A Domain-Specific Multilingual Speech Translation Corpus via Simultaneous Interpretation

Seunghee Han[†]
Learning Sciences Research Institute
Seoul National University
Seoul, Republic of Korea
seunghee.han@snu.ac.kr

Gary Geunbae Lee
Department of Computer Science and Engineering
Postech
Pohang, Republic of Korea
gblee@postech.ac.kr

Hyung Soon Kim
Department of Electronics Engineering
Pusan National University
Busan, Republic of Korea
kimhs@pusan.ac.kr

Sunhee Kim
Department of French Language Education
Seoul National University
Seoul, Republic of Korea
sunhkim@snu.ac.kr

Minhwa Chung[*]
Department of Linguistics
Seoul National University
Seoul, Republic of Korea
mchung@snu.ac.kr

ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10887990

Abstract—**This paper presents a novel multilingual speech translation corpus for complex, domain-specific content in Korean, English, Spanish, and Japanese. The corpus contains 4,000 hours of parallel speech, including 1,000 hours of Korean audio with simultaneous sight interpretations in the other three languages by 294 professionals (242 interpreters and 52 Korean voice actors). It also includes transcriptions, translations, and annotations for all languages. The Dewey Decimal Classification was adapted to balance knowledge representation, and speech tasks were conducted in a controlled studio environment to ensure data consistency. Translation, transcription, and annotation workflows were managed through a custom-built platform. The corpus captures nuanced contexts, cultural sensitivities, and domain-specific terminology, addressing linguistic challenges like structural differences between SOV (Korean, Japanese) and SVO languages (English, Spanish). Preliminary evaluations indicate its potential to enhance end-to-end speech translation models, support cross-lingual transfer learning, and tackle real-time translation issues.**

Index Terms—**Speech-to-speech translation, Speech recognition, Speech synthesis, Speech processing resources**

I. Introduction

Despite advances in machine translation, both large language models (LLMs) [1], [2] and neural machine translation (NMT) systems [3] struggle with limited-resource languages due to training data biases [4], [5]. LLMs face contextual drift as sentence complexity increases, losing coherence in long-range dependencies and producing incoherent or simplified translations. NMT models, while effective for everyday conversations, often fail to capture domain-specific content, particularly technical terminology and cultural nuances.

The demand for simultaneous speech translation (SST) systems [6], [7] is growing, particularly in specialized domains like academic and technical lectures. Advances in multilingual speech processing have spurred interest in developing systems capable of handling domain-specific language and complex contexts, such as retrieval-augmented generation (RAG) [8]. Enhancing automatic speech recognition (ASR) to effectively handle L2 speech and code-switching thus remains a critical challenge.

While significant progress in ASR and speech translation (ST) has been achieved through unsupervised pre-training [9] and semi-supervised learning [10] using unlabeled English speech datasets [11], extending these methods to limited-resource languages is still challenging due to insufficient data. The reliance on English-centric corpora [12], [13] emphasizes the need for high-quality labeled data in other languages to support robust language model training.

This work targets limited- or medium-resource languages—those with more resources than traditional low-resource languages but still insufficient for training state-of-the-art models. To bridge this gap, we present a 4,000-hour multilingual speech translation corpus created through "simultaneous sight interpretation" by professional interpreters. The corpus includes translations from Korean (source language) into English, Spanish, and Japanese. It aims to advance speech-to-speech translation systems and ASR models for limited-resource languages while providing validated benchmarks for accuracy, alignment, and future research in speech translation and cross-lingual processing.

