512-dimension Transformer decoder to our pre-trained encoders. This decoder has 34M parameters. For ST-only fine-tuning, this model is then fine-tuned on the CoVoST 2 ST dataset. A dropout probability 0.3 is used on the input embedding and all residual connections in the Transformer decoder to mitigate overfitting. For multi-modal fine-tuning, this model is fine-tuned on the CoVoST 2 ST dataset simultaneously with the MT dataset described above. Each training batch contains equal numbers of ST and MT examples, with a higher loss weight, 5.0, on the MT objective. A lower dropout probability 0.1 is used because more training data is available.[2]

4.2. Speech Classification

Fleurs-LangID: Fleurs[3] is a speech extension of the FLORES massively multilingual benchmark for MT (Goyal et al., 2021). Fleurs contains 2009 sentences from the FLORES multi-way parallel evaluation set in 102 languages. We collect read speech corresponding to these sentences, and split these utterances into train-dev-test splits with 1109 for training (around 1.3 hours of data), 400 for dev and 500 for test, per-language. We collected 2.3 utterances per sentence on average. We evaluate our pre-trained models on Speech Language Identification (LangID) on this dataset.

MINDS-14: MINDS-14 (Gerz et al., 2021) is an intent classification task from spoken data. It covers 14 intents extracted from the e-banking domain, with spoken examples in 14 language varieties. We merge monolingual datasets into a single dataset, with a 30-20-50 train-dev-test split.

Fine-tuning setup: When fine-tuning our models on speech classification we train the multi-modal and speech encoders. Speech input is fed into the speech encoder, and the outputs from the multi-modal encoder are max-pooled together before feeding into a softmax classifier. Optionally a projection layer is applied before pooling. We tune hyper-parameters on dev performance; tuning batch sizes over $\{16, 32, 64\}$, learning rates over $\{2e-6, 4e-6, 2e-5, 4e-5\}$ and projection over $\{None, model\_dim\}$. For MINDS we tune number of epochs over $\{100, 300\}$ and for Fleurs over $\{5, 10, 20\}$. We pick the run with the best dev performance and evaluate on the test split. For MINDS-14, we report the macro-averaged accuracy over all 14 languages.

[2] These hyper-parameters were found by optimizing the w2v-BERT speech-only baseline for CoVoST 2 development BLEU.

[3] Dataset to be released with another publication.

4.3. Multilingual Speech Recognition

VoxPopuli: Following Wang et al. (2021a), we evaluate on the 14 languages with more than 1-Hr of data from the VoxPopuli ASR task.

MLS-10Hr: We report results on the 10-Hr training split for the MLS task (Pratap et al., 2020).

Babel: Following Babu et al. (2021), we report results on 5 languages from the Babel-ASR task.

Fine-tuning Setup: We fine-tune our pre-trained encoders with a 2-layer LSTM (Hochreiter & Schmidhuber, 1997) as a conformer-transducer model, following Chung et al. (2021). We use a merged grapheme vocabulary based on the task-specific training set for all ASR fine-tuning experiments. We do not use language-model fusion for any experiments. For VoxPopuli and MLS we report results with multilingual fine-tuning, while we fine-tune separate models per language for Babel. Our finetuning parameters follow (Zhang et al., 2020); for the pre-trained encoder, we use a peak learning rate of $3e-4$ with $5k$ warm-up steps, while for the decoder, a peak learning rate of $1e-3$ and $1.5k$ warm-up steps. All finetuning experiments on ASR use a constant 256 batch size. In practice, these parameters worked well across several tasks and amounts of data.

4.4. Text Classification

XNLI: We also evaluate mSLAM models on the XNLI sentence-pair classification task (Conneau et al., 2018) to understand its strengths and weaknesses on text understanding tasks. We evaluate our models under both the zero-shot and translate-train-all settings (Ruder et al., 2021), and compare performance against mT5 (Xue et al., 2021a).

Fine-tuning setup: We train the multi-modal and text encoders on XNLI. We tune batch sizes over $\{16, 32\}$, learning rates over $\{2e-5, 4e-5\}$, projection over $\{None, model\_dim\}$ and number of epochs over $\{3, 5\}$.

5. Results

5.1. Multilingual Speech Translation

ST fine-tuning: Multilingual speech translation results are shown in Table 1. Removing all text and paired data from pre-training gives us our speech-only pre-training baseline, w2v-bert-51 (0.6B), which is already very competitive with the state-of-the-art, outperforming XLS-R (1B) (Babu et al., 2021), despite having fewer parameters and not using a pre-trained decoder. mSLAM-TLM adds text and a paired TLM objective to pre-training as described by Bapna et al. (2021), and actually leads to an average degradation in ST quality, potentially due to interference between the speech and text modalities alongside the additional pressure of massive multilinguality. Fortunately, mSLAM-CTC's addition of a CTC component to the TLM objective, as described in Section 3, recovers w2v-bert-51 (0.6B) performance; in fact, it improves slightly, mostly on low-resource languages. As we show in Section 6, this CTC component is essential to zero-shot cross-modal behavior.

Table 1: Speech translation - CoVoST 2 X→En summarized results in BLEU. Full per-language results are available in the Appendix Table 9.

