SPEECH-TO-SPEECH TRANSLATION FOR INTERNATIONAL TRAVELERS

Tran Ly Nhat Hao Nguyen Ngoc Canh

CSC15012 — Applications of NLP in Industry
VNUHCM — University of Science

13/4/2026

Business Problem Definition: Context & Motivation

Business Context

As international tourism rebounds, travelers face significant friction due to language barriers in high-paced environments (e.g., ordering food, medical emergencies).

Business Problem Definition: Context & Motivation

Business Context

As international tourism rebounds, travelers face significant friction due to language barriers in high-paced environments (e.g., ordering food, medical emergencies).

- **The Problem:** There are not many translation apps specifically designed for travelers. Traditional translation apps all require users to interact with on-screen buttons during the translation process, which is slow, intrusive, and unnatural during face-to-face interactions.
- **Core Objective:** To provide a hands-free, low-latency bridge for travelers navigating foreign-language environments (e.g., US, Japan).

Business Problem Definition: Context & Motivation

Business Context

As international tourism rebounds, travelers face significant friction due to language barriers in high-paced environments (e.g., ordering food, medical emergencies).

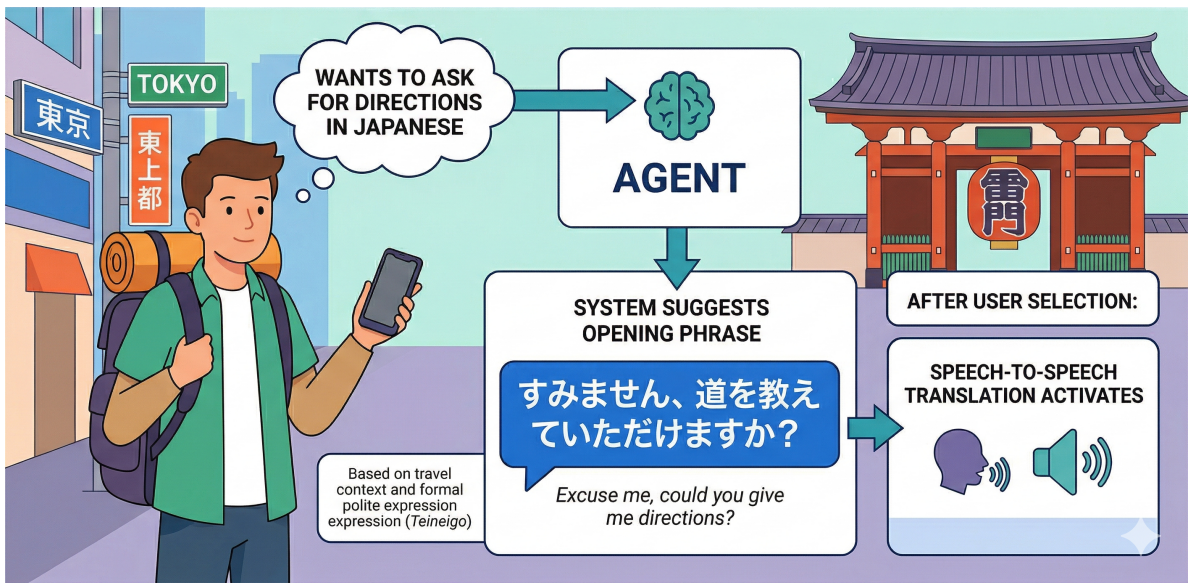


Figure: Illustration of the Agentic AI component providing proactive opening phrase suggestions. (Image generated by Gemini)

Stakeholders & Success Metrics

Target Stakeholders

- **Users:** International travelers and local service providers (shopkeepers, drivers).
- **Admins:** System maintainers responsible for model fine-tuning and resource scaling.

Success Metrics

Metric Type	Indicators
Business	User Satisfaction (CSAT) Cost per Inference Retention Rate
Technical	BLASER (Chen et al., 2023) MOS (Mean Opinion Score) Latency

Feedback Loop

User ratings will serve as a direct signal for the Admin to trigger model updates (SFT/DPO) based on real-world failure cases.

