Dataset for Dzongkha and English Speech-to-Speech Translation

Dorji Phuntsho[1*], Tashi Wangchuk[2], Sumit Adhikari[3], Tshering Norphel[4], Karma Wangchuk[5]

Information and Technology Department, College of Science and Technology, Royal University of Bhutan
E-mail: dorjiphuntsho1491@gmail.com[1*], sumitadhikari828@gmail.com[3*], tnorphel09@gmail.com[4*]
karma.cst@rub.edu[5*]

Received: 21 June 2025; Revised: 23 July 2026; Accepted: 8 August 2026; Published: 23 August 2026

Abstract
Dzongkha is a low-resource language. Lack of a parallel Speech-to-speech translation dataset is a major challenge. This paper introduces a novel method for curating the first Dzongkha-English speech parallel corpus. The proposed method uses divergent TTS models to synthesize aligned speech from translated text pairs. Furthermore, Google's gTTS is used for English, and a tonally adjusted Tacotron2 system for Dzongkha. The dataset includes 7,385 Dzongkha ASR audio-text pairs, 419,731 Dzongkha-English machine translation text pairs, 1,530 Dzongkha TTS audio-text pairs, and 240 hours of synthetic speech data each for Speech-to-Unit and Dzongkha Unit-to-Speech tasks. The dataset curation pipeline entails rigorous text normalization, 22.05kHz dual-language audio recording, and three-stage signal conditioning-noise reduction, silence truncation, and EBU R128 loudness normalization. Most notably, we use tool-assisted temporal alignment followed by human verification to ensure sub-second synchronism of utterances. By rejecting 8.7% of samples via PESQ quality thresholds, we ensure acoustic consistency. This work constructs a scalable method for producing high-fidelity parallel speech data for under-resourced languages directly facilitating S2ST model development in the lack of natural datasets.

Keywords: *Dzongkha Speech Corpus, Parallel Speech Data Synthesis, Low-Resource Speech Translation, Tonal Text-To-Speech, Speech-To-Speech Alignment, Under-Resourced Language Processing*

1. INTRODUCTION

Dzongkha, the national language of Bhutan, is a significant cultural symbol but constitutes serious challenges in the computing age. Being a tonal Sino-Tibetan language with an agglutinative morphology and a classical Tibetan-derived script, Dzongkha resists simplistic computational modeling. Due to its low resource nature, it has extra problems that make it inapplicable to the many mainstream speech technologies made for languages with significant resources, such as Mandarin or English.

In Bhutan's multilingual setting, where Dzongkha coexists with more than 19 regional dialects, this disparity is especially significant. Speech-to-speech (S2ST) translation would revolutionize government services, healthcare, and education. In low-resource environments, compound errors hinder cascaded systems that rely on text-to-speech (TTS), machine translation (MT), and automatic speech recognition (ASR). Direct S2ST offers a desirable substitute by skipping the intermediate text but remains data hungry.

Three core challenges to progress: a desert of data, with little or no speech-text aligned datasets; phonological challenge like tonal minimal pairs and dialect variation; and an infrastructural chasm, where little exists for data harvesting or model training. This paper argues for a community-driven framework for S2ST development for Dzongkha on the principles of linguistic prioritization, cross-lingual transfer, and participatory design. By contrasting cascaded and direct S2ST models via comparative analysis, it offers a framework for building equitable speech technologies in low-resource settings.

2. RELATED WORK

2.1 Cascaded Speech Translation Systems

Cascaded speech-to-speech translation (S2ST) systems decompose the translation process into three successive modules: automatic speech recognition (ASR), machine translation (MT), and text-to-speech synthesis (TTS). This modular

sequence has long been the dominance of speech translation research, particularly for high-resource language scenarios. Pioneering systems such as Verbmobil demonstrated the feasibility of this approach for various language pairs (Wahlster, 2000). An advantage of cascaded architecture is that they are modular, and thus, it is straightforward to optimize each module separately (Duong et al., 2016).

