BLASER: A Text-Free Speech-to-Speech Translation Evaluation Metric

Mingda Chen, Paul-Ambroise Duquenne, Pierre Andrews, Justine Kao, Alexandre Mourachko, Holger Schwenk*, Marta R. Costa-jussà*
Meta AI
{mingdachen,padqn,mortimer,jtk,}@meta.com
{alexmourachko,schwenk,costajussa}@meta.com

Abstract

End-to-End speech-to-speech translation (S2ST) is generally evaluated with text-based metrics. This means that generated speech has to be automatically transcribed, making the evaluation dependent on the availability and quality of automatic speech recognition (ASR) systems.

In this paper, we propose a text-free evaluation metric for end-to-end S2ST, named BLASER, to avoid the dependency on ASR systems. BLASER leverages a multilingual multimodal encoder to directly encode the speech segments for source input, translation output and reference into a shared embedding space and computes a score of the translation quality that can be used as a proxy to human evaluation. To evaluate our approach, we construct training and evaluation sets from more than 40k human annotations covering seven language directions. The best results of BLASER are achieved by training with supervision from human rating scores. We show that when evaluated at the sentence level, BLASER correlates significantly better with human judgment compared to ASR-dependent metrics including ASR-SENTBLEU in all translation directions and ASR-COMET in five of them. Our analysis shows combining speech and text as inputs to BLASER does not increase the correlation with human scores, but best correlations are achieved when using speech, which motivates the goal of our research. Moreover, we show that using ASR for references is detrimental for text-based metrics. [1]

1 Introduction

Speech-to-Speech translation seeks to translate speech segments from one language into another. Historically, it has been implemented and evaluated as a concatenation of three systems: automatic speech recognition (ASR), machine translation (MT) and text-to-speech (TTS) (Lavie et al., 1997; Lazzari, 2006). In recent years, there has been increasing interest in end-to-end approaches (Jia et al., 2019; Lee et al., 2022a). While end-to-end S2ST is becoming popular, researchers still rely on text-based metrics to evaluate model performance by automatically transcribing the generated speech segments (Jia et al., 2019). These cascaded metrics rely on ASR systems, which for a given language may not have enough quality or may not even be available (Javed et al., 2022). They are also inappropriate for languages lacking standardized writing systems (Salesky et al., 2021a), like Hokkien or Algerian Arabic.

In this work, we propose the text-free metric BLASER for S2ST evaluation, sidestepping the dependency on ASR systems. In particular, we use LASER encoders that support multiple languages and modalities including text (Heffernan et al., 2022) and speech (Duquenne et al., 2021). We use the LASER encoders to directly embed speech segments into vectors and compute a score estimating the quality of generation. We then construct training and evaluation datasets from more than 40k human annotations, covering seven language directions (Spanish↔English, French↔English, Russian→English, Hokkien→English, and English→German). We evaluate BLASER on these datasets on the popular benchmark of MusT-C (Di Gangi et al., 2019). We also benchmark several strong ASR-based metrics, e.g., ASR-SENTBLEU (i.e., sentence-level ASR-BLEU (Jia et al., 2019)) and ASR-COMET (i.e., applying COMET (Rei et al., 2020) on ASR outputs). There is a recent interest of supervised evaluation metrics that are trained on human quality scores (Rei et al., 2020). However, these human quality scores are precious and somehow limited or nonexistent, specially for

* Equal Research Leadership Contribution
[1]Code is available at https://github.com/facebookresearch/stopes

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 9064–9079
July 9-14, 2023 ©2023 Association for Computational Linguistics

low-resource languages. Therefore, we propose both an unsupervised and a supervised version of BLASER. The results show that on average both unsupervised and supervised BLASER outperform their corresponding baseline metrics. In particular, BLASER outperforms ASR-COMET significantly in five language directions and obtains comparable results in two other language directions. Our analysis reveals that, while BLASER can use both text and speech, encoding speech data give the most significant benefits. In addition, we show that replacing human-written source input and human-written reference with ASR-generated ones hurts performance of text-based metrics, which motivates the use of modality-agnostic metrics as BLASER.

2 Related Work

S2ST Evaluation. Early approaches for automatic S2ST evaluation use metrics consisting of three modules where each module is used to evaluate individual component in the cascaded S2ST pipeline: e.g., BLEU and Translation Edit Rate (Snover et al., 2006) for NMT, Word Error Rate for ASR, and Mel-Cepstral Distortion (Kominek et al., 2008) for TTS. Recent approaches have been primarily focused on adapting text-based metrics for end-to-end S2ST (Jia et al., 2019; Lee et al., 2022a). In contrast to these works, we propose a text-free metric.

MT Metrics. There is a huge amount of literature in automatic machine translation evaluation in the area of natural language processing (Papineni et al., 2002; Denkowski and Lavie, 2014; Popović, 2015, *inter alia*). Recent methods have approached this goal by using human ratings for training model-based metrics, such as COMET, BERTSCORE (Zhang* et al., 2020) and BLEURT (Sellam et al., 2020). These metrics have achieved remarkable performance on text (Freitag et al., 2021; Kocmi et al., 2021).

Speech Metrics. Our work involves computing semantic similarity of speech segments to evaluate translation quality. It is thus related to reference-based automatic evaluation metrics for TTS where the metrics seek to measure the quality of generated speech segments given reference speech segments e.g., Mel-Cepstral Distortion, Gross Pitch Error (Nakatani et al., 2008) and other model-based metrics (Bińkowski et al., 2020). Unlike our work, these metrics primarily focus on the *naturalness* of synthesized speech.

