

SpeechQE: Estimating the Quality of Direct Speech Translation

HyoJung Han
Computer Science
University of Maryland
hjhan@cs.umd.edu

Kevin Duh
HLTCOE
Johns Hopkins University
kevinduh@cs.jhu.edu

Marine Carpuat
Computer Science
University of Maryland
marine@cs.umd.edu

Abstract

Recent advances in automatic quality estimation for machine translation have exclusively focused on written language, leaving the speech modality underexplored. In this work, we formulate the task of quality estimation for speech translation, construct a benchmark, and evaluate a family of systems based on cascaded and end-to-end architectures. In this process, we introduce a novel end-to-end system leveraging pre-trained text LLM. Results suggest that end-to-end approaches are better suited to estimating the quality of direct speech translation than using quality estimation systems designed for text in cascaded systems. More broadly, we argue that quality estimation of speech translation needs to be studied as a separate problem from that of text, and release our data and models to guide further research in this space.¹

1 Introduction

Recent progress in quality estimation (QE) (Specia et al., 2010) makes it possible to automatically rate the quality of machine translation (MT) given only the input and output of an MT system. QE ratings have been found to correlate well with human judgments, sometimes as well as reference-based metrics (Kepler et al., 2019; Rei et al., 2020, 2023). However, this work has focused on text translation.

Meanwhile, the rapid development of speech technology (Radford et al., 2023; Seamless Communication et al., 2023) has expanded the use of speech translation (ST) applications in daily life, thus increasing the need to predict the reliability of their output. This raises the question of whether quality estimation for ST can be performed using a combination of state-of-the-art automatic speech recognition (ASR) and text-based QE (text-QE or MTQE) methods. However, relying on a cascade of ASR and text-QE systems presents two major issues: (1) The current top-performing ST models

¹<https://github.com/h-j-han/SpeechQE>

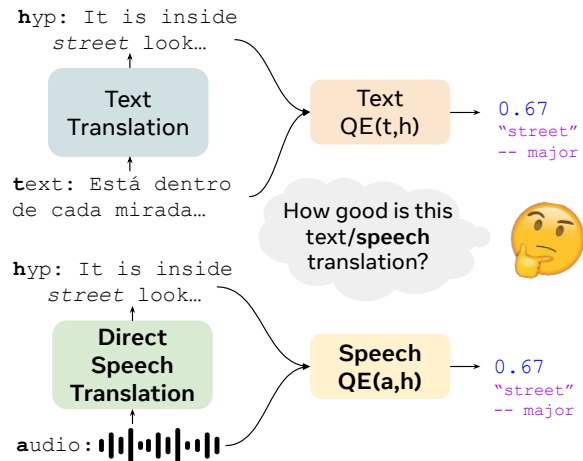


Figure 1: Quality Estimation for Speech Translation (SpeechQE) vs. Text Quality Estimation (text-QE).

directly translate the audio input into target language text without transcribing the audio, making it inefficient to run an additional ASR system to generate an input for the text-QE module. (2) ASR transcriptions of the audio input may not match the gold transcription, potentially misleading the text-QE system. Hence, we hypothesize that end-to-end approaches might be better suited for this task.

In light of these issues, we formulate the task of **quality estimation for speech translation (SpeechQE or STQE, Figure 1)** and explore both cascaded and end-to-end (E2E) systems for this task (Figure 2). While we rely on existing ASR and text-QE modules for the cascaded system, we introduce a novel E2E SpeechQE model architecture to address the lack of a dedicated end-to-end system for this task. Our design incorporates a pre-trained speech encoder and a large language model (LLM) to leverage their existing capabilities in extracting high-quality audio features and handling translation-related tasks.

To conduct a thorough evaluation, we contribute an evaluation benchmark and training data for SpeechQE from diverse ST outputs scored with

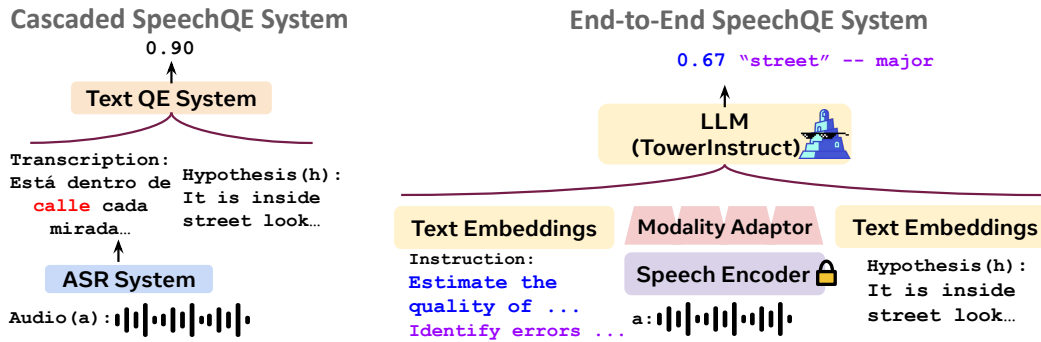


Figure 2: Comparing cascaded and end-to-end approaches to Quality Estimation for Speech Translation (SpeechQE).

reference-based metrics. Results show that E2E models outperform the cascaded system based on a state-of-the-art (SOTA) ASR module in correlation with both (1) human direct assessment ratings and (2) metric scores. Additionally, our E2E model can detect error spans to some extent in a zero-shot fashion, though the best results are still achieved by cascaded systems with SOTA ASR. Qualitative analysis highlights the robustness of E2E models against wrong speech representation in score prediction, error span detection, and severity prediction. Based on this evidence, we argue that SpeechQE should be studied as a distinct problem from text-QE.

2 Background

Quality estimation makes it possible to assess translation quality without reference translations, which is essential for practical use cases (Specia et al., 2010; Callison-Burch et al., 2012). QE signals can benefit end users by helping them decide how to rely on outputs in casual and high-risk settings alike (Specia et al., 2022; Mehandru et al., 2023). They can also benefit downstream tasks or enhance MT itself (Fernandes et al., 2022).

The QE task has been framed in various ways, including predicting sentence-level quality ratings (Callison-Burch et al., 2012) or word-level binary tags of OK/BAD (Bojar et al., 2013). While a wealth of methods have been developed for these tasks, recent work has shown the benefits of developing solutions to address them jointly. OpenKiwi (Kepler et al., 2019) streamlined QE by supporting both word-level tagging and regression toward a sentence-level score within a unified toolkit (Kim et al., 2017). It was further improved with a training recipe that better supports multilingual generalization (Rei et al., 2020, 2023). Together with the development of learned metrics for reference-based

evaluation (Rei et al., 2020; Sellam et al., 2020), this set the stage for a single or family of models that flexibly rate the quality of MT output with or without access to a reference human translation (Guerreiro et al., 2024; Juraska et al., 2023) with high correlations with human quality ratings (Fritag et al., 2023). xCOMET (Guerreiro et al., 2024) even integrates both sentence-level evaluation and error span detection capabilities while categorizing error spans, thereby enriching the quality measures.

Meanwhile, quality estimation for speech translation remains understudied. Le et al. (2016) address the task of tagging each word in an ST output as good or bad, using ASR and MT features. Their approach can be viewed as a cascaded SpeechQE system, which propagates a confidence score in a pipeline of ASR and statistical machine translation (SMT) modules. BLASER2.0 (Seamless Communication et al., 2023) produces a similarity score between a translation output and input, using SONAR sentence-embeddings that can compare either speech or text (Duquenne et al., 2023). While this enables SpeechQE, this approach was initially designed for speech-to-speech translation (Chen et al., 2023), and was exposed to only a small amount of training data with quality labels.

