International Journal of Speech Technology (2024) 27:613–635
https://doi.org/10.1007/s10772-024-10126-4

Automatic transcription system for parliamentary debates in the context of assembly of the republic of Portugal

Pedro Nascimento[1,2] · João C. Ferreira[2,3] · Fernando Batista[2,4]

Received: 1 April 2024 / Accepted: 1 July 2024 / Published online: 17 July 2024

Abstract

The transcription of parliamentary proceedings is essential for democratic governance. Traditional methods are manual and time-consuming. This work introduces an Automatic Transcription System for the Assembly of the Republic of Portugal (STAAR) that uses an automatic speech recognition model and speaker diarization technologies. STAAR was developed after analyzing existing technologies and the Assembly's specific needs, leading to an effective solution that integrates with current processes. STAAR stands out for its efficiency in transcribing debates and adapting to parliamentary language nuances. It significantly exceeded expectations by presenting a low transcription error rate, ranging from 1.7 to 11.3%, depending on the context and speech style, reducing the time required to produce the official parliamentary debates journal, and improving overall transcription efficiency. Additionally, STAAR enabled the transcription of previously undocumented parliamentary committee meetings, enhancing the documentation of parliamentary activities. This achievement marks a significant step in modernizing parliamentary processes, increasing transparency and accessibility of political information, and positions the Portuguese Parliament at the forefront of technological innovation in parliamentary debates transcription.

Keywords Automatic transcription · Parliamentary debates · Automatic speech recognition · Natural language processing · Machine learning · Large language model · Speaker diarization

1 Introduction

The evolution of democratic governance is deeply intertwined with the accessibility and transparency of its parliamentary proceedings. These sessions, often dense with policy discussions, debates, and decisions, form the backbone of the legislative process. However, the traditional means of documenting these proceedings involve manual transcription and presents significant challenges. This labor-intensive process not only demands a considerable amount of time and specialized human resources but also often leads to delays in making these vital discussions publicly accessible. Such delays, in turn, can make it harder for the public to be informed and involved in time, which is especially important for a healthy democracy.

Recognizing this challenge, this work focuses on the development of an innovative solution that leverages advanced speech-to-text and speaker diarization technologies. The Automatic Transcription System (STAAR) represents a significant step towards enhancing the efficiency, accuracy, and accessibility of parliamentary documentation in the Assembly of the Republic of Portugal. The motivation for this work is centered in the broader context of technological advancement in organizational and institutional settings. The rapid development in the field of speech recognition and natural language processing in recent years has opened new avenues for improving various business and governance processes. In particular, the application of automatic speech-to-text technologies holds immense potential in revolutionizing how parliamentary debates are transcribed and archived.

This work has three main objectives. Firstly, to identify and analyze different speech-to-text technologies for their applicability in transcribing parliamentary debates, with a special focus on the Portuguese language, which involves

✉ Pedro Nascimento
pmddn@iscte-iul.pt

[1] Assembleia da República, 1249-068 Lisbon, Portugal

[2] Instituto Universitário de Lisboa (ISCTE-IUL), 1649-026 Lisbon, Portugal

[3] Inov Inesc Inovação—Instituto de Novas Tecnologias, 1000-029 Lisbon, Portugal

[4] INESC-ID, 1000-029 Lisbon, Portugal

a comprehensive review of the state of the art in automatic parliamentary debate transcription. Secondly, to create an automatic transcription system specifically tailored to the context of the Assembly of the Republic of Portugal. This objective comprehends the challenges and requirements gathered from the team that currently undertakes manual transcriptions, aiming to develop and implement a system that not only enhances efficiency but also paves the way for the continuous evolution of transcription technology. Thirdly, to incorporate and optimize speaker diarization technology within the system. This involves developing a capability to accurately identify and attribute speech segments to individual speakers within the parliamentary debates, a critical aspect for ensuring the clarity and utility of the transcriptions in capturing the dynamic nature of parliamentary discussions.

Concerning the research novelty, the development of STAAR represents a major advancement in automatic transcription systems designed for parliamentary use. Unlike general ASR systems, STAAR is specifically built to handle the unique language and procedural details of the Portuguese Parliament. Key features include speaker diarization, which accurately identifies different speakers in a lively debate setting, and ASR technology adapted to manage interruptions, background noise, and specialized parliamentary language. By using an advanced speech recognition model, STAAR achieves exceptional transcription accuracy, far better than traditional methods. This research not only improves the efficiency and accuracy of parliamentary transcription but also sets a new standard for automated documentation in legislative settings, increasing transparency and accessibility of parliamentary proceedings.

This study describes the journey of conceptualizing, designing, and implementing STAAR, a system that not only simplifies the transcription process but also adapts to the unique linguistic nuances of parliamentary language and specific terms. The reminder of this article details the methodological approach, the development phases of STAAR, its evaluation in terms of accuracy and efficiency, and the implications of implementing such a system in the context of modern democratic processes. By doing so, it aims to contribute to the growing field of speech processing technology, particularly in the realm of governmental and parliamentary settings.

2 Methodology

The development of this work was conducted following the Design Science Research Methodology (DSRM) (Peffers et al., 2007), an approach that combines theory and practice to solve specific problems through the creation and evaluation of artifacts. DSRM is characterized by an iterative research cycle, starting with the identification and understanding of a problem, followed by the design and construction of a solution, and culminating in the evaluation of that solution in a real-world context.

As illustrated in Fig. 1, DSRM comprises several interrelated steps that guide the researcher from problem identification to the evaluation and refinement of the proposed solution, along with the three interactions that were performed in order to design, develop and refine the artifact created, STAAR.

To ensure that the work development meets the identified needs, it is essential to assess the degree to which the previously established requirements in the first stage of the DSRM have been fulfilled. The methodology chosen for this evaluation is the ISO 15504's NLPF scale (14:000 ISO &

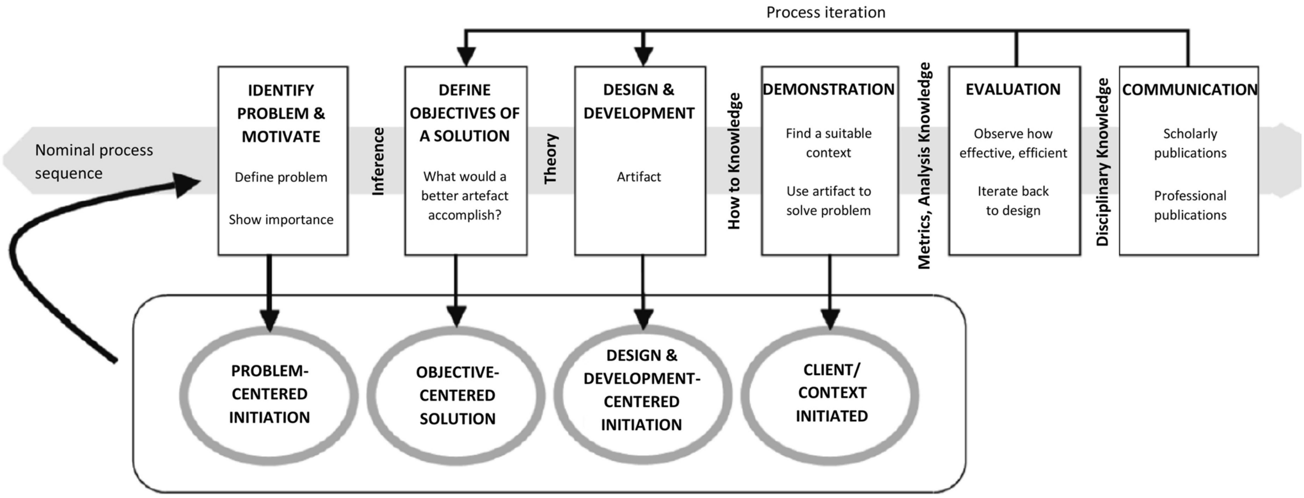

Process iteration
Nominal process sequence
IDENTIFY PROBLEM & MOTIVATE
Define problem
Show importance
Inference
DEFINE OBJECTIVES OF A SOLUTION
What would a better artefact accomplish?
Theory
DESIGN & DEVELOPMENT
Artifact
How to Knowledge
DEMONSTRATION
Find a suitable context
Use artifact to solve problem
Metrics, Analysis Knowledge
EVALUATION
Observe how effective, efficient
Iterate back to design
Disciplinary Knowledge
COMMUNICATION
Scholarly publications
Professional publications
PROBLEM-CENTERED INITIATION
OBJECTIVE-CENTERED SOLUTION
DESIGN & DEVELOPMENT-CENTERED INITIATION
CLIENT/ CONTEXT INITIATED

Fig. 1 DSRM process, Peffers et al. (Peffers et al., 2007)

IEC, 15504-2:2003, 2003), which is divided into four levels (El Emam, 1998):

- Not Achieved (N): The requirement has not been met, or it has been met in a limited way;
- Partially Achieved (P): The requirement has been partially met, but there are still significant gaps or areas needing improvement;
- Largely Achieved (L): The requirement has been largely met, with some areas needing improvement;
- Fully Achieved (F): The requirement has been completely met, with no gaps identified.