X → English	high	mid	low	all
Prior work, mBART decoder init. (Babu et al., 2021)				
XLS-R (0.3B)	30.6	18.9	5.1	13.2
XLS-R (1B)	34.3	25.5	11.7	19.3
XLS-R (2B)	36.1	27.7	15.1	22.1
Our Work: Speech Only				
w2v-bert-51 (0.6B)	35.6	25.3	13.4	20.4
Our Work: Speech + Text				
mSLAM-TLM (0.6B)	34.4	23.4	11.3	18.6
mSLAM-CTC (0.6B)	35.5	25.2	13.7	20.6
mSLAM-CTC (2B)	36.3	27.5	15.6	22.4
Our Work: Speech Only w/ joint fine-tuning				
w2v-bert-51 (0.6B)	36.4	25.9	13.8	21.0
Our Work: Speech + Text w/ joint fine-tuning				
mSLAM-TLM (0.6B)	35.5	25.3	12.3	19.8
mSLAM-CTC (0.6B)	37.6	27.8	15.1	22.4
mSLAM-CTC (2B)	**37.8**	**29.6**	**18.5**	**24.8**

The 0.6B param model is more imp IMO

ST + MT joint fine-tuning: The picture becomes more interesting as we introduce MT (text-to-text) data during fine-tuning in the bottom four lines of Table 1. On top of the speech-only w2v-bert-51 (0.6B), adding MT data produces a modest average improvement of +0.6 BLEU. However, adding MT data to mSLAM-CTC, results in a larger improvement of +1.8 BLEU, suggesting that exposure to text during pre-training makes the encoder more amenable to using text during fine-tuning. This results in a new state-of-the art for the CoVoST 21→En task, surpassing the 4× larger XLS-R (2B) by 0.3 BLEU, enabled by large gains on high-resource languages. Increasing the capacity of mSLAM-CTC to $2B$ parameters further improves performance by 2.4 BLEU.

CTC rocks!

5.2. Speech Classification

Evaluations on the MINDS-14 and Fleurs-LangID tasks are detailed in Table 2. We find that pre-training jointly with text and paired data with a TLM loss, mSLAM-TLM, improves over our speech-only baseline, w2v-bert-51 (0.6B), by 1.3% and 4.6% on MINDS-14 and Fleurs-LangID respectively. The addition of a CTC loss in mSLAM-CTC

Table 2: Speech Classification - MINDS-14 speech intent classification and Fleurs speech language identification accuracy.

Model	**MINDS-14**	**Fleurs-LangID**
Our work: Speech Only		
w2v-bert-51 (0.6B)	82.7	71.4
Our work: Speech + Text		
mSLAM-TLM (0.6B)	84.0	76.0
mSLAM-CTC (0.6B)	**86.9**	73.3
mSLAM-CTC (2B)	86.6	**77.7**

further improves accuracy by 2.9% on MINDS-14. On Fleurs-LangID, mSLAM-CTC is worse than mSLAM-TLM by around 2.7%, still maintaining an accuracy improvement of 1.9% over our speech-only baseline. Increasing mSLAM-CTC capacity to $2B$ parameters results in further 1.7% improvement over our previous best accuracy on Fleurs, while being 0.3% worse than the 600M model on MINDS-14.

5.3. Multilingual Speech Recognition

Table 3: Speech Recognition - Average Word Error Rate (WER) on the VoxPopuli, Babel and MLS-10Hr datasets. Per-language results can be found in Appendix Tables 10, 11 and 12 respectively.

Model	VoxPop	Babel	MLS
Prior work (Babu et al., 2021)			
XLS-R (0.3B)	12.8	32.0	12.8
XLS-R (1B)	10.6	**29.5**	10.9
XLS-R (2B)	-	**29.5**	11.0
Our work: Speech-only			
w2v-bert-51 (0.6B)	9.3	32.8	9.9
Our work: Speech + Text			
mSLAM-TLM (0.6B)	9.4	33.2	10.4
mSLAM-CTC (0.6B)	9.2	32.9	10.1
mSLAM-CTC (2B)	**9.1**	31.3	**9.7**

We present ASR results on VoxPopuli, Babel and MLS-10hrs in Table 3. Our speech-only pre-training baseline, w2v-bert-51 (0.6B) already outperforms XLS-R (Babu et al., 2021) on VoxPopuli and MLS-10hrs as shown in Table 3. Our mSLAM-CTC (0.6B) model slightly outperforms the speech-only baseline on VoxPopuli and slightly lags on MLS-10hrs, but both improve over published results. On Babel, our model is behind XLS-R (1B) (Babu et al., 2021); possibly due to a lack of language model fusion. mSLAM-CTC is very close to w2v-bert-51 (0.6B) and both improve over mSLAM-TLM. In conclusion, mSLAM achieves competitive ASR results without losing speech capacity across a variety of ASR tasks and languages. Increasing mSLAM-CTC capacity to $2B$ parameters results in improvements over both, the 600M model and our speech-only baseline.

5.4. Text Classification

Table 4: Text Classification - XNLI dev accuracy on English, European (bg, de, el, es, fr) and Non-European (ar, hi, ru, sw, th, tr, ur, vi, zh) languages. Full results in Appendix Table 13. Note, for mSLAM models, only 450M and $1.4B$ out of the 600M and $2B$ parameters are fine-tuned for text tasks.

Model	**English**	**Euro**	**Non-Euro**	**Avg**
Prior work: Text Only, Zero-shot (Xue et al., 2021b)				
mT5-Small (0.3B)	79.6	66.6	60.4	63.8
mT5-Base (0.6B)	84.5	77.1	69.5	73.0
Our work: Speech + Text, Zero-shot				
mSLAM-TLM (0.6B)	75.7	57.5	48.6	53.4
mSLAM-CTC (0.6B)	80.4	71.4	49.5	58.9
mSLAM-CTC (2B)	80.1	74.4	59.9	66.1
Prior work: Text Only, Translate-Train-All (Xue et al., 2021b)				
mT5-Small (0.3B)	78.3	73.6	69.2	71.3
mT5-Base (0.6B)	85.9	82.1	77.9	79.8
Our work: Speech + Text, Translate-Train-All				
mSLAM-TLM (0.6B)	74.1	69.3	64.6	66.8
mSLAM-CTC (0.6B)	81.1	76.0	65.5	70.0
mSLAM-CTC (2B)	84.1	80.5	73.7	76.1

Capacity dilution effect!

