STREAMS: Live Streaming and Micro-batching with Automated Multilingual Services

Aitor Álvarez[1], Thierry Etchegoyhen[1], Joaquín Arellano[1], Víctor Ruíz[1], Haritz Arzelus[1], David Ponce[1,2], Ander González-Docasal[1,3] and Harritxu Gete Ugarte[1,2]

[1] *Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Donostia-San Sebastián, 20009, Spain*

[2] *University of the Basque Country UPV/EHU*

[3] *Aragon Institute for Engineering Research, University of Zaragoza, 50009 Zaragoza, Spain*

Abstract

We present STREAMS, a batch and streaming platform which integrates and manages AI services for rich transcription, translation, voice synthesis and audio description. Its main goal is to provide a unified platform to enhance the processes and products of companies working in multiple sectors. STREAMS has been successfully deployed in real-life batch and multilingual streaming scenarios, offering a rich array of multilingual services with innovative methods, in particular to enhance portability, responsiveness and output readability.

Keywords

Streaming, Micro-batching, Rich Transcription, Machine Translation, Speech Synthesis, Subtitling

1. Introduction

Artificial Intelligence (AI) solutions have risen across multiple industries, in large part due to the advent of high-quality language processing models based on deep neural networks (DNN). This has notably been the case for Automatic Speech Recognition (ASR) and Neural Machine Translation (NMT), building on the power of modern DNN models and computing infrastructure [1, 2].

Despite significant progress, several important challenges remain. For instance, providing responsive systems that can accurately transcribe and/or translate continuous input streams is notably difficult, given the need to balance prompt responses and the necessary processing time to provide high-quality output. Additionally, batch content processing faces the multitude of scenarios which require specific processes to deliver usable output, such as subtitling, audio-description, or adapted speech synthesis, among others. Most current solutions are either oriented to specific scenarios, or provide minimal, and in some cases no adaptation, to use-cases that warrant dedicated processes.

To address these shortcomings, we present STREAMS, a batch and streaming platform which integrates and manages AI services for rich transcription, translation, voice synthesis and audio-description. The main goal of the platform is to support and enhance the processes and products of companies working in different sectors, e.g., audiovisual, media, animation and communication. This platform supports the combination of its technological modules into powerful multilingual pipelines for subtitling, speech-to-text, translation and/or synthetic-voice driven applications, both in batch and streaming modes.

The solution has been successfully deployed in real-life batch and multilingual streaming scenarios, offering a rich array of multilingual services to companies in varied sectors. The solution has been mainly tested for AI services in English, Spanish, French and Basque, although it can easily be extended to other languages. Several innovative methods have been developed within STREAMS, in particular for streaming and automatic subtitling, with important advances in portability and subtitle readability.

The solution has been developed within the applied research project STREAMS, partially supported by the Department of Economic Development of the Basque Government. The project started in May 2021 and finalised in December 2023, and was carried out by the following consortium: Haiko[1] (project coordinator), Jarkatza[2], MondragonLingua[3], Goiena[4], Mixer[5], Noticias de Gipuzkoa[6], Onda Vasca[7] and Vicomtech[8].

SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations, June 19-20, 2024, A Coruña, Spain.

aalvarez@vicomtech.org (A. Álvarez); tetchegoyhen@vicomtech.org (T. Etchegoyhen); jarellano@vicomtech.org (J. Arellano); vruiz@vicomtech.org (V. Ruíz); harzelus@vicomtech.org (H. Arzelus); adponce@vicomtech.org (D. Ponce); agonzalezd@vicomtech.org (A. González-Docasal); hgete@vicomtech.org (H. G. Ugarte)
0000-0002-7938-4486 (A. Álvarez); 0000-0001-7253-1693 (T. Etchegoyhen); 0000-0003-3585-6346 (A. González-Docasal)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

[1] https://haiko.es/
[2] https://jarkatza.com/
[3] https://www.mondragonlingua.com/en
[4] https://goiena.eus/
[5] https://mixer.eus/
[6] https://www.noticiasdegipuzkoa.eus/
[7] https://www.ondavasca.com/
[8] https://www.vicomtech.org/en

