LIUM Submission for IWSLT 2026 Low-Resource Speech Translation Track

Mohammad Mohammadamini
LIUM, Le Mans University
first.last@univ-lemans.fr

Marie Tahon
LIUM, Le Mans University
first.last@univ-lemans.fr

Abstract

This paper describes the LIUM submission to the IWSLT 2026 low-resource speech translation track. It proposes different data augmentation methods for low-resource speech-to-text translation, including two main pipelines: pseudo-labeling and speech synthesis. The goal is to generate parallel speech data in low-resource scenarios without relying on human-annotated speech translation data. Our submission focuses on Central Kurdish–English language pairs. The objective of this work is to explore the advantages and limitations of each data augmentation method. Our best results are obtained using the pseudo-labeling pipeline, achieving a BLEU score of 25.73 on the development set and 21.09 on the test set for Central Kurdish–English translation.

Keywords: Low-resource speech translation, Pseudo-labeling, Synthetic data, Central Kurdish.

1 Introduction

End-to-end speech translation provides a seamless way to facilitate communication across different languages. Training end-to-end (E2E) speech-to-text translation (S2TT) models requires large and diverse amount of speech data in the source language aligned with translations in the target language (Barrault et al., 2025). The lack of such resources motivates the use of cascaded S2TT configurations, which sequence automatic speech recognition (ASR) and machine translation (MT) models. Although this approach remains competitive, it can lead to error propagation and increased translation latency, thereby hindering simultaneous speech translation.

In the current work, we do not simply perform cascaded S2TT; instead, we explore different data augmentation approaches for end-to-end speech-to-text translation. To achieve this goal, we propose two approaches for generating parallel speech data from source-language speech and target-language text. The first approach combines three components sequentially: speech segmentation, ASR, and MT modules, which are enhanced to generate high-quality parallel data suitable for end-to-end speech translation. In the second pipeline, we propose the use of synthetic data by introducing a low-resource text-to-speech (TTS) setup to generate the data required for E2E S2TT. Within this pipeline, we explore different scenarios, such as starting from a raw text corpus, generating the speech modality using TTS, and translating the raw text into the target language using an MT model. In a simpler scenario, we also explore extending existing parallel corpora by adding a speech modality to the low-resource side.

Our submission focuses on Central Kurdish–English (ckb–en) language pairs. This study extends our previous work. In (Mohammadamini et al., 2025b), a framework for pseudo-labeling Central Kurdish raw audio was proposed, achieving significant results on the FLEURS (Conneau et al., 2023) dataset in end-to-end configurations. In (Mohammadamini et al., 2026), a synthetic data generation setup was used for speech translation from Central Kurdish to English and evaluated on read speech benchmarks such as FLEURS and Asosoft (Veisi et al., 2020). In the current study, we extend this line of work by exploring the proposed approaches on spontaneous speech proposed as Kurdish-Commute dataset for IWSLT 2026 low-resource shared track. We also investigate comparisons and combinations of two pipelines and additional speech synthesis scenarios. This study provides an overview of the limitations and potential of different data augmentation approaches for low-resource S2TT. Furthermore, in the current study we explore the performance of different ASR models such as Whisper (Radford et al., 2022), Seamless (Barrault et al., 2025) and Omnilingual (OmnilingualASR et al., 2025) models.

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 124–131

2 Previous work

Two main data augmentation approaches used in this research—pseudo-labeling and synthetic speech—are among the active lines of speech translation/recognition research.

In (Li et al., 2025), the impact of synthetic data for low-resource speech translation on Bemba and North Levantine languages is explored. The reported results show only marginal improvements when combining synthetic data with real speech. In another study (Mizumoto et al., 2025), the impact of synthetic data on training speech large language models (Speech LLMs) for both speech translation and recognition is investigated. The study suggests that using only synthetic data leads to limited performance, whereas combining it with real speech in a balanced setup can yield improvements.

In (Minixhofer et al., 2025), the scaling laws of using synthetic data for model training are examined. The idea is to condition TTS models in a way that resembles real speech datasets. The experiments show that having sufficiently large training data and conditioning the TTS model on real speech distributions can produce synthetic datasets whose performance approaches that of real speech data for ASR. In (Pu et al., 2025), F5-TTS models are used to generate synthetic speech from parallel English–Chinese text corpora, which is then used to fine-tune speech language models for speech-to-speech translation tasks. In (Tran et al., 2025), a guided TTS model is proposed that uses a Gradient Reversal Layer to make representations of real and synthetic speech invariant, which is then used to improve ASR models.

