Why We Need Speech to Evaluate Speech Translation

Anonymous ACL submission

Abstract

Speech translation models are increasingly capable of preserving speech-specific information (e.g., speaker gender, prosody, and emphasis), yet evaluation metrics remain blind to such phenomena. We meta-evaluate both text- and speech-based quality estimation metrics on two contrastive datasets targeting gender agreement and prosody, and find that both fall short, even when given direct access to the speech signal. We then train SpeechCOMET, a family of quality estimation models with speech encoders, and evaluate a state-of-the-art SpeechLLM as a judge. Both match or exceed text-based COMET on standard quality estimation, but neither consistently assesses speech-specific phenomena. We identify three causes: (1) speech-specific features are not reliably preserved in current encoders, (2) models tend to ignore the speech source signal, and (3) quality estimation training data contains too few relevant examples. We release all models and code, and argue that progress requires dedicated speech-specific training data and models that genuinely condition on speech.

1 Introduction

Speech translation is evolving from cascaded pipelines, which first transcribe speech before translating the text, to end-to-end models (Papi et al., 2026) that process speech directly and can, in principle, preserve speech-specific information such as speaker gender, prosody, and emphasis. The correct translation of an utterance can depend on such cues: the same words carry different meanings depending on the speech, as shown in Figure 1.

Although cases where the speech signal is decisive may represent only a small fraction of all translations, the quality of speech translation has improved substantially recently (Abdulmumin et al., 2025) to the point where such phenomena can no longer be neglected. With the rise of Speech Large Language Models (SpeechLLMs) (Xu et al., 2025; Microsoft et al., 2025) and end-to-end ST models (Communication et al., 2023; Radford et al., 2023; Ambilduke et al., 2025), recent work started to evaluate speech translation models on speaker gender agreement (Züfle and Niehues, 2025; Conti et al., 2026), emotion (Papi et al., 2026) and prosodic phenomena (Li et al., 2026), using dedicated contrastive benchmarks such as MuST-SHE (Bentivogli et al., 2020), EmotionTalk (Sun et al., 2025) and ContraProST (Tsiamas et al., 2024).

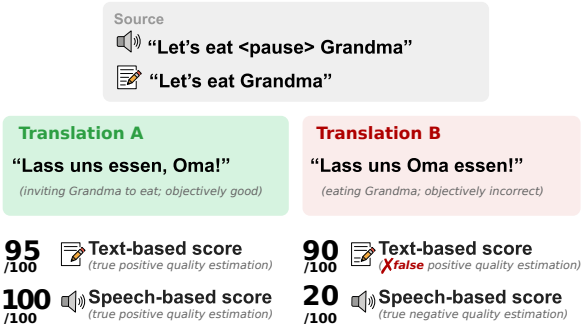

Figure 1: Speech is necessary to correctly evaluate speech translation quality: the prosodic break in the source speech disambiguates two translations.

Yet dedicated quality estimation (QE) metrics for assessing these phenomena are missing. Current evaluations rely on text-based QE models such as COMETKiwi (Rei et al., 2022), which are blind to speech-specific phenomena by construction (Figure 1). While speech-aware metrics such as BLASER (Dale and Costa-jussà, 2024) and SpeechQE (Han et al., 2024) have been proposed, they have only been evaluated on standard correlation with human scores, leaving open whether they can actually capture speech-specific phenomena.

We investigate whether existing and new metrics can close this gap. We show that existing speech-aware metrics, including BLASER and SpeechQE, are not sensitive to speech-specific phenomena despite receiving speech as input. We then introduce and train SpeechCOMET, a family of QE models

with SONAR (Duquenne et al., 2023) and Whisper (Radford et al., 2023) encoders, and additionally evaluate a state-of-the-art SpeechLLM as a judge. SpeechCOMET matches text-based COMETKiwi on general QE performance, and SpeechLLM exceeds it. However, neither SpeechCOMET nor SpeechLLM reliably assess speech-specific phenomena. We find three complementary explanations: (1) speech-specific features may not be preserved in the encoder representations at all; (2) models tend to ignore the speech source and focus on the translation; and (3) even when features are present in the encoder, the QE training signal, which contains very few speech-specific examples, is insufficient to teach the model to use them.

Our contributions are threefold: (i) we benchmark state-of-the-art metrics on speech-specific evaluation; (ii) we train and evaluate new speech-aware QE models; and (iii) we provide hypotheses and evidence of why current models fail.

We release the trained QE models and code, and argue that progress in speech translation evaluation requires moving beyond text-based proxies and investing in speech-specific benchmarks and data.

2 Current Evaluation of Speech Translation

We review existing approaches to speech translation quality estimation (QE).

2.1 Current Approaches

Text-Based Metrics. The dominant approach to evaluating speech translation quality relies on text-based QE models. For shared tasks and in recent research, WMT22 COMETKiwi DA (Rei et al., 2022) and its reference-based variant have been the standard metrics to date (Abdulmumin et al., 2025; Papi et al., 2026), applied to human reference transcripts as source input. Two practical challenges arise in this setting. First, human transcripts are rarely available. Alternatives such as automatic speech recognition (ASR) transcripts and back-translations have been explored by Cettolo et al. (2025), showing that ASR transcripts are more effective than back-translations when word error rate is low. Second, beyond source representation, standard QE models assume perfectly segmented inputs, whereas speech translation usually relies on automatic segmentation (Amrhein and Haddow, 2022; Abdulmumin et al., 2025). COMET Partial (Zouhar et al., 2025) addresses this by training on segmentation-noisy data. Yet none of these adaptations have speech access.

Speech-Based Metrics. A complementary line of work proposes metrics that operate directly on speech, avoiding ASR dependence altogether. BLASER (Chen et al., 2023) and its successor Blaser 2.0 Qe (Dale and Costa-jussà, 2024) encode the source and translation directly into a shared multimodal embedding space using SONAR encoders (Duquenne et al., 2023), originally targeted at speech-to-speech translation but extended to speech-to-text. SpeechQE TowerInstruct 7B En2de (Han et al., 2024) takes a dedicated approach, combining Whisper (Radford et al., 2023) with an LLM to estimate speech-to-text translation quality. However, BLASER and SpeechQE are evaluated only on correlation with human scores or reference-based metrics, with no evaluation on benchmarks targeting speech-specific phenomena.

We now benchmark existing metrics on both standard correlation with human judgements and on datasets targeting speech-specific phenomena. We describe the datasets, metrics, and results in turn.

2.2 Data

IWSLT 2026 Metrics dev set. We evaluate on the development set of the IWSLT 2026 Metrics Shared Task IWSLT2026 Metrics Shared Train Dev (Adelani et al., 2026), which is, to the best of our knowledge, the only speech translation dataset with human-annotated scores[1], newly released after the existing metrics were developed. It comprises ACL talks from IWSLT 2025 submissions for en-de and en-zh. The dev set contains 5.56K segments annotated with human direct assessment scores across four speech translation systems.

The source texts in this dataset are human-annotated transcripts, which are rarely available in practice. We therefore transcribe the source speech with Whisper Large V3 (Radford et al., 2023) at a WER of 24.4%, and evaluate text-based models with both human- or ASR source texts, to evaluate potential degradations (Cettolo et al., 2025).

Beyond correlation with human scores, we evaluate whether QE metrics are sensitive to speech-specific phenomena. We use two contrastive datasets that test linguistically-motivated distinctions only accessible from the speech signal:

[1]The human annotations from IWSLT 2023 are also public, we use them as part of the train set in later experiments.

		IWSLT dev								MuST-SHE			ContraProST		
		Segment τ_b (%)				System SPA (%)				PA (%)			PA (%)		
		de	**zh**	**avg**	**ASR src**	**de**	**zh**	**avg**	**ASR src**	**es**	**fr**	**it**	**de**	**es**	**ja**
TEXT	COMET-PARTIAL	11.2	12.7	11.9	-0.3	43.1	67.0	55.0	+4.7	52.4	55.7	51.8	50.0	50.0	50.0
	COMETKIWI	**32.8**	**36.4**	**34.6**	-6.0	**86.3**	**89.1**	**87.7**	-0.7	**61.5**	**60.6**	**55.2**	50.0	50.0	50.0
SPEECH	SPEECHQE	26.9	32.7	29.8	–	79.2	71.3	75.2	–	37.0	36.0	31.0	26.1	20.0	38.2
	BLASER	22.0	26.8	24.4	–	85.5	67.4	76.5	–	52.0	51.6	51.7	**51.0**	**51.5**	**51.1**

Table 1: Results for transcript- and speech-based baseline metrics. Transcript-based metrics cannot evaluate speech-specific phenomena by design; speech-based metrics can access the signal but still fail on gender and prosody.

