Journal of Engineering Interfaces, Vol 6 (2) June-December 2023. ISSN 2619-4473 E-ISSN 2619-4465, pp 1-14.

Transcription from audio to text in Municipal Sessions in Planeta Rica

Transcripción de audio a texto en Sesiones Municipales de Planeta Rica

Jaime Andrés Ruiz-Melendres **
Jorge Eliecer Gómez-Gómez **

Recibido: abril 1 de 2023 - Evaluado: abril 30 de 2023 - Aceptado: agosto 2 de 2023

Para citar este artículo / To cite this Article
J. Ruiz-Melendres, J. E. Gómez-Gómez, "Transcription from audio to text in Municipal Sessions in Planeta Rica" Interfaces Engineering Journal, vol. 4, no. 1, pp.1-14, 2023.

Abstract
The document addresses the issue of manually transcribing municipal sessions in Planeta Rica. It investigates the use of open-source tools to automate the transcription of audio to text in these sessions with the aim of improving efficiency and accuracy in this process. The importance of integrating models into the system to address different aspects and enhance transcription quality is emphasized. In this regard, two artificial intelligence models are mentioned: OpenAI's Whisper and Deezer's Spleeter. Whisper is a general-purpose speech recognition model. On the other hand, Spleeter is an audio track separation tool that utilizes pre-trained models to separate voices from any audio track. Furthermore, an architecture is developed to enable the automatic integration of these models. This architecture is based on the use of Python for managing the artificial intelligence models, while the application's backend is developed using Go and the frontend with Next.js/React. This allowed for the automation of transcriptions for Planeta Rica's municipal council sessions, improving both efficiency and precision in the process.

Keywords: AI, artificial intelligence, audio-to-text, transcription, sessions, whisper, spleeter.

*Unpublished article: "Transcription of audio to text in Municipal Sessions in Planeta Rica".

**Systems Engineering, Universidad de Córdoba, jruizmelendres88@correo.unicordoba.edu.co, ORCID: 0009-0000-7511-4131, Monteria, Colombia.

***Systems Engineering, Universidad de Córdoba, jeliecergomez@correo.unicordoba.edu.co, ORCID: 0000-0001-8746-9386, Monteria, Colombia.

Resumen: El documento aborda el tema de la transcripción manual de sesiones municipales en Planeta Rica. Se investiga el uso de herramientas de código abierto para automatizar la transcripción de audio a texto en estas sesiones con el objetivo de mejorar la eficiencia y precisión en este proceso. Se enfatiza la importancia de integrar modelos en el sistema para abordar diferentes aspectos y mejorar la calidad de la transcripción. En este sentido, se mencionan dos modelos de inteligencia artificial: Whisper de OpenAI y Spleeter de Deezer. Whisper es un modelo de reconocimiento de voz de propósito general. Por otro lado, Spleeter es una herramienta de separación de pistas de audio que utiliza modelos previamente entrenados para separar voces de cualquier pista de audio. Además, se desarrolla una arquitectura que permite la integración automática de estos modelos. Esta arquitectura se basa en el uso de Python para la gestión de los modelos de inteligencia artificial, mientras que el backend de la aplicación se desarrolla utilizando Go y el frontend con Next.js/React. Esto permitió automatizar las transcripciones de las sesiones de los concejos municipales de Planeta Rica, mejorando tanto la eficiencia como la precisión del proceso.

Palabras clave: IA, inteligencia artificial, audio a texto, transcripción, sesiones, whisper, spleeter.

1. Introduction

The municipal councils are local legislative bodies in Colombia, in charge of the legislative function in the municipalities. Their sessions are a fundamental space for debate and discussion of public policies that affect citizens. To ensure transparency and accountability, municipal council sessions must be recorded and transcribed to text. Recordings are a faithful record of what was said in each session, while transcripts allow citizens to access this information easily and quickly. [1]

The transcripts of the sessions of the municipal councils are used to prepare the minutes of each session, as mentioned in Concept 97471 of 2020 Departamento Administrativo de la Función Pública.[1] The minutes are official documents that certify the attendance of the councilors, the topics discussed at the session and the decisions taken.Therefore, these sessions are manually transcribed by one person, which is an extremely tedious and time-consuming process, delaying the output of the minutes of each session by the municipal council.

Although there are currently several alternatives to solve this problem, most of them are costly and inaccessible to public entities in small cities, as is the case of the municipal council of Planeta Rica.

Therefore, we focused on investigating how open source audio-to-text processing tools[2] can automate the creation of the records recorded in the Planeta Rica plenary sessions. To this end, several tools will be explored in order to perform the transcriptions of the audio recordings, evaluating their accuracy, efficiency and ease of use.

Motivation

Traditionally, the transcriptions of the audios obtained by each session in the plenary are made by one person, either the general secretary of the council (Mauricio Andrés Ruiz Herazo) or someone external hired by him. This was explained by the secretary general of the municipal council of Planeta Rica, which is the subject of our study.