CEUR-WS.org/Vol-3729/d19_rev.pdf

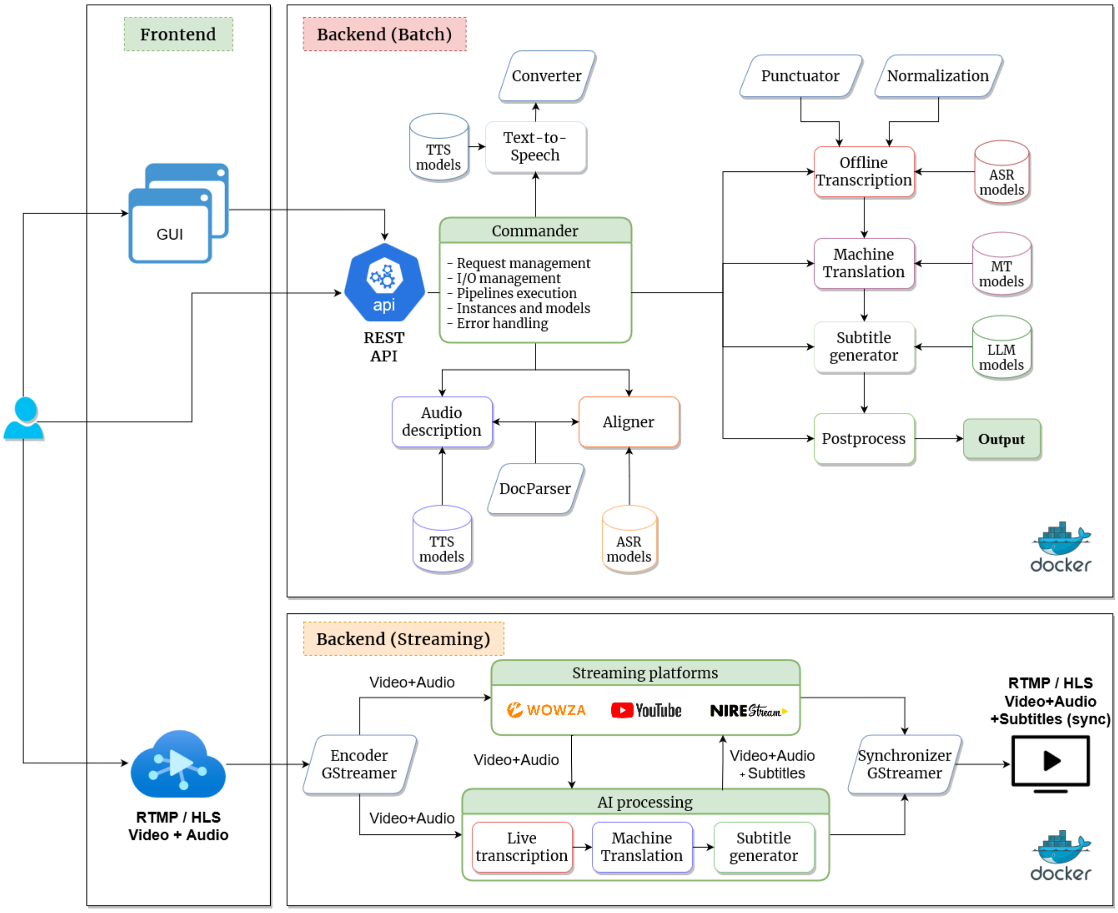


Figure 1: Core architecture of the STREAMS solution.

2. STREAMS platform

The core architecture of STREAMS is shown in Figure 1. The solution can operate in both batch and streaming modes, via independent deployments.