However, cascaded systems perform poorly in low-resource settings. Error propagation is the main problem: ASR errors, such as misrecognitions, are carried over to the MT phase and build up in the final output. In Dzongkha, where ASR systems frequently have word error rates (WER) of more than 40%, this problem is especially severe. The end translations are consequently rendered incomprehensible. They reported the same constraints on Tibetan, a linguistically similar language. Kano et al. (2022) reported that the cascaded S2ST systems had 62% inferior translation quality compared to text-based machine translation using clean transcripts.

2.2 Direct Speech-to-Speech Models

To mitigate the cascaded error problem, researchers have put forward end-to-end direct S2ST models that skip intermediate text representations by projecting source speech to target speech directly. Translatotron is among the pioneering systems proposed in this direction, merging recognition and synthesis into a unified neural architecture (Jia et al., 2019). These models have the benefit of retaining paralinguistic features such as speaker identity and prosody, which usually get lost within modular cascades.

But the success of direct S2ST models depends on access to large-scale, high-quality parallel speech corpora, which is typically not available for low-resource languages. Methods such as COSTT (Lee et al., 2022) attempt to reduce reliance on parallel data by using disentangled representations from monolingual corpora. These methods still require speech-text alignments for the source and target languages. For tone languages such as Dzongkha, direct systems are faced with additional challenges as pitch contours with lexical meaning need to be modeled.

2.3 ASR Dataset Preparation

Another largest cause of delay in the development of ASR for low-resource languages is that annotated speech corpora are in short supply. Traditional ASR pipelines rely on large-scale human-transcribed data, which are time-consuming and expensive to produce. Recent approaches such as Wav2Vec 2.0 have gone some way to reversing this dependence on manually annotated data through the application of self-supervised learning on raw audio and reducing the use of labeled data by as much as S90% (Baevski et al., 2020).

2.5 Text-To-Speech (TTS) Data Preparation

The creation of TTS systems for tone languages such as Dzongkha with fewer resource data is very problematic due to smaller annotated data and difficulty in pitch and prosody modeling. Conventional TTS designs require massive multi-speaker databases to learn the natural speech variations. A few latest few-shot adaptation methods such as AdaSpeech can do successful speaker adaptation with only one minute of speech by using conditional layer normalization and acoustic-level encoders (Chen et al., 2021). In multi-dialect synthesis, FMSD-TTS integrates dialect-conditioning and dynamic routing to generate high-quality Tibetan speech across dialects with few-shot learning (Liu et al., 2025).

3. METHODOLOGY

3.1 Dataset Collection

This study developed a parallel Dzongkha–English speech corpus to enable cascaded and direct speech-to-speech translation (S2ST) systems. Official speech content in the form of announcements, policy speeches, and official broadcasts were received from the Department of Culture and Dzongkha Development (DCDD) as Dzongkha audio recordings. English translations were prepared manually, whereas English translations were prepared via Google Translate for sentence-level alignments. While automated translation may not always capture tonal flavor or idiomatic nuance, it can facilitate scalable data creation. The corpus spans a broad range of domains-from formal text to colloquial spoken language-to promote lexical variety and model generalization.

When there was insufficient naturally aligned speech, TTS systems were employed to synthesize audio for English and Dzongkha. Text pairs were used to create high-fidelity waveforms in support of training on subcomponents such as ASR, MT, and TTS in cascaded models while data augmentation in end-to-end S2ST systems. The resulting dataset consists of native Dzongkha speech, bilingual text alignments, and parallel synthesized audio in both directions.

3.2 Speech Data Generation

To mitigate the unavailability of parallel Dzongkha–English naturally recorded speech data—essential for training end-to-end speech translation models—this work resorted to synthetic speech synthesis for both the language pairs. English was synthesized with Google's Text-to-Speech (gTTS) API using a 24 kHz British English (en-GB) voice with quality suitable for use in downstream applications like ASR and prosody modeling. For Dzongkha, a specialized Tacotron2-based TTS model was developed and trained using approximately 10 hours of native DCDD speech. The model incorporated tone-level pitch features to capture the tonal and rhythmic properties of Dzongkha well in the synthesized speech, with phonological fidelity guaranteed.