MuST-SHE. With MuST-SHE (Bentivogli et al., 2020), we evaluate whether metrics are sensitive to speaker gender agreement in speech translation. The dataset contains English speech translated to French, Spanish, and Italian, which inflect grammatical forms according to the speaker's gender. Each of the 1612 evaluation pairs consists of a gender-correct translation and a gender-incorrect alternative. A metric that uses speech should prefer translations that match the speaker's gender.

ContraProST. ContraProST (Tsiamas et al., 2024) tests sensitivity to prosodic and paralinguistic features: correct translations are paired with prosodically incorrect alternatives across en→{de, es, ja} (15,972 pairs), covering five phenomena: sentence stress, intonation, prosodic breaks, emotional prosody, and politeness. These distinctions are not recoverable from transcripts alone, making this a direct test of whether speech-based metrics capture information beyond text.

2.3 Meta-Evaluation Metrics

We adopt the meta-evaluation metrics from the IWSLT 2026 Metrics Shared Task[2], assessing our QE systems at both the segment and system level. For text-based models, we additionally evaluate the effect of using automatic transcripts as source inputs, which may contain ASR errors.

Segment-level correlation. For each source segment, we compute Kendall's τ_b (Kendall, 1945) between metric predictions and human direct assessment scores across the available system translations, and then average this across source segments. A higher τ_b indicates better agreement between the metric and human ranking of translations.

System-level correlation. At the system level, we evaluate the metric's ability to rank speech translation systems over a full dataset. We use Soft Pairwise Accuracy (SPA, Thompson et al., 2024), which measures agreement between human and automatic pairwise system rankings. SPA is 1 when the metric assigns identical pairwise confidence to every system comparison as human judgements.

Contrastive pairwise accuracy. For contrastive evaluation datasets such as MuST-SHE and ContraProST, no human scores are available. Instead, each source is paired with one correct and one incorrect translation. We measure whether the metric assigns a higher score to the correct translation:

$$\text{PA} = \frac{\sum_{k=1}^{N} \mathbf{1}[\hat{y}_k^+ > \hat{y}_k^-]}{N} \quad (1)$$

where N is the number of contrastive pairs, $\hat{y}_k^+$ is the metric score for the correct translation, and $\hat{y}_k^-$ is the score for the incorrect one. PA equals 1 when the metric always prefers the correct translation, and ties are counted as incorrect.

2.4 Evaluating Existing Metrics

We benchmark existing approaches on the datasets described above; results are shown in Table 1. Transcript-based metrics outperform speech-based ones when given human reference transcripts, but degrade noticeably under ASR input: COMET drops from 32.8% to 27.8% at the segment level for en-de and from 36.4% to 29.4% for en-zh. With ASR transcripts, transcript-based and speech-based metrics perform comparably.

For evaluations targeting speech-specific phenomena, neither family performs well. Transcript-based metrics score at chance on MuST-SHE and ContraProST by design, as they have no access to the speech signal. More surprisingly, speech-based metrics also do not perform well: despite receiving speech as input, SpeechQE and BLASER score at or below chance on both datasets.

3 New Approaches to Speech Translation Evaluation

Having established that existing metrics are not sensitive to speech-specific phenomena, we investigate

[2] iwslt.org/2026/metrics

whether this limitation is architectural. We train SpeechCOMET, an adaptation of COMETKiwi with a speech encoder, and evaluate a state-of-the-art SpeechLLM as a judge to test whether either approach can capture speech-specific phenomena.

3.1 SpeechCOMET

Architecture. Reference-free text-based QE models (Rei et al., 2022) encode the source sentence and the hypothesis separately using a shared multilingual encoder, and feed their representations into an MLP estimator. Concretely, let $\mathbf{s}_t \in \mathbb{R}^d$ and $\mathbf{h}_t \in \mathbb{R}^d$ denote the pooled source and hypothesis embeddings, respectively. The estimator receives the concatenation of four interaction features:

$$\mathbf{e} = \left[\mathbf{h}_t;\, \mathbf{s}_t;\, |\mathbf{h}_t - \mathbf{s}_t|;\, \mathbf{h}_t \odot \mathbf{s}_t\right] \in \mathbb{R}^{4d}, \qquad (2)$$

and produces a scalar quality score $\hat{y} = \mathrm{MLP}(\mathbf{e})$, where the MLP is a two-layer feed-forward network with Tanh activations.

In SpeechCOMET, we replace the text source encoder with a speech representation $\mathbf{s}_a \in \mathbb{R}^d$ that is projected to match the text encoder dimension d via a learned linear layer. The hypothesis is still encoded by a text encoder, and the same four-way interaction is applied. Two alternative fusion strategies are described and evaluated in Appendix B.1. They perform worse than the four-way interaction.

We further extend SpeechCOMET to take *both* the source speech and source text as input. The two source representations are fused into a single embedding $\mathbf{s} = f(\mathbf{s}_t, \mathbf{s}_a)$, where f is one of three strategies: element-wise average, element-wise sum, or a learned linear projection applied to their concatenation $[\mathbf{s}_t;\, \mathbf{s}_a] \in \mathbb{R}^{2d}$. The fused embedding $\mathbf{s}$ then replaces $\mathbf{s}_t$ in Eq. (2).

Speech and Text Embedding Models. We encode the hypothesis text using XLM RoBERTa Base (Conneau et al., 2019), a multilingual transformer model, choosing the base model over the larger InfoXLM-large (Chi et al., 2021) used by COMETKiwi (Rei et al., 2022) for efficiency. To enable a direct comparison at equal encoder capacity, we additionally train our best model with InfoXLM-large (SpeechCOMET^{XL}). We also train a text-only baseline using the smaller encoder to isolate the contribution of speech input.

For source speech, we investigate two encoders. SONAR (Duquenne et al., 2023) is a multilingual speech encoder that produces fixed-size sentence-level embeddings. Whisper Large V3 (Radford et al., 2023) outputs a variable-length sequence of frame-level representations. For the latter, we use only the encoder component and evaluate two aggregation strategies: average pooling and attention-based pooling over the encoder output.

Training Details. For both speech encoders, we experiment with frozen and fine-tuned variants. For SONAR, we partially fine-tune by freezing the bottom four transformer layers and updating the remainder, motivated by probing studies showing that phonetic information is concentrated in the middle to upper-middle layers (Gubian et al., 2025; Choi et al., 2026). For Whisper, full fine-tuning is memory-prohibitive, so we instead use LoRA (Hu et al., 2021) while keeping the base weights frozen. We train on a single NVIDIA A100-SXM4-40GB GPU. Hyperparameters are listed in Table 4.

Data. All models are trained on source-hypothesis pairs annotated with human direct assessment scores, where the source is either text or speech depending on the model variant. The IWSLT2026 Metrics Shared Train Dev dataset (Adelani et al., 2026) provides paired source speech and source text alongside human quality scores, comprising 229 hours of audio across 33,721 samples. As the IWSLT training set is relatively small, we investigate whether augmenting with text-quality-score pairs from Wmt Human All dataset (Kocmi et al., 2025) improves performance. The two sets differ in domain, as WMT covers text translation with gold segmentation, while IWSLT covers speech translation with automatic segmentation. WMT provides only the source text, so we synthesise speech for the 490,943 WMT samples using Kokoro 82M TTS to obtain 1,114 hours of paired audio-quality-score examples. Language pairs and further dataset details are provided in Appendix A.3. We continue to evaluate all models on the test sets described in Section 2.2.

