12
MACHINE INTERPRETING

Claudio Fantinuoli

12.1 Introduction

Language barriers pose a significant obstacle for numerous people. These barriers not only restrict access to content in daily leisure and entertainment contexts, in the circulation of news, etc., but also create difficulties in critical situations, such as accessing public services during humanitarian operations or while managing crisis scenarios, to name just a few. Previous research has extensively shown, for example, that limited proficiency in the languages of host societies significantly contributes to anxiety and stress among migrants (Ding and Hargraves, 2009); the absence of mutual understanding because of language barriers creates dangerous situations, hazards, exclusion, and other cascading effects during crises (O'Brien and Federici, 2022); and the lack of language accessibility in the information selection process impacts the framing and agenda-setting factors in newsrooms (Van Doorslaer, 2009). These challenges are not limited to the physical world; they are also becoming increasingly significant in online communication, where new challenges arise for achieving barrier-free interaction among users with diverse linguistic and accessibility needs (Kožuh and Debevc, 2018; Abarca et al., 2020).

For centuries, overcoming language barriers involved relying on human interpreters, both professional and amateur, adopting a common lingua franca (Cogo, 2015) or even creating artificial languages, like Esperanto (Fettes, 1997). While each approach presents its own advantages, it is important to acknowledge that these also come with recognised limitations. For example, human interpreters are not universally accessible or cost-effective, and their services are often reserved for specific situations deemed essential by stakeholders or mandated by law (Jaeger et al., 2019). The use of a lingua franca, such as English, is accessible only to a subset of individuals, and often, their proficiency in it is limited (Hartshorne et al., 2018). Most situations involving people from different language backgrounds remain unaddressed.

Several solutions are conceivable as a way to overcome language barriers. These include training more interpreters, training bilingual staff in multilingual workforces to act as interpreters, making interpreter provision more effective through the use of remote interpreting, etc. Among these opportunities, many have proposed speech translation technology as a

DOI: 10.4324/9781003053248-16

tool to enhance accessibility and bridge linguistic and cultural divides, offering a more inclusive and connected world (Waibel et al., 1991; Salesky et al., 2023; Anastasopoulos et al., 2022). By overcoming the exclusivity of professionals in delivering this service, machines promise to make accessibility across languages ubiquitous and more affordable (Susskind and Susskind, 2017), contributing, at least to some extent, to mitigating language barriers.

The quest for automatic speech translation is long, but it is only recently that this technology has gained momentum in research. The International Workshop on Spoken Language Translation (IWSLT),[1] the most important annual scientific conference dedicated to this topic, was founded in 2004. The first workshop described the vision of this new branch of computational science as an 'attempt to cross the language barriers between people with different native languages who each want to engage in conversation by using their mother tongue'.[2]

More recently, riding the wave of advancements in speech recognition and machine translation driven by neural networks and, lately, the introduction of large language models, MI has transitioned from research labs and conference venues to real-life applications. Despite the progress made, the pursuit of high-quality automatic speech translation is fraught with a wide range of challenges that go beyond technical and linguistic hurdles. Achieving a high standard of translation quality is just the beginning; it is also important to navigate the ethical considerations of MI, assess its suitability in various settings, and engage in deeper discussions about the implications of society having unlimited access to spoken language information from different languages and cultures.

This chapter seeks to enhance the understanding of the application of this technology beyond its mere technical aspects, targeting a broader audience rather than the experts of the field. Section 12.2 introduces the concept of MI and the terminological conventions used in different disciplines. Section 12.3 gives an overview of the development of the field. Section 12.4 delves into the technological approaches that are the base for MI. Section 12.5 presents the major challenges for machine interpretation. Section 12.6 outlines some technological perspectives, and Section 12.7 the challenges in evaluating MI applications. Section 12.8 reflects on possible ethical implications of MI in society at large. Finally, Section 12.9 concludes the chapter.

12.2 Definition and terminology

The automated process of converting spoken language from one language to another, particularly for live communication settings, is known by various terms across different disciplines. The two most common terms are *speech-to-speech translation* and *machine interpreting*.[3]

The term *speech-to-speech translation*, primarily used in computer science, refers to any technology that converts spoken words in one language into spoken words in another. The translation process can be performed offline in the case of pre-recorded audios and videos, or in a streaming format in the case of live and real-time translation (Seligman, 1997; Kidawara et al., 2020). Speech-to-speech is a specific category within *speech translation*, an even broader term that includes all forms of translation from an oral source, whether into a written form, for example, for the creation of subtitles (Xu et al., 2022; Papi et al., 2023), or another oral form, for example, for dubbing or voiceover (Saboo and Baumann, 2019; Hu et al., 2022).

The term *machine interpreting* or *machine interpretation* (MI) is predominantly used in the field of translation studies and refers to a specific subset of speech-to-speech translation

characterised by its focus on immediacy, that is, the fact that the message in translation is available only once and cannot be edited (Fantinuoli, 2023). This is similar to the distinctive features of human interpreting (Kade, 1968; Pöchhacker, 2022). The focus of MI is therefore on real-time and live scenarios, setting it apart from other forms of speech-to-speech translation. MI can be either *consecutive*, where the machine processes and translates an entire oral segment after it is spoken (such as a sentence or a longer passage), or *simultaneous*, where the translation is generated incrementally as the original speech is being delivered, that is, based on partial input. Notably, MI may involve nuanced intervention in the translation process, which may include adaptations, omissions, voice speed changes, or other modifications to tailor communication effectively for specific contexts.

While the term *speech-to-speech translation* may refer to both the underlying technology and its application, MI is primarily used only to denote the application rather than the technology itself.

12.3 History

Modern speech translation applications have only caught the world's attention in the last few years thanks to impressive improvements in the underlying technologies, but researchers have been working to develop and understand its potential for much longer than that. The history of speech-to-speech translation has always been tightly related to the progress of a series of different technologies, especially speech recognition and machine translation and, more recently, of large language and multimodal models.

While research and development in machine translation date back to the 1940s, it was only during the 1983 World Telecommunication Exhibition that multilingual speech translation appeared on the stage (Kato, 1995). This technology, integrating speech recognition and speech synthesis with machine translation, captured widespread attention at that event and initiated a series of projects and conferences in the domain. Following that, the SpeechTrans (Tomita et al., 1990) system, developed in 1988, marked another significant step in speech translation, consolidating the field inside the computer science space. In 1992, the German Federal Ministry of Education and Research launched a project called *Verbmobil*, which developed a prototype portable translation system for the language pair German–English. During the implementation of Verbmobil, a great number of scientific publications and several spin-off products were generated. Start-up companies were established by the participating universities and research centres (Wahlster, 2000). The project was characterised by its pioneering nature, and its legacy continues to resonate today.