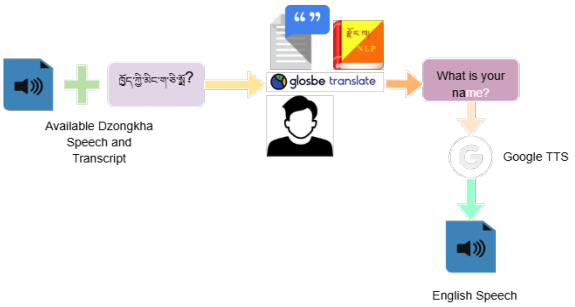

***Figure 1**: Speech Data Generation*

The resulting parallel corpus of high-quality synthetic speech for both languages was used to train two translation systems: (a) a cascaded system with ASR, MT, and TTS components, and (b) a speech-to-unit translation (S2UT)-based direct S2ST model that obviates the need for an intermediate text. This corpus can sidestep resource limitations inherent to low-resource languages and provide a scalable, linguistics-based alternative to natural parallel speech. In addition, the methodology suggests a reproducible framework that can be used for other under-resourced or endangered languages with similar data constraints.

3.3 Audio Preprocessing

All the audio files were cut into sensible durations of between several seconds and several minutes. Background noise reduction methods by spectral gating through LibROSA were employed in addition to eliminating long silences using the WebRTC Voice Activity Detector (VAD). Volume normalization was carried out across the dataset for uniformity in the recordings.

Audio files longer than 120 seconds or with perceptual quality scores lower than a PESQ (Perceptual Evaluation of Speech Quality) score of 0.85 were excluded from analysis. Furthermore, amplification was applied to normalize recordings to model-tuned levels of volume. Based on such preprocessing demands, approximately 8.7% of the audio files were excluded to maintain data quality standards.

***Figure 2**: Audio Preprocessing process*

.4 TEXT PREPROCESSING

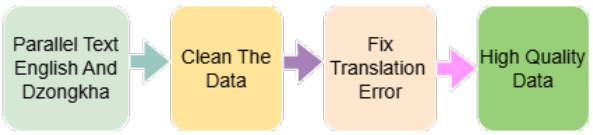

***Figure 3**: Text Preprocessing*

Prior to model training, parallel text Dzongkha–English data were preprocessed using an intensive preprocessing pipeline to ensure linguistic quality, structural integrity, and alignment accuracy—paramount parameters for enhancing machine translation (MT) performance. The initial step was cleaning sentence pairs by removing entries with missing text, non-linguistic features, and incomplete or extremely short utterances.

3.5 Mapping and Annotation

Sentence-level annotation was acquired using uncomplicated mapping methods that did not have laborious phonetic or time-aligned annotation. In the S2UT corpus, English speech and the Dzongkha speech along with their English speech translation share identical filenames and source and target speech segments can be directly paired.

All Dzongkha audio files of the Automatic Speech Recognition (ASR) corpus were also aligned with their respective transcripts, and both the files were aligned with a distinctive id such that alignment will be convenient during training.

Similarly, in the Machine Translation (MT) corpus, source text Dzongkha and target text English are aligned according to the same file names so that there is uniformity in the corpus. This light and regular annotation form facilitated efficient dataset construction for low-resource settings, without fine-grained phoneme segmentation or timing synchronization.

***Table 1**: Dzongkha Text mapping*

ID	Dzongkha Transcription
001f000002	དགེ་སློང་། དེའི་རིགས་ནི་ལུང་གིས་བསྟན་པའི་ཚུལ་ལ་འཇུག་ག་དང་མཐུན་པའི་སྡོམ་ལ་མཐུན་པ་ཅན་ཞེ།
001f000001	རྒྱལ་ཁབ་ཀྱི་སིང་པ།
001f000003	དགེ་བའི་ལས་ཀ་བསྒྲུབས།

3.6 Implementation Environment

All data synthesis, preprocessing, and audio processing activities were carried out with Python 3.8 on the Anaconda distribution. Key libraries included gTTS (v2.3.1) for English speech synthesis, LibROSA (v0.9.2) for audio analysis, and PyDub (v0.25.1) for audio file conversion and manipulation. ELAN (v6.4) and Audacity (v3.2.1) were used for annotation and waveform alignment that facilitated segment-level editing with precision and quality validation.

