The Interspeech 2026 Challenge on Transfer of Pragmatic Intent in Speech-to-Speech Translation

Nigel G. Ward[1,**], *Marcel de Korte*[1], *Javier Vazquez*[1], *Montserrat G. Molina*[1], *Vanessa Bolado*[1], *Carol Figueroa*[1], *Eliya Nachmani*[2], *John E. Ortega*[3], *Satoshi Nakamura*[4]

[1]University of Texas at El Paso, USA [2]Ben Gurion University, Israel
[3]Northeastern University, USA [4]Chinese University of Hong Kong, Shenzen, China

nigelward@acm.org, mdekorte@utep.edu, jvazquez073015@gmail.com, mgmolina1313@gmail.com, vanessa.bolado.garcia@gmail.com, cfigueroa2@utep.edu, eliyanac@bgu.ac.il, johneortega@gmail.com, snakamura@cuhk.edu.cn

Abstract

The Interspeech 2026 challenge task on Transfer of Pragmatic Intent in Speech-to-Speech Translation was motivated by the fact that current speech-to-speech translation systems do not prioritize pragmatic fidelity, and thus do not ideally support people conversing across languages. This paper describes the design, datasets, and evaluation methods for the challenge, and the results for the submissions. We find that even the best system lags human performance by 1.2 points on a 5-point scale, and note strengths and weaknesses of our evaluation methods.
Index Terms: pragmatic fidelity, subjective evaluation, automatic evaluation, dialog, conversation, prosody, benchmarking

1. Motivation

Speech-to-speech translation (S2ST) has the potential to enable people to communicate conversationally across language barriers, fostering personal fulfillment, social inclusion, and economic growth. Recent years have seen tremendous progress, but current speech-to-speech translation systems, while adequate for some purposes, are not entirely well matched to the needs involved in supporting dialog [1]. In particular, their outputs are usually insensitive to the interpersonal and pragmatic goals [2, 3, 4, 5, 6] that are present in many source utterances. For example, given the input *did you see her?* with its specific prosody — perhaps showing breathless interest, encouraging the interlocutor to continue his story, and displaying sensitivity to the complex emotions he's feeling after a break-up — the output should ideally translate not only the lexical content, but also these elements of stance and intent.

One reason that current systems tend to omit such elements is that pragmatic fidelity has never been systematically evaluated. To date, S2ST evaluations have focused on mainly on semantic fidelity and output naturalness. We accordingly created a challenge task and invited all comers to test their systems' ability to transfer pragmatic intent across languages.

Through this challenge task, we aimed to 1) measure how well current technology succeeds at this aspect of translation, 2) discover what types of pragmatic functions are not well handled by current technology, and 3) explore the validity of pragmatics-related evaluation methods for S2ST.

2. Related Work

This challenge task aligns with current interest in S2ST, and especially in expanding its scope, as seen, for example, in the recent flowering of work on speaker-property transfer and transfer of emotion and expressiveness through prosody [7].

It also aligns with recent initiatives to improve the evaluation of S2ST [8, 9, 10]. Beyond metrics for semantic fidelity and naturalness, recent work has also considered preservation of speaker identity and emotion, latency and isochrony, spatial cues, and aspects of style and expressivity [11, 12, 13, 14, 15, 16, 17, 18], however, only one pragmatic function has received significant attention: emphasis [19, 20, 21]. Despite such advances, evaluation in S2ST still relies heavily on methods derived from the more mature fields of speech synthesis and machine translation [22, 23, 24]. However, S2ST is essentially different: it involves more information than is present in text alone, especially the extra pragmatic meanings that are conveyed largely through prosody. The current challenge is the first to focus on these aspects of S2ST.

3. Task and Data

The overall task aim was for systems to produce pragmatically-faithful translations of utterances taken from conversations. There was an English-to-Spanish direction and a Spanish-to-English one (En→Es and Es→En, respectively).

Our data strategy prioritized quality over quantity. We wanted to include a diverse sampling of pragmatic functions, especially of the types that are common in natural conversations. We also wanted to ensure that the pragmatic functions were as similar as possible for every source-target pair.

