FIAT IFTA #ArchivalReads

Artificial Intelligence: an object of desire

The ‘Value Use and Copyright Commission’ (VUC) of FIAT/IFTA has decided to publish a series of interviews and articles about use of audiovisual archive materials. In this article Virginia Bazán-Gil gives an overview of how artificial intelligence technologies were tested at the RTVE Archives in the Iberspeech and Crónicas projects. Finally a roadmap is added to learn how to embrace artificial intelligence technologies for audiovisual archival work.

By Virginia Bazán-Gil – project manager, RTVE Archives, Spain - published on 13 May 2020.

The RTVE Archive

The RTVE Archive is the archive of Spain’s main public broadcaster. It is split into two sections: the RNE archive (radio) and the TVE archive (television). The radio archive was established back in 1957 and keeps the radio programs broadcasted by RNE’s six radio stations together with other external productions supporting the daily broadcasting (such as music recordings). The television archives preserves footage produced by TVE during the last 64 years including film collections.

RTVE Archive’s main mission is the preservation of the RTVE archival collection. The archival team makes sure that the content is valid for production and for commercial purposes. As a public organization RTVE also has to guarantee the accessibility to the its audiovisual heritage since it is preserving unique and historical footage.

So, how has it done that? The radio archive has been fully digitized since 2002. The digitalization of the television archive was the most important digitization project carried out by a broadcaster in Spain. Part of the process has been performed externally and it was finished in March 2013. Since then the digitalization continues in-house. The current focus is placed on the collections coming from the RTVE local delegations and on the film archive.

The team is currently dealing with the migration from LT04 and LT05 to LT07 and has finished the unification our the two main digital storages, the news library and the programs library. Finally, we are analysing the opportunities of the cloud infrastructure but only in terms of storage. In other words RTVE is not considering the idea of including the cloud in its archive daily workflows.

RTVE team uses AVID for production and a homemade development for the archive called ARCA which integrates different collections such as: photos, texts, audios, videos, etc., coming from radio and television. Our archival digital library is connected through gateways with different production systems of RTVE, including the rtve.es CMS, which allows our colleagues from the digital area to use the archive content for social media and for the RTVE website.

”As a public broadcaster RTVE has to open up its collections but at the same time the archival team is required to work for the production needs. This dual task of making RTVE’s collections accessible for general purposes as well as focusing on professional production can be a challenge.”

The long and winding road to AI: lessons learned at RTVE

The main goal of this paper is to share the RTVE archive's journey towards the automatic metadata generation. This is a trip full of chimeras, mirages and dangers but at the end, the promised land seems to be waiting for us, because once you have your archive fully digitalized, Artificial Intelligence is the first thing you think about when you make a wish concerning your future projects.

In this brief exchange of thoughts I'm starting with an introduction to the technologies for automatic metadata generation. Next, I'm highlighting its relevance in the production, broadcast and archive workflows. After that I'm identifying technological vendors, and last but not least, I'm sharing some experiences from the RTVE Archive. I will conclude with some lessons the RTVE team has learned during our on trip to automatic metadata generation.

Technologies for automatic metadata generation: Speech technologies, natural language processing and artificial vision.

Artificial vision (AV) deals with image recognition, scene grouping and segmentation and tracking of objects and people. This technology provides an image description, expressed in natural language or as tags.

Speech technologies allow voice to text transcription, the identification of the speakers and their emotions and speech segmentation and enable voice to text transcription. Speech to text is relevant for content editing as well as for broadcasting. It allows the automatic transcription of interviews during the editing process and also the generation of subtitles for broadcasting and web uploading.

Natural Language processing (NLP) can be used to enrich the content uploaded to the web through automatic classification and automatic generated tags, which also enriches the archival content. All these technologies are relevant at all stages of the creation-broadcasting-archive workflow.

Who should provide us with the technology we need?

