Journal of Information Hiding and Multimedia Signal Processing ©2026 ISSN 2073-4212
Ubiquitous International Volume **17**, Number **1**, March **2026**

Toward Speech-to-Speech Translation in Low-Resource Education: English–Vietnamese Cascade Benchmark

Trang Pham Thi Thuy[1]
[1]Faculty of Information Technology
University of Information Technology - VNUHCM
Ho Chi Minh City, VietNam
21522697@gm.uit.edu.vn

Tu Nguyen Tuan[1]
[1]Faculty of Information Technology
University of Information Technology - VNUHCM
Ho Chi Minh City, VietNam
21522744@gm.uit.edu.vn

Thuan Nguyen Dinh[1,*]
[1]Faculty of Information Technology
University of Information Technology - VNUHCM
Ho Chi Minh City, VietNam
thuannd@uit.edu.vn

Nhut Nguyen Minh[1]
[1]Faculty of Information Technology
University of Information Technology - VNUHCM
Ho Chi Minh City, VietNam
nhutnm.17@grad.uit.edu.vn

*Corresponding author: Thuan Nguyen Dinh

Received November 15, 2025, revised February 7, 2026, accepted February 12, 2026.

Abstract. *Online education is becoming increasingly widespread, yet language barriers remain a major challenge for Vietnamese learners, particularly in the field of Information Technology where most lectures and materials are delivered in English. This paper presents a modular English–Vietnamese speech translation pipeline designed to support online IT lectures. The system focuses on Automatic Speech Recognition (ASR) and Machine Translation (MT) as the initial stages of a cascade architecture. In ASR experiments, Whisper-medium achieved relatively strong performance (WER $\approx$ 3.36%, CER $\approx$ 1.57%), while Whisper-small, HuBERT, and Wav2Vec2 showed substantially higher error rates. For MT, Gemini 2.0 Flash produced the most promising results (BLEU $\approx$ 55, chrF $\approx$ 72, TER $\approx$ 38), with faster processing and broader coverage compared to mBART, NLLB, EnViT5, and OpusMT. These findings suggest the feasibility of a cascade-based approach and provide a benchmark for future extensions. In particular, integrating Text-to-Speech (TTS), voice cloning, and lip synchronization could move the system toward full speech-to-speech translation and improve accessibility for multilingual online IT education.*

Keywords: Speech Translation, English–Vietnamese, Automatic Speech Recognition (ASR), Machine Translation (MT), Cascade Pipeline, Online IT Education, Voice Cloning, Lip Synchronization

1. **Introduction.** The rapid growth of digital learning has made online teaching unavoidable for any area of study, including Information Technology. Yet, a large share of high-quality courses and lectures is still delivered in English, which creates a substantial barrier for Vietnamese learners. An off-the-shelf solution involves using machine-translated subtitles; however, this often falls short on domain-specific content. This challenge is reinforced by Editage Insights (2023) [1], which reports that non-native English learners spend about 47% more time reading and 51% more time writing academic materials than native speakers. As a result, both learning efficiency and productivity can suffer.

To address this gap, this study investigates the development of an English–Vietnamese speech translation pipeline for online IT lectures. At the current stage, the pipeline focuses on two key components: Automatic Speech Recognition (ASR) and Machine Translation (MT), aiming to generate accurate transcriptions and translations while preserving the meaning and academic nuances of the original lectures. Modern models are evaluated and compared, including Whisper [2], Wav2Vec2 [3], and HuBERT [4] for ASR, as well as mBART [5], NLLB [6], EnViT5 [7], OpusMT, and Gemini for MT.

The major contribution of this work is the design and evaluation of a modular pipeline that incorporates several state-of-the-art models with effective configurations for English–Vietnamese speech translation in online IT education. This study also forms the foundation for future extensions such as the integration of Text-to-Speech (TTS), voice cloning, and lip synchronization to enhance naturalness and preserve speaker identity in translated lectures.

2. **Related works.** In recent years, Speech-to-Speech Translation (S2ST) has attracted increased interest as an alternative to classical subtitles. Jia et al. (2019) [8] introduced *Translatotron*, one of the first end-to-end systems mapping source speech directly to target speech without using text as an intermediate layer. However, the system often produced unstable outputs and lacked naturalness. Subsequently, Jia et al. (2021) [9] developed *Translatotron 2*, which incorporated attention mechanisms and speaker encoders to better preserve voice characteristics. Nevertheless, limitations remained in voice quality and language scalability.

The cascade architecture, which is more frequently used in S2ST systems, includes Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS) modules in sequence. As Popescu-Belis et al. (2025) [10] underlined, cascade models enable specialized expertise at each stage and often provide better performance than end-to-end systems, especially in low-resource languages such as Vietnamese. Nguyen et al. (2022) [11] also reported similar findings, where cascade systems outperformed end-to-end systems in English-to-Vietnamese translation tasks, achieving significantly higher BLEU scores.

