Bulletin of Electrical Engineering and Informatics
Vol. 14, No. 4, August 2025, pp. 2837~2848
ISSN: 2302-9285, DOI: 10.11591/eei.v14i4.9241

A multicriteria comparison of end-to-end and cascade speech-to-text translation models

Maria Labied[1], Abdessamad Belangour[1], Mouad Banane[2]
[1]Laboratory of Information Technology and Modeling LTIM, Faculty of Science Ben M'Sik, Hassan II University of Casablanca, Casablanca, Morocco
[2]Laboratory of Artificial Intelligence and Complex Systems Engineering, Faculty of Legal, Economic, and Social Sciences, Hassan II University of Casablanca, Casablanca, Morocco

Article Info

Article history:

Received Aug 31, 2024
Revised Jan 29, 2025
Accepted Mar 9, 2025

Keywords:

Cascade model
End-to-end model
Multicriteria comparison
weighted sum method machine translation
Speech recognition multilingual communication
Speech-to-text translation

ABSTRACT

This paper presents a thorough examination of two prominent speech-to-text translation (STT) models: the end-to-end (E2E) model and the cascade model. STT is a critical technology in today's multilingual society, facilitating communication across language barriers. The study focuses on comparing these models using a multicriteria approach to evaluate their effectiveness in translating speech to text. The E2E model represents a unified architecture that directly translates speech into text, while the cascade model involves separate modules for speech recognition and machine translation (MT). Both models have distinct advantages and challenges, which are explored in detail. Through a multicriteria comparison, this research assesses various performance metrics and criteria to determine the strengths and weaknesses of each model. The weighted sum method is employed to assign weights to evaluation criteria, providing a systematic evaluation framework. The findings have implications for researchers and developers in STT. By understanding the comparative performance of E2E and cascade models, researchers can make informed decisions regarding model selection based on criteria such as accuracy, speed, robustness, and resource requirements. This research advances the understanding of speech translation technologies and provides a foundation for future studies to refine evaluation methodologies, explore hybrid models, and enhance translation quality.

Corresponding Author:

Maria Labied
Laboratory of Information Technology and Modeling LTIM, Faculty of Science Ben M'Sik
Hassan II University of Casablanca
Casablanca, Morocco
Email: mr.labied@gmail.com

1. INTRODUCTION

Multilingual communication is the ability to communicate with people who speak different languages. It is an essential skill in today's globalized world, where people from different cultures and backgrounds are increasingly interacting with each other. Speech translation can play an important role in multilingual communication [1], [2]. It can help to break down language barriers and facilitate communication between people who do not speak the same language. This can be beneficial in a variety of settings, such as business, travel, education, and healthcare. For example, a business traveler who does not speak the language of the country they are visiting could use speech translation to communicate with locals. A student who is learning a new language could use speech translation to practice their conversational skills.

A healthcare provider could use speech translation to communicate with patients who do not speak the same language. As the technology continues to improve, speech translation will become more accurate, reliable, and affordable. This will make it a more widely used tool for people who need to communicate with others who speak different languages. However, the accuracy of speech translation can vary depending on the quality of the audio input, the complexity of the language, and the similarity between the two languages being translated. Also, Speech translation can be computationally expensive, so it is not always possible to use it in real-time. and can be difficult to use in noisy environments.

Different approaches have been used for speech translation [3], the earlier one is the traditional speech translation approaches use cascade-based speech translation models which consist of two separate components: an automatic speech recognition (ASR) model and a machine translation (MT) model. The ASR model transcribes the spoken input into text, and the MT model then translates the text into the target language [3]. Where ASR model is typically trained on a large corpus of audio recordings and their corresponding transcripts and the MT model is typically trained on a parallel corpus, which is a collection of text pairs that consist of the same content in two different languages. Once the ASR and MT models are trained, they can be used to translate speech. The ASR model transcribes the spoken input into text, and the MT model then translates the text into the target language. Cascade-based speech translation models have several advantages. First, they are relatively straightforward to train and deploy. Second, they can be very accurate, especially for high-resource languages with large amounts of parallel training data. Third, they can be used to translate a wide variety of speech content, including news, conversations, and technical documents. However, cascade-based models also have some disadvantages. First, they can be sensitive to errors in the ASR output. If the ASR model makes a mistake, The error will propagate and the MT model will likely produce an incorrect translation [4]-[7]. Second, cascade-based models can be slow, as they require two separate models to be run. Third, cascade-based models can be difficult to adapt to new languages or domains.

The recent speech translation approaches are the end-to-end (E2E) and are known also as direct speech translation. E2E [8]-[11] speech translation approaches aim to directly translate speech from one language to another without the need for an intermediate text representation. This can be done by using a single neural network that learns to map from the acoustic features of the speech to the target language text. E2E approaches have the potential to overcome the error propagation problem of traditional cascaded systems [12]-[14], but they are more challenging to train and require more data [15]-[17]. E2E-STT has been made possible by the development of large language models (LLMs) and the availability of large datasets of speech and text. LLMs can be used to pre-train E2E speech translation models, which can help to improve the performance of the models. Large datasets of speech and text can be used to train E2E speech translation models, which can help to improve the robustness of the models to different acoustic conditions and language variations. Compared to cascade models E2E are faster and can be more easily adapted to new languages or domains. Despite the promising results of E2E approaches, there are still some challenges that need to be addressed in E2E speech translation, including data scarcity, model complexity, and performance limitations. Data scarcity arises from a lack of large, high-quality datasets of speech and text for many language pairs, hindering the training of E2E models for effective generalization. Additionally, the complexity of E2E models poses challenges in both training and deployment phases, while their performance still lags behind that of traditional cascaded systems due to difficulties in accurately modeling the complex relationships between speech and text. Addressing these challenges is essential for enhancing the effectiveness and applicability of E2E speech translation models in practical settings.