On XNLI, similar to SLAM results on GLUE, we observe decreases in performance compared to mono-modal models due to capacity dilution (see Table 4). In the translate-train-all setting, our mSLAM-CTC (0.6B) model obtains 70.0% accuracy on average compared to 79.8% for an mT5-Base model (0.6B). However, it performs comparably to the smaller mT5-Small model (0.3B) which gets 71.3%.

The task where mSLAM performs poorly compared to mT5

On zero-shot classification, we observe a bigger drop in performance when using multi-modal pre-training compared to the mT5 models. Zero-shot classification being a test-bed for the sharing of multilingual representations, we attribute this to speech interfering with the sharing of text representations across languages. Looking more closely at per-language results, the performance drops in particular for non-European languages, e.g. Thai and Chinese where the model loses around 20% accuracy. Note that the paired data used during pre-training is predominantly from European languages, and the performance of mSLAM-CTC improves significantly over mSLAM-TLM on this set of languages. We hypothesize that having in-language paired data and alignment losses could be contributing to reduced interference between speech and text for these languages, resulting in more robust representations. This is also supported by the significantly improved performance on non-European languages with the mSLAM-CTC (2B) model, where the increased capacity might alleviate some of the interference. However, there are other confounding factors in our text pre-training approach compared to standard multilingual

Hypothesis for this reduced performance

text pre-training, including the conformer architecture and fully character-level encoder pre-training, which might be contributing to these findings. We leave the study of multilingual representation alignment in joint speech-text models to future work.

6. Analysis

Best question

Do we really need text pre-training or just alignment losses? mSLAM-CTC models add two improvements over the speech-only baseline: (i) TLM and CTC alignment losses over paired data, and (ii) Pre-training with large amounts of web-text. This raises the question whether our improvements are arising from (i), (ii) or a combination of the two. To answer this question we train a mSLAM-CTC model on unlabeled speech and paired speech text data, but no unlabeled text. We evaluate this model on CoVoST ST, MINDS-14 and Fleurs-LangID and present results in Table 5. We find that the performance of the mSLAM-CTC model without text falls somewhere between our speech-only model and mSLAM-CTC on MINDS-14 and Fleurs-LangID, suggesting that the additional text pre-training data is at least partially responsible for the observed improvements on these tasks. On CoVoST-2, mSLAM-CTC without text almost matches the performance of mSLAM-CTC when fine-tuning jointly with text translation data, suggesting that a majority of the improvements in this setting arise from MT data, and the alignment loss might be enough to enable the model to benefit from text supervised data for fine-tuning.

Table 5: Comparing mSLAM-CTCmodels trained with and without unlabeled text on CoVoST-2 ST BLEU (with joint fine-tuning), MINDS-14 accuracy and Fleurs-LangID accuracy.

	CoVoST Avg.	MINDS-14	Fleurs
w2v-bert-51 (0.6B)	21.0	82.7	71.4
mSLAM-CTC (0.6B)	22.4	86.9	73.3
mSLAM-CTC (0.6B) - Text	22.2	85.0	71.9

Are cross-modal representations really aligned?

We have seen benefits from adding text to pre-training alongside a CTC loss. The importance of this CTC loss suggests that some amount of cross-modal representation alignment is necessary to take advantage of speech and text data in the same model, but can we construct an experiment to clearly demonstrate this alignment?

Zero-shot performance is one strong indicator for representation alignment. To that end, we use our joint fine-tuning infrastructure to conduct CoVoST 2, 21→En translation experiments where we fine-tune the mSLAM-CTC model on one modality (speech or text) and evaluate on the CoVoST 2 test set using the the other input modality.

Table 6: Zero-shot Performance - CoVoST 2 translation results with X→Y indicating X as the fine tuning modality and Y as the testing modality: S=Speech, T=Text. CAE is our CTC zero-shot character auto-encoding probe.

	Hours	BLEU ↑			CER ↓
Lang	Paired	S→S	S→T	T→S	S→T CAE
ar	0	13.3	0.0	0.0	82.6
fa	0	6.2	0.0	0.0	80.0
ja	0	1.6	0.0	0.0	100.0
zh	0	8.7	0.0	0.0	100.0
cy	0	6.1	0.1	0.0	24.3
mn	0	0.5	0.1	0.0	78.4
id	0	3.9	5.1	0.0	10.4
lv	0	19.4	8.2	0.0	18.4
et	0	17.2	8.3	0.0	16.5
sv	0	33.1	15.2	0.0	13.9
ca	0	33.4	16.7	0.0	10.0
ru	0	41.7	21.9	0.0	85.9
sl	6	24.9	7.8	0.0	10.6
pt	10	34.2	17.2	0.0	9.0
nl	41	32.6	16.8	0.0	11.3
ta	63	0.3	0.0	0.0	91.2
tr	69	11.7	1.7	0.0	12.6
it	79	35.0	19.7	0.0	11.2
es	140	39.1	21.2	0.0	7.9
fr	179	36.7	20.0	0.0	9.4
de	197	32.7	16.8	0.0	8.3

Cross-modal results, alongside the amount of paired data available during pre-training, are shown in Table 6. For score calibration, the S→S column shows a modality-matched scenario of fine tuning on speech and testing on speech, corresponding to the sixth row of Table 1. First, note that zero-shot cross-modal translation is possible: the S→T column shows that fine-tuning on speech and testing on text results in translation performance above 5 BLEU for 13 of 21 languages. Furthermore, 6 of those 13 languages had no paired data available during pre-training, demonstrating the power of being both multimodal and multilingual. Most surprisingly of all, Russian (ru) has an excellent zero-shot score of 21.9 BLEU, and it not only has no paired data during pre-training, but also no paired data in its Cyrillic script, yet the mSLAM model can translate it into English.