Substantial progress has also been made in individual components. For ASR, Radford et al. (2022) [2] proposed *Whisper*, a large-scale multilingual model trained on 680,000 hours of data, showing strong performance across languages. Le et al. (2024) [12] further introduced *PhoWhisper*, a fine-tuned version optimized for Vietnamese, which achieved superior performance. On the MT side, multilingual models like mBART [5], NLLB [6], and EnViT5 [7] have shown strong BLEU scores on domain-specific corpora. More recently, large language models (LLMs) such as *Gemini* have been shown to produce competitive or superior translations in challenging scenarios such as idiomatic expressions, outperforming ChatGPT and Google Translate in certain benchmarks (Haider, 2023) [13]. In addition to Whisper, self-supervised ASR approaches including Wav2Vec2 [3] and HuBERT [4] are widely used as baselines in multilingual settings.

Apart from ASR and MT researches, recent years also witnessed significant development of text-to-speech and voice cloning technologies, making complete S2ST procedures more feasible than ever. Recently, VoiceCraft (P. Peng et al., 2024) [14] proposes a codec-based generative model capable of high-quality voice editing and voice retention. Meanwhile, XTTS (Casanova et al., 2023) [15] and F5-TTS (Y. Chen et al., 2025) [16], demonstrated robust multilingual voice cloning performance with only a few seconds of reference audio, enabling natural and consistent synthesis across different languages. Studies on TTS also show potential for application to Vietnamese, such as zero-shot Vietnamese TTS (Vu, Nguyen, & Nguyen, 2025) [17], indicating increasing potential in applying these techniques in English-Vietnamese speech translation. These researches point to promising directions for integrating voice cloning and speech synthesis into speech translation applications.

However, most of these studies focus on individual components and do not actually build or evaluate a complete pipeline for practical use. This study combines several state-of-the-art models in ASR and MT into a single unified pipeline and provides a reproducible benchmark for the task of English-Vietnamese speech translation in online education. In the near future, we would like to further extend the pipeline with TTS, voice cloning, and lip synchronization to improve naturalness and provide a better overall user experience.

3. **Motivation and Contribution.** The three relevant observations that have motivated this work are: (1) Educational inequality - IT students in Vietnam do not have equal opportunities to access English language educational resources due to language barriers; (2) Technological gap - although significant advances have recently been made in ASR and MT, comprehensive benchmarking of integrated systems for English–Vietnamese speech translation is still limited; and (3) Practical needs - online IT education platforms need real-time translation solutions that preserve the technical accuracy of source content.

This research contributes on several key dimensions. First, we provide a systematic comparison of the state-of-the-art ASR and MT models for IT education English–Vietnamese speech translation tasks, represented by five ASR models: Whisper Small/Medium, Wav2Vec2 Base/Large, HuBERT Large, and five MT models: mBART, NLLB-200, EnViT5, OpusMT, Gemini 2.0 Flash. Second, we constructed a domain-specific dataset of 246 English academic lecture videos, covering IT, AI, Cloud Computing, and Computer Vision courses. Third, we designed and validated a modular cascade architecture, yielding strong performance: ASR: WER $\approx$ 3.36%; MT: BLEU $\approx$ 55. Fourth, a practical framework is given that can be applied directly to online IT education platforms. Finally, this research forms a solid basis for subsequent extensions toward full speech-to-speech translation systems with TTS, voice cloning, and lip synchronization.

4. **Proposed Methods.**

4.1. **Dataset construction.**

We built our experimental dataset from English academic lecture videos sourced from open educational platforms and well-known YouTube teaching channels. To ensure both technical quality and linguistic diversity, we applied the following selection criteria: (i) content in the IT domain with precise, discipline-specific terminology; (ii) high-quality audio with minimal background noise (bitrate $\geq$ 128 kbps); (iii) video resolution of at least 720p to enable potential lip-synchronization analysis; (iv) lecture segments ranging from 15–25 minutes to support manageable segmentation; and (v) delivery by native English speakers representing a range of speaking styles.

These criteria produced audio data that not only met the requirements for ASR but also preserved sufficient linguistic variation to ensure a reliable basis for MT evaluation. The resulting dataset serves as the foundation for training and evaluating the proposed speech translation pipeline. Figure 1 summarizes the overall data collection workflow. The overall data collection workflow is illustrated in Figure 1.

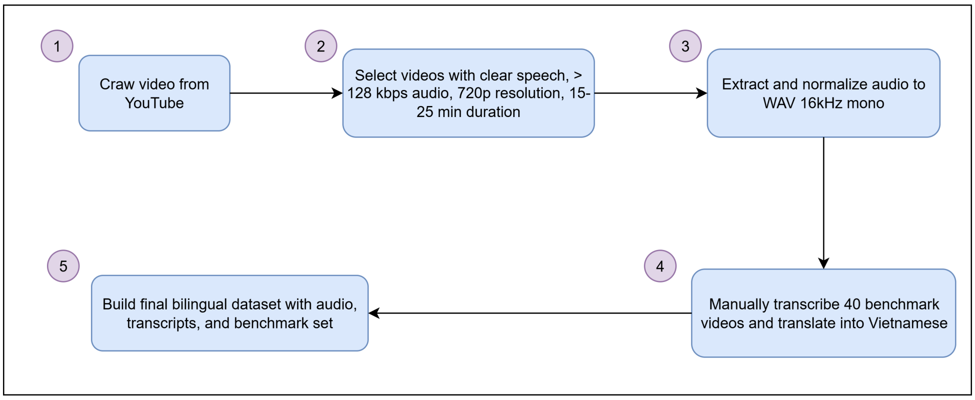
1
Craw video from YouTube
2
Select videos with clear speech, > 128 kbps audio, 720p resolution, 15-25 min duration
3
Extract and normalize audio to WAV 16kHz mono
4
Manually transcribe 40 benchmark videos and translate into Vietnamese
5
Build final bilingual dataset with audio, transcripts, and benchmark set

FIGURE 1. Workflow of the data collection process

4.2. **PreProcessing Data.**

Data preprocessing was one of the most labor-intensive stages of this study. Although the raw dataset was pre-filtered for audio and content quality, several additional steps were required before experimentation.