Over the next two decades, especially after the Consortium for Speech Translation Advanced Research was founded in 1991, various speech translation systems were created, ranging from restricted domain and vocabulary to open-domain translation systems (Waibel et al., 1991; Fügen et al., 2006). To further advance speech translation systems, the International Workshop on Spoken Language Translation (IWSLT) was founded in 2004. In July 2010, the National Institute of Information and Communications Technology (NICT) in Japan launched the world's inaugural field experiment of a network-based multilingual speech-to-speech translation system using a smartphone application.

In 2019, the European Parliament, a significant institution utilising a multilingual regime with 24 official languages, initiated an innovation project in the space of speech translation.[4] While this project was not related to MI but rather to speech-to-text translation (the goal was to improve the accessibility of plenary meetings, enabling deaf and hard-of-hearing

individuals to follow debates in real time), the initiative represents a crucial milestone in the large-scale implementation of speech translation technologies.

More recently, MI has begun to be used in real-life scenarios, such as conferences, workshops, town halls, etc.[5] While these systems were initially designed only for consecutive interpretation, recent developments have also introduced systems capable of near simultaneous interpretation.

Speech-to-speech translation is a challenging research problem. From a technological perspective, the evolution of speech technology has followed a pattern similar to that of machine translation, moving from simple, rule-based and statistical approaches to more complex, deep learning paradigms.[6] To simplify the task and make it somewhat more feasible, the first research projects addressing this topic in the 1990s worked on restricted domains, such as scheduling appointments or making orders, and moved only later to free and spontaneous speech translation. Initially, speech translation systems were based on rule-based or statistical approaches and, therefore, had all the limitations that came with them. Limitations include the inability to translate idiomatic expressions, to take into consideration context, or to produce fluent translation, to name just a few. After Google announced the development of the Google Neural Machine Translation (Wu et al., 2016) in September 2016, a significant paradigm shift from statistical approaches to deep learning and neural machine translation has been taking place, paving the way for new speech translation systems with increased accuracy. The increased availability of special datasets, such as Europarl-ST (Iranzo-Sànchez et al., 2020) or TEDx (Salesky et al., 2021), has further increased the efforts of the community to create dedicated models for speech translation tasks.

At the moment of writing, commercial systems are designed based on the cascading approach, that is, combining several components, such as speech recognition and machine translation (see Section 12.4.1). However, significant strides have been made in developing and embracing end-to-end systems, that is, systems capable of translating from one spoken language to another directly, without relying on an intermediate text representation (Jia et al., 2019b). In the future, the end-to-end paradigm will simplify architectures and bring about systems that will further improve the translation experience, allowing for example expressiveness and the retention of the features of the original (Barrault et al., 2023).

Recently, much interest has also been devoted to reducing the latency of cascading and direct systems for a range of reasons, including supporting real-time and near simultaneous translation (Sudoh et al., 2020), extending coverage of dialect and low-resource languages, and adapting register (Agarwal et al., 2023).

12.4 Current approaches

As previously explained, there are two main technical approaches used in MI: the cascading approach and the end-to-end approach. These can be seen as the extreme ends on a continuum within many possible implementations and variations.

12.4.1 Cascading approach

The cascading approach is characterised by a pipeline of variable components concatenated one after the other. Cascading approaches are based on composite systems. In such

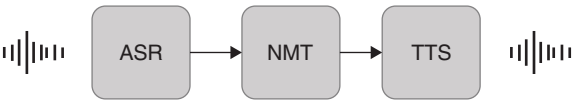

Figure 12.1 Simple cascading system for MI.

systems, the overall goal of the application is divided into multiple tasks, with each task assigned to a specialised module. The simplest cascading configuration involves automatic speech recognition (ASR), to convert speech into written text; neural machine translation (NMT), to translate the written text from one language to another; and voice synthesis or text-to-speech (TTS), to convert the written translation into an oral form (Figure 12.1 – see also Davitti, this volume).

Configurations of cascading systems can vary considerably, depending on technological advancements, design requirements, and complexity of the system. Recently, more sophisticated applications have emerged. These may have, for example, dedicated components for achieving low latency and continuous translation, or components to adapt translation strategies to the translation goals, for example, speed adaptation, speech correction or normalisation, and language simplification (for translation into plain or simplified language), etc.

Generally speaking, cascading systems have the advantage of being able to leverage robust technologies with years of development behind them. The robustness of the single components is also granted by the availability of large training sets (Sperber and Paulik, 2020). However, cascading systems suffer from two challenges. First, the extreme complexity of training and maintaining composite pipelines. Second, such systems tend to suffer from error propagation from one component to the next, that is, errors made in one stage of the process are carried forward and potentially amplified in subsequent stages. Notwithstanding the possibility to correct errors, for example, in transcription issues (Martucci et al., 2021; Macháček et al., 2023b), each of these stages can introduce new errors. Since each subsequent stage relies on the output of the previous one, errors can accumulate and worsen as the process continues.

Simultaneous machine interpreting is the most intricate form of speech translation, as it adds a new significant temporal constraint. Systems need to translate an ongoing stream of speech incrementally, without interruptions and without complete knowledge of what the speaker will say moments later. To accomplish this, speech must be segmented into meaningful chunks in real time (Waibel et al., 2012). The goal of a segmentation module is to achieve a system that balances translation accuracy and latency. *Latency* can be broadly defined as the time (in seconds) between the speaker uttering a word and the moment the translation engine delivers the same word. Achieving lower latency, which implies less context for the engine, can pose challenges in producing accurate translations. Segmentation methods range from detecting pauses in the speaker's flow and employing fixed word lengths to utilising dynamic approaches based on real-time syntactic and semantic analysis of incoming speech.

In Figure 12.2, a simultaneous model is added to the pipeline. This model allows the pipeline to achieve near simultaneous speech translation. Generally speaking, this model aims at segmenting the incoming stream of text into chunks of text, for example, based on

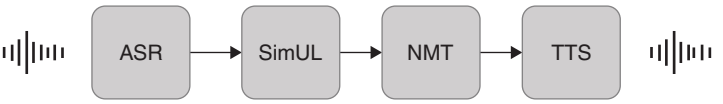

Figure 12.2 Cascading system for simultaneous translation.

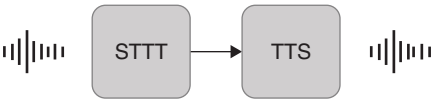

Figure 12.3 Cascading system with direct speech-to-text translation.

units of meaning. These units of meaning can then be passed to the MT engine, and thereafter to the TTS model (Kumar et al., 2013; Wang et al., 1999; Ma et al., 2019), in a way that the translation, once spoken, is meaningful and coherent.

