MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

ZeroST: Zero-Shot Speech Translation

Khurana, Sameer; Hori, Chiori; Laurent, Antoine; Wichern, Gordon; Le Roux, Jonathan

TR2024-122 September 04, 2024

Abstract

Our work introduces the Zero-Shot Speech Translation (Ze- roST) framework, leveraging the synergistic potential of pretrained multilingual speech and text foundation models. Inspired by recent advances in multimodal foundation models, ZeroST utilizes a Query Transformer (Q-Former) to seamlessly connect a speech foundation model, such as Whisper or Massively Multilingual Speech (MMS), with a text translation model like No-Language-Left-Behind (NLLB). Our proposed learning framework enables the model to perform the speech- to-text translation in a zero-shot manner, bypassing the need for explicit supervision from expensive-to-collect speech-text translation pairs during training. Our extensive experiments, notably on the Europarl-ST benchmark, demonstrate that ZeroST achieves results comparable to those of a strong cascaded translation system and significantly outperforms baseline models. This promising approach paves the way for future research in zero-shot speech translation.

Interspeech 2024

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

ZeroST: Zero-Shot Speech Translation

Sameer Khurana[1], *Chiori Hori*[1], *Antoine Laurent*[2], *Gordon Wichern*[1], *Jonathan Le Roux*[1]

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
[2]LIUM, Le Mans Université, France

{khurana,chori,wichern,leroux}@merl.com

Abstract

Our work introduces the Zero-Shot Speech Translation (ZeroST) framework, leveraging the synergistic potential of pre trained multilingual speech and text foundation models. Inspired by recent advances in multimodal foundation models, ZeroST utilizes a Query Transformer (Q-Former) to seamlessly connect a speech foundation model, such as Whisper or Massively Multilingual Speech (MMS), with a text translation model like No-Language-Left-Behind (NLLB). Our proposed learning framework enables the model to perform the speech-to-text translation in a zero-shot manner, bypassing the need for explicit supervision from expensive-to-collect speech-text translation pairs during training. Our extensive experiments, notably on the Europarl-ST benchmark, demonstrate that ZeroST achieves results comparable to those of a strong cascaded translation system and significantly outperforms baseline models. This promising approach paves the way for future research in zero-shot speech translation.

Index Terms: Zero-Shot, Automatic Speech Translation, Multilingual, Query Transformer, Foundation Models

1. Introduction

In recent years, tremendous advancements have been made in multilingual speech foundation models, with such recent works as Whisper [1] and Massively Multilingual Speech (MMS) [2], as well as multilingual text foundation models, such as GPT [3, 4, 5], T5 [6, 7], and No-Language-Left-Behind (NLLB) [8]. These foundation models form the backbone of many modern audio- and text-based natural language processing systems [9, 10, 11]. As the performance of modality-specific perceptual foundation models reaches new heights, a logical next research step is to connect unimodal foundation models to create multimodal foundation models capable of performing multimodal tasks, such as image captioning [12], speech-to-text translation [13, 14], multimodal retrieval [15, 16, 17], and others. One recent research effort in this direction is bootstrapping language-image pre-training with frozen image encoders and large language models (BLIP-2) [12]. BLIP-2 bridges the representation gap between a pre-trained vision foundation model and a pre-trained large language model (LLM). The key idea in BLIP-2 is to translate visual representations outputted by a pre-trained image encoder into “text-like” representations that can be ingested and processed by an LLM. To that end, BLIP-2 uses a Query Transformer (Q-Former), initially proposed in [18], as the bridge between the image encoder and the LLM, giving rise to a multimodal LLM.

Our Work, inspired by BLIP-2, connects a pre-trained multilingual speech foundation model, such as MMS [2] or Whisper [1], with the pre-trained multilingual text translation model NLLB [8] to perform multilingual speech-to-text translation (ST). Figure 1 shows a high-level illustration of our proposed framework. Like BLIP-2, we use a Q-Former to bridge the representation gap between the speech and text foundation models. We refer to our framework as the Zero-Shot Speech Translation (ZeroST) model since the Q-Former bridge is trained using speech-text pairs in the same language, i.e., no translation pairs are provided during the training of our framework. Nonetheless, the model can generate text translations for a speech waveform during inference in different languages. We show this on the Europarl-ST benchmark (Table 1). Our model performs comparably to a powerful cascaded translation system and significantly outperforms the baselines set for this work while never being exposed to speech-text translation examples during training.

