Speech-to-Speech Translation Evaluation: A Systematic Analysis and Multi-Dimensional Metrics

Lalaram Arya
Department of EECE
IIT Dharwad
Dharwad, India
202021004@iitdh.ac.in

Mrinmoy Bhattacharjee
Department of CSE
IIT Jammu
Jammu, India
mrinmoy.bhattacharjee@iitjammu.ac.in

S. R. Mahadeva Prasanna
Department of EECE
IIT Dharwad
Dharwad, India
prasanna@iitdh.ac.in

***Abstract*—Evaluating speech-to-speech translation (S2ST) systems remains challenging due to the multi-dimensional nature of speech, encompassing semantic accuracy, acoustic quality, and speaker characteristics. Existing approaches rely on either text-based or speech-based metrics, each capturing only partial aspects of translation quality. In this work, we conduct a systematic empirical analysis of S2ST evaluation metrics across multiple language pairs (fr–en, es–en, de–en, hi–en) using SeamlessM4T-v2 translations generated from the FLEURS dataset. We analyze n-gram, neural text, and speech-embedding metrics at both corpus and sentence levels. Our analysis shows that text-based metrics are sensitive to linguistic variation but depend on automatic speech recognition (ASR), while speech-based metrics provide more consistent scores yet exhibit limited discriminative ability. Based on these observations, we propose a multi-dimensional evaluation framework that jointly assesses semantic adequacy and acoustic naturalness. The proposed framework improves semantic alignment and achieves strong correlation with speech naturalness across language pairs, enabling more reliable evaluation of S2ST systems.**

***Index Terms*—Speech to Speech Translation, Speech Evaluation, Evaluation Metrics**

I. Introduction

Speech-to-Speech Translation (S2ST) has emerged as a promising technology for overcoming language barriers, enabling cross-lingual communication and global information access [1]. Early systems relied on cascaded pipelines [2] combining Automatic Speech Recognition (ASR) [3], Machine Translation (MT) [4], and Text-to-Speech systhesis (TTS) [5]. While effective, these pipelines suffer from error propagation, increased latency, and loss of prosodic information. Recent approaches explore direct models that map speech to speech without explicit intermediate text [6], improving translation quality, voice preservation, and latency, while varying in their reliance on text supervision [7] [8]. However, the diversity of these paradigms and the increasing realism of generated speech highlight limitations in existing evaluation methods. These methods typically assess textual semantics or acoustic quality in isolation, motivating the need for more comprehensive evaluation frameworks.

Unlike MT, which primarily evaluates semantic adequacy, S2ST quality is inherently multi-dimensional. It includes semantic correctness, acoustic naturalness, speaker identity preservation, and prosodic expressiveness, all of which affect usability and user experience. Despite this, current evaluation practices are largely dominated by text-based metrics such as ASR-BLEU [9], which rely on transcribing speech before comparison. While simple and interpretable, this approach is sensitive to transcription errors and fails to capture speech-specific attributes such as naturalness and speaker characteristics. Speech-based metrics such as BLASER [10] and BLASER 2.0 [11] address this limitation by operating directly on speech embeddings (e.g., LASER, SONAR) [10], [12]. However, they primarily focus on semantic similarity and do not explicitly model acoustic or prosodic quality. Their reliance on human-annotated data further limits scalability. Overall, existing metrics evaluate different aspects of S2ST quality in isolation, highlighting the need for more comprehensive evaluation frameworks.