Pipelines in translation processes are not inherently restricted to a unidirectional flow. Depending on the chosen architecture, feedback loops can be integrated to enhance control over the translation process. Contrary to the unidirectional approach illustrated in Figure 12.3 (which necessitates segmenting the input stream into chunks that resemble semantically coherent sentences and thus incurs a latency at least as long as the length of these sentences), the incorporation of a feedback loop can reduce latency. By reintroducing the already-translated portion of a sentence back into the system, the MT engine is prompted to complete the translation of the new segment while still considering the context of the previously translated part. This strategy not only achieves shorter latency but also improves coherence and cohesion in the translated output.

As previously mentioned, tasks and components are not rigidly defined. For example, the tasks of ASR and MT can be unified in a single component performing direct speech-to-text translation (Berard et al., 2016; Papi et al., 2023; Sethiya and Maurya, 2023), that is, a model that is able to directly convert speech signals in a language into text in another language. The translation can then be passed directly to a TTS system for audio generation.

Several additional components can be integrated into a cascading pipeline, such as text normalisation (Fügen, 2008), language identification (Singh et al., 2021), suppression of disfluencies (Fitzgerald et al., 2009), prosody transfer (Kano et al., 2018), speaker diarisation (Yang et al., 2023), and so forth.

Recently, translation has been performed not only by neural machine translation engines but also by large language models, such as GPT.[7] This enables a more comprehensive approach to the translation task, leveraging the capabilities of generative AI.

For instance, LLMs can incorporate contextual understanding for translation through methods such as in-context learning (Moslem et al., 2023) or even integrate automatic quality estimation to enhance real-time translation (cf. Kocmi and Federmann, 2023; Wang and Fantinuoli, 2024).

12.4.2 End-to-end approach

The emerging alternative to cascading systems is the end-to-end approach. This approach uses a single component to translate input audio in one language directly into output audio

in another language without generating intermediate text representations (Lee et al., 2022). Prominent projects are Translatotron[8] by Google and SeamlessM4T[9] by Meta.

While cascading systems are unable to preserve para-/non-linguistic speech characteristics that are central to human communication, such as prosody, tone of voice, pauses, and emphasis, the end-to-end approach promises to maintain such traits also in translation. In fact, end-to-end models learn to map features of language in a holistic way. They map spoken language, comprising all the aforementioned features, to translations. End-to-end models are then able to reproduce these features, at least partially (Barrault et al., 2023).

Initially, end-to-end models primarily used supervised machine learning techniques that rely on bilingual speech datasets (Jia et al., 2019a, 2019b, 2021). This approach has two limitations. On the one hand, collecting bilingual datasets, especially but not only for low-resource languages, is difficult. On the other hand, the lack of bilingual data with corresponding non-linguistic characteristics in both source and target languages makes it impossible to transfer such features to the translated speech. More recently, attempts have been made to overcome this limitation, using unsupervised machine learning with monolingual datasets (Nachmani et al., 2023).

While the end-to-end approach was initially applied only to consecutive translation, the first systems have also been proposed for the simultaneous modality (Barrault et al., 2023).

12.5 Challenges

MI encounters various challenges, drawing from the complexities inherent in speech translation (Macháček et al., 2023b) and adding unique difficulties due to its need for immediacy. These challenges broadly fall into the following categories: linguistic, cultural and communicative, and technological.

12.5.1 Linguistic challenges

The linguistic challenges in MI relate to all aspects connected with language, and especially the distinctive features of spoken language. These challenges include, but are not exhaustive to, the following:

- **Disfluencies.** Natural speech often contains filler words and sounds, such as ‘um’, ‘uh’, ‘you know’, and ‘like’.
- **Poor enunciation.** Mumbled or slurred words are difficult to interpret accurately.
- **Features of spoken grammar.** People often speak in sentences that would not be grammatically or syntactically correct in written language: changing direction mid-sentence, not completing sentences, or using non-standard grammar.
- **Proper nouns.** Identifying and transcribing names (in cascading systems) or maintaining the correct pronunciation of names (in direct systems) is a big challenge for machines.
- **Simultaneity.** The need to process incoming speech while it is unfolding is a further challenge.

The difficulty in tackling these challenges is exacerbated in the case of low-resource languages, particularly Indigenous and minority languages. The same issue applies to many

accent variations that are not covered in the training data. Therefore, providing comprehensive language coverage continues to be a significant challenge. Commercial MI systems, such as those offered by Interprefy[10] or KUDO,[11] offer support for 80 and 38 languages, respectively, at the time of writing. Variations in quality might be significant, not only among vendors, but also between languages. Addressing these challenges requires ongoing effort, resources, sophisticated algorithms, and extensive linguistic knowledge. It should be recognised that the number of languages supported in machine translation is constantly increasing, and this trend is expected to be reflected in commercial systems for MI as well.[12]

12.5.2 Cultural and communicative challenges

The cultural and communicative challenges of MI are many. Languages and cultures are understandably deeply interconnected. One of the most nuanced challenges in speech translation is handling cultural references, idioms, and colloquialisms that are specific to a particular language or culture (Yao et al., 2023). Machine translation performs poorly on culturally specific data, mostly due to the gap between the cultural norms associated with the languages (Liebling et al., 2022). The same applies to irony and sarcasm, both of which are highly dependent on cultural context.

Culturally bound information often cannot be translated directly, as its meaning is deeply rooted in the culture and history of the people who use it. A speech translation system should recognise such information and strategically translate it into an equivalent expression in the target language. To be effective, the speech translation system would be required to consider not only the semantic correspondence but also its communicative dimension, which is limited to that specific moment and communicative act.

A fully functional speech translation system would require profound cultural and contextual understanding to precisely interpret and translate cultural references. Moreover, it should possess the capability to make informed decisions regarding translation strategies, based on the language/culture pairs and the communicative context. While initial research is undergoing in this direction (cf. Yao, 2023), such advanced abilities are currently beyond the scope of existing systems.

MI also faces challenges in processing background information about a communicative event. Achieving accurate translation hinges significantly on comprehending the context of a speech or conversation. In the absence of, or with limited access to, this contextual understanding, a translation system may misinterpret the meaning or intention behind an utterance. This could lead to potentially misleading translations.

Similarly, machine translation, especially in the context of real-time speech communication, is not grounded in the communicative context (cf. Lala et al., 2019). While in some simple contexts, speech translation might functionally work even with or without limited access to broader context, in more complex and nuanced settings, systems would need a comprehensive grasp of the broader communicative setting to provide accurate translations. This involves understanding the relationships between speakers, their tone of voice, and the numerous non-verbal cues that are key to constructing meaning (Seeber, 2012). While progress has been made in aspects such as multimodality (Sulubacak et al., 2019), the limitations in grounding the translation in MI remain numerous.

12.5.3 Technical challenges

The technological challenges related to the design and deployment of speech translation systems at scale are also manifold and are common to many machine learning applications (Anastasopoulos et al., 2022; Agarwal et al., 2023).