We can identify three main groups of vendors:

- Big international corporations - they own the technology and are constantly improving it. They also manage a lot of data coming from social media.
- Companies focused on the broadcasting industry who integrate third party solutions or have their own developments based on open source technologies. These are MAM vendors or specialized companies who generate secondary workflows to process media content.
- National or international research groups constitute another relevant stakeholder group. AI needs large amounts of data and data is the jewel of archives. Identifying research groups in

artificial vision, speech technologies and NLP, and jointly setting up a research project could be an easier way to embrace AI.

Is metadata archival stuff only?

No it is not. Other business areas such as production, broadcasting or sales have been generating metadata for years for the same content according to their own needs. Recently RTVE is facing high demand for data coming from the website. Broadcasters websites demand more metadata to identify, classify, enrich and make content more accessible and easier searchable. They need more data and they need it fast.

As archivists we know that many of the areas mentioned above have their own resources to start new technological projects that are just impossible for the archive. This situation, which could be considered as a threat, is indeed a unique opportunity to find joint strategy and build internal alliances to embrace new technologies.

To identify these strategic allies but also to contribute to the efficiency and sustainability of metadata in your own organization you need to understand the life circle of data in the editing, broadcasting and archiving workflow. Why should we create new data if we can reuse data already generated by others? This is one of the main lessons we have learned over these past 3 years.

Artificial Intelligence at the RTVE Archive: Iberspeech 2018 Challenge and Cronicas Project

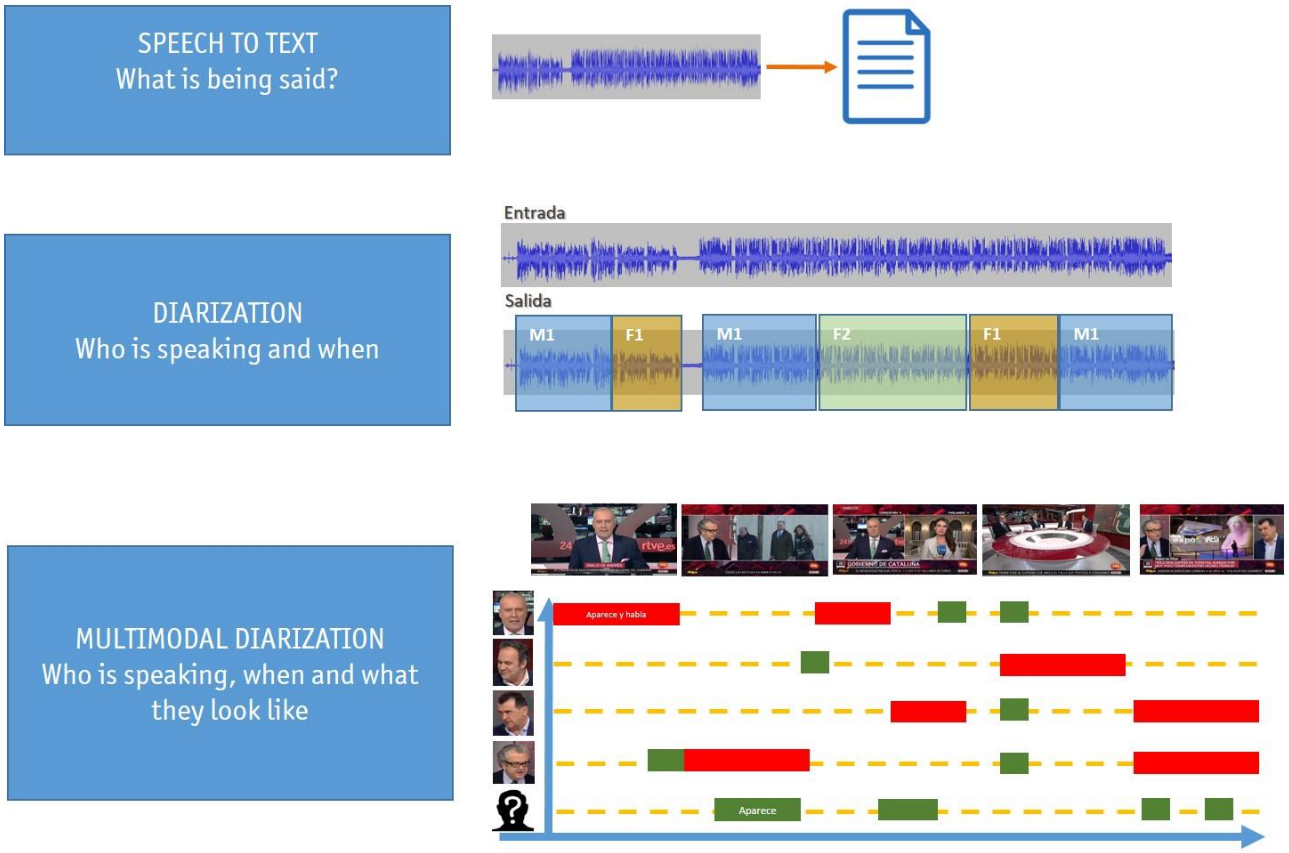
Once we had identified the technologies and their uses, the vendors and the strategic allies, the work on specific projects began. The Iberspeech Challenge and the Cronicas project were both focusing on speech technologies.