Contemporaneous to this work, Besacier et al. (2022) propose a text-free metric for comparing two speech segments in the same language. Their work limits to comparing English speech data and they do not cover multilingual S2ST evaluation. Their work is based on synthetic datasets where ratings are generated by automatic text-based measures as opposed to human annotators. Differently, we cover S2ST evaluation and we show how our metric correlates with human annotations and how it improves over text-based metrics.

Speech and/or Text Representations. There is a large body of research on learning multilingual text embeddings for various downstream tasks. LabSE (Feng et al., 2022), SentenceBERT (Reimers and Gurevych, 2019), mUSE (Yang et al., 2020) and LASER (Artetxe and Schwenk, 2019; Heffernan et al., 2022) are popular encoders that capture the semantic information of a sentence into fixed size vector representations. In the speech modality, approaches such as wav2vec 2.0 (Baevski et al., 2020a) or Hubert (Hsu et al., 2021) allow learning embeddings at acoustic-frame level.

There has recently been increased interest in aligned speech-text representations such as mSLAM (Bapna et al., 2022), MAESTRO (Chen et al., 2022b), SAMU-XLSR (Khurana et al., 2022), and LASER (Duquenne et al., 2022). While our approach could accommodate any speech representation architecture given the right pooling strategy, we chose LASER in this work for three reasons. **(1)** The encoders modules are freely-available; (2) the LASER embedding space can easily be extended to new languages at a minimal cost: contrary to most multilingual encoders, the teacher-student approach does not require the whole embedding space to be retrained after including data for the new language. This makes BLASER virtually usable for any language in the future **(3)** the embedding space could potentially be extended to any new modality meaningful to translation use cases.

3 Approach

The underlying idea of our approach is to leverage the similarity between speech segments without requiring intermediate textual representations. Compared to ASR-based metrics, the advantage of BLASER is that it is text-free. In particular, given the source input speech, the translated output speech of a S2ST model, and the reference speech segment, respectively, we embed them into vectors h_{src}, h_{mt}, and h_{ref}. These embeddings are combined and BLASER predicts a score for each translation output, where higher scores suggest better translation quality.[2]

The effectiveness of BLASER depends on the quality of vector representations encoded from speech segments: it requires rich semantic information to be encoded in the speech embeddings. In this work, we use LASER speech encoders (Duquenne et al., 2022), which we describe below. We note that our approach is generic and can be extended to other encoders.

We study BLASER under the unsupervised and the supervised settings, which allows it to exploit the information of human ratings, if available.

3.1 Background: LASER Encoders

The LASER encoder was initially trained in a sequence-to-sequence model (Schwenk and Douze, 2017) and supported 93 languages in its follow-up publications (Artetxe and Schwenk, 2019). In recent work, a teacher-student approach was applied to incorporate more languages (Heffernan et al., 2022) and to extend the model to the speech modality (Duquenne et al., 2021). All these encoders use the same teacher model and are mutually compatible. The embeddings are of dimension 1024. The reader is referred to these papers for a detailed description. These LASER encoders were successfully applied to automatically mine semantically similar sentences, in the text (NLLB Team et al., 2022) and speech domain (Duquenne et al., 2022).

3.2 Unsupervised BLASER

In the unsupervised setting, we directly compute the cosine similarities between h_{src} and h_{mt}, and h_{ref} and h_{mt}. Formally, this metric is defined as follows:

$$\text{BLASER}_{\text{u}} = \frac{\cos(h_{\text{src}}, h_{\text{mt}}) + \cos(h_{\text{ref}}, h_{\text{mt}})}{2} \quad (1)$$

where $\cos(\cdot, \cdot)$ is the cosine similarity function.

3.3 Supervised BLASER

Previous work has shown that evaluation metrics (e.g. (Rei et al., 2021)) can take advantage of human ratings for training. We follow COMET (Rei et al., 2020) and RUSE (Shimanaka et al., 2018) and use the following features:

- Element-wise source product: $h_{\text{src}} \odot h_{\text{mt}}$
- Element-wise reference product: $h_{\text{ref}} \odot h_{\text{mt}}$
- Absolute element-wise source difference: $|h_{\text{src}} - h_{\text{mt}}|$
- Absolute element-wise reference difference: $|h_{\text{ref}} - h_{\text{mt}}|$

We concatenate these features with the embeddings of references h_{ref} and translation outputs h_{mt} and then use it as input for a neural regressor to predict a scalar indicating the quality of the translated speech, as shown in Figure 1. This metric corresponds to the following equation:

$$\text{BLASER}_{\text{s}} = \text{nnet}([\text{h}_{\text{ref}}; \text{h}_{\text{mt}}; \text{h}_{\text{src}} \odot \text{h}_{\text{mt}}; |\text{h}_{\text{src}} - \text{h}_{\text{mt}}|; \text{h}_{\text{ref}} \odot \text{h}_{\text{mt}}; |\text{h}_{\text{ref}} - \text{h}_{\text{mt}}|])$$

where $\text{nnet}(\cdot)$ is a two-layer neural network and $[\cdot;\cdot]$ represents the concatenation of vectors. We note that the dimension of concatenated input vectors to the neural regressor is 6144. The entire model except the LASER encoders (which are kept

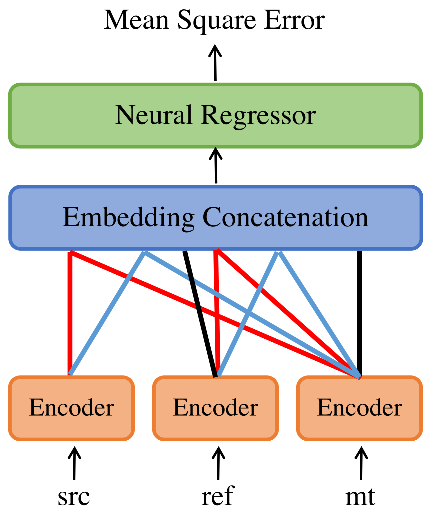

Figure 1: Diagram for supervised BLASER. The source input (src), reference (ref), and translation (mt) speech segments are embedded into vectors using pretrained speech encoders. We create different combinations of these embeddings through element-wise product (red lines), absolute element-wise difference (blue lines), and unchanged (black lines). Combinations are concatenated as input for a neural regressor.We keep the encoders fixed and train the neural regressor using the Mean Square Error.