With advances in ST technology and their growing use (Rubenstein et al., 2023), there is a need for QE to support ST scenarios where intermediate automatic speech recognition (ASR) outputs are not available, along with new evaluations to correctly gauge the effectiveness of quality estimation in speech translation.

3 SpeechQE: Task and Models

We define the task of estimating the quality of speech translation (SpeechQE or STQE²), before

²We choose to use terms SpeechQE and text-QE as main instead of alternative terms STQE and MTQE to emphasize

introducing our cascaded and E2E systems.

In this work, we focus on predicting sentence-level scores and measuring the correlation of reference ratings provided by humans or reference-based metrics (Fonseca et al., 2019). Additionally, we will explore an error span detection task (Blain et al., 2023) in Section 5.4, to broaden the scope of QE beyond holistic numerical ratings.

We refer to a **reference-based metric** as *metric*. Given a reference target text r , an MT hypothesis h and optionally the MT source text t , the *metric* rates the quality of h as a score m :

$$m = \text{metric}(h, r) \text{ or } m = \text{metric}(t, h, r) \quad (1)$$

Likewise, we refer to a **text quality estimation** system as text-QE. It produces an output score q given only a source text t and an MT hypothesis h .

$$q = \text{text-QE}(t, h) \quad (2)$$

In the **SpeechQE task** (Figure 1), given the source **audio** a and the translation hypothesis h , a system outputs the quality score q for this hypothesis:

$$q = \text{SpeechQE}(a, h) \quad (3)$$

3.1 Cascaded SpeechQE System

We first consider cascaded SpeechQE systems that output the score q_{cas} from a text-based QE system with the input of transcribed text $ASR(a)$ from an ASR system and hypothesis text h (Figure 2).

$$q_{cas} = \text{text-QE}(ASR(a), h) \quad (4)$$

While the cascaded systems offer a straightforward approach to SpeechQE, they present several issues. First, efficiency is a concern, as there are no naturally occurring intermediate ASR transcripts in the case of direct ST, necessitating additional ASR runs to generate inputs for the text-QE component. This introduces latency that may be undesirable in user-facing quality estimation applications. Second, source transcriptions produced by a separate ASR do not always accurately represent the spoken input, making the text-QE system vulnerable to the wrong speech representation. Third, there is a modality mismatch, as the text-QE component is not adapted to spoken language, which exhibits different styles or errors from written language. These challenges motivate us to explore end-to-end (E2E) SpeechQE solutions.

the contrast between speech and text and to facilitate easier reading. More discussion of the terminology in Appendix E.

3.2 End-to-End SpeechQE System

We introduce the architecture and training scheme for our E2E SpeechQE model.

Model Architecture Rather than training an integrated model from scratch, we choose to leverage a pre-trained speech encoder and a large language model (LLM) to utilize their abilities in extracting high-quality audio features and handling translation-related tasks, respectively. This approach is particularly useful when there is limited or no data available for training from scratch, as it enables the transfer of knowledge from text-based large language models (text-LLM) to the speech domain. We adopt a popular configuration for integrating speech modality into text-LLM that trains a lightweight modality adapter (Wu et al., 2023; Fathullah et al., 2023; Wang et al., 2023a,b), but the optimal architecture for SpeechQE or even broadly for integrating speech modality into text language model remains an open question.

Figure 2 shows the overview of E2E system architecture. The E2E SpeechQE model has three parts: pre-trained speech encoder, modality adapter, and pre-trained text-LLM. The speech encoder extracts the audio feature from the raw audio, where we initialize with existing competitive speech models. The modality adapter subsamples the audio features to compress the audio sequence and bridges the speech representation to the text embedding space to output speech embeddings. We fix the speech encoder for all experiments, while the weights of the adapter and text-LLM can be updated depending on the training settings. The input of the text-LLM model is the concatenation of text and audio embedding sequence.

Training Supervised SpeechQE training and evaluation requires triplets of audio inputs, ST hypotheses, and quality ratings. We build a corpus by generating hypotheses with direct ST systems of varying quality and obtain automatic quality labels from reference-based metric (§ 4.1).³ We train the E2E model with the SpeechQE task, complemented with the ASR and ST tasks which provide supervision of mapping between text and speech modality. We consider two training strategies. The first is a simple single-phase approach where we train a modality adapter (and optionally update

³This is intended to minimize any bias from the written text domain, rather than augment speech modality with TTS on existing text datasets with human scores.

CoVoST2/CV4	ASR	ST	SpeechQE
es2en	297k	79k	546k
en2de	305k	290k	589k

Table 1: Number of instances of training corpus of each speech related tasks. CoVoST2 for ST and SpeechQE, and Common Voice 4 for ASR. SpeechQE set is generated from the subset of ST by seven translation systems.

Es2En direct ST systems	CoVoST2	FLEURS
whisper-large-v3	39.05	22.45
whisper-large-v2	39.53	23.62
whisper-large	38.11	22.89
whisper-medium	37.39	21.93
whisper-small	31.27	17.78
whisper-base	16.93	11.67
whisper-tiny	7.81	6.86
En2De direct ST systems	CoVoST2	FLEURS
seamless-m4t-v2-large	43.12	32.21
seamless-m4t-large	40.55	31.41
seamless-m4t-medium	38.39	26.83
s2t-wav2vec2-large-en-de	26.98	19.92
s2t-medium-mustc-multilingual-st	8.08	13.43
s2t-small-mustc-en-de-st	7.82	12.34
s2t-small-covost2-en-de-st	14.19	9.50

Table 2: The list of seven direct ST models and their BLEU scores for generating training corpus and test benchmarks of SpeechQE.

text-LLM) with all three tasks. The second is a two-phase approach where we first train only an adapter with ASR and ST tasks while freezing text-LLM to focus solely on mapping between text and speech modality. Then, we continue training with the SpeechQE task to let the LLM learn the unseen task of QE. In the second phase, the adapter pre-trained in the previous phase can be frozen or updated, while text-LLM is always trained with LoRA (Hu et al., 2022).

We now turn to the empirical evaluation to determine whether the E2E model successfully overcomes the efficiency and modality alignment issues raised by cascaded systems.

4 Experimental Settings

In this section, we describe the construction of the SpeechQE benchmark as well as the configuration of the evaluated systems.

4.1 Building SpeechQE Benchmark

We build a training corpus and test benchmark for SpeechQE from CoVoST2 (Wang et al., 2021) which is a speech translation corpus based on Common Voice 4 ASR datasets (Ardila et al., 2020). We consider two translation directions: Spanish-to-English and English-to-German. We subsample about 80k segments from the training set and 500 from the dev and test of CoVoST2, then run seven different direct ST models to generate the ST hypotheses. The direct ST models are off-the-shelf models of a wide range of translation quality including Whisper (Radford et al., 2022) for Es2En, and Seamless-M4T (Seamless Communication et al., 2023) and Fairseq S2T (Wang et al., 2020) for En2De. The details of ST models are in Table 2.

