System Description for the TOPI Challenge: A Cascaded Approach to Pragmatic-Aware Speech Translation

Yuchen Song, Satoshi Nakamura
The Chinese University of Hong Kong, Shenzhen

January 16, 2026

Abstract

This paper describes our submission to the Transfer of Pragmatic Intent in Speech-to-Speech Translation (TOPI) challenge. We propose a cascaded Speech-to-Speech Translation (S2ST) system designed to translate Spanish source speech into English target speech while attempting to preserve pragmatic information. Our pipeline integrates *Whisper Large v3* for Automatic Speech Recognition (ASR), *Qwen2.5-72B-Instruct* for Machine Translation (MT), and *FishAudio/OpenAudio-S1-Mini* for Text-to-Speech (TTS) synthesis. While we initially experimented with projecting pragmatic vectors into the TTS model via a multi-layer perceptron (MLP) adapter, we encountered significant challenges related to the stability of the base model during fine-tuning. This report details our architecture, the rationale behind our model selection, our specific adaptation attempts using soft prompts, and an analysis of the training difficulties attributed to the base model's reinforcement learning alignment.

1 Overview of the System

Our submitted system operates as a three-stage cascade. Unlike end-to-end models which often prioritize semantic accuracy over prosodic fidelity, our modular approach allows us to select state-of-the-art components for each modality.

The workflow is as follows:

1. **ASR Module:** The source Spanish audio is transcribed into text.
2. **MT Module:** The transcribed Spanish text is translated into English text.
3. **TTS Module:** The English text is synthesized into speech. Crucially, this stage uses the original Spanish audio and the transcribed Spanish text as prompts to guide the prosody and timbre of the generated English audio.

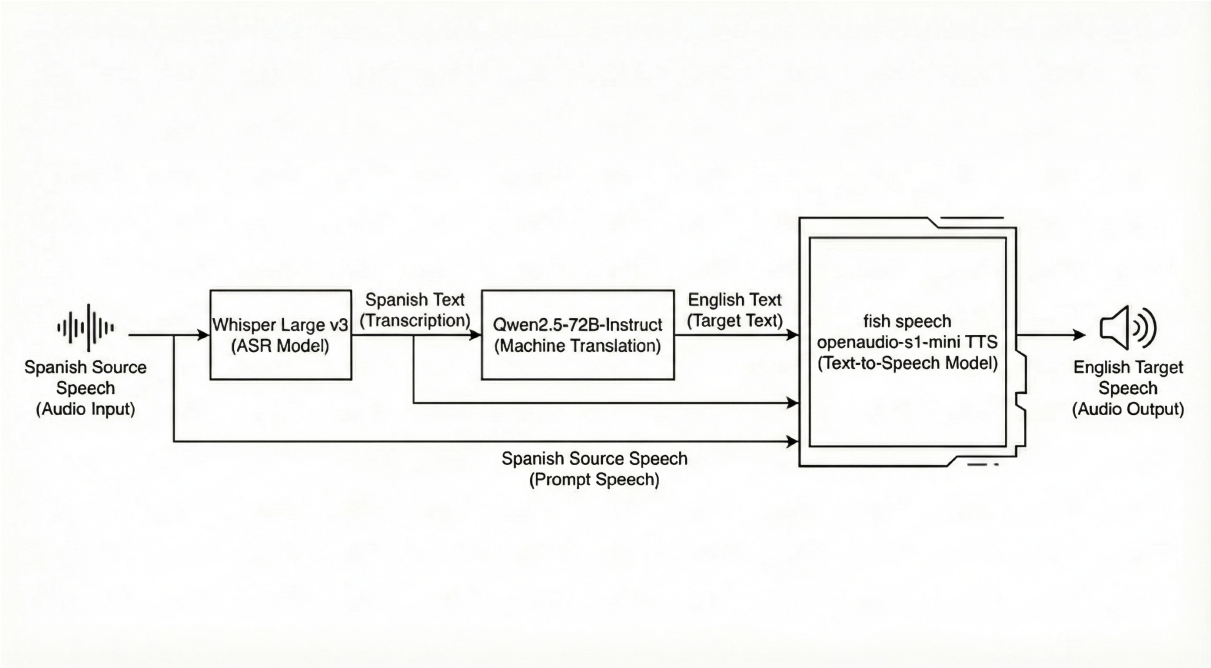


Figure 1: Block diagram of the proposed Cascaded Speech-to-Speech Translation System.

