INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# Strategies for improving low resource speech to text translation relying on pre-trained ASR models

*Santosh Kesiraju*[1], *Marek Sarvaš*[1], *Tomáš Pavliček*[2], *Cécile Macaire*[3], *Alejandro Ciuba*[4]

[1]Speech@FIT, Brno University of Technology, Czechia. [2]Phonexia, Czechia.
[3]Univ. Grenoble Alpes, France. [4]University of Pittsburgh, USA.

kesiraju@fit.vutbr.cz, xsarva00@stud.fit.vutbr.cz, tomas.pavlicek@phonexia.com,
cecile.macaire@univ-grenoble-alpes.fr, alejandrociuba@pitt.edu

## Abstract

This paper presents techniques and findings for improving the performance of low-resource speech to text translation (ST). We conducted experiments on both simulated and real-low resource setups, on language pairs English - Portuguese, and Tamasheq - French respectively. Using the encoder-decoder framework for ST, our results show that a multilingual automatic speech recognition system acts as a good initialization under low-resource scenarios. Furthermore, using the CTC as an additional objective for translation during training and decoding helps to reorder the internal representations and improves the final translation. Through our experiments, we try to identify various factors (initializations, objectives, and hyper-parameters) that contribute the most for improvements in low-resource setups. With only 300 hours of pre-training data, our model achieved 7.3 BLEU score on Tamasheq - French data, outperforming prior published works from IWSLT 2022 by 1.6 points.

**Index Terms**: speech translation, low-resource, multilingual, speech recognition

## 1. Introduction

Speech translation (ST) systems consume speech (features) from source language as input and generate text in the target language. A cascaded approach to this task involves passing speech through an automatic speech recognition (ASR) system that generates (decodes) $n$-best discrete text-hypotheses in source language, which are then passed on to a text-based machine translation (MT) system to generate the text in target language (Fig. 1a). Here, the errors from the ASR outputs are *likely* to be propagated to the MT system. End-to-end approaches aim to overcome such errors by establishing a continuous (differentiable) path from input source speech to target translations (Fig. 1b) [1, 2]. End-to-end approaches based on encoder-decoder architectures also make use of source transcriptions to provide additional supervision (Fig. 1c) [3]. There were also attempts to train a direct speech translation system without relying on source text, however such approaches were studied only on high-resource scenarios (Fig. 1d) [4]. For high resource scenarios, the ASR on source language can be trained on huge amounts of available transcribed data, and the MT can be also trained on massive parallel data. Such trained modules can be used as initializations in any of the above frameworks.

However, such a luxury is not available in low-resource scenarios, where neither source speech transcriptions, nor source to target parallel text data are available. Moreover, the amount of speech translation training data can also be very limited (e.g. $<$ 20 hours), which is also the scenario for most of the experiments and analyses in this paper. Automatic translation of

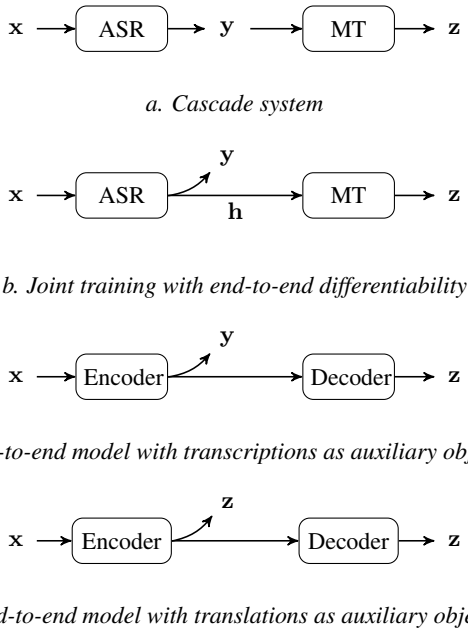


Figure 1: *Cascaded and end-to-end frameworks for speech translation.* $\mathbf{x}$ *is the input speech (features),* $\mathbf{y}$ *is the corresponding text transcriptions, and* $\mathbf{z}$ *is the target text translations.* $\mathbf{h}$ *is the hidden representation from ASR that establishes the continuous path between ASR and MT models. The ASR, MT, encoder and decoder modules can be initialized from various kinds of pre-trained models.*