The recordings can have an average duration of 20 minutes, up to more than 2 hours, depending on the debate to be held and the type of session to be held, whether normal or extraordinary. These recordings play a fundamental role in the persistence and proof of what was said in each session, which must be recorded in minutes for the council's general archive. [1]

Therefore, it is essential that one person transcribes these audio files in order to be able to record in the minutes what was said by each participant in the plenary session. However, there is a fundamental problem when it comes to optimizing the time for transcription, since transcriptions tend to be slow and very laborious.

On the other hand, in the context of technology and artificial intelligence, there are several tools that facilitate mechanical processes or automate repetitive tasks. These tools use machine learning algorithms to identify patterns and make decisions, allowing them to perform tasks more efficiently and accurately than humans. Thus, one of the approaches that has been booming in recent years is text and audio processing, developing tools that enable audio-to-text transformations and vice versa.[3] This approach is based on the idea that text and audio are ways of representing information, and that it is possible to translate between the two forms.

A possible solution to this problem is the integration of an audio-to-text transcription tool in the Municipal Council of Planeta Rica, Córdoba. This tool would make it possible to transcribe municipal sessions more quickly and accurately, which would improve the efficiency and effectiveness of the council at the time of evidencing what was said in the minutes of each session.

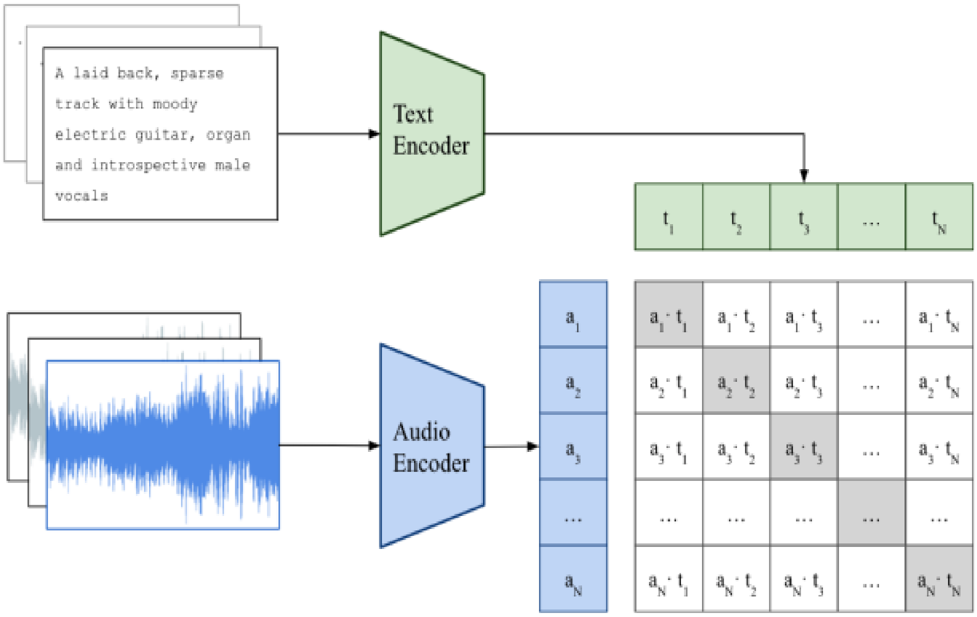
However, in the context of audio-to-text processing, it is important to keep in mind that not all tools are publicly accessible. In addition, their processing will be affected by the environment in which the tool is run, which limits their use to certain people. The implementation of OpenSource models for audio to text processing by developing a user-friendly system for the people in charge of the development of the minutes of each session in the municipal council.

Related work

This section explores different studies, articles and research that implement related technologies for similar use cases.

Contrastive Audio-Language Learning for Music [4].

In previous works, it is possible to find several systems designed to perform multimodal processing in machine learning systems.[5] These systems address multimodality[6] in various aspects, such as text-to-image or audio-to-text conversion.[4] The effort to merge audio and language reflects the search for algorithms that allow establishing an adequate dimensional relationship between concepts.[7] This relationship is crucial to achieve effective coding, as illustrated in the figure below Figure 1.


A laid back, sparse track with moody electric guitar, organ and introspective male vocals

$a_1 \cdot t_1$	$a_1 \cdot t_2$	$a_1 \cdot t_3$	...	$a_1 \cdot t_N$
$a_2 \cdot t_1$	$a_2 \cdot t_2$	$a_2 \cdot t_3$	...	$a_2 \cdot t_N$
$a_3 \cdot t_1$	$a_3 \cdot t_2$	$a_3 \cdot t_3$	...	$a_3 \cdot t_N$
...	...	...	...	...
$a_N \cdot t_1$	$a_N \cdot t_2$	$a_N \cdot t_3$	...	$a_N \cdot t_N$

Figure 1. Overview of MusCALL. An audio encoder and a text encoder are trained using contrastive loss to maximize the similarity between the representations of N aligned pairs (audio, text) within a mini-batch. At test time, the similarity between the embeddings in the learned multimodal space is used to rank the database items and perform cross-modality retrieval.
Source: Papers with Code - Contrastive Audio-Language Learning for Music. (2022, August 25).[4].