On the other hand, pseudo-labeling is another common data augmentation approach in speech translation. Several studies show that pseudo-labels generated by robust teacher models can outperform human-annotated data (Hwang et al., 2022), mainly due to the consistency of automatically generated labels. Applying effective filtering to pseudo-labels can significantly impact the performance of models trained on them. In (Khurana et al., 2021), a multi-view pruning approach is proposed, where the agreement between pruned models is used as a confidence measure for pseudo-label filtering. In (Higuchi et al., 2022), a method called momentum pseudo-labeling is introduced, consisting of a framework with a pair of online and offline models that interact and learn from each other. Some state of the art speech translation systems, such as Seamless, heavily rely on pseudo-labeling (Barrault et al., 2025). Due to the lack of robust ASR systems for low-resource languages, most efforts have focused on high-resource languages. However, in the context of low-resource languages, it has been shown that using pseudo-labeled speech for self-training can significantly improve model performance (Bartelds et al., 2023). Its effectiveness has also been demonstrated in several low-resource languages, such as Bengali (Nandi et al., 2023), Northern Sami (Getman et al., 2024), and Hindi (Bhogale et al., 2024).

3 E2E S2TT data augumentation approches

3.1 ASR+MT pipeline

The proposed pseudo-labeling pipeline is depicted in Figure 1. It aims to leverage large amounts of unlabeled Central Kurdish speech to generate training data for S2TT. First, a large corpus of natural speech is collected and automatically segmented into shorter utterances. The ASR model, Seamless Large V2 in our case, is fine-tuned, then used to produce source-language transcriptions from the raw speech, *i.e.*, $S_s \rightarrow \hat{T}_s$. These transcriptions are subsequently translated into the target language using an MT model (fine-tuned NLLB (Costa-jussà et al., 2024)), *i.e.*, $\hat{T}_s \rightarrow \hat{T}_t$. This process enables the creation of large-scale pseudo-parallel triplets $(S_s, \hat{T}_s, \hat{T}_t)$, where S_s denotes the source speech, $\hat{T}_s$ the ASR transcription, and $\hat{T}_t$ the translated text. This approach allows scaling S2TT training data by exploiting raw speech without requiring manual annotations.

To ensure the quality of the generated pseudo-labels, a set of filtering criteria is applied to $(S_s, \hat{T}_s, \hat{T}_t)$. First, partially transcribed samples are removed by enforcing a word-per-minute (WPM) constraint on $\hat{T}_s$ within the range (90–200). Very short or long utterances are also discarded based on duration and token length constraints: $S_s < 1$s or $|\hat{T}_s| < 3$, and $S_s > 30$s or $|\hat{T}_s| > 50$. ASR confidence is incorporated by removing samples whose average confidence score of $\hat{T}_s$ is below 0.9. Hallucinations in $\hat{T}_s$ or $\hat{T}_t$ are detected using repeated n-grams (more than 2 repetitions). A length ratio constraint is also applied to ensure consistency between source and target texts: $0.5 < |\hat{T}_s|/|\hat{T}_t| < 1.5$. Finally, samples where more than 50% of tokens in $\hat{T}_t$ correspond to named entities are filtered out. These criteria, empirically