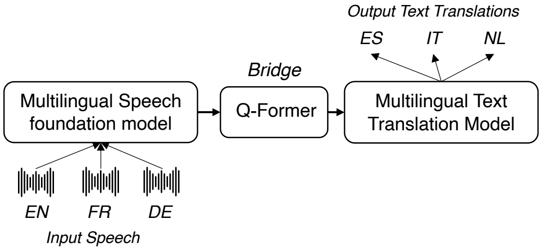
Output Text Translations
ES
IT
NL
Bridge
Multilingual Speech foundation model
Q-Former
Multilingual Text Translation Model
EN
FR
DE
Input Speech

Figure 1: *A high-level illustration of our proposed ZeroST model. A Query Transformer bridges the representation gap between a pre-trained speech foundation model and a pre-trained text translation model to translate speech into text.*

2. Proposed ZeroST Framework

2.1. Model Overview

Our proposed ZeroST model consists of three main modules: speech foundation model, Q-Former, and text translation model.

Speech Foundation Model: We use either the pre-trained Whisper-large-v3 [1] or pre-trained MMS [2] as our speech encoder. Whisper is an encoder-decoder model, of which we only use the encoder. Whisper is trained using 1M hours of weakly-labeled speech downloaded from YouTube and 5M hours of pseudo-labeled YouTube recordings. The pseudo-labels are obtained via Whisper-large-v2, a previous version of Whisper-large-v3. The model is trained using supervised learning to maximize the conditional probability $p(y|x)$, where x is a speech waveform and y is its text transcript or translation. It supports 96 spoken languages. The Whisper transformer encoder has 32 layers, an embedding dimension of 1280, and a capacity of 630M parameters. The MMS transformer encoder

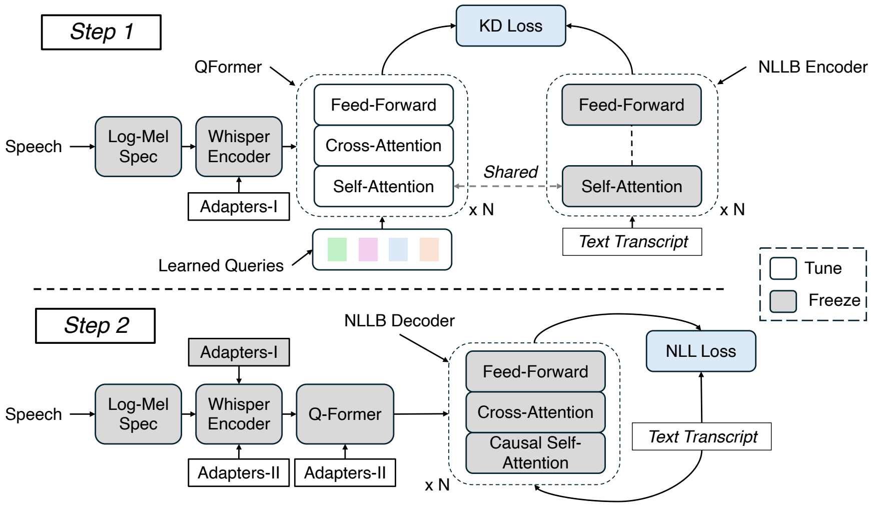

Figure 2: *An illustration of our proposed ZeroST learning process.*

used in this work has 24 layers, an embedding dimension of 1024, and a capacity of 300M parameters. It is trained via contrastive self-supervised learning (SSL) of the form proposed in wav2vec-2.0 [19] on unlabeled speech data collected from 1K+ languages from several public corpora. We use these two encoders to show that our ZeroST framework is robust to the choice of the speech encoder.

