AI-powered Speech-to-Speech Translation (En/Ar): A Comparative Study Between SeamlessM4T V2 and Microsoft Translator

Ayah Yussuf Teama[1], Ain-Shams University, Cairo, Egypt (aya.yIbrahim@alsun.asu.edu.eg)
Nagwa Younis[2], Ain-Shams University & Arab Open University, Cairo, Egypt
(nagwayounis@edu.asu.edu.eg)

Abstract

The advancement of artificial intelligence (AI) has transformed the field of interpreting, especially with AI-powered Speech-to-Text and Speech-to-Speech translation tools gaining prominence. Limited AI-powered tools provide such services for free, among which are SeamlessM4T V2 by Meta AI (demo version) and Microsoft Translator by Azure AI. This research relies on the Multidimensional Quality Metrics (MQM) framework to compare and contrast both AI tools in terms of accuracy and fluency when translating from English into Arabic. Through a qualitative analysis conducted on the sentence level, we aim to provide insights and early observations on the strengths and weaknesses of such tools, whether to recommend or discourage their use by users, and support developers with guidance to enhance the quality of interpreting. Automatic Speech Recognition (ASR) quality findings show that SeamlessM4T V2 achieved a Word Error Rate (WER) of 2.72%, while Microsoft Translator achieved a WER of 3.4%. As for translation accuracy and fluency, both systems struggled with idioms, culture-specific references, sarcasm, and proper nouns—frequently producing literal or inconsistent renderings, while SeamlessM4T V2 showed slightly superior grammatical cohesion. In speech synthesis, SeamlessM4T V2 was generally preferred for intonation and pacing, while Microsoft Translator was perceived as more robotic, especially when rendering numerals. Humor and sarcasm were challenges for both systems. The study highlights the need for developers to improve AI systems' ability to achieve pragmatic equivalence, thereby enhancing the contextual appropriateness, cultural sensitivity, and human-likeness of AI-generated speech.

Keywords: Artificial Intelligence, Speech-to-speech Translation, Automatic Speech Recognition, Text-to-speech Synthesis, AI Translation

1. Introduction

Speech-to-speech translation (S2ST), a technology that interprets speech from one language into speech in another, bridges communication barriers between people speaking different languages. AI has contributed to the enhancement of cross-linguistic communication, especially through real-time speech translation. As the world has become more connected, the demand for tools that can quickly and accurately render spoken language increases. Two leading systems in this domain include SeamlessM4T V2 powered by Meta AI and Microsoft Translator by Azure AI. Despite their technological advancements, their efficiency (as AI systems) in substituting human interpreters remains uncertain, particularly in settings requiring precision and nuance, such as political discourse. Continuous expert assessment is needed to determine when and how to replace human interpreters with AI, if possible. This study undertakes a comparative analysis of SeamlessM4T V2 by Meta AI (demo version) and Microsoft Translator by Azure AI to examine a segment of the 2010 State of the Union Address. The analysis entails three critical stages: converting speech to text (Automatic Speech Recognition), translating the text using AI, and turning that translation back into speech

(Text-to-speech synthesis). Performance is evaluated based on Word Error Rate (WER) in stage one, Multidimensional Quality Metrics (MQM) in stage two; a comprehensive framework developed by Lommel (2018) to assess machine translation quality across multiple dimensions, including accuracy and fluency, and Mean Opinion Score (MOS) (Streijl et al., 2016) in stage three. Investigating eight extracted sentences, the study aims to evaluate the capability of AI to conduct autonomous interpreting or the need for human interpreters.

Microsoft Translator utilizes a cascade approach in its speech-to-text (S2T) and speech-to-speech (S2ST) translation subsystems, which notably distinguishes it from SeamlessM4T V2, which uses an end-to-end direct approach. The cascade approach involves training separate models for each step: a) Automatic Speech Recognition (ASR) transcribes the spoken words, b) Machine Translation (MT) translates the texts into a different language, and c) Text-to-Speech (TTS) synthesis generates a speech in the target language. Accordingly, Microsoft Translator's cascaded structure gives transparency of each step with the ability to improve it independently.

On the other hand, SeamlessM4T V2 by Meta AI functions as an end-to-end multilingual system, performing both speech-to-text and text-to-speech translation. The S2ST system limits the number of steps; thus, latency and error accumulation from one stage to another are reduced. This architecture aims to eliminate such issues while putting more emphasis on tone and emotion by directly performing a target speech from the source speech without converting it first into a translatable text explicitly. Yet, it is harder to train and control as it is less flexible.

2. Research Aim

The purpose of the study is evaluating the effectiveness of two free speech-to-speech AI translation systems (English into Arabic) on selected political discourse, which are SeamlessM4T V2 by Meta AI and Microsoft Translator by Azure AI. The major objectives are the evaluation of their output regarding both accuracy and fluency of translations and naturalness of output, followed by the weaknesses and areas of strength when dealing with named entities, abbreviations and acronyms, customized terms with different meanings in different contexts, culture-specific lingo, idiomatic expressions, humor, and sarcasm. The potential benefit of carrying out such an analysis is to gain an early indication of how they might be used in real-world situations of interpreting, if and when they can replace human interpreters, and to advise developers on how to enhance these tools' performance through qualitative feedback based on sentence-to-sentence comparisons.

Research questions are as follows:
1. Is there a statistically significant difference in the Word Error Rate (WER) between the two systems?
2. What types of translation errors are most frequent in each system according to the MQM framework?
3. Which of the two systems achieves a higher rating in terms of speech naturalness (MOS)?

