Leveraging AI Technologies for Enhanced Multimedia Localization

Ashley Mondello amondello@languagescientific.com
Sahil Rasane srasane@languagescientific.com
Language Scientific, Boston, MA, USA

Alina Karakanta a.karakanta@hum.leidenuniv.nl
Leiden University Centre for Linguistics, Leiden University, the Netherlands

Laura Casanellas Laura@lcmt.solutions
LCTM Solutions, Dublin, Ireland

Abstract

As demand for multilingual video content rises, multimedia localization is becoming crucial for Language Service Providers (LSPs), offering revenue growth and new business opportunities. To cope with labor-intensive multimedia workflows and the rise in client demand for cheaper and faster multimedia localization services, LSPs are starting to leverage advanced AI applications to streamline the localization process. However, workflows and tools adopted by media service providers may not be suitable for LSPs, while the plethora of available solutions makes it hard for LSPs to choose the ones that most effectively optimize their workflows. In this presentation, we assess AI technologies that offer efficiency and cost reduction in the traditionally human-driven workflows of transcription, translation, voice-over (VO), and subtitling with the goal to offer insights into how an LSP can evaluate which tools work best for their processes.

1 Introduction

With the growing demand for multilingual video content as a tool for companies to enhance their global communication and engagement, multimedia localization is becoming an important growth vector for Language Service Providers (LSPs), presenting opportunities to boost revenues and expand to new business cases (Slator, 2024). There are clear challenges faced by LSPs in multimedia localization currently; the most significant are lengthy timelines, high execution costs, as well as difficulty sourcing and managing voice talent and video engineering resources. To cope with labor-intensive multimedia workflows and the rise in client demand for cheaper and faster multimedia localization services, LSPs are starting to leverage advanced AI applications to streamline the localization process. However, LSPs are often not prepared to adopt workflows and tools used by media service providers, while choosing tools that most effectively optimize their workflows is challenging due to the plethora of available solutions.

In this presentation, we assess AI solutions that offer efficiency and cost reduction in the traditionally human-driven workflows of transcription, translation, voice-over (VO), and subtitling with the goal of guiding LSPs in selecting the tools that work best for their processes. We investigate three categories of AI solutions for video localization workflows: open-source tools, commercial AI services and APIs, and dedicated video localisation platforms. Our evaluation examines tools for automatic transcription, machine translation, synthetic voices, and automatic subtitling in two high-demand language pairs: English to Chinese (Simplified) and English to Spanish (Latin American). We assess the tools based on criteria such as ease of use, cost, language availability and quality. Our analysis suggests that out-of-the-box solutions that offer easy integration into existing workflows are a good transition step towards adopting AI, especially for low/medium project volumes. The existence of an in-house development team and higher volumes may justify investing in tailored solutions. Still, the availability of languages is the most decisive factor in tool selection. We also show preliminary productivity gains when AI tools are applied in existing manual workflows. We conclude with recommendations for LSPs in selecting AI tools based on key aspects like price, volume of multimedia projects, language pairs and the existence or not of an internal development team.

2 Background

2.1 Traditional Multimedia workflows

Multimedia localization involves the adaptation of audiovisual content, such as videos, to make it accessible and relevant to different linguistic and cultural audiences. Traditionally, this process has been heavily reliant on human labor, encompassing various stages including transcription, translation, voice-over, and subtitling. Each stage requires specific skills and significant time investment, making the overall process labor-intensive and costly.

Subtitling Workflow The traditional subtitling workflow encompasses multiple stages. Initially, the process begins with transcription and time-coding of the video content to generate a script for translation. This script undergoes a thorough quality assurance review to ensure accuracy before advancing to the translation phase. Following translation, the content is carefully edited. Subsequent steps involve video engineering to format the subtitle lengths and burn the subtitles to the video. Finally, several rounds of video QA and verification are performed, culminating in the finalization of the video.

Voice-Over Workflow The voice-over workflow is equally rigorous, beginning with transcription, time-coding, and a quality assurance review to produce the final script. This is followed by translation and editing of the script. Once translated, the script proceeds to voice-over recording, accompanied by additional QA and necessary revisions. The process continues with video engineering to sync the individual audio segments to the video to ensure the audio and video are aligned. Finally, multiple stages of video QA and verification are performed, leading to the finalization of the video.