When using MMS, the input to the model is a zero-mean unit-variance speech waveform, while it is a 128-dimensional log-mel spectrogram when using the Whisper encoder. MMS can take variable-length speech waveforms sampled at 16 kHz as input. We restrict the maximum size to 30 s due to memory constraints. Whisper takes in padded or trimmed 30 s speech utterances as input, also sampled at 16 kHz. The pre-trained checkpoints are available via HuggingFace[1] [20] and fairseq[2] [21]. A linear projection layer transforms the output of the speech encoder before inputting to the Q-Former to match its embedding dimension.

Text Translation Model: We use NLLB [8] as the text translation model. NLLB is a transformer encoder-decoder trained on 200 x 200 text-to-text translation tasks. NLLB ranges in size from 1.3B to 54.6B parameters. We use the 1.3B parameter model available via HuggingFace[3]. The encoder and decoder have 24 layers. The decoder has slightly more parameters due to the cross-attention module. The model's embedding size is 1024 and its vocabulary has 256k BPE tokens.

Q-Former: Q-Former has the same architecture as the NLLB decoder. Unlike the NLLB decoder, the self-attention module in Q-Former is bi-directional. Q-Former's self-attention is initialized using the NLLB encoder's self-attention. Q-Former is conditioned on the output of the speech encoder via its encoder-decoder cross-attention module. The input to the Q-Former is a set of learnable embeddings referred to as queries, whose number is a hyper-parameter. We found 256 to be the optimal number of queries for our work (Table 2). Since Q-Former aims to output a representation close to the NLLB encoder's output and its decoder's input, it is natural to parameterize Q-Former with the same architecture as the NLLB encoder. This removes one source of discrepancy between the two. Next, we detail how the model's parameters are tuned.

2.2. Learning Process

We propose a two-step learning process (cf. Fig. 2 for illustration). The first step involves using the NLLB encoder as the teacher training the student Q-Former. The goal is to reduce the representation gap between the Q-Former output and the NLLB text encoder output and, consequently, the input of the NLLB text decoder. The second step involves using the NLLB decoder as the teacher training the student Q-Former. This step removes the remaining representation gap between the Q-Former's output and the decoder's input. We refer to the first step as *knowledge distillation (KD)* and the second as *negative-log-likelihood (NLL) training*. Both steps use multilingual transcribed speech data for training. We detail the training data and the two steps below.

Training Data: We collect the multilingual transcribed speech corpora from CommonVoice-v16.1 (CoVo) [22], VoxPopuli (VP) [23], and Multilingual Speech (MLS) [24] datasets. From CoVo, we collect transcribed speech in 96 languages that intersect with the languages supported by Whisper[4]. VP has data in 16 languages: English (en), German (de), French (fr), Spanish (es), Polish (pl), Italian (it), Romanian (ro), Hungarian (hu), Czech (cs), Dutch (nl), Finnish (fi), Croatian (hr), Slovak (sk), Slovene (sl), Estonian (et), and Lithuanian (lt). MLS has data in eight languages: en, es, it, pl, Portuguese (pt), nl, de, fr. The data distribution across languages is highly imbalanced. We follow [15] (Eq. 3) to up/down-sample the data per-language. The total multilingual transcribed speech data used for training is about 12k hours. We exclude the 44k hours of transcribed English data from the MLS corpus to avoid overfitting the model to the English language.

Knowledge Distillation: This step trains the Q-Former on the task of speech-to-text retrieval. Given a tuple (x, y), where x is a speech waveform and y its corresponding transcript, the combination of the speech encoder and Q-Former transforms $x \in \mathbb{R}^S$ into a set of embeddings $Q \in \mathbb{R}^{q \times d}$, where q is the number of queries. At the same time, the NLLB text encoder

[1] https://huggingface.co/openai/whisper-large-v3

[2] https://github.com/facebookresearch/fairseq/tree/main/examples/mms

[3] https://huggingface.co/facebook/nllb-200-1.3B

[4] https://github.com/openai/whisper/blob/main/whisper/tokenizer.py#L10

transforms the corresponding transcript y into a set of embeddings $T \in \mathbb{R}^{m \times d}$, where m is the number of tokens in the transcript y. Note that the number of queries q is fixed, while the number m of tokens is variable. Using Q and T, we compute two KD losses: fine-grained and global. The fine-grained loss is computed as follows:

$$Q \leftarrow \texttt{L2Norm}(\texttt{Proj}(Q)), \;\; T \leftarrow \texttt{L2Norm}(\texttt{Proj(T)}), \quad (1)$$

$$\mathcal{L}_{\text{Fine}} = \sum_i \left(1 - \texttt{max}_{j \in |T|}\, Q[i] \cdot T[j]\right), \quad (2)$$

$$= \underbrace{|Q|}_{q} - \underbrace{\sum_i \texttt{max}_{j \in |T|} Q[i] \cdot T[j]}_{\text{score}}, \quad (3)$$

where $Q[i] \in \mathbb{R}^d$ is the i^{th} query embedding, $T[j]$ is the j^{th} token embedding, `Proj` transforms the embeddings in Q and T via a linear projection followed by `Tanh` non-linearity, and `L2Norm` normalizes the input embeddings by their L_2 norm. Therefore, the dot product $Q[i] \cdot T[j]$ gives the cosine similarity between $Q[i]$ and $T[j]$. The fine-grained KD loss where each query embedding is compared to each token embedding, and the max-association (or the precision score) which is optimized over are inspired by Colbert's [25] text query-document retrieval model (See Fig. 3 in [25] for a pictorial illustration of the fine-grained loss).

To compute the global loss, we first average the embeddings in sets Q and T to get single embeddings $q \in \mathbb{R}^d$ and $t \in \mathbb{R}^d$. The global KD loss is then computed as follows:

$$q \leftarrow \texttt{L2Norm}(\texttt{Proj}(q)), \;\; t \leftarrow \texttt{L2Norm}(\texttt{Proj}(t)), \quad (4)$$

$$\mathcal{L}_{\text{Global}} = 1 - q \cdot t, \quad (5)$$

where $\mathcal{L}_{\text{Global}}$ is the cosine distance between q and t. The `L2Norm` and `Proj` layers perform the same operations as in the fine-grained KD loss. The final KD loss is computed as: $\mathcal{L}_{\text{KD}} = \beta * (\mathcal{L}_{\text{Global}} + \mathcal{L}_{\text{Fine}})$, where β is a scaling factor. Since the computation of the global and fine-grained losses involve cosine similarities, the magnitude of the KD loss can be quite small, which could lead to small magnitude gradient updates, leading to inefficient training. We upscale the loss by a factor of $\beta > 1$. The scaling factor is a hyper-parameter. An appropriate value can be found early in training by monitoring the training dynamics (loss scale, loss value, gradient norm). We found $\beta = 10$ to work well for us, leading to a stable training process.

During training, all the parameters of the Q-Former are tuned. In contrast, the parameters of the text encoder are frozen, which allows for an efficient training process since the text embeddings can be extracted offline, obviating the need to load the large NLLB text encoder during training. We freeze the speech encoder's pre-trained parameters, insert adapter layers, and fine-tune the adapter layer parameters during training. We use adapter layers of the form proposed in [26]. Two adapters are inserted in each speech encoder's layer, one after self-attention and the other after the feed-forward module.

Negative Log-Likelihood Training: This step trains the Q-Former from the previous step to generate text transcription y corresponding to a speech waveform x by conditioning the NLLB decoder on the output embeddings Q of the Q-Former. The parameters of the Q-Former are tuned to optimize the negative log-likelihood given below:

$$\mathcal{L}_{\text{NLL}} = -\sum_{n=1}^{m} \log p(y_n | y_{1:n-1}, Q) \quad (6)$$

where m is the number of tokens in the transcript y and Q the set of query embeddings. The NLLB transformer decoder estimates the conditional probability of each token y_n conditioned on the previous tokens in the sequence and the query embedding set Q. During training, the previous tokens $y_{1:n-1}$ are the ground-truth tokens (teacher-forcing), while during inference, the model is conditioned on the tokens it generates. During NLL training, the NLLB text decoder remains frozen. For the Q-Former and speech encoder, we insert adapters and only fine-tune those parameters. For the speech encoder, new adapters are inserted in sequence with the adapters used in the KD step.