Model Variants. We train the following model variants. *Text models* are trained on two datasets: WMT direct assessment data, which provides a large amount of training signal, and the IWSLT dataset using source texts. The latter enables a fair comparison with speech models, which are trained on IWSLT only. *Speech models* use source speech encoded by SONAR or Whisper and are trained in frozen and fine-tuned encoder variants. *Speech+text models* take both modalities as input and are trained with three fusion strategies (average,

		IWSLT dev								MuST-SHE			ContraProST		
		Segment τ_b (%)				System SPA (%)				PA (%)			PA (%)		
		de	**zh**	**avg**	**ASR src**	**de**	**zh**	**avg**	**ASR src**	**es**	**fr**	**it**	**de**	**es**	**ja**
TEXT	COMETKIWI	**32.8**	**36.4**	**34.6**	-6.0	**86.3**	**89.1**	**87.7**	-0.7	**61.5**	**60.6**	**55.2**	**50.0**	**50.0**	**50.0**
SpeechCOMET															
TEXT	$\text{COMETKiwi}^{\text{WMT}}_{\text{RoBERTa}}$	20.3	25.9	23.1	-1.5	44.7	67.4	56.0	+1.5	51.2	52.8	50.1	**50.0**	**50.0**	50.0
	$\text{COMETKiwi}^{\text{IWSLT}}_{\text{RoBERTa}}$	20.2	25.3	22.8	-0.5	69.9	67.4	68.7	-1.8	52.8	49.1	52.0	**50.0**	**50.0**	50.0
SPEECH	$\text{SpeechCOMET}_{\text{SONAR}}$	17.3	22.9	20.1	–	51.0	**70.9**	61.0	–	53.1	50.9	50.9	**50.0**	**50.0**	50.0
	$\text{SpeechCOMET}_{\text{Whisper}}$	16.8	18.7	17.8	–	44.9	68.5	56.7	–	51.4	50.7	**52.8**	**50.0**	**50.0**	**50.1**
SP.+TXT	SpeechCOMET	24.5	27.1	25.8	-3.8	79.7	67.4	73.5	-1.4	**55.6**	52.6	52.4	**50.0**	49.9	50.0
SP.+TXT	$\text{SpeechCOMET}^{\text{XL}}$	**32.7**	**36.0**	**34.4**	-6.0	**85.2**	67.8	**76.5**	+6.5	54.6	**58.0**	51.1	**50.0**	**50.0**	50.0
SpeechLLM															
TEXT	SpeechLLM	33.2	47.5	40.4	-5.5	**95.6**	32.6	64.1	+2.1	23.7	17.3	16.2	23.8	18.1	31.1
	+FT	39.3	55.0	47.1	-3.8	93.4	**83.7**	**88.5**	-3.4	14.0	13.8	15.4	14.0	14.0	26.1
SPEECH	SpeechLLM	26.5	37.5	32.0	–	87.5	32.7	60.1	–	38.0	32.3	32.4	**32.2**	**28.4**	**33.1**
	+FT	33.8	46.5	40.1	–	90.2	50.1	70.1	–	35.9	15.0	20.7	21.1	21.9	15.7
SP.+TXT	SpeechLLM	32.2	44.1	38.1	-4.5	90.7	32.6	61.6	-0.6	**43.1**	**37.1**	**37.8**	31.0	26.1	32.6
	+FT	**39.7**	**60.1**	**49.9**	-4.6	89.1	72.1	80.6	-7.2	8.2	6.3	6.8	17.8	8.8	20.4

Table 2: Results for SpeechCOMET and SpeechLLM models. ASR src shows the change in average τ_b/SPA when using ASR transcripts instead of human transcripts; speech-only models have no text input (–).

sum, projection). We additionally train speech+text models initialised from a text checkpoint, to investigate whether a strong text initialisation benefits multimodal models. Finally, we train *TTS-augmented models* on WMT data with synthesised speech, as well as combined IWSLT+WMT-TTS variants, to study the effect of training data size and domain on speech model quality.

3.2 SpeechLLM as Judge

As a second approach we evaluate Qwen2.5 Omni 7B (Xu et al., 2025), a multimodal large language model capable of processing text, audio, and both modalities simultaneously, as a zero-shot and fine-tuned quality estimator. Unlike SpeechCOMET, which is trained specifically for QE, the LLM is prompted to produce a scalar quality score directly, without task-specific architectural adaptations.

We evaluate three input modalities, text only, speech only, and speech+text, mirroring the SpeechCOMET setup. In the *zero-shot* setting, the model is prompted with each modality using a standard QE prompt as well as two task-specific prompt variants for MuST-SHE and ContraProST (see Figure 4). In the *fine-tuned* setting, we train one model per modality using LoRA adapters on the IWSLT dataset described earlier. Fine-tuning and inference are performed on a single NVIDIA A100-SXM4-40GB GPU. Hyperparameters and prompts are listed in Appendix A.2.

4 Results

Results for all model families are given in Table 2.

4.1 SpeechCOMET Results

Text-only: in-domain training data outweighs scale. We train two text-only SpeechCOMET variants: one on large-scale WMT text translation data ($\text{COMETKiwi}^{\text{WMT}}_{\text{RoBERTa}}$) and one on IWSLT speech translation system outputs ($\text{COMETKiwi}^{\text{IWSLT}}_{\text{RoBERTa}}$). Despite WMT being substantially larger, $\text{COMETKiwi}^{\text{IWSLT}}_{\text{RoBERTa}}$ achieves much higher system-level correlation (68.7 vs. 56.0 SPA avg), while segment-level performance is comparable (22.8 vs. 23.1 τ_b avg). As expected for text-only models, neither variant shows sensitivity to gender (MuST-SHE) or prosody (ContraProST).

Speech-only: limited sensitivity to speech-specific phenomena. Next, we replace the text encoder with a speech encoder, using either SONAR or Whisper, to encode the source speech directly. For each encoder, we experiment with freezing vs. unfreezing the encoder during fine-tuning and with augmenting training data with a TTS-synthesized variant of WMT data (see Section 3). The best-performing variant of each is reported in Table 2, with the full ablation in Table 7 in Appendix B.2. Both encoders benefit from unfreezing, but only Whisper benefits from TTS pretraining. Unlike SONAR, Whisper is an ASR model with no native sentence-level representation and requires additional speech exposure to adapt.

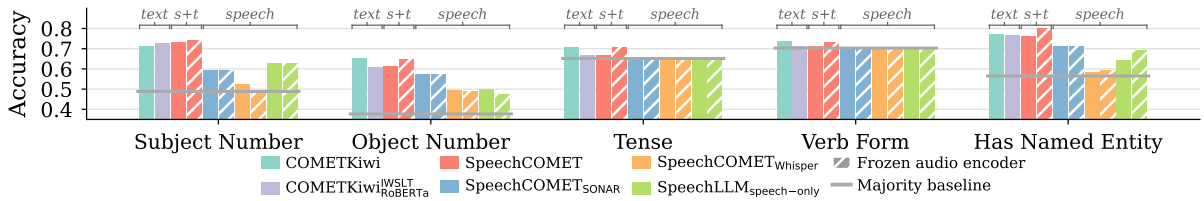

Figure 2: Probing accuracy for linguistic features on the IWSLT 2026 Metrics dataset. Scores averaged over 3 runs (std. ≤ 0.02). *Text-aware* systems consistently outperform *speech-only* systems. Fine-tuned variants (hatched) closely track their frozen counterparts (solid). SONAR-based SpeechCOMET outperforms Whisper-based variants.

The two speech models slightly underperform COMETKiwi$^{IWSLT}_{RoBERTa}$, which is trained on the same IWSLT data but with text as source, indicating that text remains a stronger signal than speech. In general, SpeechCOMET$_{SONAR}$ outperforms SpeechCOMET$_{Whisper}$ (20.1 vs. 17.8 τ_b avg) on the IWSLT test set. Neither model shows sensitivity to gender or prosody.

Speech+text: multimodal fusion improves correlation. We then combine both modalities by fusing the individual source embeddings. We ablate different fusion strategies and two initialization schemes: training from scratch and initializing from an existing text SpeechCOMET checkpoint. We report the best model, sum fusion initialized from COMETKiwi$^{WMT}_{RoBERTa}$, in Table 2; the full ablation is in Table 8 in Appendix B.3.