Stakeholders & Success Metrics

Target Stakeholders

- **Users:** International travelers and local service providers (shopkeepers, drivers).
- **Admins:** System maintainers responsible for model fine-tuning and resource scaling.

Success Metrics

Metric Type	Indicators
Business	User Satisfaction (CSAT) Cost per Inference Retention Rate
Technical	BLASER (Chen et al., 2023) MOS (Mean Opinion Score) Latency

Feedback Loop

User ratings will serve as a direct signal for the Admin to trigger model updates (SFT/DPO) based on real-world failure cases.

Development Infrastructure & Tooling

Backend

- **Django 5 + DRF** REST API
- Singleton model loader (`@lru_cache`)
- Eager background loading at startup
- `--noreload` prevents model unloading
- SQLite (lightweight, no external DB)

Frontend

- **Astro 5 + React 19** islands
- Tailwind CSS design system
- Browser `MediaRecorder` API for live mic input
- File upload *and* recording coexist

DevOps

- `Makefile` one-command setup
- Docker Compose deployment
- `ffmpeg` for audio decoding

Development Infrastructure & Tooling

Backend

- **Django 5 + DRF** REST API
- Singleton model loader (`@lru_cache`)
- Eager background loading at startup
- `--noreload` prevents model unloading
- SQLite (lightweight, no external DB)

Frontend

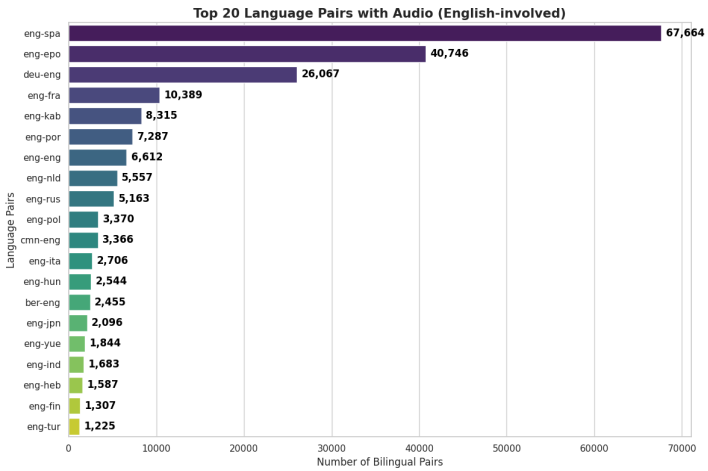
- **Astro 5 + React 19** islands
- Tailwind CSS design system
- Browser `MediaRecorder` API for live mic input
- File upload *and* recording coexist

DevOps

- `Makefile` one-command setup
- Docker Compose deployment
- `ffmpeg` for audio decoding

Data Sources

Source 1: Tatoeba



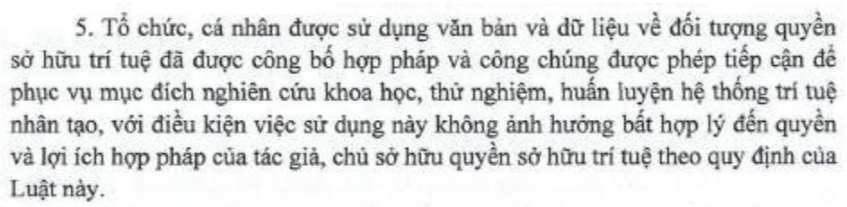
Parallel sentence pairs for multilingual evaluation.

Source 2: YouTube Multilingual Audio

- **Sources:** MrBeast, Mark Rober, NatGeo, Nick DiGiovanni, Jamie Oliver.
- **Processing:** 10–30s chunking or VAD-based pause splitting.
- **High-volume corpus:** Thousands of videos and hours of high-fidelity speech.
- **Multilingual parallelism:** Content in 40+ languages for aligned S2S training.

Legal & Licensing Considerations

- Public online audio does not automatically imply free reuse for model training.
- The project must consider copyright, licensing, and fair-use limitations.
- For a production system, data sourcing should prioritize public, licensed, synthetic, or explicitly permitted datasets.