[2] A straightforward corpus-level score could be obtained via averaging over sentence-level scores, which can be used to compare different S2ST models, similar to metrics like BLEU.

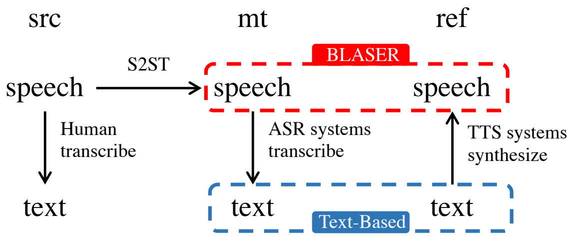

Figure 2: Diagram illustrating data sources of source input (src), reference (ref), and translation output (mt) used in this work. The source speech data and the reference text are generated/annotated by humans. We use color boxes to highlight the differences between BLASER and text-based metrics.

frozen) is trained by minimizing the Mean Squared Error between the BLASER_s predicted scores and human ratings. We choose to freeze LASER encoders because (1) we do not want to break the aligned embedding space; and (2) it allows us to extend to unseen languages more easily.

4 Experimental Framework

To show that BLASER is useful both in its unsupervised and supervised form, we compare it to several baseline metrics. In this section, we describe the experimental framework for doing this comparison, including the evaluation data, the training and implementation of both baseline and proposed metrics and their evaluation.

4.1 Data

We create training and evaluation data from MusT-C (Di Gangi et al., 2019), Multilingual TEDx (Salesky et al., 2021b), and TAT corpus (Liao et al., 2020). Given a source input from these datasets, we generate translated outputs using various S2ST models. We then conduct human evaluations to collect human ratings for generated speech segments. As the datasets do not have reference speech segments but provide human-written transcripts, we use TTS to synthesize speech data from these transcripts to facilitate fair comparison between our metrics and other reference-based textual metrics. While the use of synthesized audios is disadvantageous to BLASER,[3] current benchmarks still use human-written transcripts because of the current dependence on the text-based metrics. We expect that, in the future, S2ST benchmarks will rely on speech references and TTS will not be needed. In this case, BLASER will have additional advantage over text-based metrics that will have to apply ASR to references in addition to ASR to system outputs.

Each data instance in our dataset consists of a source input, a translation output, a reference, and a human evaluation score, where the source, translation output, and reference have both speech and text. Figure 2 summarizes the data sources of these components. As follows we describe the details of each data sources.

Human Annotations. We do not use crowd workers as human annotators and instead we use a vendor-managed pool of well-trained and qualified bilingual annotators who pass a qualification test for their language skills. Human annotators are instructed to rate semantic similarities between source input and generated speech segments[4] on a 5-point Likert scale, where higher values are better, following annotation guidelines similar to Licht et al. (2022). More details on human evaluations are in Appendix D. Each model generation has 1~18 human ratings, leading to 4k~20k annotations per language direction. We take medians of rating scores when there are more than one score associated with a particular model generation following NLLB Team et al. (2022) and Licht et al. (2022).

Speech To Speech Translation. We evaluate the translation outputs generated with the following S2ST architectures:

1. Cascaded two-stage models with speech-to-text translation and TTS. This system includes Spanish-English, English-French and Russian-to-English translation directions;

2. The model presented in Lee et al. (2022b), which represents target speech as discrete units and uses a speech-to-unit translation model to convert source speech to target units followed by a code HiFi-GAN vocoder (Park and Mulc, 2019; Polyak et al., 2021) to convert units to waveform. This system includes English-Spanish and Russian-to-English translation directions;

3. The model presented in Inaguma et al. (2022), which is similar to Lee et al. (2022b) except that it is a two-pass direct S2ST architecture

[3] For example, examples 2 and 3 in table 6 do not correctly synthesize *SMS* or *PKW*.

[4] We note that the generated speech segments could be reference speech segments coming from the TTS models or translated speech segments coming from the S2ST models.

	es→en	ru→en	hk→en	fr→en	en→de	en→es	en→fr
No. of annotators	14	16	9	4	13	13	8
No. of S2ST systems	5	4	1	1	1	4	1
No. of unique source inputs	989	1002	988	1015	2047	1000	1000
No. of annotations	20 636	17 908	6978	4545	12 282	14 817	4426
No. of train instances	2470	2004	0	0	1023	2000	0
No. of test instances	2475	2004	988	1015	1024	2000	1000
No. of annotations per instance							
maximum	6	6	18	6	6	6	6
minimum	1	1	4	1	6	1	2
average	4.2	4.5	7.1	4.5	6.0	3.7	4.4

Table 1: Dataset Statistics. We collect human annotations for speech segments generated by S2ST systems.

that first generates textual representations and predicts discrete acoustic units subsequently. This system includes the Spanish-to-English translation direction;

4. The model presented in Wang et al. (2022), which employs mBART (Liu et al., 2020) for unsupervised machine translation in their unsupervised cascaded speech-to-text translation pipeline. This system includes the Spanish-to-English translation direction.

5. The Hokkien-to-English S2ST system is three-stage cascaded: a concatenation of Hokkien to Chinese speech-to-text translation + Chinese to English machine translation + English TTS (English text-to-unit + unit vocoder from Lee et al. (2022b)).

