Speech-to-Text Translation: Cascaded or End-to-end? Multidimensional Comparative Evaluation

Lilit Kharatyan[1], Frédéric Blain[2], Gloria Corpas Pastor[3]

[1]Julius Maximilian University of Würzburg, [2]Tilburg University, [3]University of Malaga

Objectives

- Evaluation of the performance of cascaded and end-to-end speech translation models
- Assessment of non-translation aspects of models
- Identify optimal metrics for the evaluation of ST systems
- Analyze the applicability of speech translation approaches across domains

Research Questions

- Do the variances in the systems' modelling have a major influence?
- Are there systematic differences in the translation outputs between cascaded and end-to-end models?
- Are the currently available automatic metrics applicable and effective for the evaluation of speech translation systems?
- How adaptable are speech translation systems in terms of their application across various domains and settings?

Multidimensional Evaluation Framework

- Latency
- Quality of Translation
- Architectural Differences
- Explainability and Interpretability
- Skillset, Accessibility and Ease of Use
- Resource Consumption and Sustainability

Selected Models

Cascade

ASR OpenAI/ Whisper-medium	MT Facebook /wmt19-en-ru		
ASR Facebook/Wav2VeC base-960h	MT Facebook /wmt19-en-ru		
ASR Facebook/Wav2VeC base-960h	Punctuator FullStop/punctuation multilang-large	MT Facebook /wmt19-en-ru	
ASR Facebook/Wav2VeC base-960h	Grammar Corrector Google/flan-t5-large-grammar-synthesis	Punctuator FullStop/punctuation multilang-large	MT Facebook /wmt19-en-ru

End-to-End

- Facebook/S2T-Multilingual
- Facebook/xm transformer-W2V2-Transformer

Explainability and Interpretability Issues

- Translation/ Transcription Errors
- Gender Detection/ Bias
- Feature (laughter, applause, caughing) Capturing

Skillset, Accessibility and Ease of Use

- Data Scarcity
- Technical Knowledge and Skillset
- Flexibility, Versatility and Customizability

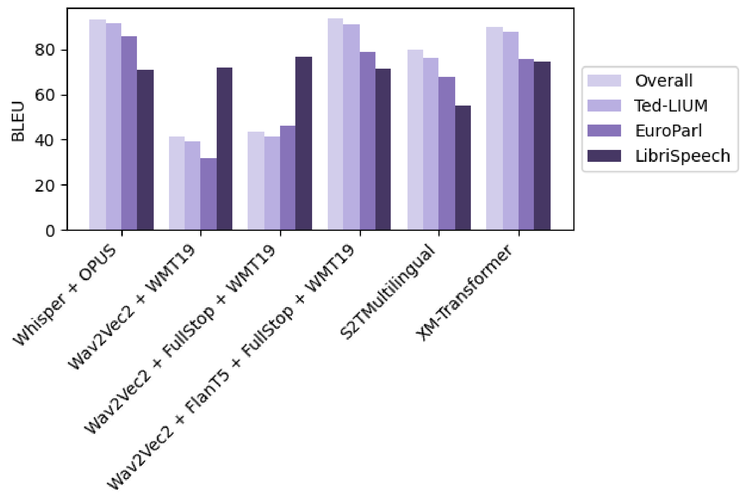
Results

Quality of Translation

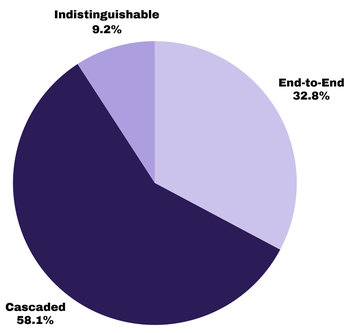
Automatic Evaluation Results for Cascaded and End-to-End Speech-to-Text Translation Models across All Datasets

№	Model Name	BLEU	BERTScore Precision / Recall / F1 Score
Cascaded Models			
1	Whisper + OPUS	92.9	**0.898** / **0.886** / **0.892**
2	Wav2Vec2 + WMT19	41.5	0.806 / 0.799 / 0.802
3	Wav2Vec2 + FullStop + WMT19	43.3	0.842 / 0.828 / 0.834
4	Wav2Vec2 + FlanT5 + FullStop + WMT19	**93.5**	0.856 / 0.850 / 0.853
End-to-end Models			
5	S2TMultilingual	80.0	0.833 / 0.828 / 0.830
6	XM-Transformer	**90.0**	**0.892** / **0.889** / **0.890**

Automatic Evaluation Results of Models across Different Genre and Datasets

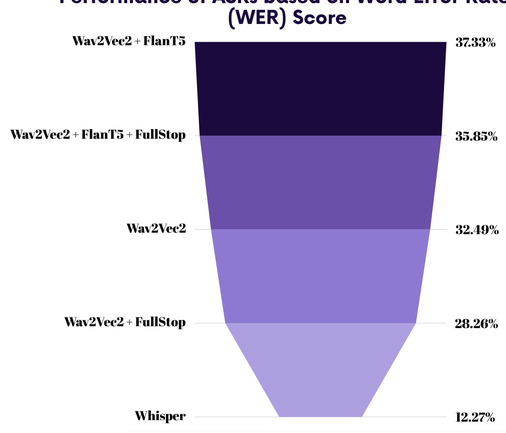


Human Evaluation Results for Cascaded and End-to-End Speech-to-Text Translation Models across All Datasets