This scale provides a precise qualitative analysis, allowing for a clear view of which requirements have been fully satisfied and which still need adjustments or improvements.

In terms of measuring the accuracy of speech recognition systems, one of the most common metrics is the Word Error Rate (WER). This metric provides an objective quantification of the quality of the transcription, allowing comparison between different systems or versions of a system and evaluating improvements over time. The WER is calculated by comparing the automatically generated transcription with a reference transcription, usually done by a human, according to the Eq. 1.

$$\text{WER} = (\text{substitutions} + \text{deletions} + \text{insertions})/\text{words in reference} \tag{1}$$

In this formula, “substitutions” refer to the number of words in the automatic transcription that need to be changed to match the reference transcription, “deletions” are words that are in the reference transcription but do not appear in the automatic transcription and “insertions” are words that appear in the automatic transcription but not in the reference transcription.

During the various iterations of the Design Science Research Methodology (DSRM) in this work, the accuracy of the automatically generated transcriptions needed to be quantified. An online tool from AmberScript (WER, 2023) was employed for this purpose. This tool allows for a direct comparison between professional manual transcriptions and those produced by automatic transcription systems, numerically quantifying the differences using the Word Error Rate (WER), allowing an objective and quantitative evaluation of the effectiveness and accuracy of the developed automatic transcription system, STAAR.

3 Related work

The field of automatic speech-to-text transcription has experienced significant advancements and increasing interest in recent years. With the proliferation of speech recognition technologies and enhancements in natural language processing algorithms, there has been substantial progress in automating the transcription of spoken language. This section reviews the literature related to automatic parliamentary debate transcription, providing a comprehensive analysis of existing technologies, methodologies, and their applications. It highlights the systematic approach taken to identify and evaluate relevant research, and underscores the importance of leveraging cutting-edge speech recognition tools to improve the efficiency and accuracy of transcribing parliamentary proceedings.

3.1 Literature review

Automatic speech-to-text transcription has been an area of research with considerable interest and rapid development in recent years. Widespread access to speech recognition technologies and advances in natural language processing algorithms have allowed for considerable progress in this field.

In order to analyze the state of the art in automatic parliamentary debate transcription, a systematic review was performed using three databases recognized for their wide coverage of scientific and technical literature: Scopus (Scopus—Document Search, 2023), IEEE Xplore (IEEE Xplore—Document Search, 2023) and WoSCC (Web of Science Core Collection—Document Search, 2023). The review was conducted using PRISMA (Page et al., 2021) methodology, which provided a structured approach to identify, select, and critically evaluate relevant research. In addition to these database, other sources of knowledge such as relevant online resources and citations from experts in the field were also included.

Appropriate search terms, specifically designed to address the research scope, were used. These terms focused on three key dimensions: ‘technology’ (related to automatic speech recognition), ‘purpose’ (involving transcription and diarization), and ‘scope’ (to consider the specifics of parliamentary debates). Based on the three identified dimensions and terms, a search string was created and used across the selected knowledge databases:

((*asr* OR *nlp* OR *speech-to-text* OR “*speech recognition*”) AND (*transcription* OR *diarization*) AND (*parliament**)).

To ensure the inclusion of recent studies in the review, only results published from 2013 to May 2023 were considered. Additionally, an additional filter was set to include studies written in Portuguese or English, due to the availability and prevalence of scientific literature in these languages in the context of research on automatic speech recognition.

The initial selection of publications was conducted, in line with the PRISMA methodology, by removing the duplicates from the searched sources. Then, an analysis of the titles and abstracts was conducted to assess their relevance and suitability for the purpose of the work. Subsequently, the availability of the full text of the articles for review was verified, including only those where this analysis was possible.

A comprehensive reading of the documents resulting from the screening was then conducted to evaluate their suitability according to the inclusion criteria, which allowed for the identification and inclusion of the most pertinent and informative studies in the field. Table 1 shows the number of studies included in each phase of this process.

Based on this analysis, Table 2 was created, summarizing the most relevant points of each study, thus contributing to a better understanding of the topic of automatic transcription and providing a solid foundation for this work.

The evolution of automatic transcription, particularly in the context of parliamentary debates, has been marked by significant technological advancements and paradigm shifts.

In the field of speech transcription, the study of De Wet et al. (Wet et al., 2016), published in 2016, describes the development of specific speech resources for South African English, a language with many nuances and variations that present unique challenges for automatic speech recognition (ASR) systems. Given these challenges, the authors focused on adapting ASR to accurately capture the peculiarities of this dialect. To this end, they used the Hidden Markov Toolkit (HTK) (HTK Speech Recognition Toolkit, 2023), a comprehensive tool that allows modeling sequences, such as speech time series. In addition to HTK, the research also relied on pronunciation dictionaries, specifically created for South African English (NCHLT and SAE), which provided the necessary linguistic support to train and refine the ASR model. The combination of these tools and resources led to significant improvements in transcription accuracy, demonstrating the effectiveness of specialized approaches in specific linguistic environments. By enhancing the detection and transcription of South African English speech, De Wet et al. (2016) demonstrated that with the right tools and resources, it is possible to adapt ASR systems to various languages and dialects.

The introduction of hybrid models addressed the challenges of purely probabilistic models. In 2017, Mansikkaniemi et al. (2017) focused on the automatic construction of a speech corpus specifically for the Finnish parliament. The authors faced several challenges in creating the corpus, including limited access to large-scale general domain training data and the need for enormous amounts of transcribed

Table 1 Number of studies included after filtering

Filter	Databases search	Web search
Search string	60	–
Unique studies	35	–
Title/Abstract analysis	22	–
Availability of full text	5	4

Table 2 Summary of most relevant points of analyzed studies

Study	Techniques	Approach[a]	Data	Language	Purpose[b]	Error rate (%)
Zhao et al. (2023)	LLM	NN	–	Multilingual	–	–
Vos and Verberne (2023)	wav2vec2.0 transformer	NN	56,300 h	English	T	17.9
Diáz-Munió et al. (2021)	4-g FairSeq transformer	P, NN	1,300 h	English	T	7.0 ~ 7.9
Aguiar de Lima and Costa-Abreu (2020)	Support Vector Machine Hidden Markov Models Convulsion and Deep Neural Networks	P, H, NN	–	Portuguese	–	–
Alumaë et al. (2018)	Kaldi Toolkit LIUM SpkDiarization Toolkit	H, NN	690 h	Estonian	D, T	8.1
Kawahara (2018)	Statistic Machine Translation HMM training with Maximum Likelihood	P, NN	200 h	Japanese	T	10.0
Mansikkaniemi et al. (2017)	Levenshtein Algorithm DNN acoustic models Speech-to-text alignment	H	2,000 h	Finnish	T	5.9 ~ 18.7
Wet et al. (2016)	Hidden Markov Toolkit Dictionaries (NCHLT and SAE)	P	105 h	South African English	T	–
Campr et al. (2014)	Gaussian Mixed Models Voice Activity Detection Maximum Expectation Models	P	30 h	Czech	D	7.2

[a] Probabilistic (P), Hybrid (H), Neural Network (NN)

[b] Diarization (D), Transcription (T)

speech data. By combining probabilistic models, such as the Levenshtein algorithm, with deep neural network (DNN)-based acoustic models, the study was able to create a transcription corpus from the data available on the website of the Parliament of Finland.