This work, although not directly related to the main research, provides valuable insight into the effectiveness of multimodal models in the audio-to-text conversion process. This, in turn, makes it easier to perform searches by recognizing words in audio tracks. This additional insight enriches our overview of how multimodal models can play an important role in improving the accessibility and usability of transcription applications.

VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text [8].

In this work, the effectiveness of self-supervised convolutional [9] Transformers [10] in image recognition can be evidenced, as well as their effectiveness in the multimodal domain, covering not only image recognition, but also text and audio recognition.The authors refer to the difficulties they faced when training models with large data sets. They highlight those important parts of the data, due to the difficulty in labeling, may be overlooked during the model training process. To address this problem, they propose a solution called 'Vatt' [8] to mitigate this drawback.

In their conclusions, Hassan Akbari and Wei-Hong Chuang, in their 2021 study [8] mention "We report new records of results on video action recognition and audio event classification and competitive performance on image classification and video retrieval. Having these results, we still see some limitations in our work. Firstly, not all videos have organic audio or speech, while our approach depends on meaningful multimodal correspondences. Besides, the text modality currently consists of speech transcripts, which are noisy and sometimes sparse.", express the difficulties they encountered when trying to convert audio files to text. They point out that these files are often noisy and sparse, which makes their transcription difficult and therefore represents a challenge in the multimodal task they tackled.

Audio to text systems and vice versa.

Research focusing on multimodality between audio and text is often oriented towards the creation of synthetic voices. In this context, systems that enable speech recognition are developed with the aim of substantially improving methods for recognizing speech and its text equivalent. [11] Furthermore, there are other systems that, in addition to speech recognition, have additional capabilities, such as the recognition and transcription of speech from English to Mandarin, as mentioned in reference. [12]

These advances contribute significantly to the improvement of similar models, resulting in substantial improvements when creating transcription systems using phoneme speech synthesis. [13, 14]. However, these models are also applicable to speech synthesis, as described in the reference. [15] In the case of text, they have the ability to generate and synthesize speech clearly and accurately. Examples of such applications include WaveNet, as mentioned in reference. [16] These applications are especially focused on generating speech from text.

System construction

In the construction of the system to effectively develop the audio to text transcription tool, this process is composed of several stages. These stages range from loading the audio file, separating tracks, reviewing the audio to be processed and transcribing the audio, with the ultimate goal of obtaining the text corresponding to the session held at the Planeta Rica council, as shown in Figure 2.