A critical constraint is latency, particularly in real-time translation scenarios, such as live broadcasting or interpersonal communication (Sarkar, 2016). *Latency* is the delay between the original spoken words and the availability of the translation. Reducing the delay between the spoken word and its translated output is essential to prevent conversation disruption, misunderstandings, or a poor user experience. In a situation where a speaker is illustrating a chart, for example, a delayed translation would not be conducive to achieving a functional understanding. In an ideal scenario, the translation should therefore be voiced quickly, matching the natural rhythm of speech, including its flow, pauses, and moments without speech.

Latency is basically determined by two elements: on the one hand, technical constraints, that is, the time which is necessary to inference the translation model(s) as well as the delay due to data transmission and processing and, on the other, linguistic constraints. Technical constraints are typical to all forms of MI, that is, both consecutive as well as simultaneous, and can be addressed by having more performant hardware and improved language models to reduce inference time or reducing the components involved in the process. Linguistic constraints apply only to simultaneous translation, for which achieving low latency while maintaining accuracy and reliability in translation presents a substantial challenge (Chang et al., 2022). Because of the progressive nature of the translation process while the original speech is unfolding, the system needs to wait for some context before processing and outputting the translation. The challenge lies in finding the right compromise: Reducing the latency means the system has less context for the translation, which can negatively impact accuracy. Conversely, allowing more latency will enhance the translation quality but will degrade the translation experience. From a linguistic perspective, latency is tied to the speech segmentation method chosen. These can range from detecting pauses in the speaker's delivery and using fixed word counts to dynamic strategies that analyse the syntax and semantics of the incoming speech in real time.

Another key concern is the reliable and scalable deployment of computationally intensive systems. Speech translation tools must consistently deliver translations under a variety of conditions, accommodate a large number of users simultaneously, and do so in real time. To ensure this level of performance, robust infrastructure and efficient algorithms are necessary, each posing its own set of complex challenges. The complexity is further compounded for cascading systems due to their architectural intricacy and the multitude of language models requiring maintenance and deployment. Currently, deploying such complex systems is feasible only in the cloud and in robust data centres. Offline deployment on the consumers' devices remains beyond reach for now, although it is foreseeable that advancements in both software and hardware will eventually support this deployment method (Wang et al., 2022).

12.6 Evaluation

Evaluating the quality of MI is an important activity. Evaluations yield insights that are crucial for various stakeholders, including developers, users, buyers, certifying agencies, and more (Han, 2022).

Quality in interpreting is a nuanced and variable concept heavily influenced by the specific needs and idiosyncrasies of end users or service purchasers (see Davitti, Korybski, Orăsan and Braun, this volume). It becomes paramount in high-stakes scenarios, where the accuracy and subtlety of translation can have significant implications. The notion of quality, therefore, is not uniformly defined and varies according to the context of use.

The task of assessing the quality of translated content is multifaceted and complex. Measuring quality is inherently challenging due to the somewhat-intangible nature of spoken language translation (Garcia Becerra, 2016a, 7). Quality perceptions can differ significantly among users, adding a layer of subjectivity to what is considered a correct and high-quality translation. Quality standards vary depending on the interpreting context. For example, in human conference interpreting, the focus is on the interpreter's output, including content, language, and delivery. While these features continue to remain important in public service settings, such as social and healthcare interpreting, other skills, such as interactional skills and discourse management, become relevant (Kalina, 2012). There seems to be a consensus that there is poor agreement on what constitutes an acceptable translation (Zhang, 2016).

Translation quality can be evaluated manually or automatically. Human evaluations provide a comprehensive view of quality by considering various aspects of communication, offering a deep understanding of the interpreting performance, as indicated by interpreting scholars (Pöchhacker, 2002; Garcia Becerra, 2016b). However, manual evaluations are labour-intensive, time-consuming, and expensive (Wu, 2011). In automatic speech translation, manual evaluation has been used to assess both accuracy and fluency (cf. Fantinuoli and Dastyar, 2022).

Automated or semi-automated metrics have been proposed as an alternative to manual evaluation in order to simplify and speed up this process. Very few studies have applied automated evaluation to human interpreting, as in the case of semantic similarity proposed by Zhang (2016). A higher number of studies have applied automated evaluation to speech-to-text translation. These use statistical metrics, such as BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020), and chrF2 (Popović, 2017), both to monitor quality evolution from a developer perspective and to compare several systems, for example, during international evaluation campaigns (Agarwal et al., 2023). Systems are often evaluated in terms of their ability to produce translations that are similar to the target language references. More recently, reference-free evaluations have been proposed, leveraging on multilingual sentence embeddings or other techniques that attempt to capture meaning (cf. Wang and Fantinuoli, 2024).

When applying automatic metrics, one important consideration that is key to understanding how reliable they are is how well they correlate with human judgment. Some scholars have come to the conclusion that, given the current quality levels of the systems, simple automatic metrics, such as COMET, can be used as proxies for quality estimation (Macháček et al., 2023a). Semantic vectors and large language models, which can directly compare the original with the translation, have also shown promise in inferring quality which better relates to human judgments (Kocmi and Federmann, 2023; Han and Lu, 2021).

12.7 Ethical aspects

Machine interpreting, like any emerging technology, presents a range of ethical challenges that require careful consideration and governance (Cath, 2018; Floridi, 2021). As a tool

designed to enhance communication and understanding across language barriers, it has the potential to significantly impact various areas of human life. For this reason, its application must be managed responsibly to ensure ethical integrity. At a very general level, we can categorise the ethical challenges of any AI solution, hence also of MI, into three primary scenarios:[13]

- **Overuse.** This occurs when MI systems are deployed without a genuine need, leading to unnecessary resource expenditure. The economic and environmental costs are significant, considering the substantial energy and computational resources required to run these systems. Overuse not only leads to wasteful expenditure but also potentially contributes to environmental degradation due to the high energy demands of current ML technologies.
- **Misuse.** This scenario arises when the technology is employed in situations where it may cause harm. The accuracy and appropriateness of MI systems vary based on several factors, including language pairs, context of the communication, cultural nuances, technical limitations, and expectations. Using MI in sensitive or high-stake environments, such as legal proceedings or medical consultations, without adequate safeguards can lead to misinterpretations with serious consequences. Such misuse is not only unethical but also potentially harmful and should be regulated to prevent adverse outcomes.
- **Underuse.** This refers to situations where MI is not utilised despite its potential to significantly enhance communication and accessibility. Underuse is considered unethical as it denies the benefits of reduced language barriers to those who could otherwise stand to gain from them. It is also economically imprudent, as the technology offers a cost-effective solution for increasing accessibility. The underuse of MI can stem from a lack of awareness, technological limitations, or resistance to change. Addressing these barriers is crucial for maximising the technology's positive impact.