First, all 240 source videos were converted to WAV format at 16 kHz, mono to ensure compatibility with modern ASR models. This step required substantial computational and storage resources.

Second, to construct a benchmark set for evaluation, 40 videos were selected from the IBM Technology channel, which features domain-rich terminology, diverse accents, and fast speaking rates. Each video was manually transcribed to create reference transcripts. Manual transcription was critical, as the quality of reference data directly affects the reliability of metrics such as WER and CER.

Third, the transcripts were translated into Vietnamese under a unified set of guidelines. Common IT terminology, especially programming languages, frameworks, and tools, was preserved in English due to its familiarity among Vietnamese developers. In contrast, fundamental academic concepts (e.g., *Database*, *System*, *Algorithm*) were translated into Vietnamese to align with educational materials. The translation style was adapted to context: natural and conversational for tutorials, but more formal and precise for academic lectures. Pronouns and sentence structures were adjusted for clarity, ensuring translations were concise and balanced with the original subtitles.

The outcome of preprocessing is a standardized dataset consisting of audio, transcripts, and bilingual translations, along with a benchmark set of 40 video/audio samples. This dataset serves as the core resource for training and evaluating the proposed English–Vietnamese speech translation pipeline.

4.3. **Data Statistics.** The final dataset was constructed from four YouTube channels focusing on technical education, covering areas such as Information Technology, Artificial Intelligence, Cloud Computing, Computer Vision, and academic lectures. In total, the dataset contains 246 videos spanning a range of lengths and topics. Table 1 summarizes the overall characteristics, and Table 2 details the distribution across individual channels.

Table 1. Overview of the dataset statistics

Attribute	**Value**
Number of source channels	4
Total number of videos	246
Average duration	27.2 minutes
Main domains	IT, AI, Cloud Computing, Computer Vision, Academic Lectures

Table 2. Video distribution across different source channels

Source Channel	**Video Count**	**Domain**	**Average Duration**
IBM Technology	164	Technology, AI, Cloud	8.5 minutes
First Principles of Computer Vision	61	Computer Vision, ML	12.3 minutes
MIT OpenCourseWare (Playlist 1)	10	Academic Lectures	45.2 minutes
MIT OpenCourseWare (Playlist 2)	11	Academic Lectures	42.8 minutes

The collection mixes short introductory pieces (e.g., IBM Technology; First Principles of Computer Vision) with long, in-depth lectures (e.g., MIT OpenCourseWare). This diversity ensures not only a wide range of linguistic features but also adequate content depth, creating a strong foundation for assessing the robustness and generalizability of ASR and MT models.

4.4. **Dataset for Experiments.** The dataset used in this study consists of 240 English academic lecture videos collected from four YouTube channels on technical education, covering topics such as Information Technology, Artificial Intelligence, Computer Vision, and academic lectures.

After preprocessing, all data were standardized and stored in four main formats to support different stages of the pipeline, as summarized in Table 3.

To enable fair model comparison, a benchmark set of 40 video/audio samples was constructed using stratified sampling. This subset reflects the characteristics of the full dataset in terms of length, domain, and speaking rate, and includes manually verified transcripts and reference translations.

These resources serve two main purposes: (i) training and evaluating ASR models using the 40 WAV files with reference transcripts, and (ii) evaluating MT models using the reference translations with BLEU, chrF, and TER metrics.

TABLE 3. Standardized dataset formats

Type	Format	Purpose
Video	MP4 (240 files)	Serves as the original raw material for annotation, transcription, and translation.
Audio	WAV, 16kHz mono (240 files)	Represents the standardized audio input used across ASR models to ensure consistent evaluation.
Transcriptions	TXT/JSON (240 files)	Contains both model-generated transcriptions and human-provided reference transcripts for ASR benchmarking.
Translations	JSON (240 files)	Provides English–Vietnamese reference translations used for machine translation evaluation.