STREAMS consists of the following main components:

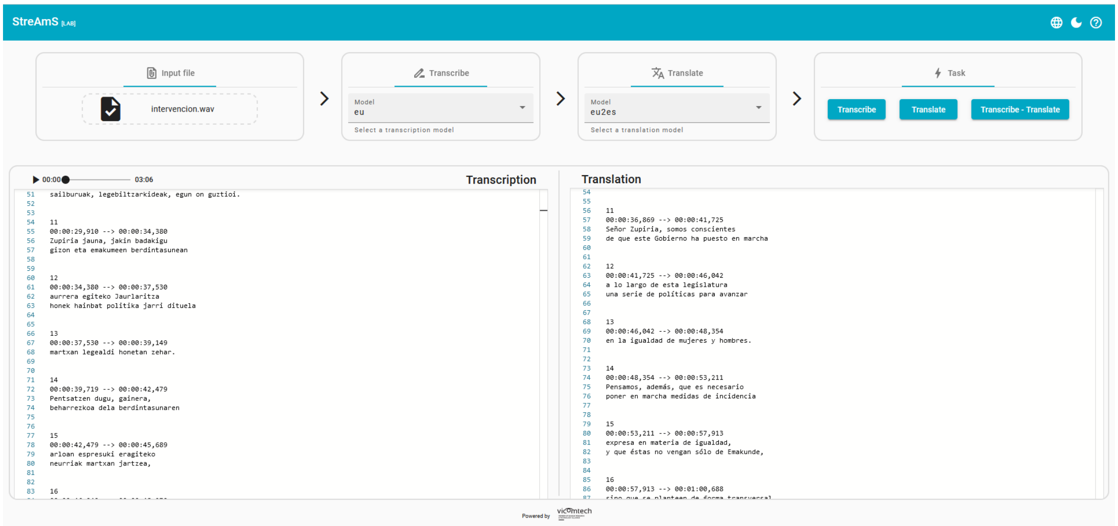
- Two web-based graphical user interface (GUI) frontends, one for transcription and translation purposes, and the other for speech synthesis. Both GUIs operate with the same backend in batch mode.
- A main REST API, which exposes the functionalities of the backend to the GUIs and/or the user for direct integration. The principal internal communications between the different modules are performed through REST APIs as well.
- A backend, which orchestrates all the functionalities of the solution, including the AI services, client requests management, pipelines definition and execution, scalability and error handling. The batch and streaming services are deployed independently, although they are both served within the same solution.

We describe each of these components in turn below.

2.1. Frontends

The STREAMS GUIs aim to facilitate the communication between the user and the backend within a batch scenario. They were designed from a usability and user experience perspective, prioritising simplicity.

The first GUI, shown in Figure 2, provides users with different input options, such as text and audio file, and allows them to select transcription and translation models to perform the desired tasks. Additionally, it integrates two main text-boxes to present the transcription and



intervencion.wav

Model: eu — Select a transcription model

Model: eu2es — Select a translation model

00:00 03:06

Transcription

```
sailburuak, legebiltzarkideak, egun on guztioi.


11
00:00:29,910 --> 00:00:34,380
Zupiria jauna, jakin badakigu
gizon eta emakumeen berdintasunean


12
00:00:34,380 --> 00:00:37,530
aurrera egiteko Jaurlaritza
honek hainbat politika jarri dituela


13
00:00:37,530 --> 00:00:39,149
martxan legealdi honetan zehar.


14
00:00:39,719 --> 00:00:42,479
Pentsatzen dugu, gainera,
beharrezkoa dela berdintasunaren


15
00:00:42,479 --> 00:00:45,689
arloan espresuki eragiteko
neurriak martxan jartzea,


16
```

Translation

```


11
00:00:36,869 --> 00:00:41,725
Señor Zupiria, somos conscientes
de que este Gobierno ha puesto en marcha


12
00:00:41,725 --> 00:00:46,042
a lo largo de esta legislatura
una serie de políticas para avanzar


13
00:00:46,042 --> 00:00:48,354
en la igualdad de mujeres y hombres.


14
00:00:48,354 --> 00:00:53,211
Pensamos, además, que es necesario
poner en marcha medidas de incidencia


15
00:00:53,211 --> 00:00:57,913
expresa en materia de igualdad,
y que éstas no vengan sólo de Emakunde,


16
00:00:57,913 --> 00:01:00,688
sino que se planteen de forma transversal
```

Figure 2: Main STREAMS GUI for multilingual transcription and translation.

translation results. It is worth noting that the transcription and translation results can be downloaded in different formats (txt, json and srt), to be used for different applications.