3.7 Dataset Split and Annotation Guidelines

The curated Dzongkha-English parallel speech dataset was split into training, and validation sets in an 80:20 ratio for further reproducibility and model development. The training set had about 219000 samples, and about 40000 samples in the validation set. It was utilized to learn the model, the validation set was kept measuring the model performance and prevent overfitting during training.

Each speech segment in Dzongkha and English translator was matched and annotated with a unique identifier for the speech segment. To keep all source text and target text, source speech and target speech consistent across the various datasets, the same method of mapping the name of the file was employed for the ASR, MT, TTS, S2U and vocoder output datasets. In the preparation of the corpus, sentence pairs that contained missing text or no complete utterances, content that was non-linguistic in nature or very short sentences were deleted. The aligned pairs were additionally examined for temporal alignment using a tool, and for semantic accuracy and sub-second synchronization by human review.

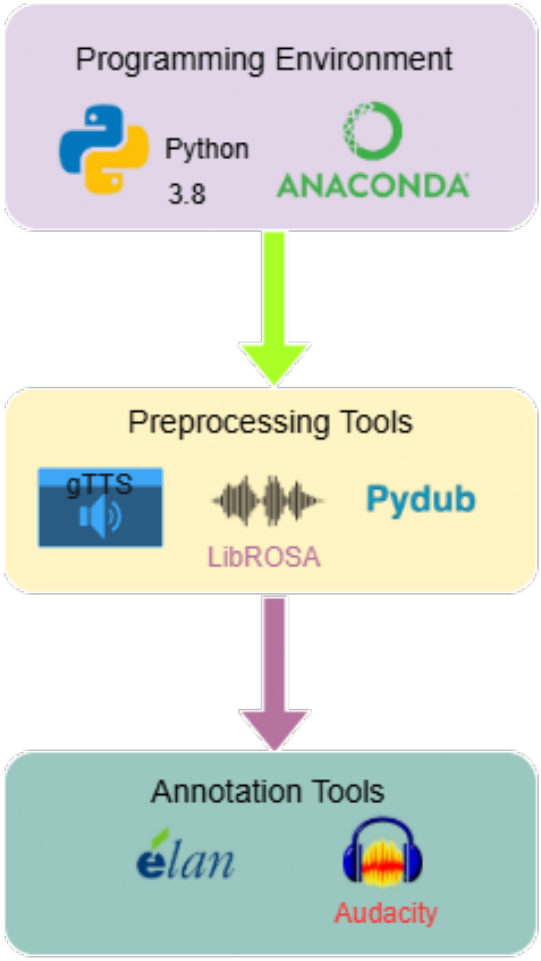

***Figure 4**: Audio implementation environment*

4. RESULTS

4.1 Dataset Composition

The final dataset included several task-specific subsets designed to enable cascaded and direct speech translation model training. The subsets included data for Machine Translation (MT), Automatic Speech Recognition (ASR), Text-to-Speech (TTS), Speech-to-Unit (S2U), and Unit-to-Speech (vocoder) tasks. The Dzongkha (Dzo) ASR dataset, for instance, included 7,385 audio-text pairs from GovTech, while the MT dataset

included 419,731 parallel text pairs from GovTech. Secondly, the Dzo TTS dataset consisted of 1,530 pairs of audio-text provided by the Department of Culture and Dzongkha Development (DCDD). Besides, the dataset comprised 240 hours of synthetic voice audio per S2U and vocoder task for encouraging speech unit modeling and waveform synthesis, respectively. The multi-task dataset played a crucial role in training both modular sub-components and end-to-end systems. Task alignment compatibility was ensured by segmentation across tasks.

Table. 2**:** *Dataset statistics used for different tasks: The Dzongkha Automatic Speech Recognition (ASR) task utilized 7,385 audio-text pairs collected from GovTech. For Machine Translation (MT), 419,731 Dzongkha-English text pairs were used, also sourced from GovTech. The Dzongkha Text-to-Speech (TTS) task used 1,530 audio-text pairs from the Department of Culture and Dzongkha Development (DCDD). Speech-to-Unit (S2U) and Dzongkha Unit-to-Speech (vocoder) tasks were trained on 240 hours of synthetic speech data.*

Task	Dataset Size
Dzongkha (ASR)	7385
Machine Translation (MT)	419,731
Dzongkha (TTS)	1530
Speech to Unit (S2U)	240 hrs
Dzongkha Unit to Speech	240 hrs

This encouraged effective training of independent ASR, MT, and TTS models, as well as coupled speech-to-speech translation (S2ST) pipelines.