Nice 👌

We tested for the same behavior with w2v-bert-51 (0.6B) and mSLAM-TLM and found no evidence of zero-shot S→T transfer. We also tested the impact of unlabeled text: average zero-shot BLEU is 9.4 for full mSLAM-CTC but only 6.4 without any unlabeled text during pre-training (not shown). The languages without paired data suffer disproportionately: Swedish (sv) drops from 15.2 to 0.7 BLEU and Russian (ru) drops from 21.9 to 4.2 BLEU.

There is still much work left to be done. Russian is somewhat of an outlier in terms of script sensitivity, all other cross-modal success stories are for predominantly European languages in the Latin script. Furthermore, some languages such as Turkish (tr) have paired data available during pre-training, but demonstrate limited zero-shot transfer. Finally,

The zero shot performance is not very well aligned in some cases where you would expect it to be better

Table 7: CTC Probing Examples - CoVoST CTC Probe with zero-shot text input to visualize zero-shot text encodings. Gold is the desired output as well as the text input. Romanization is provided by the GOST 7.79 System B standard for Cyrillic transliteration.

fr	Gold	`Certains départements sont mieux équipés que d'autres.`
	S→T (CAE)	`certains départements sont mieux équipés que d'autres`
ru	Gold	И нам следует руководствоваться им.ль.
	Romanized Gold	`I nam sleduet rukovodstvovat'sya im.`
	S→T (CAE)	`  nam sleduet rucovodstvowats    im.`

note that this zero-shot transfer does not work in the other direction: the T→S column clearly shows that a system fine-tuned only on text cannot translate speech.

Table 8: Zero-shot text translation examples. Drawn from the CoVoST 2 FrEn test set, decoded by mSLAM-CTC (0.6B).

Source	Il réalise aussi quelques courts-métrages.
Gold	He also makes short films.
S→T	He also writes a few short films either short films either short films.
Source	Il a réalisé deux courts-métrages.
Gold	He produced two short films.
S→T	He created two short short films.

Examining zero-shot text translation outputs. While the system's cross-modal capabilities are surprising, there is still a substantial drop in BLEU for zero-shot translation of text: compare the S→S column to the S→T column in Table 6. This reduced performance often manifests as repeated or empty outputs: see Table 8 for contrastive zero-shot text translation examples. This is reminiscent of oscillatory hallucinations caused by unexpected inputs (Lee et al., 2018).

Visualizing cross-modal alignment with a CTC probe.

The setup

To visualize the information available when text is input to a model fine-tuned only for speech, we create a CTC probe for mSLAM encodings. Freezing the mSLAM encoder after pre-training, we tune only the softmax parameters of a CTC decoder using a 21-language ASR objective on the CoVoST 2 data: speech is input, and the gold character-level transcription is the output. We can then decode the CoVoST 2 test set using either speech or text inputs. If speech is input, the ASR task matches the fine-tuning objective. If text is input, this represents a zero-shot character-level auto-encoding (CAE) task. We measure the success of both tasks using character-error-rate (CER).

Per-language results of this CTC probe are also shown in Table 6 as S→T CAE. First, it is notable that even with a frozen encoder and a far less powerful decoder, we still see zero-shot transfer from speech to text inputs. In fact, with the exception of Turkish (tr) and Russian (ru), zero-shot CAE performance with less than 20 CER is predictive of zero-shot translation performance greater than 5 BLEU. Russian again is an interesting case, with terrible zero-shot CAE performance, but excellent zero-shot MT performance. However, the real value of such a probe is the ability to inspect the outputs. Table 7 shows randomly selected examples for both, a typical success (French, fr) and one of our more mysterious languages (Russian, ru). For French, most of the content is retained with text input, though some capitalization and punctuation is lost. Interestingly, for Russian, its Cyrillic text input results in a partial transliteration into the Latin script. This suggests that Russian is mapped into the same encoding space as Latin script languages during pre-training, which helps explain its strong cross-modal and cross-lingual transfer behavior.

7. Conclusion

We introduced mSLAM, a multilingual pretrained model capable of representing speech and text in a shared representation space. mSLAM is trained on unlabeled speech in 51 languages with a w2v-BERT objective and character-level text in 101 languages with a SpanBERT objective. In addition to unlabeled data, we train mSLAM on small amounts of paired speech-transcript data with a novel TLM+CTC objective to encourage representation sharing across the two modalities. Downstream evaluations on CoVoST 2 Speech Translation, Speech intent classification and Speech LangID demonstrate that mSLAM improves over equivalent speech-only baselines on speech understanding tasks, while maintaining similar quality on ASR. In addition to fine-tuning with labeled speech data, mSLAM can also leverage text supervision to improve the quality of end-to-end speech tasks, as demonstrated by our experiments on CoVoST 2 Speech Translation, establishing a new state of the art on this dataset. Increasing the capacity of mSLAM to $2B$ parameters further improves quality on Speech Translation, Speech Language Identification and multilingual ASR.

On XNLI sentence-pair classification, we observe cross-lingual zero-shot performance equivalent to text-only models half the size of mSLAM on (European) languages with relatively large amounts of paired data, but severe quality degradation on languages with scarce parallel data. We notice that this degradation can be addressed to some extent by increasing the capacity of the joint speech-text model.