For radio and television archives speech to text and diarization are the most relevant speech technologies. Speech to text allows to know what is being said and it is relevant to:

- Ensure content accessibility through subtitles
- Improve production: automatic transcription of interviews
- Allow partial automation of the cataloguing process through automatic classification and tagging.

Diarization helps identify who is speaking, when and what they look like if we also apply facial recognition techniques. It also applies to:

- Indexing speakers and speeches
- Content structuring
- Speaker recognition as a help for speech to text



SPEECH TO TEXT
What is being said?

DIARIZATION
Who is speaking and when

MULTIMODAL DIARIZATION
Who is speaking, when and what they look like

Figure 1: Technologies tested in Iberspeech 2018

In this context, in March 2018 the RTVE and the University of Zaragoza Chair organized and promoted the **Iberspeech Challenge**. For this purpose RTVE made more than 500 hours of broadcast contents and subtitles available for scientists. The dataset included about 20 programs of different kinds and topics produced and broadcasted by RTVE between 2015 and 2018. The programs were selected in order to represent different challenges from the point of view of speech technologies such as the diversity of Spanish accents, overlapping dialogues, spontaneous speech, acoustic variability, background noise or specific vocabulary. All these elements have an effect on the system performance.

For the dataset the participants had to automatically identify, tag and transcribe speech segments. Three different challenges were defined: speech to text, diarization and multimodal diarization.

The main goal was to automatically transcribe 39 hours of TV programs. The WER (Word Error Rate) was used to measure the systems performance. This metric takes into account the number of insertions, substitutions and deletions in relation to the total number of words. The results show that the quality of the transcription depends on the type of program:

- WER between 16,00% and 35,00%
- Programs with interviews or spontaneous chats outside the studio are challenging for systems, with WER above 20%

- WER below 20% for many programs. WER below 10% for "Latin America in 24H", with different accents of Spanish.

For the diarization challenge the goal was to run speech segmentation and speaker clustering for 22 hours of broadcast material. The goal for the multimodal diarization was the same but also included facial recognition. The DER (Diarization Error Rate) was used to evaluate the system performance. The results showed that system performance depends on the type of content with DER between 17.22% and 39.09%. The incorrect attribution of speakers is the most common mistake. As for the multimodal diarization we obtained a better performance for the speaker diarization rather than for face diarization.

Lessons learned? The RTVE team needs to approach these results from the archive's point of view, taking into account the utility of the tested technologies and their integration in the organization's workflows.