The purpose of this study is twofold. Firstly, it aims to provide a nuanced understanding of how E2E and cascade models perform across key criteria essential for speech-to-text translation (STT) systems. These criteria include but are not limited to accuracy, computational efficiency, adaptability to diverse linguistic inputs and accents, scalability, resource utilization, and user experience. Secondly, by employing a multicriteria evaluation methodology, this research seeks to offer actionable insights for researchers, developers, and industry practitioners. The findings from this comparative analysis will contribute to the advancement of STT technologies by identifying areas for improvement and optimization in both E2E and cascade models. The outcomes of this study are anticipated to benefit a wide range of researchers involved in language technology research and development. These insights can inform decision-making processes related to the selection and implementation of STT models based on specific use cases, linguistic contexts, and performance requirements. Through this work, we aim to contribute meaningfully to the discourse on STT models, paving the way for enhanced cross-lingual communication.

Following the introduction, this paper is structured into several key sections that delve into a comprehensive analysis of E2E and cascade STT models using a multicriteria approach. In this section 2, we outline our approach to comparing the two models using a multicriteria evaluation framework, specifically the weighted scoring method (WSM). We define the evaluation criteria, assigning weights to reflect their

relative importance in the comparison process. This section elucidates the rationale behind our choice of criteria and the methodology employed for evaluating the models. Moving forward, section 3 details the criteria evaluation and analysis process, where each model is assessed based on the predefined criteria using the WSM. We present the numerical weights assigned to each criterion and calculate the overall scores for both models. Tables and visualizations are provided to summarize the evaluation results, facilitating a comprehensive understanding of the comparative analysis. In section 4, we analyze the comparative analysis results between the E2E and cascade models. We delve into the strengths and weaknesses of each model based on the evaluation criteria, identifying any significant differences in performance, usability, or adaptability. Finally, we provide our conclusions derived from our comparative study, offering valuable insights and directions for future research in STT technologies.

2. COMPARISON OF CASCADE-STT AND E2E-STT MODEL

2.1. Advantages of cascade speech-to-text translation models

The most significant advantages of cascade STT models is their modular design. This design approach divides the process into distinct ASR and MT components. Each component can be independently optimized and improved, allowing developers to use the best available technologies for both speech recognition and translation. This modularity also makes it easier to update or replace individual components without overhauling the entire system, providing flexibility in system maintenance and upgrades [1], [2]. Cascade models facilitate better error isolation and management. Since the ASR and MT stages are separate, it is easier to identify where errors occur. If there is a drop in performance, developers can pinpoint whether the issue lies in the speech recognition phase or the translation phase. This clear separation allows for more targeted troubleshooting and refinement of each component. For example, improving the ASR accuracy can be focused on independently enhancing the MT quality, making it easier to manage and rectify specific issues [3]. The flexibility of cascade models is another key advantage. Different language pairs and application scenarios might require tailored approaches in ASR or MT. Cascade models allow developers to mix and match different ASR and MT systems to best suit the specific needs of the application. This adaptability extends to handling various languages and dialects more effectively by using specialized components for each task. Le *et al*. [4] have studied the benefits of using separate ASR and MT components to optimize for different languages and scenarios, thereby enhancing flexibility and adaptability in system design. Moreover, developers can quickly incorporate advancements in ASR or MT technology into the existing system, ensuring that the translation system remains up-to-date with the latest research and development.

Cascade models benefit from the extensive research and development that has been conducted over the years in both ASR and MT fields. Each component has a robust body of existing models, datasets, and methodologies that can be leveraged to build effective systems. Li and Niehues [5] explores how pre-trained ASR and MT models can be integrated into cascade systems to quickly leverage the latest advancements in these fields, ensuring up-to-date performance and adaptability. This established foundation allows for quicker deployment and more reliable performance, as the technologies used in each stage have been tested and refined through years of academic and commercial research. In scenarios where resources are limited, cascade models can be more cost-effective. Training a full E2E model can be resource-intensive, requiring large datasets and significant computational power. Cascade models, on the other hand, can utilize pre-existing ASR and MT models, reducing the need for extensive retraining. This approach is particularly advantageous for low-resource languages or applications where obtaining sufficient training data is challenging. Kozhirbayev and Islamgozhayev [6] have demonstrated how leveraging pre-existing ASR and MT models can facilitate effective speech translation systems for specific language pairs, such as Kazakh to Russian, without the high costs associated with developing new E2E models. The study highlights the ability to incorporate advancements in ASR and MT independently, ensuring that the system remains up-to-date with the latest technologies. Another relevant study, by Zhang *et al.* [7] have explored the challenges and strategies for optimizing neural MT systems under low-resource conditions. The findings suggest that using existing components in a cascade approach can be more efficient and cost-effective compared to training full E2E models from scratch, particularly in low-resource settings. By leveraging existing components, organizations can deploy effective translation systems without the high costs associated with developing and training new E2E models. Cascade models can be particularly effective in handling complex scenarios where the translation process benefits from intermediate text representations [8]. For instance, in professional translation services, having an intermediate text allows for additional processing steps such as manual corrections or domain-specific adjustments before the final translation. This intermediate step can be critical for ensuring the highest accuracy and quality in translations, especially in specialized fields like legal or medical translations.