II. Challenges and Opportunities in Multilingual Speech Translation Corpora

LibriSpeech [12] and Multilingual LibriSpeech (MLS) [14] are commonly used for speech translation, recognition, and synthesis, where audio is aligned with text using tools like Aeneas. LibriS2S [15] expands this concept by manually aligning English and German audiobooks from Librivox, creating parallel datasets of the same works. However, LibriS2S does

† Primary Author, * Corresponding Author

This study was supported by the AI Data Project from the National Information Society Agency (NIA), Republic of Korea. The resulting multilingual speech translation corpus is publicly available through AI Hub (https://aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=data&dataSetSn=71524)

TABLE I
Overview of the Corpus Structure

			Korean	**English**	**Spanish**	**Japanese**
Demographics	# of Speaker		52	85	91	66
	Age	20s	47.48%	79.40%	71.77%	43.51%
		30s	13.24%	18.17%	28.23%	22.80%
		+40s	39.28%	2.43%	-	33.69%
	Gender	M	44.95%	16.85%	14.85%	11.94%
		F	55.05%	83.15%	85.15%	88.06%
	Domain	Humanities /Literature	40.47%	35.87%	35.58%	39.45%
		Self-Development /Practical	22.30%	22.81%	23.90%	22.76%
		Science/IT	18.49%	20.75%	21.06%	19.41%
		Social Sciences	18.74%	20.58%	19.45%	18.39%
Speech Characteristics	Duration (hours)		1,000.5	1,016.1	1,062.0	1,028.7
	# of Sentences		435,961	413,619	379,005	403,607
	# of Tokens/Characters		21,806,403	7,545,050	7,619,161	18,118,747
	Avg. Tokens (EN, SP)/ Avg, Characters (KR, JP)		50.02	18.24	20.10	44.89
	Duration/Speaker (hours)		19.24	11.95	11.67	15.59
	Avg. Duration of Speaker/Book (hours)		0.20	0.19	0.21	0.17
Annotation			Textnumber, transcription, timestamps (start time, end time)			
Metadata			Speaker id, gender, age, domain, topic (keyword), script file name (book title), script number (language code_document id_section id, language (ko, en, es, jp)			
Recording Protocols			Format (WAV), Channels (1 channel), Sampling rate (16-44kHz)			

not provide sentence-level alignment, limiting its usefulness for sentence-aligned translation tasks. While these datasets offer high-quality speech data, LibriSpeech is based on public domain texts, primarily from classic literature, introducing data bias regarding vocabulary and style. This bias limits the ability of models trained on LibriSpeech to generalize to modern or domain-specific tasks, such as technical fields like law and medicine.

Other multilingual datasets like VoxPopuli [16], CVSS [17], and SpeechMatrix [18] offer alternative solutions but also present limitations. VoxPopuli, sourced from European Parliament speeches, includes 400,000 hours of unlabeled speech, 1,800 hours of transcribed speech, and 17,300 hours of simultaneous interpretation in 23 European languages. It captures real-world political speech with natural interruptions, making it valuable for speech-to-text, speech-to-speech translation, and simultaneous interpretation research. However, its focus on European languages limits its applicability to non-European languages.

CVSS, derived from Common Voice and CoVoST 2, offers 1,900 hours of synthetic speech-to-speech translation across 21 languages, focusing on preserving speaker identity for real-time translation. However, the reliance on synthetic data lacks the natural variability needed to handle spontaneous speech patterns, reducing its robustness in specific contexts.

SpeechMatrix provides 418,000 hours of speech data across 136 language pairs, including low-resource languages, mining both real-world and synthetic speech from sources like VoxPopuli. While it supports scaling multilingual speech translation systems, its focus on European languages limits its representation of Asian and non-European languages.

III. Corpus Design and Data Collection

Building on this analysis, we paired medium-resource and high-resource languages with scalability in mind, using Korean as the source language to complement existing corpora. To create a corpus for domain-specific simultaneous speech translation, an emerging field, we used a variety of academic and specialized texts. In designing the corpus, we addressed common issues in real-world simultaneous interpretation, such as omissions, mistranslations, abbreviations, paraphrasing, and noise, to ensure its technical usability.