speech from a low-resource to high-resource language has applications in topic detection [5, 6]. In such low-resource scenarios, one can rely on transfer learning, where the ST model or parts of it are either initialized from a *target-language* ASR or MT or a speech representation model based on self-supervised learning (SSL). More specifically, in an encoder-decoder framework for speech translation, the speech encoder can be initialized from a pre-trained ASR [7] or SSL [8], whereas the decoder can be initialized from a pre-trained ASR [9] or MT [10] model. The model can then be fine-tuned using the target speech translation data. Depending on the choice of initializations, the encoder and decoder can either be aligned or misaligned, i.e., the contextual representations from encoder live in a subspace different than that of the representations in the decoder. Moreover, the vocabulary of an ASR and MT system can differ, which also contributes to the misalignment. Table 1 summarizes the various initialization options and the consequent alignments. The benefit of initializations from large pre-trained models is diminished when the fine-tuning data is very low, which can be at-


10.21437/Interspeech.2023-2506

Table 1: *Initialization options for encoder-decoder based speech translations systems.*

| Encoder init. | Decoder init. | Aligned? |
|---|---|---|
| Encoder from ASR | Decoder from ASR | Yes |
| Encoder from ASR | Decoder from MT | No |
| Encoder from SSL | Decoder from MT | No |
| Random | Random | No |

tributed to the misaligned representations during initialization.

Such a problem of misaligned initialization doesn't arise when both the encoder and decoder are initialized from a pre-trained ASR. However, the ASR models assume monotonic alignment between the input speech and target text, which is not true in the case of speech translation. Here the challenge is to learn the re-ordering with limited amount of ST training data. While there are numerous approaches and analysis on high-resource speech translation [11, 12, 13, 14], there is scope for studying these techniques in low-resource scenarios.

### 1.1. Related works

Prior works [7, 9] have shown that a speech translation system initialized from a monolingual ASR built on target language could benefit in low-resource speech translation. The authors concluded that pre-training on any language could still yield a benefit, however the use of pre-trained multilingual ASR is not fully explored in their work.

The Connectionist-temporal classification (CTC) [15] was originally proposed for ASR. The CTC model built on RNN encoder assume a monotonic alignment between the input speech (features) to the target tokens, which is not suitable for speech translation. Chuang et al. [11] have shown that transformers trained with CTC objective for speech translation can learn to reorder. This has motivated other works exploring direct speech translation with CTC as an auxiliary objective only during training [4]. More recently, Yan et al. [16, 14] have seen the benefits of joint training and decoding for speech translation. However, their models and experiments were mostly focused to mid-to-high resource language where source transcriptions are also available.

In the recent findings from IWSLT 2022 low-resource track for Tamasheq → French speech translation task, the majority of the techniques involving large multilingual SSL models (XLS-R) and pre-trained MT models (mBART) have shown very poor results [17, 18]. This motivated us to revisit the strategies for training low-resource speech translation.

### 1.2. Contributions of the paper

- Study of pre-trained multilingual ASR as initialization for low-resource speech translation with joint training and decoding with CTC objective in low-resource setups.
- Extensive analysis on the effect of various initialization, auxiliary objectives, hyperparameters and amounts of fine-tuning data, identifies the most important factors that contribute most to the improvements.
- On low-resource Tamasheq → French task, our ST model initialized from a pre-trained multilingual ASR with only 300 hours training data achieved 7.3 BLEU score, which is +1.6 points higher than the best published result from IWSLT'22.

