

Textless Speech-to-Speech Translation from Nepali to English Using Discrete Units: A Deep Learning Approach

Alina Chudali*, Asmi Adhikari, Abhishek Panthi, Ashok G.M.

Himalaya College of Engineering, Tribhuvan University, Nepal

*Corresponding Author: poudelalina86@gmail.com

Abstract

This study implements direct speech to speech translation eliminating the conventional need for Automatic Speech Recognition (ASR) and Machine Translation (MT) by leveraging HuBERT for discrete unit extraction, a Transformer-based encoder-decoder for translation, and HiFi-GAN for speech reconstruction. Inspired by recent advances in unit-based S2ST, the model is trained on the multilingual FLEURS dataset for Nepali to English language translation. It achieved a BLEU score of 42.80, demonstrating both translation accuracy and natural-sounding speech synthesis. This work offers a scalable and effective solution for low-resource languages like Nepali.

Keywords: Speech-to-Speech Translation, HuBERT, Transformer, HiFi-GAN, FLEURS Dataset, Discrete Units

1. Introduction

Conventional speech translation systems are often built upon cascaded pipelines that depend on text transcripts for automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS). Such systems perform well in high-resource settings but fall short when dealing with under-represented languages that lack sufficient annotated data or standardized orthographies.

Recent research in speech processing has introduced direct speech-to-speech translation (S2ST) systems that bypass intermediate text. These systems leverage self-supervised learning and unit-based modeling to translate directly between spoken languages, enabling applications even for unwritten or under-documented languages.

This study implements a discrete unit-based textless speech-to-speech translation framework that directly maps Nepali speech to English speech. The system utilizes log-Mel filterbank features for Nepali acoustic representation, a Transformer-based encoder-decoder architecture for sequence modeling, and HuBERT-derived discrete units as translation targets. HiFi-GAN is employed for reconstructing natural-sounding English speech from predicted unit sequences.

The main contributions of this work are:

- Development of a textless Nepali-to-English speech-to-speech translation system using discrete unit modeling.
- Demonstration of feasibility for low-resource language speech translation without intermediate text representation.

2. Literature Review

Speech-to-speech translation (S2ST) has become an essential technology for bridging communication barriers across languages. Traditional S2ST systems rely on intermediate textual representations involving Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-To-Speech

(TTS), often requiring large annotated datasets [1]. While effective for high-resource languages, these pipelines present significant challenges for low-resource languages like Nepali.

Textless S2ST addresses these limitations by directly translating source speech into target speech without relying on text. This approach is particularly advantageous for languages lacking standardized orthographies or sufficient transcribed data [2]. Unit-based methods using HuBERT or Wav2Vec 2.0 extract discrete speech units via self-supervised learning. These units act as pseudo-text tokens for translation using encoder-decoder models, with vocoders like HiFi-GAN reconstructing the final speech output [3].

End-to-end S2ST models, such as those proposed by Jia et al. [4], translate speech embeddings directly, bypassing intermediate steps. These models may also incorporate speaker embeddings to preserve voice characteristics. Alternatively, continuous representation approaches use Mel spectrograms for acoustic-to-acoustic translation, which retain prosody but face alignment challenges due to variable-length sequences [5].

Pretrained multilingual models such as Whisper and XLS-R leverage cross-lingual embeddings to improve performance across diverse language pairs, reducing the need for language-specific resources [6]. Fine-tuning these models for S2ST tasks enhances adaptability and quality.

Developing S2ST systems for Nepali remains difficult due to the scarcity of parallel corpora and the diversity in accents and dialects. As noted by Kumar et al., corpus availability remains a major bottleneck. Unsupervised and data-efficient methods like HuBERT and Wav2Vec 2.0 offer promising solutions.

Translation quality is typically assessed using BLEU, Word Error Rate (WER), and Mean Opinion Score (MOS). BLEU and WER measure objective accuracy, while MOS provides subjective evaluations of fluency and naturalness [7]. Combining these metrics offers a comprehensive evaluation of translation systems.