SpeechCOMET achieves 25.8 τ_b and 73.5 SPA avg, outperforming both unimodal variants. MuST-SHE and ContraProST performance remains at chance. Scaling the text encoder to InfoXLM-large (SpeechCOMET^{XL}), the same encoder as the off-the-shelf COMETKiwi uses, closes the gap to the COMETKiwi text baseline (34.4 vs. 34.6 τ_b), suggesting that the performance gap of the smaller models relative to COMETKiwi is due to encoder capacity rather than training procedure. However, this gain is driven entirely by the stronger text encoder, as speech sensitivity remains unchanged.

4.2 SpeechLLM as a Judge Results

We now discuss the performance of the SpeechLLM. Generally, SpeechLLM outperforms all SpeechCOMET variants and the COMETKiwi text baseline on IWSLT, with the best model reaching 49.9 τ_b avg compared to 34.4 for SpeechCOMET^{XL} and 34.6 for COMETKiwi (Table 2).

Text-only: performance improves with FT. The SpeechLLM with only text input achieves 40.4 τ_b avg without fine-tuning, and 47.1 with IWSLT fine-tuning, with similarly large gains at system level (64.1 to 88.5 SPA avg), outperforming the SpeechCOMET models. The below-chance PA scores on MuST-SHE and ContraProST are driven by ties: the model assigns identical scores to most pairs, and PA is exactly 50% once ties are excluded.

Speech-only: speech input underperforms text. With speech input, SpeechLLM performs below its text-only counterpart (32.0 vs. 40.4 τ_b avg, 60.1 vs. 64.1 SPA avg), with the gap persisting after fine-tuning (40.1 vs. 47.1 τ_b). Despite having access to the speech signal, MuST-SHE and ContraProST show the same tie-driven pattern as the text-only model, with no gender or prosody detection.

Speech+text: combining modalities yields best correlation but no speech sensitivity. Combining both modalities yields the strongest results: 49.9 τ_b avg with FT and 80.6 SPA avg with FT, surpassing all other models. Still, the model does not show sensitivity to speech-specific phenomena.

Specific prompts do not enhance speech sensitivity. Since speech sensitivity does not emerge from architecture or fine-tuning, we additionally test whether it can be elicited through task-specific prompting. However, instructing the model to attend to paralinguistic cues yields no consistent improvement on MuST-SHE or ContraProST. We report these results in Table 9 in Appendix B.

5 Analyses

Having shown that no current metric is sensitive to speech-specific phenomena, we test three complementary explanations: the features are 1) missing from encoder representations, 2) not induced during training, or 3) disregarded during QE scoring.

5.1 Are features present in the encoders?

We probe for source linguistic and acoustic features in Figure 2 and 3. For each system, we extract the source representations and train an MLP probe (Adi et al., 2017) to predict the target feature. For COMET and SpeechCOMET, we use either the

encoder's fixed-size sentence embedding (SONAR, RoBERTa for text) or the utterance representation from the pooling layer learned during QE training (Whisper). For SpeechLLMs, we use the hidden state immediately following the prompt. Details about the probing experiments are in Appendix C.1.

5.1.1 Probing Linguistic Features

We probe five linguistic features: subject number, object number (singular, plural, none in both cases), tense (present, past, none), verb form (finite, infinite, participle, none), and the presence of named entities (binary). Figure 2 shows three patterns across these probed linguistic features.

Modality gap separates text-aware from speech-only systems. There is a clear gap between input modalities: systems that have access to text at the source, whether text-only (COMETKiwi variants) or speech-and-text combined (SpeechCOMET), achieve substantially higher probing accuracy than their speech-only counterparts. This pattern holds consistently across all five features and suggests that a considerable amount of linguistically structured information present in the text signal is not recoverable from the audio signal alone, at least as encoded by the models studied here.

QE fine-tuning does not close the speech-text modality gap. Pairwise comparison of fine-tuned and frozen variants of the same architecture reveals nearly identical probing accuracies. The absence of improvement from fine-tuning indicates that the quality estimation training signal alone is insufficient to drive the encoder towards more linguistically informative representations of source speech, likely because standard QE training provides no examples that would enforce such features.

SONAR representations dominate Whisper in preserving linguistic features. SONAR provides a fixed-size utterance-level embedding pretrained to capture semantic content, whereas Whisper's utterance representation is obtained via a pooling mechanism learned from scratch during quality estimation training. The fact that the off-the-shelf SONAR embedding outperforms the jointly trained Whisper pooling reinforces the conclusion that the quality estimation objective, in its current form, does not provide sufficient signal for creating high-quality audio source representations.

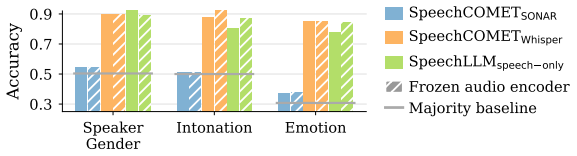

Figure 3: Probing accuracy for speaker gender on MuST-SHE and for intonation (question vs. statement) and emotion (angry, happy, neutral, surprised) on ContraProST. Scores averaged over three runs (std. ≤ 0.02). SpeechCOMET$_{\text{Whisper}}$ and SpeechLLMs enable accurate recovery of acoustic features.

5.1.2 Probing Acoustic Features

Having probed for linguistic features, we now test whether acoustic and speaker-level features can be recovered from the source representations (Figure 3). We probe three features: speaker gender, intonation, and emotion. We focus on speech-only models, and defer the full set of text-only and speech-text models (Table 2) to Appendix C.2. These features are not recoverable from text alone, so the probe must rely on acoustic information.

Overall, the trend is consistent across the three probed features: Whisper representations and the SpeechLLM hidden states support accurate recovery, while SONAR representations do not. The results have the following implications.

SONAR's semantic alignment may erase acoustic information. SpeechCOMET with the SONAR backbone, despite outperforming the Whisper variant on the earlier linguistic probes (Figure 2), now performs at or just above the majority baseline on acoustic features. A plausible reason is that SONAR's training objective, which aligns sentences with the same meaning across languages, pulls together utterances that share meaning but differ acoustically. Features like speaker gender and intonation, which are largely uncorrelated with a sentence's overall meaning, would then be suppressed as a side effect.

Whisper and SpeechLLMs encode speaker gender but do not use it. For both Whisper and the SpeechLLMs, the probe recovers speaker gender with high accuracy, confirming that this information is present in the source representation. Yet neither is sensitive to acoustic inputs in its QE outputs (Table 2). Therefore, the feature is available but unused. Unlike SONAR, where the acoustic information appears to be erased from the audio representation, here the bottleneck lies not in the encoder but in the training signal, which does not

		Segment τ_b (%)	real src	Δ random
Baselines				
TEXT	COMETKIWI-PART.		11.9	-4.3
	COMETKIWI		34.6	-24.9
SPEECH	SPEECHQE		29.8	-22.5
	BLASER		24.4	-15.4
SpeechCOMET				
TEXT	$\text{COMETKiwi}^{\text{WMT}}_{\text{RoBERTa}}$		23.1	-11.1
	$\text{COMETKiwi}^{\text{IWSLT}}_{\text{RoBERTa}}$		22.8	-12.0
SPEECH	$\text{SpeechCOMET}_{\text{SONAR}}$		20.1	-3.1
	$\text{SpeechCOMET}_{\text{Whisper}}$		17.8	-2.4
SP.+TXT	SpeechCOMET		25.8	-17.8
SP.+TXT	$\text{SpeechCOMET}^{\text{XL}}$		34.4	-30.2
SpeechLLM				
TEXT	SpeechLLM		40.4	-29.2
	+FT		47.1	-38.6
SPEECH	SpeechLLM		32.0	-28.1
	+FT		40.1	-25.5
SP.+TXT	SpeechLLM		38.1	-37.4
	+FT		49.9	-39.2

Table 3: Effect of replacing the source input with a randomly mismatched one. Scores are averaged over de and zh. Δ = random−real: large negative values indicate the model uses the source signal.

reward attending to acoustic information.

