CTC Regularization for Low-Resource Speech-to-Text Translation

Zachary Hopton and Rico Sennrich
{zacharywilliam.hopton,rico.sennrich}@uzh.ch
University of Zurich

Abstract

The challenges of building speech-to-text translation (ST) systems (e.g., a relative lack of parallel speech–text data and robustness to noise in audio) are exacerbated for low-resource language pairs. In this work, we seek to improve low-resource ST by building on previous studies that regularize ST training with the connectionist temporal classification (CTC) loss. By systematically evaluating a diverse range of linguistic annotations as CTC labels across multiple auxiliary loss configurations, we improve speech translation systems for both low- and high-resource settings. These improvements over both a standard end-to-end ST system and a speech LLM indicate a need for continued research on regularizing speech representations in ST.

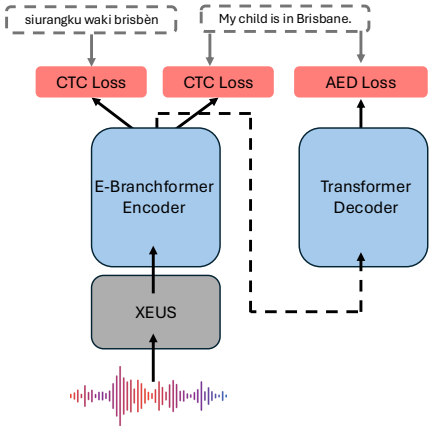

Figure 1: The **ASR-ST** synchronous CTC configuration. Ground truth sequences (transcription and translation of a Tondano utterance from He et al. (2024)) are in dashed gray boxes.

1 Introduction

Training end-to-end speech-to-text (ST) systems requires overcoming a scarcity of parallel data between modalities and the *lack of invariance* problem inherent to speech, whereby many signals can be mapped to the same phoneme (Xu et al., 2023; Appelbaum, 1996). Here, we study the feasibility of developing ST systems that push the former challenge to its limits, i.e., training ST systems for language pairs with under 10 hours of training data. Given that data sources for such extremely low-resource languages often come from linguistic fieldwork, a central question to this work is whether annotations from fieldwork can be used to improve ST models.

We leverage several approaches common in the low-resource ST literature—multilingual feature extractors, ASR pretraining, and regularization—then build on them by experimenting with labels and loss configurations for regularization with connectionist temporal classification (CTC). By directly comparing the effectiveness of using translations, transcriptions, interlinear glossings, and morphologically segmented transcriptions as CTC labels, we add to the body of work studying the most effective choice of label in CTC regularization for ST. We also go beyond single labels and compare various combinations of CTC labels using the synchronous CTC loss presented for high-resource ST by Xu et al. (2024). Our main findings are that morphological segmentation is a useful auxiliary task in low-resource ST, and that training a speech encoder with multiple regularizing objectives is also beneficial in very low-resource ST settings. All code to replicate the experiments is available[1].

2 Related Work

Because of the variation inherent to the speech modality, training end-to-end ST models presents a substantial modeling burden (Xu et al., 2023). With very small amounts of data (e.g., less than 5 hours) it is rare to achieve BLEU scores over 5.0 (Ortega et al., 2024). Still, a number of efforts have been made to improve very low-resource ST systems.

[1] https://github.com/zhopto3/lores-ctc-reg

Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026), pages 186–197

Pretraining Pretraining some or all of an ST model's parameters can benefit low-resource ST. For instance, fine-tuning ASR models with small amounts of ST data is more effective than using the same amount of data to train an end-to-end ST model from scratch, even when the ASR model is for a language unrelated to the source or target (Bansal et al., 2019; Zhang et al., 2022). Moreover, it has been shown that pretrained, multilingual feature extractors such as XLS-R improve results for low-resource ST, indicating that cross-lingual transfer learning is another benefit of pretraining (Babu et al., 2022; Chen et al., 2024).

Regularization There have also been various efforts to improve end-to-end ST systems with auxiliary loss objectives that aim to regularize encoded representations of source language speech. Tang et al. (2021) used a speech and text encoder with an auxiliary loss objective that moved speech representations of the same sample closer to the text representations. The CTC loss—which jointly predicts the probability of both a sequence and its alignment with the source (Graves et al., 2006)—has also been used to regularize ST training, using both the source-language transcription and translation as labels (Bahar et al., 2019; Dong et al., 2021; Zhang et al., 2022; Yan et al., 2023; Xu et al., 2024). Though CTC regularization for end-to-end speech models was originally applied at the final encoder layer (Kim et al., 2017), applying CTC loss to features at intermediate layers in the encoder has been shown to benefit multilingual ASR (Chen et al., 2023).