Following extensive use of hybrid models, technological evolution and advances in research led to the emergence of end-to-end neural networks, which promised to simplify the architecture of ASR systems by eliminating the need for multiple processing steps and allowing for a more direct and integrated approach to automatic transcription. Neural networks, with their ability to learn complex patterns, began to dominate the field of ASR.

In 2018, Alumaë et al. (2018) published a paper describing the development of an advanced transcription system for the Estonian language using the Kaldi toolkit (Kaldi: Kaldi, 2023), an open-source software for speech recognition that supports various types of neural networks. One of the challenges faced by the system was dealing with data recorded "in natural environments," such as interviews and meetings recorded in adverse acoustic conditions, which are common in the real world. To overcome these challenges, the system was trained with a diverse variety of speech data and employed techniques like noise reduction to improve results. The study demonstrated the effectiveness of neural networks in transcription, achieving remarkable accuracy with a Word Error Rate (WER) of 8.1%. According to the authors, the developed system had the potential to adapt to other languages, as it utilizes a combination of acoustic and linguistic models that can be trained with speech data in other languages.

Also in 2018, Kawahara (2018) focused on the automatic transcription of Japanese parliamentary meetings, but with a different approach from that used by Alumaë et al. (2018), mentioned earlier. Kawahara (2018) employed probabilistic methods, such as statistical machine translation (SMT) and HMM training based on maximum likelihood (ML) criteria, as well as exploring the capabilities of neural networks, to demonstrate how the combination of different techniques can lead to superior results. The ASR system was developed using a combination of open source with proprietary software and was trained with a large corpus of audio data from previous parliamentary meetings. The system was evaluated through long-term operation in the environment of the Japanese parliament and recorded an accuracy rate of about 90%. One of the main challenges faced by the author was the variability of speech in parliamentary meetings, which includes disfluencies, fillers, and colloquial expressions. To address this challenge, the author adopted a sustainable approach that combines automatic transcription and manual editing to generate a faithful transcription of the meeting.

At the European level, in 2021, Diáz-Munío et al. (2021) made a significant step in the area of automatic speech recognition (ASR) by creating the Europarl-ASR, a remarkable corpus not only for its size, with 1300 h of transcribed parliamentary speeches, but also for its linguistic diversity, reflecting the various languages spoken in the European Parliament (Multilingualism in the European Parliament, 2023). Working with such a diverse corpus, Díaz-Munío et al. (2021) also faced challenges related to the wide variety of speech styles, accents, linguistic nuances, and specific technical terminology, as well as cultural and historical references that can vary from one language to another. To overcome these issues, the authors used a combination of deep neural networks (DNNs) and recurrent neural networks (RNNs). The DNNs, with their ability to learn hierarchical representations of data, were crucial in capturing the acoustic features of speech. On the other hand, the RNNs, with their ability to model temporal sequences, were essential in understanding the structure and sequence of the speeches. The results of the study were promising, achieving error rates between 7 and 7.9%, demonstrating that neural networks are not only capable of accurately transcribing parliamentary debates but also of adapting and generalizing to various languages, although Portuguese was not evaluated nor is it planned for future inclusion, according to the conclusions presented.

Recently, in 2023, the Europarl-ASR was used as a training and evaluation model in a study conducted by de Vos and Verberne (2023), which created a corpus with transcriptions from the LIBE Committee of the European Parliament (LIBE | Committees | European Parliament, 2023), totaling 3.6 million words. The main focus of the study was to explore and optimize advanced neural network techniques for ASR. Instead of relying solely on the Kaldi toolkit, as in previous studies, the authors incorporated the wav2vec 2.0 transformer model (Baevski et al., 2020) into their ASR pipeline. This model, known for its ability to learn rich audio representations in an unsupervised manner, allowed for an innovative approach to transcribing parliamentary debates. The choice of wav2vec 2.0 was strategic, given the complexity and linguistic diversity of EU debates, making it essential to use a tool capable of capturing subtle nuances in speech. However, even with the incorporation of wav2vec 2.0, the obtained Word Error Rate (WER) was 14.5%, indicating that while the model is effective, there is still room for optimization.

The articles found in this systematic review, which describe the use of ASR systems, do not specifically mention transcription in the Portuguese language, which has its own unique particularities.

In this context, the work of Lima et al. (Aguiar de Lima & Costa-Abreu, 2020), published in 2020, provides a comprehensive perspective on the current state of ASR systems for this language. The study consisted of a systematic review of 101 articles published between 2012 and 2018, using the PRISMA methodology, with the aim of identifying the most used ASR techniques, as well as assessing the challenges faced in developing ASR systems for Portuguese, such as the lack of resources and the high variability of the language. The authors found that most of the scientific research on ASR for Portuguese focuses on European Portuguese, with little work conducted on Brazilian Portuguese or other variations of the language. It is also noted that, while some articles aim to collect Portuguese speech data, most do not publish it freely, making it difficult to utilize more sophisticated techniques like DNNs and CNNs, which require a large amount of data. The authors conclude that despite deep neural networks (DNN) becoming the most common technique for ASR in recent years, some innovative methods for ASR in Portuguese remain unexplored, such as transfer learning or unsupervised learning.

Another particularly interesting and current study was conducted by Zhao et al. (2023), published in 2023, which consists of a systematized literary review on techniques and innovative applications with large language models (LLMs). The research emphasizes the importance of scaling models to achieve maximum performance in natural language processing tasks. The authors assessed the current challenges in scaling linguistic models, including computational resources and data availability, which is always the most difficult to obtain, and how these challenges can be overcome through distributed training techniques and data augmentation. In terms of technology and language models evaluated, the study covers a wide range of models like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), T5 (Ma et al., 2023), and GPT (OpenAI GPT-4 Technical Report, 2023). Regarding the latter model, the authors recognize the level of excellence of GPT-4 in solving general tasks and its robustness against noise or disturbances. Empirical results from the study show that GPT-4 outperforms the other models evaluated in a wide variety of tasks, such as language comprehension. One of the conclusions in the publication is that large language models have revolutionized the field of natural language processing and have the potential to transform many other fields, including scientific research, health, and education, although with reservations regarding ethical and security issues around AI.

One of the objectives of this work is to complement the transcription of parliamentary debates with the identification of speaker changes. This process, known as diarization, involves distinguishing and segmenting an audio or video recording to determine “Who spoke when?”.

In this regard, the studies of Campr et al. (2014), from 2014, and Alumaë et al. (2018), from 2018, are relevant as they provide methodologies and experience that can be applied or adapted to improve the accuracy and effectiveness of diarization in the context of parliamentary debates.

Campr et al. (2014) utilized Gaussian Mixture Models (GMMs) to address the task of speaker diarization. GMMs are probabilistic models that can be trained to recognize and distinguish different speakers based on the unique characteristics of their speech. By combining this with Expectation Maximization (EM) algorithms, the authors were able to effectively segment the recordings and assign segments to individual speakers. One of the main challenges faced was detecting voice activity in parliamentary environments, which can have significant background noise, applause, interjections, and other interruptions. To minimize these issues, the system uses a combination of audio and video, as well as facial recognition technology, to associate individual models of audio and video modalities in an unsupervised manner. The system was evaluated on 30 h of video, specifically on broadcasts of Czech parliamentary meetings. The results show that the proposed combination of individual audio and video diarization systems results in an improvement in the Diarization Error Rate (DER), assessed at 7.2%.

The research by Alumaë et al. (2018), published in 2018, uses only audio for diarization, which is performed using the LIUM SpkDiarization toolkit (SpkDiarization, 2023). This open-source tool employs a combination of clustering and classification techniques to identify speakers in audio segments. According to the authors, this tool enabled them to achieve a 95% accuracy rate in the diarization process.

The related work analysis reveals that creating specific databases for speech-to-text transformation tasks has been a widely adopted strategy in the scientific community. While effective in many cases, this approach presents significant challenges. Building such corpora requires substantial investments in terms of human, financial, and computational resources. Moreover, the need for high-quality and massive quantities of speech data, often scarce or difficult to access, makes the process even more complex and time-consuming.

The availability of advanced language technologies, exemplified by ASR systems such as Whisper, marks a transformative era in this field. These ASR systems, trained on vast datasets and capable of generalizing to a wide variety of tasks and languages, offer a promising alternative to the traditional corpora-based approach.

