E2E VS. CASCADING ON DIFFERENT METRICS
FINAL PROJECT REPORT
SPEECH TECHNOLOGY FOR CONVERSATIONAL AI (11492/11692/18495)

Aadit Deshpande[1], *Evan Fellman*[1], *Alexander Rasskin*[1], *Sophia Topi*[1]

[1] Carnegie Mellon University

aaditd, efellman, arasskin, stopi@andrew.cmu.edu

ABSTRACT

In speech to speech translation (S2ST), both cascading and end-to-end models are competitive according to methods of scoring such as BLEU. We seek to analyze translation abilities of both cascading and end-to-end models through a more holistic approach using new metrics such as deep learning encoding similarities [1] and through synonym leniency measurements.

***Index Terms*—** Speech-to-speech translation, speech translation, cascading, end-to-end

1. INTRODUCTION

Speech-to-speech translation systems are an important area of research for the wealth of potential applications that it can offer. From students to travelers to workers to even consumers of popular entertainment, the benefits of instantaneous and accurate translation of speech from one language to another are clear. Traditionally, speech-to-speech translation systems have been cascading, where the task is broken into separate automatic speech recognition (ASR), machine translation (MT), and finally text-to-speech (TTS) tasks. Although this allows for optimization of the individual sub-tasks (and thus improved performance on the overall task), it also allows for error propagation throughout the entire system. More recent systems have taken an end-to-end approach, where one encoder-decoder system is trained end-to-end, with no intermediary representations. This approach can be less resource-intensive, and does not allow for the propagation of ASR errors, but also does not benefit from improvements in ASR or MT. Particularly, end-to-end speech translation systems are text-free, and thus cannot benefit as cascading systems can from the larger quantities of high-quality text training data and pretrained text-based models. Consequently, end-to-end systems have tended to lag behind cascading systems in performance, though recent studies have shown that that performance gap (in terms of BLEU and TER scores) is nearly closed for speech-to-text based systems [1].

Our goal is to measure any differences in performance in end-to-end and cascading speech-to-speech translation systems using a holistic analysis through metrics that take into account more than just the exact translation accuracy (such as ASR-BLEU).

2. RELATED STUDIES

Until very recently, cascaded systems have usually outperformed end-to-end systems on most available metrics [2]. More recent studies have shown that direct end-to-end systems can now improve on the BLEU scores of cascaded systems in S2ST applications, even when trained on comparable datasets [3] [4]. However, these studies still tend to rely on converting the results of S2ST back into text using ASR and using text-based metrics like BLEU for evaluation. There are some textless evaluation metrics that have been developed including BLASER2.0, a modality-agnostic model-based metric that compares input and output sentence embeddings [5]. Other metrics suffer from a lack of consideration for evaluating paraphrasing, synonyms, and prosodic information in the synthesized speech, which metrics like METEOR have attempted to address [6]. The push to move beyond simple synthesis of translated speech, and to preserve the characteristics of the input speech, has led to some end-to-end S2ST systems also making attempts to transfer timbre or speaker characteristics into the synthesized speech [7] [8] or other prosodic information [9], for what is called expressive S2ST. However, the metrics for these expressive S2ST systems are rarely applied outside of systems specifically designed for them. Our goal is to include all of these strategies of evaluation in order to assess more completely the performance differences between cascading and end-to-end S2ST systems.

3. METHOD

An end-to-end and a cascading model were evaluated on shared datasets of untranslated speech in Spanish (222 hours). Source speech was obtained from the Common Voice corpus—a massively-multilingual collection of speech utterances and their transcriptions [10]. For each model we will use the following evaluation metrics:

- **ASR-BLEU**: An ASR system is run on our output speech to generate a hypothesis transcript. BLEU is a commonly used metric that relies on calculating the n-gram similarity (exactly matches words) between the predicted and reference text. The BLEU score here is computed between our generated transcript and target transcript provided by the dataset. Using ASR-BLEU, we also keep track of the Brevity Penalty (BP) and the Hypothesis to Reference Ratio (HRR).
- **BLASER 2.0** [5]: A text-free evaluation method that calculates the similarity between input and output speech embeddings (SONAR) generated via a large multilingual and multimodal encoder. BLASER 2.0 considers semantic similarity; it has previously been shown to correlate more closely with human judgment than ASR-BLEU. [5]
- **ASR-METEOR** [6]: Similar to ASR-BLEU but using METEOR for our evaluation score. METEOR calculates sentence level similarity through four criterion: exact word, stem word, synonyms and paraphrase appearance. To do

