The Emergence of Machine Interpreting

Claudio Fantinuoli
University of Mainz

Machine interpreting (MI), also known as speech-to-speech translation, is an automated language translation process that converts spoken content from one language to another in the form of speech. Distinct from offline speech translation, which is typically employed for pre-recorded audio and video, MI's most notable feature is its immediacy: the interpretation is produced in real-time on the basis of a single presentation and intended for immediate consumption. The translation process can be either consecutive, interpreting content sentence by sentence, or simultaneous, which allows for continuous speech without interruption. MI aims at reducing language barriers and fostering global information exchange and unobstructed, fair communication.

MI is advancing rapidly, transitioning from research labs to practical applications, initially catering to recreational or casual purposes, such as for conveying information at hotel desks, and more recently, addressing professional scenarios such as live interpreting of lectures and events. The emergence of MI is fueled by the latest advancements in machine learning techniques applied to natural language processing, which have found significant adoption in industrial settings.

There are two primary approaches to machine interpreting. The end-to-end approach utilizes a single component to directly interpret input audio to output audio without generating intermediate text (Lee et al. 2022). Although this method is still experimental and not yet applied in real-life situations, it represents one end of the MI spectrum. On the other end is the cascading approach, which employs a flexible pipeline of components, typically involving speech recognition, machine translation, and voice synthesis. This is currently the prevalent method for tackling speech translation challenges (see e.g., Sperber & Paulik 2020).

Recent trends lie between these two extremes, aiming to merge some components of the cascading approach into single components. This reduces system complexity, enhances translation accuracy and naturalness, and prevents error propagation between components (Gaido et al. 2020). One such example is speech-to-text translation, which combines speech recognition and machine translation in a single language model that accepts speech as input and delivers translated text as output (Zhang et al. 2022)

Simultaneous MI is the most intricate form, as it necessitates interpreting an ongoing stream of speech incrementally, without interruptions or complete context knowledge (e.g., what the speaker will say moments later). To accomplish this, speech must be segmented into meaningful chunks in real-time. Segmentation methods range from detecting pauses in the speaker's flow and employing fixed word lengths, to utilizing dynamic approaches based on real-time syntactic and semantic analysis of incoming speech.

As MI systems have only recently emerged, research on user-centered evaluation methodologies is still in its early stages. Initial attempts have utilized written translations or human interpretations as the gold standard (Fantinuoli & Prandi 2021). Depending on the system, results have demonstrated high accuracy. However, the systems exhibit limited flexibility, performing well in specific scenarios like formal presentations but experiencing a rapid decline in quality in other situations, particularly when the spoken content is disfluent, poorly structured, or relies on meaning not solely encoded in language (see Anastasopoulos et al. 2022). Evaluation of other aspects, such as speech clarity and voice naturalness, is only beginning to gain traction (see the latest IWST evaluation campaign at https://iwslt.org/2022/speech-to-speech), while facets like human-machine interaction remain largely unexplored.

MI faces a multitude of challenges due to the complexity of human communication, which are further compounded in multilingual spoken exchanges. At present, MI depends exclusively on the information embedded in spoken language, overlooking essential communication elements such as non-verbal cues and vocal intonation. Additionally, the translation process is not anchored in the communicative event, resulting in machines lacking awareness of the context, speaker's intentions, or the interlocutors' reactions.

To address these limitations, emerging approaches are being explored, such as incorporating additional layers of information, like images, into the process (Sulubacak et al. 2019). More recently, generative language models (e.g., ChatGPT) and their ability to derive meaning from language have shown promising advances in translation, improving aspects such as text coherence, gender usage, and more (Hendy et al. 2023; Castilho et al. 2023).

As MI continues to progress, it has become increasingly evident that numerous tasks requiring a high level of human intelligence, such as interpretation, can be effectively executed by machines without them necessarily displaying intelligence themselves (e.g., Floridi 2023). However, it is important to recognize that MI may not be appropriate for all purposes, regardless of the performance quality it reaches in the near future. In scenarios where deep understanding, human empathy, and accountability are essential, human interpreters will remain irreplaceable. We are now entering an era where both humans and machines collaboratively facilitate access to multilingual content. This will require new collective efforts in providing counselling on and regulating its use according to practical and ethical considerations.

References

Anastasopoulos, A., Barrault, L. Bentivogli, L., Zanon Boito, M., Bojar, O., Cattoni, R., Currey, A. et al. 2022. "Findings of the IWSLT 2022 Evaluation Campaign." In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, 98–157. Dublin: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.iwslt-1.10.

Castilho, S., Mallon, C., Meister, R., & Yue, S. 2023/Forthcoming. "Do Online Machine Translation Systems Care for Context? What about a GPT Model?" In *24th Annual Conference of the European Association for Machine Translation (EAMT 2023), 12-15 June 2023*. Tampere: EAMT. https://events.tuni.fi/eamt23/

Fantinuoli, C. & Prandi, B. 2021. "Towards the Evaluation of Automatic Simultaneous Speech Translation from a Communicative Perspective." In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, 245–54. Bangkok: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.iwslt-1.29

Floridi, L. 2023. "AI as Agency without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4358789

Gaido, M., Savoldi, B., Bentivogli, L., Negri, M., & Turchi, M. 2020. "Breeding Gender-Aware Direct Speech Translation Systems." *arXiv:2012.04955 [cs.CL]*. http://arxiv.org/abs/2012.04955

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Jin Kim, Y., Afify, M., & Hassan Awadalla, H. 2023. "How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation." *arXiv:2302.09210 [cs.CL]*. http://arxiv.org/abs/2302.09210

Lee, A., Gong, H., Duquenne, P.-A., Schwenk, H., Chen, P.-J., Wang, C., Popuri, S. et al. 2022. "Textless Speech-to-Speech Translation on Real Data." *arXiv:2112.08352 [cs.CL]*. http://arxiv.org/abs/2112.08352

Sperber, M. & Paulik, M. 2020. "Speech Translation and the End-to-End Promise: Taking Stock of Where We Are." *arXiv:2004.06358 [cs.CL]*. http://arxiv.org/abs/2004.06358

Sulubacak, U., Caglayan, O., Grönroos, S.-A., Rouhe, A., Elliott, D., Specia, L. & Tiedemann, J. 2019. "Multimodal Machine Translation through Visuals and Speech." *arXiv:1911.12798 [cs.CL]*. http://arxiv.org/abs/1911.12798

Zhang, B., Haddow, B., & Rico Sennrich. 2022. "Revisiting End-to-End Speech-to-Text Translation from Scratch." *arXiv:2206.04571 [cs.CL]*. http://arxiv.org/abs/2206.04571