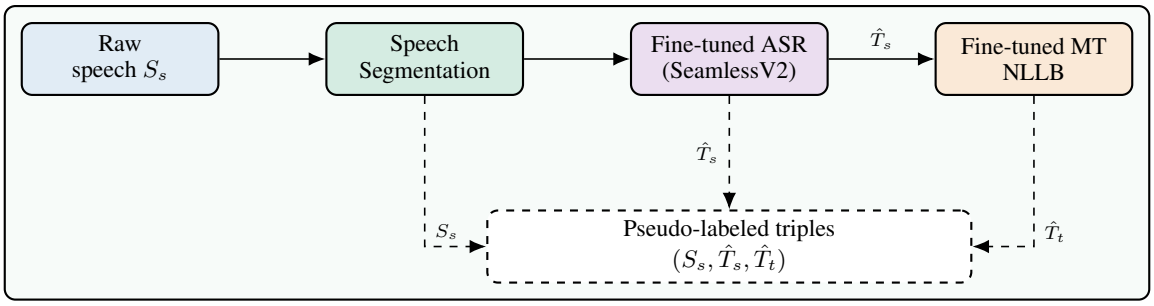

Figure 1: Pseudo-labeling pipeline composed of ASR and MT from real speech

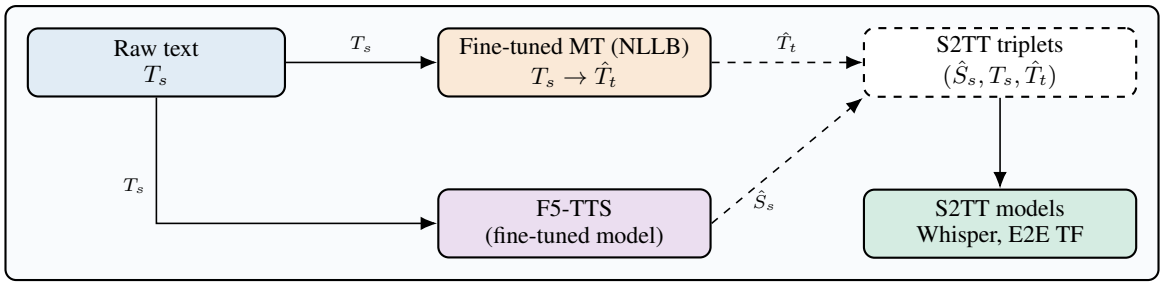

Figure 2: Synthetic data pipeline composed of TTS and MT models generating parallel speech from raw text.

optimized to ensure a high-quality pseudo-labeled dataset for training E2E S2TT systems. More details on this pipeline can be found at (Mohammadamini et al., 2025b).

3.2 TTS+MT pipeline

The synthetic data pipeline relies on TTS models to transform large-scale text corpora into speech data for S2TT training. An F5-TTS (Chen et al., 2025) model is first trained on real Kurdish speech (Mohammadamini et al., 2026), including studio-quality recordings ($\sim$ 13h from a single speaker) and additional audiobook data from multiple speakers (1 female, 1 male, $\sim$ 10h per voice), with transcriptions obtained via ASR and manually corrected.

In the first scenario, we are extending the available parallel MT datasets by synthesizing the sentences of source language. The fine-tuned TTS models are used to synthesize speech from the source text T_s, *i.e.*, $T_s \rightarrow \hat{S}_s$, while the corresponding target text T_t is already available. This process generates synthetic triplets $(\hat{S}_s, T_s, T_t)$, enabling the creation of large-scale parallel speech-text datasets for training S2TT systems.

In a second scenario, we assume access only to raw monolingual text in the source language. In this case, the same fine-tuned F5-TTS model is used to synthesize speech from the source text, $T_s \rightarrow \hat{S}_s$, while an MT system is applied to generate the target text, $T_s \rightarrow \hat{T}_t$. This approach allows transforming raw text corpora into pseudo-parallel data by constructing triplets $(\hat{S}_s, T_s, \hat{T}_t)$. Although this setting introduces additional noise due to MT errors, it is interesting in this study, because we are generating the synthetic speech from a text corpora collected from news media portals, which is a similar domain as the COMMUTE dataset used in this study which comes from Kurdish media. This experiment let us to generate domain adapted data without human intervention. The synthetic speech pipeline is depicted at Figure 2.

4 Experimental setup

4.1 ASR+MT pipeline

The experimental setup relies on a three-stage pseudo-labeling pipeline including speech segmentation, ASR, and MT. Speech segmentation is performed using an energy-based VAD, where segments are created when more than 30 consecutive 10 ms frames are silent, while preserving 15 frames of silence at boundaries. For ASR, the Seamless Large V2 model (Wav2Vec-BERT encoder with NLLB decoder) is fine-tuned on 177 hours of human-annotated Kurdish speech from Common Voice 18 and Asosoft (Veisi et al., 2022), using 80-dimensional Mel filterbank features, a batch size of 16, learning rate of $1e{-}4$, over 10 epochs. The MT component is based on NLLB 1.3B, fine-tuned on 222k parallel CKB–EN sentence pairs using