Table 2: *Statistics of speech translation data.*

| Direction | Speech translation data: hours (utterances) Training | | Dev. | | Test | |
|---|---|---|---|---|---|---|
| taq→ fr | 13.8 | (4444) | 1.9 | (581) | 2.0 | (804) |
| en→ pt | 292.5 | (184.3k) | 3.2 | (2022) | 3.7 | (2305) |
| Low-resource simulation splits | | | | | | |
| en→ pt | 50.0 | (31.5k) | 3.2 | (2022) | 3.7 | (2305) |
| en→ pt | 16.4 | (10.5k) | 3.2 | (2022) | 3.7 | (2305) |

## 2. Methodology

This section formally introduces the necessary terminology and describes the methods we followed to train ASR and ST systems. The ASR is trained on several examples of paired speech and text $(\mathbf{x}^s, \mathbf{y}^s)$ from one or more (*seen*) languages $s \in \mathcal{S}$. The speech translation systems are trained on pairs $(\mathbf{x}^u, \mathbf{z}^s)$, where the input speech $\mathbf{x}^u$ is from an *unseen* language $u \notin \mathcal{S}$, and the target translation text $\mathbf{z}^s$ is from *seen* languages $s \in \mathcal{S}$.

### 2.1. Training ASR

A transformer [19] based encoder-decoder architecture with additional CTC layer is used to train the ASR models. For multilingual ASR, we keep a separate vocabulary for each language, which results in a language-specific CTC layer at the output of the encoder, and language-specific input (embedding) and output layers in the decoder. Such an architecture allows us to decode tokens only in the desired target language. The models are trained with joint CTC and attention objective function [20]

$$\mathcal{L}_{\text{asr}}(\mathbf{x}^s, \mathbf{y}^s) = \lambda\, \mathcal{L}_{\text{ctc}}(\mathbf{x}^s, \mathbf{y}^s) + (1-\lambda)\, \mathcal{L}_{\text{att}}(\mathbf{x}^s, \mathbf{y}^s). \quad (1)$$

### 2.2. Training ST

The ST models are also based on transformer encoder-decoder architecture and are identical to the ASR models, which allows us to initialize ST models with any pre-trained ASR. More specifically, we are given speech $\mathbf{x}^u$ from a previously unseen language $u \notin \mathcal{S}$, and its translation $\mathbf{z}^s$ from a language that was already seen, $s \in \mathcal{S}$. The ST model is also trained with joint objective function

$$\mathcal{L}_{\text{st}}(\mathbf{x}^u, \mathbf{z}^s) = \alpha\, \mathcal{L}_{\text{ctc}}(\mathbf{x}^u, \mathbf{z}^s) + (1-\alpha)\, \mathcal{L}_{\text{att}}(\mathbf{x}^u, \mathbf{z}^s). \quad (2)$$

### 2.3. Decoding

A beam search based joint decoding [20] that relies on the weighted average of log-likelihoods from both the CTC and transformer decoder modules is used, that produces the most likely hypotheses according to

$$\hat{\mathbf{z}} = \arg\max_{\mathbf{z}}\ \beta\, \log p_{\text{ctc}}(\mathbf{z} \mid \mathbf{x}) + (1-\beta)\, \log p_{\text{att}}(\mathbf{z} \mid \mathbf{x}). \quad (3)$$

## 3. Experimental setup

The ST experiments were conducted on two datasets: (i) Tamasheq (taq) → French (fr) from IWSLT'22 evaluation campaign [17, 18], and (ii) English (en) → Portuguese (pt) from HOW2 dataset [21]. The latter dataset is mainly used for simulating low-resource setups with various amounts of fine-tuning data. Moreover, it also allows is to compare the performance against a typical end-to-end system exploiting source transcripts

Table 3: *Statistics of data for training ASR models.*

| Languages | ASR data: hours Training | Dev. | Test |
|---|---|---|---|
| fr | {50. . . 764} | 25.5 | 26.1 |
| pt | 50 | 10.3 | 11.1 |
| de, es, fr, it, pl, pt (6L) | 300 | 124.1 | {26.1, 11.1} |