3 Approach

3.1 CTC Regularization

The CTC loss over encoded speech representations can be jointly optimized with the attention-based encoder-decoder (AED) loss to regularize ST training (Kim et al., 2017). This formulation requires a source audio $X = x_1, ..., x_T$ and some target sequence of labels $S = s_1, ..., s_V$. The weights in an encoder H are optimized so the encoded representations of X minimize the following loss:

$$\mathcal{L}_{\text{CTC}} = \sum_{A \in A_{(X,S)}} \prod_{t=1}^{T} p(a_t|x_t), \quad (1)$$

where $A_{(X,S)}$ refers to the set of possible alignments between the source audio X and the target sequence S, and A is a sequence of target tokens and blank symbols, $a_1, ..., a_T$. A sequence is considered to align with the target output if it equals the target output after collapsing across equal, adjacent symbols and removing blank symbols; in practice dynamic programming is used to carry out this marginalization over valid alignments. See Prabhavalkar et al. (2023) and Graves et al. (2006) for a detailed review of CTC loss. Probabilities at each time step are generally calculated from a learned set of weights that take encoded speech features as input (Baevski et al., 2020; Zhang et al., 2023). Jointly, the encoder and decoder weights are optimized to minimize the autoregressive AED loss using the reference translation $Y = y_1, ..., y_L$ as the label:

$$\mathcal{L}_{\text{AED}} = -\sum_{t=1}^{T} \log P(y_t|y_{<t}, H(X)), \quad (2)$$

where H refers to an encoder model.

3.2 ST Model

Using CTC regularization, we train several ST models by fine-tuning an encoder-decoder model pretrained for multilingual ASR (Chen et al., 2024). We refer to this pretrained model as *XEUS-F*. See Appendix A for model specifications.

3.3 Data

The WAV2GLOSS:FIELDWORK dataset includes audio and interlinear glossing data compiled from fieldwork on 37 typologically and areally diverse languages (He et al., 2024). This data includes source audio transcriptions, morphological segmentations and interlinear glossings of the source audio, and translations into a higher-resource language (Figure 2). We select a sample of 5 languages from FIELDWORK for use in our experiments: Ainu (`ainu1240`), Beja (`beja1238`), Sumi Naga (`sumi1235`), Ruuli (`ruul1235`), and Tondano (`tond1251`). None of the languages we select have more than 10 hours of training data (Table 6), allowing us to compare the effectiveness of various CTC regularization strategies for benefiting extremely low-resource ST. See Appendix B for preprocessing details.

3.4 Experiments

Varying CTC Labels Previous work has used the source language transcription (Bahar et al., 2019; Dong et al., 2021) or the translation (Zhang et al., 2022; Yan et al., 2023) as the label S used

CTC Label		ainu1240	beja1238	ruul1235	sumi1235	tond1251	Mirco Avg.
Baseline	–	17.28	14.89	11.51	16.09	10.99	14.75
Morphological Segmentation	+InterCTC	20.14	17.02	13.11	18.34	13.10	**17.07***
	–InterCTC	18.54	17.74	12.93	17.96	12.15	**16.42***
Interlinear Glossing	+InterCTC	16.64	15.45	13.22	16.86	10.99	14.97
	–InterCTC	18.50	15.69	11.58	15.98	11.05	15.32*
Transcript	+InterCTC	19.89	17.79	13.10	18.15	11.91	16.85*
	–InterCTC	17.93	17.60	13.31	16.45	11.20	15.75*
Translation	+InterCTC	14.62	15.46	12.59	16.33	11.89	14.30*
	–InterCTC	12.44	13.76	10.58	14.99	10.76	12.62*

Table 1: Test set chrF2 scores for ST systems trained with various CTC labels. Baseline system is trained without CTC regularization. "+/- InterCTC" refers to whether intermediate CTC modules were used in training. **Bold** scores represent the best systems on average; underlined values represent the best systems for a given source language; *: Systems that differ significantly ($p < 0.05$) from the Baseline on average.

as the ground truth in Eq. 1. We experiment with these labels, as well as morphologically segmented transcriptions and interlinear glosses. Given that multitask learning of linguistic annotation tasks has been shown to improve low-resource MT systems (e.g., Zaremoodi et al. (2018)), we suspect that ST training may benefit from these labels in particular.