2.2 Challenges for LSPs

The traditional manual workflows present several challenges for Language Service Providers (LSPs). Firstly, the labor-intensive nature of these workflows results in high operational costs. Each stage requires specialized human resources, which increases the overall expense of the localization process. Second, due to the sequential and manual nature of the tasks, the localization process is time-consuming. Meeting tight deadlines becomes challenging, especially when handling large volumes of content or multiple language pairs simultaneously. Another challenge is resource management. Managing and coordinating the different stages of the workflow requires meticulous planning and resource allocation. The availability of skilled translators, editors, subtitlers, voice-over artists, and video engineers is critical, and any delays in one stage can impact the entire timeline. Last comes quality control, which entails ensuring consistent quality across all stages. Each step involves human intervention, which can introduce variability in the output quality. Maintaining high standards requires rigorous QA processes, further adding to the time and cost.

2.3 The Role of AI in Enhancing Efficiency

To address these challenges, the adoption of AI technologies in multimedia localization is becoming increasingly essential. AI offers several advantages that can enhance efficiency and reduce costs, first of all the automation of repetitive tasks. Tools like automatic speech recognition (ASR), automatic time-coding and machine translation (MT) can significantly reduce the time required for these tasks. In addition, AI solutions can handle large volumes of content and multiple language pairs simultaneously. This scalability is crucial for LSPs dealing with high-demand projects and tight deadlines. Consequently, by automating labor-intensive tasks, AI can significantly reduce operational costs. The reduced reliance on human resources for certain stages of the workflow allows LSPs to allocate their resources more efficiently. AI tools can also process content much faster than humans. This speed is particularly beneficial for projects with quick turnaround times, allowing LSPs to deliver localized content more rapidly.

3 Methodology

This section presents the settings to test AI tools and solutions for multimedia localisation, using workflows adopted by the company Language Scientific (LS) as a case study.

3.1 Data

To test the quality of the tools, we used previously completed multimedia projects in the life sciences domain from LS for the language pairs English to Chinese (Simplified) and English to Spanish (Latin American). These amount to several hours of content and contain videos focusing on medical topics, such as e-learning, presentations, webinars and doctor-patient discussions. Thus they contain both scripted and unscripted content, single- and multi-speaker videos and speakers with different accents. The human outputs serve as references for computing automatic quality metrics.

3.2 Tools and systems

We investigate three categories of AI solutions for video localization workflows: open-source tools (e.g. Whisper), commercial AI services and APIs (e.g. Amazon Transcribe, Google text-to-speech) and dedicated video localisation platforms (e.g. Matesub, Speechify). Our evaluation examines tools for automatic transcription with timestamp prediction, machine translation, synthetic voices, and automatic subtitling. Specifically, we assess the following tools:

- Transcription: Whisper (Radford et al., 2023), Amazon Transcribe and Matesub[1]

- Translation: Amazon Translate, ChatGPT (OpenAI, 2023), Google Translate

- Subtitling: Amazon subtitling pipeline[2], Matesub

- Voice-over: Amazon Polly[3], Google text-to-speech, Speechify[4]

3.3 Evaluation criteria

The evaluation contains the following criteria:

- Ease of use (EoU): User interface, learning curve, integration capabilities. Since ease of use is different depending on the profile and technical skills of the person operating the tool, we report ease of use for project managers and developers separately. Two project managers and two developers at LS assessed the usability of the tools as 'low', 'medium' or 'high'.

- Cost: Pricing models, total cost of ownership. Assessed as 'low', 'medium', 'high', with low pricing being most suitable for LSPs with up to 50% of revenue comprised by multimedia, medium pricing being most suitable for LSPs with between 50%-75% of revenue comprised of multimedia and high pricing being most suitable for LSPs with over 75% of revenue comprised by multimedia.

- Language coverage: Number of supported languages, dialects, and regional varieties.

- Quality: The evaluation is performed with automatic metrics and, when possible, using human ratings. The accuracy of transcription is evaluated with Word Error Rate and the translation quality using COMET (Rei et al., 2020). For voice-over, we collect human ratings from 5 native speakers on the naturalness and clarity of the generated speech. Subtitle quality, synchronization and readability is evaluated using SubER (Wilken et al., 2022), an edit-based metric which considers edits in the text, timestamps and segmentation, while we also report subtitle conformity to the formal constraints of length (42 characters per line [CPL] for Es and 16 for Zh) and reading speed (21 characters per second [CPS] for Es and 9 for Zh) (Papi et al., 2023).