the Adam optimizer, with a learning rate of $1e{-}4$, weight decay of $1e{-}3$, for 500k steps. The raw audio consists of 4,300 hours of Kurdish audiobooks, which are processed through the pipeline to generate pseudo-labeled data. After quality filtering, the final dataset contains 1.71M triplets $\langle S_s, \hat{T}_s, \hat{T}_t \rangle$, corresponding to 3,231 hours of speech and approximately 22.66M English tokens. These pseudo-labeled data are used to train E2E Fairseq-S2T model (Wang et al., 2020), a transformer-based architecture trained for 25 epochs using Adam (learning rate $2e{-}3$), and Whisper Large V3 (Radford et al., 2022) fine-tuned with AdamW (learning rate $1e{-}5$, batch size 16) for 10 epochs.

5 TTS+MT pipeline

We fine-tuned three mono-speaker Central Kurdish TTS systems based on F5-TTS (Chen et al., 2025), one for each dataset: *audiobook-M* (10h51min, 6044 utterances), *audiobook-F* (10h54min, 5572 utterances), and *studio-M* (13h35min, 6055 utterances). For each dataset, 500 samples were held out as test data and the remaining utterances were used for training. The audiobook recordings were originally in 44 kHz and were downsampled to 24 kHz mono WAV. More details about the sources of training the TTS models is reported at (Mohammadamini et al., 2026).

The TTS models were obtained by fine-tuning the English checkpoint of F5-TTS-base. Since the baseline vocabulary does not cover all Kurdish characters, two preprocessing steps were applied in both training and inference: (i) text normalization including Unicode correction, punctuation normalization, number unification, and number-to-word conversion using the Asosoft toolkit, and (ii) grapheme-to-phoneme (G2P)[1] conversion to make the text compatible with the baseline vocabulary. The three models were trained independently on 1 RTX8000 GPU (48GB) and each fine-tuning run took about 2 days.

In the first scenario, the synthetic speech was generated from parallel Kurdish–English text by randomly selecting one of the three TTS models and a prompt reference from a pool of 5000 utterances sampled from the Central Kurdish portion of Common Voice. This prompt-based setup was used to increase speaker variability in the synthesized data. Using this framework, about 340k synthetic utterances corresponding to approximately 514 hours of speech were generated, mainly from Kuvost (Mohammadamini et al., 2025a) (200k samples) and an additional bitext corpus (140k samples) (Mohammadamini et al., 2025b).

In the second scenario, 100k sentences from (Veisi et al., 2020), collected from Kurdish media, are selected. These sentences are first synthesized using the same setup as in the first scenario and then translated using a fine-tuned NLLB model.

Table 1 gives a brief description of all resources used during the experiments.

6 Results

6.1 Constrained Scenario

In the constrained scenario, we fine-tuned Whisper Large V3 for speech translation over 5 epochs using the training split of the COMMUTE-Kurdish dataset. The results are reported in Table 2. We used an initial learning rate of $1e{-}5$ and updated all model parameters during training. The model converged after the third epoch. It achieved BLEU scores of 18.85 and 14.94 on the development and test sets, respectively. All S2TT scoring are done according to the IWSLT rules: the uppercase are converted to lowercase, the punctuation marks are removed.

6.2 Data augmentation results

Table 3 presents the performance of different data generation strategies for Central Kurdish to English S2TT. When using only synthetic data generated from already available parallel texts, the performance is relatively low, reaching 10.36 BLEU on the test set. This indicates that synthetic speech alone, even with high-quality parallel text, is not sufficient to train robust S2TT models. When synthetic data is generated from raw text using a TTS and MT pipeline, performance improves to 12.69 BLEU. This indicates that targeting the domain of the COMMUTE dataset—derived from Kurdish media—via raw text sources from similar domain can benefit the model. However, the improvements remain limited due to the fully synthetic nature of both the speech and the translations.