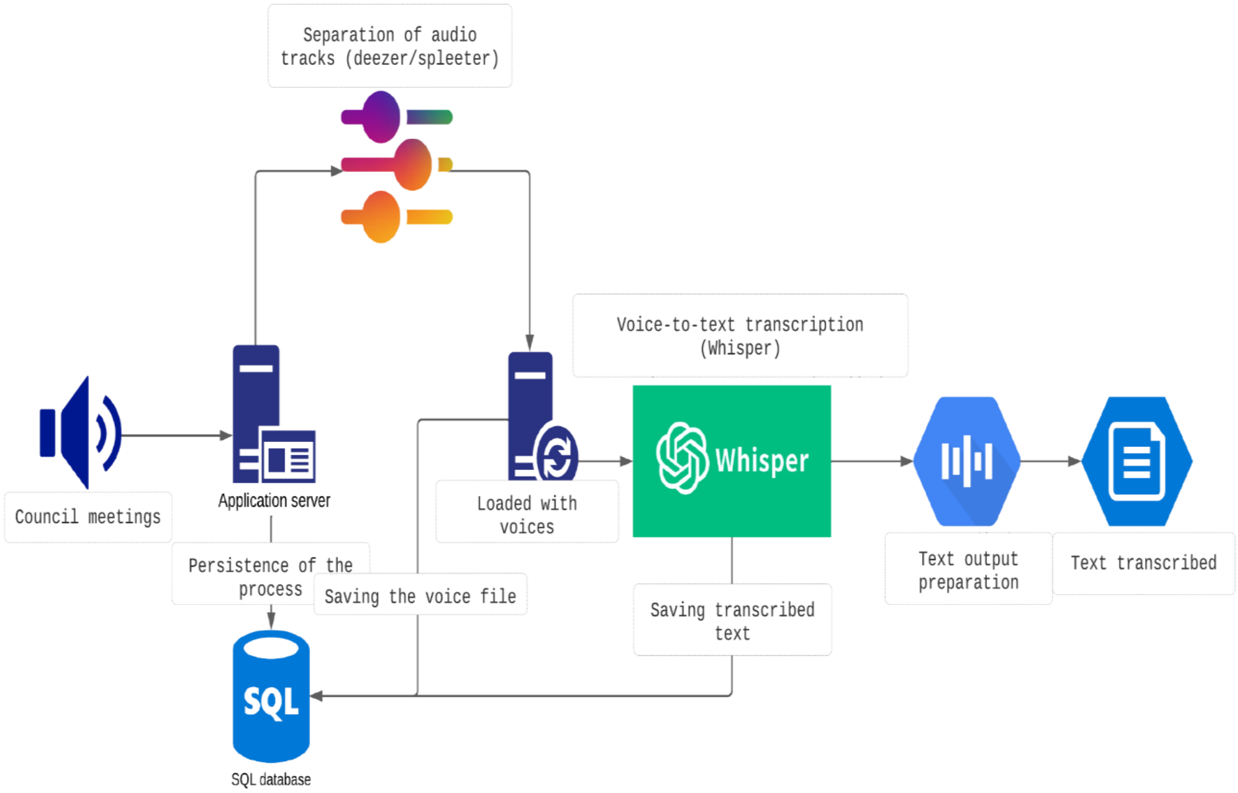


Figure 2. General diagram of the system.
Source: Authors

Deezer/Spleeter track separation tool. [17]

Deezer's Spleeter is a set of pre-trained models written in Python using the Tensor stream machine learning library used for music source separation (MSS). These models are already trained and show state-of-the-art Performance in MSS. [18]. Reviewing the study conducted on Spleeter, it was found that it is designed under the U-Net architecture. [18]

The U-Net architecture is characterized by its "U" shape and consists of an encoder-decoder structure, as shown in Figure 3 [19].

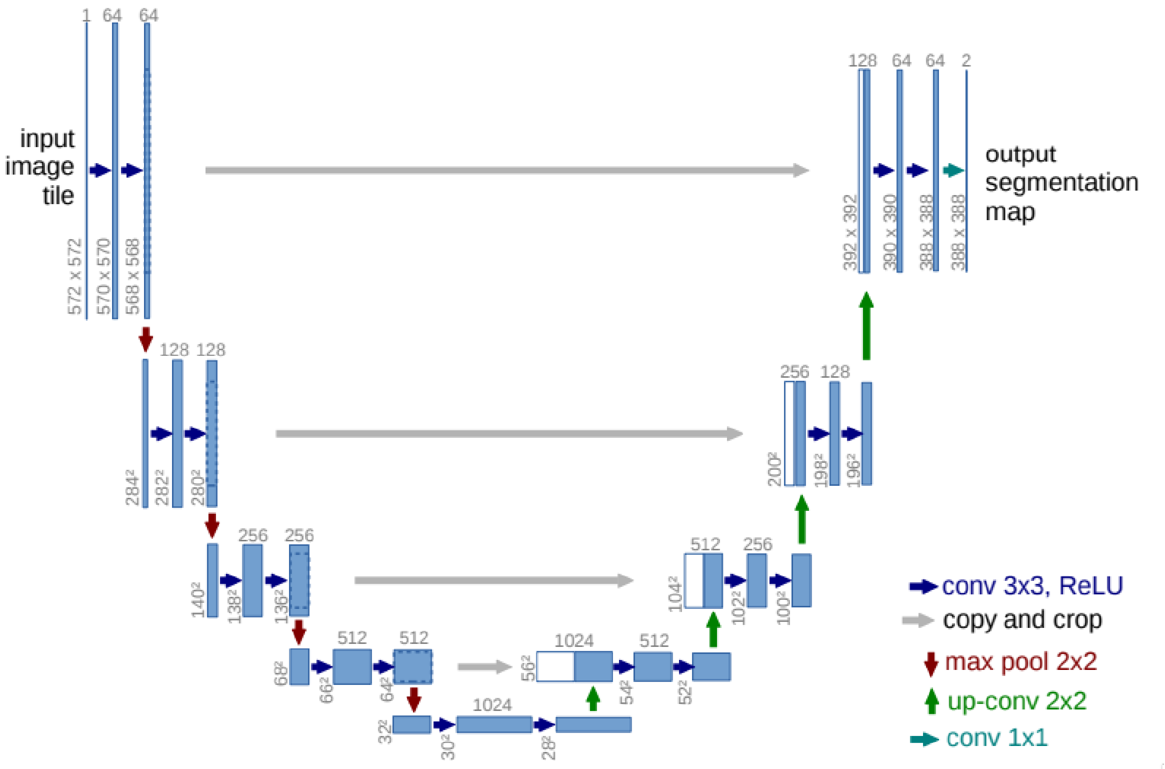


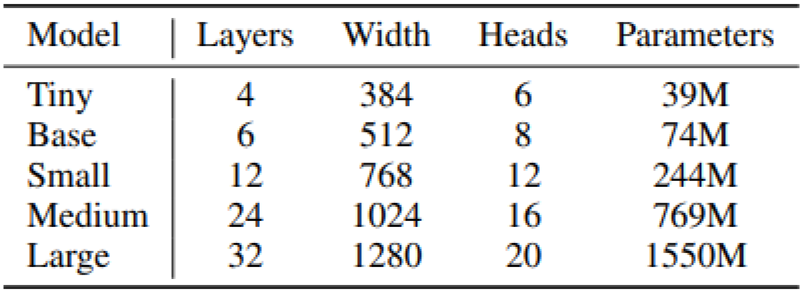
Figure 3. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.
Source: Ronneberger, O. (2015, May 18). U-NET: Convolutional Networks for Biomedical Image Segmentation.[19].