3. Literature Review

Before reviewing the literature, there is an essential distinction we need to highlight, i.e. speech-to-speech translation vs. human interpreting and why the term 'interpreting' was kept to distinguish human activity. Although both processes are remarkably similar in terms of purpose; oral input, interpreting message, and oral output, humans are completely different from the machine they trained. Language transfer that includes decoding and encoding linguistic aspects as well as searching for translation unit equivalents is known as automated speech translation, or AST. The primary goal of this linear process is to match data in datasets that humans feed into language models in machine translation (MT) systems as training material. Accordingly, the aim of S2S translation is neither communication nor is it interactive as in human interpreting, as Horváth (2021) argued "AST functions with text and not speech" (p. 179).

Horváth (2021) compared both in terms of function, role, process, communication, number of languages, language use, speech, memory, vocabulary, method, information processing, knowledge acquisition, professional awareness, and soft skills. He concluded all results in the following table (Horváth, 2021, p. 178).

	Automatic speech translation	**Interpreting**
Function	artificial language mediation decoding and encoding linguistic elements	natural language mediation, facilitating and supporting the communication process, providing a service
Role	conduit, channel	mediator
Process	no feedback from user linear, unidirectional	feedback from user multidirectional, interactive
Communication	speech recognitions speech synthesis	looking for meaning and sense on speech level situation embedded constructive
Number of Languages	limited depending on database	limited only by the number of existing languages
Language use	_	conscious and intentional for supporting the communication intent
Speech	synthetized speech	Human speech
Memory	limited only by the size of training datasets	limited
Vocabulary	limited only by the size of training datasets (the Hungarian BERT- large language model's corpus contains 3,67 billion words)	an average educated person's monolingual vocabulary contains 30,000 words (Levelt 1989)
Method	Unit matching	creating new TL form context-driven interpretation of meaning
Information processing	only verbal and not information but matching data	multimodal (verbal, visual)
Knowledge acquisition	_	before, during and after the interpreted event
Professional awareness	_	✓
Soft skills	_	✓

(Table 3.1 – Horváth's comparison, 2021, p. 178)

This comparison shaded lights on numerous differences which, to a great extent, answers why we compare an AI tool to an AI tool and do not insert a human interpretation in our experiment believing that the human element who once programmed an AI can never be replaced by an AI, especially in a profession like simultaneous interpretation.

Previous studies concerned with assessment, to the best of our knowledge, can be divided into two completely different dimensions. The first is computational-linguistic related and the second is end-user concerned. The first focuses on how to develop and well-train AI to get better results and responses. Examples of such are various.

Huang et al (2023) worked on improving speech-to-speech translation in terms of dealing with two challenges: acoustic multimodality and high latency. They proposed a new model of speech-to-speech translation, called TranSpeech, with bilateral perturbation. Bilateral Perturbation (BiP), which consists of the style normalization and information enhancement stages, was their solution to face acoustic multimodality. With reduced acoustic multimodal challenges, they established a non-autoregressive S2ST technique that predicts unit selections, achieving high accuracy in just a few iterations with an improvement of 2.9 BLEU. Also, parallel decoding minimized latency noticeably.

Gupta et al (2024) argued that researchers have recently developed direct S2ST models, which have improved decoding latency, may preserve paralinguistic and non-linguistic properties, and may be able to translate speech without the need for intermediate text generation. Nevertheless, direct S2ST still performs poorly compared to cascade models, particularly when it comes to real-world translation, and has not yet attained high-quality performance for smooth communication. Their survey compared direct S2ST models in terms of data, application, and performance, and concluded that there are many challenges to work on. cascade vs. end-to-end S2ST models, S2ST on code-mix data, discrepancy between automatic and human evaluation, multiple speakers and noise handling, multilingual and simultaneous S2ST, low-resource S2ST datasets and models, voice cloning, and faster token generation are among these challenges that researchers still need to explore.

Lin et al (2024) improved Speech Emotion Recognition (SER) via Speech-to-Speech translation. Speech Emotion Recognition (SER) is essential to better human-computer interaction by allowing computers to comprehend and react to users' emotions. However, they claimed that many languages have limited datasets, frequently with less than 10 hours of speech, which impairs SER performance, whereas high-resource languages like English, Chinese, and Russian have substantial SER datasets. They employed an SER model that can be generalized to many languages of limited resources and across different upstream models by making use of extensive datasets in the target language. To produce labeled data in the target language, they specifically used expressive Speech-to-Speech translation (S2ST) in conjunction with a unique bootstrapping data selection method.

From an end-user perspective, same as this paper, related works took a totally different direction in research and evaluation. Hashimoto et al (2011) focused on speech synthesis in speech-to-speech translation that also contains speech recognition and machine translation steps prior to speech synthesizing. They claim that Numerous methods for combining machine translation and speech recognition have been put forth. Speech synthesis has not been taken into consideration yet, although, in their paper, they proved that the fluency of the translated sentences has a significant impact on the synthesized speech's naturalness and intelligibility.

Dhawan (2022) explored Speech-to-Speech translation challenges. He stated that the latency of automatic speech-to-speech translation is a major issue. Tasks involving spoken language processing are often approached at the sentence level not utterance level. As the procedure begins after the end of a sentence is observed, speech-to-speech translation suffers from a lengthy delay that is proportionate to the input length. That is not helpful for lengthy monologues like lecture talks and is comparable to consecutive interpretation, not simultaneous.

Fitria (2023) compared Google Translate, DeepL Translator, and Microsoft Translator in Indonesian English translation. She used Grammarly Premium to focus on grammar and spelling mistakes. Her descriptive qualitative analysis showed that Google Translate had 25 writing issues, DeepL Translator had 10 writing issues, and Microsoft Translator had 26 writing issues. She concluded that no Machine Translation can provide the same quality of human translation as these writing issues affected the clarity, correctness, and delivery of the text.

