Improving Speech Translation through Data Augmentation with Data in Similar Languages

Yu-Chien Lin*
University of Illinois Urbana-Champaign
Champaign, Illinois, USA
liniris2001@gmail.com

Chia-Hua Wu*
Academia Sinica
Taipei, Taiwan
maxwu@iis.sinica.edu.tw

Yu Tsao
Academia Sinica
Taipei, Taiwan
yu.tsao@citi.sinica.edu.tw

Hsin-Min Wang
Academia Sinica
Taipei, Taiwan
whm@iis.sinica.edu.tw

***Abstract*—Previous studies have shown that utilizing training data in similar languages can reduce translation errors in machine translation (MT). However, its potential in speech translation (ST) remains underexplored. In this study, we propose a novel method to enhance cross-modal training of ST models by incorporating external ST data from the same language group without explicit language tagging. We evaluate our method based on the ConST model on the CoVoST 2 Es-En dataset, and experimental results show that the performance improves as more ST data in similar languages is introduced. The model trained with 400 hours of external ST data in similar languages improved the BLEU score by 11.42% and 10.74% compared to the baseline model with no external data and the model trained with 1 million external Es-En MT data. These results demonstrate that our approach can improve ST performance, especially under low-resource conditions.**

***Index Terms*—Speech translation, Low-Resource Language, Language group, multi-modality**

I. Introduction

Speech Translation (ST) is a cross-modal task in which a model converts speech input in a source language into text in a target language. Recent advances in end-to-end ST have achieved performance comparable to traditional cascaded systems combining automatic speech recognition (ASR) and machine translation (MT). Additionally, these end-to-end approaches have the advantage of reducing latency and error propagation [1]–[6].

One of the main challenges in training end-to-end ST models is the scarcity of data. For example, there are only a few hundred hours of data in the MuST-C corpus [7], or the data per language pair is unbalanced, as in the CoVoST 2 corpus [8], where the data range from a few to hundreds of hours per language pair. These amounts are much smaller than the available data for ASR [9]–[14] or MT [15]–[17]. How to better utilize the limited labeled ST data and other parallel MT corpora is a promising research area.

One aspect of the research explores the use of pre-trained ASR or MT models to initialize the parameters of the ST model. Another aspect is the use of ASR or MT data for multi-task assistance in ST model training [18]–[21]. There are also some studies on data augmentation for ST, such as perturbation-based methods SpecAug [22], DropDim [23], and SkinAugment [24], synthetic methods text-to-speech (TTS) [25] and back-translation [26], and STR [27]. However, these previous studies typically only performed data augmentation on a single modality and have not yet been validated on strong baselines.

Furthermore, previous research has demonstrated that utilizing training data in similar languages can significantly improve MT performance for target language pairs, especially for low-resource language pairs [28]–[32]. It would be highly beneficial if we could effectively leverage data from relevant languages in ST tasks, since ST training data is much less than what is available for MT. However, to the best of our knowledge, applying data from similar languages to enhance ST has not been extensively explored.

Therefore, in this paper, we explore data augmentation using ST training dada in similar languages when training ST models. Many current state-of-the-art (SOTA) cross-modal training methods for ST focus on modality alignment [33]–[37], and most studies are performed on the MuST-C dataset (mainly English-other language ST). However, while there are many public tools available for cross-modal alignment or word alignment between speech and text in English, adapting these tools to cross-modal tasks in other languages, especially low-resource ones, can be quite challenging. To avoid these limitations and ensure that our approach works well in low-resource settings, we choose the ConST model [38] as the backbone ST model and use the CoVoST 2 dataset (Spanish-English ST) as our experimental setting. ConST is built on a multi-task training framework employing a cross-modal contrastive learning approach. We propose data augmentation with ST training dada in similar languages to enhance cross-modal training within its multi-task training framework.

The contributions of this study are as follows:

- We explore the impact of different modalities and languages on data augmentation. Experimental results show that our method achieves good performance on the Es-En ST task on the CoVoST 2 dataset.
- We see great potential in using ST data of similar languages for data augmentation in ST model training compared to data augmentation using external MT data of the same language pair.
- Our method is expected to generalize to other low-resource ST tasks.