The preceding categorisation serves as a general framework to guide the responsible adoption of this technology. However, many factors remain to be defined in practical terms. Such factors include deciding who is in charge of defining 'critical use', or how to define metrics of acceptable translation performances. From a technical and legal standpoint, MI systems present several critical aspects that necessitate responsible management and regulation. Key areas of concern include:

- **Confidentiality.** MI interpreting applications are cloud-based and face data breach risks. Ensuring confidentiality requires robust encryption, secure storage, and strict access controls. Companies should adopt certifications like ISO 27001 and SOC2.
- **Data ownership and privacy.** MI systems process sensitive data, which raises ownership and privacy concerns. Clear policies and compliance with regulations like GDPR are crucial for safeguarding user data.
- **Appropriate use.** MI system effectiveness varies by language pair, situation complexity, and cultural nuances. Users need guidelines to ensure they avoid misuse in critical contexts. Certifications can ensure reliability in regulated areas, like courts or healthcare.
- **Liability.** Accountability for translation errors requires clarity. Balanced regulations are needed to ensure quality and innovation without eroding trust or stifling development.
- **Ethical AI and bias mitigation.** MI systems must address biases to prevent stereotypes or discrimination. Regular audits and updates are key to ensuring ethical AI practices.

In addressing these challenges, it is essential to develop a balanced approach that maximises the benefits while minimising its risks (Floridi et al., 2018). This involves continuous evaluation of the technology's impact, updating ethical guidelines, and ensuring that its deployment is aligned with societal values and needs. Stakeholders, including developers, users, and policymakers, must collaborate to establish standards and regulations that guide the responsible use of MI.

MI boldly promises to provide unrestricted access to spoken content across languages. While this ambition appears commendable and worth pursuing, it carries subtle risks that should not be overlooked. The interconnectedness of individuals, facilitated by the internet and social media, while fostering increased information exchange and knowledge accessibility, has simultaneously triggered societal polarisation and various negative consequences (Becker et al., 2019). In a similar vein, unrestricted access to information through machine interpretation could yield both advantages and disadvantages.

On the positive side, MI offers the potential for enhanced and autonomous dissemination of information and knowledge. While the exclusive provision of services by professionals provides numerous advantages, including the assurance of expertise and the high-quality standards that professionals can deliver, only machines have the possibility to make accessibility available to everyone (Susskind and Susskind, 2017).[14]

However, the ubiquity of spoken language translation also carries the risk of further radicalisation and polarisation of ideological positions. AI creates the illusion that all content can and should be made accessible in every language and culture. However, not all content created by individuals can be meaningfully translated without proper contextualisation of cultural, historical, and sociological aspects. Some content is deeply rooted in specific cultures or subcultures and only holds significance within that specific context. This makes translation – if not culturally mediated or contextualised – futile or even counterproductive. In such contexts, MI is bound to fall short or be executed poorly, potentially exacerbating misunderstandings and polarisation.

12.8 Future developments

The future trajectory of MI is difficult to predict, as it hinges significantly on advancements in various facets of language technology. While existing cascading systems are poised for quality enhancements, due to improvements in their underlying components, further evolution will be contingent on the integration of novel technologies and methodologies (Sperber and Paulik, 2020; Gaido et al., 2022), for example, end-to-end approaches (see Section 12.4.2).

Two primary developmental trajectories merit attention. The first developmental avenue involves augmenting the traditional cascading system pipeline with new components. These additions aim to address and improve the limitations inherent in current systems. Potential areas for exploration include the integration of large language models, the enhancement of information extraction from acoustic models, and the incorporation of visual elements.

LLMs and the capabilities of generative AI could serve various functions (Wu et al., 2023). Unlike traditional neural approaches to language and translation tasks, LLMs introduce the capacity for nuanced language manipulation and a certain level of understanding (Chang et al., 2023). While the extent of this 'understanding' remains a topic of debate among scholars (Bender and Koller, 2020) and hinges on the interpretation of the term itself, it is widely acknowledged that LLMs exhibit an enhanced proficiency in contextual

text interpretation. This capability can be directly applied to contextual translation (Karpinska and Iyyer, 2023) or used to pre-process source information for translation, akin to the semi-automated approach described by Korybski et al. (2022).

In the area of acoustic models, integration could focus on extracting information from a speaker's prosody (Barrault et al., 2023; Ko et al., 2023). Current cascading systems use acoustic models in their automatic speech recognition segment to transcribe spoken words into text for subsequent processing. However, valuable information – conveyed through speech elements such as intonation, prosody, and pauses – is often lost. This information is crucial, as communication extends beyond mere words to how those words are articulated. Such nuances might be implicitly captured in direct speech-to-text translation models (ten Bosch, 2003).

The second trajectory pertains to the advent of robust end-to-end systems. These systems stand to benefit from increased computational power, the availability of appropriate data, and innovative machine learning frameworks capable of yielding quality results with reduced training data requirements (Sperber and Paulik, 2020). The future of speech translation seems to be rapidly shifting from cascading systems to end-to-end systems. Recent developments show that end-to-end systems can produce translations while preserving the various features of speech. End-to-end models focus on often-overlooked elements of prosody, such as speech rate and pauses. They also retain the emotion and style of the original speech (Barrault et al., 2023; Jia et al., 2021; Nachmani et al., 2023). Not only can end-to-end models be a useful solution for dealing with languages that do not have a formal writing system (Duong et al., 2016), but they should also enable more effective streaming techniques and better source language segmentation approaches (see Section 12.4.1) that mimic the behaviour of human interpreters. Therefore, use of these models could lead to both a potential reduction in latency and a simplification of deployment and maintenance.

Visual analysis could further ground the translation process in the communicative context. Live communication typically encompasses both verbal and non-verbal elements, tailored to the situational demands and communicative goals of the participants (Lala et al., 2019; Sulubacak et al., 2019). This aspect is equally vital in multilingual communication and MI. Emerging vision systems, when combined with LLMs, demonstrate remarkable proficiency in image analysis, with live video analysis on the horizon. These systems are now capable of converting visual data into what might be termed 'situational meta-information' – essentially, information about what is happening in the communicative context: who is involved, what the features of the setting are, etc. Leveraging this data may significantly enrich the translation process, leading to enhanced accuracy and nuance.

In applications outside the simultaneous modality, virtual avatars might influence the perception of artificial interpretation further, for example, in dialogic contexts, where the embodiment of an interpreter seems to play a central role (Li et al., 2023).

12.9 Conclusions

This chapter has provided a thorough examination of MI, highlighting its significant evolution from early experimental stages to its current applications. Despite remarkable technological advancements, this technology continues to face complex challenges, particularly in accurately capturing the nuances of spoken language and contextual subtleties. There are also methodological challenges to tackle, such as the need to define shared quality criteria that should help diverse stakeholders evaluate the fit of the technology for their use cases.