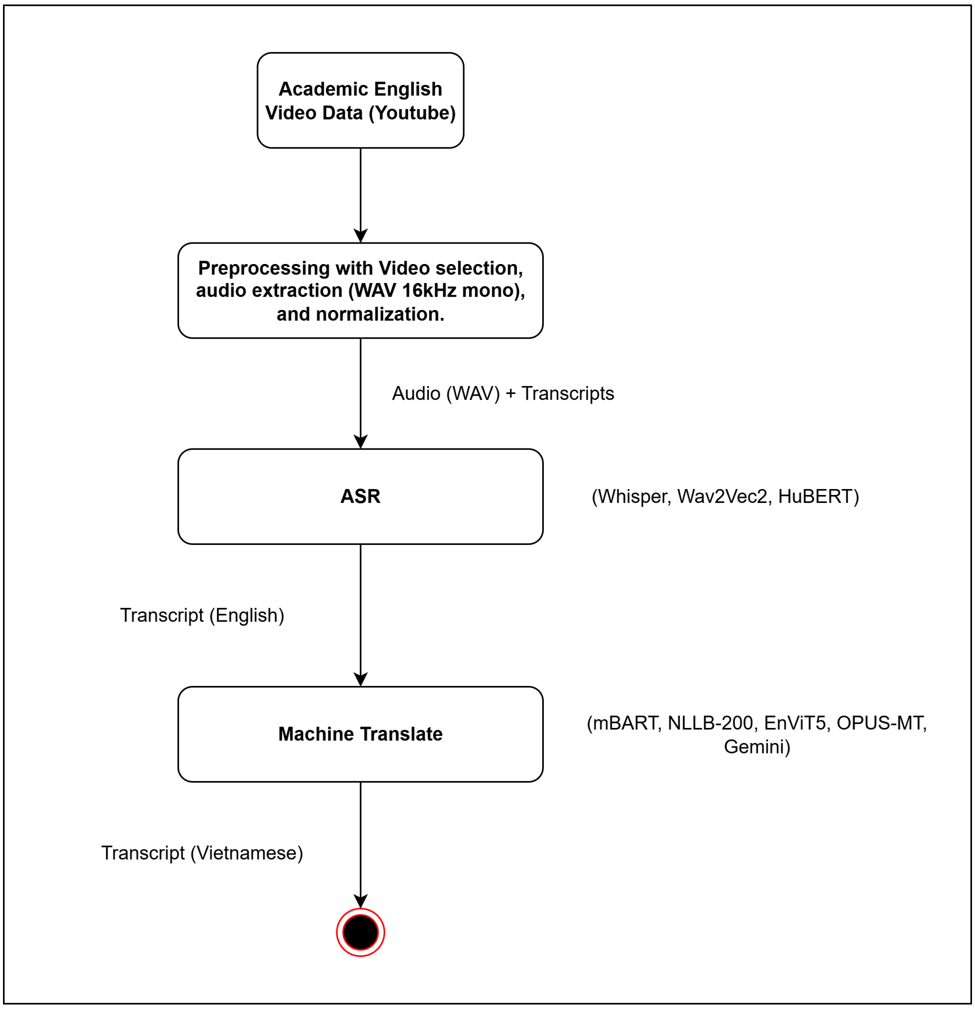
Academic English Video Data (Youtube)
Preprocessing with Video selection, audio extraction (WAV 16kHz mono), and normalization.
Audio (WAV) + Transcripts
ASR
(Whisper, Wav2Vec2, HuBERT)
Transcript (English)
Machine Translate
(mBART, NLLB-200, EnViT5, OPUS-MT, Gemini)
Transcript (Vietnamese)

FIGURE 2. Overview of the proposed cascade pipeline: English speech → ASR transcript → MT translation (Vietnamese).

4.5. **Proposed Pipeline.** We adopt a cascade architecture for English–Vietnamese lecture translation. The system operates sequentially: the Automatic Speech Recognition (ASR) module first transcribes English speech into text; the resulting transcript is then passed to the Machine Translation (MT) module to produce the Vietnamese translation. This modular design facilitates component-wise evaluation and easy swapping or upgrading of individual blocks.

In short, the input is English speech, which is converted to text by the ASR component; that transcript is then translated into Vietnamese by the MT component.

4.6. **Models Selection.** The proposed English–Vietnamese cascade pipeline comprises two core components: Automatic Speech Recognition (ASR) and Machine Translation (MT). We compare representative state-of-the-art models at each stage to identify effective configurations for online IT lectures.

4.6.1. *ASR models.*

- **Whisper (Small, Medium)** [2]: a large-scale multilingual ASR model trained on ~680k hours of weakly supervised speech–text data, noted for robustness in noisy conditions and competitive performance on low-resource languages.
- **Wav2Vec 2.0 (Base, Large)** [3]: a self-supervised framework that learns acoustic representations directly from raw waveforms and can be fine-tuned with limited labeled data.
- **HuBERT (Large)** [4]: leverages masked prediction of clustered hidden units to improve phonetic representations, often yielding competitive WER/CER on medium-to-large training sets.

4.6.2. *MT models.*

- **mBART** [5]: a multilingual sequence-to-sequence Transformer with denoising pretraining, enabling effective fine-tuning for English–Vietnamese.
- **NLLB-200** [6]: the No Language Left Behind effort supporting 200+ languages, with solid English–Vietnamese coverage.
- **MTet / EnViT5** [7]: Vietnamese-focused resources for English–Vietnamese translation across multiple domains; complementary to prior ViT5/EnViT5 work.
- **OPUS-MT** [18]: open models trained on OPUS corpora; practical and easy to deploy/fine-tune.
- **Gemini 2.0 Flash** [13]: a proprietary LLM-based baseline that has shown competitive translation quality and speed in challenging scenarios, suitable for near real-time applications.