A. Topic Selection and Text Extraction

After securing the necessary copyrights, text was randomly extracted by chapter from specialized books categorized using a domain-specific adaptation of the Dewey Decimal Classification (DDC). This random selection ensured a diverse and balanced range of content across different topics, meeting the target of 1,000 hours of Korean audio recordings. Although the final data is sentence-aligned, interpretation and translation were performed chapter-by-chapter to preserve the contextual integrity needed for complex speech-to-speech translation.

B. Data Preprocessing

Due to the nature of published texts, sentences with excessive foreign terms, numbers, or units were preprocessed and removed. Punctuation was retained to capture intonation patterns. For speech recognition, synthesis, and translation tasks, Korean sentences averaged 12-15 words, while translations ranged from 15-25 words. Meanwhile, the selected Korean

TABLE II
PARTICIPANTS GENDER DISTRIBUTION

Language	Male (%)	Female (%)	Total
Korean	19 (36.54%)	33 (63.46%)	52
English	10 (11.76%)	75 (88.24%)	85
Spanish	15 (16.48%)	76 (83.52%)	91
Japanese	6 (9.09%)	60 (90.91%)	66
Total	50 (17.01%)	244 (82.99%)	294

text was initially machine-translated (MT) to produce a literal translation, a common practice in simultaneous interpretation settings. Interpreters often refer to MT outputs when preparing for real-time interpretation, especially when they receive speaker scripts shortly before the session begins. Building on this practice, our data collection process involved interpreters providing real-time translations while considering MT outputs and listening to Korean audio narrated by professional voice actors, simulating actual simultaneous interpreting conditions. This method blends simultaneous interpretation with sight translation, ensuring contextual fidelity and technical accuracy while enhancing interpretation quality in the recording phase.

C. Recording Protocols for Simultaneous Sight Interpretation

A second interpreter listened to the same Korean audio and performed simultaneous sight interpretation in a controlled recording studio environment while referring to their peer's sight-translated output in the previous stage. This process ensured accurate and faithful interpretation of the original speech, allowing the second interpreter to correct any semantic errors or unnaturalness. Peer review guarantees high linguistic quality, which is essential for creating a dataset that supports the development of high-performance speech-to-speech translation models.

D. Data Alignment and Annotation

The interpreter who performed the sight interpretation used the authoring platform to match the recorded audio with the transcription, ensuring proper synchronization between the speech and the text. This sentence-level alignment accurately linked the spoken language with the target translations, creating a reliable dataset for training speech-to-speech translation systems. The platform supported the alignment process, combining interpreter adjustments and automated tools to ensure consistency.

E. Participant Recruitment

We recruited 294 professionals (242 interpreters and 52 Korean voice actors), to ensure linguistic proficiency and accuracy. As Table II indicates, the corpus remains skewed toward female voices. Globally, the majority of foreign language majors are women, and this trend was reflected in our participant pool.

F. Transcription Guidelines

The corpus includes the book text as an orthographic transcription, Korean voice actor recordings, and sight interpretation audio with transcriptions. Discrepancies between text and audio were addressed during the transcription validation stage by either re-recording the audio or adjusting the transcription to match the audio. This process faithfully conveys the original text's intent, supporting high-quality speech-to-speech translation models.

IV. METHODOLOGICAL INNOVATIONS

This research introduces key innovations to improve the quality and utility of multilingual speech translation corpus, particularly for handling complex, domain-specific content in limited-resource languages.

A. Simultaneous Sight Interpretation

Our workflow models real-world interpretation by integrating simultaneous interpretation and sight translation. Interpreters reference machine-translated outputs while interpreting live speech, improving contextual alignment and accuracy. This hybrid approach produces higher-quality results than traditional machine translation or standalone interpretation and resolves the asynchronicity inherent in traditional sight translation methods used in specialized fields like law and medicine.