Given the generated hypothesis text, reference text, and gold transcription text, we get automatic quality labels from (reference-based) metrics since reference-based scores are generally known to be better correlated with human judgment on translation quality than reference-free scores (Freitag et al., 2023). For training, we choose xCOMET-XL (Guerreiro et al., 2024) as metric because it is one of the best-performing submissions in the WMT23 metric shared task. The final statistics for the training dataset are in Table 1. For the test, we obtain metric scores from both xCOMET-XL and MetricX-23-XL (Juraska et al., 2023) as two distinct types of quality labels to avoid biased comparison with the cascaded system.

4.2 Cascaded Modeling

For the cascaded system, we use the same set of Whisper models that generates the Es2En ST hypothesis as the ASR module for both the Es2En and En2De cascaded experiments. For QE modules, we use the same metric models that generate reference-based quality labels in Section 4.1 but with reference-free inputs: source and hypothesis.⁴

4.3 E2E Modeling

We initialize the speech encoder from Whisper-large-v2 and freeze it for all experiments. The text-LLM is TowerInstruct-7B (Alves et al., 2024) which is continued pre-training and finetuned with instructions relevant to translation processes from

⁴We choose to report QE decoding of MetricX-23-XL instead of the dedicated QE model of MetricX-23-QE-XL as the former has higher correlations with human DA and the findings in the Results sections are the same.

$\rho = \text{corr}(\mathbf{q}, \mathbf{m})$	$\mathbf{m}_{\text{xCOMET}} = \text{xCOMET}(\text{gold } t, h, r)$	Es2En		En2De	
	$\mathbf{m}_{\text{MetricX}} = \text{MetricX}(h, r)$	$\mathbf{m}_{\text{xCOMET}}$	$\mathbf{m}_{\text{MetricX}}$	$\mathbf{m}_{\text{xCOMET}}$	$\mathbf{m}_{\text{MetricX}}$
Cascaded SpeechQE Systems Correlations $\rho_{cas} = \text{corr}(\mathbf{q}_{cas}, \mathbf{m})$					
$\mathbf{q}_{cas} = \text{xCOMET-qe}(\text{gold } t, h)$		0.929	<u>0.812</u>	0.967	<u>0.872</u>
$\mathbf{q}_{cas} = \text{xCOMET-qe}(\text{ASR}(a), h)$		<u>0.892</u>	0.782	<u>0.910</u>	0.821
$\mathbf{q}_{cas} = \text{MetricX-qe}(\text{gold } t, h)$		<u>0.834</u>	<u>0.844</u>	<u>0.908</u>	0.932
$\mathbf{q}_{cas} = \text{MetricX-qe}(\text{ASR}(a), h)$		0.803	0.803	0.854	<u>0.871</u>
$\mathbf{q}_{cas} = \text{text-BLASER2.0-qe}(\text{gold } t, h)$		<u>0.813</u>	<u>0.739</u>	<u>0.870</u>	<u>0.833</u>
$\mathbf{q}_{cas} = \text{text-BLASER2.0-qe}(\text{ASR}(a), h)$		0.776	0.711	0.813	0.771
End-to-End SpeechQE Systems Correlations $\rho_{e2e} = \text{corr}(\mathbf{q}_{e2e}, \mathbf{m})$					
$\mathbf{q}_{e2e} = \text{BLASER2.0-qe}(a, h)$		0.780	0.712	0.856	0.819
$\mathbf{q}_{e2e} = \text{TowerInstruct-Fixed+Adapter}(a, h)$		0.862	0.797	0.882	0.848
$\mathbf{q}_{e2e} = \text{TowerInstruct-LoRA+Adapter}(a, h)$		0.882	0.818	0.914	0.867
$\mathbf{q}_{e2e} = \text{TowerInstruct-LoRA+Adapter-pt}(a, h)$		0.890	0.833	0.922	0.872
$\mathbf{q}_{e2e} = \text{TowerInstruct-LoRA+Adapter-pt-Fixed}(a, h)$		0.895	0.834	0.925	0.873

Table 3: Correlations (ρ) between SpeechQE system scores (\mathbf{q}) and metric scores (\mathbf{m}) for quality of ST on CoVoST2 test. ASR is whisper-large-v3, the cutting-edge model. E2E systems outperform ASR cascaded systems and even some cascaded ones with gold transcriptions. Overlines in cascaded correlation mean that the best E2E system outperforms the corresponding cascaded system. Bolded text in E2E indicate the best score within each column.

Llama 2 (Touvron et al., 2023). This model has not trained on the task of predicting the quality score of a given translation (QE) but has trained on the error span detection task. We either freeze the TowerInstruct model or train it with LoRA ($r = 16$, $\alpha = 32$). The modality adapter consists of three 1-dimensional convolutional layers followed by a 512-dimensional bottleneck layer (Houlsby et al., 2019), following Wang et al. (2023a). The adapter is initialized randomly and unfrozen unless stated.

All our E2E models are trained on a single A6000 GPU with a batch size of 8 updated in fixed steps (140k steps for the single phase strategy, and 120k+80k steps for the two-phase strategy). In addition to the SpeechQE training set, we use Common Voice 4 and CoVoST2 for ASR and ST. We use language modeling loss with fixed instruction prompts for each task for all settings, following the chat template of TowerInstruct. More experimental details are in Appendix D including the instruction prompt templates for each task (Figure 3).

As another baseline, we use the BLASER2.0-qe to experiment with both cascaded and E2E scenarios. The inputs of E2E setting are SONAR embedding of source speech and target text, while all text embedding is for the cascaded setting.

4.4 Evaluation

We evaluate all models on the SpeechQE test set built in Section 4.1, which has two types of met-

ric labels from xCOMET-XL and MetricX-XL. A lower score of MetricX indicates better quality, while that of xCOMET and E2E systems indicates the opposite. To simplify our analysis, we multiply MetricX scores by negative one, which allows us to focus on the extent of correlation without considering the direction. We use the Spearman as the primary measurement following Blain et al. (2023).

For evaluation on quality labels by human judgement instead of metric, we compare human direct assessment (DA) score on IWSLT ACL set from Sperber et al. (2024) which is based on Salesky et al. (2023).⁵ This dataset is based on presentation videos describing their ACL papers, thus including highly technical terms and having domain mismatches between our main training corpus. It contains the source-based DA ratings of 416 hypotheses from each of the ten ST systems, resulting in a total of 4,160 instances. We include additional QE and metric models including sentence BLEU and Comet(KiWi) (Rei et al., 2022a,b, 2023).

5 Results

We first present our main results by comparing SpeechQE ratings with reference-based metrics (§ 5.1), then turn to using human ratings of translation quality (§ 5.2). We add the results of varying model sizes and architecture of the cascaded sys-

⁵<https://huggingface.co/datasets/IWSLT/da2023>

tem. (§ 5.3). Finally, we evaluate our models on a zero-shot error detection task (§ 5.4) and conduct a qualitative analysis of outputs (§ 5.5). We additionally evaluate and train our systems with out-of-domain settings (Appendix A and B).

5.1 Correlation with Reference-based Metrics

Table 3 shows correlations between metric scores as quality labels and SpeechQE system output scores, where the input of metric includes gold transcription source text and reference text.

Cascaded. For metric and text-QE scores, we cross-compare two metric scores (xCOMET and MetricX) as quality labels and two QE scores (xCOMET-qe and MetricX-qe) within cascaded configurations since the matching QE and metric model could favor the output from the model similar to its own. For example, the xCOMET is a single model for both metric and QE with different inputs, showing higher correlation values in the metric-QE model matching configuration (0.929 in Es2En) than mismatch (0.834 or 0.812).

