# A Zhuang Speech-to-Text Translation Method with Source Language-Aware Conditional Attention Mechanism


## Abstract

End-to-end speech-to-text translation aims to directly convert spoken input into text in the target language. As a minority language in China, Zhuang faces challenges such as data scarcity and limited technological support in the field of speech translation. To support research on Zhuang speech translation, we construct a recording platform that is used to compile speech-text parallel corpus. We introduce a Source Language-Aware Conditional Attention Mechanism (LACA), which incorporates source language information into the Transformer to bias attention toward Zhuang-specific linguistic features. Additionally, an Implicit Connectionist Temporal Classification Auxiliary Mechanism (ICAM) is employed during training to provide auxiliary supervision for alignment learning between speech and text representations. Experimental results demonstrate that our model achieves a BLEU score of 32.46 on Zhuang speech-to-text translation, outperforming the baseline by 4.12 BLEU points.


## 1 Introduction

In recent years, end-to-end models (Ren et al., 2020; Prabhavalkar et al., 2023; Chen et al., 2024; Zhang et al., 2024a) have become mainstream in speech translation, driving the development of large systems like GPT-4o (OpenAI, 2024), Gemini (Google, 2024), and SeamlessM4T (Communication, 2023).

Previous research on Zhuang text processing has investigated a range of technical approaches. Xian Wu (2024) employed word embeddings, relative position encoding, and a Transformer architecture for Chinese-Zhuang translation. Zhang et al. (2024b) proposed Dipmy++, incorporating a bilingual dictionary and 5,000 parallel sentences. Weiquan Zhang (2022) developed an F-BiLSTM-CNN-CRF model for named entity recognition, while Kui Ning (2019) designed a system for translating and proofreading Zhuang folk culture texts. Compared to text-based research, research on Zhuang speech remains relatively scarce. Huang et al. (2025) incorporated the SimAM attention mechanism into a Bidirectional Encoder Representations from Transformers model to perform speech classification across different Zhuang dialects. Jie Wang (2024) developed a Zhuang speech synthesis system supporting text normalization, phonetic conversion, and waveform generation.

However, research on end-to-end Zhuang speech-to-text translation (S2TT) remains limited. Most current end-to-end S2TT models are developed based on the Fairseq framework (Wang et al., 2020), such as s2t_transformer and s2t_conformer. Despite their success in high-resource settings, these models struggle to align speech and text representations under low-resource conditions due to limited supervision and alignment optimization. Moreover, the differences in speech modality between Zhuang and the languages used during pretraining limit their ability to generalize to unseen languages.

To address the above challenges, we designed a Source Language-Aware Conditional Attention mechanism (LACA) and introduced Connectionist Temporal Classification (CTC) as an implicit auxiliary supervision signal, referred to as the Implicit CTC Auxiliary Mechanism (ICAM), thereby enabling end-to-end S2TT for Zhuang. First, we introduced CTC loss during the training phase to assist the pre-trained speech encoder model wav2vec_small (Baevski et al., 2020) in learning feature alignment. Then, a one-dimensional convolutional adaptor (referred to as