Based on the above, Spleeter can be relied upon as a solid tool for audio track separation. Although it is primarily designed for music, it can also be used to separate vocals from any audio track, which significantly facilitates transcription in the next step of the process.

Whisper audio-to-text transcription tool. [20]

Whisper is a general-purpose speech recognition model developed by OpenAI.[21] This model has been trained on a large set of diverse audio data and is characterized by multitasking, meaning that it can perform tasks such as multi-language speech recognition, speech translation, and language identification. [20]

Whisper offers a range of variants, each with different capacities and sizes. These variants include Tiny, Base, Small, Medium and Large, as shown in Figure 4 below.



Model	Layers	Width	Heads	Parameters
Tiny	4	384	6	39M
Base	6	512	8	74M
Small	12	768	12	244M
Medium	24	1024	16	769M
Large	32	1280	20	1550M

Figure 4. Architecture details of the Whisper model family.
Source: Radford, A. (2022, December 6). Robust speech recognition via Large-Scale Weak Supervision.[20].

The architecture of this model, based on Transformers, has been designed to address multimodality, i.e., understanding both audio and text data. In addition, it can be trained in several languages, which allows it to generate translations between different languages with high reliability. The structure of the model is described in detail in Figure 5.

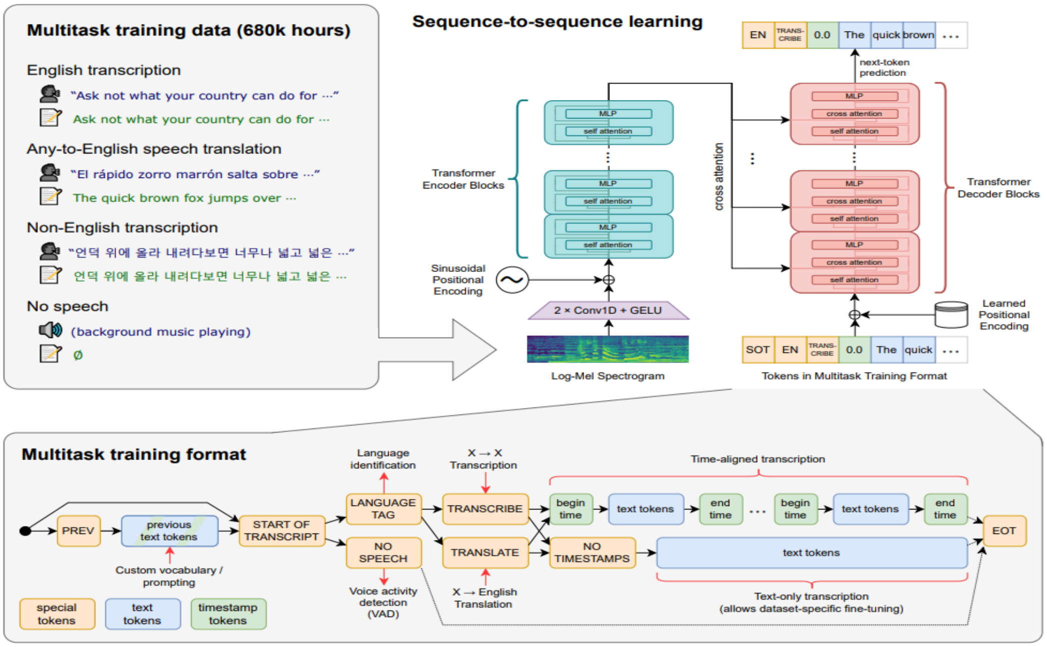

Multitask training data (680k hours)
English transcription
"Ask not what your country can do for ···"
Ask not what your country can do for ···
Any-to-English speech translation
"El rápido zorro marrón salta sobre ···"
The quick brown fox jumps over ···
Non-English transcription
"언덕 위에 올라 내려다보면 너무나 넓고 넓은 ···"
언덕 위에 올라 내려다보면 너무나 넓고 넓은 ···
No speech
(background music playing)
ø
Sequence-to-sequence learning
EN
TRANS-CRIBE
0.0
The
quick
brown
...
next-token prediction
MLP
self attention
cross attention
Transformer Encoder Blocks
Transformer Decoder Blocks
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Log-Mel Spectrogram
Learned Positional Encoding
SOT
EN
TRANS-CRIBE
0.0
The
quick
...
Tokens in Multitask Training Format
Multitask training format
Language identification
X → X Transcription
Time-aligned transcription
PREV
previous text tokens
START OF TRANSCRIPT
LANGUAGE TAG
NO SPEECH
TRANSCRIBE
TRANSLATE
begin time
text tokens
end time
...
begin time
text tokens
end time
NO TIMESTAMPS
text tokens
EOT
Custom vocabulary / prompting
Voice activity detection (VAD)
X → English Translation
Text-only transcription (allows dataset-specific fine-tuning)
special tokens
text tokens
timestamp tokens

Figure 5. Overview of our approach. A sequence-to-sequence Transformer model is trained on many different speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection. All of these tasks are jointly represented as a sequence of tokens to be predicted by the decoder, allowing for a single model to replace many different stages of a traditional speech processing pipeline. The multitask training format uses a set of special tokens that serve as task specifiers or classification targets.
Source: Radford, A. (2022, December 6). Robust speech recognition via Large-Scale Weak Supervision.[20].