Table 4: *Performance of various ASR systems in terms of word (WER) and character error rates (CER).*

| ASR | Training data (in hrs.) | Dev WER | Dev CER | Test WER | Test CER |
|---|---|---|---|---|---|
| | | French (fr) | | | |
| Mono (fr) | 50 | 39.1 | 21.2 | 42.7 | 23.9 |
| | 100 | 30.3 | 15.4 | 33.9 | 17.9 |
| | 200 | 23.8 | 11.8 | 27.4 | 14.1 |
| | 300 | 21.5 | 10.6 | 24.7 | 12.6 |
| | 764 | 16.8 | 8.1 | 19.8 | 9.9 |
| Multi (6L) | 300 | 33.0 | 16.8 | 36.4 | 19.2 |
| | | Portuguese (pt) | | | |
| Mono (pt) | 50 | 27.0 | 11.2 | 29.6 | 12.6 |
| Multi (6L) | 300 | 23.3 | 9.1 | 24.7 | 9.8 |

and source-target parallel data. Table 2 presents the ST data statistics, where the bottom half indicates the low-resource simulation splits derived from HOW2 dataset. To train multilingual ASR models, we picked a subset of 6 languages (6L) from Mozilla Common Voice v8.0, including French and Portuguese. We sampled 50 hours of transcribed data for each language, which resulted in 300 hours of training data. For monolingual ASR training, we considered the same 50hr for Portuguese. In case of French, we trained several monolingual ASR systems on various amounts of data: {50, 100, 200, 300, 764} hours. The Table 3 presents the statistics of data used for ASR training.

### 3.1. Model configuration and hyper-parameters

The input to the model is 80-dimensional filter-bank features appended with 3-dimensional pitch features extracted from the speech signal for every 25ms, with a frame shift of 10ms. The NN model is based on standard transformer encoder-decoder architecture starting with Conv2d layer with 256 output channels, kernel size $(3, 3)$, stride 2. This is followed by 12 transformer layers in the encoder and 6 in the decoder, with $d_{\text{model}} = 256$, $d_{\text{ff}} = 2048$, heads $= 4$, $\text{dropout}_{\text{ff}} = \{0.1, 0.2, 0.3\}$, $\text{dropout}_{\text{att}} = \{0.0, 0.1\}$. The models are trained for $\{100, 200\}$ epochs with 25000 warm-up steps and a peak learning rate from $\{5e-3, 1e-2\}$, using ADAM optimizer. The batch size is varied among, $\{64, 96, 128\}$ depending on the available GPU memory.

The CTC weight $\lambda$ when training ASR models was chosen from $\lambda = \{0.3, 0.5, 0.9\}$. Higher CTC weight gave lower WERs when training on low amounts (e.g. 50hr) of data. The CTC weight ($\alpha = \{0.0, 0.1, 0.5\}$) during ST training and decoding ($\beta = \{0.0, 0.1, 0.3, 0.5, 0.7, 0.9\}$) are the main hyper-parameters explored in our experiments, while keeping the rest of the network architecture the same across all the ASR and ST setups. Decoding is done with beam size 10, while best $\beta$ was chosen based on performance on development set.

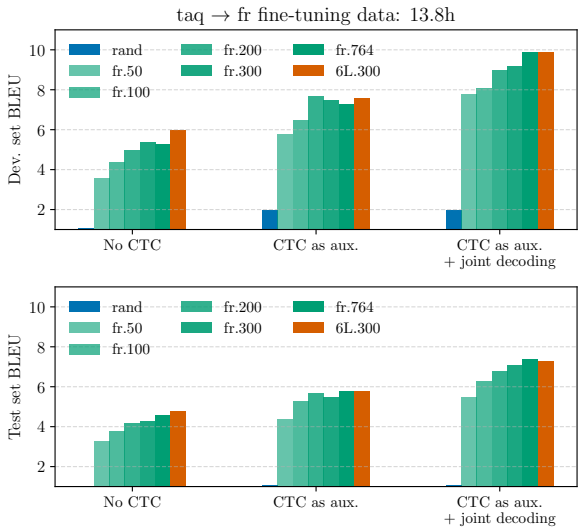


Figure 2: *Performance of ST systems on taq → fr dataset, relying on various initialization, fine-tuning and decoding schemes.*