When fine-tuned on speech translation only, mSLAM is capable of cross-modal zero-shot text translation, demonstrating strong evidence for representation alignment. Probing a frozen mSLAM encoder with a CTC head fine-tuned for ASR demonstrates high quality on text reconstruction, providing additional supporting evidence in favour of aligned speech-text representations.

The use of paired data and alignment losses results in quantitative improvements on several speech understanding tasks and reduced degradation on text understanding tasks, highlighting the need for mitigating interference in multilingual multi-modal pre-training. We hope that this work catalyzes further research towards improving and understanding universal, multi-modal pre-trained models.

References

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. Common Voice: A massively-multilingual speech corpus. In *Proceedings of Language Resources and Evaluation Conference*, 2020.

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G., Cherry, C., et al. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*, 2019.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., et al. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*, 2021.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*, 2020.

Baevski, A., Hsu, W.-N., Conneau, A., and Auli, M. Unsupervised speech recognition. *arXiv preprint arXiv:2105.11084*, 2021.

Bai, J., Li, B., Zhang, Y., Bapna, A., Siddhartha, N., Sim, K. C., and Sainath, T. N. Joint unsupervised and supervised training for multilingual asr. *arXiv preprint arXiv:2111.08137*, 2021.

Bapna, A., Chung, Y.-a., Wu, N., Gulati, A., Jia, Y., Clark, J. H., Johnson, M., Riesa, J., Conneau, A., and Zhang, Y. Slam: A unified encoder for speech and language modeling via speech-text joint pre-training. *arXiv preprint arXiv:2110.10329*, 2021.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 1–61, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5301. URL https://aclanthology.org/W19-5301.

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., and Zampieri, M. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pp. 1–55, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.1.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pp. 1–44, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-2201.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 1–46, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3001. URL https://aclanthology.org/W15-3001.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pp. 169–214, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4717. URL https://aclanthology.org/W17-4717.

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Koehn, P., and Monz, C. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Transla-*

tion: Shared Task Papers, pp. 272–303, Belgium, Brussels, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6401. URL https://aclanthology.org/W18-6401.

Caruana, R. Multitask learning. *Machine Learning*, 28 (1):41–75, Jul 1997. ISSN 1573-0565. doi: 10.1023/A:1007379606734. URL https://doi.org/10.1023/A:1007379606734.

Chung, Y.-A., Zhang, Y., Han, W., Chiu, C.-C., Qin, J., Pang, R., and Wu, Y. w2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *ASRU*, 2021.

Conneau, A. and Lample, G. Cross-lingual language model pretraining. In *NeurIPS*, 2019.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S. R., Schwenk, H., and Stoyanov, V. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. In *ACL*, 2019.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

Eriguchi, A., Johnson, M., Firat, O., Kazawa, H., and Macherey, W. Zero-shot cross-lingual classification using multilingual neural machine translation, 2018.

Gales, M. J. F., Knill, K., Ragni, A., and Rath, S. P. Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. In *SLTU*, 2014.

Gerz, D., Su, P.-H., Kusztos, R., Mondal, A., Lis, M., Singhal, E., Mrkšić, N., Wen, T.-H., and Vulić, I. Multilingual and cross-lingual intent detection from spoken data. *arXiv preprint arXiv:2104.08524*, 2021.

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzman, F., and Fan, A. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *arXiv preprint arXiv:2106.03193*, 2021.

Graves, A. and Jaitly, N. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pp. 1764–1772. PMLR, 2014.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376, 2006.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech*, 2020.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pp. 4411–4421. PMLR, 2020.

Hu, J., Johnson, M., Firat, O., Siddhant, A., and Neubig, G. Explicit alignment objectives for multilingual bidirectional encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3633–3643, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.284. URL https://aclanthology.org/2021.naacl-main.284.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

Kale, M., Siddhant, A., Constant, N., Johnson, M., Al-Rfou, R., and Xue, L. nmt5 – is parallel data still relevant for pre-training massively multilingual language models?, 2021.

Kannan, A., Datta, A., Sainath, T. N., Weinstein, E., Ramabhadran, B., Wu, Y., Bapna, A., Chen, Z., and Lee, S. Large-scale multilingual speech recognition with a streaming end-to-end model. *arXiv preprint arXiv:1909.05330*, 2019.

Kawakami, K., Wang, L., Dyer, C., Blunsom, P., and Oord, A. v. d. Learning robust and multilingual speech representations. *arXiv preprint arXiv:2001.11128*, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kudo, T. and Richardson, J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012.

Kudugunta, S. R., Bapna, A., Caswell, I., Arivazhagan, N., and Firat, O. Investigating multilingual nmt representations at scale. *arXiv preprint arXiv:1909.02197*, 2019.

Lee, K., Firat, O., Agarwal, A., Fannjiang, C., and Sussillo, D. Hallucinations in neural machine translation. In *NeurIPS 2018 Interpretability and Robustness for Audio, Speech and Language Workshop*, 2018.

Lewis, P., Oğuz, B., Rinott, R., Riedel, S., and Schwenk, H. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*, 2019.

Li, J., Selvaraju, R. R., Gotmare, A. D., Joty, S., Xiong, C., and Hoi, S. Align before fuse: Vision and language representation learning with momentum distillation, 2021.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., and Collobert, R. Mls: A large-scale multilingual dataset for speech research. *arXiv preprint arXiv:2012.03411*, 2020.

Qi, Y., Sachan, D., Felix, M., Padmanabhan, S., and Neubig, G. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 529–535, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2084. URL https://aclanthology.org/N18-2084.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Ruder, S., Constant, N., Botha, J., Siddhant, A., Firat, O., Fu, J., Liu, P., Hu, J., Neubig, G., and Johnson, M. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. *arXiv preprint arXiv:2104.07412*, 2021.

