CATENG Submission for the IWSLT 2026 Dialectal and Low-resource Speech Translation Task

Rodolfo Zevallos[1], **Marc Casals**[2], **Pol Buitrago**[2,3], **Fabrício Carraro**[3], **Guillermo Cámbara**[1] , **John E. Ortega**[4]
[1]Universitat Pompeu Fabra, Spain, [2]Universitat Politècnica de Catalunya, Spain
[3]Barcelona Supercomputing Center, Spain, [4]Northeastern University, USA
contact email: rodolfojoel.zevallos@upf.edu

Abstract

We present the **CATENG** systems submitted to the IWSLT 2026 Dialectal and Low-Resource Speech Translation shared task for the Catalan–English (CA–EN) pair. Although Catalan is not strictly low-resource, its dialectal diversity and relative under-representation in speech technology make it a challenging setting. We evaluate three unconstrained systems: two cascaded approaches combining ASR and MT, and one end-to-end model. Our primary system uses a Mamba-based ASR (ConMamba) with a fine-tuned NLLB-200 MT model, while a contrastive system replaces the ASR with Whisper-v3; we also evaluate an end-to-end SpeechT5 model with data augmentation. Experiments are conducted on the IWSLT 2026 Catalan dataset (15 hours), complemented with large-scale parallel text. Results show that cascaded systems outperform end-to-end ST, with Whisper-v3 + NLLB achieving 44.7 BLEU and 65.1 chrF. We find that performance is primarily constrained by ASR quality rather than MT capacity, and that Mamba-based ASR models provide competitive results, highlighting the importance of robust speech representations and dialectal coverage for Catalan–English speech translation.

1 Introduction

In this article, we describe the systems submitted by the **CATENG** team to the IWSLT 2026 Dialectal and Low-Resource Track for Speech Translation (ST), for the Catalan–English (CA–EN) language pair. Although Catalan is not a low-resource language in the strict demographic sense, it remains under-served by speech technology relative to English or Spanish, and its multiple regional varieties make it a natural fit for the *dialectal* component of this year's shared task. The track continues to support an *unconstrained* submission type, and, as in the previous edition, the *constrained* setting has been discontinued.

Catalan (*català*, ISO 639-1: ca) is a Western Romance language with approximately 4.1 million L1 speakers and more than 10 million people who can speak or understand it across its territories. It is spoken primarily in Catalonia, the Valencian Community (where it is locally known as Valencian), the Balearic Islands, and parts of eastern Aragon in Spain; in Andorra, where it is the sole official language; in the *Catalunya Nord* region of southern France; and in the Sardinian city of Alghero. In the Spanish autonomous communities where it is spoken, Catalan is co-official with Spanish and is widely used in education, media, and public administration. Dialectologically, Catalan is typically divided into an Eastern group (Central, Balearic, Northern/Rossellonese, and Alguerese) and a Western group (Northwestern and Valencian), with systematic differences in vowel reduction, clitic systems, and lexicon. Despite structural proximity to other Romance languages, CA–EN speech translation is non-trivial: dialectal variation, frequent Catalan–Spanish code-switching in naturalistic recordings, and the typological distance between Catalan and English all introduce real difficulty. Unlike the closely related CA–SPA direction, CA–EN cannot rely on lexical and syntactic proximity between source and target, making MT quality a decisive factor in the overall ST pipeline.

The data for the task is distributed through the public IWSLT 2026 Catalan dataset repository.[1] We describe the dataset, supplementary publicly available resources, and our preprocessing pipeline in Section 3. Because our machine-translation subsystem is central to the cascaded configurations we submit, and because the availability of large-scale CA–EN parallel text fundamentally changes the trade-offs from previous low-resource editions, we devote a dedicated section to MT experiments in

[1] https://github.com/rjzevallos/IWSLT_2026_Catalan_Dataset

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 164–170

Section 3.1. The remainder of the paper is organized as follows. Section 2 reviews related work on Catalan speech processing and on multilingual speech translation. Section 3 describes our three unconstrained submissions in detail, Section 4 reports results and discussion, and Section 5 concludes.