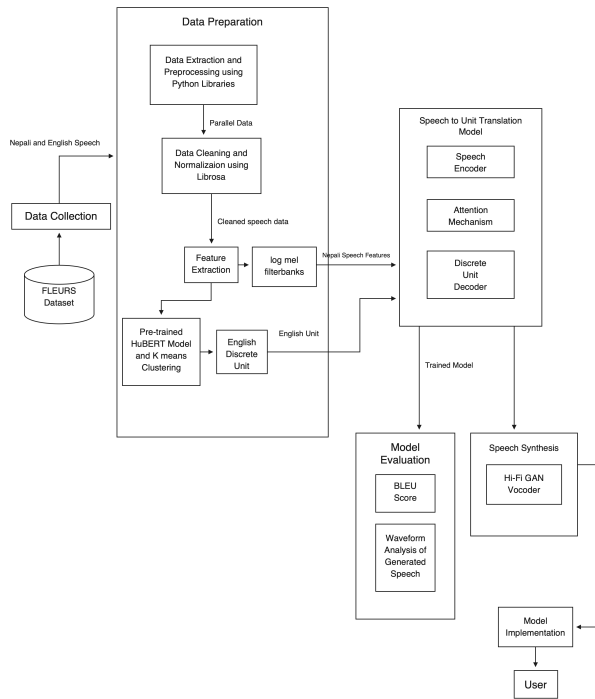


Figure 1. System Architecture Illustrating the Complete Workflow of the Textless Speech-to-Speech Translation

3. Objectives

The main objective of this project is to develop a robust and efficient system for translating Nepali speech into English speech using discrete unit-based speech-to-speech translation. Specifically, the goal is:

- To design and implement a translation system that translates spoken Nepali input directly into fluent English speech by leveraging HuBERT for unit extraction, a Transformer-based translation model, and HiFi-GAN for speech synthesis.

4. Methodology

The methodology includes systematic procedures for data acquisition, preprocessing, feature extraction, model architecture design, and training configuration.

The overall architecture of the textless speech-to-speech translation system is shown in Figure 1. It presents the complete processing pipeline, including unit extraction, translation, and speech synthesis stages, demonstrating how Nepali speech is transformed into English speech.

4.1 Data Collection and Preparation

1) Data Collection

The speech data used in this project was sourced from the multilingual FLEURS (Few-shot Learning Evaluation of Universal Representations of Speech) dataset, available on Hugging Face. This dataset includes high-quality parallel audio in WAV format, sampled at 16 kHz. Approximately 15 hours of speech

A	B
/content/drive/MyDrive/aawaj/data/english/train	
1155.wav	169920
429.wav	179520
608.wav	130560
1103.wav	223680
1315.wav	137280
823.wav	184320
606.wav	132480
687.wav	210240
14.wav	105600
1224.wav	199680
1429.wav	204480
1484.wav	136320
705.wav	195840
1064.wav	170880
1072.wav	84480
721.wav	216960
946.wav	193920
1199.wav	89280

Figure 2. Manifest File

were collected for parallel Nepali and English language, with an average utterance duration of 4–8 seconds. The dataset was split into 70% training, 15% development, and 15% testing sets to ensure balanced training and evaluation.

2) Data Preprocessing

To ensure data consistency and quality, several preprocessing steps were applied:

- **Parallel Pairing:** A custom Python script was used to extract and align parallel Nepali-English utterances based on shared IDs from the dataset's metadata. This ensured each Nepali audio was correctly paired with its English counterpart.
- **Silence Trimming and Normalization:** Silence segments were removed using SoX and Librosa. All audio files were normalized and resampled to 16 kHz, matching the requirements of downstream models.
- **Manifest File Creation:** A structured manifest file was created containing metadata such as file paths, IDs, language tags, and durations. This enabled efficient loading and alignment during training.

4.2 Feature Extraction

Feature representations differ between Nepali and English due to the model's architectural design. Nepali speech, serving as the source input, is encoded using log-Mel filterbank features that capture continuous acoustic information for the encoder. English speech, used as the target, is represented as discrete units obtained from self-supervised HuBERT embeddings clustered via k-means. This approach allows the system to map continuous source acoustics to linguistically meaningful discrete targets while supporting high-quality speech reconstruction through the vocoder.