To bridge these gaps, we conduct a systematic empirical analysis of S2ST evaluation metrics guided by four research questions: (RQ1) Do text-based and speech-based semantic metrics capture similar aspects of translation quality? (RQ2) What are their key limitations? (RQ3) Which evaluation dimensions are adequately covered or missing? and (RQ4) Under what conditions do these metrics fail and why? Through this study, we provide a comprehensive assessment of existing S2ST evaluation methods, highlighting their strengths and limitations and their coverage across different quality dimensions. Our experiments are conducted on the FLEURS multilingual benchmark [13] for Spanish (es), French (fr), German (de), and Hindi (hi) to English (en). The translated speech is generated using the SeamlessM4T-v2 [14] model. We evaluate outputs using a diverse set of metrics, including text-based semantic measures (ASR-BLEU [9], ASR-chrF++ [15], ASR-METEOR [16], ASR-COMET [17], ASR-BLEURT [18]), speech-based semantic metrics (BLASER [10], BLASER 2.0-QE and BLASER 2.0-REF [11]). We further perform empirical correlation analysis to examine relationships between these metrics. Finally, we introduce a multi-dimensional evaluation metric that jointly captures semantic and acoustic quality, enabling a unified assessment of translated speech.

Key contributions of this work are as follows:

- We conduct quantitative correlation analysis across text-based and speech-based metric families, revealing complementarities and coverage gaps between semantic and

acoustic evaluation dimensions in S2ST.
- We analyze current S2ST evaluation practices to derive evidence-based guidelines for metric selection, highlighting the need for multi-dimensional evaluation.
- We propose a multi-dimensional evaluation framework that jointly captures semantic and acoustic quality, enabling more reliable S2ST assessment.

The paper is organized as follows. Section II reviews S2ST evaluation metrics. Section III outlines the proposed multi-dimensional evaluation framework. Section IV describes the experimental details. Section V presents quantitative results and analysis. Finally, Section VI concludes the paper.

II. S2ST EVALUATION METRICS

In this section, we review existing evaluation metrics for S2ST, categorizing them based on the quality dimensions they target and the methodologies they employ.

A. S2ST Evaluation Dimensions

The quality of S2ST is inherently multi-dimensional, with each aspect contributing to overall system performance and user experience. **Semantic adequacy** measures the preservation of meaning from the source to the target language, including lexical accuracy, structural consistency, and completeness without omissions or additions. **Acoustic naturalness** reflects the perceptual quality of translated speech, characterized by clarity, pronunciation accuracy, appropriate speaking rate, and the absence of artifacts such as distortions or unnatural prosody. **Speaker preservation** assesses how well the target speech retains the source speaker's characteristics, including voice timbre, pitch range, and speaking style. **Prosody and expressiveness** captures suprasegmental aspects of speech, such as intonation, stress patterns, rhythm, timing, and emotional cues. These dimensions are largely independent. An often overlooked issue is that high semantic accuracy does not necessarily guarantee acoustic naturalness or speaker consistency in the translated speech.

B. Text-Based Evaluation Metrics

Text-based metrics evaluate S2ST outputs by first transcribing generated speech using an ASR system and subsequently applying standard text-based MT evaluation metrics. This approach remains widely adopted due to its interpretability, established benchmarks, and compatibility with existing MT evaluation frameworks.

N-gram and symbolic metrics are early text-based evaluation methods that rely on surface-level matching between hypothesis and reference translations. ASR-BLEU [9] measures n-gram overlap between ASR-transcribed output and reference text [9]. While it is simple and interpretable, it is sensitive to exact word matching and does not account for semantic equivalence. ASR-METEOR [16] improves upon BLEU by incorporating synonym matching, stemming, and paraphrase alignment. This makes it more robust to lexical variation and better aligned with human judgments. ASR-chrF++ [15] computes character n-gram overlap, making it less sensitive to tokenization and particularly effective for morphologically rich languages.

Neural text-based metrics aim to capture semantic similarity beyond surface-level matching by leveraging contextual representations. ASR-BERTScore [19] measures contextual embedding similarity between tokens, enabling soft matching for semantically equivalent translations. BLEURT [18] leverages pretrained language models fine-tuned on human annotations to provide robust quality estimation. XCOMET [17] extends neural evaluation with explainable quality estimation by identifying error spans and their severity.