Siddhant, A., Johnson, M., Tsai, H., Ari, N., Riesa, J., Bapna, A., Firat, O., and Raman, K. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8854–8861, 2020.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *NIPS*, 2017.

Wang, C., Rivière, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., and Dupoux, E. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*, 2021a.

Wang, C., Wu, A., and Pino, J. CoVoST 2 and massively multilingual speech-to-text translation. In *interspeech*, 2021b.

Wang, Z., Tsvetkov, Y., Firat, O., and Cao, Y. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. *arXiv preprint arXiv:2010.05874*, 2020.

Wu, S., Conneau, A., Li, H., Zettlemoyer, L., and Stoyanov, V. Emerging cross-lingual structure in pretrained language models. *arXiv preprint arXiv:1911.01464*, 2019.

Xu, Q., Baevski, A., Likhomanenko, T., Tomasello, P., Conneau, A., Collobert, R., Synnaeve, G., and Auli, M. Self-training and pre-training are complementary for speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3030–3034. IEEE, 2021.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. mt5: A massively multilingual pre-trained text-to-text transformer. In *NAACL*, 2021a.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, Online, June 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

Zhang, Y., Qin, J., Park, D. S., Han, W., Chiu, C.-C., Pang, R., Le, Q. V., and Wu, Y. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504*, 2020.

Zheng, R., Chen, J., Ma, M., and Huang, L. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation, 2021.

Table 9: Speech translation - CoVoST 2 X→En full results in BLEU.

	High-resource				Mid-resource					Low-resource		
X → English	fr	de	es	ca	fa	it	ru	pt	zh	tr	ar	et
Train Hours	264h	184h	113h	136h	49h	44h	18h	10h	10h	4h	2h	3h
Prior work, mBART Decoder init. (Babu et al., 2021)												
XLS-R (0.3B)	32.9	26.7	34.1	28.7	5.9	29.0	26.4	28.3	4.9	4.6	3.0	3.5
XLS-R (1B)	36.2	31.2	37.9	31.9	9.6	33.1	37.0	39.3	8.7	12.8	12.2	8.3
XLS-R (2B)	37.6	33.6	39.2	33.8	12.9	34.9	39.5	41.8	9.4	16.7	17.1	11.1
Our Work: Speech Only												
w2v-bert-51 (0.6B)	36.9	33.1	38.9	33.5	5.8	34.9	41.8	36.1	8.0	8.8	13.7	17.4
Our Work: Speech + Text												
mSLAM-TLM (0.6B)	35.7	31.6	37.8	32.4	5.5	33.6	39.9	29.4	8.7	8.2	9.0	14.9
mSLAM-CTC (0.6B)	36.7	32.7	39.1	33.4	6.2	35.0	41.7	34.2	8.7	11.7	13.3	17.2
mSLAM-CTC (2B)	37.6	33.8	39.5	34.4	8.8	36.1	43.6	42.0	7.1	19.7	15.8	18.6
Our Work: Speech Only w/ joint fine-tuning												
w2v-bert-51 (0.6B)	37.5	34.1	39.6	34.2	6.1	35.7	44.1	34.7	9.0	12.7	15.5	19.1
Our Work: Speech + Text w/ joint fine-tuning												
mSLAM-TLM (0.6B)	36.8	32.8	38.8	33.6	9.7	34.6	41.2	32.1	8.8	12.2	12.6	16.6
mSLAM-CTC (0.6B)	38.6	36.1	40.6	35.2	7.2	37.0	47.5	36.4	10.8	15.6	14.2	20.3
mSLAM-CTC (2B)	39.0	35.9	41.0	35.4	9.7	37.3	48.4	42.8	10.0	24.2	19.3	22.6

	Low-resource									Average			
X → English	mn	nl	sv	lv	sl	ta	ja	id	cy	high	mid	low	all
Train Hours	3h	7h	2h	2h	2h	2h	2h	2h	2h				
Prior work (Babu et al., 2021)													
XLS-R (0.3B)	0.4	22.0	10.3	6.0	6.6	0.2	0.6	1.4	2.5	30.6	18.9	5.1	13.2
XLS-R (1B)	0.8	28.2	24.7	16.0	16.7	0.3	1.9	10.3	8.6	34.3	25.5	11.7	19.3
XLS-R (2B)	1.6	31.7	29.6	19.5	19.6	0.5	3.5	16.5	14.0	36.1	27.7	15.1	22.1
Our Work: Speech Only													
w2v-bert-51 (0.6B)	0.3	33.8	33.9	16.0	25.5	0.3	0.9	3.5	6.2	35.6	25.3	13.4	20.4
Our Work: Speech + Text													
mSLAM-TLM (0.6B)	0.5	31.7	29.5	14.0	17.4	0.3	1.7	3.8	5.1	34.4	23.4	11.3	18.6
mSLAM-CTC (0.6B)	0.5	32.5	32.1	18.6	25.0	0.3	1.7	3.7	6.8	35.5	25.2	13.7	20.6
mSLAM-CTC (2B)	0.3	34.4	35.5	22.8	29.2	0.3	1.7	4.7	4.4	36.3	27.5	15.6	22.4
Our Work: Speech Only w/ joint fine-tuning													
w2v-bert-51 (0.6B)	0.7	34.6	31.6	13.8	23.9	0.2	1.3	4.5	7.3	36.4	25.9	13.8	21.0
Our Work: Speech + Text w/ joint fine-tuning													
mSLAM-TLM (0.6B)	0.3	33.2	26.3	15.2	19.8	0.5	1.3	3.7	5.6	35.5	25.3	12.3	19.8
mSLAM-CTC (0.6B)	0.9	36.3	31.7	19.8	25.6	0.5	2.4	6.1	7.7	37.6	27.8	15.1	22.4
mSLAM-CTC (2B)	0.8	37.6	38.5	26.8	32.3	0.6	3.3	8.8	6.7	37.8	29.6	18.5	24.8