The test data was derived from 10 conversations among 5 Spanish-English bilinguals, speaking American English and Northern Mexican Spanish. These conversations were unscripted and unprompted, and covered numerous topics, such as pets, cooking, teaching, and taking classes. Following our existing protocol [25], from these original conversations, the participants jointly selected topics where the conversation was lively or interesting in some way, and re-enacted selected utterances from those topics in the other language. These utterances were typically sentence-length, averaging 2.5 seconds, but ranging from 0.4 s to 6.0 s. The re-enactments were done carefully, with the original participants making as many attempts as needed until they both were satisfied that each utterance closely matched the original in tone and communicative intent.

The test data consisted of these utterances and re-enactments: 240 in the English-to-Spanish direction and 199 in the Spanish-to-English direction. We were prepared to also provide metadata, corrected transcripts, and utterance contexts, but no particpating team requested these.

[**]indicates the corresponding author.

For training, due to the dearth of parallel speech-to-speech data, especially conversational data, we expected teams to rely on other resources and on indirect training methods [26, 27]. However a small set of matched training data was available, namely 2893 English-Spanish utterance pairs [25, 28] collected using the same protocol, recorded in the same studio, and with speakers from the same demographic, and thus well matched and potentially useful for tuning hyperparameters, etc.

4. Subjective Evaluation Protocol

Our evaluation strategy also prioritized quality, rather than number of metrics or number of judges. Thus we had four Spanish-English bilinguals score the pragmatic fidelity of each translation: they rated each in terms of "pragmatic fidelity to the original, in terms of tone, feeling, and intent." We did not provide definitions for any of these terms, but we did explain that the target scenario was S2ST for conversational communication, so participants were implicitly guided to focus on the quality aspects relevant to successful communication. We stressed that they were to rate only pragmatic fidelity, and in particular to disregard differences in speaker characteristics and differences in the semantic content of the translations. In the end, participants appeared to interpret "pragmatic fidelity" fairly broadly, as discussed below. Ratings were on a scale from 1 (no similarity) to 5 (equivalent) and entered directly, to two significant digits, on a spreadsheet.

For each original utterance, they rated three candidate translations: two from S2ST systems and the human re-enactment. Rating one set of candidates involved ten listenings: participants first heard the original twice, then the first candidate twice, then the original, then the second candidate twice, then the original, and finally the third candidate twice. The candidates were presented in random order. We replayed utterances as often as participants requested.

Participants needed to be able to listen closely, make sensitive judgments, reflect on why they made these judgments, and maintain sustained attention. We therefore recruited participants carefully. All were students or former students of the University of Texas at EL Paso. Two had been involved in creation of the test data, so at times they were judging translations of their own utterances, including by themselves. Participants were compensated generously by local standards: $70 plus lunch for 3 hours of work.

In total, each participant judged 59 sets, for 177 total ratings. For the first six sets, we asked participants to share their perceptions and reasons for their ratings, focus-group style, and allowed them to change their ratings after discussion. After that they made their ratings independently. Finally we had a qualitative phase, where they listented again to the first ten sets and discussed the factors influencing their judgments and what they perceived as the common patterns of strengths and weaknesses.

5. Automated Evaluation Metric

The automated metric compares the system output to the re-enactment. (While, for the subjective evaluation, the latter was evaluated with the others on an equal footing, here it was taken as the gold standard.) Specifically, we use the Segura metric [29], which estimates the pragmatic similarity between any two utterances of English (respectively, of Spanish). This computes the cosine similarity between the representations of two utterances in terms of a subset of average-pooled HuBert features: 103 features (respectively 101 for Spanish). In contrast to other similarly-motivated evaluation metrics, this aims to measure pragmatic appropriateness directly, rather than via aspects of the prosody [30, 31]. We used this metric out-of-the-box, although it was not designed specifically for the task of S2ST evaluation. However it was tuned to maximize match to a large collection of human judgments of "pragmatic similarity in terms of tone, feeling, intent" [32], which is essentially the same instrument used here in the subjective evaluation. Morever, our recent exploration of its suitability for this purpose found it to often match human perceptions of similarity in S2ST [33]. However, that work also noted two weaknesses. First, while applicable to both systems that output audio and those that output features, it does not support comparisons across the two conditions: it appears to be easier to specify the necessary prosody than to actually synthesize an utterance with actual words. Second, the metric is occasionally over-sensitive to nuances of hesitancy and confidence.