3. Experiments

Evaluation Protocol: We use the Europarl-ST [27] benchmark for evaluating ZeroST. The set of spoken languages in Europarl-ST is $X = \{\text{en, fr, de, it, es, pt, pl, ro, nl}\}$. Each spoken language $L \in X$ is paired with its text translations in the other languages X_{-L}. Thus, the total translation tasks are 72. We report an evaluation score for all $L \in X$. The evaluation score is computed for a language L by averaging the BLEU-4 scores for the eight translation tasks $L \rightarrow L', L' \in X_{-L}$. Table 1 presents the main results of our work. To understand the results, we first provide details about the terminologies used in Table 1.

Training Details: Most of the results in Table 1 are obtained with the ZeroST framework trained using VoxPopuli (VP) multilingual transcribed speech corpora. We combine VP, MLS, and CoVo corpora (detailed in Section 2.2), referred to as Big, to get our best ZeroST results. All the models are trained on 8 A100 GPUs for 100k iterations except those that use Big data for training, which are trained on 64 GPUs for 400k iterations. The batch size is approximately 2.6 hours of transcribed speech, or 2.5 minutes per GPU. We use the Adam optimizer with a learning rate of 1e-4. Following [19], we use a three-phase learning rate scheduler with the setting [0.1, 0.4, 0.5], i.e., the learning rate is warmed up to 1e-4 during the first 10% of the training iterations, remains constant for the next 40%, and decays for the remaining 50%. Both the KD and NLL learning steps share the same optimization settings.

Q-Formers: We compare different Q-Former architectures with varying degrees of complexity: 1) **Q-Simple**: a Q-Former with no transformer layers. The queries are directly applied to the output of the speech encoder as follows: $Q = \texttt{softmax}(WK^T)V$, where $W \in \mathbb{R}^{256 \times d}$, $K \in \mathbb{R}^{n \times d}$, and $V = K$. W are the learnable queries, K is the output of the pre-trained speech encoder, and $Q \in \mathbb{R}^{256 \times d}$ is the output of the Q-Former. 2) **Q-Lite**: a bi-directional transformer decoder with four layers, an embedding size of 768, four attention heads in each layer, and a feed-forward layer dimension of 3072. 3) **Q-NLLB**: the same architecture as the NLLB decoder but with bi-directional self-attention. The self-attention module of Q-NLLB is initialized with the self-attention of the NLLB encoder.

Baselines and Toplines: The light-gray rows at the top of Table 1 present the baseline results, which use either NLL or KD loss for training. Our work uses the 2-step learning process described in Section 2.2. The dark-gray rows at the bottom present the toplines. The Whisper-NLLB-cascade model transcribes speech waveforms using Whisper-large-V3, which the NLLB model translates to text in the target language. The NLLB-topline uses the ground-truth text transcript for the speech utterances in the Europarl-ST benchmark and translates them into the desired target language using the NLLB model. The best our ZeroST framework can do is match this perfor-

Table 1: *Translation results (BLEU-4) on Europarl-ST. Results for each language are averaged over eight translation tasks corresponding to the other eight languages as translation targets. The light-gray rows are the baselines, and dark-gray rows the toplines.*

Sys. #	Data	Encoder	Q-Former	Loss	en	fr	de	it	es	pt	pl	ro	nl	Avg.
1	VP	MMS	Q-Simple	NLL	16.9	14.5	8.7	13.9	14.5	9.0	9.5	13.3	10.3	12.3
2	VP	MMS	Q-Simple	KD	2.3	1.8	2.1	3.5	1.9	1.4	0.9	2.2	1.6	1.9
3	VP	MMS	Q-Simple	KD→NLL	21.7	18.9	14.4	17.3	18.1	9.9	16.3	18.3	14.9	16.6
4	VP	MMS	Q-Lite	KD→NLL	23.6	21.0	15.3	19.2	20.2	9.8	17.1	20.2	17.0	18.2
5	VP	MMS	Q-NLLB	KD→NLL	25.1	23.2	17.9	21.5	21.3	10.8	19.4	23.1	19.5	20.2
6	Big	MMS	Q-NLLB	KD→NLL	30.9	23.8	18.4	21.2	21.9	22.8	20.6	21.1	19.3	22.2
7	Big	Whisper	Q-NLLB	KD→NLL	31.5	24.2	17.9	20.4	22.6	23.2	22.9	23.8	19.2	22.8
8	Whisper-NLLB-cascade				27.9	22.1	17.7	19.9	20.8	21.3	21.7	24.9	18.9	21.7
9	NLLB-Topline				33.4	24.9	19.3	23.4	23.7	23.6	24.2	26.2	20.5	24.3

mance. We draw the following insights from the results presented in Table 1.