2.2. Challenges and limitations of cascade speech-to-text translation models

The primary challenges of cascade STT models is error propagation [3], [9]. In a cascade system, the output of the ASR component is fed directly into the MT component. Any errors in the ASR output, such as misrecognized words or phrases, are propagated to the MT stage, potentially compounding errors and degrading the overall translation quality. This issue is particularly problematic when dealing with noisy environments or speakers with heavy accents, as the initial transcription inaccuracies can lead to significantly flawed translations. Cascade models often experience higher latency and longer processing times compared to E2E models [10], [11]. Since the ASR and MT components operate sequentially, the total processing time is the sum of the time taken by each component. This sequential processing can result in delays, making cascade models less suitable for real-time applications where quick response times are critical, such as live translation services or interactive voice-activated systems.

Integrating and maintaining separate ASR and MT systems can be complex and resource-intensive. Each component requires its own set of training data, optimization, and tuning, which can increase the overall development and maintenance workload. Moreover, ensuring that the outputs of the ASR system are compatible with the inputs expected by the MT system can involve additional preprocessing steps, adding further complexity to the pipeline. This integration challenge can slow down development cycles and complicate system updates. Tran *et al.* [12] point out that the distinct training requirements for ASR and MT systems can result in a mismatch between the components, further complicating the integration process. This research underscores the resource-intensive nature of maintaining separate systems and the challenges associated with ensuring seamless compatibility between them. Cascade models might suffer from inconsistencies between the language models used in the ASR and MT components. The ASR system is typically optimized to transcribe spoken language accurately, while the MT system is optimized to translate written text. These differing optimization goals can lead to mismatches where the transcribed text might not be in the optimal form for translation [13]. For example, spoken language often includes disfluencies, colloquialisms, and informal speech patterns that may not translate well if the MT system is not adequately adapted to handle these features. Developing effective cascade STT models requires extensive datasets for both ASR and MT components. Gathering and annotating large volumes of high-quality speech data for ASR, as well as parallel text corpora for MT, can be resource-intensive and costly [6]. This challenge is exacerbated for low-resource languages, where available data may be sparse or of poor quality. Additionally, the computational resources required to train and run these models can be significant, posing further constraints for smaller organizations or those with limited access to advanced hardware.

Maintaining and scaling cascade STT models can be challenging due to the need for continuous updates and improvements in both ASR and MT components. As new linguistic data becomes available or as the models need to adapt to new domains and use cases, both components must be re-evaluated and potentially re-trained. Sperber *et al*. [14] discusses the complexities involved in maintaining and updating cascade models. It emphasizes that separate training and optimization for ASR and MT components can be demanding, and ensuring compatibility between these components often requires additional preprocessing steps. Zhang *et al.* [15] points out that the need for continuous updates to incorporate new data and adapt to different domains increases the complexity and resource requirements of these models. This ongoing maintenance effort can be labor-intensive and requires specialized expertise in both speech recognition and MT. Adapting cascade models to specific domains or contexts can be difficult. Each component must be individually fine-tuned to handle domain-specific terminology and context, which can be particularly challenging if the ASR and MT systems were initially trained on general-purpose data. Tran *et al.* [12] discusses how differences in linguistic style and punctuation between spoken and written domains pose challenges for cascade models. The study highlights that without fine-tuning the models on in-domain data, the ASR and MT components may struggle to handle domain-specific inputs and outputs effectively, leading to reduced accuracy. Another study by Zhao *et al.* [16] addresses the difficulties in maintaining high translation accuracy when adapting ASR and MT components to specific domains. The study highlights the challenges of handling domain-specific terminology and context, which can significantly affect the performance of the models. This lack of domain adaptation can lead to lower accuracy and relevance in specialized applications, such as medical or legal translations, where precise terminology and context are crucial.

2.3. Advantages of end-to-end-speech-to-text translation models

One of the primary benefits of E2E-STT models is integrated learning [7], [17], [18]. In E2E-STT models, both ASR and MT tasks are handled by a single, cohesive model. This integration allows the model to directly learn the mapping from speech in one language to text in another without the need for intermediate representations typically used in cascade models. This holistic approach enables the model to optimize both ASR and MT tasks simultaneously, leading to potentially higher translation accuracy and consistency. The

integrated model can leverage the intricacies of the acoustic input directly in the translation process, fostering a deeper understanding of the source language's nuances as they are spoken. E2E-STT models significantly simplify the translation pipeline [7], [19]. Traditional cascade models require separate processing stages for ASR and MT, each with its complexities and potential points of failure where errors can propagate from one stage to the next. In contrast, E2E-STT models streamline these processes into a single flow, reducing the complexity of the system. This simplification not only makes the system easier to manage and maintain but also reduces the latency involved in processing. By eliminating the need to first transcribe speech to text and then translate the text into another language, E2E-STT models can provide faster response times, making them ideal for real-time applications like live translation and interactive language translation tools. Another significant advantage of E2E-STT models is their improved ability to handle contextual information and long-term dependencies [13], [20]. Because these models process the input speech directly to the output text, they maintain a more coherent flow of information throughout the translation process. Latif *et al.* [21] discussed how E2E-STT models, particularly those using transformer architectures, can better handle long-term dependencies and contextual information. This capability allows E2E-STT models to better preserve the context over long stretches of speech, which is often a challenge in traditional models where the context can get lost between the ASR and MT stages. Enhanced context handling ensures that the translations are not only accurate on a word-by-word basis but also coherent and contextually appropriate over entire conversations or speeches. This is particularly beneficial in scenarios involving complex dialogues or technical discussions where maintaining context is crucial for understanding the intended meaning. E2E-STT models mitigate the issue of error propagation that is prevalent in cascade models. In cascade models, errors made in the ASR phase are carried over and potentially amplified in the MT phase, leading to degraded translation quality. E2E-STT models, however, process the speech signal in a unified manner, reducing the chances of such error accumulation. This holistic approach ensures that errors are less likely to propagate through the system, thereby enhancing overall translation quality [14].