Nepali Speech: Log-Mel filterbank features were computed

from preprocessed Nepali speech audio to represent the signal in the frequency domain, mimicking human auditory perception. The process involved segmenting the waveform $S = \{s_1, s_2, \dots, s_T\}$, $s_t \in \mathbb{R}$ into short overlapping frames, then applying the Short-Time Fourier Transform (STFT) to each frame to obtain the spectral representation $X_t = \text{FFT}(s_t)$. The magnitude spectra were passed through Mel filterbanks $H_m(f)$, and the energy in each Mel band was computed as:

$$E_m = \sum_f H_m(f) \cdot |X_t(f)|^2 \quad (1)$$

Finally, a logarithmic compression was applied to approximate human loudness perception:

$$F_{\text{Nepali}} = \log(E_m) \quad (2)$$

These log-Mel features served as the input to the encoder and were aligned with English discrete units for effective speech-to-speech translation.

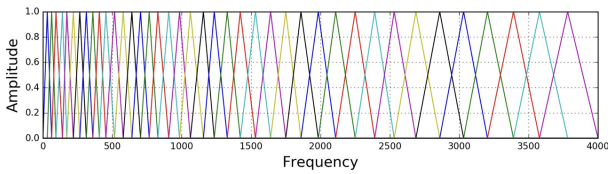


Figure 3. Filterbanks on Mel Scale

English Speech: Discrete units were obtained using a pretrained HuBERT model followed by k -means clustering. HuBERT converts raw speech $S = \{s_1, s_2, \dots, s_T\}$ into a sequence of high-level embeddings $H = \{h_1, h_2, \dots, h_T\}$, $h_t \in \mathbb{R}^d$, where T is the number of frames and d is the embedding dimension.

Each embedding h_t is then assigned to the nearest centroid c_k from a learned set of K centroids $C = \{c_1, c_2, \dots, c_K\}$ using:

$$z_t = \arg \min_{k \in \{1, \dots, K\}} \|h_t - c_k\|^2 \quad (3)$$

This results in a sequence of discrete units $Z = \{z_1, z_2, \dots, z_T\}$, which serve as symbolic representations of the English speech. These units act as the target in the translation model.

The HuBERT Base architecture consists of a convolutional feature encoder followed by a Transformer encoder. The convolutional encoder includes seven layers with strides $[5, 2, 2, 2, 2, 2, 2]$ and kernel widths $[10, 3, 3, 3, 3, 3, 2]$, generating 512 output channels to capture low-level acoustic patterns while reducing temporal resolution. The resulting features are processed by a 12-layer Transformer with an embedding dimension of 768, feed-forward network dimension of 3072, and 8 self-attention heads to model long-range contextual dependencies. Finally, a 256-dimensional projection layer maps the contextualized representations into the discrete unit space. This configuration, comprising approximately 95

million parameters, enables robust extraction of linguistically meaningful representations for downstream speech-to-speech translation.

The pairing of Nepali log-Mel filterbanks and English discrete units provided a fully parallel training signal. Since the data was aligned at the utterance level, manual alignment was unnecessary. This streamlined the training process and enabled the model to learn robust mappings between Nepali acoustic features and English unit sequences.

5. Model Training and Speech Synthesis

5.1 Model Training

The speech-to-speech translation model is trained to map Nepali input speech to English output speech using discrete units as an intermediate representation. The training data consists of parallel Nepali and English audio files aligned by unique IDs, enabling supervised learning of the mapping between Nepali speech features and English discrete units.

The Nepali audio input is first converted into log-Mel filterbank features, which capture frequency and temporal characteristics of speech. These features are generated through a pipeline involving framing, Short-Time Fourier Transform (STFT), Mel-scale filtering, and logarithmic compression, thus mimicking the human auditory perception.

The log-Mel features are fed into a Transformer-based encoder-decoder model. The encoder processes the Nepali features and generates a latent, language-agnostic intermediate representation. This representation captures the phonetic and prosodic information necessary for translation.

The decoder maps this representation to a sequence of English discrete units, extracted beforehand using a pretrained HuBERT model and k -means clustering. The model is supervised using a cross-entropy loss function, which compares the predicted units to the ground-truth units.