The second GUI for speech synthesis, shown in Figure 3, includes a main text-box for the input text and a drop-down menu to select the desired voice.

The GUIs were developed using the Angular framework[9] and deployed via a Nginx web server[10]. Figure 2 shows the graphical interface for transcription and translation purposes.

2.2. REST API

The principal REST API connected to the Commander serves as the main interface between the GUI and the backend for batch processing. Additionally, it provides an alternative way for the user to directly access all the features of the solution via http requests, allowing third party systems to be built on top of STREAMS and thus extend their functionality. The principal communication between all modules of the solution are performed via REST APIs as well.

2.3. Backends

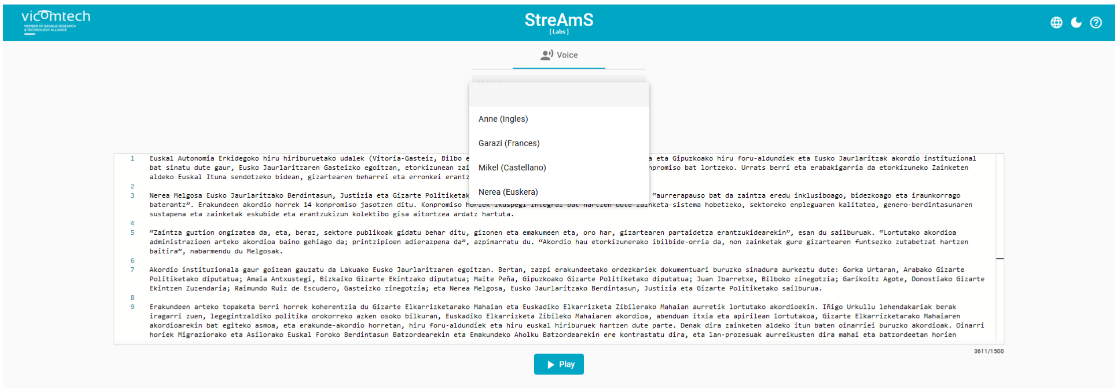
The STREAMS solution offers two types of backends depending on the operational mode under which contents are provided as input, supporting batch and streaming processing modes.

The batch operational mode operates on previously generated contents in video, audio and texts formats, generating transcriptions and translation in different output formats. In addition to providing results in text plain (txt), the STREAMS solution also generates outputs in JSON format, which includes confidence scores and time stamps at the word level for both transcription and translation, and SRT subtitling format.

The streaming operational mode supports the generation of multilingual subtitles from a video transmitted over the common RTMP and HLS streaming protocols. The streaming backend was built to manage both the input video and audio channels, generate a new subtitle channel in the original or translated languages, and construct a new output stream with the 3 channels of video, audio and subtitles perfectly synchronised in time.

Managing the input video and audio channels during the subtitle generation process enables the generation of perfectly aligned subtitles within a controlled timing of the video broadcast, which should not exceed a configurable maximum delay. Depending on the content and speech rates, a maximum delay of between 15 and 30 seconds with respect to the original video was sufficient in our tests to guarantee good quality and stable results over time, without information loss. Based on the experience of the companies in the consortium that specialise in streaming, these delays were considered perfectly acceptable in streaming applications designed for web broadcasts. Finally, the output video streams can be transmitted over the same RTMP and HLS streaming protocols, being thus compatible with the most important commercial video streaming platforms.