Despite methodological differences, both n-gram and neural text-based metrics share key limitations. They rely on ASR transcriptions, causing error propagation and obscuring translation errors. They also ignore acoustic quality, speaker traits, and prosody in S2ST outputs.

C. Speech-Based Evaluation Metrics

Speech-based metrics evaluate S2ST directly on speech embeddings without relying on ASR transcription. This avoids error propagation and enables a direct assessment of speech translation quality.

BLASER [10] introduced one of the first text-free S2ST evaluation metrics by leveraging LASER [20], a multilingual encoder that produces language-agnostic embeddings in a shared semantic space. The method extracts fixed-dimensional embeddings from source reference and target speech, computes similarity in this space, and employs a regression model trained on human judgments to predict translation adequacy. BLASER demonstrates improved correlation with human judgments compared to ASR-BLEU while eliminating ASR dependency. However, it has several limitations, including reliance on reference speech, limited language coverage, and restriction to semantic evaluation.

BLASER 2.0 [11] extends this framework in two key ways. First, it adopts SONAR [12], a more scalable multilingual encoder with broader language coverage and improved representation quality. Second, it introduces a reference-free quality estimation (QE) variant that predicts translation quality using only source and target speech. Despite these improvements, BLASER 2.0 remains focused on semantic adequacy and does not capture acoustic quality or speaker characteristics. Furthermore, these models are trained using human annotations, which are costly and difficult to obtain at scale.

III. PROPOSED MULTI-DIMENSIONAL EVALUATION

This section outlines a Multi-dimensional evaluation framework with a unified multi-head architecture for jointly assessing *semantic adequacy* and *speech naturalness* in S2ST systems. As shown in Fig. 1, the model comprises two task-specific heads, a semantic head and a naturalness head, each producing an independent score from a shared model.

A. Semantic Evaluation Head

The semantic evaluation head is inspired by the architecture of BLASER 2.0-QE [11] and is designed to assess the semantic adequacy of translated speech in a reference-free setting.

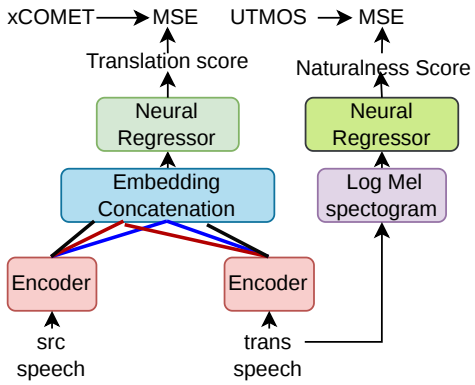

Fig. 1. Overview of the proposed Multi-dimensional evaluation framework

A pretrained SONAR [12] speech encoder is used to extract fixed-dimensional embeddings for the source and translated speech, denoted as $\mathbf{e}_{\text{src}}$ and $\mathbf{e}_{\text{trans}}$, respectively. To capture semantic relationships, interaction features are constructed as $\left[\mathbf{e}_{\text{src}}; \mathbf{e}_{\text{trans}}; \mathbf{e}_{\text{src}} \odot \mathbf{e}_{\text{trans}}; |\mathbf{e}_{\text{src}} - \mathbf{e}_{\text{trans}}|\right]$. These features are passed to a neural regressor that predicts a semantic quality score, trained using a Mean Squared Error (MSE) loss with supervision distilled from xCOMET [17] scores. This design reduces reliance on human-annotated data and provides a scalable approach to semantic evaluation.

B. Naturalness Evaluation Head

The naturalness evaluation head estimates the perceptual quality of translated speech. Log-Mel spectrograms [21], denoted as $\mathbf{M} \in \mathbb{R}^{T \times F}$, are extracted from the translated speech and used as input to a neural regressor that outputs a scalar naturalness score. This head is also trained using an MSE objective, with supervision obtained via distillation from a pretrained UTMOS [22] model.