4.7. **Evaluation Metrics.** To evaluate the performance of our proposed speech-to-speech translation pipeline, we adopted standard metrics commonly used in automatic speech recognition and machine translation tasks. These measures provide a reliable basis for assessing system accuracy and allow meaningful comparison with other state-of-the-art approaches.

4.7.1. *Word Error Rate (WER).* measures the proportion of word-level errors (substitutions, deletions, insertions) required to transform the system output into the reference transcription:

$$\mathrm{WER} = \frac{S + D + I}{N}, \tag{1}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of correctly recognized words, and $N = S + D + C$ is the total number of words in the reference.

4.7.2. *Character Error Rate (CER).* similar to WER but operates at the character level. It is particularly useful for languages where word boundaries are ambiguous or where character-level precision is critical:

$$\mathrm{CER} = \frac{S_c + D_c + I_c}{N_c}, \tag{2}$$

where S_c is the number of substitutions at the character level, D_c is the number of deletions at the character level, I_c is the number of insertions at the character level, C_c is the number of correctly recognized characters, and $N_c = S_c + D_c + C_c$ is the total number of characters in the reference.

4.7.3. *Sentence Error Rate (SER).* measures the proportion of sentences that are recognized incorrectly, regardless of how many word-level errors they contain:

$$\mathrm{SER} = \frac{S_{\mathrm{err}}}{S_{\mathrm{total}}}, \tag{3}$$

where S_{err} is the number of sentences containing at least one error (substitution, insertion, or deletion), and S_{total} is the total number of sentences in the evaluation set.

4.7.4. *Bilingual Evaluation Understudy (BLEU).* [19] measures the n-gram overlap between the system output and one or more reference translations:

$$\mathrm{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{4}$$

where p_n is the modified n-gram precision (the proportion of n-grams in the hypothesis that also appear in the reference), w_n is the weight assigned to n-grams (typically uniform, e.g., 0.25 for 1–4 grams), and BP is the brevity penalty applied when the hypothesis translation is shorter than the reference.

4.7.5. *chrF Score.* [20] evaluates machine translation quality at the character level. Unlike BLEU, which relies on word n-gram precision, chrF combines character n-gram precision and recall:

$$\mathrm{chrF}_\beta = \frac{(1+\beta^2)\cdot chrP \cdot chrR}{\beta^2 \cdot chrP + chrR}, \tag{5}$$

where $chrP$ is the character n-gram precision (proportion of character n-grams in the hypothesis that also appear in the reference), $chrR$ is the character n-gram recall (proportion of character n-grams in the reference that are found in the hypothesis), and β is a weighting factor controlling the importance of recall relative to precision.

Typical settings: $\beta = 1$ (balanced), $\beta > 1$ (favor recall), $\beta < 1$ (favor precision).

4.7.6. *Translation Edit Rate (TER).* [21] measures the number of edits (insertions, deletions, substitutions, and shifts) required to transform the system output into the closest reference translation, normalized by the average length of the reference:

$$\mathrm{TER} = \frac{\text{number of edits}}{\text{average number of reference words}} \tag{6}$$

4.8. **Modular Design and Expansion Prospects.** The cascade architecture offers flexible interchangeability of models at each stage, allowing the creation of a controlled and transparent benchmark. This is especially valuable for online IT education, where diverse technical terminology is common and Vietnamese remains a relatively low-resource language. In addition, the modular design lays the groundwork for a future speech-to-speech pipeline, where the translated Vietnamese transcript can be directly passed to a TTS system. By combining with voice cloning and lip-syncing technologies, this extension aims to deliver a more natural, intuitive, and learner-friendly translation experience.

5. **Experiments.**

5.1. **Automatic Speech Recognition (ASR).** We evaluated five ASR models on the 40-file benchmark set. Within this comparison, Whisper Medium achieved the best performance (WER ≈ 3.36%, CER ≈ 1.57%), while Whisper Small also produced relatively strong results (WER ≈ 5.20%). In contrast, HuBERT Large and Wav2Vec2 (Base, Large) showed error rates exceeding 60%, indicating limited suitability for specialized online lecture scenarios.

Table 4. ASR performance on the 40-file benchmark set

Model	**Params**	**WER (%)**	**CER (%)**	**Remarks**
Whisper Medium	769M	∼3.36	∼1.57	Best performance in this comparison
Whisper Small	244M	∼5.20	∼2.03	Reasonably strong
HuBERT Large	316M	∼61.59	∼11.32	High error rate
Wav2Vec2 Large	317M	∼64.86	∼22.46	Very high error rate
Wav2Vec2 Base	95M	∼61.56	∼20.42	Very high error rate

5.2. **Machine Translation (MT).** Table 5 presents MT performance. Among the tested models, Gemini 2.0 Flash obtained the highest scores (BLEU ≈ 55, chrF ≈ 72, TER ≈ 38), reflecting relatively good translation quality in academic contexts. Within open-source models, NLLB-200 achieved the strongest BLEU (≈48), followed by mBART (≈45), EnViT5 (≈42), and OpusMT (≈35).

These results indicate that Gemini provides the strongest performance among the tested models, while NLLB-200 represents the most competitive open-source alternative.