E2E. Among four E2E models, LoRA training the text-LLM with a fixed pre-trained speech adapter (*TowerInstruct-LoRA+Adapter-pt-Fixed*) performs the best in all language pairs and metric types. The simplest training of fixing LLM and updating only the adapter with all three tasks in a single phase (*TowerInstruct-Fixed+Adapter*) shows the lowest correlations followed by similar methods but LoRA training the text-LLM (*TowerInstruct-LoRA+Adapter*). This suggests that a separate training phase for mapping speech-to-text perception is critical and that the weight updates are necessary when a text-LLM is not fine-tuned for the target task and therefore lacks the required capabilities. In this case, TowerInstruct is not fine-tuned with QE tasks, therefore, updating it is necessary. All variants of our E2E system outperform BLASER2.0, perhaps due to its limited exposure to diverse translation quality at training time.

E2E vs Cascaded. The end-to-end SpeechQE systems consistently outperform the cascaded system which included the SOTA ASR system (whisper-large-v3). The best E2E system not only outperforms ASR-based cascades, but cascaded systems that use gold transcriptions in all QE(row)-metric(column) mismatched settings of both language pairs. For instance, 0.834 of E2E versus 0.812 of xCOMET-qe(gold t, h) cascaded in Es2En

IWSLT23-ACL En2De Test set $\rho = corr(\mathbf{x}, \mathbf{d})$	Human DA score d
Metric and Human DA correlation $\rho = corr(\mathbf{m}, \mathbf{d})$	
$m = \text{xCOMET}(\text{gold } t, h, r)$	0.557
$m = \text{MetricX}(h, r)$	0.539
$m = \text{wmt22-comet-da}(\text{gold } t, h, r)$	0.544
$m = \text{sentBLEU}(h, r)$	0.336
Cascaded SpeechQE and Human DA $\rho = corr(\mathbf{q}_{cas}, \mathbf{d})$	
$q = \text{xCOMET-qe}(\text{gold } t, h)$	0.544
$q = \text{MetricX-qe}(\text{gold } t, h)$	0.556
$q = \text{wmt23-cometkiwi-da-xl}(\text{gold } t, h)$	0.576
$q = \text{wmt22-cometkiwi-da}(\text{gold } t, h)$	0.580
$q = \text{xCOMET-qe}(\text{ASR}(a), h)$	0.485
$q = \text{MetricX-qe}(\text{ASR}(a), h)$	0.495
$q = \text{wmt23-cometkiwi-da-xl}(\text{ASR}(a), h)$	0.503
$q = \text{wmt22-cometkiwi-da}(\text{ASR}(a), h)$	0.486
$q = \text{text-BLASER2.0-qe}(\text{ASR}(a), h)$	0.428
E2E SpeechQE & Human DA correlation $\rho = corr(\mathbf{q}_{e2e}, \mathbf{d})$	
$q = \text{BLASER2.0-qe}(a, h)$	0.420
$q = \text{TowerInst-LoRA+Adapter-pt}(a, h)$	0.492
$q = \text{TowerInst-LoRA+Adapter-pt-Fixed}(a, h)$	0.509

Table 4: Correlations (ρ) between human direct assessment scores (\mathbf{d}) from IWSLT23-ACL and metric/QE scores (\mathbf{m} or \mathbf{q}) for English-to-German speech translation. E2E SpeechQE scores correlate better with human labels than cascaded approaches.

MetricX column. Similarly, BLASER2.0 with the E2E setting of speech input and text output outperforms the cascade system with the text input-output setting (text-BLASER2.0).

Overall, the correlation analysis underscores the advantage of end-to-end SpeechQE systems over cascaded ones. The strong correlations with metric scores across various configurations indicate its reliability as a measurement for quality estimation in automatic speech translation tasks, highlighting the potential of end-to-end approaches.

5.2 SpeechQE Correlation with Human DA

In Table 4, we compare the output quality scores from SpeechQE systems with human direct assessment (DA) scores from the IWSLT-ACL test set, instead of metric scores as in the previous sections. We use the ASR output provided by Salesky et al. (2023).⁶ Overall correlations in the IWSLT-ACL setting are lower compared to the prior section.

⁶We tried Whisper ASR systems, but the output quality was not acceptable, likely due to the IWSLT23-ACL set being out-of-domain and covering highly technical NLP topics. The ASR provided is Azure API speech-to-text service, which we believe performs comparably to SOTA ASR models.

$\rho = \text{corr}(\mathbf{q}, \mathbf{m} \text{ or } \mathbf{d})$	CoVoST2 Es2En Test				IWSLT23
	$m_{\text{xCOMET-XL}}$	$m_{\text{xCOMET-XXL}}$	$m_{\text{MetricX-XL}}$	$m_{\text{MetricX-XXL}}$	En2De d
Cascaded Model with XXL Size vs E2E speech-LLM					
$q_{cas} = \text{ASR (1.5B)} \rightarrow \text{xCOMET-XL-qe (3.5B)}$	0.892	0.800	0.782	0.788	0.485
$q_{cas} = \text{ASR (1.5B)} \rightarrow \text{xCOMET-XXL-qe (10.7B)}$	0.787	0.873	0.708	0.734	0.486
$q_{cas} = \text{ASR (1.5B)} \rightarrow \text{MetricX-XL-qe (3.7B)}$	0.803	0.758	0.803	0.766	0.495
$q_{cas} = \text{ASR (1.5B)} \rightarrow \text{MetricX-XXL-qe (13B)}$	0.700	0.677	0.652	0.694	0.502
Cascaded text-LLM vs E2E speech-LLM					
$q_{cas} = \text{ASR (1.5B)} \rightarrow \text{text-TowerInstruct-LoRA (7B)}$	0.852	0.816	0.780	0.785	–
$q_{e2e} = \text{TowerInstruct-LoRA+Adapter-pt-Fixed (7.5B)}$	0.895	0.827	0.834	0.834	0.509

Table 5: Impact of model size and architecture choices. The table reports correlations (ρ) between SpeechQE system scores (\mathbf{q}) and either metric scores (\mathbf{m}) or human direct assessment scores (\mathbf{d} , right-most column). Regardless of the size of the text-QE model, the E2E SpeechQE system mostly outperforms the cascaded system. Also, the cascaded system with a similar architecture of text-LLM shows lower performance than E2E SpeechQE system.

We hypothesize that this may be due in part to the out-of-domain nature of this test set (NLP technical talks), and to the fact that the direct assessment task performed by human judges differs from the tasks performed to obtain the gold ratings that informed our QE and metric model (MQM and WMT DA).

Metric vs Gold-QE. The best correlation between human DA and cascaded text-QE with gold transcription (0.580) shows a higher coefficient than the best metric-human correlation (0.557), unlike the assumptions that metric scores would better correlate with human scores as in Freitag et al. (2023). This could result from the annotation process, such as source-based DA, where annotators are shown the source text and the translated target text but not the reference text, or they are shown re-segmented translation system output along with the previous and next system outputs as described in Sperber et al. (2024).

E2E vs Cascaded. The best E2E SpeechQE system outperforms all ASR cascaded systems in correlation with human DA. The ASR + WMT23-CometKiWi combination shows the highest correlation among the ASR-based configurations (0.503), but it is still slightly lower than the best E2E system (0.509). Notably, this best E2E system is also the top performer in the previous section. Overall, the data suggests that the best-practice E2E system is more effective in aligning with human judgments on translation quality compared to all cascaded systems with ASR.