6. Challenge Task Logistics

We originally offered three conditions of participation. One was designed to enable participation by researchers with insights in how specific prosodic patterns serve various functions in the two languages, but who might lack extensive modeling skills. In this condition, teams were to map from a 100-feature description of the prosody of an utterance in one language to a 100-feature description of a pragmatically equivalent utterance in the other [34]. However no teams took up this challenge, so in the end there was participation in only two conditions:

C1. S2ST Systems. In this condition, the input and output were both audio.

C2. Feature-Mapping Systems. In this condition, both input and output were feature vectors that captured the key pragmatics-related aspects of utterances. There were the features mentioned above: 103 features for the English utterances and 101 for the Spanish respectively. Thus, for the English-to-Spanish direction, systems were required to map from 103 features to 101, and conversely. For this condition, to lower the barriers to entry, we released a Jupyter notebook with a baseline system, described below. For this we modified the original code [33] to make it easier to swap in new models, or, with a little more effort, to try new loss functions or training data.

As seen in Figure 1, the S2ST systems were evaluated both by human judgments and by the automated metric. However, as noted above, the results of the latter are not comparable across the two conditions.

Submission of feature-mapping systems were processed through CodaBench [35] and the scores automatically computed, but audio submissions were handled through email.

7. System Descriptions

Although we received numerous expressions of interest, including from large corporate teams, and several informal submissions, in the end only 4 teams made official submissions, all university teams. We attributed the modest amount of participation to the novelty of the task, to the relatively small size of the S2ST community, and to the tight timeline, with only seven weeks from the announcement of the challenge to the submission due date. Thus we evaluated:

CUHK-SZ (C1, Es→En) The system from the Chinese University of Hong Kong, Shenzen [36] was a cascaded pipeline using Whisper for speech recognition, Qwen2.5-72B-Instruct to translate the Spanish text to English, and

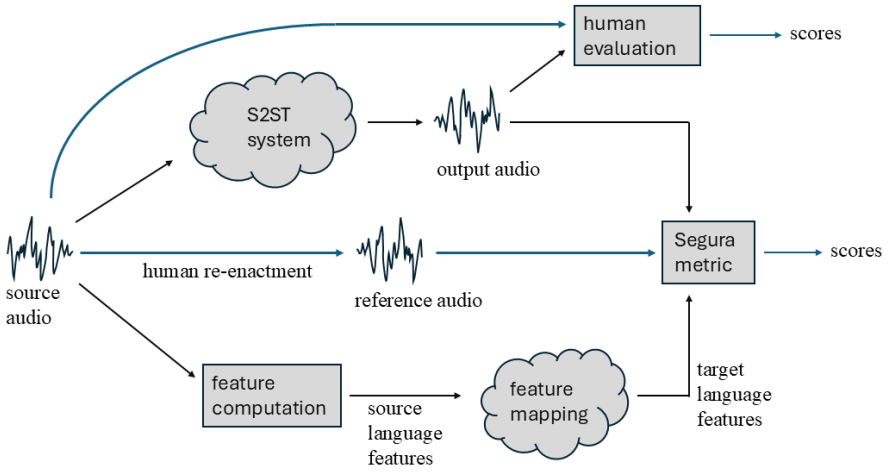

Figure 1: *Summary of conditions and evaluations.*

	average	std. dev
human	4.59	.36
CUHK-SZ	3.46	.63
Seamless	2.98	.64

Table 1: *Spanish to English, audio-to-audio, human ratings of pragmatic fidelity*

system	Segura
CUHK-SZ	0.7270
Seamless	0.7175

Table 2: *Spanish-to-English, audio-to-audio, automatic metric*

FishAudio/OpenaAudio-S1-Mini to synthesize the English speech. The key innovation was augmenting the synthesizer input with two prompts: the Spanish source speech and the recognized Spanish text.

MUC (C2 En→Es) The system from Minzu University of China (with one team member from the Chinese University of Hong Kong, Shenzen, who was not part of the team for the other submission) [37] improved the baseline feature-mapping system by replacing the multi-layer perceptron with a transformer, and by augmenting the loss function to include the Euclidean distance between target and output in terms of 100 prosodic features, specifically, the "Marco" metric' [31].