Table 10: Speech recognition - VoxPopuli ASR results in terms of WER.

	en	de	it	fr	es	pl	ro	hu	nl	cs	sl	fi	hr	sk	Avg
Labeled data	543h	282h	91h	211h	166h	111h	89h	63h	53h	62h	10h	27h	43h	35h	
Prior work (Babu et al., 2021)															
XLS-R (0.3B)	10.2	13.0	19.2	12.6	9.8	9.6	7.9	11.6	14.8	10.5	24.5	14.2	12.3	8.9	12.8
XLS-R (1B)	8.8	11.5	15.1	10.8	8.2	7.7	7.3	9.6	12.5	8.7	19.5	11.3	10.0	7.1	10.6
Our work: Speech-only															
w2v-bert-51 (0.6B)	7.2	9.0	15.8	9.2	8.6	6.5	7.6	8.4	10.5	7.0	15.8	9.3	9.1	6.0	9.3
Our work: Speech + Text															
mSLAM-TLM (0.6B)	7.3	8.9	15.6	9.3	8.7	6.5	8.5	8.4	10.5	7.1	15.8	9.0	10.0	6.2	9.4
mSLAM-CTC (0.6B)	7.1	8.9	15.6	9.3	8.6	6.5	8.5	8.1	10.3	7.0	14.2	9.2	9.1	5.9	9.2
mSLAM-CTC (2B)	7.0	8.7	15.4	9.4	8.4	6.4	7.8	8.4	10.5	6.8	15.1	8.7	9.1	6.0	**9.1**

Table 11: Speech recognition - BABEL ASR baselines in five languages, reporting WER.

Model	as	tl	sw	lo	ka	Avg
Number of pretraining hours	55h	76h	30h	59h	46h	-
Number of fine-tuning hours	55h	76h	30h	59h	46h	-
Prior work (with LM) (Babu et al., 2021)						
XLS-R (0.3B)	42.9	33.2	24.3	31.7	28.0	32.0
XLS-R (1B)	40.4	30.6	21.2	30.1	25.1	29.5
XLS-R (2B)	**39.0**	**29.3**	**21.0**	**29.7**	**24.3**	**28.7**
Our work: Speech-only, no LM						
w2v-bert-51 (0.6B)	42.8	32.9	26.7	30.6	31.1	32.8
Our work: Speech + Text, no LM						
mSLAM-TLM (0.6B)	43.0	32.7	27.6	30.9	31.8	33.2
mSLAM-CTC (0.6B)	42.7	32.6	27.1	30.7	31.4	32.9
mSLAM-CTC (2B)	41.1	31.1	25.1	29.9	29.1	31.2

Table 12: Speech recognition - Multilingual LibriSpeech (MLS) ASR baselines in 8 languages, reporting WER.

Model	en	de	nl	fr	es	it	pt	pl	Avg
Number of training hours	10	10	10	10	10	10	10	10	-
Prior work (monolingual fine-tuning) (Babu et al., 2021)									
XLS-R(0.3B)	15.9	9.0	13.5	12.4	8.1	13.1	17.0	13.9	12.8
XLS-R(1B)	12.9	7.4	**11.6**	10.2	7.1	12.0	15.8	10.5	10.9
XLS-R(2B)	14.0	7.6	**11.8**	10.0	6.9	12.1	15.6	9.8	11.0
Our work: Speech Only (multilingual fine-tuning)									
w2v-bert-51 (0.6B)	12.7	7.0	12.6	8.9	5.9	10.3	14.6	**6.9**	9.9
Our work: Speech + Text (multilingual fine-tuning)									
mSLAM-TLM (0.6B)	13.9	7.2	13.0	9.9	5.8	10.7	**14.2**	8.4	10.4
mSLAM-CTC (0.6B)	13.3	7.0	12.5	9.7	**5.5**	10.5	**14.1**	8.5	10.1
mSLAM-CTC (2B)	**11.9**	**6.6**	12.4	**8.5**	5.8	**9.8**	15.2	7.7	**9.7**

Table 13: Text Classification - XNLI dev accuracy for all 15 languages. For mSLAM models, only $450M$ and $1.4B$ out of the $600M$ and $2B$ parameters are fine-tuned for XNLI.