Karosekali et al (2024) compared the used translation techniques of humans and Microsoft Translator (Windows 11) when translating folklore from Indonesian into English. They found that human translators used 11 translation techniques while Microsoft Translator used 8 translation techniques. Human translators used adaptation, amplification, description, established equivalent, generalization, linguistic amplification, linguistic compression, literal, modulation, reduction, and transposition. Microsoft Translator used adaptation, borrowing, Calque, description, established equivalent, literal, modulation, and reduction.

Alkodimi et al (2024) compared the translation of literary texts between Arabic and English and their back translation with and without the aid of AI. They divided 80 English-major undergraduate students into 4 groups: two control and two experimental. The results showed that compared to students employing conventional techniques, students utilizing AI technologies were able to generate superior translations and back translations. This supported the concept of human-AI collaboration in the translation industry.

Hạnh (2024) interviewed a hundred sophomore students majoring in translation about AI because they are potential translators to come. Most students saw AI as an assisting tool to enhance the quality and speed of translation. AI is exceptional at doing routine repetitive jobs quickly and consistently, but it is insensitive to crucial elements like cultural variations and idioms that a skilled translator can handle with ease. Additionally, the development of human translators should focus on AI-compatible skills like foresight, critical and creative thinking, and interpersonal communication.

Doshi (2024) discussed the loss in AI translation when translating Jain scriptures from Hindi to English to preserve Jainism, an ancient Indian religion. He examined the advantages and disadvantages of the available AI models, emphasizing problems such as "semantic distortion, loss of context, cultural misinterpretation, and linguistic errors." He also highlighted the significance of protecting spiritual and cultural heritage, as well as the ethical ramifications of AI translation errors. After comparing Transformer and Seq2Seq models, he concluded that the parallel processing, long-range dependency handling, and effective domain-specific knowledge integration of the Transformer model made it better than Seq2Seq model in this task.

4. Research Methods

4.1. Theoretical Framework

This paper is built on a Three-dimensional integrated framework. The comparative study between SeamlessM4T V2 and Microsoft Translator is investigated against an eclectic approach; Word Error Rate (WER), Multidimensional Quality Metrics (MQM), and Mean Opinion Score (MOS). Further explanation is provided in the supplementary online resource at (https://docs.google.com/document/d/1HmiMnkz05SQzEt1koooccfFE9MXr1nJL/edit0?usp=drive_link&ouid=116265779930492647496&rtpof=true&sd=true)

4.2. Data Selection

Eight speech segments were extracted from former U.S. President Barack Obama's 2010 State of the Union Address, each with a maximum duration of 15 seconds. The choice of this source material is based on specific considerations. First, authenticity and relevance: it is a natural political speech in an official international setting, and AI systems can be evaluated with conditions resembling the practice of professional human interpreters. Second, linguistic diversity: there are general statements, numerals, proper nouns, abbreviations, acronyms, customized terms, idiomatic expressions, culture-specific references, humor, and sarcasm; thus, the speech will reflect a realistic range of potential linguistic challenges. Third, controlled variables: using a single speech by the same speaker ensured consistency in terms of accent, speech rate, level of formality, tone, and thematic focus, which particularly focused the experiment on the differences in the systems rather than the variation of inputs. Fourth, ease of segmentation: the audio could be readily divided at natural pauses and transitions marked by applause. The official transcript of the speech was obtained from the National Archives (White House) to serve as the reference text for ASR evaluation.

4.3 Pre-processing

The original video material was downloaded from YouTube and converted into WAV audio format (44.1 kHz, 16-bit). The selected speech segments were manually segmented with an open-source audio editor capable of segmenting utterances.

4.4 System Settings

Both speech-to-speech translation systems were tested. A) Meta SeamlessM4T V2 (Demo): this system was accessed through the Hugging Face web interface; its source language was set to English (en-US), and the target language was set to Modern Standard Arabic (ar). Automatic punctuation was enabled. Text-to-speech (TTS) output used the default voice settings. B) Azure Microsoft Translator: this system was accessed through the Microsoft Translator mobile application; its source language was set to English (United States), and the target language was set to Modern Standard Arabic. TTS output used the default neural Arabic voice. For comparative purposes, both systems were used on the same pre-processed WAV files, and no manual transcription or text normalization was applied.

4.5 Evaluation Procedures

4.5.1 Automatic Speech Recognition (ASR) Evaluation

Using the official transcript, the ASR output from each system was assessed using the Word Error Rate (WER) metric. The formula used was: WER = (S + D + I) / N ×

100, where S = substitutions, D = deletions, I = insertions, and N = total number of words in the reference transcript.

4.5.2 Machine Translation (MT) Evaluation using MQM

Translations were evaluated using the Multidimensional Quality Metrics (MQM) framework (Lommel, 2018), focusing on accuracy and fluency. Evaluating each translation segment was done through marking errors by type and severity (Critical, Major, Minor, Null). Penalty points that are used to score translations are associated with each severity level. 100 points (critical), 10 points (major), 1 point (minor), and 0 points (null) are the default penalties. The final quality or fluency score is determined using the formula below:

Score = 1 - (Total Penalties / Word Count).

4.5.3 Text-to-Speech (TTS) Evaluation using MOS

The Arabic TTS outputs were evaluated for naturalness using the Mean Opinion Score (MOS) method (Streijl et al., 2016). Sixteen expert raters with backgrounds in Arabic linguistics and interpretation participated. Audio samples were presented in a controlled acoustic environment. Raters scored each clip on a 5-point Likert scale (1 = Poor, 5 = Excellent). MOS scores were averaged, and experts' comments were analyzed.