5.2 Source Reliance of the QE Models

Automated metrics are notorious for focusing primarily on the translation only and disregarding the sources (Sun et al., 2020; Zouhar et al., 2023). To quantify how much the models rely on their source input when assigning QE scores, we replace every source with a randomly mismatched one and measure the drop in segment τ_b in Table 3.

COMETKIWI and the speech baseline SPEECHQE show strong source sensitivity, dropping by -24.9 and -22.5 pp respectively, yet the latter is still not able to use speech-specific features (discussion in Section 6).

In contrast, $\text{SpeechCOMET}_{\text{SONAR}}$ and $\text{SpeechCOMET}_{\text{Whisper}}$ drop by only -3.1 and -2.4 pp respectively, indicating that these models largely ignore the speech input. On the other hand, for the speech+text SpeechCOMET models, the drop with random source (-17.8 and -30.2 pp) is substantial, but shuffling only the text reproduces it almost entirely while shuffling only the speech has a negligible effect (Table 11 in Appendix C.3), confirming these models condition on text.

Manual inspection of 50 examples where speech shuffling causes a drop reveals segmentation errors where the speech segment boundaries do not align with the translation (e.g., audio *"Right."* paired with MT *"Ansätze symmetrisch, richtig?* EN: The approaches are symmetric, right?"), suggesting that the small residual speech sensitivity reflects misalignments rather than genuine speech conditioning.

Finally, SpeechLLM shows large drops across all modalities (-25 to -39 pp), including speech-only, confirming that these models do condition on the source speech, yet the probing results show that speech-specific features go unused.

6 Discussion and Recommendations

Three barriers to speech-specific evaluation. Our results point to three interacting barriers to speech-specific QE. *First*, source encoders discard relevant information: SONAR suppresses acoustic attributes by design, while Whisper representations are weak on linguistic features. *Second*, even when features are present, models do not use them, suggesting QE training fails to ground predictions in the source. *Third*, QE fine-tuning does not help, likely due to standard QE corpora lacking examples where speech-specific variation impacts quality.

Practical guidance for practitioners. For practitioners evaluating speech translation quality, we recommend SpeechLLM (+FT) when computational cost is not a constraint, as it achieves the strongest segment-level correlation across all modalities. When a lighter model is needed the off-the-shelf COMETKiwi remains the strongest text-based option. However, none of the evaluated models should be used to assess speech-specific translation quality, such as gender agreement or prosodic appropriateness: all models score at chance on this.

7 Conclusion

In this work, we evaluate speech translation quality estimation across regression-based and LLM-based approaches and find that, despite strong overall correlation with human judgements, no current metric is sensitive to speech-specific phenomena such as gender agreement and prosody. Addressing this requires dedicated training data that explicitly targets such phenomena and encoders that capture richer acoustic representations than those induced by the quality estimation objective alone. We hope that our work can serve as a stepping stone for future research on speech-aware quality estimation and encourage the development of benchmarks and models that treat speech translation evaluation as a genuinely multimodal problem.

8 Limitations

Correlation results are based on a single test set (IWSLT 2026 Metrics dev set), as no other human-annotated speech translation evaluation datasets are currently available. Similarly, our evaluation of speech-specific phenomena is limited to gender agreement and prosody, and all datasets use English as the source language, as no comparable resources exist for other phenomena or languages. This highlights the need for broader evaluation resources to ensure that findings generalise across languages and domains.

9 Ethical Considerations

This work uses only existing publicly available datasets and pre-trained models. We foresee no direct ethical concerns arising from this research.

References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Mark Fishel, Marco Gaido, Dávid Javorský, Marek Kasztelnik, and 33 others. 2025. Findings of the IWSLT 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 412–481, Vienna, Austria (in-person and online). Association for Computational Linguistics.

David Ifeoluwa Adelani, Antonios Anastasopoulos, Victor Agostinelli, Luisa Bentivogli, Ondřej Bojar, Sebastien Bratières, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, Marcello Federico, Marco Gaido, Mahendra Gupta, HyoJung Han, Ali Hatami, David Javorský, Yejin Jeon, Marek Kasztelnik, Danni Liu, and 26 others. 2026. Speech translation and metrics in 2026: Findings of the iwslt campaign. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, California, US. Association for Computational Linguistics.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, and 43 others. 2023. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Kshitij Ambilduke, Ben Peters, Sonal Sannigrahi, Anil Keshwani, Tsz Kin Lam, Bruno Martins, Andre Martins, and Marcely Zanon Boito. 2025. From tower to spire: Adding the speech modality to a translation-specialist LLM. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19658–19673, Suzhou, China. Association for Computational Linguistics.

Chantal Amrhein and Barry Haddow. 2022. Don't discard fixed-window audio segmentation in speech-to-text translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 203–219, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.

Mauro Cettolo, Marco Gaido, Matteo Negri, Sara Papi, and Luisa Bentivogli. 2025. How to evaluate speech translation with source-aware neural MT metrics. *CoRR*, abs/2511.03295.

Mingda Chen, Paul-Ambroise Duquenne, Pierre Andrews, Justine Kao, Alexandre Mourachko, Holger Schwenk, and Marta R. Costa-jussà. 2023. BLASER: A text-free speech-to-speech translation evaluation metric. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9064–9079, Toronto, Canada. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Kwanghee Choi, Eunjung Yeo, Cheol Jun Cho, David Harwath, and David R. Mortensen. 2026. [b]=[d]-[t]+[p]: Self-supervised speech models discover phonological vector arithmetic. *Preprint*, arXiv:2602.18899.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong,

Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, and 46 others. 2023. Seamless: Multilingual expressive and streaming speech translation. *Preprint*, arXiv:2312.05187.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Lina Conti, Dennis Fucci, Marco Gaido, Matteo Negri, Guillaume Wisniewski, and Luisa Bentivogli. 2026. Voice, bias, and coreference: An interpretability study of gender in speech translation. *Preprint*, arXiv:2511.21517.

David Dale and Marta R. Costa-jussà. 2024. BLASER 2.0: a metric for evaluation and quality estimation of massively multilingual speech and text translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16075–16085, Miami, Florida, USA. Association for Computational Linguistics.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoit Sagot. 2023. SONAR: sentence-level multimodal and language-agnostic representations.

Michele Gubian, Ioana Krehan, Oli Danyi Liu, James Kirby, and Sharon Goldwater. 2025. Analyzing the relationships between pretraining language, phonetic, tonal, and speaker information in self-supervised speech models. *CoRR*, abs/2506.10855.

HyoJung Han, Kevin Duh, and Marine Carpuat. 2024. SpeechQE: Estimating the quality of direct speech translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21852–21867, Miami, Florida, USA. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Maurice George Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Yi Li, Rui Zhao, Ruiquan Zhang, Jinsong Su, Daimeng Wei, Min Zhang, and Yidong Chen. 2026. Plast: Towards paralinguistic-aware speech translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(38):31805–31813.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, and 57 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Sara Papi, Javier Garcia Gilabert, Zachary Hopton, Vilém Zouhar, Carlos Escolano, Gerard I. Gállego, Jorge Iranzo-Sánchez, Ahrii Kim, Dominik Macháček, Patricia Schmidtova, and Maike Züfle. 2026. Hearing to translate: The effectiveness of speech modality integration into llms. *Preprint*, arXiv:2512.16378.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Haoqin Sun, Xuechen Wang, Jinghua Zhao, Shiwan Zhao, Jiaming Zhou, Hui Wang, Jiabei He, Aobo Kong, Xi Yang, Yequan Wang, Yonghua Lin, and Yong Qin. 2025. Emotiontalk: An interactive chinese multimodal emotion dataset with rich annotations. *Preprint*, arXiv:2505.23018.

Shuo Sun, Francisco Guzmán, and Lucia Specia. 2020. Are we estimating or guesstimating translation quality? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages

6262–6267, Online. Association for Computational Linguistics.

Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1222–1234, Miami, Florida, USA. Association for Computational Linguistics.