6. The English-to-German S2ST system is the MLLP-VRAIN system (Iranzo-Sánchez et al., 2022) from IWSLT 2022 (Anastasopoulos et al., 2022), which is a cascaded system of separate ASR, MT, and TTS models.

Automatic Speech Recognition. For ASR, we use the open-sourced implementation in FAIRSEQ (Ott et al., 2019),[5] that provides strong models built on top of the unsupervised pretrained wav2vec (Schneider et al., 2019) or XLSR (Conneau et al., 2020a) models. In particular, for English and Russian, we use wav2vec 2.0 large (Baevski et al., 2020b) finetuned with CTC loss (Graves et al., 2006). For Hokkien, Spanish, French, and German, we use the ASR models released in Chen et al. (2022a), Grosman (2021b), Grosman (2021a), and Grosman (2022), respectively.

Text to Speech. For TTS, we use the toolkit released by Wang et al. (2021a), which provides a set of recent state-of-the-art speech synthesis models.

The language directions in the final dataset are Spanish-English and French-English in both directions (i.e., en→es, es→en, en→fr, and fr→en), Russian to English (ru→en), Hokkien to English (hk→en) and English to German (en→de). We split the data into training and test sets when there is enough data available (i.e., at least one thousand data instances for a language direction). We also make sure that there is no overlapping source inputs between train and test sets. Table 1 summarizes the dataset statistics.

4.2 Baseline Metrics

We consider a variety of baseline metrics, including BLEU and CHRF+ (Popović, 2017), which are standard metrics to evaluate textual similarities. While BLEU is by nature corpus-level, here we use the *sentence*-level version due to the insufficient amount of human annotations. To differentiate these two versions, we denote the sentence-level BLEU as SENTBLEU. We also benchmark BERTSCORE (Zhang* et al., 2020) and COMET, which are popular model-based metrics that correlate well with human judgments on textual data (Kocmi et al., 2021).[6] We extend these metrics to speech data by using ASR systems to transcribe the machine-translated speech segments. We prepend "ASR-" to the beginning of the names of these metrics to indicate the use of ASR systems. Table 2 summarizes the differences among the metrics.

Specifically, we use BLEU[7] and CHRF+[8] as im-

[5]https://github.com/facebookresearch/fairseq/blob/ust/examples/speech_to_speech/asr_bleu

[6]Multilingual BLEURT (Pu et al., 2021) reports similar performance as COMET on WMT metrics tasks and therefore we decided to only include COMET in our experiments.

[7]SacreBLEU signature: nrefs:1|case:mixed|eff:yes|tok:13a|smooth:exp|version:2.2.0

[8]SacreBLEU signature:

Baseline Metrics	req. train	req. ASR
ASR-SENTBLEU	✗	✓
ASR-CHRF+	✗	✓
ASR-BERTSCORE	✓	✓
ASR-COMET	✓	✓
Proposed Metrics		
BLASER_{u}	✗	✗
BLASER_{s}	✓	✗

Table 2: Comparisons between baseline and proposed metrics regarding the dependency of training data and ASR systems. We use "ASR-" to indicate that the metric depends on ASR systems to transcribe speech segments.

plemented in SacreBLEU (Post, 2018).[9] We normalize the reference text before computing ASR-SENTBLEU and ASR-CHRF+ to match the lower-cased and punctuationless ASR output. We use the official implementations for BERTSCORE[10] and COMET.[11] To form competitive baselines, we also train COMET from scratch on our training data ($\text{COMET}_{\text{retrain}}$) and the concatenation of our training data and the direct assessments from WMT 15-19 metrics tasks (Stanojević et al., 2015; Bojar et al., 2016, 2017; Ma et al., 2018, 2019) ($\text{COMET}_{\text{retrain with WMT}}$).

4.3 Training and Evaluation

LASER Encoders. We use the speech LASER encoders released in Duquenne et al. (2022) except for English and Hokkien.[12]. For Hokkien speech LASER encoder, we followed the training procedure presented in (Chen et al., 2022a) using the same pretrained model and training data. For the English speech LASER encoder, we fine-tuned XLSR 2B (Babu et al., 2021) on several ASR datasets including CoVoST2 (Wang et al., 2021c), Common Voice (Ardila et al., 2020), EuroparlST (Iranzo-Sánchez et al., 2020), MusT-C (Di Gangi et al., 2019), Voxpopuli (Wang et al., 2021b) and Librispeech (Panayotov et al., 2015).

Training Setup and Hyperparameters. For BLASER_{s}, the regressor has two hidden layers of sizes 3072 and 1536, similar to COMET. We keep the LASER encoders fixed during training. We use a learning rate of 5×10^{-5} and employ learning rate annealing with a linear schedule. When training COMET, we follow the official implementation and fine-tune the entire model from the XLM-R-LARGE model checkpoint (Conneau et al., 2020b). For both BLASER_{s} and COMET, we train them for 20 epochs. We standardize the human ratings in our training set by subtracting them with a mean and a variance computed based on the entire training set.

Computational Cost. We trained BLASER_{s} using 1 Quadro GV100 and the training takes less than one hour. We used 4 Tesla V100 to train COMET and the training takes more than two days.

Evaluation. We compute Pearson's correlation at the sentence level between the automatic and human rating scores. Given that our test sets are relatively small, we perform statistical significance test using the bootstrap method from Koehn (2004).[13]

5 Experimental Results and Analysis

In this section we report the main results of our proposed metric BLASER, on two different settings (unsupervised and supervised) and we compare it to widely used baseline text-based metrics. Additionally, we report an analysis at various levels, including the impact of evaluating using different modalities and a qualitative inspection of several examples to observe scores of various metrics for particular examples.