All the text is tokenized using MOSES toolkit. We retain the true-case and punctuation for both ASR and ST experiments, which allowed us to use the same vocabulary of tokens for both ASR and ST models. Unigram-based segmentation method [22] from SentencePiece [23] was used to learn the sub-word vocabulary of 1000 tokens for each language. The subword segmentation algorithm was trained only on the text transcripts from ASR training data. In case of random initialization of ST models, the ST training data was used of learning the segmentation.

### 3.2. Training details

All the monolingual ASR models have 27.93M parameters, whereas the multilingual ASR has 31.78M. Depending on the size of training data, it took between 6 - 30 hours on a single GPU to train these models. The ST models were initialized from pre-trained ASR models in two ways: (i) retain the CTC layer and perform joint training, (ii) discard the CTC layer to perform standard training with attention loss. In case of initialization from multilingual ASR model, the parameters of non-target language do not get updated. All our experiments were conducted on a custom clone[1] of ESPnet2 framework [3].

## 4. Results and discussion

This section presents the results of ASR and ST systems. Since we trained ASR models on true case text with punctuation, the word-error-rates (WER) would be slightly higher than if we were to train on lower case text. Hence, we report both WER and character error rate (CER) for ASR systems. The ST systems were evaluated using 4-gram BLEU with the help of sacrebleu [24] library[2]. We additionally report CHRF2[3], an F-score based on character $n$-grams [25].

Table 4 presents the results of our ASR systems in terms of word and character error rates (WER, CER). In case of French (fr), we can see that the multilingual ASR model performs

[1] https://github.com/BUTSpeechFIT/espnet/tree/main/egs2/iwslt22_low_resource/st1

[2] nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1

[3] nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1

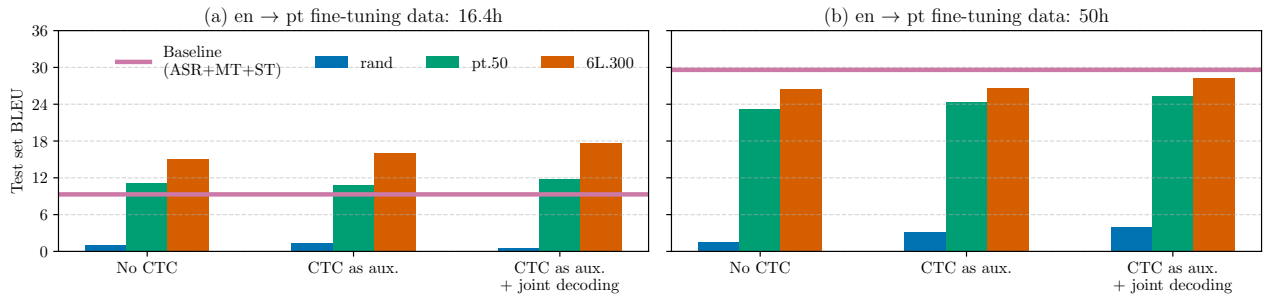


Figure 3: *Effect of various initialization and amounts of ST fine-tuning data.*

worse than the best monolingual ASR in terms of WER. The difference in CER is a bit lower. This is caused by smaller model capacity ($d_{\text{model}} = 256$). However, we still report results with this model, as it would be comparable to the monolingual counter-parts in terms of architecture and parameters.