5.3 Cascaded Model Size and Architecture

Is the dominance of E2E over cascaded models due to the E2E parameter size rather than its end-to-end nature? We address this question by varying the model size and architectural similarity between the cascaded and E2E SpeechQE system.

Cascaded with XXL Size. In Table 5, we evaluate cascaded systems based on bigger text-QE models—*text-TowerInstruct-qe*(7B), *xCOMET-XXL-qe* (10.7B), and *Metric-23-XXL-qe* (13B)—resulting in cascaded SpeechQE systems whose total size is bigger than that of E2E (e.g. total 14.5B of cascaded MetricXXL vs 7.5B of E2E). We also extend the size of metric models in the CoVoST2 comparison. The larger text-QE system generally correlates better with human quality score than smaller cascaded system (rightmost column); however, the performance is still below that of the E2E. Similarly in CoVoST2 test results, the E2E system outperforms the cascaded system regardless of the size of the text-QE model, except for the case where *xCOMET-XXL* metric favors the QE scores of the same model.

Overall, E2E models tend to show a higher correlation than the cascaded systems with similar/bigger-sized text-QE models, showing the advantages of the E2E system extend across efficiency considerations.

Cascaded with text-LLM. We LoRA fine-tune the *TowerInstruct* model in Spanish-to-English direction with similar training methods to E2E SpeechQE model but only with text modality input. This produces a text-based QE model based on the same *TowerInstruct-7B* model as the E2E

ESD for ST	Precision	Recall	F1 Score
<i>Cascaded Systems</i>			
txt-ESD(gold t, h)	0.438	0.591	0.503
txt-ESD(w-large-v2(a, h))	0.434	0.550	0.485
txt-ESD(w-medium(a, h))	0.429	$\overline{0.540}$	0.478
txt-ESD(w-small(a, h))	$\underline{0.413}$	0.535	$\overline{0.466}$
txt-ESD(w-base(a, h))	0.385	0.550	0.453
<i>End-to-End Systems</i>			
<i>TowerInst-Fixed+Adt</i> (a, h)	0.411	0.542	0.467

Table 6: Zero-shot error span detection for speech translation (SpeechESD) on CoVoST2 Spanish-to-English test. Even without being explicitly trained by the SpeechESD task, E2E model performs decently suggesting that text-LLM ability is transferable to speech LLM in a zero-shot manner.

SpeechQE model. Pairing it with ASR results in a cascaded SpeechQE system with 8.5B parameters as opposed to 7.5B for the E2E system. Yet, the E2E system still outperforms this version of cascaded model. Besides the efficiency advantage, we can also conclude that the improvements are coming from the E2E nature of the approach rather than the LLM-based solution, reaffirming that E2E system is better suited for SpeechQE task than the cascaded system.

5.4 Zero-Shot Error Span Detection for ST

Simply providing the quality score may offer a straightforward indication of translation quality, but it can be difficult to interpret when trying to identify specific issues (Lu et al., 2024). To broaden the scope of QE beyond overall numerical ratings, we further explore an error span detection (ESD) for ST task (SpeechESD) that predicts the error span within the hypothesis (Blain et al., 2023).

We test our E2E model in a zero-shot manner where SpeechESD is an unseen task during the speech adaptation. Since the TowerInstruct is fine-tuned from its base model with several translation-related tasks including error span detection, we can see how effectively the method of injecting speech modality generalizes the capability of text-LLM to speech LLM without explicitly training the target speech task. We evaluate quantitatively in this section and also qualitatively in Section 5.5.

Experimental Settings. We use the error span output of the xCOMET metric function as reference-based error span labels and compare the E2E and cascaded system where TowerInstruct is

a text-ESD model.⁷ We use the same test set as SpeechQE. The input of the ESD task is source and hypothesis as in the QE task. We calculate the F1 score following Blain et al. (2023). For the E2E model, we only run the model that fixes the text-LLM, as the model performs exclusively on a few trained tasks when the weights of text-LLM are updated with those tasks. Also, we build an additional SpeechQE train set from FLEURS train set (Conneau et al., 2022) and include it into a single phase SpeechQE training to have better meaningful results in ESD, especially in qualitative analysis.

E2E vs Cascaded. We show F1 score, recall, and precision in Table 6. Cascaded systems show the best performance in SpeechESD indicating that they remain the preferred choice for achieving the highest performance when we do not have speech training data for the target task. Still, even without being explicitly trained by the SpeechESD task, the E2E model performs decently by outperforming cascaded with medium-quality ASR in recall and cascaded with whisper-small in F1-score. This suggests that text-LLM ability is transferable to speech LLM in a zero-shot manner.

5.5 Example Analysis

We analyze the examples of how E2E and cascaded SpeechQE systems score the speech translation quality and detect the error spans. Table 7 shows examples of Spanish-to-English speech translation from whisper-large-v2 and quality estimations of SpeechQE systems, where the ASR model of the cascaded system is whisper-medium. We use xCOMET metric outputs of scores, error spans, and severity as the quality and error labels, similar to the setting of Section 5.1 and 5.4.

The example translation has two major errors in “Calpaniado” and “camp”, which are supposed to be translated into “Carpanedo” and “championship”. However, the cascaded system estimates the quality of this translation as high as 0.93, and could not detect the error spans or its severity correctly. These issues primarily arise because ASR incorrectly transcribed the name “Calpaniado” as “Calpaniado” and the word “campeonato” (meaning “championship”) as “campamento” (meaning “camp”) In contrast, E2E SpeechQE system is not affected by these issues and correctly detects those major errors. We

⁷We did not compare with text-xCOMET-qe in this case as we are not training SpeechESD explicitly like SpeechQE and xCOMET-qe output are similar to that of xCOMET-metric.

Spanish-to-English ST Example		
Gold transcription	<u>Carpanedo</u> participó en dos carreras individuales del <u>campeonato</u> aparte de la competencia del miércoles.	
ASR	Calpaniado participó en dos carreras individuales del campamento , aparte de las competencias del miércoles.	
Hypothesis	Calpaniado participated in two individual races of the camp , apart from the Wednesday races.	
Reference	Beyond Wednesday’s event, <u>Carpanedo</u> competed in two individual races at the <u>Championships</u> .	
Systems	SpeechQE Scores	Error Span Detection
Quality/Error Span Labels	0.611	Calpaniado – major, of the camp – major, <i>aces</i> –major
Cascaded Predictions	0.932	camp –minor, <i>ace</i> –minor
E2E Predictions	0.497	Calpaniado – major, camp – major

Table 7: Example of Spanish-to-English speech translation and quality estimations of SpeechQE systems. Bolded text represents the wrong ASR or ST spans while underlined indicates the correct ones. Cascaded SpeechQE incorrectly estimates the translation quality of the hypothesis due to speech recognition error, while E2E could correctly catch the errors in the ST.

discuss another example of En2De in Appendix C.

This example shows that the E2E system is more robust to speech representation error in estimating quality and indicating the error spans for ST.

6 Related Work

Recent work has explored how to inject additional modalities into a model pre-trained on a single modality. Various configurations have been proposed to meet different demands including speech modality into text-LLM (Wu et al., 2023; Wang et al., 2023a,b), visual modality into text-LLM (Liu et al., 2023; Li et al., 2023), visual modality into speech foundation model (Seo et al., 2023; May et al., 2023; Han et al., 2024), and audio-visual modalities into text-LLM (Zhang et al., 2023).