* These authors contributed equally to this work.

ISBN: 978-9-46-459362-4

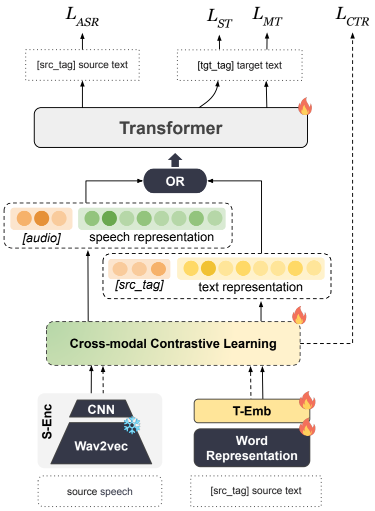


Fig. 1: Architecture diagram of cross-modal multi-task training in speech translation.

II. Method

Our work mainly extends the ConST model [38]. As shown in Figure 1, the framework comprises four sub-modules: a *speech encoder* (S-Enc), a *word embedding layer* (T-Emb), a *Transformer encoder*, and a *Transformer decoder*. The architecture provides a common framework for three tasks, including ST, MT, and ASR. Input can be fed into the speech encoder (speech input) or the word embedding layer (text input) along with a leading language ID tag, allowing seamless adaptation to ST, MT, and ASR tasks, all of which are involved in the model training process. Language ID tags are also used as beginning-of-sequence (BOS) tokens for the Transformer decoder to generate the corresponding text. For example, if the input is Spanish speech corresponding to "*Fue de origen belga.*", the ASR task uses [es] as the BOS token and expects to output "*[es] Fue de origen belga.*". To translate the input Spanish speech into English text, [en] is used as the BOS token, and the decoded output "*[en] He was from Belgium.*" is expected. In the inference stage, the model performs the ST task, which takes speech in the source language as input and produces text in the target language as output.

The above multi-task training requires an ST corpus consisting of *speech-transcription-translation* triples, denoted as $\mathcal{D} = \{(\mathbf{s}, \mathbf{x}, \mathbf{y})\}$, where $\mathbf{s}$ represents speech in the source language, $\mathbf{x}$ is the corresponding transcription, and $\mathbf{y}$ is the translation in the target language. In multi-task training, ST is the primary task, and ASR and MT are secondary tasks. Given the training sets $\mathcal{D}_{ST} = \{(\mathbf{s}, \mathbf{y})\}$, $\mathcal{D}_{ASR} = \{(\mathbf{s}, \mathbf{x})\}$, and $\mathcal{D}_{MT} = \{(\mathbf{x}, \mathbf{y})\}$, the loss functions of these three tasks are defined as follows:

$$\mathcal{L}_{ST} = -\sum_{(\mathbf{s},\mathbf{y})} \sum_{i=1}^{|\mathbf{y}|} \log p(y_i | y_{1:i-1}, \mathbf{s}), \tag{1}$$

$$\mathcal{L}_{ASR} = -\sum_{(\mathbf{s},\mathbf{x})} \sum_{i=1}^{|\mathbf{x}|} \log p(x_i | x_{1:i-1}, \mathbf{s}), \tag{2}$$

$$\mathcal{L}_{MT} = -\sum_{(\mathbf{x},\mathbf{y})} \sum_{i=1}^{|\mathbf{y}|} \log p(y_i | y_{1:i-1}, \mathbf{x}). \tag{3}$$

Many studies [33]–[35], [38]–[41] have shown that during the model training stage, bringing the representation of speech and its transcripts closer while pushing the representations of speech and other speech transcripts away can improve ST performance. We consider each *speech-transcript* pair $(\mathbf{s}, \mathbf{x})$ in a batch as a positive example and the pairs of $\mathbf{s}$ with the remaining $N-1$ transcripts in the batch $\{(\mathbf{s}, \mathbf{x}_j)\}_{j=1}^{N-1}$ as the corresponding negative examples. The multi-class N-pair contrastive loss [42] is defined as:

$$\mathcal{L}_{\text{CTR}} = -\sum_{(\mathbf{s},\mathbf{x})} \log \frac{\exp(\text{sim}(u(\mathbf{s}), v(\mathbf{x}))/\tau)}{\sum_{\mathbf{x}' \in \mathcal{A}} \exp(\text{sim}(u(\mathbf{s}), v(\mathbf{x}'))/\tau)}, \tag{4}$$

where $\mathcal{A} = \{\mathbf{x}\} \cup \{\mathbf{x}_j\}_{j=1}^{N-1}$, τ is the temperature parameter, $\text{sim}(u, v)$ is the cosine similarity function, and u and v are mean functions of speech and transcript representations calculated as $u(\mathbf{s}) = \text{MeanPool}(S\text{-Enc}(\mathbf{s}))$ and $v(\mathbf{x}) = \text{MeanPool}(\text{T-Emb}(\mathbf{x}))$.

The overall training loss function combines this contrastive loss with the cross-entropy losses of ST, MT, and ASR, and is expressed as follows:

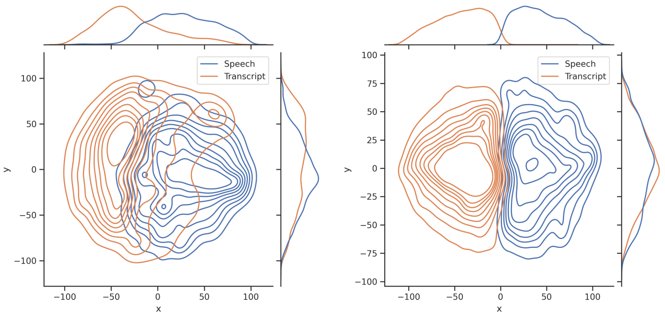
$$\mathcal{L} = \mathcal{L}_{\text{ST}} + \mathcal{L}_{\text{ASR}} + \mathcal{L}_{\text{MT}} + \lambda \mathcal{L}_{\text{CTR}} \tag{5}$$

where λ is a hyper-parameter that controls the contribution of the contrastive loss.

A. Data Augmentation with Data in Similar Languages

In natural language processing (NLP) and related fields, data augmentation that takes into account linguistic similarities and differences across languages can significantly improve model performance [43]. An effective strategy is to classify languages into groups or families based on common linguistic features such as grammar, syntax, and vocabulary. Common language families include Indo-European, Sino-Tibetan, and Afro-Asiatic, each of which contains multiple languages with varying degrees of similarity.

Within the framework of our Spanish-English ST experiments on the CoVoST 2 dataset [8], we incorporate ST data from other languages into the original training data as a form of data augmentation. Specifically, we collected data for the French-English (264 hours), Italian-English (44 hours), Portuguese-English (10 hours), and Catalan-English (136 hours) ST tasks, for a total of 454 hours, called external Language Group Speech Translation data (**external LG ST data**). French, Italian, Portuguese, and Catalan belong to the same Indo-European/Italic branch. We hypothesize that their linguistic proximity to Spanish might have a positive impact on the performance of the Spanish-English ST model.



(a) Unified tagging (b) Language-specific tagging

Fig. 2: Bivariate KDE contour plots of the representation of speech and transcripts under different language tagging strategies (unified vs. language-specific) used in data augmentation using external LG ST data. Blue lines represent speech embeddings, and orange lines represent transcript embeddings. The visualization is derived from the output of the Transformer encoder on the Es-En test set, reduced to 2D using t-SNE.

To compare "data augmentation using external ST data in similar languages" and "data augmentation using external MT data", we leveraged the OPUS-100 corpus [44] as an external source of MT data to enable cross-task data augmentation. One million Spanish-English sentence pairs selected from the OPUS-100 corpus are used as external Spanish-English Machine Translation data (**external Es-En MT data**).

Note that if using external Es-En MT data in data augmentation, the additional input is Spanish text, and [src_tag] is always [es]. However, when using external LG ST data in data augmentation, the language ID of the additional input can be other than [es]. In our experiments, if [es] is used uniformly as [src_tag] for input in different languages, it is called "unified", and if different language IDs are used as [src_tag] for input in different languages, it is called "language-specific". A unified audio tag [audio] is used for input speech in different languages.