Specifically, we fine-tune XEUS-F for Xx→En ST using our sample of five source languages from FIELDWORK. For each language pair, we fine-tune several ST models, taking a different sequence label as the CTC ground truth in each. We jointly minimize the losses in Eq. 1 and 2, as well as K CTC losses calculated with intermediate encoder features (Chen et al., 2023, 2024). All CTC modules in a given fine-tuned model are optimized with the same label. The final joint CTC-AED loss used is as follows:

$$\mathcal{L}_{\text{joint}} = \lambda(w\left(\frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_{\text{CTC-}k}\right) + (1-w)\mathcal{L}_{\text{CTC}}\Bigg) + (1-\lambda)\mathcal{L}_{\text{AED}} \quad (3)$$

where $K = 3$ (intermediate CTC modules at layers 3, 6, and 9 of the encoder), $w = 0.3$, and $\lambda = 0.3$, as in Chen et al. (2024). For comparison, we also train ST models where $K = 0$, i.e., without intermediate CTC modules.

Synchronous CTC Regularization In addition to studying the impact of intermediate CTC modules on low-resource ST, we compare various configurations of the synchronous CTC loss presented by Xu et al. (2024) (Figure 1). This consists in using the encoder's output features as input to B CTC modules, each with a different ground truth label S:

$$\mathcal{L}_{\text{ML}} = \lambda(\frac{1}{B}\sum_{b=1}^{B}\mathcal{L}_{\text{CTC-}b}) + (1-\lambda)\mathcal{L}_{\text{AED}} \quad (4)$$

We weight all CTC modules equally when calculating the loss and use a value of $\lambda = 0.5$. Using the loss in Eq. 4, we create several conditions: ***ASR-SEG*** ($B = 2$, S_1: transcription, S_2: morphological segmentation), ***ASR-ST*** ($B = 2$, S_1:transcription, S_2: translation), and ***ALL*** ($B = 4$, S_1: transcription, S_2: translation, S_3: morphological segmentation, S_4: interlinear glossing). To isolate the effects of this loss configuration, we do not use intermediate CTC modules in the Synchronous CTC Regularization experiments.

Throughout the results, we make comparisons to a **Baseline** model, which refers to XEUS-F fine-tuned for ST in a given language pair with only the AED loss (Eq. 2) and no CTC modules. See Appendices C and D for details on fine-tuning, inference, and evaluation. We evaluate using chrF2 (Popović, 2016). All significance testing is carried out using the SacreBLEU implementation of paired bootstrap resampling, with $N = 1000$ and a significance threshold of 0.05 (Post, 2018; Koehn, 2004).

4 Results

4.1 Varying the CTC Label

When experimenting with the label used for all CTC modules during ST fine-tuning, we find that using the transcription or the morphologically segmented transcription as CTC labels yields the largest improvements over the baseline model in terms of chrF2 (Table 1). In models with intermediate CTC modules, there is no significant difference

	ainu1240	beja1238	ruul1235	sumi1235	tond1251	Micro Avg.
Baseline	17.28	14.89	11.51	16.09	10.99	14.75
ASR-SEG	20.31	17.11	12.12	18.0	11.92	**16.76***
ASR-ST	20.14	17.0	12.55	16.98	12.06	16.52*
ALL	21.0	16.36	12.66	16.93	10.79	16.54*

Table 2: Test set chrF2 for varied Synchronous CTC configurations. Baseline system is trained without CTC regularization. *: Systems that differ significantly ($p < 0.05$) from Baseline on average.

	FIELDWORK	CoVoST 2
ASR-ST	**16.52**	**56.85**
+ASR, -ST	15.87	55.92
-ASR, +ST	14.03	53.16
-ASR, -ST	14.75	54.88

Table 3: Ablating CTC labels from the ***ASR-ST*** configuration. Values are micro-averaged test set chrF2 scores.

between transcriptions and morphologically segmented transcriptions ($p > 0.05$), but in models without intermediate CTC modules, morphological segmentation significantly outperforms transcripts ($p < 0.001$). Improvements from interlinear glosses are not significant with intermediate CTC modules, and on average, translations as CTC labels resulted in significantly worse models than the baseline. We report BLEU scores in Appendix E (Papineni et al., 2002).