BLCU (C2 En→Es) The system from Beijing Language and Culture University [38] exploited a retrieval approach: its core module retrieved the 70 most similar English utterances in the training data and then predicted the Spanish features as the similarity-weighted average of those of the 70 translations.

PAPTAN (C2, both directions) The team from the Université Paris Cité [39] created a system that combined ideas from Seamless and Hibiki [17, 40].

Baseline (C2, both directions) was a simple multi-layer perceptron operating over the provided features, and was trained on the matched data [25, 33].

Seamless (C1, Es→En) [17] was included to have another point of comparison, since it is open-source and since it performed best in an earlier evaluation [33].

8. System Performance: Audio Condition

Table 1 shows the subjective evaluation results for the 59 sets. All differences were significant by a matched-pairs t-test at $p < 0.01$. While the CUHK-SZ submission outperformed Seamless, its translations were rated higher than Seamless's only 77% of the time, suggesting that the two systems may have complementary strengths. Interestingly, for one input, CUHK-SZ translation was rated better than the human re-enactment.

Table 2 shows the results for the automatic metric. According to this also, CUHK-SZ outperforms Seamless, but this difference was not significant.

8.1. Qualitative Observations

In the qualitative phase, the most commonly mentioned weakness was a translation omitting some kind of pragmatic function. To probe further, we did a follow-on exploration: we roughly identified the non-semantic aspects we heard in a sample of the original utterances; informally grouped them into common factors, in the spirit of thematic analysis [41]; and tabulated how often these were carried over into the translations.

Across the 32 sets we examined, we counted the originals to be expressing 107 functions (60 different functions), with the average utterance expressing 3.3 functions. In contrast CUHK-SZ conveyed only 30, and Seamless 26: 28% and 24%, respectively. This aligns with the subjective impression, also voiced by the judges, that both systems had a general tendency to output neutral, read-style speech, devoid of pragmatic intent.

Some aspects were handled relatively well by one or both systems, notably marking questions and making negative assessments, although neither was common in this data. Of the more common aspects, marking lexical emphasis and using a story-telling style or an explaining style were often transfered, although still omitted around half the time. Other common aspects were handled only rarely, including marking a realization, showing hesitancy, thoughtfulness, or amusement, grounding a new referent, and using a conversational style. One common function, positive assessment, was never successfully transfered; nor were a long tail of rarely-seen aspects.

One thing that we had not anticipated was the significance of stylistic differences, as also noted by the judges. In contrast to pragmatic-function differences, stylistic differences gener-

system	score
MUC, submission 1	0.8601
baseline	0.8574
MUC, submission 2	0.8534
PAPTAN	0.6259

Table 3: *English-to-Spanish, feature-mapping averages, Segura metric*

system	score
baseline	0.8054
PAPTAN	0.5613

Table 4: *Spanish-to-English, feature-mapping averages, Segura metric.*

ally reflect the ongoing dialog activity, rather than intents local to a single utterance. The style of these utterances was generally low-key, soft-spoken, and casual in style, often superimposed with more specific styles, such as storytelling, humorous, evocative of a scene, or bidding for empathy. In contrast, the style of the systems' outputs were generally quite neutral.

Apart from weaknesses of omission, another form of weakness, much less common, involved the presence in the system output of a pragmatic function that was not present in the input. In some cases the technical causes of this were obvious: some were related to speech recognition errors, and some related to prosody, including misinterpretation of input prosody and glitches of output prosody, such as realizing a slight contrast as strong emphasis. Within this category, there seemed to be a tendency for systems to struggle with the shortest utterances, such as *ohhh*, and with the longest ones, especially those including reported speech or a within-utterance change of speaker intention.

Examples appear at [https://www.cs.utep.edu/topi/].]

9. System Performance: Features Condition

For the feature-to-feature mapping systems, Tables 3 and 4 show the performance according to the automated metric. Only one system outperformed the baseline, and that by a very small amount. Incidentally, the Spanish-to-English direction seems to be the harder task, as previously observed and discusse in [33].

10. Evaluation Protocol and Metric Properties

	correlations
human vs human, pairwise, avg	0.613
Segura vs avg human	0.512
Marco vs avg human	0.258

Table 5: *Spanish to English, agreement over 53 ratings*

A secondary goal of the challenge was to understand the strengths and weaknesses of our subjective evaluation protocol and of automated metrics.