III. Experimental Setup

A. Dataset

The CoVoST 2 dataset [8] is a multilingual ST corpus containing translations from 21 languages to English and from English to 15 languages. Our study focuses specifically on the non-English-English (X-En) translation direction. The dataset covers 21 languages in 9 distinct language families. For the purpose of our analysis, we selected the Spanish-English (Es-En) ST task as a case study. The dataset for this task includes 113 hours of training data, 22 hours of validation data, and 23 hours of testing data. As described in Section II-A, we performed data augmentation in model training using training data for the French (Fr)-, Italian (It)-, Portuguese (Pt)- and Catalan (Ca)-En ST tasks.

B. Baseline Systems

This study uses the ConST model [38] as the backbone model and one of the baseline models. It uses cross-modal training and has approximately 150 million parameters. The baseline models also include three models from the CoVoST 2 benchmark [8]. The ASR model and the ST model share the same Transformer encoder-decoder architecture [46], where there are 12 encoder layers and 6 decoder layers. A convolutional downsampler is applied to reduce the length of speech inputs by $3/4$ before they are fed into the encoder. The MT model uses a Transformer *base* architecture with 3 encoder and 3 decoder layers, a dropout rate of 0.3, and shared embeddings for the encoder/decoder inputs and decoder outputs

Transformer-ST is a Transformer-based ST model trained from scratch without any pre-training. The Transformer-ST+ASR pre-trained model uses ASR pre-training before fine-tuning for the ST task. Transformer-ST model parameters are about 31M. Cascade ST is a model that sequentially connects two independent components: ASR and MT. Revisit ST [45] represents an enhanced and optimized version of the Transformer-based ST model with 51 million parameters.

C. Experiment Details

We conducted experiments on the ConST model[1], which is built on the Fairseq framework [47]. Performance is measured by case-sensitive, de-tokenized BLEU scores calculated using sacreBLEU [48].

When training the model without using external LG ST data, we applied SentencePiece [49] with a shared vocabulary of 10,000 tokens to process bilingual text, following the approach in [38]. In a data augmentation setting using external LG ST data, we switched to SentencePiece with byte-pair encoding (BPE) [50], expanding the vocabulary size to 30,000 tokens. In all experiments, λ in Eq. 5 was set to 1.0.

The Wav2vec 2.0 model used in S-Enc is pre-trained on the Librispeech dataset [9] without any downstream fine-tuning. Following the Wav2vec 2.0 module, the two convolutional neural network (CNN) layers were configured with a kernel size of 5, a stride size of 2, and a hidden size of 512. The Transformer component used a base configuration with 6 layers each for the encoder and decoder, a hidden size of 512, 8 attention heads, and 2048 feed-forward network (FFN) hidden states. Pre-layer normalization was applied to ensure stable training. Through these configurations, the entire model has approximately 150 million parameters.

IV. Experimental Results

The experimental results in Table I show the performance of various models and data augmentation strategies on the CoVoST 2 Es-En test set. The baseline ConST model (A1) has a BLEU score of 26.28, serving as a reference to compare different models. Among them, the Transformer-ST model trained without any auxiliary data showed limited performance with a BLEU score of 12.00, while the inclusion of ASR

[1] https://github.com/ReneeYe/ConST.git

TABLE I: BLEU scores of different models on the CoVoST-2 Es-En test set. *: results from [8] or [45].

ID	Model	# MT data	# ST data	BLEU (%)	
		augmented	augmented	Score ↑	Gain ↑
	Transformer-ST from scratch [8]*	0	0	12.00	-
	Transformer-ST + ASR pre-trained [8]*	0	0	23.00	-
	Cascaded ST [8]*	0	0	27.40	-
	Revisit ST [45]*	0	0	15.70	-
A1	Baseline (ConST) [38]	0	0	26.28	-
A2	w/ external Es-En MT data	1 million	0	26.44	(+0.16)
A3	w/ external LG ST data (Part of Fr-En)	0	10 hours	24.55	(−1.73)
A4	w/ external LG ST data (Part of Fr-En)	0	50 hours	25.77	(−0.51)
A5	w/ external LG ST data (Part of Fr-En)	0	100 hours	26.71	(+0.43)
A6	w/ external LG ST data (Fr-En)	0	264 hours	27.31	(+1.03)
A7	w/ external LG ST data (Fr-En + Ca-En)	0	400 hours	**29.28**	(+3.00)
A8	w/ external LG ST data (Fr-En + Ca-En + It-En)	0	444 hours	28.52	(+2.24)
A9	w/ external LG ST data (all)	0	454 hours	29.02	(+2.74)