4.2 Alignment Accuracy

To evaluate alignment fidelity between text and speech, a visual inspection was done on a randomly selected sample of 500 sentence pairs. The inspection had a 94% accuracy in alignment, i.e., the proportion of samples with both temporal and semantic similarity within a ±0.2-second tolerance. The rest of 6% consisted mostly of insignificant timing adjustments or partial mismatches over coarticulated speech boundaries.

5. CONCLUSION AND FUTURE WORK

This work suggested a scalable approach to building a parallel Dzongkha–English speech corpus to support cascaded and direct speech-to-speech translation (S2ST) models. By the combination of speech synthesis, alignment techniques, and human validation, the built corpus bridges critical gaps in low-resource language processing and supports modular (ASR, MT, TTS) and end-to-end (S2UT) architectures. This work contributes to inclusive AI by making it possible to create speech technology for underrepresented languages. Subsequent work will focus on training and evaluating S2ST models according to metrics like BLEU, METEOR, character error rate (CER), and mean opinion score (MOS) with particular emphasis on tone and prosody. Extension to dialectal and domain-specific data are foreseen, in addition to transfer learning from close languages like Tibetan. Additional work includes optimizing models for real-time deployment, local community engagement in participatory data collection, and improving tonal modeling through prosody-sensitive architectures. Combined, these include building a deployable, culture-aware Dzongkha–English S2ST system and offering an architecture transferable to other low-resource languages.

6. REFERENCE

Anderson, O. (n.d.). *Ethical Considerations in Machine Translation : Bias , Fairness , and Accountability Sources of Bias in Machine Translation*.

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (n.d.). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. https://github.com/pytorch/fairseq

Chen, M., Duquenne, P. A., Andrews, P., Kao, J., Mourachko, A., Schwenk, H., & Costa-Jussà, M. R. (2023). BLASER: A Text-Free Speech-to-Speech Translation Evaluation Metric. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *1*, 9064–9079. https://doi.org/10.18653/v1/2023.acl-long.504

Communication, S., Chung, Y., Meglioli, M. C., Dale, D., Duquenne, P., Elsahar, H., Gong, H., Heffernan, K., Hoffman, J., Klaiber, C., Li, P., Licht, D., Maillard, J., Rakotoarison, A., Ram, K., Wenzek, G., Ye, E., Akula, B., Chen, P., … Wang, S. (n.d.). *SeamlessM4T: Massively Multilingual & Multimodal Machine Translation*.

Dugan, L., Wadhawan, A., Spence, K., Callison-Burch, C., McGuire, M., & Zordan, V. (2023). Learning When to Speak: Latency and Quality Trade-offs for Simultaneous Speech-to-Speech Translation with Offline Models. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2023-Augus*, 5265–5266.

Hira, M., Goel, A., & Gupta, A. (2024). *CrossVoice: Crosslingual Prosody Preserving Cascade-S2ST using Transfer Learning*. 1–8. http://arxiv.org/abs/2406.00021

Huang, S.-F., Lin, C.-J., Liu, D.-R., Chen, Y.-C., & Lee, H. (2021). *Meta-TTS: Meta-Learning for Few-Shot Speaker Adaptive Text-to-Speech*. https://doi.org/10.1109/TASLP.2022.3167258

Inaguma, H., Popuri, S., Kulikov, I., Chen, P. J., Wang, C., Chung, Y. A., Tang, Y., Lee, A., Watanabe, S., & Pino, J. (2023). UnitY: Two-pass Direct Speech-to-speech Translation with Discrete Units. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *1*, 15655–15680. https://doi.org/10.18653/v1/2023.acl-long.872

Jia, Y., Ramanovich, M. T., Remez, T., & Pomerantz, R. (n.d.). *Translatotron 2: High-quality direct speech-to-speech translation with voice preservation*. https://google-research.github.io/