A significant improvement is observed when using pseudo-labeled data obtained through ASR+MT pipeline, achieving 17.93 BLEU using fine-tuned Whisper model. This reflects that pseudo-labeling provides more realistic data and better acoustic variability compared to purely synthetic data. Training models from scratch us-

[1] https://github.com/AsoSoft/AsoSoft-Library

Data Type	Source	Size / Duration	Task
IWSLT Shared Task data	COMMUTE-Kurdish	train (9h10min, dev (9h16min), test (11h9min)	Constrained scenario training, Evlauation
Human-annotated ASR data	Common Voice 18 and Asosoft (Veisi et al., 2022)	163.5k examples, 177 hours	Fine-tuning ASR models: Seamless, Whisper, Wav2Vec-BERT CTC
TTS training data	GigaNet and TTS4ALL (Mohammadamini et al., 2026)	13h studio male, 10h audiobook male, 10h audiobook female	fine-tuning F5-TTS
Pseudo-labeled data	Audiobooks (Mohammadamini et al., 2025b)	3231hours	Fine-tuning Whisper for E2E S2TT; training Fairseq-S2T from scratch for both E2E S2TT and ASR
Synthetic parallel data	Kuvost (Mohammadamini et al., 2025a) and previous research (Mohammadamini et al., 2025b)	340k examples	Fine-tuning Whisper for E2E S2TT
Synthetic raw data	Asosoft text corpus (Veisi et al., 2020)	100k sentences	Fine-tuning Whisper for E2E S2TT

Table 1: Summary of datasets used for ASR, TTS, and speech translation experiments.

Set	BLEU	ChrF++
Dev	18.85	41.89
Test	14.94	38.82

Table 2: Constrained situation by Whisper V3 Large model, fine-tuned on the training part of COMMUTE-Kurdish dataset.

ing pseudo-labeled data with Fairseq-S2T (Wang et al., 2020) further increases performance, reaching 21.09 BLEU on the test set. This highlights the advantage of leveraging large-scale pseudo-labeled and language specific models trained from scratch having enough large-scale data. Finally, combining synthetic and pseudo-labeled data results in a performance drop to 16.24 BLEU, indicating that the naive combination of these two data sources does not necessarily lead to complementary gains.

6.3 ASR results

The results obtained for several ASR models are presented in Table 4. In all experiments the same preprocessing are done before scoring which include the removal of Unicode characters unification, punctuation marks, conversion of numbers to alphabetic form. The first experiment evaluates currently released Omnilingual models (OmnilingualASR et al., 2025). Omni-1B-CTC corresponds to a wav2vec-CTC model with 1B parameters, while Omni-1B-LLM refers to an Omnilingual model powered by a large language model. The Omnilingual models are used as baseline and not finetuned on Kurdish data. The three other models, namely W2V-BERT-CTC (Chung et al., 2021), Whisper-L-V3, and Seamless-L-V2, are fine-tuned as part of this study. All models are fine-tuned on 177 hours of human-annotated Kurdish speech (163.5k samples from Common Voice 18 and Asosoft). As shown, the Seamless model significantly outperforms the others.

The final experiment involves Fairseq-S2T, trained on 3,231 hours of pseudo-labeled data generated by the Seamless model—the same data used for training the E2E S2TT models. As shown, this data-driven knowledge distillation approach yields results close to those of the Seamless model, with only a slight performance degradation. The intuition behind this experiment stems from the S2TT results presented in the previous section. As observed, due to the high quality of the pseudo-labeled transcriptions, the language-specific Transformer model trained from scratch on this data ranks second among other models, despite having only 268 million parameters.

6.4 Cascade model results

The cascade model results are obtained using the best-performing ASR system, namely Seamless Large V2, combined with a fine-tuned NLLB model. The results are reported in Table 5. Compared to the Fairseq-S2T model, the cascade approach exhibits almost the same results as our best E2E model.

7 Discussion

Our results on E2E S2TT, using two data augmentation methods—pseudo-labeling and synthetic

Data Setup	System	Dev		Test	
		BLEU	ChrF++	BLEU	ChrF++
Synthetic (parallel)	Whisper-L-V3	10.51	31.92	10.36	30.65
Synthetic (raw)	Whisper-L-V3	13.26	35.13	12.69	33.96
Pseudo-labeled	Whisper-L-V3	17.98	43.01	17.93	41.47
Pseudo-labeled	Fairseq-S2T	**25.73**	**54.28**	**21.09**	**49.48**
Synthetic + Pseudo-labeled	Whisper-L-V3	16.37	40.78	16.24	39.26

Table 3: Comparison of different data generation strategies for E2E S2TT. Synthetic (parallel) refers to the extension of parallel bitext corpora using TTS models. Synthetic (raw) refers to a TTS+MT pipeline, where S2TT data is created using TTS and MT. Pseudo-labeled refers to an ASR+MT pipeline.