While the conventional approach of creating specific corpora remains relevant in the research and development of ASR systems, the rise of ASRs, trained akin to Large Language Models (LLMs), offer a more agile and cost-effective pathway for implementing automatic transcription solutions.

For the Portuguese Parliament and similar institutions, this evolution represents a valuable opportunity to improve the accessibility and efficiency of their parliamentary records, benefiting both the institution and the public at large.

3.2 Whisper, an automatic speech recognition model

In September 2022, OpenAI, the company behind ChatGPT (2023) and DALL-E (DALL·E3, 2023), released its new ASR model, Whisper (2023), a multilingual and multitasking system that is increasingly performing at a near-human level. Innovatively, Whisper supports transcription in multiple languages, as well as the translation of these languages into English. OpenAI details that the model was trained on 680,000 h of supervised data, equivalent to over 77 years of continuous audio. According to the study published in 2022 by Radford et al. (2022), Whisper achieves human-level robustness and accuracy when operated on English speech.

OpenAI states that while several other models have been trained to perform well on specific datasets or test sets, with Whisper, they shifted the focus to supervised pre-training using larger datasets created by combining extracted and filtered data from extensive compiled datasets and filtering usable data through smart heuristics. These models tend to be more robust and generalize more effectively to real-world use cases. While some models may outperform Whisper in individual test sets, by scaling its minimally supervised pre-training, the OpenAI team achieved very impressive results without using the self-supervision and self-training techniques commonly employed by current state-of-the-art ASR models.

Another significant advantage of Whisper is that it is free from subscriptions or licensing fees, allowing individuals, companies, organizations, and independent researchers to utilize its capabilities without substantial investments in licenses or expensive products. Whisper can be run locally, offering greater control over the processed data, which is especially important in scenarios where data privacy is a primary concern. Audio data does not need to be sent to external servers for processing, providing a significant advantage in terms of privacy and security. Additionally, the ability to run locally allows for greater customization and adaptation of the model to meet specific needs.

3.3 WhisperX, aligned timestamps and diarization

Large-scale, lightly supervised automatic speech recognition models like Whisper have shown remarkable results in speech recognition across various domains and languages. However, the timestamps associated with each transcription tend to be imprecise and are not available at the word level. Additionally, their use in extensive audio via buffer transcription precludes batch inference due to its sequential nature. To overcome these challenges, Bain et al. (2023) developed and released WhisperX, a voice recognition system with temporal accuracy that, using Whisper, provides word-level temporal records through vocal activity detection and forced phoneme alignment, as illustrated in Fig. 2.

WhisperX not only enables automatic speech recognition but also supports diarization, a process of segmenting and identifying different speakers in an audio recording. The goal of speaker diarization is to divide the audio stream into homogeneous segments, where each segment corresponds to a specific speaker or speaker turn. Essentially, it aims to answer the question “Who spoke when?” throughout an audio recording. This diarization process is conducted using pyannote.audio, an open-source toolkit written in Python by Bredin et al. (2019).

4 Development and implementation

The development and implementation of STAAR for the Assembly of the Republic of Portugal required a meticulous and structured approach to address the unique challenges of transcribing parliamentary proceedings. This section outlines the comprehensive process followed to gather requirements, design a modular architecture, and implement a robust transcription framework. It highlights the collaboration with key stakeholders, the identification of critical needs, and the iterative development phases that ensured STAAR met the specific demands of the parliamentary context. Through a series of formal meetings and technical evaluations, the project aimed to create an efficient and

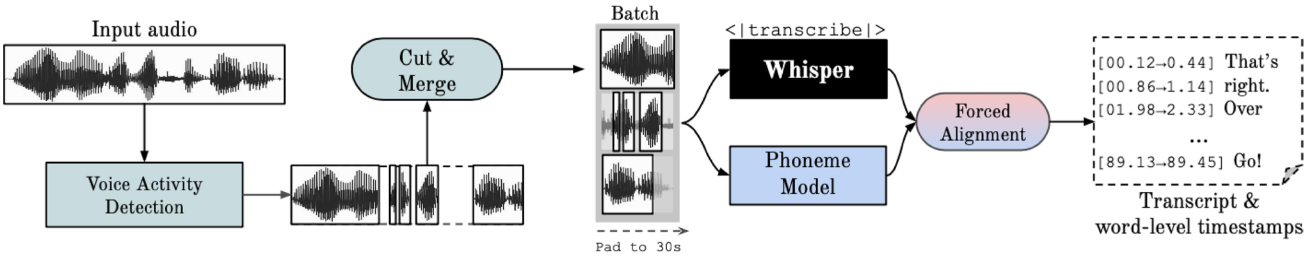


Fig. 2 WhisperX speech-to-text process diagram using Whisper (Bain et al., 2023)

reliable system that enhances the accessibility and accuracy of transcribed parliamentary records.

4.1 Requirements

According to the Design Science Research Methodology (DSRM), identifying requirements is a vital step in creating a system. Between December 2022 and February 2023, five formal meetings were held with key stakeholders, including members of the team that creates the Assembly of the Republic (AR)'s official journal (DAR), the head of the team that records session video and audios, and leadership from the involved areas. Additionally, various contacts were made with the transcription team to better understand the functioning and method of transcribing parliamentary debates.

From these meetings, contacts, previous experiences with other transcription software, and knowledge of AR's information systems, a comprehensive assessment of the specific needs and challenges of the parliamentary context was conducted. This led to a detailed list of requirements for the work, outlined in Table 3. In this list, linguistic, technical, and operational aspects were considered, including the speed and accuracy of transcription, the ability to process existing audios, the use of existing resources and the flexibility to expand to transcribe other parliamentary activities like committee meetings. Another important requirement was the ability to manage specific acoustic challenges, such as background noise, applause, or simultaneous speakers.

4.2 STAAR framework

To address the identified requirements, a modular architecture was designed to optimize the transcription process of parliamentary interventions. This architecture was structured in four sequential stages: Audio Collection, Audio Processing, Text Treatment, and Transcript Storage, as illustrated in Fig. 3. Each stage was designed to address a specific set of needs and challenges, ensuring a smooth and efficient transition from audio to text.

4.2.1 Audio collection

This stage marks the beginning of the process, responsible for gathering audios from various sources including plenary sessions, parliamentary committees, and other relevant events. The Audio Collection lays the groundwork for the subsequent audio-to-text conversion. It involves connecting to various audio sources, determining if the audios already have transcriptions, and if not, making them available for the next step, Audio Processing. The module for this purpose, **Audio Collector**, was developed in Python (The Official Home of the Python Programming Language, 2023) language, known for its efficiency and versatility in data manipulation tasks.

The audio collection process starts by connecting to various sources where the audios are stored, such as network shares. These audios are organized into folders named after

Table 3 List of requirements for the automatic transcription system

#	Problem	Requirement
1	Manual transcription time	Automatic transcription of audios, eliminating the need for manual typing
2	Legibility vs. Verbatim	Create legible and understandable records, not verbatim transcripts
3	Text editing	Allow editing and formatting of transcribed text directly in Word
4	Quick availability	Ensure that transcriptions for plenary sessions are available quickly, ideally within 10 min after audio is released
5	Existing equipment compatibility	Use with existing parliamentary equipment for text review
6	Hardware control (pedal)	Support a hardware pedal to control audio (pause, start)
7	Multiple speakers	Identify speaker changes and associate text accordingly
8	Multiple audio sources	Capable to obtain audio from various sources
9	Ad hoc transcriptions	Allow transcription of audios manually introduced by users
10	Portuguese model	Consider the specifics of Portuguese language
11	Pre-implementation testing	Allow evaluations and tests before full implementation
12	Expansion	Initially for plenary sessions, allow expansion to committee meetings
13	Customizable dictionaries	Enable creation and customization of replacement dictionaries for automatic correction of specific terms and jargon
14	Accuracy	Aim for a low WER (max 15%) to ensure accuracy and reliability
15	Data security	Process transcriptions using AR's resources without cloud solutions
16	Remote work capability	Allow access to the transcripts from various locations
17	Access levels	Provide different access levels to ensure data security and confidentiality
18	Background noise handling	Handle background noise, such as applause or interruptions, to ensure transcription accuracy in challenging acoustic environments