> *"In the last 5 years speech technologies have improved exponentially, even though there is still a great variability in the error rate, which limits the possibility of using speech to text in some collections."*

The **Crónicas project** aimed to answer to a long-term demand from journalists: automatic transcription of interviews. The main goal of the project was to determine the utility of the speech to text in the production of new content. Different tools and their possible integration with AVID (RTVE's production system) were tested. We also wanted to determine which WER is admissible for our journalists. In cooperation with Crónicas Newsroom, we defined a number of system requirements:

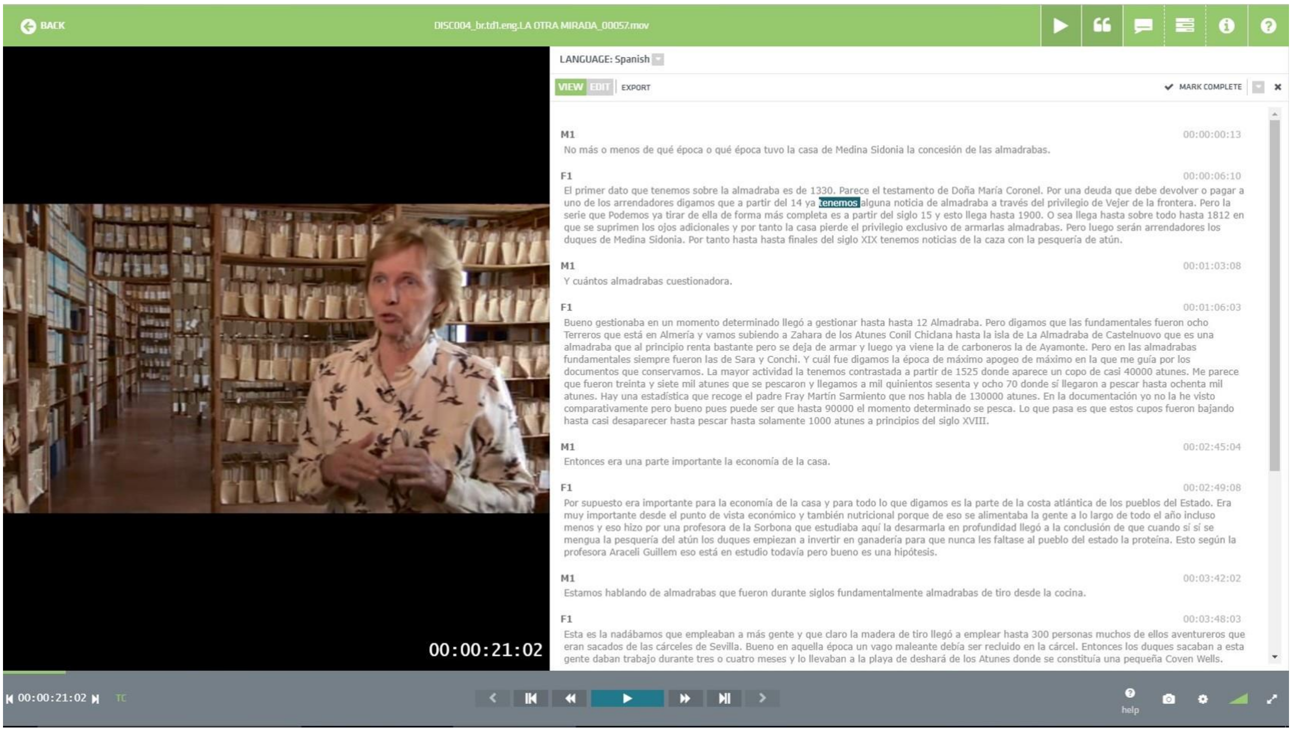
- Cloud solution
- Intuitive and user friendly file uploading and processing
- Allowing multiple users working at the same time
- Keeping the original time codes (a must)
- Segmentation by questions and answers
- Capitalization and punctuation marks
- Easy text edition
- WER around 15%

A content sample was selected to be processed, taking into account relevant issues such as: variety of language models and audio quality, spontaneous speech, overlapping dialogues, different Spanish accents. For this test we processed around 25 clips of raw material produced by the Crónicas team and we tested two different tools: VSN and Limecraft.