4.6 Data Availability

All ASR outputs, translation files, and TTS audio samples are stored in an open-access repository and can be accessed via this URL: (https://drive.google.com/drive/folders/1xv1EBtlgTeiTXYiXNsgirk7ipjj3J2Sn?usp=sharing).

5. Results

5.1. ASR Results

As for the quality of both ASR systems, both ASR systems performed with high accuracy, but SeamlessM4T V2 slightly outperformed Microsoft Translator with a lower WER of 2.72% compared to 3.4%.

5.2. Translation Quality Findings

Regarding the fluency and accuracy of AI translation offered by both tools, eight observations have emerged from the qualitative analysis conducted. In general sentences, both AI-powered systems produced generally accurate and fluent translations. While SeamlessM4T V2 showed a major grammatical fluency error, Microsoft Translator had minor and null-level issues related to accuracy and punctuation that did not significantly affect meaning or speech quality. Secondly, both tools managed to recognize and translate the named entity correctly. However, SeamlessM4T V2 had minor errors affecting accuracy and fluency, while Microsoft Translator's only issue was a null-level whitespace error with no impact on the spoken output. Thirdly, although both tools handled the acronym similarly by leaving it untranslated, the translation of the remaining parts of the sentences raised important issues; SeamlessM4T V2 made major errors in spelling and literal translation, while Microsoft Translator produced critically unintelligible numeral renderings, cohesion issues, and a severe mistranslation that significantly impacted both fluency and accuracy. Fourth, there were major accuracy and fluency errors as a result of both tools' inability to translate customized terms like "Wall Street" and "Main Street" in their context, which gave them customized meanings. Though slightly better, SeamlessM4T V2 produced an ambiguous passive construction that further impeded clarity. Microsoft Translator further distorts meaning due to a critical ASR mis-transcription. This further demonstrated how ASR accuracy affects the final output. Fifth, both AI tools were able to accurately translate culturally specific terms into their equivalents. However, Microsoft Translator only had a null-level punctuation error that did not affect the spoken output, and SeamlessM4T V2 had a slight fluency problem because of an odd collocation. Sixth, both AI systems made a critical accuracy error when rendering idioms literally, and SeamlessM4T V2 also made a minor fluency error in Arabic grammar. Seventhly, both AI tools failed to capture the humor and its tone intended in a sentence. SeamlessM4T V2 produced critical accuracy and fluency errors that rendered the output unintelligible. Microsoft Translator delivered a more acceptable, though still humorless, translation. Eighthly, both AI tools failed to convey the sarcastic meaning of the sentence. SeamlessM4T V2 committed several major and critical errors in accuracy and fluency that rendered the output confusing, while Microsoft Translator produced an even more problematic translation due to severe misrecognition of pronouns and sentence structure, resulting in critical errors that significantly distorted the intended message. This means that humor and sarcasm, if not comprehended by a machine, perplex it and make the context vague, resulting in unexpected errors.

5.3. TTS Evaluation

In terms of speech synthesis quality, the comparison showed that SeamlessM4T V2 offered a more natural, human-like, and coherent auditory experience compared to Microsoft Translator, which was praised for its clarity, intelligibility, and confident tone. Yet, Microsoft Translator was criticized for being robotic and monotonous. Nevertheless, both tools performed poorly in expressing humor and sarcasm and lacked the prosodic detail and contextual information required to produce emotive speech in communication.

6. Discussion

The results are based on a three-tiered analysis, each level of analysis is discussed in detail separately in the following sections. Although SeamlessM4T V2 is an end-to-end model, its demo version via the public demo interface unveils all steps for comprehensive comparison with Microsoft Translator, the cascade model.

6.1.1. Automatic Speech Recognition (ASR)

Because this paper focused on certain sentences with distinct natures, the data is limited. Yet, the accuracy of the ASR systems of SeamlessM4T V2 and Microsoft Translator can still be measured through WER metric.

	S1	S2	S3	S4	S5	S6	S7	S8
Seamless	0	1D	1S	0	1S	0	0	1S
Microsoft	1S	0	2D	1D	0	0	0	1S
Total reference words	22	11	33	20	20	18	9	14

(Table 6.1.1 – WER per sentence)

Not only was the Word Error Rate (WER) formula used for calculating wrongly transcribed words, but also punctuation marks as they add clarity and precision to the sentence and accordingly can affect the meaning. While human interpreters do not rely on punctuation, ASR punctuation materially affects downstream S2S quality as the input is not purely the speech, but the transcribed version of it based on ASR system's quality.

Based on the above data and the provided digital appendix, WER can be calculated through this formula to compare both AI tools. As for SeamlessM4T V2, its ASR suffered 1 deletion and 3 substitutions in all eight sentences subjected to analysis: WER= (0+1+3)/147 = 0.027 (2.72% error rate). As for Microsoft Translator, its ASR committed 3 deletions and 2 substitutions in total. Accordingly, its WER= (0+3+2)/147 = 0.034 (3.4% error rate). As a result, we can conclude that SeamlessM4T V2's ASR is slightly more accurate than Microsoft Translator's. As a fact, a WER below 5% is typically considered excellent for most ASR systems, thus both tools have distinguished scores which indicate extremely high accuracy and minimal errors.

6.1.2. AI Translation

An in-depth qualitative analysis is conducted at the sentence level to measure the effectiveness of AI translation of general sentences, named entities, abbreviations and acronyms, customized terms with different meanings in different contexts, culture-specific lingo, idiomatic expressions, humor, and sarcasm. For further explanation, please see the supplementary material available at (https://docs.google.com/document/d/1TDVJsLcu5gZr4G3qOAQZIpsZ3pJBJrZZ/edit?usp=drive_link&ouid=116265779930492647496&rtpof=true&sd=true)

The result of comparing AI translation quality generated by SeamlessM4T V2 and Microsoft Translator, according to Multidimensional Quality Metrics (MQM), is highlighted in the following table, yet no totals or final scores are calculated.