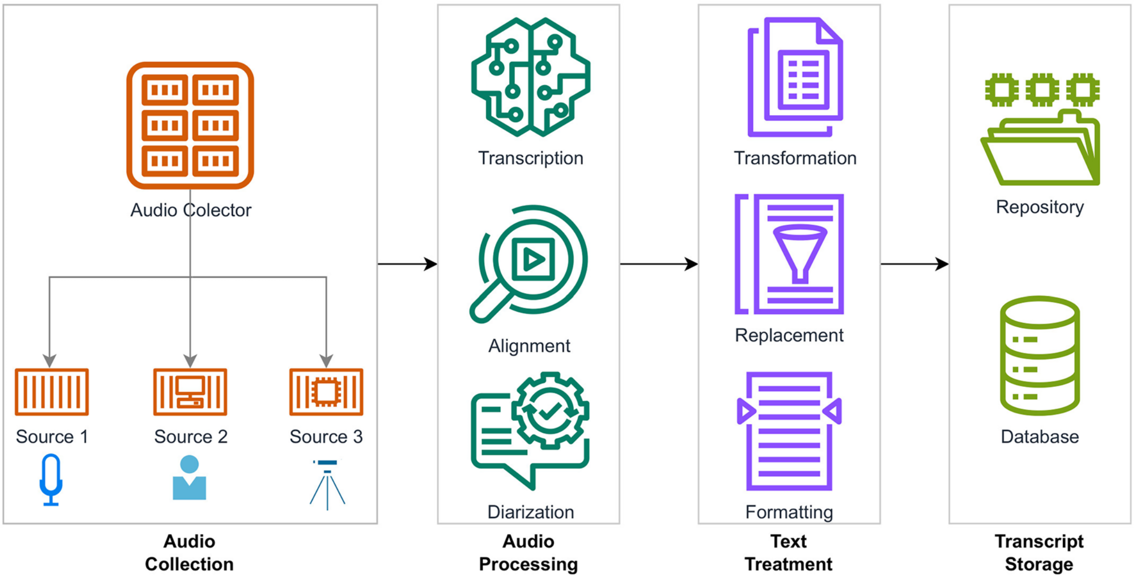


Fig. 3 STAAR's framework, with four stages

the corresponding body, be it plenary, committee, events, or others. Plenary audios are segmented into 15-min fragments, with a typical plenary session lasting about 3 h. Thus, every quarter of an hour, the recording system generates a new audio file.

Once connected to the sources, the system identifies files modified in the last 30 days to accommodate potential delays in audio availability. The identified audios go through a normalization process to standardize file names, using the modification date and time for consistent naming and avoiding duplicates.

The system then checks the repository of transcribed texts to see if the audio has already been processed. If a corresponding transcription is found, the audio is discarded; otherwise, it is transferred to a central repository for transcription. This check prevents duplication and ensures only non-transcribed audios are processed.

Finally, this verification and collection cycle is executed every 30 s to ensure that audios are collected and made available for transcription almost immediately, meeting the need for quick transcriptions in the parliamentary environment.

4.2.2 Audio processing

The Audio Processing is the core stage where the transcription from audio to text is conducted, transforming spoken discourse in parliamentary interventions into cohesive and highly accurate textual transcriptions. Audio processing is conducted in three distinct modules: Transcription, Alignment, and Diarization. Each of these phases play an essential role in ensuring the accuracy and utility of the final transcription.

The **Transcription** module is the heart of the Automatic Transcription System (STAAR). Its main purpose is to transform parliamentary interventions, recorded in audio format, into faithful textual transcriptions. The primary complexity of this module lies in the accurate comprehension and interpretation of spoken discourse. For this task, the decision was made to apply Artificial Intelligence (AI)-based ASR models, based on research conducted in this work, which claim to enable a smooth transcription of spoken discourse into written format. ASR systems, trained with a wide range of linguistic data, can recognize specific words, phrases, and intonations. By analyzing the sound and contextual patterns, the ASR translates the audio into text, creating a textual representation of the original speech. The chosen model for this function was the Whisper ASR system developed by OpenAI (Whisper, 2023). As mentioned in Sect. 3.2, Whisper offers high accuracy and robustness in handling diverse speech patterns, accents, and background noises, which are crucial for the dynamic and often noisy environment of parliamentary sessions. Moreover, prior to the commencement of this work, several preliminary tests were conducted with other ASR tools, including Microsoft Azure, Speechmatics, Telegram, Calligraphus, Audimus Server, and the Word Transcribe feature. These tests concluded that Whisper would be the most promising solution. Amongst the various models offered by Whisper, the large-v2 model was selected for its ability to provide transcriptions for Portuguese, with impressive low word error rate, ranging from 4.3 (FLEURS) and 6.8 (CommonVoice9) (Radford et al., 2022). Python programming was used to check for the presence of audios in the temporary central repository. When an audio is identified, it is submitted to Whisper, which, by analyzing the first 30 s of the audio, detects the language and performs the transcription based on that language. This Whisper feature is particularly useful because, although most of the speeches are in Portuguese language, there are occasions when other languages are used, making Whisper's multilingual capability an asset. Throughout the process, information such as the start and end date/time of the transcription process is collected for later database entry, for the purpose of evaluating the system's performance.

The **Alignment** module is responsible for synchronizing each word with the original audio by establishing precise timestamps for every transcribed word. This strict alignment ensures that the text and audio remain synchronized, preserving the exact temporal sequence of the transcribed audio. While the ASR Whisper model creates transcriptions with timestamps, the alignment between sound and times is not rigorous. Precise alignment is required to maintain the exact temporal sequence of the transcribed audio. Given the gaps in Whisper's sound-to-timestamp alignment, it was necessary to find solutions to this problem. The adopted solution was to use WhisperX (Bain et al., 2023), an extension of Whisper, available on GitHub (Whisper, Approach. Png at Main Openai & Whisper, 2023), which offers several advantages over Whisper. WhisperX not only addresses the abovementioned problems but also performs transcriptions in significant less time, even with Whisper's "large-v2" model. It has lower computational requirements, reduces errors related to hallucinations during the voice activity detection (VAD) process, and has speaker recognition and diarization functions that Whisper does not offer. The alignment process is executed through a function developed in Python code, created based on the product documentation. Metadata is also collected in this process to calculate the performance associated with the alignment.

The **Diarization** module distinguishes the voices present in the audio, assigning identifiers to different speakers. While it does not identify the exact name of the speaker, counting the number of speakers greatly aids in the transcription review process, especially since the audios/texts can be quite lengthy, and having a context for each speaker is significantly helpful. For diarization, WhisperX is used again, which, by providing exact timestamps, makes it possible to attribute speakers to each text segment found. The diarization process is a function of the Python code developed, created based on the documentation of WhisperX, and where final metadata is collected to calculate the performance associated with this stage.

WhisperX allows saving the results of the transcription, alignment, and diarization phases in different formats, including "json", "srt", "tsv", "txt", and "vtt", which can be used by various platforms, such as for transcription or subtitling. For the purpose of transcribing parliamentary debates, at the end of the Audio Processing stage, only the "json" and "srt" files are saved, as exemplified in Fig. 4. The "json" format is the most complete as it contains the transcription with timestamps at the level of each word in a structured manner, while the "srt" only has timestamps at the sentence level, being the format used in the next phase of the transcription process.