B. Peer Review Integrated

Simultaneous sight-interpreting serves as a peer-review process. After the initial sight translation, a second interpreter reviews the output while simultaneously sight-interpreting, correcting errors, and ensuring faithfulness to the original speech. This step improves interpretation and transcription (or translation) quality, which is essential for training reliable speech-to-speech models that need to handle complex linguistic features like nuances, idiomatic expressions, and cultural references.

C. Multi-Stage Alignment and Annotation

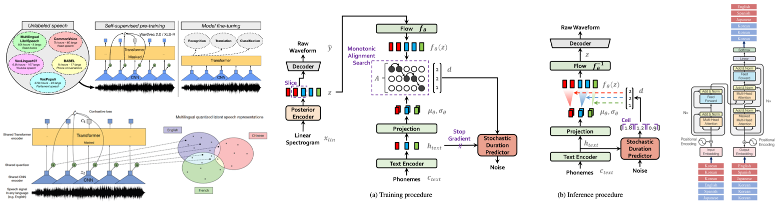
IDs were automatically assigned based on the upload order of the Korean text, with corresponding IDs then applied to Korean and foreign audio and transcriptions. Interpreters aligned transcriptions with audio, and external reviewers verified accuracy, ensuring a reliable and well-mapped dataset for multilingual speech-to-speech translation tasks.

D. Domain-Specific Content Categorization

The corpus uses a domain-specific Dewey Decimal Classification (DDC) adaptation to ensure a broad range of topics. This enables models to handle domain-specific language challenges across technical, academic, and specialized fields (See Table I).

E. Strategic Language Pairing

Unlike large-scale English-centric corpora, the Korean language lacks sufficient data for unsupervised or semi-supervised learning, especially for speech processing and machine translation. Thus, the collected data is intended for fine-tuning models pre-trained on high-resource languages for speech recognition, synthesis, and machine translation tasks. English, Spanish, and



(a) Training procedure

(b) Inference procedure

Fig. 1. Model algorithms [19], [20], [21], [22]

Japanese were strategically included for knowledge distillation-based transfer learning. Japanese shares an SOV syntactic structure with Korean, while English and Spanish follow SVO structures, providing a basis for exploring linguistic differences in translating complex, domain-specific content. This setup also allows us to analyze how interpreters manage these structural differences and make semantic choices to convey equivalent meanings across languages.

V. Quality Assurance

This corpus underwent rigorous internal evaluations and external validation by the Telecommunications Technology Association (TTA), Korea's leading ICT standards organization.

A. Schema Validation and Data Alignment

Content, organized by a domain-specific adaptation of the Dewey Decimal Classification (DDC), demonstrates diverse and balanced topic representation across languages, as shown in the Data Structure Table. Both schema validation and data alignment were completed with 100% accuracy, ensuring that the corpus meets high consistency standards, crucial for speech recognition and translation tasks.

B. Transcription and Translation Quality

Based on sampled validation, Korean speech transcription achieved 99.72% accuracy, demonstrating robust speech recognition. Translation accuracy was 90.37% (Korean-English), 91.20% (Korean-Spanish), and 92.50% (Korean-Japanese), validated by expert assessments, with 1–5 scores converted to a 100-point scale to ensure high performance for linguistically demanding tasks.

$$\mathit{Transcription\ Accuracy}\ (\%) = 100\times\left(1-\frac{Ic+Dc+Sc}{\text{Total number of true syllables}}\right)$$

- Ic = Number of extra syllables added beyond the true syllables
- Dc = Number of syllables deleted from the true syllables
- Sc = Number of syllables replaced in the true syllables

C. Data Fitness for Purpose

We conducted preliminary tests using state-of-the-art models to validate the dataset's utility for speech recognition, synthesis, and machine translation (see Figure 1 and Table III). For ASR, WER decreased from 6.46% (Spanish) and 7.38% (Japanese), measured using 60% of the dataset, to 5.40% and 4.47% after fine-tuning. In speech synthesis, MOS improved from 3.75 (10% dataset) to 4.12 (100% dataset) for Korean. For machine translation, baseline BLEU scores, obtained using 10% of the dataset, increased from 5.25 (Korean-English) and 18.87 (Korean-Japanese) to 20.31 and 57.10, respectively, after fine-tuning with the full dataset. These results underscore the corpus's effectiveness in multilingual speech processing and its potential for fine-tuning models in complex, domain-specific tasks.