1) **Is NLL or KD training alone sufficient for ZeroST?** According to Table 1, neither is. Keeping the training data, speech encoder, and Q-Former architecture fixed, we observe that the proposed 2-step learning process (row 3) outperforms the NLL-only baseline (row 1) by 4.3 BLEU points and KD-only baseline (row 2) by 14.7 BLEU points, implying that our 2-step sequential learning process is crucial for ZeroST.

2) **Comparing different Q-Formers:** We observe in Table 1 that Q-NLLB (row 5) outperforms Q-Simple (row 3) by 3.6 BLEU points and Q-Lite (row 4) by 2 BLEU points. Nonetheless, it is remarkable that Q-Simple, being such low complexity, achieves quite decent performance. Q-NLLB enjoys improved translation performance across the board for all source languages.

3) **Impact of training data size:** Using Big data improves over VP training data by a couple of BLEU points, as seen in the Avg. results of row 6 vs. row 5 in Table 1. The average performance on the eight translation tasks pt→$X_{-\text{pt}}$ improves dramatically with the change in training data. This is due to VP training data not having transcribed speech for pt language. This exposes a limitation of our ZeroST framework: it cannot generalize to unseen source spoken languages.

4) **Comparing speech encoders:** We observe that using the Whisper speech encoder instead of MMS leads to slightly better overall translation performance (22.8 vs. 22.2 Avg. BLEU-4). This is expected given Whisper's larger size.

5) **Comparison with toplines:** Our final results (row 7) are better than the Whisper-NLLB-cascade model and a couple of points worse than the NLLB-topline. Our model performs better than cascade on all translation tasks except for ro→$X_{-\text{ro}}$ translation tasks where the results are worse (23.8 vs. 24.9). Our model uses in-domain (w.r.t. Europar-ST benchmark) VoxPopuli data for training, which could explain away this result. Future work should test the out-of-domain transfer capabilities of the proposed ZeroST model.

Generating text translations in target languages unseen during training: In this experiment, we train the ZeroST model on transcribed speech in en, fr, de, it, and es languages in the VoxPopuli corpora. During inference, we use ZeroST model to perform the following 15 translation tasks: X ={en, fr, de, it, es}→Y ={pl, ro, nl}. Note that the speech and text in languages in X are seen during training, while languages in Y are unseen during ZeroST model training. We are translating speech in X to text in Y. In this scenario, the ZeroST model achieves an average BLEU-4 score of 13.5 on the 15 translation tasks. This is comparable with the 14.1 average BLEU-4 achieved by the ZeroST model trained on transcribed speech in all the languages in the VoxPopuli transcribed speech corpora, including languages in Y. We hypothesize that our model can generate text translations in unseen target languages due to the complete freezing of the NLLB text decoder during the second step (NLL training) of our proposed ZeroST learning process.

Table 2: *Impact of the number of queries on ZeroST performance. We report average BLEU-4 on Europarl-ST.*

#Queries	16	64	128	256	512
BLEU-4	8.9	11.1	14.2	16.6	16.9

Impact of Varying The Number of Queries: Table 2 ablates over the number of queries fed to the Q-Former. We use the same setting as row 3 in Table 1. We observe diminishing returns as we increase the number of queries.

4. Conclusions

This work presents a promising approach for zero-shot speech-to-text translation. Inspired by recent works on combining unimodal foundation models for multimodal tasks such as BLIP-2 [12], we propose the ZeroST model, which connects Whisper, a pre-trained multilingual speech foundation model, with No-Language-Left-Behind (NLLB), a transformer-based multilingual text-to-text translation model. We bridge the gap between the above-mentioned foundation models using a Query Transformer and train it using a proposed two-step learning process that uses NLLB as the teacher. Zero-shot translation results on Europarl-ST verify our claim that zero-shot multilingual speech-to-text translation is possible using only multilingual transcribed speech data. We achieve better results than a strong cascade and are comparable to the top line. Future work could explore the following issues. Although a preliminary analysis of our model's capability to translate speech into text in languages not seen during the model training showed encouraging results, a more thorough treatment is required to draw a definite conclusion. Another issue is that our model could only translate speech included in the multilingual transcribed speech corpora used to train it. Currently, this is a limitation of our model.