Ioannis Tsiamas, Matthias Sperber, Andrew Finch, and Sarthak Garg. 2024. Speech is more than words: Do speech-to-text translation systems leverage prosody? In *Proceedings of the Ninth Conference on Machine Translation*, pages 1235–1257, Miami, Florida, USA. Association for Computational Linguistics.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Vilém Zouhar, Shehzaad Dhuliawala, Wangchunshu Zhou, Nico Daheim, Tom Kocmi, Yuchen Eleanor Jiang, and Mrinmaya Sachan. 2023. Poor man's quality estimation: Predicting reference-based MT metrics without the reference. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1311–1325, Dubrovnik, Croatia. Association for Computational Linguistics.

Vilém Zouhar, Maike Züfle, Beni Egressy, Julius Cheng, and Jan Niehues. 2025. Early-exit and instant confidence translation quality estimation. *Preprint*, arXiv:2502.14429.

Maike Züfle and Jan Niehues. 2025. Contrastive learning for task-independent SpeechLLM-pretraining. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8469–8490, Vienna, Austria. Association for Computational Linguistics.

A Training Details

A.1 SpeechCOMET

Table 4 lists the hyperparameters used for all SpeechCOMET models. All models share the same optimizer and estimator architecture; the Whisper LoRA and SONAR fine-tuning settings apply only to the respective encoder variants. All models are trained on a single NVIDIA A100-SXM4-40GB GPU; speech models take approximately 30 hours to train, while text-only models require around 20 hours.

Training	
Optimizer	AdamW
Learning rate (estimator)	1.5×10^{-5}
Learning rate (encoder)	1×10^{-6}
Layerwise LR decay	0.95
Batch size	2
Gradient accumulation	8
Max epochs	20
Gradient clipping	1.0
Encoder warm-up	30% of epoch 0 frozen
Dropout	0.1
Early stopping patience	2 (val Kendall τ)
Architecture	
Text encoder pooling	layerwise attention (sparsemax)
Aggregation	average pooling
Estimator hidden sizes	[2048, 1024]
Estimator activation	Tanh
Whisper LoRA	
LoRA rank r	8
LoRA α	16
LoRA dropout	0.0
SONAR Fine-tuning	
Frozen layers (bottom)	4 of 24

Table 4: Hyperparameters for SpeechCOMET models. The effective batch size is 16. All models are trained on a single NVIDIA A100-SXM4-40GB GPU.

A.2 SpeechLLM Details

We evaluate Qwen2.5 Omni 7B (Xu et al., 2025) in both zero-shot and fine-tuned settings across three input modalities for the QE task: text, audio, and audio+text. For zero-shot evaluation we assess all three modalities of the unmodified base model. In addition to the standard QE prompt, we evaluate two task-specific prompt variants: a gender-aware prompt for MuST-SHE and a prosody-aware prompt for ContraProST. All prompts are shown in Figure 4.

For fine-tuning, we train one model per input modality using LlamaFactory (Zheng et al., 2024) with LoRA adapters (Hu et al., 2021). Fine-tuning and inference are performed on a single NVIDIA A100-SXM4-40GB GPU. Fine-tuning takes around 13 hours. Detailed hyperparameters are listed in Table 5.

A.3 Training Data

IWSLT training set. The Iwslt2026 Metrics Shared Train Dev dataset combines three sources

Standard QE

System: You are an evaluator. Given the source and/or audio and a translation, respond with only a single float score between 0 and 1 indicating translation quality. Output nothing else.

User (text): Source: *{src}*
Translation: *{mt}*

User (audio): *[audio waveform]*
Translation: *{mt}*

User (audio+text): *[audio waveform]*
Source: *{src}*
Translation: *{mt}*

Gender-aware QE (MuST-SHE)

System: You are a translation quality evaluator specialising in gender accuracy. Listen carefully to the speaker's voice in the audio to determine their gender. The speaker's gender must be correctly reflected in the translation through gendered forms such as adjectives, pronouns, and verb agreement. Score the translation from 0 to 1: give a high score if the gender is correctly reflected, a low score if it is wrong or ambiguous. Output only a single float score, nothing else.

User (text): Source: *{src}*
Translation: *{mt}*

User (audio): *[audio waveform]*
Translation: *{mt}*

User (audio+text): *[audio waveform]*
Source: *{src}*
Translation: *{mt}*

Prosody-aware QE (ContraProST)

System: You are a translation quality evaluator specialising in prosodic meaning. The speaker's prosody in the audio — including intonation, stress, rhythm, pauses, and emotional tone — conveys meaning that must be accurately reflected in the translation (e.g. whether an utterance is a question or statement, which word is emphasised, the level of politeness or emotion). Score the translation from 0 to 1 based on how well it captures the prosodic meaning of the source speech. Output only a single float score, nothing else.

User (text): Source: *{src}*
Translation: *{mt}*

User (audio): *[audio waveform]*
Translation: *{mt}*

User (audio+text): *[audio waveform]*
Source: *{src}*
Translation: *{mt}*

Figure 4: System and user prompts used for Qwen2.5-Omni-7B QE. The system prompt varies across the three settings: standard translation quality, gender accuracy, and prosodic meaning. The user prompt is identical across all three settings and varies only by input modality.

Fine-tuning	
LoRA rank r	16
LoRA α	32
LoRA target	all (attention + MLP)
Epochs	2
Learning rate	1×10^{-4}
LR scheduler	cosine
Warmup ratio	0.1
Batch size	1
Gradient accumulation	8
Precision	bf16
Max input length	2048
Early stopping patience	3
Eval interval	500 steps
Inference	
Decoding	greedy
Max new tokens	16
Precision	bf16

Table 5: Fine-tuning and inference hyperparameters for Qwen2.5-Omni-7B.

of human direct assessment annotations, all paired with source speech: IWSLT 2023 (Agarwal et al., 2023), where evaluators had access to one preceding and one following segment for context; and WMT 2024 (Kocmi et al., 2024) and WMT 2025 (Kocmi et al., 2025), both evaluated on human-segmented audio. It covers 17 language pairs: en-ar, en-bho, en-cs, en-de, en-es, en-et, en-hi, en-is, en-it, en-ja, en-mas, en-ru, en-sr, en-uk, en-zh, cs-de, and cs-uk.

WMT augmentation. The Wmt Human All dataset aggregates human assessments from multiple WMT editions. We exclude the WMT 2024 and WMT 2025 subsets that overlap with the IWSLT training set. The remaining data covers text translation systems with gold sentence segmentation and 49 language pairs: bn-hi, cs-de, cs-en, cs-uk, de-cs, de-en, de-fr, en-ar, en-bho, en-cs, en-de, en-es, en-et, en-ha, en-hi, en-hr, en-is, en-it, en-iu, en-ja, en-kk, en-ko, en-liv, en-mas, en-pl, en-ru, en-sr, en-uk, en-zh, fi-en, fr-de, gu-en, ha-en, he-en, hi-bn, is-en, ja-en, ja-zh, kk-en, liv-en, lt-en, pl-en, ps-en, ru-en, sah-ru, uk-en, xh-zu, zh-en, and zu-xh.

TTS synthesis. Since WMT provides only source text, we synthesise speech using Kokoro 82M for eight languages (English, Japanese, Chinese, Spanish, French, Hindi, Italian, Portuguese) and Mms Tts as a fallback for the remaining languages. Voice selection is deterministic per source segment. Identical source texts receive the same audio. The resulting dataset contains 1,114 hours.

B Detailed Results

B.1 SpeechCOMET Fusion Strategies

Table 6 compares three strategies for fusing the SONAR audio embedding $\mathbf{s}_a$ with the translation representation $\mathbf{h}_t$ before the regression head. All variants use XLM-RoBERTa-base with SONAR audio input, fine-tuned on IWSLT.

Four-way interaction. Following the original COMET estimator design, the two representations are combined as $[\mathbf{h}_t;\ \mathbf{s}_a;\ \mathbf{h}_t \odot \mathbf{s}_a;\ |\mathbf{h}_t - \mathbf{s}_a|]$, with SONAR audio replacing the text source embedding. This is the architecture used for SpeechCOMET in the main experiments and achieves the best segment-level correlation among the three variants.

Additive. Following the COMETKiwi unified-metric design, the audio embedding is added to the MT CLS representation before the regression head, collapsing both modalities into a single vector without explicit interaction terms.