When injecting the speech modality into text-LLM, the main challenges are aligning long speech signals to corresponding text sequences with the same semantic contents, while avoiding overfitting to default training tasks like ASR and ST. Several methods of compressing and aligning the speech and text sequence include the use of convolutional layer (Wang et al., 2023a), CTC compression (Wu et al., 2023; Pan et al., 2023), and random down-sampling (Wang et al., 2023b). Many mention the problem of task overfitting to homogeneous fixed instruction training on limited tasks. They suggest training on many diverse tasks (Chu et al., 2023; Tang et al., 2024) or tuning on diverse speech instructions with TTS-augmented instruction datasets (Wang et al., 2023b; Pan et al., 2023).

However, most of these works focus on ASR, ST, QA, and general instruction following within

speech comprehension tasks (Gaido et al., 2024). This paper initiates their application to the understudied SpeechQE problem.

7 Conclusion

This work focused on the task of SpeechQE, evaluating the quality of speech translation using both cascaded systems and end-to-end systems. We developed an E2E SpeechQE model, proposing methods for corpus creation, training strategies, and architectural design. Our findings indicate that E2E systems are generally better suited to estimate the quality of direct speech translation. Additionally, we examined the error span detection task for ST finding that E2E speech model transfer ability from text-based LLM while cascaded systems with state-of-the-art ASR still hold advantages in performance. We conclude that SpeechQE needs dedicated attention separate from text-QE, due to the growing use cases of ST and the significant potential for further improvements in this field.

Quality estimation in the speech domain opens up a wide range of potential applications. In addition to the promise of helping people use speech translation systems more reliably in their daily lives, quality estimation can enhance speech translation itself, for instance by enabling prefix-to-prefix quality estimation for re-translation and simultaneous speech translation. We contribute data, code, and models to support future work that broadens the scope of the translation-related tasks for the speech domain.

Limitations

This work assumes that we can use quality evaluation schemes designed for text translation and port them directly to speech to distill the quality estimation ability while adapting it to the speech domain. However, some errors might matter more when translating text than when translating speech (e.g., punctuation, capitalization), while speech inputs might raise new issues (e.g., segmentation). In future work, we encourage the collection of quality annotations specifically designed for speech translation and look forward to investigating how to transfer knowledge from text-QE systems in those settings.

Our E2E models are trained with an A6000 GPU with 8 instances per batch updating up to 200k steps. Training with larger number of GPUs and batch size, as is often the case with speech LLM training, could show better performance in SpeechQE.

Our training tasks include ASR, ST, and SpeechQE with fixed instructions which interfere with the success of downstream zero-shot tasks like error span detection. Further augmenting the training tasks with speech instruction tuning and diverse speech question answering tasks could enhance the performance of ESD.

We experimented with two language pairs, English-to-German and Spanish-to-English, both of which are European languages. We could expand language diversity in future work by including non-European languages, which would help assess the generalizability and robustness of our models across different linguistic and cultural contexts.

We have explored a single type of architecture for speech LLM. Investigating various architectural approaches could help better understand their impact on performance and robustness in SpeechQE performance and transferability of knowledge.

Acknowledgments

This work was supported, in part, by the Human Language Technology Center of Excellence at Johns Hopkins University. We also extend our gratitude to the team of the SCALE 2023 workshop on Translation of Conversational Speech, whose findings and resources gave us a headstart on this project. Finally, we thank the anonymous reviewers, Nishant Balepur, Xinchun Yang, Dayeon Ki, and the members of the CLIP lab at UMD for their insightful and constructive feedback.

References

- Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. [Tower: An open multilingual large language model for translation-related tasks](#). In *First Conference on Language Modeling*.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. [Common voice: A massively-multilingual speech corpus](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.
- Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. [Findings of the WMT 2023 shared task on quality estimation](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. Association for Computational Linguistics.
- Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. [Findings of the 2013 Workshop on Statistical Machine Translation](#). In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.
- Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. [Findings of the 2012 workshop on statistical machine translation](#). In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.
- 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.
- Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. [Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models](#). *Preprint*, arXiv:2311.07919.
- Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. [FLEURS: Few-shot learning evaluation of universal representations of speech](#). *arXiv preprint arXiv:2205.12446*.

- Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. [Sonar: Sentence-level multimodal and language-agnostic representations](#). *Preprint*, arXiv:2308.11466.
- Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Jun-teng Jia, Yuan Shangguan, Ke Li, Jinxi Guo, Wenhan Xiong, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. 2023. [Prompting large language models with speech recognition abilities](#). *Preprint*, arXiv:2307.11795.
- Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. [Quality-aware decoding for neural machine translation](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.
- Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. [Findings of the WMT 2019 shared tasks on quality estimation](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.
- Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. [Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.
- Marco Gaido, Sara Papi, Matteo Negri, and Luisa Bentivogli. 2024. [Speech translation with speech foundation models and large language models: What is there and what is missing?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14760–14778, Bangkok, Thailand. Association for Computational Linguistics.
- Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. [xcomet: Transparent Machine Translation Evaluation through Fine-grained Error Detection](#). *Transactions of the Association for Computational Linguistics*, 12:979–995.
- HyoJung Han, Mohamed Anwar, Juan Pino, Wei-Ning Hsu, Marine Carpuat, Bowen Shi, and Changhan Wang. 2024. [XLAVS-R: Cross-lingual audiovisual speech representation learning for noise-robust speech perception](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12896–12911, Bangkok, Thailand. Association for Computational Linguistics.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. [Parameter-efficient transfer learning for NLP](#). In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. [MetricX-23: The Google submission to the WMT 2023 metrics shared task](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.
- Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. [OpenKiwi: An open source framework for quality estimation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.
- Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. [Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation](#). In *Proceedings of the Second Conference on Machine Translation*, pages 562–568, Copenhagen, Denmark. Association for Computational Linguistics.
- Ngoc-Tien Le, Benjamin Lecouteux, and Laurent Besacier. 2016. [Joint ASR and MT features for quality estimation in spoken language translation](#). In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. [BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. [Error analysis prompting enables human-like translation evaluation in large language models](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8801–8816, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

- Avner May, Dmitriy Serdyuk, Ankit Parag Shah, Otavio Braga, and Olivier Siohan. 2023. [Audio-visual fine-tuning of audio-only asr models](#). *Preprint*, arXiv:2312.09369.
- Nikita Mehandru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine Khoong, Marine Carpuat, and Niloufar Salehi. 2023. [Physician detection of clinical harm in machine translation: Quality estimation aids in reliance and backtranslation identifies critical errors](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11633–11647, Singapore. Association for Computational Linguistics.
- Matteo Negri, Marco Turchi, José G. C. de Souza, and Daniele Falavigna. 2014. [Quality estimation for automatic speech recognition](#). In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1813–1823, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.
- Jing Pan, Jian Wu, Yashesh Gaur, Sunit Sivasankaran, Zhuo Chen, Shujie Liu, and Jinyu Li. 2023. [Cosmic: Data efficient instruction-tuning for speech in-context learning](#). *Preprint*, arXiv:2311.02248.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#). *arXiv preprint*.
- 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, ICML’23*. JMLR.org.
- 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. 2022a. [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.
- Ricardo Rei, Nuno M. Guerreiro, Josã© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. [Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. [CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirović, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Frank. 2023. [Audiopalm: A large language model that can speak and listen](#). *Preprint*, arXiv:2306.12925.
- Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. 2023. [Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology](#). In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 62–78, Toronto, Canada (in-person and online). Association for Computational Linguistics.
- Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilya Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. [SeamlessM4T: Massively Multilingual & Multimodal Machine Translation](#). *Preprint*, arXiv:2308.11596.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. [BLEURT: Learning robust metrics for text generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

- Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. 2023. [Avformer: Injecting vision into frozen speech models for zero-shot av-asr](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22922–22931.
- L. Specia, C. Scarton, and G.H. Paetzold. 2022. *Quality Estimation for Machine Translation*. Synthesis Lectures on Human Language Technologies. Springer International Publishing.
- Lucia Specia, Dhvaj Raj, and Marco Turchi. 2010. [Machine translation evaluation versus quality estimation](#). *Machine Translation*, 24(1):39–50.
- Matthias Sperber, Ondřej Bojar, Barry Haddow, Dávid Javorský, Xutai Ma, Matteo Negri, Jan Niehues, Peter Polák, Elizabeth Salesky, Katsuhito Sudoh, and Marco Turchi. 2024. [Evaluating the IWSLT2023 speech translation tasks: Human annotations, automatic metrics, and segmentation](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6484–6495, Torino, Italia. ELRA and ICCL.
- Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. [SALMONN: Towards generic hearing abilities for large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiohu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. [Fairseq S2T: Fast speech-to-text modeling with fairseq](#). In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39, Suzhou, China. Association for Computational Linguistics.
- Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. [CoVoST 2 and Massively Multilingual Speech Translation](#). In *Proc. Interspeech 2021*, pages 2247–2251.
- Chen Wang, Minpeng Liao, Zhongqiang Huang, Jintian Lu, Junhong Wu, Yuchen Liu, Chengqing Zong, and Jiajun Zhang. 2023a. [Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing](#). *Preprint*, arXiv:2309.00916.
- Mingqiu Wang, Wei Han, Izhak Shafran, Zelin Wu, Chung-Cheng Chiu, Yuan Cao, Nanxin Chen, Yu Zhang, Hagen Soltau, Paul K. Rubenstein, Lukas Zilka, Dian Yu, Golan Pundak, Nikhil Siddhartha, Johan Schalkwyk, and Yonghui Wu. 2023b. [SIm: Bridge the thin gap between speech and text foundation models](#). In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.
- Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linqun Liu, and Yu Wu. 2023. [On decoder-only architecture for speech-to-text and large language model integration](#). In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.
- Hang Zhang, Xin Li, and Lidong Bing. 2023. [Video-LLaMA: An instruction-tuned audio-visual language model for video understanding](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore. Association for Computational Linguistics.

Out-of-Domain Test set (FLEURS)	Es2En	
	m_{xCOMET}	m_{MetricX}
Cascaded SpeechQE Systems $\rho_{cas} = \text{corr}(\mathbf{q}_{cas}, \mathbf{m})$		
xCOMET-qe(gold t, h)	0.945	<u>0.849</u>
xCOMET-qe(whspr-large-v3(a, h))	0.919	0.824
xCOMET-qe(whspr-large-v2(a, h))	0.919	0.825
xCOMET-qe(whspr-medium(a, h))	<u>0.906</u>	0.813
xCOMET-qe(whspr-small(a, h))	0.895	0.804
xCOMET-qe(whisper-base(a, h))	0.852	0.776
<hr/>		
MetricX-qe(gold t, h)	<u>0.855</u>	<u>0.893</u>
MetricX-qe(whspr-large-v3(a, h))	0.834	0.858
MetricX-qe(whspr-large-v2(a, h))	0.833	0.860
MetricX-qe(whspr-medium(a, h))	0.815	0.840
MetricX-qe(whspr-small(a, h))	0.791	0.810
MetricX-qe(whspr-base(a, h))	0.709	0.726
<hr/>		
End-to-End SpeechQE Systems $\rho_{e2e} = \text{corr}(\mathbf{q}_{e2e}, \mathbf{m})$		
TowerInst-LoRA+Adapter-pt(a, h)	0.897	0.858
TowerInst-LoRA+Adt-pt-Fixed(a, h)	0.892	0.849
Adding FLEURS to E2E Training		
TowerInst-LoRA+Adapter-pt(a, h)	0.904	0.872
TowerInst-LoRA+Adt-pt-Fixed(a, h)	0.906	0.873

Table 8: Correlations on out-of-domain (OOD) test set of Spanish-to-English FLEURS. Cascaded shows better audio domain robustness than E2E as E2E models are trained on limited data. Still, E2E outperforms gold-cascaded when compared with cross QE-metric cascade configuration in different model families. We also experiment with additional FLEURS training, which increases (now in-domain) FLEURS test correlation score.

A Robustness to Out-of-Domain Test Sets

We also explore how the SpeechQE systems are robust to the domain changes. We build a test set with FLEURS (Conneau et al., 2022) for out-of-domain (OOD) evaluation following the same protocol as an in-domain test set. Table 8 shows correlations between SpeechQE system score and metric score on the out-of-domain test set of FLEURS.

Effect of ASR quality in Cascaded. We present cascaded results with a wide quality range of ASR, from whisper-large v3 to whisper-base. The correlations are proportional to the ASR performances, while gold cascaded is an upper bound.

Robustness Effect of Training E2E Adapter with Target Task In contrast to Section 5.1, the best-performing E2E model is the model that updates the pre-trained adapter weight in the final training stage with the SpeechQE task. We note that the training of the adapter and the final E2E model is

based solely on Common Voice audio, where the adapter is trained with ASR and ST tasks and the final E2E model is only trained with the SpeechQE.

We conclude that E2E models become more robust to audio domain shift if the speech adapter is trained with the target task—SpeechQE in this case—instead of being frozen.

E2E vs Cascaded. The results suggest that cascaded systems have better domain robustness when comparing the correlation between matching QE and metric models like the pair of ASR + xCOMET-qe and xCOMET metric scores. In those cases, the E2E system (e.g. 0.858 in MetricX) only outperforms the cascaded system with medium-quality ASR systems (e.g. 0.840 with whisper-medium ASR). This advantage is likely due to ASR systems being trained on a broader domain of audio corpora, whereas E2E systems are limited to Common Voice domain. Nevertheless, the E2E system shows competitive correlations in settings with non-matching QE and metric models (e.g., xCOMET-qe and MetricX metric), outperforming the cascaded systems of gold transcription and text-QE.

B Adding FLEURS set to E2E Training

Training a model on a single speech domain may lead to learning domain-specific speech representation, such as particular accents or speaking styles. We experiment with an additional SpeechQE training set to verify whether the conclusion from single-domain experiments holds in broader settings. We create an additional SpeechQE training set from the FLEURS dataset (20k), which is relatively small compared to CoVoST2 (more than 500k). We include it into a single phase SpeechQE training, which is the same corpus setting described in Section 5.4. We present the evaluation results on CoVoST2 and IWSLT23-ACL in Table 9 and on FLEURS in Table 8, specifically in the last two rows of each table.