2 Starting Point: Base Models and Architecture

We initially considered fine-tuning the *SeamlessM4T* system. However, we observed that Seamless primarily optimizes for semantic precision in its decoder, often neglecting the cloning of prosody, timbre, and paralinguistic features. Given the competition's focus on pragmatic information, we pivoted to a cascade of specialized models:

2.1 ASR: Whisper Large v3

For the first stage, we utilize OpenAI's **Whisper Large v3**. It is currently widely regarded as the most robust open-source model for multilingual transcription, ensuring high-fidelity capture of the source Spanish content.

2.2 MT: Qwen2.5-72B-Instruct

For text translation, we selected **Qwen2.5-72B-Instruct**. Among open-source Large Language Models (LLMs), it offers an optimal balance between model size and translation performance, specifically for Spanish-to-English tasks.

2.3 TTS: FishAudio OpenAudio-S1-Mini

The core of our generation strategy relies on **FishAudio/OpenAudio-S1-Mini**. We selected this model for several reasons:

- **LLM-based Architecture:** Built upon the Llama backbone, it possesses strong cross-lingual semantic understanding capabilities, trained on hundreds of thousands of hours of multilingual data.
- **Pragmatic Alignment via GRPO:** The model has undergone post-training alignment using **Group Relative Policy Optimization (GRPO)**. Unlike standard next-token prediction which focuses on local accuracy, GRPO optimizes the model towards global,

sequence-level rewards derived from speaker verification systems. Specifically, the reward function prioritizes high cosine similarity between the generated and prompt audio embeddings. To maximize this reward in a cross-lingual setting, the policy is implicitly incentivized to clone not only the static vocal timbre but also dynamic **pragmatic behaviors** (such as prosodic rhythm, stress patterns, and speaking rate). This reinforcement learning process effectively "forces" the model to retain the source speaker's behavioral identity across language boundaries to satisfy the similarity objective.

3 Adaptations: Pragmatic Vector Injection

To explicitly incorporate the 103-dimensional pragmatic vector provided by the competition task, we designed an adaptation module to inject this information into the TTS generation process.

3.1 MLP Projector

We hypothesized that a simple linear mapping would be insufficient to capture the non-linear relationship between the pragmatic feature space and the semantic space of the LLM. Therefore, we implemented a two-layer Multi-Layer Perceptron (MLP) projector:

$$\mathbf{h}_1 = \text{GELU}(\mathbf{W}_1\mathbf{v} + \mathbf{b}_1) \tag{1}$$

$$\mathbf{e}_{raw} = \mathbf{W}_2\mathbf{h}_1 + \mathbf{b}_2 \tag{2}$$

Where:

- $\mathbf{v} \in \mathbb{R}^{103}$ is the input pragmatic vector.
- $\mathbf{W}_1 \in \mathbb{R}^{512\times103}$ projects the features to a higher dimension to capture interactions.
- $\mathbf{W}_2 \in \mathbb{R}^{D\times512}$ maps the intermediate representation to the model's hidden dimension size D.

3.2 Virtual Token Extension (Soft Prefix)

To prevent the pragmatic signal from being overwhelmed by long text sequences, we adopted a strategy inspired by "Prefix-Tuning." The projected vector is reshaped/transformed into K virtual tokens (where $K = 5$). This allows the model to accept the 103-dimensional floating-point vector as a forceful *soft prefix* injected directly into the transformer's attention mechanism, theoretically guiding the generation style.

4 Training and Tuning Challenges

Despite the theoretical soundness of our architecture, we encountered significant challenges during the training phase, leading to the decision to submit the zero-shot cascade results while documenting our fine-tuning attempts as a negative result analysis.

4.1 Data and Setup

We utilized the competition-provided dataset, which consists of a few hours of speech. Our goal was to fine-tune the S1-Mini model to respect the injected pragmatic vectors.

4.2 The Failure of LoRA and GRPO Sensitivity

Our initial attempt used Low-Rank Adaptation (LoRA) to avoid catastrophic forgetting given the small dataset size. However, we discovered that the base model's weights, having been refined via Group Relative Policy Optimization (GRPO), were extremely "sharp."

The GRPO process likely resulted in a parameter distribution that is highly sensitive to perturbation. Although LoRA only updates a low-rank subset of weights, these slight behavioral changes were sufficient to disrupt the "correct reasoning/generation modes" established by the reinforcement learning stage. This resulted in:

- **Mode Collapse:** The model quickly lost the ability to generate coherent speech.
- **Catastrophic Forgetting:** After only a few training steps, the model's generalization capability vanished.