Table 5. MT performance on the 40-file benchmark set

Model	BLEU	chrF	TER	Remarks
Gemini 2.0 Flash	∼55	∼72	∼38	Best performance in this comparison
NLLB-200 (1.3B)	∼48	∼65	∼45	Strongest open-source option
mBART	∼45	∼62	∼47	Stable, Transformer-based
EnViT5	∼42	∼60	∼50	Vietnamese-specific design
OpusMT	∼35	∼55	∼55	Lowest performance

5.3. **Error Analysis.** To understand model behavior beyond aggregate metrics, we conducted comprehensive error analysis on our benchmark set, examining both quantitative error distributions and qualitative error patterns.

5.3.1. *ASR Error Distribution Analysis.* Our evaluation of 40 IBM Technology videos (15-35 minutes) revealed significant performance variations across STT models. Table 6 presents detailed statistical analysis including standard deviation and error ranges.

Table 6. Detailed ASR performance statistics on benchmark set

Model	Avg WER (%)	Std WER (%)	Min WER (%)	Max WER (%)	SER (%)
Whisper Medium	3.36	2.48	0.90	13.66	100
Whisper Small	5.20	3.12	1.20	15.44	100
Wav2Vec2 Base	61.56	11.97	35.53	88.77	100
Wav2Vec2 Large	64.86	12.43	38.12	91.23	100
HuBERT Large	61.59	10.24	23.59	71.96	100

Key observations:

- Whisper Medium's standard deviation of 2.48% indicates sensitivity to audio quality conditions, with performance ranging from 0.90% (studio recording) to 13.66% (conference audio) - a 14.9× degradation factor.
- All models achieved 100% Sentence Error Rate (SER), indicating that achieving perfect sentence-level transcription remains challenging even for state-of-the-art systems.
- Traditional self-supervised models (Wav2Vec2, HuBERT) showed consistently high error rates (>60% WER), suggesting limited applicability for specialized IT lecture transcription without domain-specific fine-tuning.

5.3.2. *Common ASR Error Patterns.* Manual analysis of transcription outputs identified four primary error categories:

(A) Technical Terminology Errors: Domain-specific IT terms were frequently misrecognized by traditional models. Representative examples include:

- *neuroplasticity* → *neural plasticity* (Whisper Small: word segmentation error)
- *LLM* (Large Language Model) → *mM* or *mMs* (multiple models: acronym misrecognition)
- *Kubernetes* → *communities* (Wav2Vec2: phonetic confusion)

Whisper models demonstrated superior technical vocabulary handling, likely due to their large-scale pretraining on diverse web content including technical discussions and documentation.

(B) Punctuation and Segmentation Issues: The 100% SER across all models primarily resulted from punctuation placement errors rather than word-level mistakes. Examples include missing commas, incorrect capitalization, and difficulty detecting natural sentence boundaries in continuous speech.

(C) Audio Quality Impact: We analyzed the correlation between audio conditions and transcription accuracy:

Table 7. Audio quality impact on transcription accuracy

Audio Condition	WER Increase	Example Cause
Background music	+2.3%	Overlapping frequencies
Multiple speakers	+4.7%	Voice overlap, unclear attribution
Low-quality microphone	+3.5%	Frequency cutoff, distortion
Ambient room noise	+1.8%	Low-level interference
Video compression artifacts	+2.1%	Audio degradation from MP4 encoding

(D) Processing Reliability: Our MongoDB-based pipeline tracked processing status for all 240 files, achieving 100% success rate after implementing audio reconversion strategies for initially corrupted WAV files (8 instances, 3.3% of dataset).

5.3.3. *MT Error Patterns.* Translation error analysis revealed five major categories:

(A) Technical Terminology Inconsistency: The same English technical term was occasionally translated differently within a single lecture. For example, "database" appeared as both "cơ sở dữ liệu" (Vietnamese translation) and "database" (preserved in English) depending on context. Table 8 shows coverage and quality trade-offs across models.

Table 8. MT model coverage and quality comparison

Model	Coverage (%)	BLEU	chrF	TER
Gemini 2.0 Flash	100	55	72.0	38
NLLB-200 (1.3B)	95	46	64.0	46
mBART	100	45	62.0	47
EnViT5	100	42	60.0	50
OpusMT	100	35	55.0	55

(B) Contextual and Semantic Errors: Lower-performing modelshad problems functioning in scenarios involving metaphoric statements. For instance, the neuroscience phrase "neurons that fire together, wire together" was correctly translated by Gemini as "Các nơ-ron kích hoạt cùng nhau, kết nối với nhau" (preserving the metaphor), while OpusMT produced "liên kết với nhau" (a generic "connect together", losing the "wire" metaphor).

(C) Chunking and Segmentation Artifacts: OpusMT's sliding window approach (required due to 512-token architectural limit) introduced redundant phrase repetition at chunk boundaries in 18% of translations and context loss between chunks in 12% of translations.

(D) Model Size vs. Quality Correlation: We observed strong positive correlation (r = 0.89) between model parameter count and BLEU score, ranging from 77M parameters (OpusMT, BLEU 35) to 1.3B parameters (NLLB-200, BLEU 46) and > 100B estimated parameters (Gemini, BLEU 55).