First, adding the FLEURS domain shows higher correlations on the FLEURS domain as anticipated (last two rows of Table 8). In contrast, it reduces performance on the CoVoST2 domain but still outperforms the cascaded SpeechQE systems (Table 9). Interestingly, the correlation between the human score of IWSLT-ACL and the SpeechQE system score (rightmost column in Table 9) shows that adding even a small set from another domain slightly increases the alignment with human judgments. Although this improvement may not be

$\rho = corr(\mathbf{q}, \mathbf{m} \text{ or } \mathbf{d})$	CoVoST2 Es2En		CoVoST2 En2De		IWSLT23
	m _x COMET	m _{MetricX}	m _x COMET	m _{MetricX}	En2De d
Cascaded SpeechQE Systems Correlations					
	$\rho_{cas} = corr(\mathbf{q}_{cas}, \mathbf{m})$			$\rho_{cas} = corr(\mathbf{q}_{cas}, \mathbf{d})$	
$\mathbf{q}_{cas} = \text{xCOMET-qe}(ASR(a), h)$	0.892	0.782	0.910	0.821	0.485
$\mathbf{q}_{cas} = \text{MetricX-qe}(ASR(a), h)$	0.803	0.803	0.854	0.871	0.495
End-to-End SpeechQE Systems Correlations					
	$\rho_{e2e} = corr(\mathbf{q}_{e2e}, \mathbf{m})$			$\rho_{e2e} = corr(\mathbf{q}_{e2e}, \mathbf{d})$	
$\mathbf{q}_{e2e} = \text{TowerInstruct-LoRA+Adapter-pt}(a, h)$	0.890	0.833	0.922	0.872	0.492
$\mathbf{q}_{e2e} = \text{TowerInstruct-LoRA+Adapter-pt-Fixed}(a, h)$	0.895	0.834	0.925	0.873	0.5085
Adding FLEURS to E2E Training					
$\mathbf{q}_{e2e} = \text{TowerInstruct-LoRA+Adapter-pt}(a, h)$	0.893	0.828	0.916	0.868	0.501
$\mathbf{q}_{e2e} = \text{TowerInstruct-LoRA+Adapter-pt-Fixed}(a, h)$	0.888	0.826	0.920	0.871	0.5091

Table 9: CoVoST2 and IWSLT23-ACL results of the E2E models trained on a single-domain of CoVoST2 corpus (first two rows of E2E section) and multi-domain corpus including CoVoST2 and FLEURS (last two rows). Adding the FLEURS domain decreases performance on the CoVoST2 domain but slightly improves in correlation with IWSLT23-ACL human direct assessment scores, while still outperforming the cascaded SpeechQE system.

statistically significant, it suggests that training on multiple speech domains (CoVoST2 + FLEURS) increases robustness against domain shifts during testing (as IWSLT ACL is also out-of-domain).

In conclusion, the findings from single-domain experiments remain valid after incorporating the FLEURS set into training, while also indicating increased robustness to domain shifts.

C Additional Examples in En2De

Table 10 shows examples of English-to-German speech translation results from s2t-medium-mustc-multilingual-st in Table 2. The translation has several major errors and both cascaded and E2E systems are able to detect the errors. However, the cascaded system incorrectly predicts the severities as minor and ends up estimating the quality score to be 0.852. One could be partly due to an ASR error where it incorrectly transcribed “GBP” as “GPP”, which might trigger the cascaded system to set its severity as a minor for the translation of “GP”.

D Additional Experimental Details

For E2E training, we use a learning rate of 5e-5 and a weight decay of 0.05. For LoRA training, we update qlkvlo projection in each attention layer with the rank of $r = 16$ and a scaling parameter of $\alpha = 32$. The size of the resulting E2E SpeechQE model is about 8.5B given that TowerInstruct text-LLM is 7B and whisper-large-v2 is 1.5B. For decoding, we use a temperature of 0.1 and set the maximum new tokens up to 500. The presented numbers in all tables are a single run for cascaded

where the outputs do not change with the same input and the mean of three runs for E2E. We use off-the-shelf models from the huggingface hub and use torch and transformer libraries for the implementation.

E Discussion of the Task Terminology

In the research area of machine translation (MT), the term QE traditionally stands for machine translation quality estimation, though the more precise acronym is MTQE. Also, MT typically indicates text-to-text translation, while ST refers to speech-to-text translation. Given the implications of QE, we add “speech” to indicate the task of quality estimation for **speech** translation, where the more accurate acronym would be STQE. We use SpeechQE for speech translation quality estimation and text-QE for machine translation quality estimation as main wordings instead of (more accurate) alternatives of STQE and MTQE to emphasize the contrast between speech and text and to facilitate easier reading. While SpeechQE could be ambiguous considering that it can be QE either for ASR or ST, previous works on ASR quality estimation (Negri et al., 2014; Rubenstein et al., 2023) use the phrase “ASR-QE”, which safely distinguishes them from STQE or SpeechQE.

English-to-German ST Example		
Gold transcription	The official Falklands currency is the Falkland pound (FKP) whose value is set equivalent to that of one British pound (GBP).	
ASR	The official Falklands currency is the Falkland Pound, FKP, whose value is equivalent to that of a British Pound, <u>GPP</u> .	
Hypothesis	Die offizielle Fäklins Währung ist ein Fäklin Pfund, FKP, der uns wertvoll ist, genauso wie ein britischer Pfund, GP .	
Reference	Die offizielle Währung der Falklandinseln ist das Falkland Pound (FKP), dessen Wert in Einklang mit dem Wert des Britischen Pfunds (GBP) festgelegt wird.	
Systems	SpeechQE Scores	Error Span Detection
Quality/Error Span Labels	0.539	“e Fäklins W” – major, “hrung ist ein Fäklin Pfund, FKP, der uns wertvoll ist, genauso wie ein britischer Pfund, GP .” – major,
Cascaded Predictions	0.852	“e Fäklins Währung” – minor, “ein Fäklin Pfund” – minor, FKP – minor, “der uns wertvoll ist, genauso” – minor, “britischer Pfund, GP ” – minor
E2E Predictions	0.550	Fäklins – major, FKP – major, “uns wertvoll ist” – major, “genauso wie” – major, “britischer Pfund – major, GP – major

Table 10: Example of English-to-German speech translation and quality estimations of SpeechQE systems. Both cascaded and E2E SpeechQE systems could detect errors. However, the cascaded system estimates the severity lower than that of the metric labels partly due to ASR error while E2E could estimate the quality closely to labels.

```

# QE4ST task, training and testing
Given the German translation of the speech, estimate the quality of the translation as a
score between 0 to 1.
English: [[audio input]]
German translation: Wir modellieren den grasweisen, obstruktiven Summize-Ansatz mit zwei
verschiedenen Methoden.
# desired output in training or example output in testing
0.851

# ASR task, training
Transcribe the following audio from English into English text.
Spanish: [[audio input]]
Spanish: Durante la ocupación trabajo en teatro y filmes.

# ST task, training
Translate the following audio from Spanish into English text.
Spanish: [[audio input]]
English: During the occupation, he worked in theaters and movies.

# Error Span Detection for ST task, only testing
You are an annotator for machine translation quality. Your task is to identify errors and
assess the quality of the translation.
Source (Spanish): [[audio input]]
Translation (German): Calpaniado participated in two individual races of the camp, apart
from the Wednesday races.
Each error may consist of several consecutive words and must be categorized as either
'minor' or 'major'. Minor errors refer to smaller imperfections, and purely subjective
opinions about the translation while major errors impact the usability or understandability
of the content.
Based on the above source and translation pair, list the errors you find. If you find no
errors, simply output 'Translation has no errors.'
# example output
Calpaniado -- major
camp -- major

```

Figure 3: Prompt template of SpeechQE (quality estimation for speech translation), ASR, ST, and SpeechESD (error span detection for ST) task.