- **Encoder:** Converts log-Mel filterbanks into an intermediate latent representation by capturing linguistic and acoustic cues.
- **Intermediate Representation:** A dense, abstract embedding that bridges the source and target languages.
- **Decoder:** Translates the intermediate representation into a sequence of English discrete units.
- **Loss Function:** The model is optimized using cross-entropy loss between predicted and target unit sequences.

The system is implemented using the Fairseq framework, employing the `speech_to_speech` task with the `--target-is-code` option to indicate unit-level outputs. The model is optimized using the Adam optimizer with an inverse square root learning rate schedule and a warm-up phase. Regularization techniques such as dropout and gradient clipping are applied to ensure stable and efficient training.

5.2 Speech Synthesis

Once the Transformer decoder outputs the English discrete unit sequence, the speech synthesis module reconstructs the corresponding audio waveform. This process is performed in two stages:

- **Duration Prediction and Upsampling:** A duration predictor estimates the temporal alignment of each discrete unit to ensure natural prosody. The discrete units are then upsampled based on these durations.
- **Vocoder (HiFi-GAN):** The upsampled sequence is passed to the HiFi-GAN vocoder, which generates a high-quality English speech waveform. HiFi-GAN excels at producing intelligible and natural-sounding speech, making it ideal for real-time synthesis applications.

This synthesis mechanism ensures that the final output preserves fluency, clarity, and expressive timing, delivering natural speech directly from discrete unit sequences.

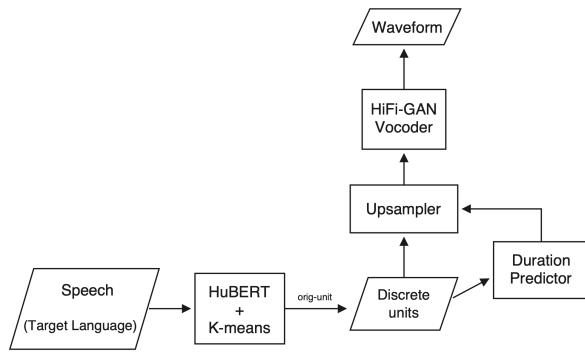


Figure 4. Mechanism of Speech Synthesis from Discrete Units

6. Results and Analysis

6.1 Model Evaluation

The model was trained using mini-batch optimization to achieve stable learning and efficient use of computational resources. Training was carried out for 60 epochs with a batch size of 10, allowing the model to gradually improve the mapping between Nepali speech features and English discrete units. The Adam optimizer was used for parameter updates due to its reliability in training deep learning models, with adaptive moment parameters $\beta_1 = 0.9$ and $\beta_2 = 0.98$ to maintain steady convergence. HuBERT embeddings were further converted into symbolic units using k -means clustering with $K = 100$ clusters, which served as targets for translation. These training settings contributed to the decreasing loss trend shown in Fig. 5, indicating stable training and good generalization performance.

The training loss (blue line) steadily decreased, indicating effective learning of acoustic-to-unit mappings. The validation loss (red line) followed a similar decreasing pattern and stabilized in later epochs. The best model was selected at

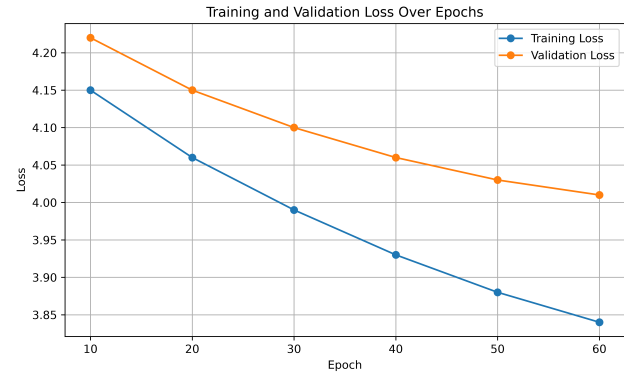


Figure 5. Loss Trend During Model Training

the point of minimum validation loss to ensure generalization performance.

Although a small gap between training and validation loss was observed, this is expected given the moderate dataset size (15 hours) and does not indicate severe overfitting.