	Word Count	**Penalties**
Seamles	**24+12+37+20+18+12+9+10**	**0+1+10+30+0+100+110+111**
Seamles	**24+12+37+20+18+12+9+10**	**10+1+10+20+1+1+100+1**
Microso	**28+12+31+19+19+12+8+11**	**0+0+100+120+0+100+0+300**
Microso	**28+12+31+19+19+12+8+11**	**1+0+210+20+0+0+0+100**

(Table 6.1.2 –total final scores)

Calculating a total score for each tool to reach a final result for the quality of translation to determine which to use and recommend in the future is not complicated, as the MQM formula is clear; Score = 1 - (Total Penalties /Word Count). However, calculating a total score based on quite limited data, of 8 sentences only, would be inequitable. Comparing both tools and conducting a deep qualitative analysis for each is already conducted under each sub-section of 6.2 AI Translation Analysis. A thorough and analytical reading of the above comparisons while examining each sentence with close attention to detail may help the reader to gain insights on the quality of AI translation critically and attentively and may also offer recommendations for users based on strengths and observations for developers to enhance weaknesses. Accordingly, we choose not to build a final decision to prefer one tool to the other relying on extreme scores.

6.1.3. Text-to-speech (TTS) synthesis

16 expert respondents were consulted to compare between the two TTS systems used by SeamlessM4T V2 and Microsoft Translator using a Mean Opinion Score questionnaire designed by the authors. This survey focused primarily on naturalness, clarity, and overall user experience when hearing the target language.

SeamlessM4T V2 was noted for its "naturalness" and "human-like qualities". Experts highlighted that it has a superb tendency to pause appropriately thus improving the flow of speech. Avoiding some robotic characteristics attributed to Microsoft Translator, SeamlessM4T V2 managed to create a more engaging auditory experience. Also, clearer pronunciation of certain words was noticed if compared to Microsoft Translator's pronunciation. Although this questionnaire was not designed to evaluate the quality of the translation itself, some experts gave comments on translation that we cannot ignore. SeamlessM4T V2 was praised for being more accurate in translation and coherent in delivering messages. However, many flaws were detected such as speed issues, as the tool reads too fast, some inevitable robotic elements, inaccurate word stress and strange pronunciation of names, which occasionally fell short of expectations.

On the other hand, the intelligibility of Microsoft Translator was of note, due to its ability to comprehensibly deliver fragmented sentences with clearer sound quality, which contributed to a more confident tone. The tool was also praised for its intonation throughout the speech. Certain experts recognized the tool's ability to excel in the pronunciation of long vowels and specific words, such as "Wall Street" and "Tampa". Despite the above strengths, Experts frequently described its output as "robotic" and "unnatural, particularly in its handling of numbers where it often rendered them in a monotonous fashion" that detracted from the listening experience. An overall lack of expressiveness was recognized, making the speech sound mechanical rather than human-like. Additionally, syntactic errors due to a lack of diacritical marks were challenges facing the audience. Although this questionnaire was not for evaluating the quality of interpreting the source languages, yet respondents pointed out that the interpretation of the message was literal leading to awkward phrasing that was difficult to comprehend. Many experts expressed struggling to identify significant advantages over SeamlessM4T V2.

An identified limitation realized in both tools was an inaptitude of oral communication of humor and sarcasm. As a result, emphasized intonation, timing, and contextual emphasis are inadequate or not at all implemented in SeamlessM4T V2 and Microsoft Translator outputs. Some experts observed that the natural and formal intonation of both tools eliminated the warmth and pleasantness of the voice, which people use to convey such fine shades of emotion. For instance, it was challenging to distinguish between a sarcastic statement and a literal statement resulting in sarcastic statements being rendered as literal. Likewise, due to a lack of voice inflection and timing, what might be joking phrases resulted in monotonous and poorly paced ones. This inability to express humor and sarcasm made both tools' output less natural and less fulfilling whereas emotional expression significantly influences communication. To remedy this deficiency, substantial improvements in the prosodic modeling and contextual understanding of the tools would be needed.

7. Conclusion

The capabilities and constraints of AI-powered interpreting systems in the understudied language direction of English to Arabic are clarified by this work. We have assessed SeamlessM4T V2 and Microsoft Translator using the Word Error Rate (WER) to measure the ASR efficiency, the MQM methodology to reveal their translation correctness and fluency, and the Mean Opinion Score (MOS) to evaluate to what extent they can sound like humans, providing developers and users with useful advice. Although these findings contribute to greater understanding of AI-powered tools designed for interpreting or speech-to-speech translation, they also reinforce the need to enhance these tools' reliability and performance through further improvements in the technologies. Launching from our observations, future research can expand by incorporating larger datasets and experimenting in real-time interpreting settings.

Limitations

This study is considered a pilot study; therefore, its findings cannot be generalized on a wide scale. The fundamental limitation of this research is the length of speech segments that were used to experience both AI-powered tools because there is a time constraint of 15 seconds maximum at a time. This may not capture important observations of longer speech segments which are typical for real-life situations in interpreting. Further, the study focuses on translation from English into Arabic only, so the findings of the study cannot be generalized to the other language pairs. A further limitation is the use of only qualitative analysis at the level of individual sentences which may well overlook important contextual factors and accumulation of errors in extended discourse. In addition, the demo versions of SeamlessM4T V2 and used version Microsoft Translator at the time of experimenting which were applied to this research are now earlier released versions of these applications and surely have since then been updated to provide even better services. For this reason, the outcomes of this research might not capture the current state of each tool's abilities. Hence, there is a continuous need to undertake ongoing investigation and research on the latest versions.