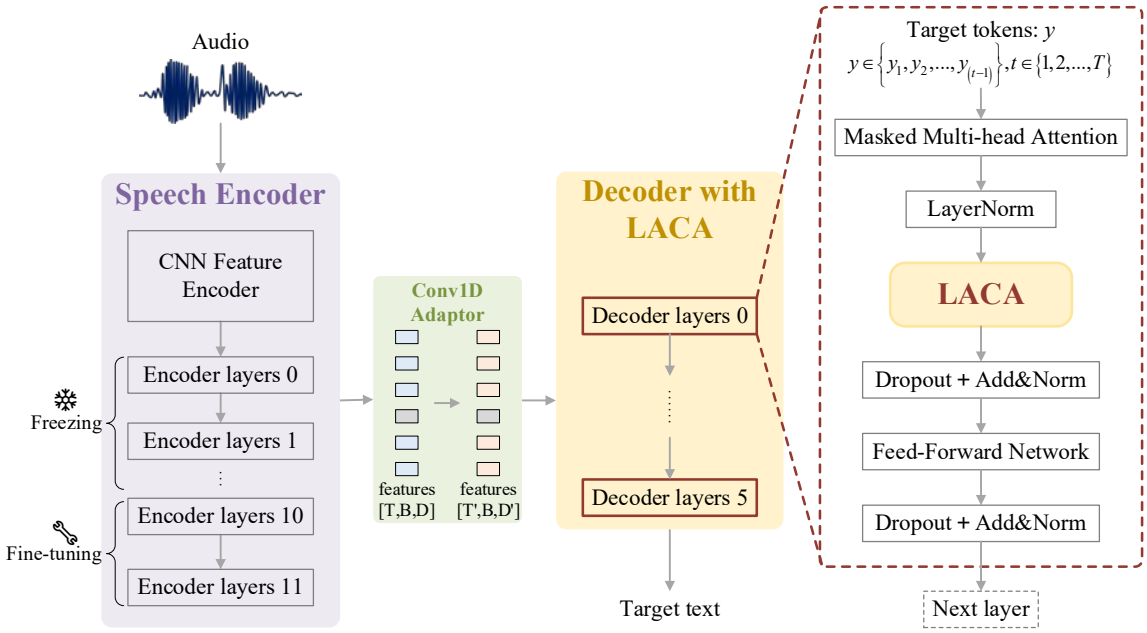


Figure 1: The overall architecture of our model. Raw speech is first encoded with the assistance of ICAM to enhance feature alignment. The resulting representations are then passed to the decoder, where LACA dynamically adjusts attention weights based on source language features.

Conv1D adaptor) was employed to transform the encoder output, serving as an interface to ensure compatibility with the decoder. Finally, LACA integrates language embeddings to enable the decoder to dynamically adjust attention weights based on the source language features during cross-attention. In addition, we developed a Zhuang translation and recording platform, and released several demos showcasing Zhuang S2TT[1].

## 2 Methods

To realize Zhuang S2TT, we designed LACA based on the xm_transformer [2] model and incorporated the ICAM. This section presents the overall framework as illustrated in Figure 1. The model consists of a speech encoder, a Conv1D adaptor, and a Transformer decoder.

### 2.1 Speech Encoder

Following prior approaches (Le et al., 2023; Xu et al., 2024; Yan et al., 2023), we incorporated CTC loss during training to facilitate alignment between speech and textual representations. The auxiliary CTC loss is computed as:

$$\mathcal{L}_{CTC} = CTC_{loss}\left(\hat{Y}, Y_{target}\right), \quad (1)$$

where $Encoder(X)$ denotes the output sequence of the speech encoder given the input audio $X$, and $\hat{Y} = Linear_{CTC}\left(Encoder(X)\right)$ is the predicted probability sequence obtained by applying a linear projection to the encoder output. $Y_{target}$ is the ground-truth sequence. Unlike conventional CTC-based models, this auxiliary CTC branch is excluded during inference and does not affect the main model architecture. To train the decoder, we adopt the standard cross-entropy loss between the predicted sequence $Y_{pred}$ and the reference sequence $Y_{target}$:

$$\mathcal{L}_{Decoder} = CrossEntropy\left(Y_{pred}, Y_{target}\right). \quad (2)$$

The overall training objective combines the decoder loss and the auxiliary CTC loss, weighted by $\lambda_{CTC}$:

$$\mathcal{L}_{total} = \mathcal{L}_{Decoder} + \lambda_{CTC} \cdot \mathcal{L}_{CTC}. \quad (3)$$

To mitigate overfitting in low-resource settings, as suggested by Yang et al. (2022; 2023), we fine-tuned only the last two layers of the encoder while keeping the first ten layers frozen.

### 2.2 Conv1D Adaptor

The Conv1D Adaptor is a lightweight feature transformation module that applies one-dimensional convolution to the output of the speech encoder, adjusting the feature dimension to

[1] http://510.english-gxu.net/

[2] https://github.com/facebookresearch/fairseq/tree/main/fairseq/models/speech_to_text