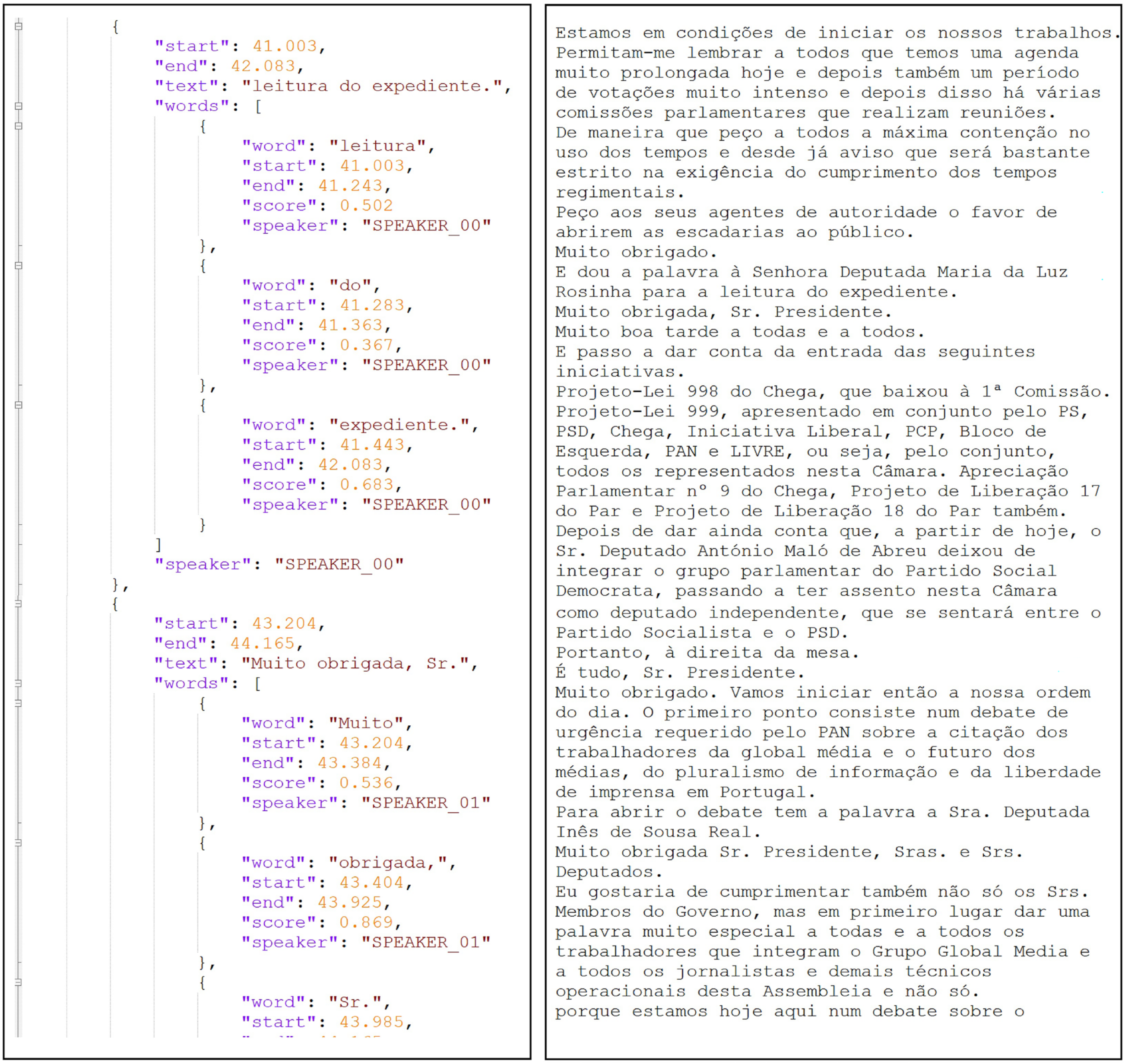
4.2.3 Text treatment

This phase transforms raw transcriptions from the Audio Processing stage into structured, readable documents ready for review. This process involves several activities and techniques to ensure the clarity and accuracy of the final text, by producing a document that closely resembles what the AR transcription team produces. As shown in Fig. 5, the text obtained from the Audio Processing stage is hard to read due to excessive unnecessary information for users creating the Assembly of the Republic's Journal (DAR). After the Text Treatment stage, the text becomes much more legible, similar to DAR format, and ready for review. As Fig. 6 shows, the text is grouped by speaker and the parliamentary jargon terms that were replaced are highlighted in blue, among other changes detailed in this section.

The purpose of the **Transformation** module is to convert the raw text from the Audio Processing stage into a suitable format. This process, programed in Python language, removes unnecessary information like IDs, timestamps, and random phrases that occur when the audio has long audio pauses, focusing on extracting only the content relevant to the preparation of the DAR. At the end of this process, a word count of the transcribed text is conducted to provide valuable statistical data for future analyses.

The **Replacement** module corrects and adjusts incorrectly transcribed words or phrases, especially due to the unique nature of parliamentary language. It uses linguistic rules and a specific dictionary tailored to parliamentary standards for term replacement and capitalization, as shown in Table 4. The process, executed by Python programing, involves two types of text files for case-sensitive and insensitive replacements, improving transcription accuracy and readability while minimizing manual correction efforts. When a replacement is made, the replaced word or expression is prefixed and suffixed with "***" to identify which words were replaced. This process not only improves the accuracy of transcriptions but also saves valuable time for the transcription team, minimizing the need for manual intervention in recurrent errors.

The Formatting module focuses on transforming the transcription files for parliamentary use, converting them into the necessary AR format. Key activities include converting transcriptions to docx format, chosen for its Microsoft Word compatibility and ease of editing, which aligns with existing parliamentary tools and facilitates remote access and



```
        {
            "start": 41.003,
            "end": 42.083,
            "text": "leitura do expediente.",
            "words": [
                {
                    "word": "leitura",
                    "start": 41.003,
                    "end": 41.243,
                    "score": 0.502
                    "speaker": "SPEAKER_00"
                },
                {
                    "word": "do",
                    "start": 41.283,
                    "end": 41.363,
                    "score": 0.367,
                    "speaker": "SPEAKER_00"
                },
                {
                    "word": "expediente.",
                    "start": 41.443,
                    "end": 42.083,
                    "score": 0.683,
                    "speaker": "SPEAKER_00"
                }
            ]
            "speaker": "SPEAKER_00"
        },
        {
            "start": 43.204,
            "end": 44.165,
            "text": "Muito obrigada, Sr.",
            "words": [
                {
                    "word": "Muito",
                    "start": 43.204,
                    "end": 43.384,
                    "score": 0.536,
                    "speaker": "SPEAKER_01"
                },
                {
                    "word": "obrigada,",
                    "start": 43.404,
                    "end": 43.925,
                    "score": 0.869,
                    "speaker": "SPEAKER_01"
                },
                {
                    "word": "Sr.",
                    "start": 43.985,
```

```
Estamos em condições de iniciar os nossos trabalhos.
Permitam-me lembrar a todos que temos uma agenda
muito prolongada hoje e depois também um período
de votações muito intenso e depois disso há várias
comissões parlamentares que realizam reuniões.
De maneira que peço a todos a máxima contenção no
uso dos tempos e desde já aviso que será bastante
estrito na exigência do cumprimento dos tempos
regimentais.
Peço aos seus agentes de autoridade o favor de
abrirem as escadarias ao público.
Muito obrigado.
E dou a palavra à Senhora Deputada Maria da Luz
Rosinha para a leitura do expediente.
Muito obrigada, Sr. Presidente.
Muito boa tarde a todas e a todos.
E passo a dar conta da entrada das seguintes
iniciativas.
Projeto-Lei 998 do Chega, que baixou à 1ª Comissão.
Projeto-Lei 999, apresentado em conjunto pelo PS,
PSD, Chega, Iniciativa Liberal, PCP, Bloco de
Esquerda, PAN e LIVRE, ou seja, pelo conjunto,
todos os representados nesta Câmara. Apreciação
Parlamentar nº 9 do Chega, Projeto de Liberação 17
do Par e Projeto de Liberação 18 do Par também.
Depois de dar ainda conta que, a partir de hoje, o
Sr. Deputado António Maló de Abreu deixou de
integrar o grupo parlamentar do Partido Social
Democrata, passando a ter assento nesta Câmara
como deputado independente, que se sentará entre o
Partido Socialista e o PSD.
Portanto, à direita da mesa.
É tudo, Sr. Presidente.
Muito obrigado. Vamos iniciar então a nossa ordem
do dia. O primeiro ponto consiste num debate de
urgência requerido pelo PAN sobre a citação dos
trabalhadores da global média e o futuro dos
médias, do pluralismo de informação e da liberdade
de imprensa em Portugal.
Para abrir o debate tem a palavra a Sra. Deputada
Inês de Sousa Real.
Muito obrigada Sr. Presidente, Sras. e Srs.
Deputados.
Eu gostaria de cumprimentar também não só os Srs.
Membros do Governo, mas em primeiro lugar dar uma
palavra muito especial a todas e a todos os
trabalhadores que integram o Grupo Global Media e
a todos os jornalistas e demais técnicos
operacionais desta Assembleia e não só.
porque estamos hoje aqui num debate sobre o
```