5. Acknowledgements

This work was partially performed using HPC resources from GENCI–IDRIS (Grant AD011012527) and has received funding from the DGA/AID RAPID COMMUTE project.

6. References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023.

[2] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni *et al.*, "Scaling speech technology to 1,000+ languages," *arXiv preprint arXiv:2305.13516*, 2023.

[3] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding with unsupervised learning," OpenAI, Tech. Rep., 2018.

[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI, Tech. Rep., 2019.

[5] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt *et al.*, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *JMLR*, vol. 21, no. 140, pp. 1–67, 2020.

[7] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," *arXiv preprint arXiv:2010.11934*, 2020.

[8] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam *et al.*, "No Language Left Behind: Scaling human-centered machine translation," *arXiv preprint arXiv:2207.04672*, 2022.

[9] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi *et al.*, "SUPERB: Speech processing Universal PERformance benchmark," *arXiv:2105.01051*, 2021.

[10] H.-S. Tsai, H.-J. Chang, W.-C. Huang, Z. Huang, K. Lakhotia, S.-w. Yang, S. Dong, A. Liu *et al.*, "SUPERB-SG: Enhanced Speech processing Universal PERformance benchmark for semantic and generative capabilities," in *Proc. ACL*, May 2022.

[11] J. Shi, D. Berrebbi, W. Chen, H.-L. Chung, E.-P. Hu, W. P. Huang, X. Chang, S.-W. Li *et al.*, "ML-SUPERB: MultiLingual Speech Universal PERformance benchmark," *arXiv preprint arXiv:2305.10615*, 2023.

[12] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. ICML*, 2023.

[13] X. Li, C. Wang, Y. Tang, C. Tran, Y. Tang, J. Pino, A. Baevski, A. Conneau *et al.*, "Multilingual speech translation with efficient finetuning of pretrained models," *arXiv preprint arXiv:2010.12829*, 2020.

[14] S. Khurana, N. Dawalatabad, A. Laurent, L. Vicente, P. Gimeno, V. Mingote, and J. Glass, "Cross-lingual transfer learning for low-resource speech translation," in *Proc. SASB*, 2024.

[15] S. Khurana, A. Laurent, and J. Glass, "SAMU-XLSR: Semantically-aligned multimodal utterance-level cross-lingual speech representation," *IEEE J. Sel. Top. Signal Process.*, 2022.

[16] P.-A. Duquenne, H. Schwenk, and B. Sagot, "Sentence-level multimodal and language-agnostic representations," *arXiv preprint arXiv:2308.11466*, 2023.

[17] H.-F. Wang, Y.-J. Shih, H.-J. Chang, L. Berry, P. Peng, H.-y. Lee, H.-M. Wang, and D. Harwath, "SpeechCLIP+: Self-supervised multi-task representation learning for speech via CLIP and speech-image data," *arXiv preprint arXiv:2402.06959*, 2024.

[18] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. ECCV*, 2020.

[19] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[20] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[21] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," 2019.

[22] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders *et al.*, "Common Voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2020.

[23] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino *et al.*, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proc. ACL*, Aug. 2021.

[24] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," in *Interspeech 2020*, Oct. 2020.

[25] O. Khattab and M. Zaharia, "Colbert: Efficient and effective passage search via contextualized late interaction over bert," in *Proc. ACM SIGIR*, 2020.

[26] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in *Proc. ICML*, 2019.

[27] J. Iranzo-Sánchez, J. A. Silvestre-Cerdà, J. Jorge, N. Roselló, A. Giménez, A. Sanchis, J. Civera, and A. Juan, "Europarl-ST: A multilingual corpus for speech translation of parliamentary debates," *arXiv preprint arXiv:1911.03167*, 2019.