TABLE II: Performance comparison of using a unified token or language-specific tokens in data augmentation using ST data in similar languages.

Model	# ST data augmented	[src_tag]	BLEU (%)↑
Baseline	0 hours	unified	26.28
w/ external LG ST data	454 hours	unified	**29.02**
w/ external LG ST data	454 hours	language-specific	28.30

pre-training significantly improved the score to 23.00. The Cascaded ST model has the highest score of 27.40, exceeding the baseline ConST model we adopted. The Revisit ST model performed mediocrely, with a score of 15.70.

When focusing on comparing different data augmentation methods, we first see that adding the Es-En MT data (A2) results in a slight performance improvement of 0.16, reaching a BLEU score of 26.44. However, leveraging external LG ST data (A3 to A9) shows a more substantial effect. With only 10 hours of Fr-En ST data, the performance drops to 24.55, but the BLEU score continues to improve as the amount of Fr-En ST data increases. The best performance is achieved with 400 hours of external LG ST data (A7), reaching a BLEU score of 29.28, a significant improvement of 3.00 over the baseline. Beyond 400 hours, the incremental benefit from additional data becomes smaller. The main reason is that the external LG ST data is not Es-En ST data after all, and too much may obscure the real Es-En ST data during model training. Overall, these results highlight the effectiveness of data augmentation using external LG ST data for improving ST performance and demonstrate that data augmentation using external LG ST data is not "more is better".

V. Analysis

A. Do We Really Need Separate Tags for Each Language?

As shown in Table II, when external LG ST data are used for data augmentation, the BLEU score using a "unified" tag is 29.02, which is slightly better than the score using "language-specific" tags (28.30). Figure 2 shows bivariate kernel density estimation (KDE) contour plots of the representation of speech and transcripts under different language tagging strategies (unified vs. language-specific) used in data augmentation using external LG ST data. As shown in Figure 2a), unified tagging leads to better cross-modal fusion, resulting in more coherent and consistent representation between data modalities. In contrast, Figure 2b) shows that language-specific tagging yields scattered and fragmented distributions, indicating weaker cross-modal alignment. The better cross-modal consistency of unified tagging in Figure 2 aligns with its better BLEU score in Table II.

VI. Conclusion

The end-to-end ST model faces major challenges due to data scarcity. Previous studies mainly addressed addressed this problem through ASR or MT pre-training, or using MT data for data augmentation. This study introduces a novel approach to improve speech translation performance by leveraging ST data in similar languages, reducing reliance on source language-specific data, especially in low-resource scenarios. The method shows strong adaptability in environments where source speech data are scarce or transcription costs are high.

In our experiments, the baseline model achieved a BLEU score of 26.28 on the CoVoST 2 Es-En test set. The commonly used MT-based data augmentation method only slightly improved the BLEU score to 26.44. In comparison, our approach, augmented directly with speech translation data in similar languages, significantly boosted the BLEU score by 3 points to 29.28. These results highlight the effectiveness of our approach in enhancing speech translation performance, especially for applications in resource-constrained languages.

Future work will explore automatic data selection techniques to further minimize reliance on manual annotation and prior language knowledge. This enhancement aims to strengthen the robustness of our approach in different language scenarios and expand its applicability.

VII. Acknowledgements

This work was partially supported by the National Science and Technology Council of Taiwan under Grant No. NSTC 112-2221-E-001-009-MY3.

References

[1] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn, "An attentional model for speech translation without transcription," in *Proc. NAACL-HLT*, 2016.