Model	ASR-BLEU	Hypothesis	Reference	Note
Cascaded	52.56	Very much of the reconstruction of the metropolitan cathedral.	Viceroy amar embarked on the reconstruction of the metropolitan cathedral.	Unnatural sentence is rewarded with unfairly high BLEU score
Cascaded	32.16	For the Spanish Civil War he assailed the United Kingdom.	He was exiled to the United Kingdom following the Spanish Civil War.	High BLEU score despite semantic mismatch (”assailed” vs. ”exiled”).
Cascaded	40.05	Cots are almost at the same height.	You can find the keys almost at the same height.	Fails to capture meaning of sentence.
E2E	5.86	Lets have it in wine.	A glass of wine will improve your night.	Low BLEU score despite high semantic similarity (paraphrase).
E2E	5.07	For other dogs so men and rusted in skin.	To another dog with that bone so fleshless and stiff.	Low BLEU score despite high semantic similarity (paraphrase).
E2E	50.00	My harmon.	My master.	Imperfect BLEU score despite near-perfect semantic match (synonym).

Table 1. Qualitative analysis of the E2E and Cascaded S2ST results on the CVSS-C Spanish-to-English dataset. Using these examples we show the drawbacks of relying on ASR-BLEU as the sole metric of translation quality. It suffers from several issues like ignoring synonyms, grammar, and naturalness.

this, it utilizes universal language parameters (learned from set of 100,000 rankings from WMT12), WordNet (language-specific data), and paraphrase tables.

- **COMET** [11]: PyTorch-based framework for training MT evaluation models that can function as metrics. Our chosen COMET model (unbabel/wmt22-comet-da) was trained on collection of human judgements (adequacy direct assessments) collected by organizers of the WMT News Translation Shared Task. [11]

4. EXPERIMENTS

4.1. Datasets

Three datasets were used for our S2ST systems:

- OWSM (cascading): 180k hours of labeled speech to text with translation for all sources (non-English speech to English text). [12]
- LJ Speech (cascading): 24 hours of single speaker text to speech data (English text to English speech) [13]
- CVSS-C (end2end): 1k hours of speech to speech with translation for all sources (non-English speech to English speech) [14]

4.2. Baselines

For the cascading model baseline, we used the untuned pre-trained OWSM 3.1 model from ESPNet [12] for translating the non-English speech to English text (S2T). We used the fastspeech2 conformer from ESPNet and the fastspeech2 conformer hifigan vocoder [15] [16].

For the end-to-end model, we used an untuned discrete-unit S2ST ESPNet model [17] pre-trained on the source speech, a Spanish-to-English subset of the CVSS-C dataset (speech and transcription data from commonvoice) [14]. The target speech came from CoVOST2 [18]. The output speech is synthesized using the Parallel WaveGAN huBERT vocoder [17].

For evaluation of any metrics that used text instead of speech, we first used the Branchformer LibriSpeech ASR model [19] to convert the output of our S2ST systems into text.

As baselines, we tested both the cascading and e2e systems on the validation set to prevent our experimental decisions to be impacted by performance on the test set. We withheld testing on the test set until the very end to prevent our experimental decisions to have unintentional train-test bleeding biases. We forward fed both our cascading model and end-to-end model on our validation set, from CVSS-C, then forward fed both through the ASR model, Branchformer LibriSpeech. From here, we compare to the given golden transcriptions by CVSS-C using the BLEU metric. The results from this baseline experiment are shown in Table 2. Link to code: `https://github.com/Aadit3003/s2st-cascading-e2e`.