Joint encoding. The SONAR audio embedding is appended as a virtual token to the MT token embeddings before the XLM-R transformer layers, enabling full cross-attention between the audio representation and every MT subword. The resulting CLS representation is fed directly to the regression head. Despite the richer interaction, this variant yields substantially lower segment-level correlation, suggesting that injecting a fixed-size audio embedding as a single token is difficult for the transformer to exploit effectively.

B.2 Training Strategies for SpeechCOMET

Table 7 reports the full ablation over encoder fine-tuning and TTS pretraining strategies for Sonar and Whisper speech encoders.

SONAR. Unfreezing the SONAR encoder consistently improves segment-level correlation over the frozen variant (20.1 vs. 18.5 τ_b avg for direct IWSLT fine-tuning). Neither WMT text pretraining nor TTS-synthesised WMT pretraining improves over direct IWSLT fine-tuning, with most pretraining variants performing below the unfrozen IWSLT baseline. This suggests that SONAR's fixed-size sentence-level representations already provide a good initialisation for quality estimation, and additional pretraining on out-of-domain data adds little.

	IWSLT dev						MuST-SHE			ContraProST		
	Segment τ_b (%)			System SPA (%)			PA (%)			PA (%)		
Architecture	**de**	**zh**	**avg**	**de**	**zh**	**avg**	**es**	**fr**	**it**	**de**	**es**	**ja**
four-way	**14.6**	**22.5**	**18.6**	**37.2**	80.4	**58.8**	**56.0**	**51.9**	51.3	50.0	**50.0**	**50.0**
additive	13.5	22.3	17.9	35.3	67.4	51.4	54.6	49.7	52.2	50.0	**50.0**	**50.0**
joint encoding	3.9	6.9	5.4	19.9	**88.6**	54.2	51.6	50.3	**52.8**	**50.6**	49.2	**50.0**

Table 6: Ablation of audio–text integration strategies for speech QE. *Four-way*: standard COMET estimator input $[\mathbf{h}_t; \mathbf{s}_a; \mathbf{h}_t \odot \mathbf{s}_a; |\mathbf{h}_t - \mathbf{s}_a|]$, with SONAR audio replacing the source text embedding (used in the main model). *Additive*: SONAR audio embedding summed with the unified CLS MT embedding, following the COMETKiwi unified-metric architecture. *Joint encoding*: SONAR audio embedding injected as a virtual token into XLM-R self-attention. All models are speech-only (SONAR audio input) fine-tuned on IWSLT.

		IWSLT dev						MuST-SHE			ContraProST		
		Segment τ_b (%)			System SPA (%)			PA (%)			PA (%)		
Model	**Variant**	**de**	**zh**	**avg**	**de**	**zh**	**avg**	**es**	**fr**	**it**	**de**	**es**	**ja**
SONAR encoder													
IWSLT FT	frozen	14.6	22.5	18.5	37.2	**80.4**	58.8	56.0	51.9	51.3	50.0	50.0	50.0
	unfrozen	**17.3**	**22.9**	**20.1**	**51.0**	70.9	**61.0**	53.1	50.9	50.9	50.0	50.0	50.0
WMT text pretrain → IWSLT FT	frozen	15.6	19.3	17.4	41.1	68.3	54.7	53.1	53.0	50.1	**50.1**	50.0	50.0
	unfrozen	16.7	21.2	18.9	46.2	73.6	59.9	53.5	52.3	50.5	50.0	50.0	50.0
WMT TTS pretrain (frozen)	pretrain only	10.7	16.6	13.7	23.6	71.5	47.6	53.7	52.1	**54.6**	49.9	50.0	50.0
	→ IWSLT FT (frozen)	13.8	18.8	16.3	32.7	67.7	50.2	53.7	52.3	54.4	50.0	49.9	50.0
WMT TTS pretrain (unfrozen)	pretrain only	11.6	15.4	13.5	26.3	72.1	49.2	50.9	50.9	53.4	50.0	**50.2**	49.7
	→ IWSLT FT (frozen)	15.1	20.0	17.6	32.7	67.9	50.3	**56.4**	51.7	53.0	50.0	50.0	50.1
	→ IWSLT FT (unfrozen)	15.7	21.8	18.7	37.7	68.6	53.1	54.1	**54.0**	53.2	49.9	50.0	**50.3**
Whisper encoder													
IWSLT FT (avg pool)	frozen	10.6	18.6	14.6	39.6	67.6	53.6	50.5	46.9	49.1	50.0	49.9	50.0
	LoRA	7.7	16.5	12.1	35.2	67.4	51.3	54.1	51.6	49.1	**50.1**	50.0	49.9
IWSLT FT (attn pool)	frozen	6.6	16.7	11.6	44.4	67.4	55.9	**55.8**	50.9	51.9	**50.1**	49.9	50.0
	LoRA	7.9	**19.4**	13.7	42.5	67.4	54.9	55.6	51.9	**53.4**	50.0	50.1	50.0
WMT TTS pretrain (LoRA)	pretrain only	11.9	13.1	12.5	**46.1**	59.9	53.0	49.5	**53.3**	48.1	49.9	**50.2**	49.9
	→ IWSLT FT (frozen)	**16.8**	18.7	**17.8**	44.9	**68.5**	**56.7**	51.4	50.7	52.8	50.0	50.0	**50.1**
	→ IWSLT FT (LoRA)	16.1	19.4	17.7	41.6	**68.5**	55.1	50.3	50.5	52.4	49.9	49.9	50.0

Table 7: Ablation of pretraining strategies for SpeechCOMET. → denotes sequential training stages. TTS pretraining uses WMT data with text-to-speech synthesis.

Whisper. Without pretraining, Whisper performs substantially below SONAR regardless of pooling strategy or adapter configuration (best: 14.6 τ_b avg). TTS pretraining on WMT data followed by frozen IWSLT fine-tuning is the strongest configuration (17.8 τ_b, 56.7 SPA avg), confirming that Whisper benefits from exposure to additional speech data before task-specific training. MuST-SHE and ContraProST scores remain near chance across all variants for both encoders.

B.3 Fusion Strategies for SpeechCOMET (Speech+Text)

Table 8 ablates embedding fusion strategy and SONAR encoder state for SpeechCOMET models combining speech and text source representations.

Fusion strategy. Sum fusion outperforms both average and concatenation on segment-level correlation. Concatenation performs worst (16.6 τ_b avg). Average fusion (20.8 τ_b) falls below sum (22.8 τ_b).

Initialisation from text checkpoint. Initialising the multimodal model from the WMT-pretrained text SpeechCOMET checkpoint before IWSLT fine-tuning yields a substantial gain over training from scratch (25.8 vs. 22.8 τ_b avg). Freezing vs. unfreezing SONAR during this stage makes little difference. MuST-SHE scores remain near chance across all variants and ContraProST shows no sensitivity.

		IWSLT dev						MuST-SHE			ContraProST		
		Segment τ_b (%)			System SPA (%)			PA (%)			PA (%)		
Fusion	**SONAR**	**de**	**zh**	**avg**	**de**	**zh**	**avg**	**es**	**fr**	**it**	**de**	**es**	**ja**
avg	frozen	17.0	24.5	20.8	64.5	67.7	66.1	52.8	53.0	50.7	**50.1**	**50.1**	**50.0**
concat	frozen	12.2	21.1	16.6	49.0	**75.7**	62.3	52.4	48.6	51.3	50.0	50.0	**50.0**
sum	frozen	20.9	24.7	22.8	70.2	67.4	68.8	54.8	**55.4**	**53.0**	50.0	50.0	**50.0**
	WMT text pretrain → FT (frozen)	**24.5**	**27.1**	**25.8**	79.7	67.4	73.5	55.6	52.6	52.4	50.0	49.9	**50.0**
	WMT text pretrain → FT (unfrozen)	23.4	**27.1**	25.2	**80.2**	67.4	**73.8**	**56.0**	54.4	**53.0**	50.0	50.0	**50.0**

Table 8: Ablation of embedding fusion strategies for TEXT + SPEECH SpeechCOMET with SONAR encoder. *Fusion* controls how speech and text embeddings are combined; *SONAR* indicates whether the encoder is frozen or unfrozen during fine-tuning. Text init initialises from the WMT-pretrained text SpeechCOMET checkpoint.