4 Results

4.1 Transcription

For transcription, the tools we compared are Whisper, Amazon Transcribe and Matesub. We only tested tools that output timestamps, since these are vital for synchronization both in subtitling and

[1] https://matesub.com/

[2] https://aws.amazon.com/transcribe/subtitling/

[3] https://aws.amazon.com/polly/

[4] https://speechify.com/

	Whisper	Amazon	Matesub
EoU - Dev	High	Med	High
EoU - PM	Low	Med	High
Cost	Low	Low	Med
Lang. cov.	99	102	85
Quality			
WER ↓	**7.32**	8.38	7.80
CPL↑	63.0%	32.4%	**100%**
CPS↑	62.8%	**86.3%**	73.8%

Table 1: Evaluation of transcription tools. Ease of Use (EoU) for the developer and project manager, language coverage (Lang. cov.) in number of languages and quality scores: Word error rate (WER), percentage of subtitles conforming to the maximum length of 42 CPL and maximum reading speed of 21 CPS. Best scores in bold.

	Google	Amazon	ChatGPT
EoU - Dev	High	High	Med
EoU - PM	High	High	Low
Cost	Low	Low	Med
Lang. cov.	134	75	99
Quality			
COMET Es	**89.5**	88.3	88.6
COMET Zh	**80.3**	79.7	80.0

Table 2: Evaluation of translation tools. Ease of Use (EoU) for the developer and project manager, language coverage (Lang. cov.) in number of languages and quality scores: COMET for Spanish (Es) and Chinese (Zh). Best scores in bold.

voice-over. The evaluation is shown in Table 1. In terms of ease of use, Whisper scores high for the developer, but low for the PM. Even though it is straightforward to use by persons with programming skills, the majority of PMs may not be familiar with operating a computer terminal. Amazon and Matesub offer a friendly user interface and thus their EoU for the PM is higher. Whisper has a low cost, since it only requires a computer with some computational power to run on and no subscription. Amazon comes next, with a pay-as-you-go model, while Matesub requires a subscription with a dedicated number of minutes available per month.

When it comes to quality, Whisper has the lowest WER on LS projects (7.32), followed by Matesub (7.8) and Amazon (8.38). It is worth mentioning that the Matesub timed transcription is different than that of Amazon and Whisper in terms of form, as shown by the conformity to length (CPL) and reading speed (CPS). Matesub, being a subtitle tool, generates short segments, conforming 100% to the length constraint of 42 CPL, while the mean line length for Amazon and Whisper is 57 and 49 respectively. Generating short subtitles comes at the expense of reading speed, with Amazon having a better conformity of reading speed than Matesub (86.3% vs 73.8%). To conclude, the timed transcriptions of Matesub are more suitable for subtitling projects, while Amazon and Whisper generate longer segments, which make them ideal for voice-over projects, which need to maintain longer units to improve prosody of synthetic outputs.

4.2 Translation

Translation for transcribed video content poses challenges compared to text translation, such as oral style and partial inputs (subtitles or incomplete sentences). The evaluation for Google Translate, Amazon Translate and ChatGPT for translation is shown in Table 2. Google and Amazon score similarly in terms of EoU and cost, since they are both well integrated in most CAT tools and offer APIs or UI to obtain the translations. ChatGPT has a lower ease of use both for developer and PM, and a higher cost. It should also be noted that it is a general purpose LLM and not a dedicated translation system. While most providers are expanding their language support in MT, language availability is still higher for Google. Translation quality for the content commonly translated in LS multimedia projects, as shown by COMET, is higher for Google, followed by ChatGPT and Amazon. While all three tools produced similar quality, our evaluation determined that, currently, Google and Amazon are the most suitable options for LSPs based on their high EoU and low pricing compared to ChatGPT.

	Amazon	Matesub
EoU - Dev	High	Med
EoU - PM	Med	High
Cost	Low	Med
Lang. cov.	75	85
Quality		
SubER Es	**55.9**	59.01
CPL↑	33.5%	**97.8%**
CPS↑	66.5%	**78.5%**
SubER Zh	**82.6**	198.1
CPL↑	62.6%	**100%**
CPS↑	**98.3%**	95.3%

Table 3: Evaluation of subtitling tools. Ease of Use (EoU) for the developer and project manager, language coverage (Lang. cov.) in number of languages and quality scores: Subtitle edit rate (SubER), percentage of subtitles conforming to the maximum length of 42 CPL for Es and 16 for Zh and maximum reading speed of 21 CPS for Es and 9 for Zh. Best scores in bold.