Fig. 4 Transcript example, after Audio Processing, in "json" (left) and "txt" (right) format

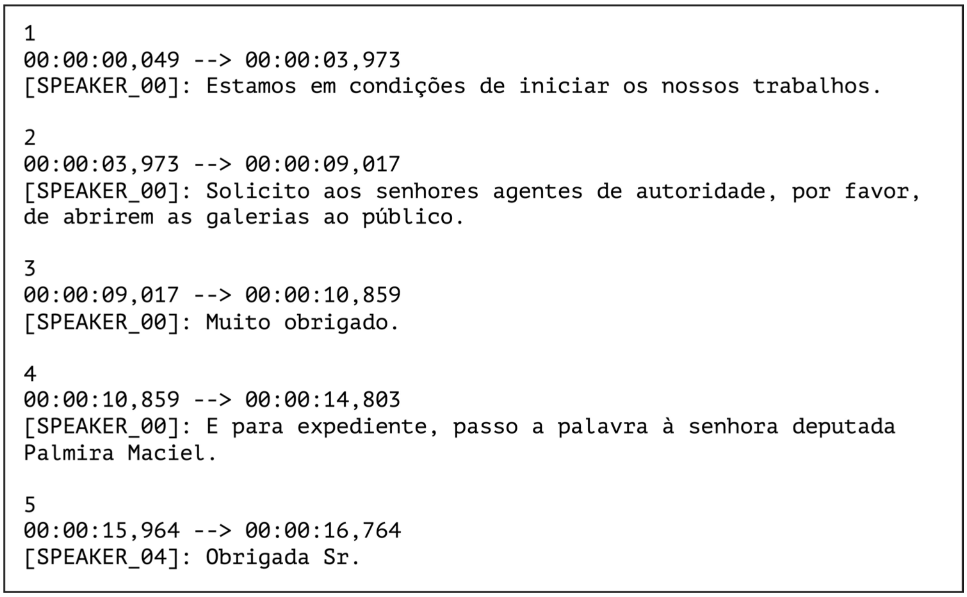
collaboration, identification of substituted terms (marked in blue for easy review), and final adjustments like text justification and font styling. These steps enhance the document's readability and ensure it meets parliamentary standards, both in accuracy and aesthetics.

4.2.4 Transcript storage

The **Storage** module serves as a central repository where processed transcriptions are stored, organized, and managed. A file-sharing network system was chosen for its simplicity and effectiveness in facilitating user access to transcriptions generated by STAAR. This network sharing approach offers significant advantages in terms of accessibility and ease for users, allowing direct access to the transcription repository without additional software installations.

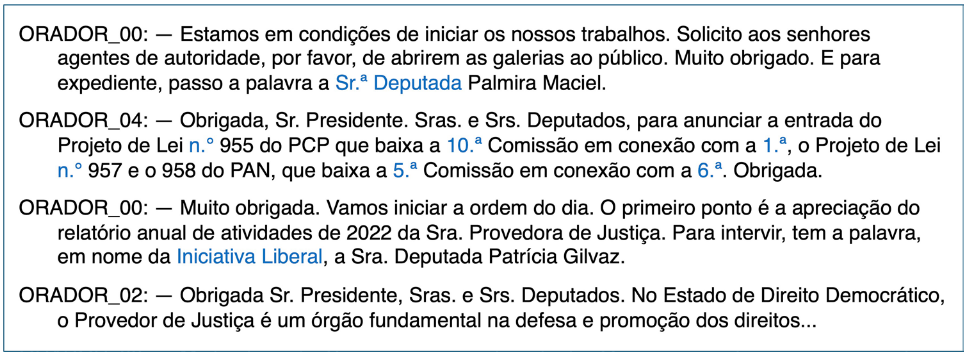
This minimizes technical and operational barriers and easily integrates into the existing work environment. The repository is structured by audio type—plenary, committees, and events—with subfolders for specific committees/events and dates. Each transcribed file, stored in docx format, is named in the "yyyy-dd-mm_hh-mm-ss" format for easy reference. Effective access controls have been implemented for the repository, ensuring that only authorized

Fig. 5 Transcript example, before the Text Treatment, in “srt” format



```
1
00:00:00,049 --> 00:00:03,973
[SPEAKER_00]: Estamos em condições de iniciar os nossos trabalhos.

2
00:00:03,973 --> 00:00:09,017
[SPEAKER_00]: Solicito aos senhores agentes de autoridade, por favor,
de abrirem as galerias ao público.

3
00:00:09,017 --> 00:00:10,859
[SPEAKER_00]: Muito obrigado.

4
00:00:10,859 --> 00:00:14,803
[SPEAKER_00]: E para expediente, passo a palavra à senhora deputada
Palmira Maciel.

5
00:00:15,964 --> 00:00:16,764
[SPEAKER_04]: Obrigada Sr.
```

Fig. 6 Transcript example, after the Text Treatment, in “docx” format



ORADOR_00: — Estamos em condições de iniciar os nossos trabalhos. Solicito aos senhores agentes de autoridade, por favor, de abrirem as galerias ao público. Muito obrigado. E para expediente, passo a palavra a Sr.ª Deputada Palmira Maciel.

ORADOR_04: — Obrigada, Sr. Presidente. Sras. e Srs. Deputados, para anunciar a entrada do Projeto de Lei n.° 955 do PCP que baixa a 10.ª Comissão em conexão com a 1.ª, o Projeto de Lei n.° 957 e o 958 do PAN, que baixa a 5.ª Comissão em conexão com a 6.ª. Obrigada.

ORADOR_00: — Muito obrigada. Vamos iniciar a ordem do dia. O primeiro ponto é a apreciação do relatório anual de atividades de 2022 da Sra. Provedora de Justiça. Para intervir, tem a palavra, em nome da Iniciativa Liberal, a Sra. Deputada Patrícia Gilvaz.

ORADOR_02: — Obrigada Sr. Presidente, Sras. e Srs. Deputados. No Estado de Direito Democrático, o Provedor de Justiça é um órgão fundamental na defesa e promoção dos direitos...

Table 4 Example of terms to be replaced by the Replacement module

Search for	Replace by
administração pública	Administração Pública
décima primeira comissão	11.ª Comissão
22.º governo	XXII Governo
senhor primeiro-ministro	Sr. Primeiro-Ministro
hemiciclo	Hemiciclo

individuals access specific transcriptions, protecting data integrity and confidentiality. While an API was considered for programmatically obtaining transcription files in different formats for integration with other platforms (like video captioning systems), it was not implemented in the current phase, as STAAR’s primary focus is on creating the Assembly of the Republic’s Journal, which requires meticulous transcription editing. Future API implementation may enhance STAAR’s utility and interoperability.

For the **Database** module, Microsoft SQL Server was chosen for its robust and reliable environment, utilizing existing resources. The database structure facilitates efficient queries and provides a clear view of the transcription process from audio collection to transcribed text generation. The ‘transcripts” table stores information related to each transcription by STAAR, with fields and types described in Table 5.

Throughout the transcription process, metadata associated with the audio and transcription are collected and inserted into the “transcript” database table at the end of the process.