TABLE III
Model Performance Comparison

Task	Model	Evaluation Metric	Baseline Performance	Fine-tuning Results
ASR (Spanish)	Wav2vec 2.0	WER	6.46%	5.40%
ASR (Japanese)	Wav2vec 2.0	WER	7.38%	4.47%
TTS (Korean)	VITS	MOS	3.75	4.12
MT (Korean-English)	Transformer	BLEU	5.25	20.31
MT (Korean-Japanese)	Transformer	BLEU	18.87	57.10

VI. Limitations and Future Work

This corpus was designed to reflect real-world simultaneous interpretation practices and cognitive processes within the data processing workflow. However, because it was not sourced from live simultaneous interpretation sessions, it falls short of fully capturing the spontaneity of real-time interpretation. The dataset also exhibits a gender imbalance, with an over-representation of female voices, reflecting trends in language programs. Furthermore, using professional interpreters and voice actors for high-quality data collection poses scalability challenges. Future efforts will address these limitations by diversifying the dataset to include a wider range of male voices and expanding the language coverage. Additionally, intonation marking will be incorporated to enable a more detailed analysis of prosodic features.

VII. Conclusion

Given the scarcity of domain-specific data, constructing a 4,000-hour parallel speech dataset with Korean as the source language is a significant contribution. Rooted in Korean culture, the source texts differentiate this dataset from Western corpora, such as the Gutenberg Project, with a balanced topic distribution ensured by the Dewey Decimal Classification (DDC) system. Including professional voice recordings, it supports applications in speech recognition, translation, and synthesis, particularly for L2 and code-switching tasks.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, ”Language Models are Few-Shot Learners,” in Advances in Neural Information Processing Systems, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., Curran Associates, Inc., 2020, pp. 1877-1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[2] Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu, W. Liu, Z. Wu, W. Gong, J. Liang, Z. Shang, P. Sun, W. Liu, X. Ouyang, D. Yu, H. Tian, H. Wu, and H. Wang, ”ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation,” arXiv preprint, arXiv:2107.02137, 2021. [Online]. Available: https://arxiv.org/abs/2107.02137.

[3] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin, ”Beyond English-Centric Multilingual Machine Translation,” arXiv preprint, arXiv:2010.11125, 2020. [Online]. Available: https://arxiv.org/abs/2010.11125.

[4] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, ”On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?,” in Proc. 2021 ACM Conf. on Fairness, Accountability, and Transparency (FAccT ’21), Virtual Event, Canada, 2021, pp. 610–623. doi: 10.1145/3442188.3445922.

[5] A. V. K. Shanbhogue, R. Xue, S. Saha, D. Zhang, and A. Ganesan, ”Improving Low Resource Speech Translation with Data Augmentation and Ensemble Strategies,” in Proc. 20th Int. Conf. on Spoken Language Translation (IWSLT 2023), Toronto, Canada (in-person and online), Jul. 2023, pp. 241-250. Association for Computational Linguistics. [Online]. Available: https://aclanthology.org/2023.iwslt-1.21. doi: 10.18653/v1/2023.iwslt-1.21.

[6] H. Tan and S. Sakti, ”Contrastive Feedback Mechanism for Simultaneous Speech Translation,” in Interspeech 2024, 2024, pp. 852-856. doi: 10.21437/Interspeech.2024-2426.