5.1 Main Results

We report unsupervised and supervised results in Table 3. We note that results that fail to pass the significance test are neither better nor worse significantly than the corresponding baseline.

Generally, model-based metrics perform significantly better than string-based ones. Among the unsupervised metrics, BLASER_{u} performance improves significantly over ASR-SENTBLEU and ASR-CHRF+ for all language directions except for en→es, showing the capabilities of BLASER in capturing semantic information even when human annotations are absent.

Among the supervised metrics, we see that BLASER_{s} almost always performs better than the official ASR-BERTSCORE and ASR-COMET.

nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.2.0

[9] https://github.com/mjpost/sacrebleu

[10] We use language-specific configurations recommended in https://github.com/Tiiiger/bert_score

[11] We use the "wmt20-comet-da" model from https://github.com/Unbabel/COMET

[12] https://github.com/facebookresearch/fairseq/tree/ust/examples/speech_matrix

[13] https://github.com/neubig/util-scripts/blob/master/paired-bootstrap.py

	es→en	ru→en	hk→en	fr→en	en→de	en→es	en→fr	average
Unsupervised Metrics								
ASR-SENTBLEU	0.3226	0.1588	0.2863	0.3277	0.1179	0.4937	0.4462	0.3076
ASR-CHRF+†	0.3910	0.2324	0.3356	0.3927	0.1469	**0.5967**	0.5267	0.3746
BLASER$_u$	**0.4970***	**0.4326***	**0.4940***	**0.4744***	**0.3148***	0.5843	**0.6356***	**0.4904**
Supervised Metrics								
ASR-BERTSCORE	0.4332	0.3511	0.4885	0.4184	0.2031	0.6127	0.6216	0.4469
ASR-COMET	0.5238	0.3988	0.5138	0.5693	0.2428	0.7126	0.6559	0.5167
ASR-COMET$_{\text{retrained}}$	0.5618	0.4265	0.4485	0.5210	0.2921	0.7489	0.6123	0.5159
ASR-COMET$^{\dagger}_{\text{retrained with WMT}}$	0.5340	0.4348	0.5314	0.5659	0.2635	0.7308	0.6436	0.5291
BLASER$_s$	**0.5774***	**0.5347***	**0.6059***	**0.5730**	**0.3297***	**0.7512**	**0.7146***	**0.5838**

Table 3: Pearson's correlation on the test set. Best results in bold. Results marked with $*$ pass the significance test with with p-value < 0.05 when compared against the baseline metric marked by † in the same category.

	es→en	ru→en	hk→en	fr→en	en→de	en→es	en→fr	average
ASR-SENTBLEU	0.3226	0.1588	0.2863	0.3277	0.1259	0.4929	0.4393	0.3076
Δ	−0.0222	−0.0244	−0.0033	−0.0161	**−0.1161**	−0.0467	−0.0341	−0.0376
ASR-CHRF+	0.3910	0.2324	0.3356	0.3927	0.1673	0.6032	0.5177	0.3771
Δ	−0.0195	−0.0204	0.0017	−0.0125	**−0.1201**	−0.0757	−0.0206	−0.0382
ASR-COMET	0.5238	0.3988	0.5138	0.5693	0.2428	0.7126	0.6559	0.5167
Δ	−0.0164	−0.0443	−0.0602	−0.0185	**−0.0929**	−0.0281	−0.0057	−0.0380

Table 4: Pearson's correlation on the test set. "Δ" rows show the performance differences when using transcripts produced by ASR systems instead of humans for the source input and reference. Negative differences indicate performance drops. We highlight the results for en→de as they are severely affected by the change.

When compared to the stronger baseline ASR-COMET$_{\text{retrained with WMT}}$, BLASER$_s$ is better than the baseline significantly in four language directions and they are comparable in the other three directions.

We also find that BLASER can generalize training signal to languages where there is no training data available. Specifically, if we compare BLASER$_s$ to BLASER$_u$, we see that BLASER$_s$ always improves over the unsupervised version. Also, for the language directions where there is no training data (i.e., hk→en, fr→en, en→fr), BLASER$_s$ still beats BLASER$_u$. Additionally, we observe that hk→en and ru→en are two of the language directions for which BLASER$_s$ shows significant improvements over ASR-COMET, confirming the zero-shot capabilities of our proposed methods in comparison to existing metrics.

5.2 Analysis

Impact of Human-Written vs ASR-transcriptions. To investigate the impact of using transcripts generated by ASR systems rather than human-written inputs and references, we replace the human-written source input and reference with the ones generated by ASR systems. We note that in this case, all the transcripts are obtained via ASR systems, simulating an evaluation setting where only audio data is available. We show the results in Table 4 where we find that the human-written transcripts are less helpful on those to-English language directions than the from-English ones. We hypothesize that this is in part due to the quality of ASR systems as these ASR-based metrics depend more on references than source inputs and English ASR systems tend to be of better quality than the non-English ones (Khare et al., 2021).

Impact of Using Source and Reference. We investigate the impact of using source and reference speech segments when computing BLASER scores. We evaluate this impact on BLASER$_u$ by reporting the performance of individual terms in Equation 1. See the results in Table 5. In general, we find the source input generates better correlations with human ratings than reference. Combining the two leads to the best performance.