Abbreviations

AI – Artificial intelligence
ASR – Automatic Speech Recognition
AST – Automated speech translation
MOS – Mean Opinion Score
MQM – Multidimensional Quality Metrics
MT – Machine Translation
SER – Speech Emotion Recognition
S2S – Speech-to-Speech
S2ST – Speech-to-Speech Translation
S2T - Speech-to-Text
TTS – Text-to-Speech

WER – Word Error Rate

References

Alkodimi, K. A., Alqahtani, O. A., & Al-Wasy, B. Q. (2024). Human-AI collaboration in translation and back translation of literary texts. *Mağallaẗ Al-dirāsāt Al-iğtimāʿiyyaẗ*, *30*(2), 173–192. https://doi.org/10.20428/jss.v30i2.2404

Dhawan, S. (2022). Speech To Speech Translation: Challenges and Future. *International Journal of Computer Applications Technology and Research*, *11*(03), 36–55. https://doi.org/10.7753/ijcatr1103.1001

Doshi, A. V. & Carroll Independent School District. (2024). The Loss in AI Translation. In *Computational Linguistics* [Journal-article]. https://www.researchgate.net/publication/383037053

Fitria, T. N. (2023). Performance of Google Translate, Microsoft Translator, and DeepL Translator: Error Analysis of Translation Result. *Al-Lisan*, *8*(2), 115–138. https://doi.org/10.30603/al.v8i2.3442

Gupta, M., Dutta, M., & Maurya, C. K. (2024, November 13). *Direct Speech-to-Speech Neural Machine Translation: A Survey*. arXiv.org. https://arxiv.org/abs/2411.14453

Hạnh Nguyễn, T. T. & Nguyen Thi Tuyet Hanh. (2024). The Influence of AI on Translation: A Transformative Change in the Language Industry. In *Int. J. Adv. Multidisc. Res. Stud.* (pp. 346–351) [Journal-article]. https://www.researchgate.net/publication/385907075

Hashimoto, K., Yamagishi, J., Byrne, W., King, S., & Tokuda, K. (2011). An analysis of machine translation and speech synthesis in speech-to-speech translation system. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2011*, 5108–5111. https://doi.org/10.1109/icassp.2011.5947506

Horváth, I. (2021). Speech translation vs. Interpreting. *Language Studies and Modern Humanities*, *3*(2), 174–187. https://doi.org/10.33910/2686-830x-2021-3-2-174-187

Huang, R., 1, Liu, J., 1, Liu, H., 1, Ren, Y., Zhang, L., He, J., Zhejiang University, & ByteDance. (2023). TRANSPEECH: SPEECH-TO-SPEECH TRANSLATION WITH BILATERAL PERTURBATION. In *ICLR 2023* [Conference-proceeding].

Hunt, M.J. (1990). *Figures of merit for assessing connected-word recognisers*. Speech Commun. 9(4), 329–336. https://doi.org/10.1016/0167-6393(90)90008-W

Karosekali, S. C. B., Sipayung, K. T., & Saragi, C. N. (2024). Comparison of Translation Techniques Between Human and Microsoft Translator (Windows 11) in Translating Folklores. *Visi Sosial Humaniora*, *5*(1), 1–9. https://doi.org/10.51622/vsh.v5i1.2284

Levelt, W. J. M. (1989) Speaking: From intention to articulation. Cambridge: MIT Press, 566 p. (In English)

Lewis, W. (2020) AI and interpreting. Will Lewis on automated speech translation: Scale, uses & edge cases. YouTube, 13 November. [Online]. Available at: https://www.youtube.com/watch?v=KihPeHh0wyo (accessed 23.11.2024). (In English)

Levenshtein, V. I. (1966). *Binary codes capable of correcting deletions, insertions, and reversals.* Soviet Physics Doklady, 10(8), 707–710.

Lin, H., Lin, Y., Chou, H., & Lee, H. (2024, September 17). *Improving Speech Emotion Recognition in Under-Resourced Languages via Speech-to-Speech Translation with Bootstrapping Data Selection*. arXiv.org. https://arxiv.org/abs/2409.10985

Lommel, A. (2018). Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies. In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation Quality Assessment: From Principles to Practice* (Vol. I, pp. 109-127). Springer International Publishing.

Nemeskey, D. M. (2020). *Natural language processing methods for language modeling* (PhD thesis). Eötvös Loránd University. Retrieved from https://hlt.bme.hu/en/resources/hubert

Remarks by the President in State of the Union Address. (2016, January 9). whitehouse.gov. https://obamawhitehouse.archives.gov/the-press-office/remarks-president-state-union-address

SeamlessM4T v2 - a Hugging Face Space by facebook. (n.d.). https://huggingface.co/spaces/facebook/seamless-m4t-v2-large

Streijl, Winkler, & Hands. (2016). *Mean Opinion Score (MOS) revisited: Methods and applications, limitations and alternatives*. Retrieved March 19, 2024, from https://stefan.winkler.site/Publications/mmsj2016.pdf

The Obama White House. (2010, January 28). *The 2010 State of the Union address* [Video]. YouTube. https://www.youtube.com/watch?v=L1PWQtCDaYY