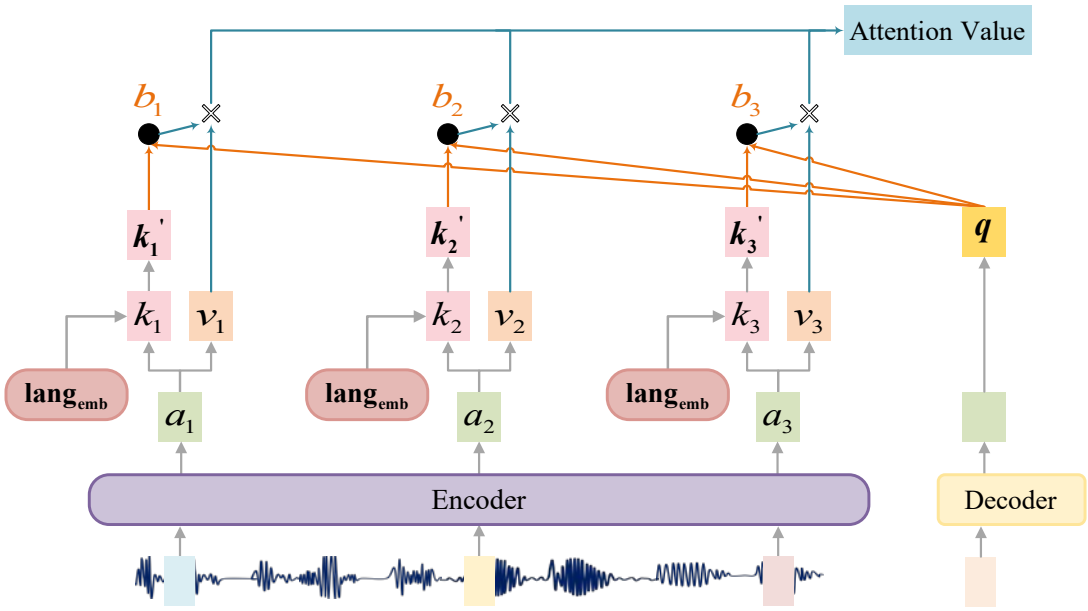


Figure 2: Source Language-Aware Conditional Attention Mechanism. $\text{lang}_{\text{emb}}$: The expanded source language embedding. ●: Inner product. ×: Weighted operation.

match the input requirements of the Transformer decoder. It also performs temporal downsampling via convolutional stride. To mitigate the high information density of speech features, a Gated Linear Unit (GLU) is applied following the convolution, facilitating the selective transmission of the most relevant temporal features to the decoder.

### 2.3 Decoder with LACA Mechanism

We proposed the LACA mechanism as illustrated in Figure 2. Let the decoder-derived query matrix be defined as:

$$Q = [q_1, q_2, \dots, q_n] \in R^{n \times d_k}, \quad (4)$$

where $n$ is the number of decoding steps and $d_k$ is the attention dimension. The encoder output is used to construct the key and value matrices:

$$K = [k_1, k_2, \dots, k_m] \in R^{m \times d_k}, \quad V = [v_1, v_2, \dots, v_m] \in R^{m \times d_k}, \quad (5)$$

where $m$ is the number of encoder steps. To inject language-specific information, we define a language embedding vector $\boldsymbol{e}_L \in R^{B \times d_k}$, where $B$ is the batch size. This vector is expanded along the temporal dimension to match the encoder output:

$$lang_{emb} = Expand(\boldsymbol{e}_L) \in R^{m \times B \times d_k}. \quad (6)$$

The expanded bias is then added element-wise to the original key to obtain the modified key:

$$K' = K + lang_{emb}. \quad (7)$$

Finally, the attention is calculated using the standard scaled dot-product formulation:

$$Attention(Q, K', V) = softmax\left(\frac{QK'^T}{\sqrt{d_k}}\right)V. \quad (8)$$

Compared with existing language-aware methods (Xu et al., 2023; Tian et al., 2022), the proposed LACA differs significantly in both the position and manner of language information injection. By introducing the language embedding as a bias term into the key vectors of the cross-attention module, LACA stably modulates the attention distribution. This lightweight design enhances training robustness under low-resource conditions and enables seamless extension to multilingual speech translation tasks.

## 3 Experimental Setup

This section introduces the experimental setup for the Zhuang end-to-end S2TT task, including the dataset, baseline system, evaluation metrics, and other implementation details.

### 3.1 Dataset

This paper focuses on the Dejing dialect of Zhuang (Lv, 2019). We developed a Zhuang speech recording platform[3] on a private server and collected speech from native speakers. All recordings were manually verified by linguistic experts. In addition, part of the dataset was sourced from news broadcasts on Jingxi TV, and

[3] http://510.english-gxu.net/record

The Data Collection, Recording, and Display Platform for the Chinese Language Resources Protection Project[4]. During data preprocessing, all audio recordings were standardized to a sampling rate of 16 kHz and converted to mono (Huang et al., 2025), followed by data augmentation. The dataset consists of both isolated vocabulary items and complete sentences, with a total duration of 19.23 hours. To evaluate the generalization capability of the model, we conducted validation on the Thai and English subsets of the Common Voice (Ardila et al., 2020).