We first computed the inter-annotator agreement. This ranged from .48 to .80, averaging 0.61, as seen in 5. From the qualitative discussion, there was no evidence that the differences reflected real disagreements, but rather they appeared to be mostly be driven by variation in what each judge was most sensitive to. We also found that the judges did not report difficulty in rating the pragmatic aspects as independent of other considerations, although they did of course notice the occasional semantic lapses and audio artifacts. Overall the subjective evaluation protocol appears fairly solid.

Yet one issue did surface: from the qualitative discussion, we realized that some of the factors that influenced judges' ratings, while certainly relevant to pragmatic fidelity, might not be relevant to real use cases. For example, various issues (notably with discourse markers such as *ohhh*, with hesitation markers and disfluencies, with indicating the intent to continue, and with style differences) may be less important in practice, since, in actual technology-mediated communication, some of these may be not used, not important, or redundant to postural and gaze cues, etc. Thus there is scope to refine the criteria and/or the data selection for future evaluations. However this should be done wisely, based on examination of the specific needs of specific use cases, perhaps building on work like [42, 43, 44].

Regarding the quality of the automatic metric, Table 5 suggests that the Segura metric, while certainly not suitable as the sole basis for evaluation, may have utility as an initial quality estimator.

The table also shows the correlation with evaluations done using the Marco metric [31]. This metric compares two utterances by the Euclidean distance over 100 features, with appropriate weights, which describe various aspects of utterance prosody known to have pragmatic import in various languages, including pitch range, creakiness, breathiness, lengthening, intensity, etc, each over windows that together span the whole utterances. It did not perform well..

11. Conclusions and Prospects

In this paper we have presented a novel challenge task for speech-to-speech translation, with unique aims, data, and evaluation methods. Although small in scale, interesting findings emerged.

Regarding performance, we find that the gap between the best system and human performance is 1.2 points, on a 5-point scale, and thus there is still lots of room for improvement. The performance gap between a top open-source system and the top research system is 0.5 points, suggesting that some improvement may be readily deployable. Further, while existing systems have many strengths, there are many pragmatic functions that are not translated reliably, as are many aspects of style.

Regarding evaluation, we find that a simple protocol for obtaining human judgments of pragmatic fidelity is reliable, with judges' ratings mutually correlating at around 0.61. We also find that automated metric correlates 0.51 with human judgments.

To support future evaluation of methods for improving pragmatic fidelity, we release the inputs and target outputs (the re-enactments), and to support further research on metrics and protocols, we release the outputs of two systems and the judges' quality ratings, all at [redacted].

Overall, our hope is that these findings and resources will inform linguistic inquiry and the design of S2ST models, loss functions, and training data. Over the long term, this should enable the development of systems able to better translate pragmatic intent, and thereby better support speakers who need to go beyond just conveying information, to explain, convince, teach, entertain, and do all the things that are so easy in one's own language, but currently so difficult across language barriers.

Acknowledgments: We thank the National Science Foundation for support under IIS-2348085, "Modeling Prosody for Speech-to-Speech Translation, " and R. Emiliano Puchaicela for help with the thematic analysis.

Use of Generative AI Disclosure: None

12. References

[1] D. J. Liebling, M. Lahav, A. Evans, A. Donsbach, J. Holbrook, B. Smus, and L. Boran, "Unmet needs and opportunities for mobile translation AI," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2020, paper 134.

[2] A. Popescu-Belis, "Dimensionality of dialogue act tagsets: An empirical analysis of large corpora," *Language Resources and Evaluation*, vol. 42, pp. 99–107, 2008.

[3] E. Couper-Kuhlen and M. Selting, *Interactional Linguistics*. Cambridge University Press, 2018.

[4] M. Weisser, "Speech acts in corpus pragmatics: Making the case for an extended taxonomy," *International Journal of Corpus Linguistics*, vol. 25, no. 4, pp. 400–425, 2020.

[5] H. Bunt and V. Petukhova, "Semantic and pragmatic precision in conversational AI systems," *Frontiers in Artificial Intelligence*, vol. 6, p. 896729, 2023.

[6] anonymous, "SpeechKit: Open-source infrastructure for speech evaluation," in *Interspeech*, 2026, under review.