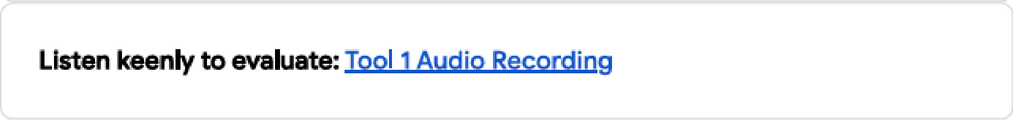
Young, S., Evermann, G., Gales, M., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., & Woodland, P. (2002). *The HTK book (for HTK version 3.4)*. Cambridge University Engineering Department. Retrieved from https://www.danielpovey.com/files/htkbook.pdf

Appendices

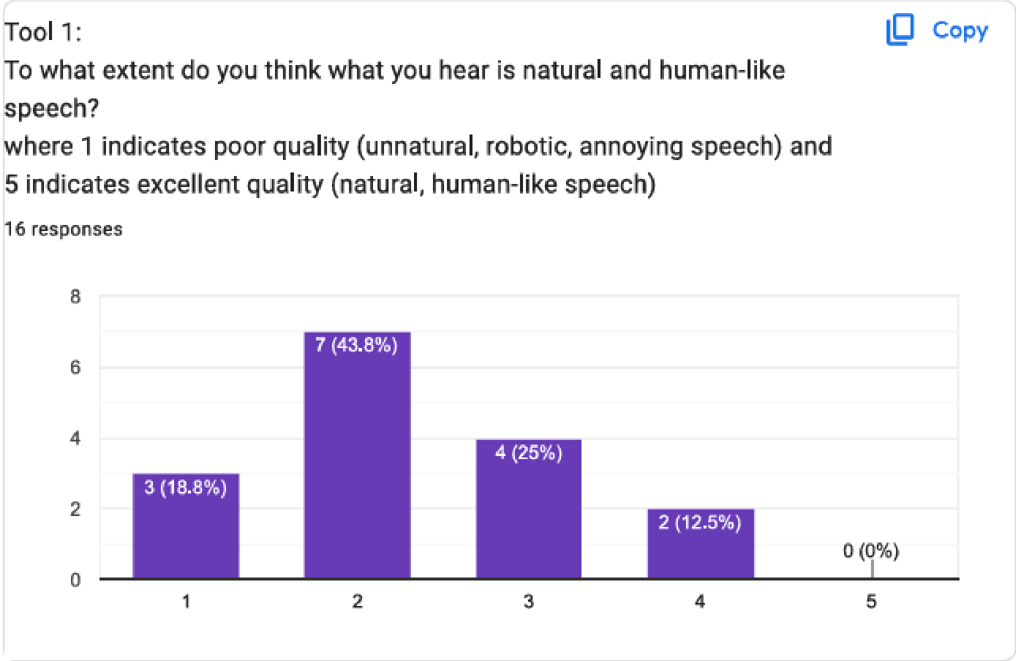
1. Mean Opinion Score (MOS) for Artificial Intelligence (AI) Text-to-Speech (TTS) Synthesis

For an extended details, refer to the online resource at: (https://drive.google.com/file/d/1zZsmdUCYOCPDCtK5NcsPNyvJbkz_P7gG/view?usp=drive_link)

For a summary of responses:
Tool 1: Microsoft Translator
Tool 2: SeamlessM4T V2



Listen keenly to evaluate: Tool 1 Audio Recording



Tool 1:
To what extent do you think what you hear is natural and human-like speech?
where 1 indicates poor quality (unnatural, robotic, annoying speech) and
5 indicates excellent quality (natural, human-like speech)

16 responses

Comparing Tool 1 to Tool 2, what are the strengths of Tool 1 in text-to-speech (TTS) synthesis?
15 responses

Clearer in sound quality, without echo

Tool 1 provides robotic performance. No areas of strength In comparison, none
Can't find strengths

The pronunciation of some words such as Wall Street, Tampa and is better. Tool 1 has a confident tone, resulting in clearer and quieter audio production.
More natural Formality and pace
The delivery of the fragmented sentences messages is quite intelligible.

It uses good intonation that rises and falls where it should.
It speaks at a rather good speed that is not too fast or too slow.

In the last sentence, it pronounces the words correctly using the correct diacritic marks, However, I am not sure it pronounced the name of the channel correctly.

I didn't feel like there were any strength points. To be honest, it felt very robotic and unnatural. It did deliver the message and was comprehensible; however, it sounds very machine-like, especially at rendering numbers. Additionally, it was too monotonous the whole time.

The sound is clearer; the pace is better Voice clarity
pronunciation of long vowels is better in 1

NA

Comparing Tool 1 to Tool 2, what are the weaknesses of Tool 1 in text-to-speech (TTS) synthesis?
16 responses

Less accurate, clearly machine-generated Literal interpretation; weird rendition of numbers

التراكيب اللغوية غير سلمية وغير مفهومة، الأرقام تنطق بطريقة غير صحيحة،، والكثير من الكلمات مبهمة غير واضحة المعنى
Incorrect and consequently unintelligible syntactic structures, weird articulated figures, vague lexis.

Problems in pronunciation and intonation

The voice is more robotic.

Tool 1 produced a more robotic tone, particularly when reading numbers. Nothing
Less natural with an extra air puff at the end of each segment

The tone and mechanic style seem robotic along with the poor translation of the text, which makes the text quite hard to appreciate its content.

The voice is more robotic in Tool 1 than in Tool 2.
It has an annoying way of reading out numbers especially with Zeroes.
Both tools committed the same mistake when dealing with metaphors esp. in the first sentence comparing Wall Street to the Main Street and in rendering “under my belt” literally. Tool 1 mispronounced . in sentence 2.

Unlike tool 2, tool 1 sounded very unnatural and monotonous, as if it was a machine, and it also articulated some Arabic words in a very odd way. One of the most disturbing things about tool 1 was the way it rendered numbers as well, where it said “1 point 8 point zero zero zero zero zero... etc” instead of trillion. The voice used also sounded like every typical “Siri” or “Alexa”.