2 Related Work

In this section, we first review prior work on Catalan speech processing (Section 2.1) and then discuss multilingual speech translation approaches relevant to our unconstrained submissions (Section 2.2).

2.1 Catalan Speech Processing

Although Catalan has a longer tradition of language-technology support than most languages featured in the IWSLT low-resource track, sustained investment in Catalan *speech* technology is comparatively recent. Much of the recent acceleration is attributable to Projecte AINA, a public initiative led by the Barcelona Supercomputing Center to produce open resources and models for Catalan, including ASR corpora, TTS voices, and foundation models. Community-driven efforts have been equally decisive: as of Common Voice v16.1, Catalan ranks as the most represented language in the corpus in terms of both recorded and validated hours (Armentano-Oller et al., 2024), providing the raw material that underpins most open Catalan ASR systems today (Ardila et al., 2020).

For ASR, the ParlamentParla corpus (Külebi et al., 2022), which derives from plenary sessions of the Parlament de Catalunya and comprises roughly 211 hours of clean and 400 hours of "other"-quality segments, has become the standard benchmark for spontaneous and semi-spontaneous Catalan speech. Dialectal coverage has also improved substantially in the most recent releases. The LaFresCat corpus (Peiró-Lilja et al., 2026, 2024) provides a multi-accent Catalan speech dataset with Central, Northwestern, Valencian, and Balearic varieties for TTS, and Projecte AINA has more recently released an accent-stratified ASR benchmark derived from Catalan Common Voice v17 (Projecte AINA, 2024), in which validated recordings are split by the five main accents (Balearic, Central, Northern, Northwestern, and Valencian) and by gender. Together, these resources make it possible, for the first time, to evaluate Catalan speech systems under a controlled dialectal protocol — a prerequisite for the framing of this year's task.

Large multilingual models such as XLS-R (Babu et al., 2021) and Whisper (Radford et al., 2023) include Catalan in their pretraining or weakly-supervised mixtures, and the latter delivers competitive zero-shot Catalan ASR out of the box. Multilingual ST benchmarks such as FLEURS (Conneau et al., 2023) and CoVoST 2 (Wang et al., 2021) include Catalan as a source or target language; CoVoST 2 in particular provides CA–EN pairs that have served as a reference point for Catalan-to-English ST.

For machine translation, the CA–EN pair is substantially better resourced than most pairs featured at IWSLT. Publicly available parallel corpora include OPUS subsets (OpenSubtitles, TED, Europarl, DGT) and JW300 (Agić and Vulić, 2019), complemented by Catalan resources released through AINA. Foundation MT models such as NLLB-200 (NLLB Team et al., 2022) support Catalan directly, and Iberian-focused LLMs released in 2025 by the Barcelona Supercomputing Center — notably the Salamandra family, a set of decoder-only models (2B, 7B, and 40B parameters) trained from scratch on 35 European languages (Gonzalez-Agirre et al., 2025), and its translation-specialised variant SalamandraTA (Barcelona Supercomputing Center, Language Technologies Unit, 2025) — have further raised the bar for zero- and few-shot translation quality between Catalan and English. A recurring observation in this literature is that the real bottleneck for CA–EN ST is not the availability of parallel text but the coverage and dialectal balance of the speech side of the pipeline, together with the handling of Catalan–Spanish code-switching in naturalistic recordings. These considerations directly motivate our submissions.

To our knowledge, this is the first edition of the IWSLT dialectal and low-resource ST track to feature Catalan as a source language, and we are not aware of a prior IWSLT system-description paper targeting Catalan-to-English speech translation, so there is no direct submission baseline to compare against within this venue.

2.2 Multilingual Speech Translation

Multilingual training is a well-established strategy for improving performance on lower-resourced language pairs through cross-lingual transfer. While transfer is most commonly framed as pairing a low-

resource language with one or more high-resource languages, it can also be beneficial among typologically related languages regardless of resource level. Chen et al. (2023) trained multilingual ASR systems across 102 languages, each in a low-resource setting, and obtained state-of-the-art results on FLEURS (Conneau et al., 2023). Whisper (Radford et al., 2023) and OWSM (Peng et al., 2023) scaled multilingual ASR and ST jointly through large supervised training mixtures, producing models that transfer strongly to underrepresented languages and dialects, and notably supporting translation *into English* as a first-class target.