Integration of Models in the System.

The integration of multiple models in our system arises from the need to address various aspects and for the management and processing of the sessions held in the municipal council. In this path we employed to perform a management of all models with Python [22] where natively these machine learning models are developed, thus preserving the workflow with each model. On the other hand, for the management of the entire application (Backend) [23], i.e., sessions, documents, audios, texts, users and other processes that encompass a cloud system will be handled by Go [24], and for the management of interfaces and user experience (Frontend) [23], it will go hand in hand with Nextjs/React [25], as we can see in Figure 6.

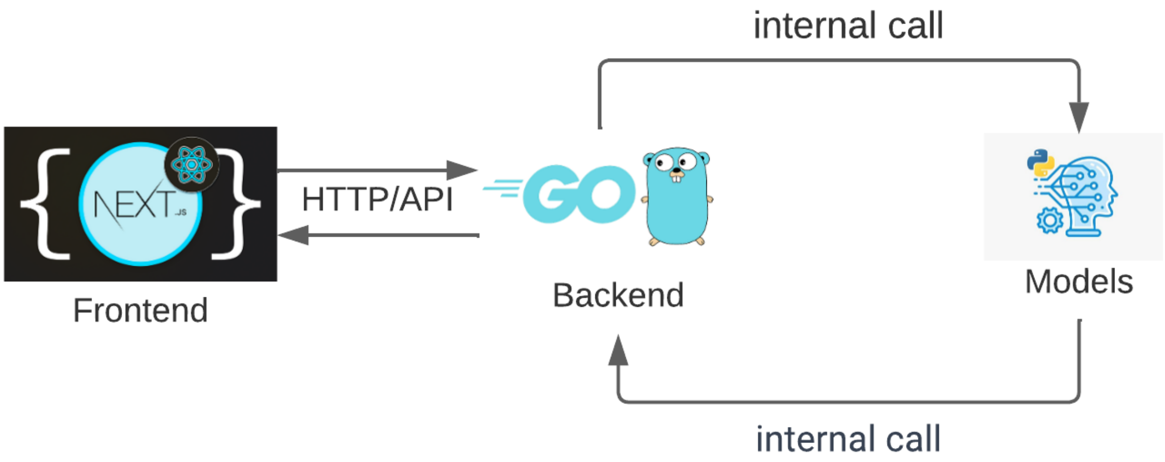


Figure 6. System architecture.
Source: Authors

Analysis and results

Track separation speed with Deezer/Spleeter

In order to analyze the speed of track separation, a study was carried out with the purpose of isolating the voices from other possible sounds present in the room. To achieve this purpose, specific simulated track separation scenarios were designed, each of which was composed of two audio tracks. These scenarios were repeated a total of 196 times, resulting in a set of 392 separate audio tracks.

To evaluate the results, the arithmetic mean equation is used, which states that the mean is calculated by adding all the results and dividing by the number of iterations:

$$\boldsymbol{Media_{seconds}} = \frac{\sum_{i+1}^{N} \ (\boldsymbol{Time_{end}} - \boldsymbol{Time_{start}})_i}{\boldsymbol{N}}$$

Where N represents the total number of iterations and $(\boldsymbol{Time_{end}} - \boldsymbol{Time_{start}})_i$ is the result of each individual iteration, where $\boldsymbol{Time_{start}}$ is the start time of the track separation process, and $\boldsymbol{Time_{end}}$ is the time when the process has finished at the time of making the separation.

In this way, we can generate a table representing each of the iterations in detail in Table I.

Table I. Average Processing Time per File - Spleeter.