The VSN MAM integrates AI solutions from Google, Microsoft and from other technology providers such as Amazon or Etiqmedia. The RTVE team only tested technology from Azure (Microsoft) and Google. Although journalists found the time needed for uploading and processing reasonable, both

AI solutions made segments without semantic meaning lasting less than one minute. From a user point of view, these segments make the correction process more difficult and interfere with their recalling of the interview during the edition process. Furthermore Google does not use capitalization or punctuation marks. Azure does but most of the sentences are finished with a question mark. Some additional tests were performed with high audio quality and proxies with Google. The WER was higher for the high quality video rather than the proxies.

A second interaction was made using Limecraft. This is a cloud solution based on the Speechmatics speech to text recognition system which was identified as one of the best systems in the Iberspeech 2018 Challenge. During the tests it was acknowledged that the time for uploading and processing was reasonable. Questions and answers were not always clear-cut (diarization error) but the segments were easier to understand and the transcription included capitalization and punctuation marks. The WER was about 20%-30% lower than the WER returned by Azure or Microsoft and users could edit transcriptions in the interface and also generate subclips to share with AVID solutions.



M1 00:00:00:13
No más o menos de qué época o qué época tuvo la casa de Medina Sidonia la concesión de las almadrabas.

F1 00:00:06:10
El primer dato que tenemos sobre la almadraba es de 1330. Parece el testamento de Doña María Coronel. Por una deuda que debe devolver o pagar a uno de los arrendadores digamos que a partir del 14 ya tenemos alguna noticia de almadraba a través del privilegio de Vejer de la frontera. Pero la serie que Podemos ya tirar de ella de forma más completa es a partir del siglo 15 y esto llega hasta 1900. O sea llega hasta sobre todo hasta 1812 en que se suprimen los ojos adicionales y por tanto la casa pierde el privilegio exclusivo de armarlas almadrabas. Pero luego serán arrendadores los duques de Medina Sidonia. Por tanto hasta hasta finales del siglo XIX tenemos noticias de la caza con la pesquería de atún.

M1 00:01:03:08
Y cuántos almadrabas cuestionadora.

F1 00:01:06:03
Bueno gestionaba en un momento determinado llegó a gestionar hasta hasta 12 Almadraba. Pero digamos que las fundamentales fueron ocho Terreros que está en Almería y vamos subiendo a Zahara de los Atunes Conil Chiclana hasta la isla de La Almadraba de Castelnuovo que es una almadraba que al principio renta bastante pero se deja de armar y luego ya viene la de carboneros la de Ayamonte. Pero en las almadrabas fundamentales siempre fueron las de Sara y Conchi. Y cuál fue digamos la época de máximo apogeo de máximo en la que me guía por los documentos que conservamos. La mayor actividad la tenemos contrastada a partir de 1525 donde aparece un copo de casi 40000 atunes. Me parece que fueron treinta y siete mil atunes que se pescaron y llegamos a mil quinientos sesenta y ocho 70 donde sí llegaron a pescar hasta ochenta mil atunes. Hay una estadística que recoge el padre Fray Martín Sarmiento que nos habla de 130000 atunes. En la documentación yo no la he visto comparativamente pero bueno pues puede ser que hasta 90000 el momento determinado se pesca. Lo que pasa es que estos cupos fueron bajando hasta casi desaparecer hasta pescar hasta solamente 1000 atunes a principios del siglo XVIII.

M1 00:02:45:04
Entonces era una parte importante la economía de la casa.