Qualitative Analysis. To get a sense of the qualitative differences between BLASER and text-based scores, and better understand what kind of nuances are captured, we manually inspect sample sen-

	es→en	ru→en	hk→en	fr→en	en→de	en→es	en→fr	average
$\cos(h_{\text{ref}}, h_{\text{mt}}) + \cos(h_{\text{src}}, h_{\text{mt}})$	**0.4970**	**0.4326**	**0.4940**	**0.4744**	**0.3148**	0.5843	**0.6356**	**0.4904**
$\cos(h_{\text{ref}}, h_{\text{mt}})$	0.4392	0.2855	0.4051	0.4144	0.1388	0.4516	0.5588	0.3848
$\cos(h_{\text{src}}, h_{\text{mt}})$	0.4392	0.4182	0.4723	0.4450	0.2654	**0.6411**	0.6215	0.4718

Table 5: Pearson's correlation on the test set. Best results are in bold. We evaluate the contributions of two individual terms in BLASER$_{\text{u}}$ (Equation 1) to the final performance.

source input	translation output	reference	HR	BR	CT	BU
The pollution in Santiago, which is one of the most polluted capitals historically in Latin America, has dropped substantially.	die verschmutzung in santiago einem der am stärksten verschmutzten hauptstädte lateinamerikas ist erheblich gesungen (the pollution in santiago one the at strongest polluted capital cities latin america is significantly sung)	Die Umweltverschmutzung in Santiago, das historisch gesehen eine der Städte mit der höchsten Umweltverschmutzung in ganz Lateinamerika ist, ist viel geringer geworden.	4.5	0.2	0.9	4.0
And for those of us that are in the know, we know that's text-speak, or SMS language.	diejenigen von uns die das kennen wissen das ist zum spracher (those from us the the know to know the is for the speaker)	Diejenigen von uns, die das kennen, wissen: Das ist SMS-Sprache.	2.5	0.0	0.9	78.6
So, when I say, "Oh, Aaron is..." It's because Aaron still is.	wenn ich aron sehe liegt das daran dass aron es immer noch ist (if I aron see located the to it that aron it always still is)	Wenn ich also sage: „Oh, Aaron ist ...“, dann sage ich das, weil Aaron immer noch ist.	3.5	-0.1	0.9	12.9

Table 6: The examples from the en→de test set and the corresponding scores given by different metrics. HR=Human Ratings. BR=BLASER$_{\text{s}}$. CT=ASR-COMET. BU=ASR-SENTBLEU. Sentences in parenthesis are gloss for translation outputs.

tences. A selection is presented in Table 6. In each of these examples, the text and generated audio perfectly match, discarding any influence potentially introduced by the ASR model. In cases where the output vocabulary does not perfectly match the reference but is still valid, BLASER seems able to capture the semantics and produce a meaningful score. In the first example, ASR-SENTBLEU is very much impacted by the vocabulary mismatch, while BLASER and ASR-COMET yield high scores, in line with human evaluation. BLASER also seem to detect clear mistranslations better than either of ASR-COMET or ASR-SENTBLEU. In the second example, the end of the output sentence makes little sense. Only BLASER accounts for this properly and produces a score aligned with human judgment. In the third example, ASR-COMET returns a high score despite the mistranslated verb which heavily changes the meaning of the sentence.

6 Conclusion and Future Work

We have introduced BLASER, a text-free metric to evaluate speech-to-speech translation, which avoids the dependency on ASR models required by popular text-based metrics currently used in S2ST. We explored BLASER in both unsupervised and supervised settings. Experimental results in seven language directions show that BLASER outperforms or is comparable to strong text-based metrics in terms of correlation with human scores at the sentence-level. Moreover, our metric is effective in zero-shot scenarios.

As for future work, we want to explore the use of speech references generated by humans and the impact of synthesized references. We also want to evaluate BLASER at the system-level with a much larger number of S2ST systems, and explore different approaches to aggregate the sentence-level scores from BLASER and we want to explore different speech and text representations as alternative to LASER.

Limitations

We are evaluating S2ST in an artificial setting given that we have to synthesize the text references. In fact, since there was no metric capable of evaluating the quality in speech, there was no motivation to build such benchmarks either (the chicken-and-egg problem). However, we expect that next benchmarks for the task will have speech references because of the rise of end-to-end S2ST systems and their quality increase. BLASER paves the way so that we can take advantage of such benchmarks when they appear.

Our metric works at the sentence-level, by embedding the entire sentence into an intermediate space. We ignore how sensitive BLASER is to the length of the sentence, which is a key aspect when we want to extend to the corpus-level metric in the future. Moreover, we are aware that sometimes sentence embedding do not discriminate different numbers or words that belong to the same word family, which may disregard impactful errors such as the change of a number in the translation output.

Ethical considerations

Translation quality scores were provided by bilingual raters as mentioned in Section 4. They were all paid a fair rate. We can not open-source the data form our experiments given that our sources are shared under *no-derivative* license. Small human evaluation detailed in appendix D was done by volunteers.

References

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4218–4222. European Language Resources Association.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020a. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020b. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau. 2022. mslam: Massively multilingual joint pre-training for speech and text. *arXiv preprint arXiv:2202.01374*.

Laurent Besacier, Swen Ribeiro, Olivier Galibert, and Ioan Calapodescu. 2022. A textless metric for speech-to-speech comparison. *arXiv preprint arXiv:2210.11835*.

Mikołaj Bińkowski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C. Cobo, and Karen Simonyan. 2020. High fidelity speech synthesis with adversarial networks. In *International Conference on Learning Representations*.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231, Berlin, Germany. Association for Computational Linguistics.

Peng-Jen Chen, Kevin Tran, Yilin Yang, Jingfei Du, Justine Kao, Yu-An Chung, Paden Tomasello, Paul-Ambroise Duquenne, Holger Schwenk, Hongyu Gong, et al. 2022a. Speech-to-speech translation for a real-world unwritten language. *arXiv preprint arXiv:2211.06474*.