Its pace is slow and sound Robotocs.some Arabic words are articulated inaccurately. The intonation is incorrect.

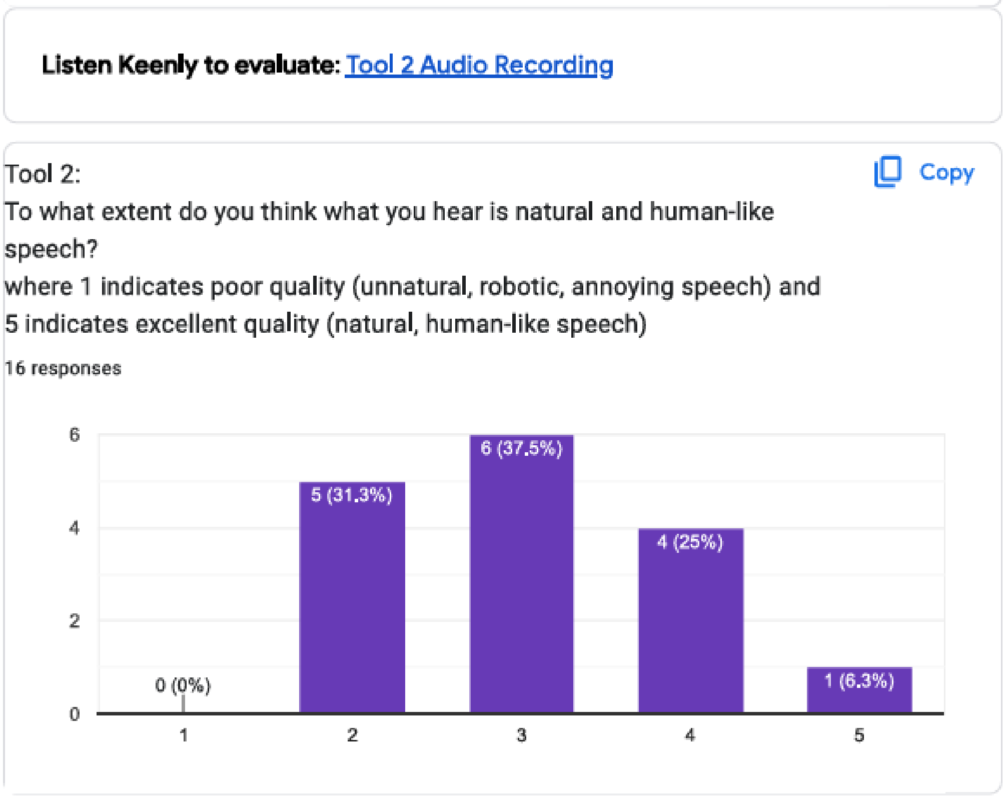
The pronunciation of some words was wrong Pauses
Syntactic errors due to lack of diacritical marks

Robotic Intonation

Pronunciation of numbers

Robotic (especially with numbers)

Tool 1: Microsoft Translator
Tool 2: SeamlessM4T V2



Listen Keenly to evaluate: Tool 2 Audio Recording

Tool 2:
To what extent do you think what you hear is natural and human-like speech?
where 1 indicates poor quality (unnatural, robotic, annoying speech) and
5 indicates excellent quality (natural, human-like speech)

16 responses

Comparing Tool 2 to Tool 1, what are the strengths of Tool 2 in text-to-speech (TTS) synthesis?
16 responses

More accurate, translation more coherent More natural and human- like speech
Higher degree of intelligibility at levels of syntactic structures and lexical items, specifically figures

I see it is less problematic in terms of pronunciation and .

The voice is clearer and not as robotic as the first one.

The tone of Tool 2 is livelier than that of Tool 1, producing more natural and human-like synthesis.

None

Sounds more natural

I see that the tone and choices of pauses seem to be quite human-like

Tool 2 uses a more human voice than tool 1. It reads numbers way better than tool 1. It pronounced . correctly.

It was not monotonous. Instead, it sounded very natural like a human who changes his/her voice and tone while talking.

Its pace sounds more natural and not mechanic. Better/correct pronunciation of few words Better pauses
Pronunciation is clearer on using tool 2 and more correctthe

Tone is more realistic and looks like a human voice

Comparing Tool 2 to Tool 1, what are the weaknesses of Tool 2 in text-to-speech (TTS) synthesis?
16 responses

Sounds more robotic

Similar to Tool 1, rendition is rather literal in some parts.

In comparison, none. But still both sound far from being natural. Same, but less frequent, problems
Some words get mispronounced. For example, was pronounced with t at the end even though there was no addition afterwards to pronounce it like that.

An echo is sometimes produced, particularly at the start of new sentences, which affects the audio quality.

Seems unnatural, the speech juncture Speedy and less accurate than Tool 1
Like tool 1, the text delivered by the second seems robotic, and the poor translation made it harder to appreciate the content of the text.

It is too fast to be well understood.
It makes more mistakes in diacritic marks than tool 1 such as ,la . The last sentence is totally lost.
Both tools committed the same mistake when dealing with metaphors esp. in the first sentence comparing Wall Street to the Main Street and in rendering "under my belt" literally.

I did't feel like there were any points of weakness. In fact, the voice itself used sounded like a human, not like a machine. It rendered the numbers naturally and uttered all Arabic words like a typical Arabic speaker. I believe it was perfect.

Few words are articulated inaccurately.
Some noise in the background; some parts are so fast, they seem unnatural Word stress
Names pronunciation

In both the annoying quality of the voice exists

Some echo