Model	**en**	**ar**	**bg**	**de**	**el**	**es**	**fr**	**hi**	**ru**	**sw**	**th**	**tr**	**ur**	**vi**	**zh**	**Avg**
Prior work: Text Only, Zero-shot (Xue et al., 2021b)																
mT5-Small (0.3B)	79.6	62.2	67.8	64.8	65.8	68.4	66.2	59.0	65.3	55.4	63.2	58.9	54.5	61.8	63.4	63.8
mT5-Base (0.6B)	84.5	71.2	76.9	75.6	76.3	79.0	77.7	66.9	74.9	63.6	70.0	69.2	64.8	72.0	72.5	73.0
Our work: Speech + Text, Zero-shot																
mSLAM-TLM (0.6B)	75.7	47.3	56.7	55.1	52.2	60.9	62.8	48.6	58.5	46.0	46.9	51.3	47.2	50.7	41.0	53.4
mSLAM-CTC (0.6B)	80.4	46.5	69.8	72.1	67.5	74.7	72.9	42.0	68.7	45.5	42.9	48.7	44.2	63.3	43.3	58.9
mSLAM-CTC (2B)	80.1	61.1	73.3	74.7	72.7	76.0	75.3	59.4	70.9	52.2	56.8	63.9	59.0	65.9	50.1	66.1
Prior work: Text Only, Translate-Train-All (Xue et al., 2021b)																
mT5-Small (0.3B)	78.3	68.8	73.5	73.2	73.4	74.4	73.5	67.4	71.1	67.2	71.1	69.9	63.6	70.5	72.9	71.3
mT5-Base (0.6B)	85.9	78.8	82.2	81.6	81.4	83.0	82.1	77.0	81.1	74.8	78.6	78.4	73.3	78.9	80.2	79.8
Our work: Speech + Text, Translate-Train-All																
mSLAM-TLM (0.6B)	74.3	64.2	68.7	69.5	69.2	70.2	71.4	64.5	65.4	63.4	65.6	65.9	62.4	67.3	64.4	67.1
mSLAM-CTC (0.6B)	81.1	63.5	76.7	76.0	73.1	77.8	76.4	63.6	73.1	64.1	64.9	66.8	60.5	68.4	64.5	70.0
mSLAM-CTC (2B)	84.1	80.2	80.1	78.7	82.9	80.5	74.4	72.1	76.8	71.7	73.8	76.2	69.8	75.9	72.8	76.1

Summary

Introduction

- Pre-training has been a huge success in almost every field. Multi-lingual pre-trained models have demonstrated large quality gains on a variety of multilingual NLP tasks. Similar improvements have been observed on speech tasks as well.
- One key advantage of multilingual pre-trained models is the ability to overcome data skew across languages to improve quality on low resource languages. These attention based models can learn cross-lingually aligned representations of text or speech in a shared representation space with a shared set of parameters.
- Given that this cross-lingual concept works, the next natural step is to build a model that could learn cross-lingual cross-modal representations of speech and text from different languages. One of the benefits of this model is that it can enable transfer learning across two modalities, extremely beneficial in scenarios where we have limited amount of text and speech data
- These kind of models have one major problem though, they suffer from inference and capacity dilution, a case that was observed in SLAM
- This paper extends SLAM by adding CTC loss on the paired speech-text data to address the above problem. The authors name this method as mSLAM that has been pre-trained with speech from 51 languages and text from 101 languages.

Background

- Multimodal pre-training
    - SLAM is a multimodal speech and text prertraining method
    - Single Conformer model that combines
        - SpanBert and w2v-BERT self-supervised losses -> leverage unlabeled monomodal data
        - TLM loss and speech-text matching loss - > use supervised speech recognition data
    - It has already been shown earlier that pre-trained speech representations are very similar to that of text. SLAM tries to leverage this similarity for cross-modal transfer
    - SLAM shows improvements on speech translation, and speech recognition but perfrmance degradation is observed on text based downstream tasks, a classic tansfer-interference tradeoff
    - SLAM focuse only on Enlgish, hence cross-modal transfer is harder to notice in that case. mSLAM on the other hand exploits both cross-modal, and cross-lingual transfer by simultaneously training on both modalities in large number of languages
- Multi-lingual training
    - Multi-lingual models strongly outperform monolingual pretrained models
    - They have also shown to improve speech understanding task as well
    - The parameter sharing in Transformers lead to multilingual representations even when there is no lexical overlap across languages used in training
    - Leveraging text can potentially cre- ate connections between speech representations across lan- guages through shared text anchor embeddings of identical character strings

Pre-training Method

- Architecture and Objectives
  - Extends on SLAM-TLM to build mSLAM
    - Pretraining on speech unlabeled data with w2v-BERT, text with spanBERT, and paired speech-transcript data with TLM
    - Skips speech-text-matching task as it didn't provide any added advantage
  - mSLAM is different from SLAM in the following ways:
    - Character vocabulary containing 4096 tokens spanning 101 languages instead of 32k token sentence-piece tokenization, resulting in longer sequence lengths
    - Length of masked spans is increased from 5 to 20 tokens fro the spanBERT objective
    - CTC is applied on the speech portion paired input using character level transcripts as the target
    - CTC is alos applied to TLM. The input consists of a concatenated masked speech and masked text sequence with the CTC loss applied to the speech portion of the output
    - The softmax vocabulary and the parameters used for the speech part is shared with softmax used for training spanBERT objective with text as the input
    - This sharing of softmax vocab and parameters in the above step along with CTC loss ensures that there is a stronger alignment between speech and the text representation learnt by the model

- Datasets used in Pre-training
  - Unlabeled Speech data
    - 429K hours of unlabeled speech data in 51 languages
    - 372k hours of speech data spanning 23 languages from VoxPopuli
    - Read speech data in 25 languages drawn from the v6.1 release of Common Voice
    - 50k hours of read books data in eight European languages from Multilingual Libri Speech
    - 1k hours of telephonic conversation data spanning 17 African and Asian languages
  - Unlabeled Text data
    - mC4 dataset spanning 101 languages.
    - Lower resource languages are upsampled by using temperature based sampling and setting T=3.0
  - Paired Speech-transcript data
    - 2.4k hours of paired speech and transcript data spanning 32 languages
    - VoxPopuli: Approximately 1.3k hours of speech and tran- script data spanning 14 languages.
    - Multilingual LibriSpeech (MLS): 10 hour training splits of the paired data for each of the 8 MLS languages
    - Babel: 1k hours of speech and transcript data spanning 17 languages from the Babel ASR task

- Downstream Tasks
  - Multilingual Speech translation
  - Speech classification
  - Multilingual Speech recognition
  - Text Classification