4.2 Synchronous CTC Regularization

We find that all Synchronous CTC configurations significantly improve ST performance over the baseline model on average (Table 2). The ***ASR-SEG*** yields significant improvements over both the ***ASR-ST*** and ***ALL*** conditions.

Whereas Xu et al. (2024) explore only the ***ASR-ST*** setting, we find that the synchronous CTC loss is also beneficial when other labels are used. Xu et al. proposed that the main benefits of the loss come from training the model to encode language agnostic representations of source speech, but our findings broaden the utility of the loss and find that using it to encode task-agnostic and linguistically informed features is also beneficial for ST. Still, the ***ASR-ST*** configuration is noteworthy because it combines the two most readily available annotations and outperforms the use of translations or transcriptions alone. We therefore further investigate whether the benefits of ***ASR-ST*** generalize to a high-resource setting.

4.3 High-Resource ST

Using the Catalan, German, Spanish and French data from CoVoST 2 (Wang et al., 2020a,b), we fine-tune XEUS-F for Xx→En ST with the ***ASR-ST*** configuration (see Table 7 for data description). We then ablate each CTC label and compare to a model with no CTC regularization. No intermediate CTC modules are used. We include the analogous low-resource settings' values for comparison.

On average, fine-tuning with the ***ASR-ST*** configuration yields the highest chrF2 (Table 3). Fine-tuning with only the transcription label is beneficial on average, but using only the translation label for CTC regularization hurts performance relative to the baseline. Within low- and high-resource settings, all decreases in performance from the ***ASR-ST*** system are significant (all $p < 0.05$).

Previous work has found that despite breaking the CTC loss's assumption of monotonicity between the source and target sequence, models trained with the CTC loss can effectively carry out ST (Chuang et al., 2021). Moreover, using translation has been shown to be beneficial as a regularizing CTC task in MT and ST (Zhang et al., 2022; Yan et al., 2023). Our findings diverge from this, as they indicate significantly worse performance than our baseline when using translation as the CTC label, albeit with small magnitude differences. This discrepancy with previous work might be caused by our choice in the CTC module's relative weight with the AED loss. Unlike Zhang et al. and Yan et al., we fine-tune a model pretrained for ASR, so the lack of correspondence between the pretraining task and the CTC regularization task may contribute to this finding as well.

4.4 Speech LLM Comparison

Large language models (LLMs) adapted for the speech modality have become increasingly popular for ST (Gaido et al., 2024). Synchronous CTC regularization is most readily applicable to encoder-decoder ST models, so we compare XEUS-

	Avg. chrF2
Baseline	14.75
ASR-ST	**16.52**
Gemini 2.0–0 Shot	13.94
Gemini 2.0–3 Shot	13.65

Table 4: Test set chrF2 scores micro-averaged across our sample of languages from FIELDWORK. Baseline system is XEUS-F fine-tuned without CTC regularization.

F fine-tuned with synchronous CTC regularization to Google's Gemini 2.0 Flash (Comanici et al., 2025). Though the exact language composition of Gemini's training data is not known, recent work has examined the model's performance in low-resource ST settings (Beyene et al., 2025; Dauvet et al., 2025). Using the same subset of FIELDWORK languages, we explore whether encoder-decoder models with CTC regularization are still competitive for very-low-resource ST. See Appendix F for a description of the prompt.

We find that the fine-tuned encoder-decoder outperforms Gemini for ST, even when the model is trained without any CTC Regularization (Table 4). The findings are in line with recent work showing that speech LLMs struggle with low-resource ST and ASR (Beyene et al., 2025; Fong et al., 2025). When pretrained with comparable data, Lam et al. (2025) actually found that encoder-decoder models consistently perform on par with or better than decoder-only ST and ASR models.

5 Conclusion

We set out to study novel labels and configurations for CTC regularization to get the most out of small amounts of ST data. In doing so, we found that morphologically segmented transcriptions can be more beneficial than using translations or transcriptions as CTC labels. We also find evidence that extends the utility of Synchronous CTC to low-resource ST, as simultaneously training ST encoders to produce representations that are beneficial for several CTC tasks ultimately improved ST performance. We hope this encourages further work on the role auxiliary training objectives can have in training ST systems for very low-resource language pairs.