For the Catalan-to-English setting, the most relevant axis of transfer on the speech side is Romance-internal: Spanish, Italian, French, Portuguese, and Occitan share lexical and morphosyntactic structure with Catalan, and multilingual encoders trained on these languages tend to produce representations that generalize well to Catalan and its dialects with minimal adaptation. On the translation side, the abundance of *-to-English data in virtually every multilingual MT and ST model means that the English decoder is typically the strongest component of any cascade, which shifts the burden of quality onto the ASR stage and onto how faithfully the Catalan transcript, including dialectal forms and Spanish borrowings, is preserved before translation. Our unconstrained submissions exploit these properties in two complementary ways: through cascaded pipelines built on Whisper and NLLB, both of which benefit from massive multilingual pretraining and strong English generation, and through direct end-to-end ST fine-tuning of a pretrained speech model on Catalan audio paired with English translations.

3 Catalan-English

In this section, we present our experiments for the Catalan–English (CA–EN) dataset provided in the IWSLT 2026 Dialectal and Low-Resource Speech Translation track. The dataset consists of speech recordings in Catalan covering multiple dialectal varieties, along with their corresponding English translations. In addition to the speech translation data, the organizers provide a substantially larger set of Catalan speech annotated with transcriptions.

To complement the speech resources, we leverage publicly available parallel text for Catalan–English machine translation. In particular, we make use of OPUS collections, including models such as the Helsinki-NLP `opus-mt-ca-en`, which are trained on diverse parallel corpora and provide strong baselines for CA–EN translation. These resources are preprocessed using standard normalization and subword segmentation techniques (e.g., SentencePiece), and have been shown to achieve high translation quality on benchmarks such as Tatoeba.

Overall, the combination of speech data from the shared task and large-scale parallel text from external resources allows us to effectively train both cascaded and end-to-end speech translation systems. Compared to previous low-resource settings, the availability of Catalan–English parallel data shifts the main challenge from text translation to robust speech recognition across dialects and domains.

We present the three submissions for *unconstrained* task only, as this year the constrained task has been abandoned:

1. A **primary unconstrained** system consisting of a Mamba ASR model (Zevallos Salazar et al., 2025) fine-tuned with unconstrained data and cascaded with the best performing NLLB MT system from our case study;

2. A **contrastive 1 unconstrained** system consisting of a Whisper (Radford et al., 2023) ASR model fine-tuned with the unconstrained data and cascaded with the best performing NLLB MT system from our case study;

3. A **contrastive 2 unconstrained** system consisting of a SpeechT5 model (Ao et al., 2022) fine-tuned for speech translation with two data augmentation techniques. (Zevallos et al., 2022).

We present the experimental settings and results for unconstrained systems starting off with the MT case studies in Section 3.1. Then, we describe the task further in Section 3.2. Primary, Contrastive 1 and Constrastive 2 descriptions are found in Sections Sections 3.3, 3.4 and 3.5, respectively. Afterwards, we offer results and discussion in Section 4.

3.1 Machine Translation

All of our MT systems are based on the 1.3B-parameter version of NLLB-200 (NLLB Team et al., 2022), which we fine-tune for Catalan–English translation.[2] During fine-tuning, we set

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

the maximum input and output sequence lengths to 128 tokens. Each model is trained for 10 epochs with a batch size of 8 for both training and evaluation, and decoding is performed with beam search using 5 beams. We save checkpoints every 10,000 steps and fix the random seed to 65 to ensure reproducibility.

3.2 Unconstrained Setting

In line with the IWSLT 2026 Catalan–English shared task, we rely on the dataset released specifically for this edition, which is publicly available in our repository.[3] The dataset comprises approximately 15 hours of Catalan speech paired with corresponding English translations, covering a diverse range of speakers and acoustic conditions.