This modular design can be naturally extended to additional evaluation dimensions, such as speaker similarity and prosody, by incorporating dedicated task-specific heads within the same unified framework.

IV. Experimental Details

This section describes the dataset, translation model, experimental setup, and evaluation metrics used in our experiments.

A. Dataset

We use the FLEURS [13] benchmark, a multilingual parallel speech corpus covering 102 languages with high-quality recordings and transcriptions, enabling both text and speech-based evaluation. We consider French, Spanish, German, and Hindi as source languages, with English as the target, spanning Romance, Germanic, and Indo-Aryan language families. The proposed model is trained on the FLEURS training and validation splits. To ensure fair comparison across language pairs, we sample 400 test utterances per pair for evaluation. Dataset statistics are summarized in Table I.

TABLE I
FLEURS evaluation statistics. Total duration (hr) is cumulative; duration (s) and words are averaged per sample.

Lang	Samples	Total Dur. (hr)	Avg Dur. (s)	Avg Words
fr	400	1.75	9.6	17.3
es	400	1.67	10.2	17.8
de	400	1.82	9.7	16.9
hi	400	1.73	9.9	17.1

B. Translated Speech Generation

We generate translated speech using SeamlessM4T-v2 [14], a multilingual S2ST system supporting over 100 languages. For each source utterance, the model produces target speech without intermediate text during inference. SeamlessM4T-v2 is based on the UnitY architecture [23], consisting of a speech encoder (wav2vec 2.0 with Conformer layers), a multilingual encoder–decoder translation module, and a speech decoder that generates discrete units, which are converted to waveforms using a HiFi-GAN [24] vocoder.

C. Experimental Setup

We use the pretrained SeamlessM4T-v2 model [14] ($\sim$2.3B parameters) for direct S2ST. Inference uses temperature 1.0 without beam search, with audio processed at 16 kHz and long inputs ($\sim$ 30s) handled via automatic chunking. The semantic and naturalness heads use MLPs with dimensions $[1024, 256]$. The model is trained for 100 epochs with batch size 8 using AdamW [25] (learning rate 5×10^{-4}, linear warm-up). Experiments are conducted on NVIDIA A100 GPUs (80GB) with FP16.

D. Evaluation Metrics

We evaluate S2ST outputs using text- and speech-based metrics. For text-based evaluation, translated speech is transcribed using Whisper Large v3 [26], and BLEU [9], chrF++ [15], METEOR [16], BLEURT [18], and XCOMET [17] are computed against FLEURS [13] references. Speech-based evaluation uses BLASER [10] and BLASER 2.0 (QE, REF) [11]. The proposed framework is evaluated using Spearman rank correlation (ρ) with the corresponding supervision signals.

V. Results and Analysis

This section presents quantitative findings from our systematic evaluation of text and speech-based metrics, including performance, score distributions, and evaluation coverage.

A. Overall Performance by Language Pair

Table II summarizes corpus-level scores across metrics for each language pair.

1) Text-based semantic metrics: Text-based metrics exhibit varying degrees of sensitivity across language pairs. ASR-BLEU shows high variability (std: 7.70), indicating higher sensitivity to linguistic differences but reduced cross-language consistency. ASR-chrF++ shows similarly high variability (std: 9.33), reflecting sensitivity to language variation. Both

TABLE II
CORPUS-LEVEL SCORES ACROSS TEXT AND SPEECH-BASED METRICS FOR EACH LANGUAGE PAIR.