### 3.2 Baseline systems and metrics

In this paper, s2t_transformer (Wang et al., 2020) is used as the baseline system. It is a Transformer based sequence-to-sequence model for S2TT. In the experiment, accuracy on the validation set during the training and BLEU score during inference are selected as evaluation metrics.

### 3.3 Implementation details

The model was trained on a single NVIDIA GeForce RTX 4090 using the Adam optimizer with a learning rate of 2e-4. The weight of the CTC loss was set to 0.3.

## 4 Results and Discussion

With a limited Zhuang speech-text parallel corpus, our model still achieved a BLEU score of 32.46 on the test set. In the following sections, we compare our model with the baseline, and analyze the contribution of each component.

### 4.1 Comparison with the Baseline System

We compared the performance of different models on the Zhuang dataset for S2TT task. As shown in Table 1, although the validation accuracy of our model was comparable to that of the baseline system, it achieved a BLEU score that was 4.12 points higher than the baseline.

To assess the generalization performance of the model on S2TT tasks, we conducted validation using the Common Voice dataset. As indicated in Table 2, the model achieved a validation accuracy of 86.89% and a BLEU score of 24.71 on the Thai dataset, demonstrating its cross-linguistic generalization capability. On the English dataset, since English and Zhuang belong to different language families and exhibit differences in speech feature patterns, our approach was sensitive to these differences, resulting in a lower validation accuracy. Nevertheless, the BLEU score of 25.05 suggests that the model-generated sentences exhibited local word order similarity to the reference translations and preserved a coherent syntactic structure at the sentence level.

| Model | accuracy(dev) | BLEU |
|---|---|---|
| s2t_transformer | 96.73% | 28.34 |
| s2t_conformer | 85.20% | 14.70 |
| Ours | 96.35% | 32.46 |

Table 1: S2TT Performance of different models on the Zhuang dataset.

| Dataset | accuracy(dev) | BLEU |
|---|---|---|
| Thai | 86.89% | 24.71 |
| English | 31.17% | 25.05 |
| Zhuang | 96.35% | 32.46 |

Table 2: S2TT Performance of our model on Zhuang, Thai, and English datasets.

### 4.2 Ablation study

To further evaluate the effectiveness of ICAM and LACA, we conducted ablation experiments, with the results summarized in Table 3. The integration of ICAM led to improvements in both validation accuracy and BLEU score. Upon incorporating LACA, despite a 0.42% drop in validation accuracy, the model attained a BLEU score of 32.46, marking a 2.42-point improvement and reflecting a favorable trade-off between accuracy and translation quality.

| Model | accuracy(dev) | BLEU |
|---|---|---|
| xm_transformer | 96.59% | 30.04 |
| xm_transformer+ICAM | 96.77% | 30.42 |
| xm_transformer+ICAM +LACA | 96.35% | 32.46 |

Table3: The results of the ablation study.

## 5 Conclusion

In this study, we developed a Zhuang translation and recording platform and introduced the LACA mechanism and the ICAM training strategy to enhance Zhuang speech-to-text translation (S2TT) task. Experimental results show that our method achieves accurate and effective performance in Zhuang S2TT. This work contributes to both low-resource speech translation research and the preservation of minority language heritage.

[4] https://zhongguoyuyan.cn/index

## Limitations

Although our method yields a satisfactory high BLEU score on the Zhuang dataset, performance on other languages remains moderate, indicating that the generalization capability still needs to be improved. In future work, we aim to improve performance in speech-to-text translation tasks across a wider range of languages. Furthermore, we plan to explore speech-to-speech translation both among Zhuang dialects and between Zhuang and other languages as an extension of this research. To facilitate these advancements, we will collect and construct a larger-scale Zhuang speech-text parallel corpus to mitigate the limitations posed by data scarcity.

## References


Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association. https://aclanthology.org/2020.lrec-1.520/.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems,* pages 12449–12460, Vancouver, Canada. Curran Associates, Inc. https://dl.acm.org/doi/abs/10.5555/3495724.3496768.