4.3 Subtitling

The evaluation of the Amazon subtitling pipeline and Matesub is shown in Table 3. In Amazon, subtitles are generated in a two step process, combining two services; transcription with timestamps (see Sec. 4.1) and machine translation (see Sec. 4.2). They can be performed by uploading and downloading input/output files in a user interface. In Matesub, the video is uploaded in the platform and the subtitling guidelines and target languages are selected, making it easier to use by PMs who are familiar with the requirements of subtitling, but not as straightforward for developers.

In terms of subtitling quality on LS projects, Amazon has a better SubER than Matesub. The high SubER for Zh is due to the fact that LS subtitling projects allow a higher CPL than 16, which is the maximum subtitle length Matesub models are trained to produce. However, Amazon has a very low CPL conformity (33.5 vs 97.8 for Es and 62.6 vs 100 for Zh). As also noted in the results for transcription, Matesub subtitles have better conformity to the constraints of length and reading speed, and therefore the tool is more suitable for subtitling projects.

	Google	Amazon	Speechify
EoU - Dev	Med	Med	Med
EoU - PM	Low	Med	High
Cost	Low	Low	Med
Lang. cov.	58	38	130
Quality (Naturalness & clarity)			
Es-fem	3.25	3.5	**3.63**
Es-male	**4**	3.25	3.63
Zh-fem	3.25	4.5	**5**
Zh-male	3.25	-	**5**

Table 4: Evaluation of synthetic voice tools for voice-over. Ease of Use (EoU) for the developer and project manager, language coverage (Lang. cov.) in number of languages and quality scores: Averaged naturalness and clarity scores from 5 native speakers of Zh and Es for female and male voices. Best scores in bold.

4.4 Voice-over

The evaluation of Amazon, Google and Speechify synthetic voices for voice-over generation is shown in Table 4. Google and Amazon have a medium to low EoU. Voice generation is performed through an API or user interface where text is pasted. Because voice-over has to be synchronized with the video, it has to be generated sentence by sentence and not as a large chunk of text, which is time consuming for the PM. For this reason, PM's EoU is lower for Google and Amazon. Google had a demanding set up process for the API because of the modular structure of Google cloud, but once set up, it was relatively easy to use, hence the medium rating. Speechify allows for uploading a timed .srt file, which performs synchronization automatically. Speechify has also an integrated voice editor, which allows a PM to adjust the speed, prosody and synchronization of the generated voice samples.

Language coverage is an issue in voice-over, since both Google and Amazon support a limited number of languages and language varieties. In addition, very few languages have models for both female and male voices, which is often a requirement for voice-over when the persons are on screen. Such is the case with the Chinese male voice for Amazon. Chinese voices were not available in Google's GUI but could be used through the API. Another issue is that the language may be available but at a low model quality. For example, Google offers different model types: standard, neural, wavenet, studio, in an ascending order of quality.

In terms of naturalness and clarity of speech, Google scores higher for the Spanish male voice (4), while Speechify for the female (3.63). It is worth noting that this rating is higher than the rating for the human female voice from the reference project, which scored an average of 3.25. For Chinese, all participants rated the Speechify voices with the highest score in terms of naturalness and clarity (5), showing that, for some languages and project types, synthetic voices may be a feasible alternative. To conclude, for voice-over projects, language/model type availability is the most decisive factor when selecting provider. Speechify has high quality of synthetic voices and an integrated editor, while Amazon and Google can be good for occasional projects, but require video synchronization as an extra step.

5 Preliminary productivity evaluation

To evaluate productivity gains of using AI in the multimedia localization process, we conducted a series of real-life scenario tests using various AI tools. Our initial test, covered in this paper, involved subtitling and voice-over of an 11-minute video with two speakers (male, female), replacing specific steps in traditional workflows with AI tools without the integration of workflow automation. The primary goal was to assess the productivity impact of low-level AI integration for LSPs beginning their AI adoption journey.

5.1 Testing process

The tests involved replacing human-driven steps with AI tools while maintaining all quality assurance steps with human resources to ensure the highest level of quality. The replacements included:

Subtitling Workflow: 1) Replacing human transcription with Amazon Transcribe, 2) Replacing human translation with Amazon Translate.