Limitations

The extent to which we could study the various annotations' effectiveness as labels for CTC regularization is limited by the pretrained model's tokenizer. The pretrained tokenizer likely split the morphological segmentations and interlinear glossing along arbitrary lines. Using an ST model trained from scratch would allow for training a more task-agnostic tokenizer, for example working from the byte or character level. Still, given the relatively small BPE vocabulary of our model (6,500 items), transcripts and morphemes in our low-resource source languages, English translations, and interlinear glosses were ultimately tokenized to characters or otherwise very small units. For instance, the Tondano transcription "PA'AYANGEN NèOKI" is tokenized as [_, P, A, ‘, A, Y, A, NG, EN, _, N, è, O, K, I], while its English translation is tokenized as [_,CH,I,L,D,RE,N,',S,_,T,O,Y,S]. This being the case, we do not have reason to believe that this particular tokenizer biased performance towards or against a given label.

Though the ***ASR-SEG*** configuration yields the numerically highest score on average, we are only able to explore the impact of the ***ASR-ST*** configuration in a high-resource setting. This limitation might be remedied by future work studying whether automatic interlinear glossing or morphological segmentation can be used to synthesize these annotations.

Acknowledgments

RS was funded by the Swiss National Science Foundation (project MUTAMUR; no. 213976).

References

Irene Appelbaum. 1996. The lack of invariance problem and the goal of speech perception. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1541–1544. IEEE.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: self-supervised cross-lingual speech representation learning at scale. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 2278–2282. ISCA.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Parnia Bahar, Tobias Bieschke, and Hermann Ney. 2019. A comparative study on end-to-end speech to text translation. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019*, pages 792–799. IEEE.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 58–68, Minneapolis, Minnesota. Association for Computational Linguistics.

Luel Hagos Beyene, Vivek Verma, Min Ma, Jesujoba O Alabi, Fabian David Schmidt, Joyce Nakatumba-Nabende, and David Ifeoluwa Adelani. 2025. msteb: Massively multilingual evaluation of llms on speech and text tasks. *arXiv preprint arXiv:2506.08400*.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

William Chen, Brian Yan, Jiatong Shi, Yifan Peng, Soumi Maiti, and Shinji Watanabe. 2023. Improving massively multilingual ASR with auxiliary CTC objectives. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

William Chen, Wangyou Zhang, Yifan Peng, Xinjian Li, Jinchuan Tian, Jiatong Shi, Xuankai Chang, Soumi Maiti, Karen Livescu, and Shinji Watanabe. 2024. Towards robust speech representation learning for thousands of languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10205–10224, Miami, Florida, USA. Association for Computational Linguistics.

Shun-Po Chuang, Yung-Sung Chuang, Chih-Chiang Chang, and Hung-yi Lee. 2021. Investigating the reordering capability in CTC-based non-autoregressive end-to-end speech translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1068–1077, Online. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit S. Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 81 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. FLEURS: few-shot learning evaluation of universal representations of speech. In *IEEE Spoken Language Technology Workshop, SLT 2022, Doha, Qatar, January 9-12, 2023*, pages 798–805. IEEE.

Jonah Dauvet, Min Ma, Jessica Ojo, and David Ifeoluwa Adelani. 2025. Reassessing speech translation for low-resource languages: Do LLMs redefine the state-of-the-art against cascaded models? In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 149–160, Suzhuo, China. Association for Computational Linguistics.

Qianqian Dong, Rong Ye, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2021. Listen, understand and translate: Triple supervision decouples end-to-end speech-to-text translation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12749–12759. AAAI Press.

Seraphina Fong, Marco Matassoni, and Alessio Brutti. 2025. Speech llms in low-resource scenarios: Data volume requirements and the impact of pretraining on high-resource languages. *arXiv preprint arXiv:2508.05149*.

Marco Gaido, Sara Papi, Matteo Negri, and Luisa Bentivogli. 2024. Speech translation with speech foundation models and large language models: What is there and what is missing? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14760–14778, Bangkok, Thailand. Association for Computational Linguistics.

Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, volume 148 of *ACM International Conference Proceeding Series*, pages 369–376. ACM.