Language	ASR-BLEU	ASR-chrF++	ASR-METEOR	ASR-xCOMET	ASR-BLEURT	BLASER	BLASER 2.0-Ref	BLASER 2.0-QE
fr–en	37.32	62.27	0.62	0.76	0.74	3.26	4.42	4.46
es–en	29.17	56.32	0.56	0.71	0.73	3.25	4.39	4.57
de–en	25.07	50.28	0.52	0.63	0.70	3.21	4.25	4.32
hi–en	18.95	40.41	0.45	0.58	0.61	3.13	3.82	3.91
Std. Dev.	7.70	9.33	0.07	0.08	0.06	0.06	0.28	0.29

metrics rely on surface-level overlap, which contributes to their increased variability across linguistically diverse pairs. In contrast, ASR-METEOR (std: 0.07), ASR-BLEURT (std: 0.06), and ASR-xCOMET (std: 0.08) exhibit lower variability, indicating more consistent behavior across languages. These metrics incorporate semantic-aware matching, which likely contributes to reduced sensitivity to surface variation and improved cross-lingual consistency.

2) Speech-based semantic metrics: Speech-based metrics exhibit varying levels of variability across language pairs. BLASER shows low variability (std: 0.06), indicating high consistency but comparatively limited discriminative power, while BLASER 2.0-Ref (std: 0.28) and BLASER 2.0-QE (std: 0.29) show higher variability, suggesting improved sensitivity. The small differences between Ref and QE variants indicate close agreement between reference-based and reference-free evaluation. Despite operating on speech, these metrics primarily capture semantic similarity learned from human judgments and show limited sensitivity to acoustic characteristics. Across all evaluated metrics, Hindi–English consistently achieves the lowest scores, possibly reflecting the greater linguistic differences between Hindi and English compared with the European language pairs.

B. Score Distribution and Discriminative Power

Beyond aggregate statistics, score distributions provide insight into a metric's ability to differentiate translation quality at the sentence level. Figure 2 shows the distribution of scores for all metrics on the fr–en evaluation set. Among text-based metrics, ASR-BLEU exhibits the most unstable behavior, with a highly skewed distribution (Mean = 19.94, Median = 13.08, Std = 20.99), indicating poor reliability at the sentence level. ASR-chrF++ shows a more balanced distribution, though it still exhibits high variance (Std = 22.73), indicating sensitivity to score fluctuations. Neural metrics such as ASR-BLEURT (Std = 0.26) and ASR-xCOMET (Std = 0.21) exhibit moderate variance with noticeable skew, capturing a broader quality spectrum while maintaining sensitivity to lower-quality outputs. ASR-METEOR shows the most symmetric distribution (Mean = 0.50, Median = 0.49, Std = 0.23), indicating balanced scoring behavior.

Among speech-based metrics, BLASER exhibits an extremely concentrated distribution (Mean = 3.26, Median = 3.26, Std = 0.02), indicating high scoring consistency but very limited discriminative power. In contrast, BLASER 2.0-QE and BLASER 2.0-REF show broader spreads (Std = 0.31 and 0.38, respectively), enabling improved differentiation across samples while maintaining similar distributional behavior. The close alignment between their distributions further indicates strong agreement between reference-based and reference-free evaluation. Overall, text-based metrics, particularly neural approaches, offer greater variability and finer-grained differentiation, albeit with increased instability. Whereas speech-based metrics provide consistent but weakly discriminative scores.