The corpus has been carefully curated to ensure high-quality alignment between speech and text, and is organized into predefined training, validation, and test splits to facilitate reproducibility and fair comparison across systems. In contrast to previous editions, particular attention has been given to phonetic diversity and speaker variability, which are critical factors for robust speech translation in Catalan.

Additionally, we provide normalized transcriptions and consistent preprocessing pipelines, enabling seamless integration with both ASR and end-to-end speech translation frameworks. These design choices aim to reduce noise introduced by data inconsistencies and allow models to better capture linguistic and acoustic patterns.

All the data resources described in this section are used to fine-tune our end-to-end speech translation models, allowing us to adapt pretrained architectures to the specific characteristics of Catalan–English translation. This setup enables us to systematically evaluate the impact of dataset scale and quality on downstream performance.

3.3 Primary System

The Primary system for the unconstrained setting follows a cascaded speech translation architecture, where the output of an automatic speech recognition (ASR) model is passed to a machine translation (MT) system. For the ASR component, we employ a Catalan ConMamba model [4], a sequence-to-sequence architecture based on the Mamba framework. Unlike traditional Transformer-based approaches, ConMamba relies on state-space modeling rather than attention mechanisms, allowing it to efficiently capture long-range dependencies in speech signals.

The ASR model is trained on a large-scale Catalan speech corpus comprising over 4,900 hours of audio data. Training is performed in a fully supervised setting using four GPUs over approximately 48 hours. The model uses a unigram tokenizer and operates without an external language model during decoding, relying instead on greedy inference. This setup allows us to evaluate the intrinsic modeling capacity of the Mamba-based architecture for speech recognition.

The trained ConMamba model achieves a word error rate (WER) of 8.81 on Catalan speech, demonstrating strong transcription performance despite the absence of language model support. Once the speech is transcribed, the resulting Catalan text is passed to the NLLB-based machine translation system described in Section 3.1, which produces the final English translation. This cascaded configuration enables us to combine a high-quality ASR system with a robust multilingual MT model for effective Catalan–English speech translation.

3.4 Contrastive 1 System

The Contrastive 1 system also follows a cascaded ASR+MT architecture. For the ASR component, we use a Catalan ASR model based on Whisper large-v3 and converted to the `faster-whisper` framework for efficient inference.[5] The original model was fine-tuned from `openai/whisper-large-v3` using 710 hours of Catalan speech from the 3CatParla corpus, making it a strong starting point for Catalan ASR. The model is intended for Catalan transcription and produces plain-text outputs without punctuation.

Starting from this Catalan-adapted Whisper-v3 checkpoint, we further fine-tune the model on the IWSLT 2026 Catalan–English training data. Since the shared-task corpus provides Catalan speech with transcriptions and English translations, we use the Catalan transcriptions as ASR targets during this adaptation step. This allows the model to specialize to the acoustic conditions, speaker distribution, and dialectal variation present in the IWSLT 2026 dataset.

[3] https://github.com/rjzevallos/IWSLT_2026_Catalan_Dataset

[4] https://huggingface.co/langtech-veu/ConMamba-small-ca

[5] https://huggingface.co/projecte-aina/faster-whisper-large-v3-ca-3catparla

During fine-tuning, the model is trained in supervised ASR mode, mapping Catalan speech inputs to Catalan text. At inference time, we decode the audio with the language fixed to Catalan and the task set to transcription. The resulting Catalan hypotheses are then passed to the same fine-tuned NLLB-based MT system described in Section 3.1, which generates the final English translations. Because this system shares the same MT backend as the Primary system, differences in final ST performance mainly reflect the quality of the ASR front-end.

This contrastive configuration allows us to compare a Transformer-based multilingual ASR model, already strongly adapted to Catalan through large-scale 3CatParla training, against the Mamba-based ConMamba ASR used in the Primary system.