6.2 Evaluation Using BLEU Score

The BLEU score was computed on the 15% held-out test set consisting of unseen parallel utterances. Since the system generates English speech output rather than text, an Automatic Speech Recognition (ASR) model was used to transcribe both the system's output English speech and the corresponding actual English speech into textual form.

Specifically, the evaluation procedure consisted of the following steps:

- The trained speech-to-speech translation model generated English speech output from Nepali input speech.
- The obtained output English speech was transcribed into text using a pretrained ASR model.
- The reference English speech from the dataset was also transcribed using the same ASR model to ensure consistency.
- The predicted transcriptions were compared with the reference transcriptions using the BLEU metric to measure translation accuracy.

The proposed system achieved a BLEU score of **42.80**, indicating that the machine has learned meaningful mappings.

6.3 Waveform Analysis of Generated Speech

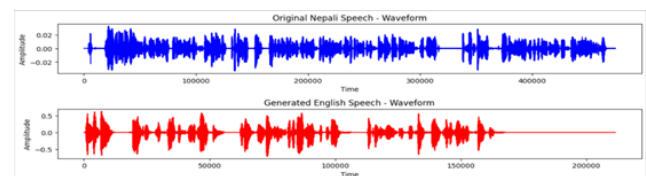


Figure 6. Waveform Analysis of Synthesized English Speech

To further analyze audio quality, waveform visualization of the generated English speech was performed (Figure 6). This analysis helps identify timing consistency, abrupt silences, or waveform distortions. The waveform showed stable and

smooth transitions, confirming that the output maintains natural temporal structure and amplitude variation.

7. Conclusion

This work successfully presents a textless speech-to-speech translation (S2ST) system that enables direct Nepali-to-English speech translation without relying on textual or phonemic intermediaries. The architecture integrates HuBERT for discrete unit extraction, a Transformer-based encoder-decoder model for translation, and HiFi-GAN for speech synthesis.

The system achieved a BLEU score of 42.80 indicating strong translation accuracy and natural speech generation. These results validate the effectiveness of discrete unit-based approaches for speech translation, especially in low-resource language scenarios like Nepali. The study contributes to multilingual speech processing by demonstrating a scalable and language-agnostic framework for direct speech translation.

8. Discussion

While the system shows promising results, several limitations remain:

- **Loss of Speaker Identity:** The discrete unit-based approach struggles to retain speaker-specific characteristics, resulting in loss of voice identity during synthesis.
- **Data Scarcity for Nepali:** The limited availability of large-scale, high-quality Nepali speech datasets restricts model performance and robustness.
- **Expressive Speech Handling:** The system has difficulty capturing emotional tone, pitch variation, and other expressive features, which can affect the naturalness of the output.

Future Work

To address the above limitations, future enhancements may include:

- **Data Expansion:** Incorporating more diverse and domain-specific speech datasets for Nepali can improve generalization.
- **Model Fine-Tuning:** Fine-tuning the HuBERT model on language-specific data may yield more accurate unit representations.
- **Improved Vocoding:** Replacing HiFi-GAN with advanced vocoders such as WaveGlow or Parallel WaveGAN can enhance synthesis quality.

Acknowledgements

The authors are grateful to Himalaya College of Engineering for this wonderful opportunity.

References

- [1] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," *ACL*, 2002.
- [2] A. Lee, H.-C. Lee, P.-A. Duquenne, and H. Schwenk, "Textless speech-to-speech translation on real data," in *Proceedings of the 2021 Conference on Neural Information Processing Systems (NeurIPS)*, 2021.
- [3] P.-J. Chen and K. Tran, "Speech-to-speech translation for a real-world unwritten language," 2023.
- [4] Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," in *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.
- [5] C. Lee, Y. Zhang, W. Han, D. S. Park, Y. Saraf, A. Narayanan, K. Kirchhoff, and Y. Wu, "Continuous representations for cross-lingual speech translation," in *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022.
- [6] R. Ma, M. Qian, and Y. Fathullah, "Cross-lingual transfer learning for speech translation," 2024.
- [7] E. Salesky and J. Mader, "Assessing evaluation metrics for speech-to-speech translation," 2021.
- [8] S. Nakamura, "Overcoming the language barrier with speech translation technology," *Science & Technology Trends - Quarterly Review*, vol. 31, Apr. 2009.