Taiqi He, Kwanghee Choi, Lindia Tjuatja, Nathaniel Robinson, Jiatong Shi, Shinji Watanabe, Graham Neubig, David Mortensen, and Lori Levin. 2024. Wav2Gloss: Generating interlinear glossed text from speech. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 568–582, Bangkok, Thailand. Association for Computational Linguistics.

Kwangyoun Kim, Felix Wu, Yifan Peng, Jing Pan, Prashant Sridhar, Kyu Jeong Han, and Shinji Watanabe. 2022. E-branchformer: Branchformer with enhanced merging for speech recognition. In *IEEE*

Spoken Language Technology Workshop, SLT 2022, Doha, Qatar, January 9-12, 2023, pages 84–91. IEEE.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 4835–4839. IEEE.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Tsz Kin Lam, Marco Gaido, Sara Papi, Luisa Bentivogli, and Barry Haddow. 2025. Prepending or cross-attention for speech-to-text? an empirical comparison. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2994–3006, Albuquerque, New Mexico. Association for Computational Linguistics.

John E. Ortega, Rodolfo Joel Zevallos, Ibrahim Said Ahmad, and William Chen. 2024. QUESPA submission for the IWSLT 2024 dialectal and low-resource speech translation task. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 125–133, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, pages 2613–2617. ISCA.

Maja Popović. 2016. chrF deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 499–504, Berlin, Germany. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Rohit Prabhavalkar, Takaaki Hori, Tara N Sainath, Ralf Schlüter, and Shinji Watanabe. 2023. End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:325–351.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Masao Someki, Kwanghee Choi, Siddhant Arora, William Chen, Samuele Cornell, Jionghao Han, Yifan Peng, Jiatong Shi, Vaibhav Srivastav, and Shinji Watanabe. 2024. Espnet-ez: Python-only espnet for easy fine-tuning and integration. In *IEEE Spoken Language Technology Workshop, SLT 2024, Macao, December 2-5, 2024*, pages 863–870. IEEE.

Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. 2021. Improving speech translation by understanding and learning from the auxiliary text translation task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4252–4261, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020a. CoVoST: A diverse multilingual speech-to-text translation corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4197–4203, Marseille, France. European Language Resources Association.

Changhan Wang, Anne Wu, and Juan Pino. 2020b. Covost 2: A massively multilingual speech-to-text translation corpus. *Preprint*, arXiv:2007.10310.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. Espnet: End-to-end speech processing toolkit. In *Proc. Interspeech*, pages 2207–2211.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. 2017. Hybrid

ctc/attention architecture for end-to-end speech recognition. *IEEE J. Sel. Top. Signal Process.*, 11(8):1240–1253.

Chen Xu, Xiaoqian Liu, Erfeng He, Yuhao Zhang, Qianqian Dong, Tong Xiao, Jingbo Zhu, Dapeng Man, and Wu Yang. 2024. Bridging the gaps of both modality and language: Synchronous bilingual ctc for speech translation and speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12176–12180. IEEE.

Chen Xu, Rong Ye, Qianqian Dong, Chengqi Zhao, Tom Ko, Mingxuan Wang, Tong Xiao, and Jingbo Zhu. 2023. Recent advances in direct speech-to-text translation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23.

Brian Yan, Siddharth Dalmia, Yosuke Higuchi, Graham Neubig, Florian Metze, Alan W Black, and Shinji Watanabe. 2023. CTC alignments improve autoregressive translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1623–1639, Dubrovnik, Croatia. Association for Computational Linguistics.

Shu-wen Yang, Heng-Jui Chang, Zili Huang, Andy T Liu, Cheng-I Lai, Haibin Wu, Jiatong Shi, Xuankai Chang, Hsiang-Sheng Tsai, Wen-Chin Huang, and 1 others. 2024. A large-scale evaluation of speech foundation models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198.

Poorya Zaremoodi, Wray Buntine, and Gholamreza Haffari. 2018. Adaptive knowledge sharing in multi-task learning: Improving low-resource neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 656–661, Melbourne, Australia. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Rico Sennrich. 2022. Revisiting end-to-end speech-to-text translation from scratch. In *International conference on machine learning*, pages 26193–26205. PMLR.

Biao Zhang, Barry Haddow, and Rico Sennrich. 2023. Efficient CTC regularization via coarse labels for end-to-end speech translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2264–2276, Dubrovnik, Croatia. Association for Computational Linguistics.