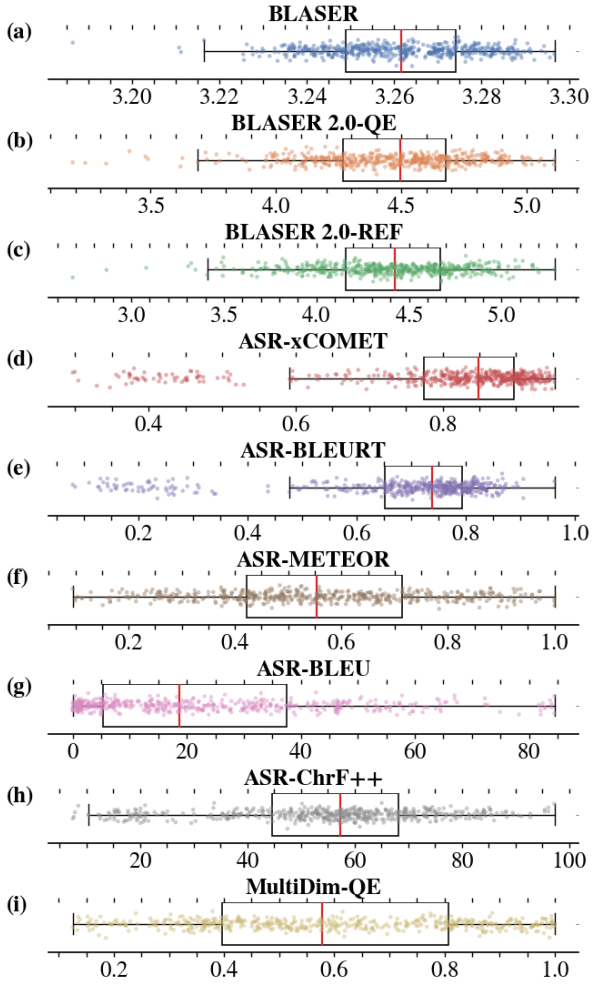

Fig. 2. Sentence-level score distributions for evaluation metrics. Each subplot shows score distributions with median (red line) and quartile range (box).

TABLE III
SPEARMAN RANK CORRELATION (ρ) FOR SEMANTIC (XCOMET-BASED) AND ACOUSTIC NATURALNESS (UTMOS-BASED) PREDICTIONS ACROSS SELECTED LANGUAGE PAIRS.

Model	fr–en	es–en	de–en	hi–en
BLASER 2.0-QE	0.571	0.536	0.529	0.512
MultiDim-QE	**0.634**	**0.621**	**0.614**	**0.593**
ρ MultiDim-Nat	0.913	0.924	0.901	0.892

C. Multi-Dimensional Evaluation

Table III reports Spearman correlation for both semantic (xCOMET-based) and acoustic naturalness (UTMOS-based) predictions across four language pairs. The proposed MultiDim-QE model consistently outperforms BLASER 2.0-QE, achieving an average improvement of ~ 0.08, with lower performance observed for certain language pairs such as Hindi–English. In parallel, the MultiDim-Nat model achieves high correlation with UTMOS (all ≥ 0.89), indicating strong agreement with acoustic naturalness.

As shown in Fig. 2 (i), the MultiDim-QE score distribution (Mean = 0.59, Median = 0.58, Std = 0.24) exhibits a balanced spread, compared to the higher variance of surface-based metrics such as ASR-BLEU and ASR-chrF++, while showing greater dispersion than highly concentrated speech-based metrics such as BLASER.

VI. CONCLUSION

In this work, we presented a systematic empirical analysis of evaluation metrics for S2ST, examining both text- and speech-based approaches across multiple language pairs. Our findings suggest varying strengths and limitations across metric families. Text-based metrics exhibit higher variability and sensitivity to language differences but suffer from instability and reliance on ASR. However, speech-based metrics provide more consistent scoring but show limited discriminative power and primarily capture semantic similarity. Our analysis highlights a critical gap in current evaluation practices, namely the limited assessment of acoustic characteristics essential for realistic S2ST systems. To address this, we propose a multi-dimensional evaluation framework that jointly assesses semantic adequacy and speech naturalness. The proposed framework improves semantic alignment over existing baselines while achieving strong correlation with acoustic naturalness across all language pairs. However, our evaluation is limited to a single S2ST system (SeamlessM4T-v2), a single corpus (FLEURS), and automated supervision signals rather than direct human judgments. Future work will extend this framework to additional languages and evaluation dimensions, including speaker similarity and prosodic expressiveness, as well as validate the proposed metrics against human judgments and evaluate their generalization across diverse S2ST systems and speech conditions.

REFERENCES

[1] S. Dhawan, "Speech to speech translation: Challenges and future," *International Journal of Computer Applications Technology and Research*, vol. 11, no. 3, pp. 36–55, 2022.