5. Tổ chức, cá nhân được sử dụng văn bản và dữ liệu về đối tượng quyền sở hữu trí tuệ đã được công bố hợp pháp và công chúng được phép tiếp cận để phục vụ mục đích nghiên cứu khoa học, thử nghiệm, huấn luyện hệ thống trí tuệ nhân tạo, với điều kiện việc sử dụng này không ảnh hưởng bất hợp lý đến quyền và lợi ích hợp pháp của tác giả, chủ sở hữu quyền sở hữu trí tuệ theo quy định của Luật này.

Vietnamese IP law amendment relevant to data reuse.

Model Selection: SeamlessM4T

Why SeamlessM4T?

A single unified model that supports **all four translation modes** — T2T, S2T, T2S, S2S — across 100+ languages, eliminating the need for separate ASR + MT + TTS pipelines.

Model	Size	CPU T2T	GPU T2T
`seamless-m4t-medium`	∼5 GB	1–5 s	< 1 s
`seamless-m4t-large`	∼10 GB	3–15 s	1–3 s

Table: Warm inference latency comparison (text-to-text).

Trade-off: We chose **medium** for demo practicality — loads faster, fits in 8 GB RAM, and still delivers strong translation quality across our 12 supported languages.

Model Selection: SeamlessM4T

Why SeamlessM4T?

A single unified model that supports **all four translation modes** — T2T, S2T, T2S, S2S — across 100+ languages, eliminating the need for separate ASR + MT + TTS pipelines.

Model	Size	CPU T2T	GPU T2T
`seamless-m4t-medium`	∼5 GB	1–5 s	< 1 s
`seamless-m4t-large`	∼10 GB	3–15 s	1–3 s

Table: Warm inference latency comparison (text-to-text).

Trade-off: We chose **medium** for demo practicality — loads faster, fits in 8 GB RAM, and still delivers strong translation quality across our 12 supported languages.

Inference Optimization

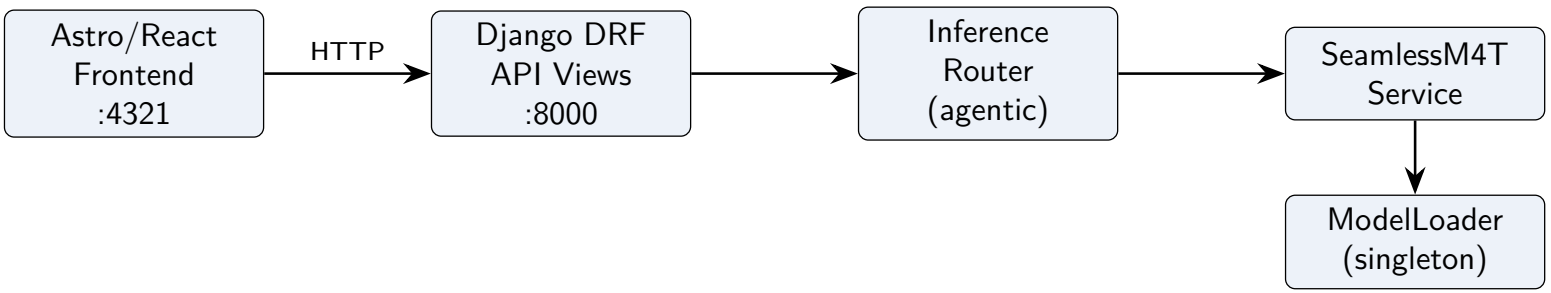
Problem: Naïve integration caused 30–60 s cold starts and redundant computation per request.

Optimizations applied:

1. **Eager loading:** load model in the background at server startup.
2. **–noreload:** prevent Django autoreload from restarting the process.
3. **Single bundle fetch:** retrieve model + processor once per request.
4. **Pre-prepared audio:** preprocess once and reuse across steps.
5. **Conditional `gc.collect()`:** run only on CUDA, skip on CPU.

Result: Warm text-to-text inference reaches **1–5 s** on CPU after a one-time **15–30 s** cold start.

System Architecture & Deployment



Four Translation Modes

- Text → Text
- Speech → Text
- Text → Speech
- Speech → Speech

Deployment Options

- `make dev` (local)
- Docker Compose
- GPU server (CUDA)

Agentic AI Component: Intent-Driven Orchestration

The Role of the Agent

Instead of a passive pipeline, the Agent acts as a **Decision Engine** that orchestrates translation based on user intent and environmental context.