Model	ASR-BLEU	BP	HRR
E2E	15.007	0.822	0.929
Cascading	18.428	0.872	0.953

Table 2. Comparison of the out-of-the-box performance of E2E and Cascading S2ST Models on the CVSS-C Spanish-to-English dev set. BP is Brevity Penalty, HRR is Hypothesis to Reference ratio, ASR-BLEU is the standard speech to speech translation metric.

In Table 2 above, we show our ASR-BLEU results for the baseline out-of-the-box systems. We hypothesize that ASR-BLEU is lower due to the prevalence of high quality text training data and high quality pretrained and generalizable text based models which gives the cascading method an advantage. From our error analysis and results from the brevity penalty metric, our E2E method tends to generate shorter sentences. We found that the semantic quality of E2E results can be better than that of the cascading system, but still receive lower BLEU scores. We show some of the results from our error analysis in Table 1. Table 1 shows the necessity for using alternate and more holistic metrics of translation quality such as measurements of semantic meaning or even measuring how ”natural” the translation appears. ASR-BLEU does not reward sentences that are semantically similar if they do not share sufficient n-grams, for example. To investigate this, in our experiments, we will score the results of the e2e system and the baseline and fine-tuned cascading systems with multiple metrics, to test our hypothesis that the end-to-end system would perform better on more forgiving, semantically-based, or human-judgment-aligned metrics.

Model	ASR-BLEU	BP	HRR	COMET	METEOR	BLASER2.0
E2E-oob	14.901	0.82	0.928	0.538	0.283	3.188
casc-oob	**17.692**	**0.88**	**0.975**	**0.619**	**0.338**	**3.604**
casc-ft-best	15.062	0.709	0.785	0.599	0.323	3.435
casc-ft-1-epoch	14.930	0.705	0.784	0.593	0.318	3.386
casc-ft-5-epoch	14.383	0.702	0.784	0.601	0.314	3.428
casc-ft-low-lr	13.031	0.636	0.722	0.570	0.281	3.298

Table 3. Comparison of the performance of E2E and Cascading S2ST Models on the CVSS-C Spanish-to-English test set. *E2E-oob* is out-of-the-box end-to-end model. *casc-oob* is out-of-the-box cascading model. casc-ft models are the fine-tuned (via LoRA) cascading models: *casc-ft-best* (learning rate 1e-5, epoch 3); *casc-ft-1-epoch* (learning rate 1e-5, epoch 1); *casc-ft-5-epoch* (learning rate 1e-5, epoch 5); and *casc-ft-low-lr* (learning rate 1e-7, epoch 1). BP is Brevity Penalty (scale 0-1), HRR is Hypothesis to Reference ratio (scale 0-1); ASR-BLEU (scale 1-100), COMET (scale 0-1), METEOR (scale 0-1), and BLASER2.0 (scale 1-5) are metrics described in Section 3.0 above.

Ex	Model	ASR-BLEU	COMET	METEOR	BLASER	Hypothesis	Reference
1	e2e	32.56	0.535	0.499	2.881	ay to cut the ides	i need to cut the noodles
2	e2e	32.47	0.917	0.807	4.291	if you believe i can fly	if i believed i can fly
3	e2e	8.30	0.460	0.232	3.041	and the noble marquises no marquises it is **wan to human**	will the noble marquis even ignore who **juan de guzman** is
4	casc-oob	25.85	0.860	0.605	4.129	son of **juan remon bernados** and **dolores filigara**	son of **juan ramon bernandez** and **dolores filgueira**
5	e2e	4.46	0.503	0.135	2.739	we have to have it we will have to have it	we got him **ricardo** we got him
6	casc-oob	16.23	0.695	0.368	4.098	i love you very much **ricardo**	it hurts so much **ricardo**
7	e2e	12.79	0.860	0.460	3.932	well i must return to my ship	well i need to go back to repair my ship
8	e2e	23.17	0.672	0.448	3.905	not a toy	it is not a flying toy

Table 4. Examples for qualitative analysis of errors seen in the results of different models.