3.5 Contrastive 2 System

The Contrastive 2 system for the unconstrained setting is based on SpeechT5 (Ao et al., 2022), a unified encoder–decoder Transformer model designed for both speech and text processing. SpeechT5 is pretrained on 960 hours of speech data from LibriSpeech, together with large-scale unlabeled text corpora, enabling the model to learn shared cross-modal representations. The architecture consists of a 12-layer Transformer encoder and a 6-layer Transformer decoder with a model dimension of 768 and 12 attention heads. The speech encoder incorporates convolutional layers to capture local acoustic patterns, while the model operates in a shared embedding space for speech and text, facilitating direct speech-to-text generation.

We fine-tune SpeechT5 for direct Catalan–English speech translation using the official SpeechT5 fine-tuning recipe.[6] In this setup, the model directly maps Catalan speech inputs to English text outputs without relying on an intermediate transcription stage. Training is performed on the speech translation data provided by the IWSLT 2026 task.

To improve robustness, we apply data augmentation using the *nlpaug* toolkit (Ma, 2019), including noise injection, temporal distortion, and signal perturbation. This effectively doubles the available training data by generating synthetic variants of the original audio. The augmented dataset is used to fine-tune the model under the same hyperparameter configuration as the original SpeechT5 training setup. During inference, the model generates translations autoregressively without the use of external language models.

This end-to-end configuration allows SpeechT5 to jointly model acoustic, linguistic, and translation information within a single framework, providing a direct alternative to cascaded ASR+MT systems.

[6] https://github.com/microsoft/SpeechT5/tree/main/SpeechT5

4 Results and Discussion

Table 1 reports the official results of our three unconstrained submissions for the Catalan–English shared task. Among our systems, the best performance is obtained by Contrastive 1, the cascaded Whisper-v3 ASR + NLLB MT system, which reaches 44.7 BLEU and 65.1 chrF. Our Primary system, based on ConMamba ASR + NLLB MT, follows closely with 43.2 BLEU and 64.3 chrF. The Contrastive 2 end-to-end SpeechT5 system obtains 41.3 BLEU and 63.1 chrF. Overall, all three submissions achieve strong results, with a spread of only 3.4 BLEU between the highest- and lowest-scoring systems.

Two main findings emerge from these results. First, in our setting, cascaded systems outperform the direct end-to-end speech translation model. Both cascades surpass SpeechT5 by a clear margin in BLEU and chrF, suggesting that the combination of a strong multilingual ASR front-end and a separately optimized MT model remains a highly competitive strategy for Catalan–English ST. This is especially plausible in our task because Catalan ASR benefits from robust multilingual pretraining, while the text translation component can leverage comparatively strong CA–EN MT resources. The decomposition of the task into transcription followed by translation therefore appears to be advantageous.

Second, the comparison between our two cascaded submissions indicates that the ASR component is the main differentiating factor. Since both systems use the same NLLB-based MT backend, the gain of Contrastive 1 over the Primary system can be attributed primarily to the quality of the Whisper-v3 transcriptions. Although the gap is modest (1.5 BLEU and 0.8 chrF), it is consistent across both metrics and suggests that Whisper-v3 currently provides a stronger off-the-shelf foundation for this task than our ConMamba configuration. At the same time, the Primary system remains highly competitive, which is encouraging given that Mamba-based speech models are still relatively

Team **CATENG** BLEU and chrF Scores			
Unconstrained 2026			
System	**Description**	**BLEU ↑**	**chrF ↑**
primary	ConMamba ASR + NLLB MT	43.2	64.3 (64.3 ± 3.5)
contrastive 1	Whisper-v3 ASR + NLLB MT	44.7	65.1 (65.1 ± 3.6)
contrastive 2	SpeechT5	41.3	63.1 (63.1 ± 3.2)

Table 1: Official results of the CATENG submissions to the IWSLT 2026 Catalan–English speech translation task under the unconstrained setting. We report BLEU and chrF scores for the three submitted systems: the Primary ConMamba+NLLB cascade, the Contrastive 1 Whisper-v3+NLLB cascade, and the Contrastive 2 end-to-end SpeechT5 system.

new compared to Transformer-based alternatives.