#	Name	Repetition	Calculation Average (Mean) (seconds)	Duration of the track(hh:mm:ss.mm)
1	audioTest1	196	6230.0	00:11:14.53
2	audioTest2	196	3579.0	00:04:21.47

Source: Authors

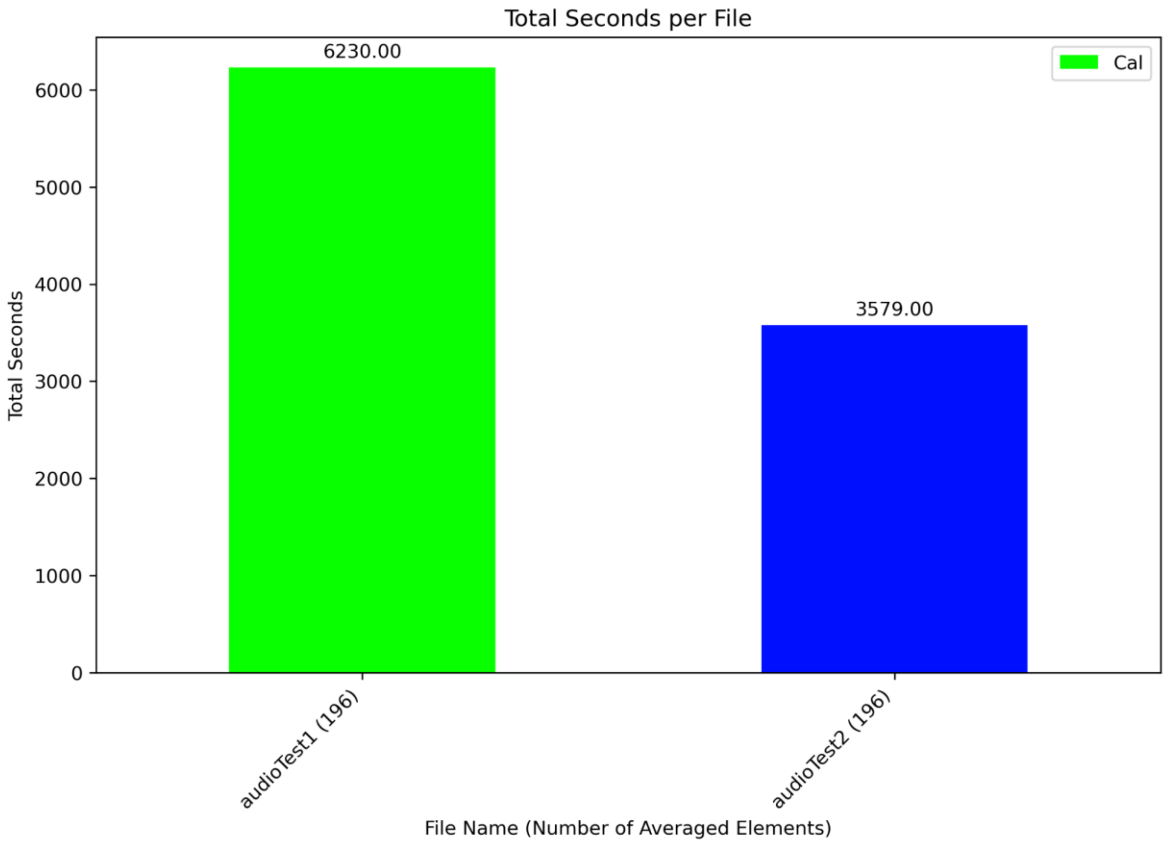


Figure 7. Spleeter run result graph.
Source: Authors

As we can see, the average conversion time for each audio was 6,230.00 seconds for the first audio, which has a duration of 11 minutes and 14 seconds, and 3,579.00 seconds for the second audio, which has a duration of 4 minutes and 21 seconds.

For the first audio track, we can calculate the average separation time using the following equation: $\frac{Total}{N} = T$ What it would be: $\frac{6230.0}{196} = 31.78$, this means that, on average, the separation of voices and other sound waves for audio track 1, which has a duration of 11 minutes and 14 seconds, can take approximately 31.78 seconds.

This information gives us an idea of the average time required to perform track separation on the first audio.

Audio to text transcription with Whisper.

In order to analyze the speed and effectiveness of audio transcription, a test will be conducted using the two aforementioned audios. In this evaluation, the time required to transcribe each audio will be investigated and a score will be assigned to the text obtained in comparison with the original audio content.

During this research, key problems that can arise when transcribing long municipal council sessions will be identified. In addition, solutions and improvements will be proposed to address these challenges and optimize the transcription process.

Table II. Average Processing Time per File - Whisper.

#	Name	Repetition	Calculation Average (Mean) (seconds)	Duration of the track(hh:mm:ss.mm)
1	audioTest1	1	1317.0	00:11:14.53
2	audioTest2	1	485.0	00:04:21.47

Source: Authors

As can be seen in Table 2, the average time for audioTest1, which has a duration of 11 minutes and 14 seconds, was approximately 22 minutes. This indicates a considerably high time, which implies that there are still fragments of audio to be processed by the model. In contrast, audioTest2, with a duration of 4 minutes and 21 seconds, had an average time of approximately 8 minutes. This substantial difference between processing times provides clues to possible areas of improvement that could be explored in future research.

However, when reviewing the transcription, an obvious problem in the transcription process could be identified. In the case of audioTest1, multiple voices were found, some of which present difficulties in terms of tonality and clarity of speech. This may be due to the quality of the microphone used during the recording session, resulting in the speaker's words becoming less audible and clear to the listener. This lack of clarity in the recording makes it difficult for both humans and the transcription model to understand when interpreting and transcribing the words spoken during the session.