4.3. Experimental Setups

We implemented our systems using the models and ESPNet recipes described above in Sections 4.1 and 4.2. We then chose to fine-tune our cascading system on CVSS-C data (for better comparison between the e2e and cascading systems) using LoRA [20]. We used a LoRA rank of 4 and trained for 10 epochs. In total, we trained on A6000 and A100_80GB GPUs for 7-9 hours. For three of the fine-tuned models we used a learning rate of 1e-5 (*casc-ft-best, casc-ft-1-epoch, casc-ft-5-epoch*). We also included a model with a learning rate of 1e-7 (*casc-ft-low-lr*). The fine-tuned model with the best performance, labeled *casc-ft-best* in Table 3 above, was epoch 3 with a learning rate of 1e-5.

We selected a randomized test set from the CVSS-C data of size 1000, and ran it through our e2e system and multiple versions of our cascading system. We then ran all of the metrics described in Section 3.0 above on the resulting translations. The results are summarized in Table 3 above.

4.4. Results and Discussion

Overall, we saw that the cascading system out-of-box performed the best across every metric. We also saw that every cascading model (except the casc-ft-low-lr model) performed better than the e2e model on every metric we evaluated other than ASR-BLEU (COMET, METEOR, and BLASER2.0). This appears to disprove our hypothesis that our e2e system would produce semantically better results that would be reflected in metrics other than ASR-BLEU. Fine-tuning the cascading system did appear to decrease the performance. Even after adjusting the learning rate and other hyperparameters, we were unable to achieve any improvement. We hypothesize that this is a result of overfitting on training data, as well the training data in CVSS-C being of a lower quality than the original training data in OWSM.

Like in our initial experiments, we see translations from our systems whose BLEU score is somewhat misleading. In examples 1,2 in Table 4 above, we can still see that, despite having similar BLEU scores, the translation in example 2 would clearly be better according to human judgment. However, this is now reflected by our other metrics: the COMET, METEOR, and BLASER scores are all far higher for the translation in example 2 than the translation in example 1, even though the BLEU scores are about the same. This proves that BLEU is not a perfect metric, and that our other metrics are capturing something closer to human judgment.

Another point of interest is that, even when the ASR-BLEU scores were relatively similar (E2E-oob had ASR-BLEU of 14.901 and Casc-ft-1-epoch had ASR-BLEU of 14.930), the cascading system still performed significantly better on the other metrics as well. Even though the Casc-ft-1-epoch's ASR-BLEU was only 0.194% higher than the e2e-oob model, its COMET score was 9.7% higher, its METEOR score was 11.6% higher, and its BLASER2.0 score was 6% higher than the e2e system. This indicates not only that the non-BLEU metrics are measuring something that BLEU is not (otherwise we would expect systems that get similar BLEU scores to also get similar scores on the other metrics), but that the fine-tuned cascading system is performing better on those measures than the e2e-oob.

We conducted a qualitative error analysis to gain more insight into the differences between e2e and cascading systems (see Table 4

above). We found the following:

- **Named entity recognition**: The cascading models tended to be better at recognizing named entities (see examples 3,4 and 5,6 in Table 4 above) and, even if the ASR spelled them incorrectly, they were less likely to be mistaken for another entirely different phrase, thus scoring better on all metrics.
- **Summarization**: As discussed in our hypothesis, the e2e system did tend to produce shorter sentences (as seen in lower BP/HRR in Table 3). When these shorter sentences still consisted of appropriate translations, although they scored well in other metrics like BLASER2.0 and COMET, they still had poor BLEU scores (see examples 7,8 in Table 4).
- **Brevity**: The e2e-oob system had a lower BP and HRR than the e2e model, indicating its sentences were on average shorter (see examples 7,8 in Table 4). However, the fine-tuned cascading systems had significantly lower BPs and HRRs than either the cascading-oob or e2e system, despite having similar ASR-BLEU scores to the e2e model. These systems having similar BLEU scores but different BPs indicates that, though the fine-tuned cascading systems gave shorter translations, the actual proportion of matching n-grams must have increased.