Core Agentic Behaviors:

- **Autonomous Context Routing:** Detect target language and situational context (e.g., dining, emergency, directions).
- **Proactive Suggestion Engine:** Suggest culturally appropriate opening phrases from recognized intent (e.g., "I'm lost").

Backend Implementation

`InferenceRouter` checks model readiness, language support, audio duration, and memory pressure before inference.

Continual Learning Strategy

The Feedback Loop

1. **Collect & Filter:** Pool samples rated as **"Good"** for high-quality SFT data.
2. **Refine "Bad" cases:**
   - **Large models:** regenerate corrected audio/text at scale.
   - **Manual collection:** handle niche or high-priority failures.
3. **Evolve:** Apply **LoRA fine-tuning** on the merged high-quality dataset.
4. **Deploy:** Run **A/B testing** against the baseline before release.

Monitoring Strategy

Monitoring endpoints:

- `/api/health/` — full diagnostics: model status, device, memory, timing.
- `/api/readiness/` — lightweight 200/503 probe for automation and deployment checks.
- Per-request `timing` field — latency breakdown for preprocessing, transcript generation, and translation.

Operational Goal

Detect failures early, track latency drift, and support safer model updates over time.

Data Privacy & Model Robustness

Privacy by Design

- Audio processed in-memory; temp files deleted after each request.
- No persistent storage of user audio beyond generated output.
- Model runs **entirely on-device** — zero third-party API calls.
- All inputs validated at API boundary (serializers + router).

Robustness Strategy

- **English as Pivot:** For weak language pairs (e.g., Vie–Jpn), bridge via English to leverage high-density semantic embeddings.
- **Confidence Thresholds:** Fallback to "Clarification Mode" if pivot uncertainty is high.
- **Input Guards:** Audio duration limit (120 s), file size limit (25 MB), extension validation.

Data Privacy & Model Robustness

Privacy by Design

- Audio processed in-memory; temp files deleted after each request.
- No persistent storage of user audio beyond generated output.
- Model runs **entirely on-device** — zero third-party API calls.
- All inputs validated at API boundary (serializers + router).

Robustness Strategy

- **English as Pivot:** For weak language pairs (e.g., Vie–Jpn), bridge via English to leverage high-density semantic embeddings.
- **Confidence Thresholds:** Fallback to "Clarification Mode" if pivot uncertainty is high.
- **Input Guards:** Audio duration limit (120 s), file size limit (25 MB), extension validation.

Ethics & Fairness: Responsible AI Framework

Ethics Impact Statement

- **Beneficiaries:** International travelers and local businesses.
- **Potential Harm:** Users with strong regional accents or speech impairments may face exclusion if the model is not robust.

Bias & Fairness Risks

- **Language Bias:** Performance gaps between high-resource and low-resource languages.
- **Accent Bias:** Potential inaccuracy across Vietnamese regional dialects.

Mitigation Strategies

Responsible AI Mitigation

1. **PII Anonymization:** Scrub personal information from audio before any admin or model review.
2. **Diverse Training Data:** Use multilingual speech data with broader accent and speaking-style coverage.
3. **Evaluation Protocol:** Track BLASER, MOS, and per-language performance to detect degradation early.

Current Limitations

Current Limitations

- Turn-based translation (**2–15 s** per utterance on CPU), not live streaming.
- Frontend language list is maintained separately from backend config.
- Generated audio is served via Django debug mode; production needs object storage.

Honest Positioning

This system is a **near-real-time, turn-based travel translation assistant**. It is practical for asynchronous interactions where speakers take turns. After warmup, each utterance translates in **2–15 s** on CPU or under **3 s** on GPU.

Future Work

Planned Improvements

- **WebSocket streaming** for progressive output delivery.
- **INT8 quantization** for faster CPU inference.
- **Speaker diarization** for multi-participant scenarios.
- **Audio chunking** for inputs beyond the 120 s cap.
- **Multi-session support** for concurrent translations.

Longer-Term Direction

Move from turn-based translation toward more responsive, scalable, and multi-user speech translation in realistic travel settings.