[9] https://angular.io/
[10] https://www.nginx.com/



Euskal Autonomia Erkidegoko hiru hiriburuetako udalek (Vitoria-Gasteiz, Bilbo e a eta Gipuzkoako hiru foru-aldundiek eta Eusko Jaurlaritzak akordio instituzional bat sinatu dute gaur, Eusko Jaurlaritzaren Gasteizko egoitzan, etorkizunean zai npromiso bat lortzeko. Urrats berri eta erabakigarria da etorkizuneko Zainketen aldeko Euskal Ituna sendotzeko bidean, gizartearen beharrei eta erronkei erantz

Nerea Melgosa Eusko Jaurlaritzako Berdintasun, Justizia eta Gizarte Politiketak “aurrerapauso bat da zaintza eredu inklusiboago, bidezkoago eta iraunkorrago baterantz”. Erakundeen akordio horrek 14 konpromiso jasotzen ditu. Konpromiso horiek ikuspegi integral bat hartzen dute zainketa-sistema hobetzeko, sektoreko enpleguaren kalitatea, genero-berdintasunaren sustapena eta zainketak eskubide eta erantzukizun kolektibo gisa aitortzea ardatz hartuta.

“Zaintza guztion ongizatea da, eta, beraz, sektore publikoak gidatu behar ditu, gizonen eta emakumeen eta, oro har, gizartearen partaidetza erantzukidearekin”, esan du sailburuak. “Lortutako akordioa administrazioen arteko akordioa baino gehiago da; printzipioen adierazpena da”, azpimarratu du. “Akordio hau etorkizunerako ibilbide-orria da, non zainketak gure gizartearen funtsezko zutabetzat hartzen baitira”, nabarmendu du Melgosak.

Akordio instituzionala gaur goizean gauzatu da Lakuako Eusko Jaurlaritzaren egoitzan. Bertan, zazpi erakundeetako ordezkariek dokumentuari buruzko sinadura aurkeztu dute: Gorka Urtaran, Arabako Gizarte Politiketako diputatua; Amaia Antxustegi, Bizkaiko Gizarte Ekintzako diputatua; Maite Peña, Gipuzkoako Gizarte Politiketako diputatua; Juan Ibarretxe, Bilboko zinegotzia; Garikoitz Agote, Donostiako Gizarte Ekintzen Zuzendaria; Raimundo Ruiz de Escudero, Gasteizko zinegotzia; eta Nerea Melgosa, Eusko Jaurlaritzako Berdintasun, Justizia eta Gizarte Politiketako sailburua.

Erakundeen arteko topaketa berri horrek koherentzia du Gizarte Elkarrizketarako Mahaian eta Euskadiko Elkarrizketa Zibilerako Mahaian aurretik lortutako akordioekin. Iñigo Urkullu lehendakariak berak iragarri zuen, legegintzaldiko politika orokorreko azken osoko bilkuran, Euskadiko Elkarrizketa Zibileko Mahaiaren akordioa, abenduan itxia eta apirilean lortutakoa, Gizarte Elkarrizketarako Mahaiaren akordioarekin bat egiteko asmoa, eta erakunde-akordio horretan, hiru foru-aldundiek eta hiru euskal hiriburuek hartzen dute parte. Denak dira zainketen aldeko itun baten oinarriei buruzko akordioak. Oinarri horiek Migraziorako eta Asilorako Euskal Foroko Berdintasun Batzordearekin eta Emakundeko Aholku Batzordearekin ere kontrastatu dira, eta lan-prozesuak aurreikusten dira mahai eta batzordeetan horien

Figure 3: STREAMS GUI for speech synthesis.

Both batch and streaming backends are composed of several modules which support the features of the solution. Whereas the backend for batch processing includes all the modules, the streaming backend is focused on generating good quality and well-timed subtitles, thus integrating three main modules dedicated to live transcription, machine translation and subtitles generation. All the main modules are described in turn in the following subsections.

2.3.1. Offline and Live transcription

The Offline and Live transcription modules are based on Vicomtech's propietary Transkit software library, which implements a different logic depending on the operational mode. The library offers easy access to speech transcription functionalities through a REST API, supports concurrent processing, can be deployed as a standalone application or in scalable mode with automatic request traffic balancing, and includes dynamic management of decentralised transcription instances.