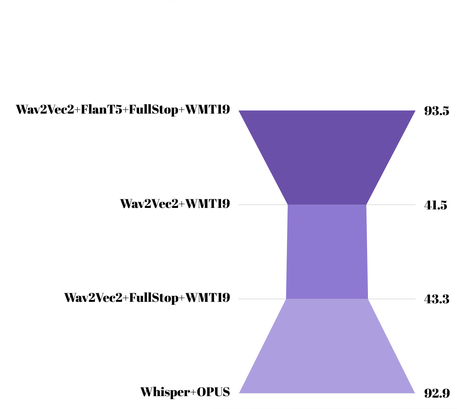


Quality of Transcription

Performance of ASRs based on Word Error Rate (WER) Score

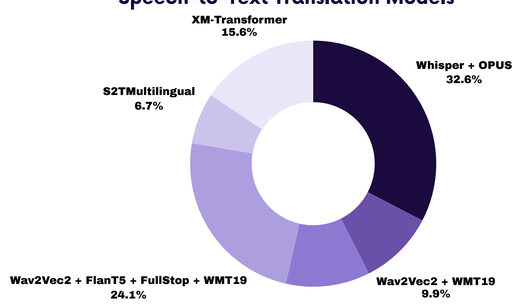


Automatic Evaluation Results for ASR and MT Components in Cascaded Models

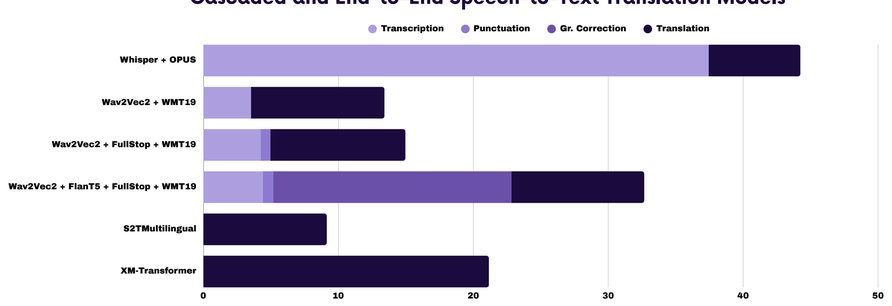


Latency Analysis

Average Latency Information of Cascaded and End-to-End Speech-to-Text Translation Models

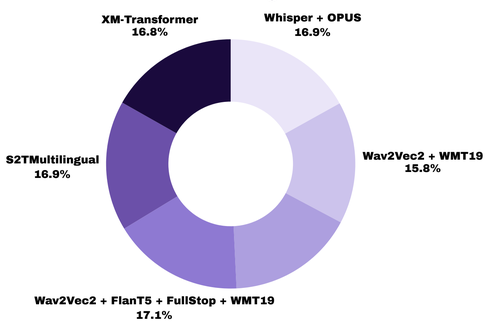


Detailed Latency Information of Separate Components of Cascaded and End-to-End Speech-to-Text Translation Models

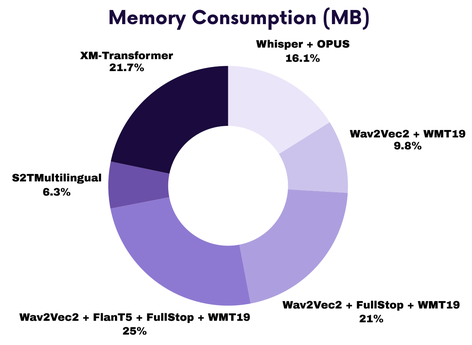


Resource Consumption and Sustainability

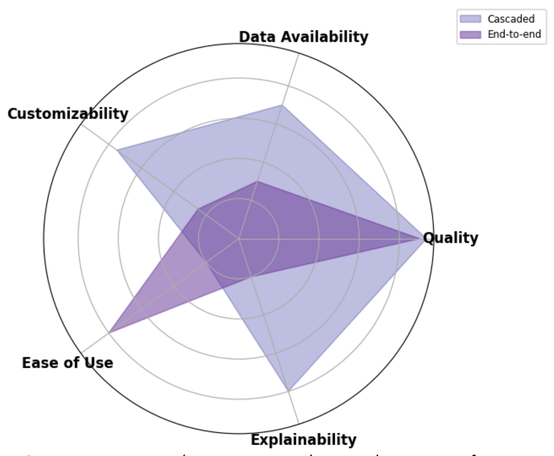
CPU Usage



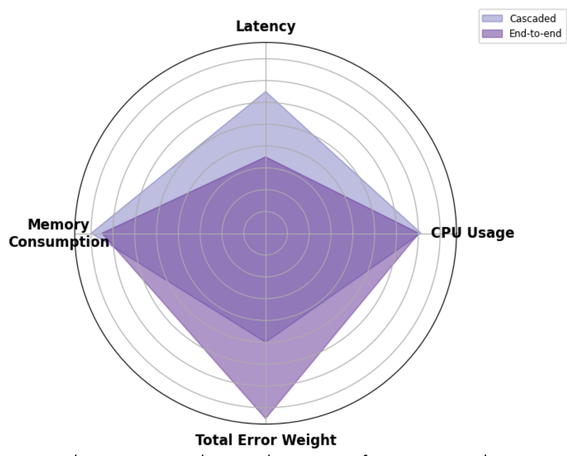
Memory Consumption (MB)



Conclusions



Higher percentage/ coverage indicates better performance in these categories.



Lower percentage/ coverage indicates better performance in these categories.

Supported by: Tilburg University, EM TTI, Universidad de Málaga