5. CONCLUSION

Our hypothesis, that a more holistic approach to metrics - using metrics that aim to measure correlation to indicators of human judgment, or be more forgiving in ways that humans are with synonyms and paraphrasing and semantics - would result in improved performance for e2e systems relative to cascading systems, did not appear to be true. This study did show that the ASR-BLEU scores do not always perfectly correlate with other metrics, proving that there are good reasons to use metrics other than just ASR-BLEU when evaluating the quality of S2ST systems. However, it appears that, when comparing very similar e2e and cascading S2ST systems, cascading systems still outperform their e2e counterparts, and the e2e systems are unable to overcome the cascading systems' advantages, including the large quantities of training data available for text-based models and high quality pretrained text-based models. Future studies could explore if this performance gap across all metrics is true for languages with different characteristics (low-resource languages, more agglutinative languages, or more dissimilar languages) or when using more similar datasets.

6. REFERENCES

[1] L. Bentivogli, M. Cettolo, M. Gaido, A. Karakanta, A. Martinelli, M. Negri, and M. Turchi, "Cascade versus direct speech translation: Do the differences still make a difference?," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Association for Computational Linguistics, 2021, vol. 1: Long Papers, pp. 2873—-2887.

[2] Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," 2019.

[3] Xinjian Li, Ye Jia, and Chung-Cheng Chiu, "Textless direct speech-to-speech translation with discrete speech representation," 2022.

[4] Ye Jia, Yifan Ding, Ankur Bapna, Colin Cherry, Yu Zhang, Alexis Conneau, and Nobuyuki Morioka, "Leveraging unsupervised and weakly-supervised data to improve direct speech-to-speech translation," 2022.

[5] Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang, "Seamlessm4t: Massively multilingual i& multimodal machine translation," 2023.

[6] Michael Denkowski and Alon Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, and Lucia Specia, Eds., Baltimore, Maryland, USA, June 2014, pp. 376–380, Association for Computational Linguistics.

[7] Kun Song, Yi lei, Peikun Chen, Yiqing Cao, Kun Wei, Yongmao Zhang, Lei Xie, Ning Jiang, and Guoqing Zhao, "The npu-msxf speech-to-speech translation system for iwslt 2023 speech-to-speech translation task," 2023.

[8] Kun Song, Yi Ren, Yi Lei, Chunfeng Wang, Kun Wei, Lei Xie, Xiang Yin, and Zejun Ma, "Styles2st: Zero-shot style transfer for direct speech-to-speech translation," 2023.

[9] Wen-Chin Huang, Benjamin Peloquin, Justine Kao, Changhan Wang, Hongyu Gong, Elizabeth Salesky, Yossi Adi, Ann Lee, and Peng-Jen Chen, "A holistic cascade system, benchmark, and human evaluation protocol for expressive speech-to-speech translation," 2023.

[10] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," 2020.

[11] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie, "Comet: A neural framework for mt evaluation," 2020.

[12] Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, Wangyou Zhang, Yui Sudo, Muhammad Shakeel, Jee weon Jung, Soumi Maiti, and Shinji Watanabe, "Reproducing whisper-style training using an open-source toolkit and publicly available data," 2023.

[13] Keith Ito and Linda Johnson, "The lj speech dataset," `https://keithito.com/LJ-Speech-Dataset/`, 2017.

[14] Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen, "Cvss corpus and massively multilingual speech-to-speech translation," 2022.

[15] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," 2022.

[16] Pengcheng Guo, Florian Boyer, Xuankai Chang, Tomoki Hayashi, Yosuke Higuchi, Hirofumi Inaguma, Naoyuki Kamo, Chenda Li, Daniel Garcia-Romero, Jiatong Shi, Jing Shi, Shinji Watanabe, Kun Wei, Wangyou Zhang, and Yuekai Zhang, "Recent developments on espnet toolkit boosted by conformer," 2020.

[17] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Yalta, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proceedings of Interspeech*, 2018, pp. 2207–2211.

[18] Changhan Wang, Anne Wu, and Juan Pino, "Covost 2 and massively multilingual speech-to-text translation," 2020.

[19] Kwangyoun Kim, Felix Wu, Yifan Peng, Jing Pan, Prashant Sridhar, Kyu J. Han, and Shinji Watanabe, "E-branchformer: Branchformer with enhanced merging for speech recognition," 2022.

[20] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "Lora: Low-rank adaptation of large language models," 2021.