Zhehuai Chen, Yu Zhang, Andrew Rosenberg, Bhuvana Ramabhadran, Pedro Moreno, Ankur Bapna, and Heiga Zen. 2022b. Maestro: Matched speech text representations through modality matching. *arXiv preprint arXiv:2204.03409*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020a. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul-Ambroise Duquenne, Hongyu Gong, Ning Dong, Jingfei Du, Ann Lee, Vedanuj Goswani, Changhan Wang, Juan Pino, Benoît Sagot, and Holger Schwenk. 2022. Speechmatrix: A large-scale mined corpus of multilingual speech-to-speech translations. *arXiv preprint arXiv:2211.04508*.

Paul-Ambroise Duquenne, Hongyu Gong, and Holger Schwenk. 2021. Multimodal and multilingual embeddings for large-scale speech mining. In *Advances in Neural Information Processing Systems*, volume 34, pages 15748–15761. Curran Associates, Inc.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Jonatas Grosman. 2021a. Fine-tuned French Voxpopuli wav2vec2 large model for speech recognition in French. `https://huggingface.co/jonatasgrosman/wav2vec2-large-fr-voxpopuli-french`.

Jonatas Grosman. 2021b. Fine-tuned XLSR-53 large model for speech recognition in Spanish. `https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-spanish`.

Jonatas Grosman. 2022. Fine-tuned XLS-R 1B model for speech recognition in German. `https://huggingface.co/jonatasgrosman/wav2vec2-xls-r-1b-german`.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. *arXiv preprint arXiv:2205.12654*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Hirofumi Inaguma, Sravya Popuri, Ilia Kulikov, Peng-Jen Chen, Changhan Wang, Yun Tang, Ann Lee, Shinji Watanabe, and Juan Pino. 2022. Unity: Two-pass direct speech-to-speech translation with discrete units. *arXiv preprint*.

Javier Iranzo-Sánchez, Javier Jorge Cano, Alejandro Pérez-González-de Martos, Adrián Giménez Pastor, Gonçal Garcés Díaz-Munío, Pau Baquero-Arnal, Joan Albert Silvestre-Cerdà, Jorge Civera Saiz, Albert Sanchis, and Alfons Juan. 2022. MLLP-VRAIN UPV systems for the IWSLT 2022 simultaneous speech translation and speech-to-speech translation tasks. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 255–264, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchís, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 8229–8233. IEEE.

Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. Towards building ASR systems for the next billion users. In *Proceedings of AAAI*.

Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Z. Chen, and Yonghui Wu. 2019. Direct speech-to-speech translation with a sequence-to-sequence model. In *INTERSPEECH*.

Shreya Khare, Ashish Mittal, Anuj Diwan, Sunita Sarawagi, Preethi Jyothi, and Samarth Bharadwaj. 2021. Low Resource ASR: The Surprising Effectiveness of High Resource Transliteration. In *Proc. Interspeech 2021*, pages 1529–1533.

Sameer Khurana, Antoine Laurent, and James Glass. 2022. Samu-xlsr: Semantically-aligned multimodal utterance-level cross-lingual speech representation. *arXiv preprint arXiv:2205.08180*.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

John Kominek, Tanja Schultz, and Alan W Black. 2008. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In *SLTU*, pages 63–68.

Alon Lavie, A. Waibel, Lori Levin, M. Finke, Donna Gates, Marsal Gavalda, Torsten Zeppenfeld, and Puming Zhan. 1997. Janus-iii: speech-to-speech translation in multiple languages. pages 99 – 102 vol.1.

Gianni Lazzari. 2006. TC-STAR: a speech to speech translation project. In *Proceedings of the Third International Workshop on Spoken Language Translation: Plenaries*, Kyoto, Japan.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, and Wei-Ning Hsu. 2022a. Direct speech-to-speech translation with discrete units. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3327–3339, Dublin, Ireland. Association for Computational Linguistics.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. 2022b. Textless speech-to-speech translation on real data. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 860–872, Seattle, United States. Association for Computational Linguistics.

Yuan-Fu Liao, Chia-Yu Chang, Hak-Khiam Tiun, Huang-Lan Su, Hui-Lu Khoo, Jane S. Tsay, Le-Kun Tan, Peter Kang, Tsun-guan Thiann, Un-Gian Iunn, Jyh-Her Yang, and Chih-Neng Liang. 2020. Formosa speech recognition challenge 2020 and taiwanese across taiwan corpus. In *2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 65–70.

Daniel Licht, Cynthia Gao, Janice Lam, Francisco Guzman, Mona Diab, and Philipp Koehn. 2022. Consistent human evaluation of machine translation across language pairs. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 309–321, Orlando, USA. Association for Machine Translation in the Americas.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Tomohiro Nakatani, Shigeaki Amano, Toshio Irino, Kentaro Ishizuka, and Tadahisa Kondo. 2008. A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments. *Speech Communication*, 50(3):203–214.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau

Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kyubyong Park and Thomas Mulc. 2019. Css10: A collection of single speaker speech datasets for 10 languages. *Interspeech*.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. In *Proc. Interspeech 2021*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. 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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Elizabeth Salesky, Julian Mäder, and Severin Klinger. 2021a. Assessing evaluation metrics for speech-to-speech translation. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 733–740. IEEE.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021b. Multilingual tedx corpus for speech recognition and translation. In *Proceedings of Interspeech*.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

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.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758, Belgium, Brussels. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association* *for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon, Portugal. Association for Computational Linguistics.

Changhan Wang, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Ann Lee, Peng-Jen Chen, Jiatao Gu, and Juan Pino. 2021a. fairseq sˆ2: A scalable and integrable speech synthesis toolkit. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 143–152, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Changhan Wang, Hirofumi Inaguma, Peng-Jen Chen, Ilia Kulikov, Yun Tang, Wei-Ning Hsu, Michael Auli, and Juan Pino. 2022. Simple and effective unsupervised speech translation. *arXiv preprint arXiv:2210.10191*.

Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Miguel Pino, and Emmanuel Dupoux. 2021b. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 993–1003. Association for Computational Linguistics.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021c. Covost 2 and massively multilingual speech translation. In *Interspeech*, pages 2247–2251.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Cross-Modal Data Analysis

Considering that LASER can conveniently encode text and speech data into a shared embedding space, we conduct experiments involving both text and speech data with the text encoders from Heffernan et al. (2022) for BLASER$_s$. In particular, we embed the source input, translation output, and reference using either the speech or text LASER encoders. That is, a data instance formed by embeddings from speech data will result in four instances in this new setting due to different modality combinations. We then evaluate the models on the speech data in our test set. The results in Table 7 show that combining supervision from different modalities does not help improve model performance. It is likely because the embedding space is shared between text and speech and therefore adding textual embeddings do not provide extra information.

B Cross-Modal Supervision Analysis

We also look into the benefits of leveraging speech embeddings by comparing several supervised configurations for BLASER$_s$. We report these results in Table 8 where we experiment with different modality combinations during training and testing. The results show that the best results on average are the ones using speech modality for the source input, translation output, and reference. Interestingly, every time that we replace speech with text in the modality combinations, we see performance drops. We find that replacing reference speech segment with text leads to the slightest performance drop, which is likely due to the fact that they are synthesized and thus do not provide extra information than text. We also find that replacing speech data with text for the source input and translation output makes BLASER$_s$ similar or even worse than ASR-COMET$_{\text{retrained with WMT}}$, confirming the benefits of using speech data for evaluation S2ST systems.

C Cross-Modal Evaluation Analysis

We additionally evaluate BLASER$_s$ on different modality combinations when training on speech data only. See the results in Table 9. We find that training on speech data only still allows BLASER to obtain similar performance when replacing the reference speech segments with text.

D Human Evaluation

We provide instructions for human evaluations in Table 10.

	es→en	ru→en	hk→en	fr→en	en→de	en→es	en→fr	average
Speech-only	0.5774	**0.5347**	**0.6059**	**0.5730**	0.3297	**0.7512**	**0.7146**	**0.5838**
Combined	**0.5791**	0.5295	0.5988	0.5459	**0.3348**	0.7456	0.7037	0.5767

Table 7: Pearson's correlation on the test set. Best results are in bold. We compare BLASER$_s$ when training with speech data only and training with both speech and text data. For testing, we always evaluate models on speech data.

Modalities			es→en	ru→en	hk→en	fr→en	en→de	en→es	en→fr	average
(Speech,	Speech,	Speech)	**0.5774**	**0.5347**	**0.6059**	**0.5730**	0.3297	**0.7512**	**0.7146**	**0.5838**
(Speech,	Speech,	Text)	0.5541	0.5164	0.5754	0.5425	**0.3675**	0.7485	0.6688	0.5676
(Speech,	Text,	Text)	0.5460	0.4866	0.5616	0.4741	0.3393	0.7372	0.6285	0.5390
(Text,	Text,	Text)	0.4555	0.4094	0.5350	0.4505	0.2710	0.6544	0.5882	0.4806

Table 8: Pearson's correlation on the test set. Best results are in bold. (x, y, z) indicates the modality used for source input (x), translation output (y), and reference (z). We train and evaluate BLASER$_s$ on the same modality combinations.

Modalities			es→en	ru→en	hk→en	fr→en	en→de	en→es	en→ru	average
(Speech,	Speech,	Speech)	**0.5774**	0.5347	0.6059	**0.5730**	0.3297	**0.7512**	**0.7146**	**0.5838**
(Speech,	Speech,	Text)	0.5588	**0.5403**	**0.6093**	0.5587	**0.3426**	0.7500	0.6978	0.5796

Table 9: Pearson's correlation on the test set. Best results are in bold. (x, y, z) indicates the modality used for source input (x), translation output (y), and reference (z). We train BLASER$_s$ on speech data only and evaluate the model with references either in speech or text modalities.

Task Descriptions	• You will be provided with a pair of audio snippets. • The pair will be in two different languages. • Your task is to assess: (1) if audio1 is coherent; (2) if audio2 is coherent; and (3) how well the pair of audios correspond to each other on a scale from 1-5. • When rating semantic similarity, please ignore minor typos, grammatical errors, and pronunciation errors if they do not affect your understanding of the audio segments.
Rating Instructions	1. The two sentences are not equivalent, do not share any details, but may be related as pertaining to similar or even different topics. 2. The two sentences are not equivalent, but share some details. However, some important information differs/is missing, which alters the intent/meaning. 3. The two sentences are mostly equivalent, but some unimportant details differ. 4. The two sentences are equivalent paraphrases of each other. They mean the same with no major or minor differences in meaning, despite potential differences in expression. 5. The two sentences are exactly and completely equivalent in meaning and usage expression (e.g., formality level, style, multiword expression)

Table 10: Instructions for human evaluations.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
section of its own name

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
in abstract and introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

section 3 and 4

☑ B1. Did you cite the creators of artifacts you used?
section 3 and 4

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
we are going to share our code and license details after anonymity period

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
section 4 and 5

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
we are relying on an external dataset, we refer to the sources for those details

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
section 4 reports coverage of domains and languages

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
section 4

C ☑ Did you run computational experiments?

section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
section 4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
section 4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
section 4 and 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
section 4

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
section 5 appendix B

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
appendix B

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
section 5

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
ethics section

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
it is quite standard protocol in the community

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
for the small annotation in appendix B, we relied on volunteers that do not necessarily want to share this info