[7] M. Gupta, M. Dutta, and C. K. Maurya, "Direct speech-to-speech neural machine translation: A survey," *Speech Communication*, vol. 175, 2025, paper id: 103317.

[8] S. Agrawal, A. Anastasopoulos, L. Bentivogli, O. Bojar, C. Borg, M. Carpuat, R. Cattoni, M. Cettolo, M. Chen, W. Chen *et al.*, "Findings of the IWSLT 2023 evaluation campaign," in *Proceedings of the 20th International Conference on Spoken Language Translation*, 2023, pp. 1–61.

[9] M. Chen, P.-A. Duquenne, P. Andrews, J. Kao, A. Mourachko, H. Schwenk, and M. R. Costa-jussà, "Blaser: A text-free speech-to-speech translation evaluation metric," in *ACL*, 2023.

[10] W.-C. Huang, B. Peloquin, J. Kao, C. Wang, H. Gong, E. Salesky, Y. Adi, A. Lee, and P.-J. Chen, "A holistic cascade system, benchmark, and human evaluation protocol for expressive speech-to-speech translation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[11] J. Iranzo-Sánchez, J. Iranzo-Sánchez, A. Giménez, and J. Civera, "Going beyond your expectations in latency metrics for simultaneous speech translation," in *Findings of the Association for Computational Linguistics: ACL*, 2025, pp. 18 205–18 228.

[12] K. Deng, W. Chen, X. Chen, and P. Woodland, "SimulS2S-LLM: Unlocking simultaneous inference of speech LLMs for speech-to-speech translation," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 16 718–16 734.

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

[14] R. Zhou, A. Ito, and T. Nose, "Preserving speaker information in direct speech-to-speech translation with non-autoregressive generation and pre-training," *Computer Speech & Language*, 2025, paper id: 101902.

[15] 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," *Advances in Neural Information Processing Systems*, vol. 37, pp. 89 682–89 705, 2024.

[16] K. Song, Y. Ren, Y. Lei, C. Wang, K. Wei, L. Xie, X. Yin, and Z. Ma, "StyleS2ST: Zero-shot style transfer for direct speech-to-speech translation," in *Interspeech*, 2023, pp. 42–46.

[17] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, M. Duppenthaler, P.-A. Duquenne, B. Ellis, H. Elsahar, J. Haaheim *et al.*, "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint arXiv:2312.05187*, 2023.

[18] T. Chen, Q. Wang, R. He, and S. Gollakota, "Spatial speech translation: Translating across space with binaural hearables," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025.

[19] Q. T. Do, S. Sakti, and S. Nakamura, "Sequence-to-sequence models for emphasis speech translation," *IEEE/ACM Transactions On Audio, Speech, And Language Processing*, vol. 26, no. 10, pp. 1873–1883, 2018.

[20] M. de Seyssel, A. D'Avirro, A. Williams, and E. Dupoux, "Emphassess: A prosodic benchmark on assessing emphasis transfer in speech-to-speech models," in *Empirical Methods in Natural Language Processing*, 2024, pp. 495–507.

[21] X. Chen, Y. Song, and S. Nakamura, "StressTransfer: Stress-aware speech-to-speech translation with emphasis preservation," 2025, arXiv 2510.13194.

[22] W.-C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS challenge 2022," in *Interspeech*, 2022.

[23] G. Bailly, E. André, E. Cooper, B. Cowan, J. Edlund, N. Harte, S. King, E. Klabbers, S. Le Maguer, Z. Malisz *et al.*, "Hot topics in speech synthesis evaluation," in *Speech Synthesis Workshop*, 2025, pp. 1–7.

[24] N. Moghe, A. Fazla, C. Amrhein, T. Kocmi, M. Steedman, A. Birch, R. Sennrich, and L. Guillou, "Machine translation meta evaluation through translation accuracy challenge sets," *Computational Linguistics*, vol. 51, no. 1, pp. 73–137, 2025.

[25] N. G. Ward, J. E. Avila, E. Rivas, and D. Marco, "Dialogs re-enacted across languages, version 2," University of Texas at El Paso, Department of Computer Science, Tech. Rep. UTEP-CS-23-27, 2023.

[26] E. Nachmani, A. Levkovitch, Y. Ding, C. Asawaroengchai, H. Zen, and M. T. Ramanovich, "Translatotron 3: Speech to speech translation with monolingual data," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 686–10 690.