A Xeus-F

	XEUS Layer	Encoder Layer	Decoder Layer
Num. Attention Heads	8	8	8
Hidden State Size	1024	512	512
Feed-Forward Network Size	4096	2048	2048
Convolutional Kernel Dim.	31x31	31x31	—

Table 5: Configuration of each layer in the XEUS-F feature extractor (XEUS), encoder, and decoder. Convolutional kernel dimension only applies to the E-branchformer architecture of XEUS and the encoder. All values are from Chen et al. (2024).

XEUS-F is a model trained in Chen et al. (2024). It consists in a 12-layer E-branchformer encoder and a 6-layer transformer decoder (Kim et al., 2022; Vaswani et al., 2017). The input to the encoder are features from XEUS, a feature extractor trained with several masked prediction objectives on speech from over 4,000 languages (Chen et al., 2024). XEUS-F was trained for joint language identification and multilingual ASR using the FLEURS dataset (Conneau et al., 2022). Across all 102 languages in FLEURS, there are 987 hours of training data. XEUS consists of 577M parameters, while the downstream encoder and decoder total 100M parameters, making a total of 677M parameters in XEUS-F. Weights in the feature extractor were frozen during XEUS-F's ASR training. The layer configurations in the encoder and decoder are described in Table 5.

The model uses a multilingual BPE vocabulary of 6,500 items trained using the FLEURS transcriptions (Sennrich et al., 2016). All Latin characters are uppercase except for those with diacritics.

B Data

surface	mina	tura	mo	-no	a	hine	ipe.
underlying	mina	tura	mo	-no	a	hine	ipe
gloss	laugh	together.with	quiet	-ADV	sit.SG	and	eat
transcription	mina tura mono a hine ipe.						
translation	He sat down smiling and ate his dinner.						

Figure 2: A sample data point from the Ainu corpus in FIELDWORK (He et al., 2024). Data points are annotated with their surface and underlying morphological segmentations (the uttered allomorph versus the abstract underlying form of the morpheme, respectively), an interlinear gloss describing the lexical meaning or grammatical function of a unit, a transcription, and a translation. When morphological segmentations are used as CTC labels throughout our experiments, we refer to the underlying segmentations.

	Hours (train/dev/test)	Number of Samples (train/dev/test)
ainu1240	7.11/0.27/0.86	6711/248/749
beja1238	1.53/0.07/0.21	5257/241/733
ruul1235	0.92/0.08/0.18	1886/244/390
sumi1235	0.40/0.10/0.30	939/246/727
tond1251	0.22/0.17/0.50	303/249/713

Table 6: Description of the train, development, and test set of FIELDWORK following removal of samples that are empty after preprocessing. We provide both durations (in hours) and the number of parallel samples in each split.

Preprocessing We focus on Xx→En ST, so we use OpenLID-v2 to filter out any samples from the FIELDWORK data without English translations (Burchell et al., 2023). In line with the tokenizer of the pretrained XEUS-F model we fine-tune for ST, we uppercase all text. We remove explicative material appearing in brackets in English translations.

	Hours (train/dev/test)	Number of Samples (train/dev/test)
Catalan	135.55/18.95/20.21	95854/12730/12730
French	264.27/21.75/23.30	207372/14760/14760
German	184.29/20.65/21.55	127824/13511/13511
Spanish	113.10/21.81/22.71	79013/13221/13221

Table 7: Description of the train, development, and test set of our sample of languages from CoVoST 2. We provide both durations (in hours) and the number of parallel samples in each split.

C Fine-Tuning for ST

When further fine-tuning XEUS-F for Xx→En ST, we generally follow the hyperparameters used to fine-tune XEUS-F for ST in Chen et al. (2024). That is, we fine-tune with a constant learning rate of $1e^{-4}$ using the Adam optimizer (Kingma and Ba, 2015). If matching Chen et al.'s (2024) batch size of 32 is not possible because of memory constraints, we use a smaller batch size with gradient accumulation to carry out weight updates. We apply SpecAugment to extracted features using the same configurations used during XEUS-F pretraining (Park et al., 2019). This includes time warping, applying frequency masks between 0 and 30 frequency bins wide, and applying time masks between 0 and 40 time steps wide. We fine-tune the encoder and decoder but freeze the parameters in XEUS. Models are trained for a maximum of 100 epochs, with an early stopping mechanism that ends training if 5 epochs pass without improvement in the validation loss. The five models with the highest token-level accuracy on the development set are retained and averaged.