4.3 Full Fine-Tuning and Degradation

We subsequently attempted full fine-tuning, hypothesizing that global weight updates might "smooth" the transition better than LoRA. While this restored the ability to output meaningful speech, performance remained significantly degraded compared to the pre-trained checkpoint.

A persistent issue was observed: **Length Generalization Failure**. As sentence length increased, the intelligibility of the generated audio dropped sharply towards the end of the utterance.

4.4 Control Experiment

To isolate the cause, we removed the pragmatic adapter and attempted to fine-tune the model on clean Librispeech data. The same degradation pattern (loss of intelligibility at the end of long sentences) persisted. We attribute this to the inherent difficulty of fine-tuning this specific model architecture without access to the original closed-source training stack or maintenance support.

5 Other Aspects

The system relies on the following configuration:

- **Prompting Strategy:** We use the source Spanish audio as the acoustic prompt and the source Spanish text as the context prompt for the TTS model to encourage cross-lingual prosody transfer.
- **Hyperparameters:** For the MLP adapter, hidden dimension = 512, Virtual Tokens $K = 5$.

6 Conclusions and Future Work

We have presented a cascaded system for Spanish-to-English speech translation. Our zero-shot system leverages the strong priors of Whisper, Qwen, and FishAudio to achieve reasonable quality. However, our attempts to explicitly control pragmatics via vector injection and fine-tuning were hindered by the high sensitivity of the RL-aligned base model, leading to degradation in stability and intelligibility.

Future Work: Given the lack of maintenance and fine-tuning stability in the current TTS backbone, we plan to migrate our development to **CosyVoice 3**. CosyVoice 3 offers a complete

training codebase, making it highly amenable to fine-tuning. Furthermore, its proven cross-lingual capabilities suggest it will be a more robust candidate for integrating pragmatic control vectors in future iterations of this task.

7 Experiments and Results

We randomly selected 30 samples from the dral dataset as the test set.

7.1 Objective Evaluation

Table 1 presents the objective evaluation results of different systems compared to the Ground Truth (GT) audio. We employ ECAPA-TDNN to measure Speaker Verification (SV) cosine similarity, Pragmatic Similarity by calculating the cosine similarity of pragmatic vector, and Word Error Rate (WER) to evaluate content accuracy.

Table 1: Objective evaluation results. SV denotes speaker similarity based on ECAPA-TDNN. Best results (excluding GT) are highlighted in **bold**.

System	**SV** (↑)	**Pragmatic Sim** (↑)	**WER** (↓)
GT Audio	0.391	-	0.147
Fish Speech	0.360	**0.800**	0.171
CosyVoice 3 (Base)	**0.430**	0.665	0.074
CosyVoice 3 (RL)	0.410	0.589	**0.046**

7.2 Analysis of RL Finetuning Strategies

A critical observation from Table 1 is the divergence in performance trade-offs between different model variants, particularly regarding the impact of Reinforcement Learning (RL) finetuning.

The Stability-Expressiveness Trade-off in CosyVoice 3: Comparing the CosyVoice 3 Base model with its RL-finetuned version, we observe a significant improvement in intelligibility, with WER dropping from 0.074 to 0.046. However, this gain comes at the cost of paralinguistic fidelity, as indicated by the drop in Pragmatic Similarity from 0.665 to 0.589.

We attribute this phenomenon to the specific optimization objective of the RL stage in CosyVoice 3. The reward function is designed to maximize robustness and ASR confidence to prevent common autoregressive artifacts (e.g., repetition, hallucination). This creates an "alignment tax": while the model becomes more stable and precise in content generation, the RL process suppresses high-variance prosodic features—such as hesitations, varying speech rates, or emotional stressors—which are essential for high pragmatic similarity but are often penalized as "instability" by standard ASR-based reward models.

Contrast with Fish Speech: In contrast, Fish Speech achieves the highest Pragmatic Similarity (0.800) but suffers from a higher WER (0.171). This is because its alignment strategy prioritizes the retention of dynamic prosody and instruction following over strict content stability.

These results demonstrate that the choice of reward function in RL finetuning acts as a pivotal lever, steering the model either towards industrial-grade stability (CosyVoice 3 RL) or expressive pragmatic preservation (Fish Speech/Base models).