The ST models initialized from pre-trained ASR are fine-tuned in two ways (i) no CTC (ii) CTC as auxiliary objective. Once the ST model is fine-tuned, the beam-search based decoding can use either CTC score (joint decoding) or not. The Fig. 2 illustrates the performance of ST system relying on various initializations, fine-tuning and decoding schemes. We can observe three things from the Fig. 2

1. CTC as auxiliary objective for translation helps across various initializations. Joint CTC decoding gives further benefits.
2. Target language ASR models (fr.50, fr.100, ..., fr.764) act as good initializations (which was also observed in prior works [7]) for speech translation.
3. The multilingual model trained on 300 hours of speech (6L 300h), which includes only 50 hours of target French data, performs better than most of the French monolingual models trained on much larger data. This suggests that even if the target-language has low-to-moderate amount of transcribed speech, one can rely on a multilingual ASR model.

Table 5 compares our best systems (from Fig 2) with the results reported in the findings of IWSLT'22 [17].

With the low-resource simulation experiments (en → pt), we aim to identify saturation of benefits from pre-trained ASR, given source language transcriptions and source → target parallel data. We trained two source language (en) ASR models on 16.4 and 50 hours of transcribed speech, respectively (Table 2). Then, we trained two en → pt MT systems on the corresponding parallel sentences (10.5k, 31.5k). We used the speech encoder from ASR and decoder from MT model to initialize an ST model, which was then fine-tuned on 16.4 and 50 hours respectively. During this fine-tuning, we use source language transcripts as targets for CTC objective function (Fig 1c). This baseline is represented by (ASR+MT+ST). Fig. 3 shows the BLEU score on test set for all kinds of initializations. Under the low-resource setup of 16.4h, we can see that models based on target-language pre-trained ASR outperform the baseline by a decent margin. In case of mid-resource setup, i.e., with 50 hours of data, the gap reduces to 1 BLEU score. Both Fig 2 and 3 have same trends, that CTC as auxiliary objective for translation and joint decoding is beneficial. We also experimented with various CTC weights ($\alpha$) during training. While in most of the low-resource setups, $\alpha = 0.1$ seemed to give best result. As the amount of ST fine-tuning data increased, we observed that higher CTC weight $\alpha = 0.5$ yielded better results. However, a further investigation on the influence of pre-trained multilingual ASR models in high-resource setups is required.

Table 5: *Performance of ST systems on taq → fr.* †*The findings of IWSLT [17] reports* CHRF++*, however their sacrebleu signature (footnote 30) with option* `nw:0` *suggests that it is* CHRF*, with an unknown $\beta$. Hence, the numbers cannot be compared.*

| System | Dev. BLEU | Dev. CHRF2 | Test BLEU | Test CHRF2 |
|---|---|---|---|---|
| Wav2vec2 (taq) + ST [17] | 8.3 | - | 5.7 | 31.4† |
| ASR + ST [18] | 6.4 | - | 5.0 | - |
| XLS-R + mBART [17] | - | - | 2.7 | 24.3† |
| Mono (fr 764h) + ST | 9.9 | 35.2 | **7.4** | 30.9 |
| Multi (6L 300h) + ST | 9.9 | 34.9 | **7.3** | 30.5 |

## 5. Conclusion

In this paper, we revisited several strategies for improving low-resource speech translation. We combined recent findings from joint-training and decoding in ASR and direct speech translation techniques and studied them with-respect-to various initializations in low-resource scenarios. Our experiments reconfirmed prior works that target-language ASR acts as good initialization for downstream speech translation. In addition, we found that pre-trained multilingual ASR is a viable alternative and performs better than the monolingual ASR in a majority of the settings. Finally, with only 300 hours of pre-training, our approaches achieved 7.3 BLEU score on low-resource Tamasheq - French dataset, outperforming prior works from IWSLT 2022.

In the future, we would like to study the effect of multilingual ST fine-tuning, as it should provide additional supervision, thus help the overall translation quality. Another important direction relates to quantifying misaligned representations when initializing modules from different modalities.

