MULTILINGTOOL, Development of an Automatic Multilingual Subtitling and Dubbing System

Ander Corral[1], **Xabier Sarasola**[1], **Iker Manterola**[2], **Josu Murua**[2], **Itziar Cortes**[2], **Igor Leturia**[1], **Xabier Saralegi**[1]
[1]Orai NLP Technologies/Elhuyar
{a.corral,x.sarasola,i.leturia,x.saralegi}@orai.eus
[2]Elhuyar Foundation
{i.manterola,j.murua,i.cortes}@elhuyar.eus

Abstract

In this paper, we present the MULTILINGTOOL project, led by the Elhuyar Foundation and funded by the European Commission under the CREA-MEDIA2022-INNOVBUSMOD call. The aim of the project is to develop an advanced platform for automatic multilingual subtitling and dubbing. It will provide support for Spanish, English, and French, as well as the co-official languages of Spain, namely Basque, Catalan, and Galician.

1 Introduction

Over the past two decades, the European audiovisual industry has undergone significant transformation due to advancements in information and communication technologies and shifts in consumer behavior, leading to a market predominantly controlled by a few large corporations. These changes have raised concerns regarding the sustainability of content production by European entities and the maintenance of the continent's cultural diversity. However, artificial intelligence (AI) provides a promising solution for smaller firms, enabling them to access enhanced subtitling and dubbing services in various languages. This technology helps expand their reach and visibility across Europe, thus bolstering the industry's diversity and resilience.

The MULTILINGTOOL project (CREA-PJG/101093511), led by the Elhuyar Foundation and financed by the CREA-MEDIA2022-INNOVBUSMOD call, commenced in 2022 and is scheduled to conclude in March 2025. Its primary objective is to develop an innovative automatic subtitling and dubbing platform specifically designed for the audiovisual sector. This platform can perform **automatic dubbing in multiple languages**, including English, French, Spanish, Basque, Galician, and Catalan. It integrates three core technologies: Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS). The platform also features **customizable dubbing voices**, allowing users to tailor the voices to enhance their final audiovisual content. Additionally, it includes a web-based interface that enables both users and content companies to efficiently **manage and review** multilingual audiovisual content, which includes subtitles, translations, and dubbing.

2 Platform specifications

A modular architecture has been developed to support multilingual transcription, neural machine translation, and personalized speech synthesis use cases. Emphasizing modularity is key to achieving robust and easily adaptable software.

2.1 ASR module

We have fine-tuned Whisper-based (Radford et al., 2023) systems for all languages involved in the project, except for English as it already obtains competitive enough results. Data augmentation techniques have been used to enhance the adaptability and robustness of the ASR module against a diverse range of acoustic phenomena encountered in real-world scenarios, such as speed and volume perturbations and a diverse set of out-of-speech signals and artifacts, including music, background noise, chatter, telephone codecs, and reverberation. We opted for the small version of Whisper to

strike a balance between performance and resource utilization. We tested our fine-tuned systems on the FLEURS standard benchmark (Conneau et al., 2022). The quality of the transcriptions has been measured in terms of word error rates (WER) as shown in Table 1.

Language	FLEURS
es	7.31
en	6.53
fr	12.37
ca	10.75
gl	14.49
eu	15.34

Table 1: WER results of the Whisper-based fine-tuned systems for the multilingual ASR module.

2.2 NMT module

A multilingual NMT module involving the six languages of the project was developed. Due to the lack of sufficient volume of training samples for some of the translation directions, a Spanish-centric pivoted translation approach has been considered, where translating from one language to another is done via Spanish. Systems were trained using the Transformer (Vaswani et al., 2017) base architecture. Data augmentation techniques were applied for general system robustness against ASR module's casing and punctuation errors and input perturbations. Additionally, we adapted the systems for an informal/speaking register leveraging back-translation for more in-domain data. All the systems were evaluated on the Flores200 test set by using the BLEU metric as reported in Table 2.

2.3 TTS module

A multispeaker cross-lingual speech synthesis system has been created to use custom speakers for dubbing in multiple languages. Our system is based on Fastpitch (Łańcucki, 2021) and we added a language embedding to make a multispeaker multilingual Fastpitch. The language embedding is added to the input of the encoder similar to the speaker embedding in the multispeaker Fastpitch. To evaluate the model we selected 30 sentences in English and we synthesized them with speakers with recordings in different languages in a cross-lingual way. For the English speaker, we translated the 30 sentences to French and we made the cross-lingual synthesis in French. We evaluated the resulting speech with neural network based Mean Opinion Score (MOS) and Speaker Encoder Cosine Similarity (SECS). The results are shown in Table 3.

	Flores200	
Language pair	$\rightarrow$	$\leftarrow$
es-en	26.7	24.9
es-fr	26.3	22.8
es-gl	13.4	18.3
es-ca	22.7	24.1
es-eu	21.6	23.6

Table 2: BLEU scores for all the translation directions developed for the multilingual NMT module.

lang	ref	cross-lingual voice (en*)	
	MOS	MOS	SECS
ca	3.21±0.11	3.82±0.09	0.39
es	4.11±0.08	4.20±0.05	0.36
en	4.36±0.04	3.46±0.16	0.59
eu	3.42±0.11	3.95±0.08	0.35
fr	3.13±0.03	3.90±0.07	0.33
gl	3.84±0.10	4.15±0.07	0.68

Table 3: MOS and SECS scores of the multispeaker cross-lingual speech synthesis system. *The English speaker synthesized French sentences.

References

Conneau, Alexis, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *arXiv preprint arXiv:2205.12446*.

Łańcucki, Adrian. 2021. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592. IEEE.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.