5.4. **Overall Pipeline Integration.** Based on these results, the pipeline that integrates Whisper Medium (ASR) and Gemini 2.0 Flash (MT) delivers relatively stable performance in terms of accuracy and processing speed. When implemented for the entire set of 240 videos, the selected configuration produced transcripts and translations that were accurate and stable enough for practical use in online IT education.

6. Discussion.

6.1. **Performance Insights and Model Selection.** Our comprehensive and detailed evaluation clearly demonstrates that the choice of model has significant implications for translation quality. Specifically, the Whisper Medium model achieved a Word Error Rate (WER) of just 3.36%, a significant 18.3x improvement over the Wav2Vec2 model (WER 61.56%). This significant difference highlights the value of performing weakly supervised large-scale pretraining on highly diverse web data. Furthermore, we observed a significant 14.9x performance degradation, with the same whisper model dropping from an optimal 0.90% to a significantly suboptimal 13.66% when the audio conditions were affected. This result underscores the critical importance of maintaining high audio quality throughout the entire educational content production process.

For machine translation, the Gemini 2.0 Flash model achieved the highest scores (BLEU 55, chrF 72) among all evaluated models. Furthermore, the strong performance of open-source NLLB-200 (BLEU 46, 95% coverage) suggests it as a viable alternative for deployments where API costs, data privacy, or online operation are constrained. The 20-point gap in BLEU between best (Gemini) and worst (OpusMT) models illustrates the substantial quality variation across current MT systems.

6.2. **Cascade Architecture Trade-offs.** The cascade architecture provides several advantages for educational applications:

(1) **Transparency** - the intermediate ASR transcripts allow for careful inspection and manual correction;

(2) **Flexibility** - each component can be upgraded independently;

(3) **Resource Efficiency** - the pipeline takes advantage of existing pretrained ASR and MT models, reducing the need for costly end-to-end training;

However, cascade architectures introduce error propagation risks. As noted by Min et al. [22], errors such as pronunciation disparities or semantic differences in cascaded speech-to-text pipelines are particularly prone to propagate through the system. This reinforces that high-quality ASR is essential for reliable end-to-end translation.

6.3. **Implications for Vietnamese IT Education.** Our results demonstrate the technical feasibility of automated English-Vietnamese lecture translation using current models. Processing the complete 240-video dataset required approximately 2 hours on consumer hardware (Apple M1 MacBook Pro), indicating practical viability for batch translation of online course platforms.

However, several challenges remain: (1) the 100% Sentence Error Rate indicates that sentence-level perfect accuracy is rare even with state-of-the-art systems; (2) terminology inconsistency within long videos may confuse learners; (3) lack of cultural and pedagogical adaptation limits effectiveness for complete beginners. For these reasons, human review remains important for critical educational content.

6.4. **Future Extensions and Research Directions.** This study establishes the foundation for a complete speech-to-speech translation system. Three expansion directions are worth investigating:

Text-to-Speech Integration: Recent Vietnamese TTS advances (e.g., XTTS [15], PhoWhisper TTS [17]) enable high-quality audio synthesis. Integrating TTS would transform MT's output text-based translations into audio, benefiting auditory learners and improving accessibility.

Voice Cloning: Preserving the original lecturer's voice characteristics in translated audio (using models like F5-TTS [16] or VoiceCraft [14]) could enhance engagement and maintain instructor presence, addressing the impersonal nature of synthesized voices.

Lip Synchronization: For video content, aligning lip movements with translated audio would create natural viewing experiences. Recent advances in visual dubbing (e.g., Wav2Lip, SyncNet) show promise but require substantial additional research for robust multilingual application.

More critical future work, beyond technical extensions, involves the following: (1) large-scale human evaluation using Vietnamese IT students to measure comprehension and learning outcomes; (2) domain-specific fine-tuning of models in the domain of IT terminology;

7. **Limitations.** Although this study demonstrated the feasibility of cascade-based English-Vietnamese speech translation for IT education, several limitations must be discussed.

7.1. **Technical and Methodological Limitations.** The evaluation in the paper was limited to single-speaker English audio, while multi-speaker segments (∼15%) were processed without speaker attribution. All experiments were conducted on consumer-grade hardware, thus restricting the use of large-scale models such as NLLB-200 3.3B and excluding real-time performance testing.

Translation quality was evaluated by assessed using automatic metrics including BLEU, chrF and TER. Reference translations were prepared manually from English transcripts, carefully keeping the fidelity and consistency, though large-scale human evaluation was not conducted. The absence of pedagogical adaptation may limit accessibility to beginner-level learners.

7.2. **Limitations of the Dataset.** The dataset is heavily dominated by IBM Technology content (∼68.3%), with studio-quality audio and neutral American accents. Model performance on informal, accented, or non-native English remains untested. Furthermore, the dataset does not include multimodal elements such as slides or code, which limits its applicability to visually oriented or interactive learning materials.

7.3. **Scope Exclusions.** This study implemented a Speech-to-Text-to-Translation pipeline. A full S2ST system would include several other modules, such as TTS, voice cloning, lip synchronization, real-time streaming translation, and adaptive learning interfaces. These aspects remain for further development to add more interactivity and naturalness to multilingual IT education.