Training E2E-STT models as a single unified system offers several efficiency advantages over traditional cascaded models that separately train ASR and MT components. In E2E-STT models, both ASR and MT tasks are optimized simultaneously using shared parameters and unified loss functions. This unified approach can lead to more coherent training dynamics and potentially better generalization across tasks, as the entire model is optimized towards a common goal rather than optimizing individual components that may not align perfectly with each other. The cross-modal progressive training strategy discussed in various studies highlights that E2E-STT models benefit from a training approach that can utilize data more efficiently. This strategy can lead to better performance, especially in multilingual settings where data for certain language pairs might be scarce [17]. By training the model in an E2E manner, you can exploit the natural relationships between the speech recognition and translation tasks, potentially improving the overall efficiency and effectiveness of the model. These models also tend to reduce the complexity involved in managing multiple separate systems, thereby simplifying the training pipeline. Additionally, the shared learning process can accelerate improvements, as enhancements in the model's capabilities directly benefit both the recognition and translation tasks [11]. E2E-STT models can handle rare words and phrases more effectively because they are trained on the direct mapping from speech to text. This direct training allows the model to learn specific pronunciation patterns and linguistic nuances that might be lost in a two-step cascade process. This is particularly useful for languages with high lexical variety or specialized vocabularies, ensuring more accurate translations even for less common terms [7]. For languages or dialects where certain phrases or terminology are rarely used or documented in written form, E2E models trained on diverse and comprehensive speech datasets can provide more accurate translations. They manage this by leveraging the full context of the spoken language, which includes intonation, emphasis, and other speech-specific characteristics that are often indicators of meaning and are not typically available in text-based training data.

E2E-STT models are known for their scalability and adaptability, which are crucial attributes, especially when dealing with multiple languages and dialects. The architecture of these models allows for the addition of new languages or dialects by simply retraining with appropriate data. This eliminates the need to individually adjust multiple components within the system, streamlining the adaptation process and enhancing the model's ability to manage multilingual content effectively. E2E-STT models employ a unified architecture that can be efficiently scaled to accommodate new languages or dialects. This is facilitated by the model's ability to learn directly from speech to text, leveraging shared encoder and decoder components across different languages. Such an approach not only simplifies adding new languages but also enhances the model's performance through transfer learning, where knowledge from one language can aid in processing others. This is particularly beneficial for low-resource languages, where data scarcity often poses a significant challenge [22]. Moreover, the adaptability of E2E-STT models is further supported by their design for easy integration with pre-trained components, which can be fine-tuned on specific tasks or languages, thereby accelerating the training process and improving the model's effectiveness across diverse linguistic datasets [23].

2.4. Challenges and limitations of end-to-end-speech-to-text translation models

One of the primary challenges facing E2E-STT models is data scarcity, particularly for low-resource languages. Unlike major languages that have vast amounts of available training data, low-resource languages suffer from a lack of sufficient audio recordings and corresponding translations needed for training robust models. This scarcity impedes the model's ability to learn effective translations and can lead to poorer performance. To address this issue, researchers often resort to techniques such as transfer learning, where a model trained on high-resource languages is adapted to work with less common languages. Additionally, data augmentation methods such as synthetic speech generation and simulating varied acoustic environments can help enrich the training datasets, providing more examples for the model to learn from [24]. E2E-STT models also demand significant computational resources for both training and deployment. These models, particularly those using advanced architectures like transformers, require extensive processing power due to the complexity of their neural networks and the large amounts of data they process. Training these models involves high-dimensional optimizations over millions of parameters, often necessitating powerful GPUs or TPUs and substantial memory capacity. This high computational demand can make it challenging for smaller organizations or researchers with limited access to resources to develop and train E2E-STT models. For deployment, particularly in real-time applications, the computational requirements also pose a barrier, as efficient processing and minimal latency are critical [11], [17], [25]. Another critical challenge in E2E-STT models is the risk of error propagation within a single integrated model. Unlike cascade models, where errors can be isolated and addressed in separate stages, E2E-STT models process speech directly into text through a unified architecture. This integration means that errors introduced at any point in the speech recognition process can directly affect the translation quality. Misrecognition or misinterpretations by the encoder can lead to incorrect translations that are hard to correct without an intermediate correction stage. To mitigate these issues, sophisticated training techniques such as adversarial training, which introduces potential errors during training to improve the model's resilience, and reinforcement learning strategies, which optimize the model's decisions in complex environments, are employed. Additionally, incorporating robust feedback mechanisms to refine model outputs based on real-world usage can help improve accuracy over time [11], [17].

3. METHOD

In the context of comparing E2E and cascade STT models, the multicriteria approach ensures a balanced assessment that considers various aspects of model performance. The WSM [26] is a popular technique within the multicriteria approach. WSM involves assigning weights to each criterion based on their relative importance and then calculating a weighted sum of the scores for each model. This method provides a systematic and objective way to aggregate the performance metrics into a single composite score, facilitating a clear comparison between the models. WSM involves assigning weights to each criterion based on their relative importance and then calculating a weighted sum of the scores for each model. This method provides a systematic and objective way to aggregate performance metrics into a single composite score, facilitating a clear comparison between the models. The process begins with the selection of key performance criteria, followed by the assignment of weights to reflect their relative importance. Each model is then evaluated and scored against these criteria, with scores derived from empirical performance data or experimental results. The scores are multiplied by their respective weights to obtain weighted scores, which are then summed to produce a total score for each model. The model with the highest total score is considered the better-performing model according to the chosen criteria and weights. To proceed, the next section will define the specific criteria used in this multicriteria evaluation, detailing their significance and the rationale behind their selection.