[2] A. Bérard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," in *Proc. NeurIPS Workshop on End-to-end Learning for Speech and Audio Processing*, 2016.

[3] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly translate foreign speech," in *Proc. Interspeech*, 2017.

[4] L. C. Vila, C. Escolano, J. A. Fonollosa, and M. R. Costa-Jussa, "End-to-end speech translation with the transformer," in *Proc. IberSPEECH*, 2018.

[5] H. Inaguma, K. Duh, T. Kawahara, and S. Watanabe, "Multilingual end-to-end speech translation," in *Proc. ASRU*, 2019.

[6] C. Han, M. Wang, H. Ji, and L. Li, "Learning shared semantic space for speech-to-text translation," in *Proc. ACL*, 2021.

[7] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "MuST-C: A multilingual speech translation corpus," in *Proc. NAACL*, 2019.

[8] C. Wang, A. Wu, J. Gu, and J. Pino, "CoVoST 2 and massively multilingual speech translation." in *Proc. Interspeech*, 2021.

[9] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015.

[10] Y. Yang, Z. Song, J. Zhuo, M. Cui, J. Li, B. Yang, Y. Du, Z. Ma, X. Liu, Z. Wang *et al.*, "GigaSpeech 2: An evolving, large-scale and multi-domain ASR corpus for low-resource languages with automated crawling, transcription and refinement," *arXiv preprint arXiv:2406.11546*, 2024.

[11] S. Li, Y. You, X. Wang, Z. Tian, K. Ding, and G. Wan, "MSR-86K: An evolving, multilingual corpus with 86,300 hours of transcribed audio for speech recognition research," in *Proc. Interspeech*, 2024.

[12] Y. Lin, M. Cheng, F. Zhang, Y. Gao, S. Zhang, and M. Li, "VoxBlink2: A 100k+ speaker recognition corpus and the open-set speaker-identification benchmark," in *Proc. Interspeech*, 2024.

[13] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng *et al.*, "WenetSpeech: A 10000+ hours multi-domain mandarin corpus for speech recognition," in *Proc. ICASSP*, 2022.

[14] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proc. LREC*, 2020.

[15] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," in *Proc. Lrec*, 2012.

[16] M. Bañón, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. Forcada, A. Kamran, F. Kirefu, P. Koehn *et al.*, "ParaCrawl: Web-scale acquisition of parallel corpora," in *Proc. ACL*, 2020.

[17] P. Koehn, H. Khayrallah, K. Heafield, and M. L. Forcada, "Findings of the WMT 2018 shared task on parallel corpus filtering," in *Proc. WMT18*, 2018.

[18] Q. Dong, M. Wang, H. Zhou, S. Xu, B. Xu, and L. Li, "Consecutive decoding for speech-to-text translation," in *Proc. AAAI*, 2021.

[19] C. Wang, Y. Wu, S. Liu, M. Zhou, and Z. Yang, "Curriculum pre-training for end-to-end speech translation," in *Proc. ACL*, 2020.

[20] Y. Tang, H. Gong, N. Dong, C. Wang, W.-N. Hsu, J. Gu, A. Baevski, X. Li, A. Mohamed, M. Auli *et al.*, "Unified speech-text pre-training for speech translation and recognition," in *Proc. ACL*, 2022.

[21] C. Xu, R. Ye, Q. Dong, C. Zhao, T. Ko, M. Wang, T. Xiao, and J. Zhu, "Recent advances in direct speech-to-text translation," in *Proc. IJCAI*, 2023.

[22] P. Bahar, A. Zeyer, R. Schlüter, and H. Ney, "On using specaugment for end-to-end speech translation," in *Proc. IWSLT*, 2019.

[23] H. Zhang, D. Qu, K. Shao, and X. Yang, "DropDim: A regularization method for transformer networks," *IEEE Signal Processing Letters*, vol. 29, pp. 474–478, 2022.

[24] A. D. McCarthy, L. Puzon, and J. Pino, "SkinAugment: Auto-encoding speaker conversions for automatic speech translation," in *Proc. ICASSP*, 2020.