Voice-over Workflow: 1) Replacing human transcription with Amazon Transcribe, 2) Replacing human translation with Amazon Translate, 3) Replacing voice-over recording with Amazon Polly for Spanish and Google Text-to-Speech for Chinese.

The workflows were evaluated by comparing the time and effort required for both traditional and AI-assisted processes.

5.2 Findings

The integration of AI tools resulted in significant time savings and efficiency improvements. In the **Subtitling Workflow**, the Traditional Workflow required **19** hours of human labor per language for the 11-minute video, while the AI-Assisted Workflow was reduced to **8** hours of human labor, saving 11 hours per project. Time gains were recorded in transcription, from 4 to 2 hours, in translation from 10 hours to 2 hours, while a gain from 5 hours to 4 hours was also reported in video engineering.

In the **Voice-over Workflow**, the Traditional Workflow required **24** hours of human labor per language, while the AI-Assisted Workflow was reduced to **12** hours, saving 12 hours per project. Time gains were recorded in transcription, from 4 to 2 hours, in translation from 10 hours to 2 hours, and voice-over engineering from 9 hours to 7 hours.

The evaluation revealed that AI-assisted workflows can reduce the required labor hours by over 50% in both subtitling and voice-over processes. The quality and availability of AI tools, however, vary depending on the language pair, underscoring the importance of selecting appropriate tools based on specific project requirements.

6 Recommendations for LSPs regarding AI tool selection

When deciding which AI tools to integrate into their workflows, LSPs should follow a systematic approach that takes into account the following aspects of their own production workflows: the volume of multimedia projects per year, the availability of a research and implementation budget, whether there is time to test different solutions beforehand, the main language pairs, whether there are engineers in the team who can work with open source solutions and, finally, whether the linguists assigned to multimedia projects are familiar with the tools or whether they need to be trained.

Our exploration of AI solutions for multimedia projects revealed that commercial AI services, which offer an all-in-one solution by the same provider and require medium technical skills, can be a good starting point in integrating AI and more suitable for low/medium project volumes. For high volumes, open source tools such as Whisper can prove a worthy investment, but require staff with technical skills, as well as equipment with some computational power. Tailored solutions also suit higher volumes. Multimedia platforms, such as Matesub for subtitling and Speechify for voice-over, offer high quality and low management effort, but the cost is slightly higher and linguists need to be trained in using the tools.

When selecting AI tools, LSPs should not only consider the factors previously mentioned but also assess whether their multimedia projects are suitable for AI integration. The efficiency gains from using AI in multimedia projects can be significantly influenced by the complexity of the source material. For example, videos featuring multiple speakers, extensive on-screen text, or embedded PowerPoint presentations present synchronization challenges when processed with AI tools. Furthermore, the target audience of these videos must be taken into account, especially in VO projects. Despite considerable advancements in synthetic voice technology, AI-generated voices remain distinguishable from human voices. LSPs must carefully evaluate how the intended audience might react to synthetic voices when considering AI's role in their projects.

7 Conclusion

The integration of AI technologies into traditional multimedia localization workflows offers significant advantages in terms of efficiency and cost reduction, all while maintaining high-quality standards. LSPs can harness these tools to optimize their processes and effectively meet growing client demands. However, it is important to consider the initial effort required to incorporate AI solutions into existing workflows. Depending on the chosen tools, this integration may demand varying levels of resource training, technical support, and budget allocation before realizing the anticipated time and cost savings. Ongoing research should explore a broader spectrum of AI tools and refine evaluation criteria to support comprehensive tool selection strategies for LSPs. We hope this presentation equips LSPs to navigate the evolving landscape of multimedia localization, enabling them to meet client demands with efficiency and effectiveness, while upholding high-quality standards.

References

OpenAI (2023). ChatGPT (Mar 14 version) [large language model].

Papi, S., Gaido, M., Karakanta, A., Cettolo, M., Negri, M., and Turchi, M. (2023). Direct speech translation for automatic subtitling. *Transactions of the Association for Computational Linguistics*, 11:1355–1376.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Slator (2024). 2024 Language Industry Market Report — Language AI Edition.

Wilken, P., Georgakopoulou, P., and Matusov, E. (2022). SubER - a metric for automatic evaluation of subtitle quality. In Salesky, E., Federico, M., and Costa-jussà, M., editors, *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 1–10, Dublin, Ireland (in-person and online). Association for Computational Linguistics.