The library is composed of 7 technological modules connected via configurable pipelines. The modules correspond to an audio transcoder which integrates the FFmpeg[11] tool, an acoustic segmenter based on the Voice Activity Detector module proposed by [3], an Automatic Speech Recognition (ASR) module built on top of Kaldi [4], a module for automatic punctuation and capitalisation [5], a rule-based text normaliser, and a final postprocessing module in charge of generating the different output formats.

Although all these components can be used for batch processing, 3 modules are mainly used for streaming, which correspond to live transcription, punctuation and capitalisation, as well as normalisation. Unlike the offline engine, the live transcription engine implements an online decoding process, in which the recognition models are previously loaded in memory and the features are processed in real time, without having to wait until the audio has finished to start decoding.

2.3.2. Machine Translation

The Machine Translation (MT) module is based on Vicomtech's proprietary Itzuli Translator engine, a robust and scalable text translation system, which can be deployed under Kubernetes orchestration[12] or as a standalone platform in a dedicated server. The module integrates MarianNMT [6] by default in its own backend to perform efficient NMT inference.

On top of the Itzuli MT engine, a dedicated layer was developed to support different processing scenarios. The module can thus handle input delivered in raw text format, or as arrays of transcribed words with their corresponding timestamps for streaming scenarios. Input management, including preprocessing and buffering in streaming scenarios, is handled by the MT module, which can return either a fully postprocessed text or arrays of translated word with an approximated time distribution.

The STREAMS MT module also provides support for end-to-end batch translation of subtitle files in SRT format. In this scenario, source text is automatically extracted from the original file, preserving the time codes, sentences are reconstructed from the original subtitles to provide better sentence-level translation, and the translated text is reinserted into the original time code structure via an information redistribution algorithm. The translated text is thus distributed into each target subtitle according to the original character distribution in the corresponding source subtitle, with additional mini-

[11] https://ffmpeg.org/

[12] https://kubernetes.io/

mal heuristics to enforce key subtitling constraints in the target language, e.g., enforcing subtitle breaks on final punctuation markers or preventing subtitle lines with single words.

2.3.3. Subtitle generator

The main purpose of the subtitle generator component is to provide quality subtitles that adhere to standards of readability and usability, featuring appropriate segmentation and persistence, among other criteria. This module can operate in batch or streaming scenarios, returning formatted subtitles from either arrays of words with their timestamps or plain text.

The module includes a portable and efficient segmentation component, to provide quality subtitles that balance grammatical segmentation and adequate persistence. It includes a character-counting method, which ensures that subtitles respect the maximum number of characters per line. The second method for subtitle segmentation is a simple rule-based approach, which uses character counting as well as enforcing segmentation on punctuation marks. These methods are included as fail-safe in scenarios where minimal latency is critical, but are typically discarded in favour of more advanced methods within the platform in most scenarios.

Two additional classes of methods are provided for quality subtitle generation within STREAMS. The first one relies on supervised models, based on Conditional Random Fields (CRF) as proposed by [7], and DNNs. These models provide quality segmentation but require an offline training phase on annotated data. The second type of method is fully unsupervised and is based on the masking prediction of a Masked Language Model (MLM)[13], as proposed by [9]. This approach provides coverage for all languages supported by the MLM model and delivers quality segmentation at a minimal cost. Within STREAMS, this approach was further extended to integrate other subtitling constraints, such as the optimal number of characters per second, subtitle persistence, or the length balance between lines in a subtitle. The importance of each constraint can be configured to generate subtitles that adhere to the standards of specific broadcasters.

2.3.4. Speech synthesis

The speech synthesis module is based on Vicomtech's proprietary VicomTTS software library for neural speech synthesis, which supports concurrent processing, deployment under local or scalable modes with automatic request traffic balancing, as well as dynamic management of decentralised synthesis instances, via REST API access.

The module consists of 3 main components, each one in charge of a specific part of the process: pre-processing the input text to normalise numbers and special characters, transforming the input text into a waveform via Text-To-Speech (TTS) technology , and finally, formatting the output into the requested audio format.