## 6. Acknowledgements

The work was supported by Horizon 2020 Marie Skłodowska-Curie grant ESPERANTO, No. 101007666, Czech National Science Foundation (GACR) project NEUREM3 No. 19-26934X, Czech Ministry of Interior project VK01020132. Computing on IT4I supercomputer was supported by the Czech Ministry of Education, Youth and Sports through the e-INFRA CZ (ID:90254). This work was inspired by insights gained from JSALT 2022, which was supported by Amazon, Microsoft and Google.

# 7. References


[1] H. K. Vydana, M. Karafiát, K. Zmolikova, L. Burget, and J. Černocký, "Jointly Trained Transformers Models for Spoken Language Translation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7513–7517.

[2] S. Dalmia, B. Yan, V. Raunak, F. Metze, and S. Watanabe, "Searchable Hidden Intermediates for End-to-End Models of Decomposable Sequence Tasks," in *Proc. of the NAACL: HLT*. Online: ACL, Jun. 2021, pp. 1882–1896. [Online]. Available: https://aclanthology.org/2021.naacl-main.151

[3] H. Inaguma, S. Kiyono, K. Duh, S. Karita, N. Yalta, T. Hayashi, and S. Watanabe, "ESPnet-ST: All-in-One Speech Translation Toolkit," in *Proc. of the 58th Annual Meeting of the ACL: System Demonstrations*. Online: ACL, Jul. 2020, pp. 302–311. [Online]. Available: https://aclanthology.org/2020.acl-demos.34

[4] B. Zhang, B. Haddow, and R. Sennrich, "Revisiting End-to-End Speech-to-Text Translation From Scratch," in *International Conference on Machine Learning*, ser. Proc. of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, July 2022, pp. 26 193–26 205.

[5] S. Strassel and J. Tracey, "LORELEI Language Packs: Data, Tools, and Resources for Technology Development in Low Resource Languages," in *Proc. of LREC*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 3273–3280. [Online]. Available: https://aclanthology.org/L16-1521

[6] S. Bansal, H. Kamper, A. Lopez, and S. Goldwater, "Cross-Lingual Topic Prediction For Speech Using Translations," in *Proc. of ICASSP*. IEEE, May 2020, pp. 8164–8168.

[7] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," in *Proc. of the NAACL:HLT*, J. Burstein, C. Doran, and T. Solorio, Eds. ACL, June 2019, pp. 58–68.

[8] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Proc. of Interspeech*, H. Ko and J. H. L. Hansen, Eds. ISCA, September 2022, pp. 2278–2282. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-143

[9] M. C. Stoian, S. Bansal, and S. Goldwater, "Analyzing ASR Pretraining for Low-Resource Speech-to-Text Translation," in *Proc. of ICASSP*. IEEE, May 2020, pp. 7909–7913.

[10] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual Denoising Pre-training for Neural Machine Translation," *Transactions of the ACL*, vol. 8, pp. 726–742, 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.47

[11] S.-P. Chuang, Y.-S. Chuang, C.-C. Chang, and H.-y. Lee, "Investigating the Reordering Capability in CTC-based Non-Autoregressive End-to-End Speech Translation," in *Findings of the ACL: ACL-IJCNLP 2021*. Online: ACL, Aug. 2021, pp. 1068–1077. [Online]. Available: https://aclanthology.org/2021.findings-acl.92

[12] L. Bentivogli, M. Cettolo, M. Gaido, A. Karakanta, A. Martinelli, M. Negri, and M. Turchi, "Cascade versus Direct Speech Translation: Do the Differences Still Make a Difference?" in *Proc. of the 59th Annual Meeting of the ACL and the 11th IJCNLP*. Online: ACL, Aug. 2021, pp. 2873–2887. [Online]. Available: https://aclanthology.org/2021.acl-long.224