Xi Chen, Songyang Zhang, Qibing Bai, Kai Chen, and Satoshi Nakamura. 2024. LLaST: Improved End-to-end Speech Translation System Leveraged by Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6976–6987, Bangkok, Thailand. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.416.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. SeamlessM4T: Massively Multilingual & Multimodal Machine Translation. *arXiv preprint arXiv:2308.11596.*

Gemini Team Google: Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh et al. (1037 additional authors not shown). 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530.*

Min Huang, Xuejun Zhang and Wenkang Chen. 2025. Classification of Zhuang Dialect combined with

Bert and SimAM. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, Hyderabad, India. IEEE. doi: 10.1109/ICASSP49660.2025.10890693.

Phuong-Hang Le, Hongyu Gong, Changhan Wang, Juan Pino, Benjamin Lecouteux, Didier Schwab. 2023. Pre-training for speech translation: CTC meets optimal transport. In *Proceedings of the 40th International Conference on Machine Learning*, pages 18667–18685, Honolulu, Hawaii USA. Proceedings of Machine Learning Research. https://proceedings.mlr.press/v202/le23a.html.

Songsong Lv. 2019. Evolution of Proximal Demonstratives in Dejing Dialect of Zhuang Language——A Perspective Based on Linguistic Contact. *Journal of Guangxi Normal University (Philosophy and Social Sciences Edition)*, 55(04):108–118. doi: 10.16088/j.issn.1001-6597.2019.04.014.

Kui Ning. 2019. Zhuang Intelligent Translation System 1.0 Based on a Parallel Corpus. Master's thesis, Guangxi University, Nanning, China, April.

OpenAI: Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn et al. (318 additional authors not shown). 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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. doi: 10.1109/TASLP.2023.3328283.

Yi Ren, Jinglin Liu, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020. SimulSpeech: End-to-End Simultaneous Speech to Text Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3787–3796, Online. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.350.

Jinchuan Tian, Jianwei Yu, Chunlei Zhang, Yuexian Zou, Dong Yu. 2022. LAE: Language-Aware Encoder for Monolingual and Multilingual ASR. In *23th Annual Conference of the International Speech Communication Association*, pages 3178–3182, Incheon, Korea. International Speech Communication Association. doi: 10.21437/Interspeech.2022-923.

Xian Wu. 2024. Research on Chinese–Zhuang Machine Translation Based on Neural Networks. Master's thesis, Guangxi University for Nationalities, Nanning, China, March. doi: 10.27035/d.cnki.ggxmc.2024.000933.

Jie Wang. 2024. Research on Zhuang Language Speech Synthesis Technology Based on End-to-End. Master's thesis, Guangxi University for Nationalities, Nanning, China, March. doi: 10.27035/d.cnki.ggxmc.2024.000424.

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. doi: 10.18653/v1/2020.aacl-demo.6.

Chen Xu, Xiaoqian Liu, Erfeng He, Yuhao Zhang, Qianqian Dong, Tong Xiao. 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*, pages 12176–12180, Seoul, Korea. IEEE. doi: 10.1109/ICASSP48485.2024.10445849.

Haoran Xu, Jean Maillard, and Vedanuj Goswami. 2023. Language-Aware Multilingual Machine Translation with Self-Supervised Learning. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 526–539, Dubrovnik, Croatia. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.38.

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. doi: 10.18653/v1/2023.eacl-main.119.

Hao Yang, Jinming Zhao, Gholamreza Haffari, and Ehsan Shareghi. 2022. Self-supervised Rewiring of Pre-trained Speech Encoders:Towards Faster Fine-tuning with Less Labels in Speech Processing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1952–1959, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.141.

Hao Yang, Jinming Zhao, Gholamreza Haffari, and Ehsan Shareghi. 2023. Investigating pre-trained audio encoders in the low-resource condition. In *24th Annual Conference of the International Speech Communication Association*, pages 1498–1502, Dublin, Ireland. International Speech Communication Association. doi: 10.21437/Interspeech.2023-343.

Shaolei Zhang, Qingkai Fang, Shoutao Guo, Zhengrui Ma, Min Zhang, and Yang Feng. 2024a. StreamSpeech: Simultaneous Speech-to-Speech Translation with Multi-task Learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8964–8986, Bangkok, Thailand. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.485.

Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024b. Teaching Large Language Models an Unseen Language on the Fly. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8783–8800, Bangkok, Thailand. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.519.

Weiquan Zhang. 2022. Zhuang Named Entity Recognition Based on Deep Learning. Master's thesis, Guangxi Normal University, Nanning, China, June. doi: 10.27036/d.cnki.ggxsu.2022.001249.