System	Dev		Test	
	CER	WER	CER	WER
Omni-1B-CTC	15.09	47.02	15.10	47.23
Omni-1B-LLM	11.10	36.98	10.74	37.26
Whisper-L-V3	13.26	35.64	14.02	37.24
W2V-Bert-CTC	9.25	31.49	9.12	31.88
Seamless-L-V2	**6.70**	**17.31**	**6.98**	**19.76**
Fairseq-S2T	8.33	24.47	8.69	25.78

Table 4: ASR results: Omnilingual models are used without fine-tuning, The Whisper Large V3, Wav2Vec Bert CTC and Seamless Large V2 models are trained on 177 hours of human annotated speech, The Fairseq-S2T Transformer model is trained on pseudo-labeled data generated by Seamless model.

Set	BLEU	ChrF++
Dev	24.77	48.77
Test	22.39	46.28

Table 5: Cascaded approach results (Seamless + NLLB)

data—highlight the limitations of synthetic data in real-world applications such as spontaneous speech translation. The previous studies (Mohammadamini et al., 2026, 2025b) have shown that in more controlled settings, such as read or studio-recorded speech, synthetic data can achieve performance comparable to pseudo-labeled data. However, in spontaneous speech scenarios, the performance of models trained on synthetic data degrades significantly. This degradation can be attributed to the various sources of variability present in spontaneous speech but absent in synthetic data, including speech disfluencies, prosodic variations, code-switching, dialectal diversity (particularly in languages such as Kurdish), and domain shifts associated with conversational speech.

On the other hand, the pseudo-labeling pipeline yields more promising results. Our experiments show that having a robust ASR system is crucial for the effectiveness of the pseudo-labeling pipeline. Although the results obtained with this approach are not yet sufficient for real-world applications, further improvements can be achieved by applying more selective data filtering to both ASR and MT outputs.

8 Conclusion

In this paper, we explore two main data augmentation approaches for low-resource speech translation. The first approach is pseudo-labeling of raw speech using ASR and MT models. The second approach consists of using TTS to synthesize speech from raw or parallel text. Our experimental results highlight the limitations of synthetic speech in real-world applications such as spontaneous speech translation. In contrast, the pseudo-labeling pipeline proves to be more effective, achieving performance comparable to cascaded models while reducing latency. This study could be further extended by designing TTS models conditioned on real-world scenarios or by explicitly modeling the variabilities present in spontaneous speech. According to the results reported in the IWSLT 2026 Findings, our systems outperformed other submitted systems on the Central Kurdish-to-English task in both the constrained and open evaluation settings (Adelani et al., 2026). These results demonstrate the effectiveness of the data augmentation approaches proposed in this paper, particularly the superiority of the pseudo-labeling pipeline in generalizing to real-world spontaneous speech translation.

References

David Ifeoluwa Adelani, Victor Agostinelli, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Sebastien Bratières, Marine Carpuat, Roldano Cattoni,

Mauro Cettolo, Lizhong Chen, Marcello Federico, Marco Gaido, Mahendra Gupta, HyoJung Han, Ali Hatami, David Javorský, Yejin Jeon, Marek Kasztelnik, Antoine Laurent, and 33 others. 2026. Speech translation and metrics in 2026: Findings of the iwslt campaign. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, California, US. Association for Computational Linguistics.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, et. all., and SEAMLESS Communication Team. 2025. Joint speech and text machine translation for up to 100 languages. *Nature*, 637(8046):587–593.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–729, Toronto, Canada. Association for Computational Linguistics.

Kaushal Santosh Bhogale, Deovrat Mehendale, Niharika Parasa, Sathish Kumar Reddy G, Tahir Javed, Pratyush Kumar, and Mitesh M. Khapra. 2024. Empowering low-resource language asr via large-scale pseudo labeling. In *Interspeech 2024*, pages 2519–2523.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2025. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. *Preprint*, arXiv:2410.06885.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. *Preprint*, arXiv:2108.06209.