3.1. Criteria selection

The selection of comparison criteria is guided by the shared attributes among speech translation models. We outline the key criteria that must be considered when opting to choose between the use of the cascade-STT model or E2E-STT speech translation model:

a. C1=translation accuracy: this criterion evaluates how accurately the speech is translated into text in the target language. It includes aspects such as word choice, grammar, and overall fidelity to the original speech.
b. C2=model complexity: this criterion assesses the complexity of the models used in E2E speech translation compared to cascade models. It includes considerations of computational requirements, training time, and model architecture complexity.
c. C3=latency and real-time performance: this criterion examines the speed at which the translation is performed, especially in real-time applications. Lower latency indicates faster and more efficient translation, which is crucial for applications like live captioning or instant translation services.

d. C4=training data requirements: this criterion considers the amount and quality of training data needed for E2E systems compared to cascade models. It evaluates how well each approach handles data scarcity or variability in training datasets.
e. C5=code-switching and multilingual capabilities: this criterion assesses how well the systems handle code-switching or multilingual speech inputs. It examines the ability of the models to accurately translate speech that contains multiple languages or dialects seamlessly.
f. C6=robustness to noise and distortions: this criterion evaluates the model's performance in noisy environments or when dealing with speech distortions. It assesses the robustness of the models in maintaining translation accuracy under challenging acoustic conditions.
g. C7=resource efficiency: this criterion considers the resource efficiency of each approach, including memory usage, processing power, and energy consumption. It examines how efficiently the models operate in resource-constrained environments.
h. C8=adaptability and customization: this criterion assesses the ease of adapting and customizing the models for specific languages, dialects, or domains. It includes considerations of transfer learning, fine-tuning, and model adaptability.
i. C9=scalability and generalization: this criterion evaluates how well the models scale with increased data or language complexity. It assesses their ability to generalize to new languages or dialects beyond the training dataset.
j. C10=context awareness: evaluate how well the models incorporate contextual information from the speech input to improve translation accuracy and understanding.
k. C11=speaker adaptation: assess the ability of the models to adapt to different speakers' accents, speech patterns, and individual characteristics for personalized translations.
l. C12=error handling and correction: evaluate the mechanisms or strategies used by each model to detect and correct errors in the translated text, such as grammatical errors or mistranslations.
m. C13=domain adaptation: evaluate the models' ability to adapt to specific domains or specialized vocabularies, such as technical terms, medical jargon, legal language.
n. C14=long-term dependencies: examine how effectively the models handle long-term dependencies in speech, such as maintaining context over longer utterances or conversations**.**

3.2. Scores definition

The implementation of the WSM involves creating a multi-criteria matrix where the columns correspond to speech translation models and the rows correspond to the criteria, each assigned a specific weight. The scores for each criterion are derived from the detailed comparisons provided in the earlier sections. For evaluating the performance of E2E and cascade models across the different criteria we use five scores, defined as:

a. Score 1: the model performs significantly below expectations. It fails to meet the basic requirements for the criterion, demonstrating major deficiencies. For instance, in terms of translation accuracy, this would mean numerous errors in grammar, word choice, and overall comprehension. In terms of latency, it would indicate a very high delay, making real-time applications unfeasible.
b. Score 2: the model performs below average but meets minimum acceptable standards. It has several notable weaknesses and only partially satisfies the criterion. For example, it produces translations that are somewhat understandable but contain frequent mistakes. In terms of noise robustness, it might struggle considerably in moderately noisy environments.
c. Score 3: the model performs at an average level, adequately meeting the criterion with some minor issues. It is acceptable for general use but not exceptional. For example, translation accuracy would be generally reliable with occasional errors. Latency would be noticeable but not excessively disruptive for real-time applications.
d. Score 4: the model performs above average, exceeding expectations for the criterion with only minor shortcomings. It delivers strong results with few errors or issues. For instance, translations are mostly accurate and clear, and latency is low enough for smooth real-time use. Robustness to noise would be effective in most environments, with only slight degradation in performance.
e. Score 5: the model performs exceptionally well, fully meeting or exceeding the criterion in all respects. It demonstrates superior performance with minimal to no issues. For example, translation accuracy would be very high with rare errors, and latency would be minimal, making it highly suitable for real-time applications. Noise robustness would be excellent, maintaining high performance even in very noisy environments.

3.3. Weights assignment

For the evaluation of E2E and cascade models based on the selected criteria, we have assigned weights to each criterion. These weights reflect the prioritization of each criterion, ensuring that critical

aspects such as translation accuracy, latency, and robustness to noise are given higher importance in the evaluation process. The following weights have been assigned, along with the reasons for each assignment:

a. C1=translation accuracy (weight=10): translation accuracy is paramount because the primary goal of the models is to produce accurate translations. High translation accuracy ensures that the intended meaning and nuances of the original speech are preserved, making this criterion crucial.
b. C2=model complexity (weight 5): while important, model complexity is secondary to performance metrics such as accuracy and latency. A less complex model that performs well is preferable, but high complexity can be justified if it significantly enhances performance.
c. C3=latency and real-time performance (weight 10): low latency is essential for real-time applications like live captioning and instant translation services. High latency can disrupt user experience, making this criterion equally important as translation accuracy.
d. C4=training data requirements (weight 10): the effectiveness of the models can be heavily influenced by the amount and quality of training data. Models that perform well with less or lower-quality data are more versatile and practical, especially in data-scarce environments.
e. C5=code-switching and multilingual capabilities (weight 5): handling code-switching and multilingual inputs is important but typically less critical than core performance metrics. This criterion is essential for models used in multilingual environments.
f. C6=robustness to noise and distortions (weight 10): robustness in noisy environments is vital for practical use, as real-world conditions often include background noise and speech distortions. High robustness ensures consistent performance.
g. C7=resource efficiency (weight 5): efficient use of resources such as memory and processing power is important for deploying models on devices with limited capabilities. However, it is not as critical as core performance metrics like accuracy and latency.
h. C8=adaptability and customization (weight 5): the ability to adapt and customize models for specific languages, dialects, or domains is valuable but secondary to fundamental performance aspects. Customization enhances the utility of models in specialized applications.
i. C9=scalability and generalization (weight 9): scalability and the ability to generalize to new languages or dialects are crucial for extending the use of models beyond their initial training. High scalability indicates the model's robustness and versatility.
j. C10=context awareness (weight 10): incorporating contextual information improves translation accuracy and understanding, making this a highly important criterion. Context-aware models can produce more accurate and coherent translations.
k. C11=speaker adaptation (weight 5): adapting to different speakers' accents and speech patterns is important for personalized translations. However, it is considered less critical than core performance metrics like accuracy and latency.
l. C12=error handling and correction (weight 10): effective mechanisms for detecting and correcting errors are essential for maintaining high-quality translations. This criterion ensures reliability and consistency in the translated text.
m. C13=domain adaptation (weight 5): the ability to adapt to specific domains or specialized vocabularies is important for applications requiring technical or specialized language. However, it is secondary to broader performance metrics.
n. C14=long-term dependencies (weight 5): handling long-term dependencies ensures coherence over longer utterances or conversations. While important, it is considered less critical than immediate performance metrics like accuracy and latency.

3.4. Weighted scores matrix

In this section, we detail the comparative evaluation of cascade-STT and E2E-STT models based on a set of predefined criteria. These criteria are weighted based on their importance to the overall effectiveness of the translation models. After assigning the scores for each model on each criterion as illistrated in Table 1, we provided in Table 2 the weighted scores for both cascade-STT and E2E-STT models, calculated by multiplying the scores assigned to each criterion by their respective weights. The final scores indicate the overall performance of each model type, with a higher score suggesting better performance relative to the weighted criteria. This quantitative assessment helps illustrate the strengths and weaknesses of each model type in a clear and structured manner, providing a basis for selecting the most appropriate model based on specific needs and conditions.

Table 1. WSM scoring matrix

	E2E-STT	Cascade-STT
C1	3	4
C2	2	3
C3	2	4
C4	3	4
C5	2	4
C6	3	2
C7	3	2
C8	4	2
C9	3	2
C10	2	4
C11	3	4
C12	3	2
C13	4	3
C14	2	4

Table 2. WSM weighted scores matrix

	Weights	Cascade-STT	E2E-STT
C1	10	30	40
C2	5	10	15
C3	10	20	40
C4	10	30	40
C5	5	10	20
C6	10	30	20
C7	5	15	10
C8	5	20	10
C9	5	15	10
C10	10	20	40
C11	5	15	20
C12	10	30	20
C13	5	20	15
C14	5	10	20
Scores	100	275	320

4. RESULTS AND DISCUSSION

The results of the multicriteria comparison between cascade-STT and E2E-STT models, as reflected by the WSM, offer a rich dataset for analysis. This evaluation spans fourteen criteria ranging from translation accuracy to long-term dependencies, offering a detailed view of each model's strengths and weaknesses. A central observation from the weighted scores matrix and radar chart (Figure 1) is the overall higher performance of E2E-STT models across most criteria. Notably, in translation accuracy (C1), E2E-STT scores significantly higher than cascade-STT, illustrating its superior capability in rendering speech into text with higher fidelity. This is paramount as it underlines the effectiveness of E2E models in understanding and translating the nuances of language, including grammar and word choice, without the intermediate steps that might introduce errors or ambiguities in cascade models. In terms of model complexity (C2) and training data requirements (C4), E2E-STT again shows an advantage, although the differences are not as stark. The relatively simpler and more data-efficient nature of E2E models may be attributed to their direct approach to translation, bypassing the need for separate models for speech recognition and translation as in the cascade approach. This streamlined architecture potentially reduces computational overhead and simplifies training processes.

However, the comparison in latency and real-time performance (C3) reveals a notable disparity favoring E2E-STT, which scores twice as high as cascade-STT. This underscores the E2E model's suitability for applications requiring real-time translation, such as live captioning or instant translation services, where quick turnaround is crucial. Conversely, the robustness to noise and distortions (C6) is one area where cascade-STT models outperform their E2E counterparts. This suggests that while E2E models excel in cleaner environments, the layered processing of cascade-STT might offer better resistance against acoustic challenges like background noise or varying speech qualities. Interestingly, adaptability and customization (C8) and domain adaptation (C13) show closer scores between the two models. These criteria are essential for applications involving specific jargon or multiple dialects. Both model types appear reasonably flexible, although neither dominates clearly, indicating a potential area for future enhancement, particularly in the development of more adaptable E2E systems. Regarding resource efficiency (C7), E2E-STT models score lower, reflecting higher resource consumption which might be a concern in resource-constrained environments. This aspect ties in with the broader trade-off between performance and operational cost, an important consideration for deploying these technologies at scale. The evaluation of long-term dependencies

(C14) again favors E2E-STT, aligning with its higher scores in context awareness (C10). This suggests that E2E models are better equipped to handle extended speech inputs, maintaining context over longer conversations which is crucial for coherent translations in complex dialogues or technical discussions. In summary, the analysis delineates the conditions under which each model excels. E2E-STT models stand out in their overall performance, particularly in accuracy, efficiency in training, and real-time applications. However, the cascade models remain relevant, especially in adverse acoustic conditions and where robustness against noise is required.