We fine-tune with automatic mixed precision using ESPnet-EZ (Someki et al., 2024; Watanabe et al., 2018). The library s3prl is used to integrate the pretrained XEUS feature extractor with the downstream XEUS-F model (Yang et al., 2021, 2024). The loss functions used for fine-tuning are detailed in 3.4. We fine-tune all models using a single GPU. If the required memory allowed for it in a given experiment, this GPU was a 32 GB V100 GPU. Otherwise, we used an 80 GB A100 GPU.

D Inference and Evaluation

We run inference using beam search with a beam size of 10. There are various algorithms for combining tokens' posterior probability distributions as predicted by the CTC module and the decoder of models trained with joint CTC-attention loss (Yan et al., 2023; Watanabe et al., 2017). However, as in Zhang et al. (2022), we do not use the CTC module during decoding. We use length-normalized scores during beam search, and we do not inform decoding with external language models. Inference is carried out on a single 32 GB V100 GPU.

For brevity, we report corpus-level chrF2 scores throughout Section 4, as in Chen et al. (2024). Scores are calculated after removing language tags output by XEUS-F, using the SacreBLEU library (Post, 2018). We also report corpus-level BLEU scores from our experiments (Papineni et al., 2002) in Appendix E.

E BLEU Evaluation

Condition		ainu1240	beja1238	ruul1235	sumi1235	tond1251	Micro Avg.
Baseline	–	1.99	3.41	0.42	0.55	0.23	2.11
Morphological Segmentation	+InterCTC	2.74	3.72	0.39	1.58	0.41	**2.57***
	–InterCTC	2.63	4.38	0.44	1.67	0.39	**2.73***
Interlinear Glossing	+InterCTC	1.83	3.95	0.36	0.82	0.28	2.04
	–InterCTC	2.79	3.43	0.29	0.83	0.33	2.48*
Transcript	+InterCTC	2.74	4.30	0.46	1.39	0.39	**2.57***
	–InterCTC	2.50	4.39	0.43	0.71	0.30	2.55*
Translation	+InterCTC	1.33	3.71	0.29	0.61	0.21	1.80*
	–InterCTC	1.41	3.26	0.52	0.60	0.24	1.99

Table 8: Test set BLEU scores for ST systems trained with various CTC labels. Baseline system is trained without CTC regularization. "+/- InterCTC" refers to whether intermediate CTC modules were used in training. **Bold** scores represent the best systems on average; underlined values represent the best systems for a given source language; *: Systems that differ significantly ($p < 0.05$) from the Baseline on average.

	ainu1240	beja1238	ruul1235	sumi1235	tond1251	Micro Avg.
Baseline	1.99	3.41	0.42	0.55	0.23	2.11
ASR-SEG	3.06	3.97	0.37	1.01	0.36	**2.78***
ASR-ST	2.74	4.46	0.15	0.94	0.26	2.50*
ALL	3.21	3.62	0.43	0.57	0.23	2.54*

Table 9: Test set BLEU scores for ST systems trained with various Synchronous CTC configurations. Baseline system is trained without CTC regularization; *: Systems that differ significantly ($p < 0.05$) from Baseline on average.

	FIELDWORK	CoVoST 2
ASR-ST	2.50	**30.40**
+ASR, -ST	**2.68**	29.44*
-ASR, +ST	1.94*	26.19*
-ASR, -ST	2.11*	28.25*

Table 10: Ablating CTC labels from the ***ASR-ST*** configuration. Values are micro-averaged test set BLEU scores.

	Average
Baseline	2.11
ASR-ST	**2.50**
Gemini 2.0–0 Shot	0.59
Gemini 2.0–3 Shot	0.52

Table 11: Test set BLEU scores micro-averaged across our sample of languages from FIELDWORK. Baseline system is XEUS-F fine-tuned without CTC regularization.

F Speech LLM Comparison

When carrying out ST inference with the speech language model `gemini-2.0-flash`, we used a prompt similar to that described by Beyene et al. (2025). In the zero-shot condition, we prompted the model with:

"You are a translation expert. Listen to the following audio in {LANGUAGE} and translate it to English. Return only the translated sentence."

In the three-shot condition, we randomly selected three samples from the source language's validation set and provided them to the models with their English translation as examples in the prompt.