The performance of the end-to-end SpeechT5 system is nevertheless noteworthy. Although it ranks below both cascaded approaches, it still achieves over 41 BLEU, showing that direct ST is viable for Catalan–English when supplemented with data augmentation and additional speech resources. However, our results indicate that, under the present data conditions, the end-to-end approach does not yet match the robustness of the cascade. One likely reason is that end-to-end ST must learn acoustic modeling, source-language normalization, and target-language generation jointly from substantially less paired speech-translation data than is available to the separate ASR and MT components.

5 Conclusion

This paper summarizes our submission to the CA–EN IWSLT 2026 evaluation campaign for low-resource speech translation. We explored two cascaded-based systems, combining different ASR with an MT model, and an end-to-end speech translation system.

The work reinforces the comparison between end-to-end and cascaded systems. In particular, the system based on Whisper-v3 ASR and NLLB MT achieved the best performance, highlighting the importance of strong multilingual pre-trained models. In comparison, the other cascaded system that used ConMamba as an ASR but maintained the NLLB model achieved worse results, which emphasizes the importance of a good-quality ASR. The end-to-end system Speech T5, despite slightly underperforming the cascaded systems, demonstrates the potential of this research branch.

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738, Dublin, Ireland. Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Carme Armentano-Oller, Montserrat Marimon, and Marta Villegas. 2024. Becoming a high-resource language in speech: The Catalan case in the Common Voice corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

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. 2021. XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Barcelona Supercomputing Center, Language Technologies Unit. 2025. SalamandraTA: A translation-specialised variant of salamandra. Model release, Hugging Face `BSC-LT/salamandrata-7b`.

William Chen, Brian Yan, Jiatong Shi, Yifan Peng, Soumi Maiti, and Shinji Watanabe. 2023. Improving

massively multilingual ASR with auxiliary CTC objectives. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

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 *Proc. IEEE Spoken Language Technology Workshop (SLT)*.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Iñigo Pikabea, Jorge Palomar, Júlia Falcão, Lucía Tormo, and 4 others. 2025. Salamandra technical report. *Preprint*, arXiv:2502.08489.

Baybars Külebi, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. 2022. ParlamentParla: A speech corpus of Catalan parliamentary sessions. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 125–130, Marseille, France. European Language Resources Association.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

NLLB Team, 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, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Alex Peiró-Lilja, Carme Armentano-Oller, José Giraldo, Wendy Elvira-García, Ignasi Esquerra, Rodolfo Zevallos, Cristina España-Bonet, Martí Llopart-Font, Baybars Külebi, and Mireia Farrús. 2026. Lafrescat: A studio-quality catalan multi-accent speech dataset for text-to-speech synthesis. *Computer Speech & Language*, page 101945.

Alex Peiró-Lilja, Martí Llopart-Font, Carme Armentano-Oller, José Giraldo, and Ignasi Esquerra. 2024. LaFresCat: A Catalan multi-accent speech dataset for text-to-speech. In *Proc. IberSPEECH 2024*.

Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, Wangyou Zhang, Yui Sudo, Muhammad Shakeel, Jee-weon Jung, Soumi Maiti, and Shinji Watanabe. 2023. Reproducing Whisper-style training using an open-source toolkit and publicly available data. In *Proc. ASRU*.

Projecte AINA. 2024. Common Voice benchmark catalan accents. Derived from Catalan Common Voice v17, stratified by accent (Balearic, Central, Northern, Northwestern, Valencian) and gender.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Changhan Wang, Anne Wu, and Juan Pino. 2021. CoVoST 2 and massively multilingual speech translation. In *Proc. Interspeech*.

Rodolfo Zevallos, Núria Bel, Guillermo Cámbara, Mireia Farrús, and Jordi Luque. 2022. Data Augmentation for Low-Resource Quechua ASR Improvement. In *Interspeech 2022*, pages 3518–3522.

Rodolfo Joel Zevallos Salazar, Martí Cortada García, Sarah Solito, Carlos Daniel Hernández Mena, Alexandre Peiró Lilja, and Francisco Javier Hernando Pericás. 2025. Assessing the performance and efficiency of mamba asr in low-resource scenarios. In *Interspeech 2025: Rotterdam, The Netherlands, 17-21 August, 2025*, pages 5198–5202. International Speech Communication Association (ISCA).