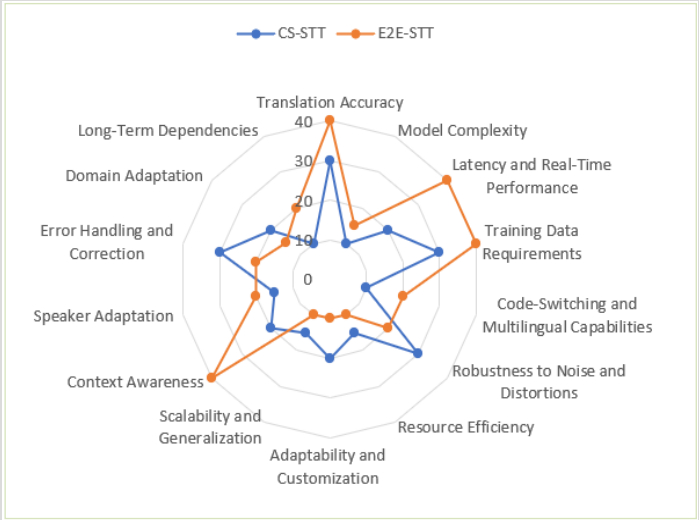


Figure 1. Cascade-STT and E2E-STT comparison models across key criteria

The findings underscore the importance of context in selecting a STT model. For applications requiring high accuracy and speed, such as live communication aids or multimedia processing, E2E-STT models are more suitable. However, for applications where robustness against noise and the need for specific linguistic customizations are critical, cascade-STT models may prove more effective. Ultimately, this research not only clarifies the operational contexts in which each model type excels but also suggests avenues for future research and development. Hybrid models that combine the rapid processing capabilities of E2E systems with the error handling and modular flexibility of cascade systems could potentially overcome the current limitations of each system type. Continued advancements in machine learning and computational hardware are expected to further enhance the capabilities of STT systems, driving innovations that could eventually merge the best features of both model types.

5. CONCLUSION

The comprehensive analysis conducted in this study meticulously delineates the distinct advantages and limitations inherent to both E2E-STT and cascade-STT models through a comprehensive multicriteria evaluation. The insights derived from the WSM evaluation reveal that while E2E-STT models exhibit superior performance in certain key areas such as translation accuracy, real-time processing, and handling of long-term dependencies, they also face significant challenges including high resource consumption and complexity in error correction. These models, therefore, shine in environments where rapid and accurate translation is paramount and where resources are abundant to support their computational demands. On the other hand, cascade-STT models, with their modular architecture, offer robustness, particularly in noisy conditions, and flexibility through easier adaptability to new languages and specialized domains. This makes them particularly valuable in settings where modular upgrades and domain-specific customizations are necessary. Despite their slower processing time and potential for error propagation between modules, their ability to isolate and correct errors in individual components remains a significant advantage. We aim that our study contributes significantly to the field of computational linguistics and offers a foundational perspective for developers and researchers aiming to optimize or choose between these two prevalent models. As speech-to-text technology continues to evolve, the insights from this analysis will help steer the development of more sophisticated, efficient, and adaptable speech translation systems.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Maria Labied	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓			
Abdessamad Belangour		✓		✓	✓					✓	✓	✓	✓	✓
Mouad Banane	✓	✓	✓	✓	✓	✓	✓	✓		✓			✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The authors confirm that the data supporting the findings of this study are available within the article.