[27] Z. Zheng, X. Sun, T. Dinh, A. Yanamandra, A. Jain, Z. Liu, S. Hadap, V. Bhat, M. Aggarwal, G. Medioni, and D. Harwath, "RosettaSpeech: Zero-shot speech-to-speech translation from monolingual data," 2025, arXiv 2511.20974.

[28] N. G. Ward, J. E. Avila, E. Rivas, and D. Marco, *Dialogs Reenacted Across Languages*. Linguistic Data Consortium, 2024, lDC Catalog No. LDC2024S08.

[29] N. G. Ward, A. Segura, A. Ceballos, and D. Marco, "Towards a general-purpose model of perceived pragmatic similarity," in *Interspeech*, 2024.

[30] W.-C. Huang, B. Peloquin, J. Kao, C. Wang, H. Gong, E. Salesky, Y. Adi, A. Lee, and P.-J. Chen, "A Holistic Cascade System, Benchmark, and Human Evaluation Protocol for Expressive Speech-to-Speech Translation," in *ICASSP*, 2023.

[31] N. G. Ward, D. Marco, and O. Fuentes, "Which prosodic features matter most for pragmatics?" in *IEEE ICASSP*, 2025.

[32] N. G. Ward and D. Marco, "A collection of pragmatic-similarity judgments over spoken dialog utterances," in *Linguistic Resources and Evaluation Conference (LREC-COLING)*, 2024, p. 154–163.

[33] J. Vazquez, "HuBert-based models and evaluation strategies for pragmatically-faithful speech to speech translation," Master's thesis, University of Texas at El Paso, 2025.

[34] J. E. Avila and N. G. Ward, "Towards cross-language prosody transfer for dialog," in *Interspeech*, 2023.

[35] Z. Xu, S. Escalera, A. Pavão, M. Richard, W.-W. Tu, Q. Yao, H. Zhao, and I. Guyon, "Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform," *Patterns*, vol. 3, no. 7, 2022, article 100543.

[36] Y. Song and S. Nakamura, "System description for the TOPI chappenge: A cascaded approach to pragmatics-aware speech translation," 2026, manuscript.

[37] D. Li, Y. Zhao, J. Wang, Q. Zhang, S. Li, and J. Cai, "CPMT-Marco: A cross-lingual pragmatic mapping transformer with enhanced Marco loss," 2026, manuscript.

[38] X. Luo, S. Jiang, S. Yang, D. Ke, Y. Xie, and J. Zhang, "R-APM: Retrieval-augmented pragmatic mapper for cross-lingual prosody transfer," 2026, manuscript.

[39] N. Ballier, L. Contreras Roa, M. Lelandais, B. Namdarzadeh, S. Patin, A. Saloev, C. Valdez, J.-B. Yunès, L. Zhu, and M. Zimina-Poirot, "The PAPTAN submission to the Interspeech 2026 challenge: The importance of dialectal variation," 2026, interspeech 2026 challenge paper submission.

[40] T. Labiausse, L. Mazaré, E. Grave, P. Pérez, A. Défossez, and N. Zeghidour, "High-fidelity simultaneous speech-to-speech translation," 2025, arXiv preprint arXiv:2502.03382.

[41] S. K. Ahmed, R. A. Mohammed, A. J. Nashwan, R. H. Ibrahim, A. Q. Abdalla, B. M. M. Ameen, and R. M. Khdhir, "Using thematic analysis in qualitative research," *Journal of Medicine, Surgery, and Public Health*, vol. 6, p. paper id: 100198, 2025.

[42] P. Hudelson and F. Chappuis, "Using voice-to-voice machine translation to overcome language barriers in clinical communication: an exploratory study," *Journal of General Internal Medicine*, vol. 39, no. 7, pp. 1095–1102, 2024.

[43] M. Sperber, M. de Seyssel, J. Bao, and M. Paulik, "Toward machine interpreting: Lessons from human interpreting studies," in *Empirical Methods in Natural Language Processing*. ACL, 2025, pp. 23 338–23 353.

[44] K. Lee and N. Lee, "Applying cultural-historical activity theory to understand Korean tourists' experiences with language translation applications," *Current Issues in Tourism*, pp. 1–18, 2025.