The chapter also underscores the importance of addressing ethical considerations in the deployment of live speech translation technologies, emphasising the need for responsible use to maximise benefits while minimising potential risks.

The future of MI looks promising, with potential advancements in computational power and machine learning techniques, which could enhance its efficiency and accuracy. Overall, MI stands at a pivotal point. It has the capacity to significantly impact global communication, provided its development and application are managed with careful consideration of its technological, evaluative, and ethical dimensions.

Notes

1 https://iwslt.org.
2 https://www2.nict.go.jp/astrec-att/workshop/IWSLT2004/archives/000196.html.
3 For an in-depth analysis of this terminology, please refer to Pöchhacker (2024).
4 See 'call for tender' at: https://etendering.ted.europa.eu/cft/cft-display.html?cftId=5249.
5 See, for example, https://slator.com/live-speech-to-speech-ai-translation-goes-commercial.
6 See Poibeau (2017) for a detailed history of machine translation.
7 https://openai.com.
8 https://blog.research.google/2019/05/introducing-translatotron-end-to-end.html.
9 https://ai.meta.com/blog/seamless-m4t.
10 www.interprefy.com.
11 www.kudo.ai.
12 See, for example, https://blog.google/products/translate/google-translate-new-languages-2024/ and https://ai.meta.com/research/no-language-left-behind/.
13 See Floridi et al. (2018) for the general theoretical framework used here.
14 Susskind and Susskind (2017, 33) note that '[m]ost individuals and organizations find it challenging to afford the services of top-tier professionals', and that the use of AI might extend accessibility, where now only a limited number of people can actually avail themselves of these services.

References

Abarca, V.M.G., Palos-Sanchez, P.R., Rus-Arias, E., 2020. Working in Virtual Teams: A Systematic Literature Review and a Bibliometric Analysis. *IEEE Access* 8, 168923–168940.

Agarwal, M., Agrawal, S., Anastasopoulos, A., Bentivogli, L., Bojar, O., Borg, C., . . . Zevallos, R., 2023. Findings of the IWSLT 2023 Evaluation Campaign. In Salesky, E., Federico, M., Carpuat, M., eds. *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), Association for Computational Linguistics*, pp. 1–61.

Anastasopoulos, A., Barrault, L., Bentivogli, L., Zanon Boito, M., Bojar, O., Cattoni, R., . . . Watanabe, S., 2022. Findings of the IWSLT 2022 Evaluation Campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, 98–157.

Barrault, L., Chung, Y.-A., Meglioli, M.C., Dale, D., Dong, N., Duquenne, P.-A., . . . Wang, S., 2023. *SeamlessM4T: Massively Multilingual & Multimodal Machine Translation*. URL https://arxiv.org/abs/2308.11596

Becker, J., Porter, E., Centola, D., 2019. The Wisdom of Partisan Crowds. *Proceedings of the National Academy of Sciences* 116(22), 10717–10722.

Bender, E.M., Koller, A., 2020. Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 5185–5198.

Berard, A., Pietquin, O., Servan, C., Besacier, L., 2016. *Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation*. NIPS Workshop on End-to-End Learning for Speech and Audio Processing, December, Barcelona, Spain.

Cath, C., 2018. Governing Artificial Intelligence: Ethical, Legal and Technical Opportunities and Challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 37(2133).

Chang, C.-C., Chuang, S.-P., Lee, H.-Y., 2022. Anticipation-Free Training for Simultaneous Machine Translation. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*. Association for Computational Linguistics, Dublin, Ireland, 43–61.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., . . . Xie, X., 2023. A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology* 15(3), 1–4.

Cogo, A., 2015. English as a Lingua Franca: Descriptions, Domains and Applications. In Bowles, H., Cogo, A., eds. *International Perspectives on English as a Lingua Franca: Pedagogical Insights*, International Perspectives on English Language Teaching. Palgrave Macmillan UK, London, 1–12.

Ding, H., Hargraves, L., 2009. Stress-Associated Poor Health Among Adult Immigrants with a Language Barrier in the United States. *Journal of Immigrant and Minority Health* 11(6), 446–452.

Duong, L., Anastasopoulos, A., Chiang, D., Bird, S., Cohn, T., 2016. An Attentional Model for Speech Translation Without Transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, San Diego, CA, 949–959.

Fantinuoli, C., 2023. The Emergence of Machine Interpreting. *European Society for Translation Studies* 62, 10.

Fantinuoli, C., Dastyar, V., 2022. Interpreting and the Emerging Augmented Paradigm. *Interpreting and Society* 2(2), 185–194.

Fettes, M., 1997. Esperanto and Language Awareness. In Van Lier, L., Corson, D., eds. *Encyclopedia of Language and Education: Knowledge About Language*, Encyclopedia of Language and Education. Springer Netherlands, Dordrecht, The Netherlands, 151–159.

Fitzgerald, E., Hall, K., Jelinek, F., 2009. Reconstructing False Start Errors in Spontaneous Speech Text. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. Association for Computational Linguistics, 255–263.

Floridi, L., 2021. *The End of an Era: From Self-Regulation to Hard Law for the Digital Industry*. Springer, Rochester, NY.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., . . . Vayena, E., 2018. AI4People – an Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28(4), 689–707.

Fügen, C., 2008. *A System for Simultaneous Translation of Lectures and Speeches* (Unpublished doctoral dissertation). University of Karlsruhe.

Fügen, C., Kolss, M., Bernreuther, D., Paulik, M., Stuker, S., Vogel, S., Waibel, A., 2006. Open Domain Speech Recognition & Translation: Lectures and Speeches. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings* 1, I–I).

Gaido, M., Papi, S., Fucci, D., Fiameni, G., Negri, M., Turchi, M., 2022. Efficient Yet Competitive Speech Translation: FBK@IWSLT2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*. Association for Computational Linguistics, Dublin, Ireland, 177–189 (in-person and online), May.

Garcia Becerra, O., 2016a. Do First Impressions Matter? The Effect of First Impressions on the Assessment of the Quality of Simultaneous Interpreting. *Across Languages and Cultures* 17(1), 77–98.

Garcia Becerra, O., 2016b. Survey Research on Quality Expectations in Interpreting: The Effect of Method of Administration on Subjects' Response Rate. *Meta* 60.

Han, C., 2022. Interpreting Testing and Assessment: A State-of-the-Art Review. *Language Testing* 39(1), 30–55.

Han, C., Lu, X., 2021. Interpreting Quality Assessment Re-Imagined: The Synergy Between Human and Machine Scoring. *Interpreting and Society* 1(1), 70–90.