[25] J. Zhao, G. Haffari, and E. Shareghi, "Generating synthetic speech from SpokenVocab for speech translation," in *Proc. EACL*, 2023.

[26] B. B. Odoom, N. Robinson, E. Rippeth, L. Tavarez-Arce, K. Murray, M. Wiesner, P. McNamee, P. Koehn, and K. Duh, "Can synthetic speech improve end-to-end conversational speech translation?" in *Proc. AMTA*, 2024.

[27] T. K. Lam, S. Schamoni, and S. Riezler, "Sample, Translate, Recombine: Leveraging audio alignments for data augmentation in end-to-end speech translation," in *Proc. ACL*, 2022.

[28] F. Philippy, S. Guo, and S. Haddadan, "Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review," in *Proc. ACL*, 2023.

[29] R. Oji and J. Kunz, "How to tune a multilingual encoder model for germanic languages: A study of peft, full fine-tuning, and language adapters," in *Proc. NoDaLiDa and Baltic-HLT*, 2025.

[30] A. Chronopoulou, D. Stojanovski, and A. Fraser, "Language-family adapters for low-resource multilingual neural machine translation," in *Proc. LoResMT*, 2023.

[31] X. Tan, J. Chen, D. He, Y. Xia, T. Qin, and T.-Y. Liu, "Multilingual neural machine translation with language clustering," in *Proc. EMNLP*, 2019.

[32] S. Bala Das, D. Panda, T. Kumar Mishra, B. Kr. Patra, and A. Ekbal, "Multilingual neural machine translation for Indic to Indic languages," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 5, pp. 1–32, 2024.

[33] S. Ouyang, R. Ye, and L. Li, "WACO: Word-aligned contrastive learning for speech translation," in *Proc. ACL*, 2022.

[34] Y. Zhou, Q. Fang, and Y. Feng, "CMOT: Cross-modal mixup via optimal transport for speech translation," in *Proc. ACL*, 2023.

[35] Y. Zhang, K. Kou, B. Li, C. Xu, C. Zhang, T. Xiao, and J. Zhu, "Soft alignment of modality space for end-to-end speech translation," in *Proc. ICASSP*, 2024.

[36] W. Wang, S. Ren, Y. Qian, S. Liu, Y. Shi, Y. Qian, and M. Zeng, "Optimizing alignment of speech and language latent spaces for end-to-end speech recognition and understanding," in *Proc. ICASSP*, 2022.

[37] Z. Zhang, L. Zhou, J. Ao, S. Liu, L. Dai, J. Li, and F. Wei, "SpeechUT: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training," in *Proc. EMNLP*, 2022.

[38] R. Ye, M. Wang, and L. Li, "Cross-modal contrastive learning for speech translation," in *Proc. NAACL*, 2022.

[39] S. R. Indurthi, S. Chollampatt, R. Agrawal, and M. Turchi, "CLAD-ST: Contrastive learning with adversarial data for robust speech translation," in *Proc. EMNLP*, 2023.

[40] H. Zhang, N. Si, Y. Chen, W. Zhang, X. Yang, D. Qu, and W.-Q. Zhang, "Improving speech translation by cross-modal multi-grained contrastive learning," *ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1075–1086, 2023.

[41] Q. Fang and Y. Feng, "Understanding and bridging the modality gap for speech translation," in *Proc. ACL*, 2023.

[42] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Proc. NeurIPS*, 2016.

[43] B. Li, Y. Hou, and W. Che, "Data augmentation approaches in natural language processing: A survey," *AI Open*, vol. 3, pp. 71–90, 2022.

[44] B. Zhang, P. Williams, I. Titov, and R. Sennrich, "Improving massively multilingual neural machine translation and zero-shot translation," in *Proc. ACL*, 2020.

[45] B. Zhang, B. Haddow, and R. Sennrich, "Revisiting end-to-end speech-to-text translation from scratch," in *Proc. ICML*, 2022.

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017.

[47] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proc. NAACL*, 2019.

[48] M. Post, "A call for clarity in reporting BLEU scores," in *Proc. WMT*, 2018.

[49] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proc. EMNLP*, 2018.

[50] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. ACL*, 2016.