F1 00:02:49:08
Por supuesto era importante para la economía de la casa y para todo lo que digamos es la parte de la costa atlántica de los pueblos del Estado. Era muy importante desde el punto de vista económico y también nutricional porque de eso se alimentaba la gente a lo largo de todo el año incluso menos y eso hizo por una profesora de la Sorbona que estudiaba aquí la desarmarla en profundidad llegó a la conclusión de que cuando sí sí se mengua la pesquería del atún los duques empiezan a invertir en ganadería para que nunca les faltase al pueblo del estado la proteína. Esto según la profesora Araceli Guillem eso está en estudio todavía pero bueno es una hipótesis.

M1 00:03:42:02
Estamos hablando de almadrabas que fueron durante siglos fundamentalmente almadrabas de tiro desde la cocina.

F1 00:03:48:03
Esta es la nadábamos que empleaban a más gente y que claro la madera de tiro llegó a emplear hasta 300 personas muchos de ellos aventureros que eran sacados de las cárceles de Sevilla. Bueno en aquella época un vago maleante debía ser recluido en la cárcel. Entonces los duques sacaban a esta gente daban trabajo durante tres o cuatro meses y lo llevaban a la playa de deshará de los Atunes donde se constituía una pequeña Coven Wells.

Figure 2: Screenshot of one of the speech to text tools tested

Although the user perception of the technology has improved during the test, the Crónicas team considered that the cost-benefit rate was not good enough to adopt the technology.

They considered that the speech correction process would negatively affect both understanding and the editorial process. The same could be said about capitalization, punctuation marks and speaker segmentation. Furthermore, the integration with video editing tools such as AVID was not perceived as positive.

From the archive’s point of view working with journalists on the editorial process will allow the archive to preserve also the transcripts done in the production stage and to have them used not only for searches but also for fact checking and news verification. That’s why technology solutions integrated with the edition tools are needed.

During this test the team realized that the journalists’ expectations towards the accuracy of the speech to text solutions were not realistic. Also, they don’t want to transcribe the interviews manually either – that’s why we have started working with another documentary program and hopefully we will get some good results from this new project in the coming months.

Main conclusions regarding speech technologies

In the last 5 years the speech recognition systems have improved, current WER are between 5% and 50% and that is why applying the technology ’everywhere’ is not possible.

The performance of speech to text depends on three key factors:

- acoustic environment,
- oral expression of the speaker
- syntactic context.

If you can’t understand what a person is saying, then the speech recognition system can’t do it either! Think about a group of people chatting: using just one microphone, spontaneous speech, with speech overlapping, using colloquial expressions – all these factors are challenging for the technology.

The system’s performance depends on the type of content. You need a balance between the real needs in terms of the archival work, and the expectations in certain areas regarding quality. Human quality performance is the main enemy of AI.

The degree of tolerance for errors varies depending on the scenario for which the metadata is generated. Error rates above 15% may be tolerable for archiving when there’s no other information, but they could be inadmissible for broadcasting, specially if we are talking about politics or other sensitive topics.

A roadmap to embrace Artificial Intelligence

After our experience I would like to leave you with my roadmap to embrace AI:

- Analyse your own needs as an archive and be ready to understand the needs of other business areas. Where is metadata generation a priority? Focus on your institutional priorities concerning content accessibility and its reuse.
- Identify technological vendors and research groups working with the technologies you would like to integrate in your workflows.

- Benchmark and test the technology. There is a sea of possibilities but also limitations.
- Create synergies with other divisions in the organisation that could lead the adoption of the technology. Remember: a change like this is only possible together with some allies.
- Analyse your metadata model: is it ready to assume metadata coming from AI?
- Adapt users' expectations to real technology performance.
- Make yourself visible - the power of metadata belongs to us, the archivists. Become the oracle for the others in your company.

And one last suggestion: don't use technology randomly or you take up the risk of using economic, technological and human resources without any guarantee of getting good results.

This article is based on a presentation given at the FRAME Access training programme in 2019.