Hartshorne, J.K., Tenenbaum, J.B., Pinker, S., 2018. A Critical Period for Second Language Acquisition: Evidence from 2/3 Million English Speakers. *Cognition* 177, 263–277.

Hu, C., Tian, Q., Li, T., Wang, Y., Wang, Y., Zhao, H., 2021. *Neural Dubber: Dubbing for Videos According to Scripts*. Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Iranzo-Sànchez, J., Silvestre-Cerdà, J.A., Rosello, N., Sanchis, A., . . . Juan, A. 2020. *Europarl-ST: A Multilingual Corpus for Speech Translation of Parliamentary Debates*. Proceedings of the ICASSP2020 Conference.

Jaeger, F.N., Pellaud, N., Laville, B., Klauser, P., 2019. Barriers to and Solutions for Addressing Insufficient Professional Interpreter Use in Primary Healthcare. *BMC Health Services Research* 19(1), 753.

Jia, Y., Johnson, M., Macherey, W., Weiss, R.J., Cao, Y., Chiu, C.-C., . . . Wu, Y., 2019a. Leveraging Weakly Supervised Data to Improve End-to-End Speech-to-Text Translation. URL https://doi.org/10.48550/arXiv.1904.06037

Jia, Y., Ramanovich, M. T., Remez, T., Pomerantz, R., 2021. Translatotron 2: Robust Direct Speech-to-Speech Translation. arxiv.org/abs/2107.08661

Jia, Y., Weiss, R.J., Biadsy, F., Macherey, W., Johnson, M., Chen, Z., Wu, Y., 2019b. Direct Speech-to-Speech Translation with a Sequence-to-Sequence Model. URL https://doi.org/10.48550/arXiv.1904.06037

Kade, O., 1968. *Zufall und Gesetzmäßigkeit in der Übersetzung*. Verlag Enzyklopädie Edition, Leipzig.

Kalina, S., 2012. Quality in Interpreting. In *Handbook of Translation Studies Online*, vol. 3. John Benjamins Publishing Company, 134–140.

Kano, T., Takamichi, S., Sakti, S., Neubig, G., Toda, T., Nakamura, S., 2018. An End-to-End Model for Crosslingual Transformation of Paralinguistic Information. *Machine Translation* 32(4), 353–368.

Karpinska, M., Iyyer, M., 2023. Large Language Models Effectively Leverage Document-Level Context for Literary Translation, but Critical Errors Persist. URL https://doi.org/10.48550/arXiv.2304.03245

Kato, Y., 1995. The Future of Voice-Processing Technology in the World of Computers and Communications. *Proceedings of the National Academy of Sciences* 92(22), 10060–10063.

Kidawara, Y., Sumita, E., Kawai, H., eds., 2020. *Speech-to-Speech Translation. SpringerBriefs in Computer Science*. Springer, Singapore.

Ko, Y., Fukuda, R., Nishikawa, Y., Kano, Y., Sudoh, K., Nakamura, S., 2023. Tagged End-to-End Simultaneous Speech Translation Training Using Simultaneous Interpretation Data. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*. Association for Computational Linguistics, Toronto, Canada, 363–375.

Kocmi, T., Federmann, C., 2023. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. URL https://doi.org/10.48550/arXiv.2302.14520

Korybski, T., Davitti, E., Orasan, C., Braun, S., 2022. A Semi-Automated Live Interlingual Communication Workflow Featuring Intralingual Respeaking: Evaluation and Benchmarking. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 4405–4413.

Kozuh, I., Debevc, M., 2018. Challenges in Social Media Use Among Deaf and Hard of Hearing People. In Dey, N., Babo, R., Ashour, A.S., Bhatnagar, V., Salim Bouhlel, M., eds. *Social Networks Science: Design, Implementation, Security, and Challenges*. Springer International Publishing, Cham, 151–171.

Kumar, V., Sridhar, R., Chen, J., Bangalore, S., Ljolje, A., Chengalvarayan, R., 2013. Segmentation Strategies for Streaming Speech Translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 230–238.

Lala, C., Madhyastha, P., Specia, L., 2019. Grounded Word Sense Translation. In Bernardi, R., Fernandez, R., Gella, S., Kafle, K., Kanan, C., Lee, S., Nabi, M., eds. *Proceedings of the Second Workshop on Shortcomings in Vision and Language*. Association for Computational Linguistics, Minneapolis, MN, 78–85.

Lee, A., Gong, H., Duquenne, P.-A., Schwenk, H., Chen, P.-J., Wang, C., . . . Hsu, W.-N., 2022. Textless Speech-to-Speech Translation on Real Data. URL https://doi.org/10.48550/arXiv.2112.08352

Li, R., Liu, K., Cheung, A.K.F., 2023. Interpreter Visibility in Press Conferences: A Multimodal Conversation Analysis of Speaker – Interpreter Interactions. *Humanities and Social Sciences Communications* 10(1), 1–12.

Liebling, D., Heller, K., Robertson, S., Deng, W., 2022. Opportunities for Human-Centered Evaluation of Machine Translation Systems. In Carpuat, M., de Marneffe, M.-C., Meza Ruiz, I.V., eds. *Findings of the Association for Computational Linguistics: NAACL 2022. Association for Computational Linguistics*, 229–240.

Ma, M., Huang, L., Xiong, H., Zheng, R., Liu, K., Zheng, B., . . . Wang, H., 2019. STACL: Simultaneous Translation with Implicit Anticipation and Controllable Latency Using Prefix-to-Prefix

Framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 3025–3036.
Macháček, D., Bojar, O., Dabre, R., 2023a. MT Metrics Correlate with Human Ratings of Simultaneous Speech Translation. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*.
Macháček, D., Polák, P., Bojar, O., Dabre, R., 2023b. Robustness of Multi-Source MT to Transcription Errors. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*.
Martucci, G., Cettolo, M., Negri, M., Turchi, M., 2021. Lexical Modeling of ASR Errors for Robust Speech Translation. In *Interspeech 2021*. ISCA, 2282–2286.
Moslem, Y., Haque, R., Kelleher, J.D., Way, A., 2023. Adaptive Machine Translation with Large Language Models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 227–237.
Nachmani, E., Levkovitch, A., Ding, Y., Asawaroengchai, C., Zen, H., Ramanovich, M.T., 2023. Translatotron 3: Speech to Speech Translation with Monolingual Data. URL https://doi.org/10.48550/arXiv.2305.17547
O'Brien, S., Federici, F.M., eds., 2022. *Translating Crises*. Bloomsbury Academic.
Papi, S., Gaido, M., Negri, M., 2023. Direct Models for Simultaneous Translation and Automatic Subtitling: FBK@IWSLT2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*. Association for Computational Linguistics, Toronto, Canada, 159–168.
Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311–318.
Pöchhacker, F., 2002. Quality Assessment in Conference and Community Interpreting. *Meta* 46(2), 410–425.
Pöchhacker, F., 2022. Interpreters and Interpreting: Shifting the Balance? *The Translator* 28(2), 148–161, Routledge.
Pöchhacker, F., 2024. Is Machine Interpreting Interpreting? In *Translation Spaces*. John Benjamins Publishing Company.
Poibeau, T., 2017. *Machine Translation*. The MIT Press Essential Knowledge Series. The MIT Press.
Popović, M., 2017. chrF++: Words Helping Character N-Grams. In Bojar, O., et al., eds. *Proceedings of the Second Conference on Machine Translation*. Association for Computational Linguistics, Copenhagen, Denmark, 612–618.
Saboo, A., Baumann, T., 2019. Integration of Dubbing Constraints into Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*. Association for Computational Linguistics, Florence, Italy, 94–101.
Salesky, E., Darwish, K., Al-Badrashiny, M., Diab, M., Niehues, J., 2023. Evaluating Multilingual Speech Translation Under Realistic Conditions with Resegmentation and Terminology. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)* (In-Person and Online). Association for Computational Linguistics, Toronto, Canada, 62–78.
Salesky, E., Wiesner, M., Bremerman, J., Cattoni, R., Negri, M., Turchi, M., . . . Post, M., 2021. *The Multilingual TEDx Corpus for Speech Recognition and Translation*. Proceeding of the Interspeech 2021 Conference.
Sarkar, A., 2016. The Challenge of Simultaneous Speech Translation. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Keynote Speeches and Invited Talks*. Seoul, South Korea, 7.
Seeber, K.G., 2012. Multimodal Input in Simultaneous Interpreting: An Eyetracking Experiment. In Zybatov, L., Petrova, A., Ustaszewski, M., eds. *Proceedings of the 1st International Conference TRANSLATA, Translation & Interpreting Research: Yesterday – Today – Tomorrow*. Peter Lang, 341–347.
Seligman, M., 1997. Interactive Real-Time Translation via the Internet. In *AAAI*.
Sethiya, N., Maurya, C.K., 2023. End-to-End Speech-to-Text Translation: A Survey. URL https://doi.org/10.48550/arXiv.2312.01053
Singh, G., Sharma, S., Kumar, V., Kaur, M., Baz, M., Masud, M., 2021. Spoken Language Identification Using Deep Learning. *Computational Intelligence and Neuroscience*. URL https://doi.org/10.1155/2021/5123671

Sperber, M., Paulik, M., 2020. Speech Translation and the End-to-End Promise: Taking Stock of Where We Are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Sudoh, K., Kano, T., Novitasari, S., Yanagita, T., Sakti, S., Nakamura, S., 2020. *Simultaneous Speech-to-Speech Translation System with Neural Incremental ASR, MT, and TTS*. URL https://arxiv.org/abs/2011.04845

Sulubacak, U., Caglayan, O., Grönroos, S.-A., Rouhe, A., Elliott, D., Specia, L., Tiedemann, J., 2019. *Multimodal Machine Translation Through Visuals and Speech*, November 2019. URL https://arxiv.org/abs/1911.12798

Susskind, R., Susskind, D., 2017. *The Future of the Professions: How Technology Will Transform the Work of Human Experts*. Oxford University Press edition.

ten Bosch, L., 2003. Emotions, Speech and the ASR Framework. *Speech Communication* 40(1), 213–225.

Tomita, M., Tomabechi, H., Saito, H., 1990. Speech Trans: An Experimental Real-Time Speech-to-Speech. *Language Research* 24(4), 663–672.

Van Doorslaer, L., 2009. How Language and (Non-)Translation Impact on Media Newsrooms: The Case of Newspapers in Belgium. *Perspectives* 17(2), 83–92.

Wahlster, W., ed., 2000. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, Berlin.

Waibel, A., Cho, E., Niehues, J., 2012. Segmentation and Punctuation Prediction in Speech Language Translation Using a Monolingual Translation System. In *IWSLT*.

Waibel, A., Jain, A.N., McNair, A.E., Saito, H., Hauptmann, A.G., Tebelskis, J., 1991. JANUS: A Speech-to-Speech Translation System Using Connectionist and Symbolic Processing Strategies. URL https://doi.org/10.1109/ICASSP.1991.150456

Wang, H., Gao, W., Li, S., 1999. Utterance Segmentation of Spoken Chinese. *Chinese Journal of Computers* 22, 1009–1013.

Wang, X., Fantinuoli, C., 2024. Exploring the Correlation Between Human and Machine Evaluation of Simultaneous Speech Translation. *Proceedings of the 25th Annual Conference of the European Association for Machine Translation* 1, 325–334.

Wang, Y., Wang, J., Zhang, W., Zhan, Y., Guo, S., Zheng, Q., Wang, X., 2022. A Survey on Deploying Mobile Deep Learning Applications: A Systemic and Technical Perspective. *Digital Communications and Networks* 8, 1–17. URL https://doi.org/10.1016/j.dcan.2021.06.001

Wu, S.-C., 2011. *Assessing Simultaneous Interpreting: A Study on Test Reliability and Examiners' Assessment Behavior* (PhD thesis). Newcastle University.

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., . . . Dean, J., 2016. Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation. URL https://doi.org/10.48550/arXiv.1609.08144

Wu, Z., Li, Z., Wei, D., Shang, H., Guo, J., Chen, X., . . . Jiang, Y., 2023. Improving Neural Machine Translation Formality Control with Domain Adaptation and Reranking-Based Transductive Learning. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*. Association for Computational Linguistics, Toronto, Canada, 180–186 (in-person and online). URL https://doi.org/10.18653/v1/2023.iwslt-1.13

Xu, J., Buet, F., Crego, J., Bertin-Lemée, E., Yvon, F., 2022. Joint Generation of Captions and Subtitles with Dual Decoding. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*. Association for Computational Linguistics, Dublin, Ireland, 74–82 (in-person and online). URL https://doi.org/10.18653/v1/2022.iwslt-1.7

Yang, M., Kanda, N., Wang, X., Chen, J., Wang, P., Xue, J., . . . Yoshioka, T., 2023. DiariST: Streaming Speech Translation with Speaker Diarization. URL https://doi.org/10.48550/arXiv.2309.08007

Yao, B., Jiang, M., Yang, D., Hu, J., 2023. *Empowering LLM-Based Machine Translation with Cultural Awareness*. URL https://ar5iv.labs.arxiv.org/html/2305.14328

Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y., 2020. BERTScore: Evaluating Text Generation with BERT. URL https://doi.org/10.48550/arXiv.1904.09675

Zhang, X., 2016. Semi-Automatic Simultaneous Interpreting Quality Evaluation. *International Journal on Natural Language Computing* 5, 1–12. URL https://doi.org/10.5121/ijnlc.2016.5501