[7] K. Deng and P. Woodland, ”Label-Synchronous Neural Transducer for E2E Simultaneous Speech Translation,” in Proc. 62nd Annu. Meeting Assoc. for Comput. Linguistics (Volume 1: Long Papers), Bangkok, Thailand, Aug. 2024, pp. 8235-8251. [Online]. Available: https://aclanthology.org/2024.acl-long.448.

[8] H. Yang, M. Zhang and D. Wei, ”IRAG: Iterative Retrieval Augmented Generation for SLU,” 2024 20th IEEE International Colloquium on Signal Processing & Its Applications (CSPA), Langkawi, Malaysia, 2024, pp. 30-34, doi: 10.1109/CSPA60979.2024.10525270.

[9] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, ”wav2vec 2.0: a framework for self-supervised learning of speech representations,” in Proc. 34th Int. Conf. on Neural Information Processing Systems (NeurIPS 2020), Vancouver, BC, Canada, 2020, pp. 1-12.

[10] J. Pino, Q. Xu, X. Ma, M. J. Dousti, and Y. Tang, ”Self-Training for End-to-End Speech Translation,” in Proc. Interspeech 2020, 2020, pp. 1476-1480. doi: 10.21437/Interspeech.2020-2938.

[11] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, ”Libri-Light: A Benchmark for ASR with Limited or No Supervision,” ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7669-7673, doi: 10.1109/ICASSP40776.2020.9052942.

[12] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, ”Librispeech: An ASR corpus based on public domain audio books,” 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015, pp. 5206-5210, doi: 10.1109/ICASSP.2015.7178964.

[13] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, ”MuST-C: a Multilingual Speech Translation Corpus,” in Proc. 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2019, pp. 2012-2017.

[14] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, ”MLS: A Large-Scale Multilingual Dataset for Speech Research,” in Proc. Interspeech 2020, 2020, pp. 2757-2761, doi: 10.21437/Interspeech.2020-2826.

[15] P. Jeuris and J. Niehues, ”LibriS2S: A German-English Speech-to-Speech Translation Corpus,” in Proc. Thirteenth Language Resources and Evaluation Conf., Marseille, France, 2022, pp. 928-935.

[16] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, ”VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation,” in Proc. 59th Annu. Meeting of the Association for Computational Linguistics and the 11th Int. Joint Conf. on Natural Language Processing (Volume 1: Long Papers), Online, 2021, pp. 993-1003.

[17] Y. Jia, M. Tadmor Ramanovich, Q. Wang, and H. Zen, ”CVSS Corpus and Massively Multilingual Speech-to-Speech Translation,” in Proc. Thirteenth Language Resources and Evaluation Conf., Marseille, France, 2022, pp. 6691-6703.

[18] P.-A. Duquenne, H. Gong, N. Dong, J. Du, A. Lee, V. Goswami, C. Wang, J. Pino, B. Sagot, and H. Schwenk, ”SpeechMatrix: A Large-Scale Mined Corpus of Multilingual Speech-to-Speech Translations,” in Proc. 61st Annu. Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada, 2023, pp. 16251-16269.

[19] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, ”XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale,” in Interspeech 2022, 2022, pp. 2278-2282. doi: 10.21437/Interspeech.2022-143.

[20] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, ”Unsupervised Cross-Lingual Representation Learning for Speech Recognition,” in Interspeech 2021, 2021, pp. 2426-2430. doi: 10.21437/Interspeech.2021-329.

[21] J. Kim, J. Kong, and J. Son, ”Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech,” in Proc. ICML 2021, 2021. [Online]. Available: https://arxiv.org/abs/2106.06103.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ”Attention is All You Need,” in Advances in Neural Information Processing Systems, vol. 30, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[23] Y. Jia, Y. Ding, A. Bapna, C. Cherry, Y. Zhang, A. Conneau, and N. Morioka, ”Leveraging unsupervised and weakly-supervised data to improve direct speech-to-speech translation,” arXiv preprint, arXiv:2203.13339, 2022. [Online]. Available: https://arxiv.org/abs/2203.13339.