7.4. **Generalizability and Ethical Considerations.** The results are only applicable for the language pair of English-Vietnamese, formal IT education, and offline batch processing. Extension to other domains, educational levels, or real-time systems may require further validation.

All the video content was collected from public YouTube channels under Fair Use provisions for nonprofit educational research (17 U.S.C. §107, USA). No original media were redistributed. Only derived textual data were analyzed. Proper attribution was maintained for all source materials (IBM Technology, MIT OpenCourseWare, Columbia University).

8. **Conclusions.** This study has developed and benchmarked a modular English–Vietna-mese speech translation pipeline designed for online IT education. The system integrates Automatic Speech Recognition (ASR) and Machine Translation (MT) components to generate accurate, context-preserving translations of the academic lecture content.

Through quantitative evaluation on 240 educational videos, Whisper Medium achieved the lowest error rates: WER $\approx$ 3.36%, CER $\approx$ 1.57%, while Gemini 2.0 Flash reached the highest translation performance: BLEU $\approx$ 55, chrF $\approx$ 72, TER $\approx$ 38. The proposed combination of these models reveals demonstrates the technical feasibility of cascade-based English–Vietnamese speech translation for educational applications.

The contributions of this research are: establishing a reproducible benchmark of ASR and MT models in a low-resource language setting; providing a standardized dataset and evaluation framework for future studies; and validating the effectiveness of modular integration for domain-specific translation in IT education.

While there are still limitations with respect to real-time performance, speaker diversity, and cultural adaptation, the proposed pipeline provides a pragmatic foundation for advancing accessibility in multilingual learning environments. Future work will implement TTS, voice cloning, lip synchronization, and real-time streaming to move closer to full speech-to-speech translation and to evaluate pedagogical impact via large-scale user assessment.

Acknowledgment. This work was supported by the Scientific Research Support Fund of VNUHCM - University of Information Technology. The authors thank IBM Technology, MIT OpenCourseWare, and Columbia University for making educational content publicly available. All content was used under Fair Use provisions for nonprofit educational research with full attribution to original creators.

REFERENCES

[1] E. Insights, The challenges of being a non-native English speaker in academia, *Editage*, 2023.

[2] A. Radford et al., Robust speech recognition via large-scale weak supervision, *Proc. of the 40th International Conference on Machine Learning (ICML)*, 2022.

[3] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[4] W. N. Hsu, B. Bolte, Y. H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, HuBERT: Self-supervised speech representation learning by masked prediction of hidden units, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[5] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, Multilingual denoising pre-training for neural machine translation, *Transactions of the Association for Computational Linguistics*, 2020.

[6] M. R. Costa-jussà, J. Cross et al., No language left behind: Scaling human-centered machine translation, *arXiv preprint arXiv:2207.04672*, 2022.

[7] C. Ngo, H. Tran, H. Nguyen, N. Minh, L. Phan, T. H. Trinh, and M. T. Luong, MTET: Multi-domain translation for English and Vietnamese, *arXiv preprint arXiv:2210.05610*, 2022.

[8] 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, *Proc. Interspeech*, 2019.

[9] Y. Jia et al., Translatotron 2: High-quality direct speech-to-speech translation with voice preservation, *Proc. of the 39th International Conference on Machine Learning (ICML)*, 2021.

[10] A. Popescu-Belis et al., Speech-to-speech translation pipelines for conversations in low-resource languages, *Proc. of the 20th Machine Translation Summit (MT Summit XX)*, 2025.

[11] L. T. Nguyen, L. T. Nguyen, L. Doan, M. Luong, and D. Q. Nguyen, A high-quality and large-scale dataset for English-Vietnamese speech translation, *Proc. Interspeech*, 2022.

[12] T. T. Le, L. T. Nguyen, and D. Q. Nguyen, Phowhisper: Automatic speech recognition for Vietnamese, *Proc. of the International Conference on Learning Representations (ICLR)*, 2024.

[13] A. Haider, Analyzing the performance of Gemini, ChatGPT, and Google Translate in rendering English idioms into Arabic, *FWU Journal of Social Sciences*, 2023.

[14] P. Peng et al., VoiceCraft: Zero-shot speech editing and text-to-speech in the wild, *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[15] E. Casanova et al., XTTS: A massively multilingual zero-shot text-to-speech model, *Proc. Interspeech*, 2024.

[16] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen, F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching, *Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[17] T. Vu, L. T. Nguyen, and D. Q. Nguyen, Zero-shot text-to-speech for Vietnamese, *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[18] J. Tiedemann and S. Thottingal, OPUS-MT: Building open translation services for the world, *Proc. of the 22nd Annual Conference of the European Association for Machine Translation*, Geneva, 2020.

[19] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, BLEU: A method for automatic evaluation of machine translation, *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, USA, 2002.

[20] M. Popović, chrF: Character n-gram F-score for automatic MT evaluation, *Proc. of the 10th Workshop on Statistical Machine Translation*, Lisbon, Portugal, 2015.

[21] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, A study of translation edit rate with targeted human annotation, *Proc. of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, Cambridge, Massachusetts, USA, 2006.

[22] A. Min, C. Hu, Y. Ren, and H. Zhao, When End-to-End is Overkill: Rethinking Cascaded Speech-to-Text Translation, 2025.