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

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Yaroslav Getman, Tamas Grosz, Katri Hiovain-Asikainen, and Mikko Kurimo. 2024. Exploring adaptation techniques of large speech foundation models for low-resource asr: a case study on northern sámi. In *Interspeech 2024*, pages 2539–2543.

Yosuke Higuchi, Niko Moritz, Jonathan Le Roux, and Takaaki Hori. 2022. Momentum pseudo-labeling: Semi-supervised asr with continuously improving pseudo-labels. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1424–1438.

Dongseong Hwang, Khe Chai Sim, Zhouyuan Huo, and Trevor Strohman. 2022. Pseudo label is better than human label. In *Interspeech 2022*, pages 1421–1425.

Sameer Khurana, Niko Moritz, Takaaki Hori, and Jonathan Le Roux. 2021. Unsupervised domain adaptation for speech recognition via uncertainty driven self-training. pages 6553–6557.

Zhaolin Li, Yining Liu, Danni Liu, Tuan Nam Nguyen, Enes Yavuz Ugan, Tu Anh Dinh, Carlos Mullov, Alexander Waibel, and Jan Niehues. 2025. KIT's low-resource speech translation systems for IWSLT2025: System enhancement with synthetic data and model regularization. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 212–221, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Christoph Minixhofer, Ondřej Klejch, and Peter Bell. 2025. Scaling Laws for Synthetic Speech for Model Training. In *Interspeech 2025*, pages 3189–3193.

Tomoya Mizumoto, Atsushi Kojima, Yusuke Fujita, Lianbo Liu, and Yui Sudo. 2025. Is Synthetic Data Truly Effective for Training Speech Language Models? In *Interspeech 2025*, pages 1808–1812.

Mohammad Mohammadamini, Daban Jaff, Sara Jamal, Ibrahim Ahmed, Hawkar Omar, Darya Sabr, Marie Tahon, and Antoine Laurent. 2025a. Kuvost: A large-scale human-annotated English to Central Kurdish speech translation dataset driven from English common voice. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 106–109, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Mohammad Mohammadamini, Meysam Shamsi, and Marie Tahon. 2026. Central Kurdish text-to-speech and its application in speech-to-text translation. In *Language Resources and Evaluation Conference (LREC)*, Palma, Spain.

Mohammad Mohammadamini, Aghilas Sini, Marie Tahon, and Antoine Laurent. 2025b. Scaling pseudo-labeling data for end-to-end low-resource speech translation (the case of Kurdish language). In *Interspeech 2025*, pages 898–902.

Rabindra Nath Nandi, Mehadi Menon, Tareq Muntasir, Sagor Sarker, Quazi Sarwar Muhtaseem, Md. Tariqul Islam, Shammur Chowdhury, and Firoj Alam. 2023. Pseudo-labeling for domain-agnostic Bangla automatic speech recognition. pages 152–162.

OmnilingualASR, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, and 14 others. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages. *Preprint*, arXiv:2511.09690.

Yu Pu, Xiaoqian Liu, Guangyu Zhang, Zheng Yan, Wei-Qiang Zhang, and Xie Chen. 2025. Empowering Large Language Models for End-to-End Speech Translation Leveraging Synthetic Data. In *Interspeech 2025*, pages 26–30.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356.

Minh Tran, Debjyoti Paul, Yutong Pang, Laxmi Pandey, Jinxi Guo, Ke Li, Shun Zhang, Xuedong Zhang, and Xin Lei. 2025. R2S: Real-to-Synthetic Representation Learning for Training Speech Recognition Models on Synthetic Data. In *Interspeech 2025*, pages 3194–3198.

Hadi Veisi, Hawre Hosseini, Mohammad MohammadAmini, Wirya Fathy, and Aso Mahmudi. 2022. Jira: a central kurdish speech recognition system, designing and building speech corpus and pronunciation lexicon. *Language Resources and Evaluation*, 56(3):917–941.

Hadi Veisi, Mohammad MohammadAmini, and Hawre Hosseini. 2020. Toward kurdish language processing: Experiments in collecting and processing the asosoft text corpus. *Digital Scholarship in the Humanities*, 35(1):176–193.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. Fairseq S2T: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39, Suzhou, China. Association for Computational Linguistics.