[2] T. Etchegoyhen, H. Arzelus, H. Gete, A. Alvarez, I. G. Torre, J. M. Martín-Doñas, A. González-Docasal, and E. B. Fernandez, "Cascade or direct speech translation? a case study," *Applied Sciences*, vol. 12, no. 3, 2022.

[3] S. Alharbi, M. Alrazgan, A. Alrashed, T. Alnomasi, R. Almojel, R. Alharbi, S. Alharbi, S. Alturki, F. Alshehri, and M. Almojil, "Automatic speech recognition: Systematic literature review," *IEEE Access*, vol. 9, pp. 131 858–131 876, 2021.

[4] Z. Tan, S. Wang, Z. Yang, G. Chen, X. Huang, M. Sun, and Y. Liu, "Neural machine translation: A review of methods, resources, and tools," *AI Open*, vol. 1, pp. 5–21, 2020.

[5] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, and J. Weber, "Xtts: A massively multilingual zero-shot text-to-speech model," in *Interspeech*, 2024, pp. 4978–4982.

[6] Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," in *Interspeech*, 2019, pp. 1123–1127.

[7] Y. Jia, M. T. Ramanovich, T. Remez, and R. Pomerantz, "Translatotron 2: High-quality direct speech-to-speech translation with voice preservation," in *ICML*, 2022, pp. 10 120–10 134.

[8] C. Le, Y. Qian, D. Wang, L. Zhou, S. LIU, X. Wang, M. Yousefi, Y. Qian, J. Li, and M. Zeng, "TransVIP: Speech to speech translation system with voice and isochrony preservation," in *NeurIPS*, 2024, pp. 89 682–89 705.

[9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of ACL*, 2002.

[10] H. Le, A. Baevski, M. Auli *et al.*, "Blaser: A text-free speech-to-speech translation evaluation metric," in *Proceedings of Interspeech*, 2023.

[11] H. Le *et al.*, "Blaser 2.0: Scaling speech translation evaluation to 200+ languages," in *Proceedings of ACL*, 2024.

[12] P.-A. Duquenne, H. Schwenk, and B. Sagot, "Sonar: Sentence-level multimodal and language-agnostic representations," *CoRR*, vol. abs/2308.11466, 2023.

[13] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *SLT*, 2023, pp. 798–805.

[14] L. Barrault *et al.*, "Seamlessm4t: Massively multilingual and multimodal machine translation," *CoRR*, vol. abs/2308.11596, 2023.

[15] M. Popović, "chrf: Character n-gram f-score for automatic mt evaluation," in *WMT*, 2015, pp. 392–395.

[16] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of ACL Workshop*, 2005.

[17] N. M. Guerreiro, R. Rei, D. v. Stigt, L. Coheur, P. Colombo, and A. F. T. Martins, "xCOMET: Transparent machine translation evaluation through fine-grained error detection," *TACL*, vol. 12, pp. 979–995, 2024.

[18] T. Sellam, D. Das, and A. Parikh, "Bleurt: Learning robust metrics for text generation," in *Proceedings of ACL*, 2020.

[19] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *ICLR*, 2020.

[20] P.-A. Duquenne, H. Gong, N. Dong, J. Du, A. Lee, V. Goswami, C. Wang, J. Pino, B. Sagot, and H. Schwenk, "SpeechMatrix: A large-scale mined corpus of multilingual speech-to-speech translations," in *ACL*, Jul. 2023, pp. 16 251–16 269.

[21] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[22] T. Saeki, S. Takamichi, and H. Saruwatari, "Utmos: Utokyo-sarulab system for voicemos challenge 2022," in *Interspeech*, 2022, pp. 4526–4530.

[23] A. Lee *et al.*, "Direct speech-to-speech translation with discrete units," in *ACL*, 2022, pp. 332–345.

[24] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *NeurIPS*, 2020.

[25] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.

[26] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023.