[13] V. A. K. Tran, D. Thulke, Y. Gao, C. Herold, and H. Ney, "Does Joint Training Really Help Cascaded Speech Translation?" in *Proc. of Conference on EMNLP*. Abu Dhabi, United Arab Emirates: ACL, Dec. 2022, pp. 4480–4487. [Online]. Available: https://aclanthology.org/2022.emnlp-main.297

[14] B. Yan, S. Dalmia, Y. Higuchi, G. Neubig, F. Metze, A. W. Black, and S. Watanabe, "CTC alignments improve autoregressive translation," in *Proceedings of the 17th Conference of the EACL*. Dubrovnik, Croatia: ACL, May 2023, pp. 1623–1639. [Online]. Available: https://aclanthology.org/2023.eacl-main.119

[15] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proc. of the 23rd ICML*, ser. ICML '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 369–376. [Online]. Available: https://doi.org/10.1145/1143844.1143891

[16] B. Yan, P. Fernandes, S. Dalmia, J. Shi, Y. Peng, D. Berrebbi, X. Wang, G. Neubig, and S. Watanabe, "CMU's IWSLT 2022 Dialect Speech Translation System," in *Proc. of the 19th IWSLT*. Dublin, Ireland (in-person and online): ACL, May 2022, pp. 298–307. [Online]. Available: https://aclanthology.org/2022.iwslt-1.27

[17] A. Anastasopoulos, L. Barrault, L. Bentivogli, M. Zanon Boito, O. Bojar, R. Cattoni, A. Currey, G. Dinu, K. Duh, M. Elbayad, C. Emmanuel, Y. Estève, M. Federico, C. Federmann, S. Gahbiche, H. Gong, R. Grundkiewicz, B. Haddow, B. Hsu, D. Javorský, V. Kloudová, S. Lakew, X. Ma, P. Mathur, P. McNamee, K. Murray, M. Nădejde, S. Nakamura, M. Negri, J. Niehues, X. Niu, J. Ortega, J. Pino, E. Salesky, J. Shi, M. Sperber, S. Stüker, K. Sudoh, M. Turchi, Y. Virkar, A. Waibel, C. Wang, and S. Watanabe, "Findings of the IWSLT 2022 Evaluation Campaign," in *Proc. of the 19th IWSLT*. Dublin, Ireland (in-person and online): ACL, May 2022, pp. 98–157. [Online]. Available: https://aclanthology.org/2022.iwslt-1.10

[18] M. Zanon Boito, F. Bougares, F. Barbier, S. Gahbiche, L. Barrault, M. Rouvier, and Y. Estève, "Speech Resources in the Tamasheq Language," in *Proc. of LREC*. Marseille, France: European Language Resources ACLociation, Jun. 2022, pp. 2066–2071. [Online]. Available: https://aclanthology.org/2022.lrec-1.222

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[20] S. Karita, N. E. Y. Soplin, S. Watanabe, M. Delcroix, A. Ogawa, and T. Nakatani, "Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration," in *Proc. of Interspeech*, 2019, pp. 1408–1412.

[21] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, "How2: A Large-scale Dataset For Multimodal Language Understanding," in *Proc. of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS, 2018. [Online]. Available: http://arxiv.org/abs/1811.00347

[22] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in *Proc. of the 56th Annual Meeting of the ACL*. Melbourne, Australia: ACL, Jul. 2018, pp. 66–75. [Online]. Available: https://aclanthology.org/P18-1007

[23] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing," in *Proc. of the 2018 Conference on EMNLP: System Demonstrations*. Brussels, Belgium: ACL, Nov. 2018, pp. 66–71. [Online]. Available: https://aclanthology.org/D18-2012

[24] M. Post, "A Call for Clarity in Reporting BLEU Scores," in *Proc. of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: ACL, Oct. 2018, pp. 186–191. [Online]. Available: https://aclanthology.org/W18-6319

[25] M. Popović, "chrF: character n-gram F-score for automatic MT evaluation," in *Proc. of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: ACL, Sep. 2015, pp. 392–395. [Online]. Available: https://aclanthology.org/W15-3049