		MuST-SHE PA (%)			ContraProST PA (%)		
Type	**Model**	**es**	**fr**	**it**	**de**	**es**	**ja**
SPEECH	SpeechLLM	31.9	28.3	32.2	**31.6**	**28.4**	**31.6**
	speech prompt	38.0	32.3	32.4	15.3	19.4	16.8
	+FT	35.9	15.0	20.7	20.8	22.1	15.3
TEXT	SpeechLLM	29.4	25.9	23.6	23.8	18.6	29.6
	speech prompt	23.7	17.3	16.2	18.9	18.5	15.7
	+FT	14.0	13.8	15.4	14.1	14.1	25.0
SP.+TXT	SpeechLLM	38.0	30.9	34.5	30.5	26.4	31.3
	speech prompt	**43.1**	**37.1**	**37.8**	22.8	26.8	21.4
	+FT	8.2	6.3	6.8	18.0	9.3	19.7

Table 9: Effect of speech-aware prompting on SpeechLLM across input modalities. The speech prompt instructs the model to consider paralinguistic cues; +FT denotes fine-tuning on IWSLT.

B.4 SpeechLLM Prompt Sensitivity

Table 9 reports MuST-SHE and ContraProST PA for SpeechLLM with and without a speech-aware prompt across all input modalities. All values fall below 50%, reflecting the tie-heavy prediction pattern described in Section 4: the model assigns identical scores to most pairs, and PA equals 50% once ties are excluded.

C Analysis Details

C.1 Probing Setup

Training details of the probing classifiers are presented in Table 10.

C.1.1 Linguistic Feature Probing Data

The linguistic probing dataset is derived from the IWSLT 2026 Metrics Shared Task dataset, which pairs English and Czech source speech and text with MT hypotheses and human quality scores across several language pairs. As the morphological tagger is English-only, we restrict probing to English sources, yielding 31,965 training and 5,556 development examples.

We annotate the source side with spaCy's `en_core_web_trf` model and extract five sentence-level features. "Verb Form" and "Tense" are read from the `ROOT` token of the dependency parse. "Subject Number" and "Object Number" give the gram-

Hyperparameter	**Value**
Architecture	
Classifier	MLP
Input dimension	768 for (speech)COMET 3584 for Qwen2.5-Omni
Hidden dimension	256
Activation	ReLU
Output layer	Linear (→ #classes)
Dropout	0.1
Optimisation	
Optimizer	Adam
Learning rate	10^{-3} by default 3×10^{-4} (Qwen2.5-Omni)
Batch size	256
Data	
Linguistic feat.	IWSLT 2026 Metric (English source) training data sampled with 10%
Speaker gender	MuST-SHE v1.2 (category-1 only) 50/50 stratified split for train/test
Intonation	ContraProST "prosodic breaks" subset 80/20 stratified split for train/test
Emotion	ContraProST "emotional prosody" subset 80/20 stratified split for train/test

Table 10: Probing classifier configuration.

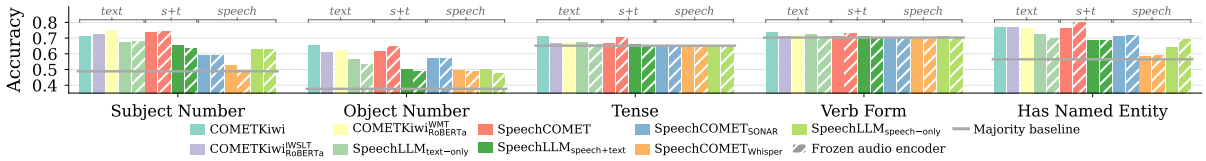

Figure 5: Probing accuracy for linguistic features on the IWSLT 2026 Metrics dataset, grouped by input modality (text-only, speech+text, speech-only).

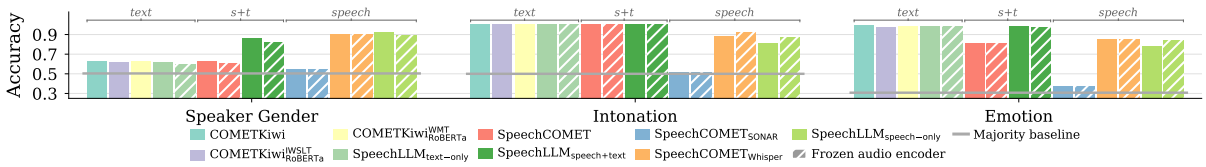

Figure 6: Probing accuracy for acoustic features on MuST-SHE (speaker gender) and ContraProST (intonation, emotion). Text models are unable to recover *speaker gender* accurately, so the speech-only systems must rely on acoustic cues. For *intonation* and *emotion*, the dataset's source transcripts expose the feature (trailing punctuation and typographic prosody markers respectively), which explains the high text-only accuracy on these two.

matical number (singular/plural) of the first `nsubj` and `obj` dependents. "Has Named Entity" indicates whether spaCy recognises any entity span, covering both proper entities (e.g., PERSON, ORG, GPE) and numeric or temporal expressions (e.g. DATE, CARDINAL, MONEY). Sentences for which a label is undefined (e.g., no object for Object Number) are excluded from that feature only.

C.1.2 Acoustic Feature Probing Data

Speaker Gender. Derived from MuST-SHE v1.2, which provides the English source, a reference translation, a gender label (He/She), and a category label for three directions (En→Es/Fr/It). We keep only category-1 segments, where gender is lexically recoverable from the utterance, and deduplicate by (talk ID, source text, speaker). Stratified splitting yields two 526-segment halves (He: 265, She: 261).

Intonation. Constructed from the ContraProST Intonation subset: 173 English sentences each recorded as a declarative (".") and a yes/no question ("?"), for 346 files. We deduplicate by audio path across the three target-language files (En→De/Es/Ja). Splitting at the sentence level (both recordings in the same partition), an 80/20 split gives 276 training (138/138) and 70 development (35/35) examples. Note that the statement/question distinction surfaces in the source text as *trailing punctuation*, which is why text-based models reach near-perfect accuracy.

Emotion. Also from ContraProST, using the Emotional Prosody subset, where each sentence is recorded under multiple emotions. Of the six original classes we keep the four most frequent (Neutral, Angry, Happy, Surprised), as the other two classes contain too few samples. Splitting at the sentence level, an 80/20 split gives 964 training (Neutral: 301, Angry: 292, Happy: 251, Surprised: 120) and 247 development (76/75/64/32) examples. As with intonation, the transcripts carry class-discriminative typography: neutral utterances are plain text, while angry, happy, and surprised ones use asterisks (`*word*`), underscores (`_word_`), capitalisation, and repeated punctuation (e.g. ?!?!), again allowing high text-based accuracy.

C.2 Additional Probing Results

The probing results for the full set of models are presented in Figure 5 and Figure 6, corresponding to linguistic and acoustic features respectively.

C.3 Source Modality Ablation for SpeechCOMET (Speech+Text)

Table 11 disentangles the contribution of each input modality for the two speech+text SpeechCOMET models by separately replacing audio or text with a randomly mismatched source. For both models, replacing only the text source reproduces nearly the full drop observed when both modalities are shuffled (-17.2 pp for SpeechCOMET, -30.6 pp for SpeechCOMET^{XL}), while replacing only the audio has a negligible effect (-1.4 and -0.3 pp respectively). The pattern is consistent across both

encoder sizes, confirming that the multimodal models condition almost exclusively on the text source.

Model	**real src**	**Δ both**	**Δ audio**	**Δ text**
Segment τ_b avg (%) over de & zh				
SpeechCOMET	25.8	-17.8	-1.4	-17.2
SpeechCOMET^{XL}	34.4	-30.2	-0.3	-30.6

Table 11: Source-modality ablation for speech+text SpeechCOMET models. Δ audio only: audio replaced with mismatched audio, text kept real. Δ text only: text replaced with mismatched text, audio kept real. Δ both: both modalities replaced. Scores are segment τ_b averaged over de and zh.