REFERENCES

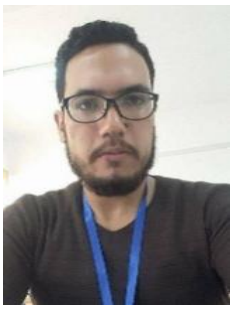
[1] C. Xu *et al*., "Recent advances in direct speech-to-text translation," *arXiv*, 2023, doi: 10.48550/arXiv.2306.11646.
[2] L. Barrault *et al*., "SeamlessM4T-Massively Multilingual & Multimodal Machine Translation," *arXiv*, 2023, doi: 10.48550/arXiv.2308.11596.
[3] N.-T. Le, B. Lecouteux, and L. Besacier, "Disentangling asr and mt errors in speech translation," *arXiv*, 2017, doi: 10.48550/arXiv.1709.00678.
[4] H. Le, J. Pino, C. Wang, J. Gu, D. Schwab, and L. Besacier, "Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 3520-3533, doi: 10.18653/v1/2020.coling-main.314.
[5] Z. Li and J. Niehues, "Efficient Speech Translation with Pre-trained Models," in *36th Conference on Neural Information Processing Systems*, 2022.
[6] Z. Kozhirbayev and T. Islamgozhayev, "Cascade Speech Translation for the Kazakh Language," *Applied Sciences*, vol. 13, no. 15, p. 8900, Aug. 2023, doi: 10.3390/app13158900.
[7] B. Zhang, B. Haddow, and R. Sennrich, "Revisiting End-to-End Speech-to-Text Translation From Scratch," *Proceedings of the 39 th International Conference on Machine Learning*, Baltimore, Maryland, USA, PMLR 162, 2022, doi: 10.48550/arXiv.2206.04571.
[8] X. Li, Y. Jia, and C. -C. Chiu, "Textless Direct Speech-to-Speech Translation with Discrete Speech Representation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10096797.
[9] N. Ruiz and M. Federico, "Assessing the impact of speech recognition errors on machine translation quality," in *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track*, 2014, vol. 1, pp. 261–274.
[10] Y. Jia *et al*., "Leveraging Weakly Supervised Data to Improve End-to-end Speech-to-text Translation," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 7180-7184, doi: 10.1109/ICASSP.2019.8683343.
[11] N. Sethiya and C. K. Maurya, "End-to-End Speech-to-Text Translation: A Survey," *arXiv*, 2023, doi: 10.48550/arXiv.2312.01053.
[12] V. A. K. Tran, D. Thulke, Y. Gao, C. Herold, and H. Ney, "Does Joint Training Really Help Cascaded Speech Translation?," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 4480–4487, doi: 10.18653/v1/2022.emnlp-main.297.
[13] T. Etchegoyhen *et al*., "Cascade or Direct Speech Translation? A Case Study," *Applied Sciences*, vol. 12, no. 3, pp. 1–24, 2022, doi: 10.3390/app12031097.
[14] M. Sperber, G. Neubig, J. Niehues, and A. Waibel, "Attention-Passing Models for Robust and Data-Efficient End-to-End Speech Translation," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 313–325, Nov. 2019, doi: 10.1162/tacl_a_00270.
[15] Y. Zhang *et al*., "Rethinking and improving multi-task learning for end-to-end speech translation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 10753–10765, doi: 0.18653/v1/2023.emnlp-

main.663.
[16] X. Zhao, H. Sun, Y. Lei, and D. Xiong, "Regularizing cross-attention learning for end-to-end speech translation with ASR and MT attention matrices," *Expert Systems with Applications*, vol. 247, p. 123241, 2024, doi: 0.1016/j.eswa.2024.123241.
[17] R. Ye, M. Wang, and L. Li, "End-to-end speech translation via cross-modal progressive training," *Computation and Language*, 2021, doi: 10.21437/Interspeech.2021-1065.
[18] H. Inaguma, K. Duh, T. Kawahara, and S. Watanabe, "Multilingual End-to-End Speech Translation," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Singapore, 2019, pp. 570-577, doi: 10.1109/ASRU46091.2019.9003832.
[19] A. Hussein, B. Yan, A. Anastasopoulos, S. Watanabe, and S. Khudanpur, "Enhancing End-to-End Conversational Speech Translation Through Target Language Context Utilization," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Republic of, 2024, pp. 11971-11975, doi: 10.1109/ICASSP48485.2024.10446102.
[20] J. Pino, Q. Xu, X. Ma, M. J. Dousti, and Y. Tang, "Self-training for end-to-end speech translation," *Computation and Language*, 2020, doi: 10.21437/Interspeech.2020-2938.
[21] S. Latif, A. Zaidi, H. Cuayahuitl, F. Shamshad, M. Shoukat, and J. Qadir, "Transformers in speech processing: A survey," *arXiv*, 2023, doi: 10.48550/arXiv.2303.11607.
[22] J. Zhao, H. Yang, E. Shareghi, and G. Haffari, "M-adapter: Modality adaptation for end-to-end speech-to-text translation," *arXiv*, 2022, doi: 10.48550/arXiv.2207.00952.
[23] C. Wang *et al*., "Fairseq S2T: Fast speech-to-text modeling with fairseq," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, 2020, pp. 33–39, doi: 10.18653/v1/2020.aacl-demo.6.
[24] A. Wu, C. Wang, J. Pino, and J. Gu, "Self-supervised representations improve end-to-end speech translation," *arXiv*, 2020, doi: 10.48550/arXiv.2006.12124.
[25] C. Han, M. Wang, H. Ji, and L. Li, "Learning shared semantic space for speech-to-text translation," *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 2214–2225, doi: 10.18653/v1/2021.findings-acl.195.
[26] M. Labied and A. Belangour, "Automatic Speech Recognition Features Extraction Techniques: A Multi-criteria Comparison," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 12, no. 8, pp. 177-182, 2021, doi: 10.14569/IJACSA.2021.0120821.

BIOGRAPHIES OF AUTHORS

Maria Labied is Ph.D. student in computer science at the Laboratory of Information Technology and Modeling (LTIM), Faculty of sciences Ben M'sik, Hassan II University, Casablanca, Morocco. Her research interests include Natural language processing, speech signal processing, and machine translation. As a researcher in this field, she has published many research papers in journals indexed in databases such as Web of Science and SCOPUS. She can be contacted at email: mr.labied@gmail.com.

Abdessamad Belangour is a full professor at the Faculty of Sciences Ben M'Sik, Hassan II University of Casablanca, Morocco. He specializes in software engineering, artificial intelligence, and data science, with a particular focus on model-driven engineering and its applications in big data, business intelligence, and cloud computing. Dr. Belangour has contributed significantly to the development of ontology-based models for data lakes and has explored the integration of semantic technologies in real-time systems. He can be contacted at email: belangour@gmail.com.



Mouad Banane is an Associate Professor at Hassan II University. Laboratory of Artificial Intelligence and Complex Systems Engineering (AICSE). He obtained his PhD at Faculty of Sciences Ben M'Sik, Hassan II University of Casablanca, Morocco. His research fields: model-driven engineering, semantic web, artificial intelligence, and big data. He can be contacted at email: mouad.banane-etu@etu.univh2c.ma.