Another difficulty that could be observed is that, when using audio files as long as 11 minutes, the quality of the transcription and the length of the transcribed parts decrease significantly. This results in not being able to transcribe the entire audio effectively.

Conclusions

The step-by-step process followed in this research has yielded significant results in building the architecture and workflow to achieve the goal of audio-to-text transcription of Planeta Rica city council sessions. From the literature review to the testing of the models and the search for efficiency in the completion of the transcriptions, each stage has contributed significantly to the success of this project.

However, we cannot overlook the difficulties that arise when trying to set up an efficient system for model execution. The choice of the right hardware can have a significant impact on the execution times of each model for a specific task. In addition, meticulous monitoring of each step in the process influences the quality of the final result. In this way, a cleaner and clearer text output can be achieved, tailored to the specific objectives to be achieved at a given point in time.

On the other hand, it is important to highlight the potential of these models to streamline various processes in public entities and organizations that are often faced with tedious, time-consuming and complicated tasks, such as audio-to-text transcription. These models become an invaluable tool, as they greatly simplify the process. Instead of involving a series of steps that include writing, reading, writing, reviewing and proofreading, thanks to these artificial intelligence models, the whole process can be simplified to a single review and proofreading stage.

Therefore, the combination of these models, supported by a robust architecture and processes that facilitate their use and effectiveness, as in the case of transcription models such as Whisper, is very useful. This provides a wide range of applicability in audio and text processing. It is strongly recommended to implement processes that simplify audio management when audio is extensive. If possible, the use of more efficient recording systems is also suggested, ensuring better audio quality for processing by the models and a more satisfying listening experience for the listeners.

References

[1] Departamento Administrativo de la Función Pública (2020, marzo 20) - Concept 97471 of 2020.Disponoble en: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=127682

[2] Shelley (2023, September 19). How much will AI cost in 2022? Developers.

[3] A. Samsukha, (2023, June 1) The rise of Speech AI: a Game-Changer in the Tech world (n.d.). Nasscom | the Official Community of Indian IT Industry.

[4] I. Manco, E. Benetos, E. Quinton, y G. Fazekas,“Papers with Code - Contrastive Audio-Language Learning for Music” Agosto 2022.

[5] P. Flach,“Machine Learning: The Art and Science of Algorithms that Make Sense of Data” 2012.

[6] MultiComp Lab (2017, October 4). Multimodal Machine Learning | MultiComp. MultiComp | MultiComp Lab's Mission Is to Build the Algorithms and Computational Foundation to Understand the Interdependence Between Human Verbal, Visual, and Vocal Behaviors Expressed During Social Communicative Interactions.

[7] C. Chen, D. Han, and J. Wang, "Multimodal EncoderDecoder Attention Networks for Visual Question Answering," IEEE Access, pp. 1-1, 2 2020.

[8] Papers with Code - VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text (2021, April 22).

[9] R. Merritt, (2022, April 19). What is a Transformer Model? [9] NVIDIA Blog. Official NVIDIA Latin America Blog.

[10] What are convolutional neural networks? IBM (n.d.).

[11] Chan, W. (2015, August 5). Listen, attend and spell. arXiv.org.

[12] D. Amodei, (2015, December 8). Deep Speech 2: End-to-End speech recognition in English and Mandarin. arXiv.org.

[13] J. Llisterri, (n.d.). The synthesis units.

[14] Olivier M. Emorine and Pierre M. Martin. 1988. The MULTIVOC text-to-speech conversion system. In Proceedings of the second conference on applied natural language processing (ANLC '88). Association for Computational Linguistics, USA, 115-120.

[15] MHTTS: Fast Multi-head Text-to-speech For Spontaneous Speech With Imperfect Transcription (2022, October 1). IEEE Conference Publication | IEEE Xplore.

[16] Van Den Oord, A. (2016, September 12). WaveNet: a generative model for raw audio. arXiv.org.

[17] R. Hennequin, A. Khlif, F. Voituret, M. Moussallam, “Spleeter: a fast and efficient music source separation tool with pre-trained models. Journal of Open-Source Software, 5(50), 2154,2020.

[18] C.A. Louis, C. A. Ancy, “Research on DNN Methods in Music Source Separation Tools with emphasis to Spleeter. International Research Journal on Advanced Science Hub, 3(Special Issue 6S), 24-28, 2021.

[19] O. Ronneberger, “U-NET: Convolutional Networks for Biomedical Image Segmentation”. Computer Vision and Pattern Recognition, 2015.

[20] A. Radford, J. Wook Kim, T. Xu, G. Brockman, C. McLeavey, y I. Sutskever, “Robust speech recognition via Large-Scale Weak Supervision”.2022.

[21] OpenAi About (n.d.).

[22] Python documentation (n.d.).

[23] Maldeadora (2018). What is Frontend and Backend: characteristics, differences and examples. Platzi.

[24] Documentation - The Go Programming Language (n.d.).

[25] Docs (n.d.). Next.js.