The core TTS technology is based on the Tacotron-2 architecture [10]. This model consists of a sequence-to-sequence model, which includes an encoder, a decoder with attention and a final post-processing convolutional neural network (CNN). For the STREAMS solution, one synthetic voice was developed and integrated for each of the 4 languages emphasised within the project, enabling the generation of audio content based on input text. It is worth noting that combining this TTS module with the ASR and MT modules, a pipeline for speech translation purposes could be easily created using the proposed architecture, among other interesting applications.

2.3.5. Audio-description

Audio-description (AD) is a growing media access service which complements subtitling by providing audio information for key visual elements in media content, a particularly useful feature for blind and/or visually impaired people. This information is commonly generated manually by dubbers in recording studios, using narrative and neutral voices to describe the scenes and their visual elements.

The STREAMS solution incorporates an AD module which enables the automatic generation of this information using the TTS module, given that the synthetic voices were built on content with a narrative and journalistic style. Taking as input a script with the time-marked descriptions of the scenes, the AD module generates a master audio of the same length of the script and/or content, including each audio segment at the corresponding timestamp. The DocParser sub-module developed for the platform (see Figure 1) aids in processing scripts in various formats and extracting information as needed.

2.3.6. Aligner

The Aligner is a complementary module that enables generating subtitles or time-marked word transcriptions from an audio input and its corresponding script.

This module was developed and integrated in response to the need expressed by several entities, such as dubbing companies, to generate subtitles from pre-existing script and audio. Using the acoustic models and the forced-recognition mode of the Offline Transcription module, the Aligner module generates well-timed subtitles by performing an alignment between the text and the audio, thus recovering the timestamps at the word level. The

[13] We use multilingual BERT by default [8].

DocParser sub-module is used to process the input scripts and extract the texts properly as well.

3. Conclusions

We described STREAMS, a batch and streaming platform which integrates and manages AI services for rich transcription, translation, voice synthesis and audio description. STREAMS is meant to enhance the processes and products of companies working in multiple sectors and has been successfully deployed in real-life batch and multilingual streaming scenarios during the project, offering a rich array of multilingual services with innovative methods, in particular to enhance portability, responsiveness and output readability.

Acknowledgments

STREAMS was partially funded by the Basque Business Development Agency, SPRI, under grant agreement ZL-2023/00211.

The authors would also like to thank the companies in the consortium that have contributed with their knowledge and experience to the project and the STREAMS solution.

References

[1] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, S. Watanabe, End-to-end speech recognition: A survey, IEEE/ACM Transactions on Audio, Speech, and Language Processing (2023).

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 6000–6010.

[3] H. Dinkel, S. Wang, X. Xu, M. Wu, K. Yu, Voice activity detection in the wild: A data-driven approach using teacher-student training, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 1542–1555.

[4] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., The kaldi speech recognition toolkit, in: IEEE 2011 workshop on automatic speech recognition and understanding, CONF, IEEE Signal Processing Society, 2011.

[5] A. González-Docasal, A. García-Pablos, H. Arzelus, A. Álvarez, Autopunct: A bert-based automatic punctuation and capitalisation system for spanish and basque, Procesamiento del Lenguaje Natural 67 (2021) 59–68.

[6] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev, A. F. T. Martins, A. Birch, Marian: Fast neural machine translation in C++, in: Proc. of ACL 2018, 2018, pp. 116–121.

[7] A. Alvarez, C.-D. Martínez-Hinarejos, H. Arzelus, M. Balenciaga, A. del Pozo, Improving the automatic segmentation of subtitles through conditional random field, Speech Communication 88 (2017) 83–95.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[9] D. Ponce, T. Etchegoyhen, V. Ruiz, Unsupervised subtitle segmentation with masked language models, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 771–781.

[10] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, R. A. Saurous, Tacotron: Towards End-to-End Speech Synthesis (2017) 4006–4010. doi:10.21437/Interspeech.2017-1452.