4.3 Infrastructure and resources

The success of any technological system relies not only on its design and development but also on the robustness and efficiency of the infrastructure supporting it. This section details the technical and operational aspects that constitute the backbone of the Automatic Transcription System for the

Table 5 Number of studies included after filtering

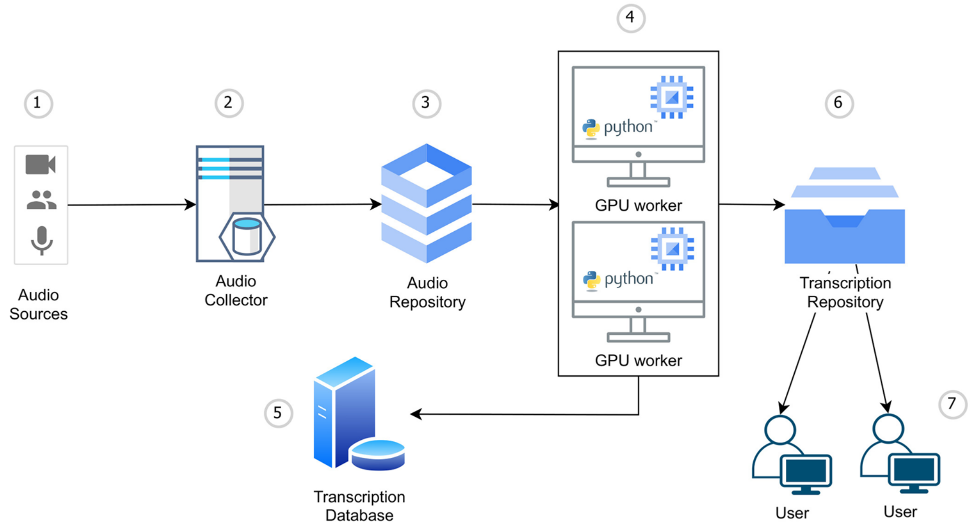
Field	Type	Description
audio_file_name	NVARCHAR	Name of audio file that was transcript
audio_file_path	NVARCHAR	Path where the audio file was stored
audio_date_creation	DATETIME	Date/time of audio file creation
audio_duration	INT	Duration of audio file, in seconds
transcript_file_name	NVARCHAR	Name of transcription file produced
transcript_file_path	NVARCHAR	Path where the transcript file was stored
process_start	DATETIME	Date/time when process was initiated
process_end	DATETIME	Date/time when process was concluded
process_duration	INT	Duration of total process, in seconds
transcript_duration	INT	Duration of transcript module, in seconds
alignment_duration	INT	Duration of alignment module, in seconds
diarization_duration	INT	Duration of diarization module, in seconds
word_count	INT	Count of words in transcript
worker_hostname	NVARCHAR	Name of machine that executed the transcript
version	NVARCHAR	Application version used in transcription
org	NVARCHAR	Organizational unit associated with transcript

Assembly of the Republic (STAAR). Figure 7 illustrates the interaction and flow of information, providing a clear visualization of the structure and functioning of its various components.

The components involved in the transcription process are:

1. Audio Source: This is where all audios from various sources are initially obtained;
2. Audio Collector: A specialized module that interacts directly with the Audio Source, in order to filter and select relevant audios for transcription;
3. Audio Repository: A file server, with scalable storage capacity, which is dedicated to securely store audios for the selected transcriptions;
4. GPU Worker: A GPU-based infrastructure that efficiently uses deep learning models at the core of the Whisper ASR system. This advanced hardware accelerates transcription and ensures accurate and effective audio analysis, even in high-demand scenarios like parliamentary debates transcription. GPUs, specifically GeForce RTX 3060 with 12 GB GDDR6 (already available at AR), were used. These robust, high-performance GPUs proved efficient, even with Whisper's most demanding modules. Python code was developed to coordinate audio transcription across different workstations, optimizing resource use and avoiding duplications. These devices manage all the tasks of Audio Processing and Text Treatment stages, ensuring com-

Fig. 7 STAAR resources and information flow



pliance with requirement #15 by processing everything internally without relying on public clouds;

5. Transcription Database: An existing AR infrastructure database engine stores all relevant information collected during the process. It acts as a centralized registry, facilitating management and analysis of transcribed files. This database will be vital for future API development for STAAR integration with other systems;
6. Transcription Repository: After transcription, the resulting documents are stored on this file server. File sharing over the network with strict access control mechanisms ensures that only authorized users access the transcriptions (requirement #17);
7. Users: Users access debate transcriptions from the Transcription Repository over the network and work with existing AR productivity tools at their workstations, meeting requirement #5.

4.4 Implementation iterations

STAAR was developed according to the Design Science Research Methodology (DSRM, chosen for its iterative and design-focused approach, which aligns perfectly with the objectives of developing a solution tailored to the needs of the parliamentary environment. Throughout this process, illustrated in Fig. 8, three distinct iterations were executed to refine and optimize the final solution.

The first iteration of STAAR (Dec 2022–Mar 2023) focused on identifying needs and establishing objectives, leading to the development of a prototype to assess if AI models like Whisper met quality standards. Google Colab (2023) was used for proof of concept, due to its intuitiveness and cost-effectiveness, leveraging its advanced computational resources like GPUs. This phase, although manual and not user friendly, validated STAAR's central idea in using ASR Whisper and transcription engine and provided insights for subsequent iterations.

The second iteration (Apr–Aug 2023) aimed at optimizing and automating the transcription process. An automated system for audio collection was implemented, reducing manual intervention and potential human errors. Transcriptions, again done by Whisper, were optimized for precision and speed using AR's infrastructure, including dedicated GPUs. A custom lexicon was introduced to improve transcription accuracy and reduce manual corrections. The system was configured to provide transcriptions in the docx format for easy access, editing, and sharing.

The third iteration (Sep–Oct 2023) focused on refinement and innovation, introducing the advanced WhisperX model, which significantly reduced transcription time and computational resource usage. Timestamp alignment at the word level and speaker diarization were implemented, enhancing transcription readability and context comprehension. The lexicon was revised and expanded to address identified gaps.

5 Results

This section delves into the effectiveness and impact of the STAAR automatic transcription system through a detailed demonstration and evaluation process. It outlines the steps taken to validate the performance of the system in real-world conditions, involving key stakeholders and rigorous testing phases. This section highlights the progressive iterations of the system, its practical application in the parliamentary context, and the metrics used to assess its performance and utility. By analyzing both quantitative data and user feedback,

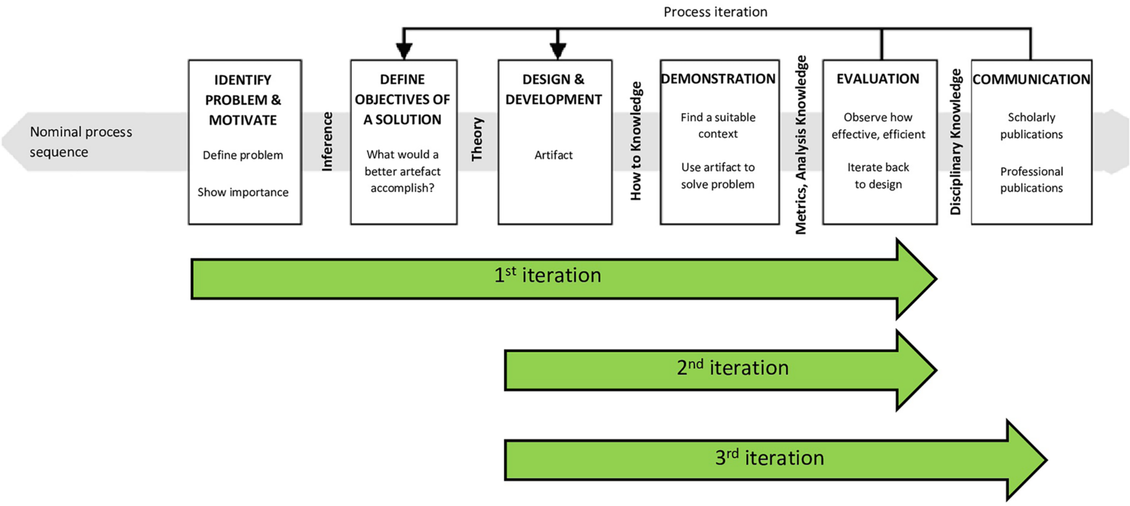


Fig. 8 Iterations represented in the DSRM process

this section provides a comprehensive overview of STAAR's success and areas for future improvement.

5.1 Demonstration

The Demonstration phase, a key step in the Design Science Research Methodology (DSRM) (Peffers et al., 2007), tests the artifact (in this case, the STAAR automatic transcription system) in real conditions to validate its effectiveness and utility. Planned to involve key stakeholders, this phase ensures the developed solution meets the needs identified in the initial DSRM stage.

In late 2021, a dedicated workgroup, the Automatic Transcription Working Group (GT-TA), was formally nominated to evaluate a commercial transcription solution that was presented to AR. It comprised a variety of professionals with diverse skills and responsibilities, all sharing an interest in automatic transcription:

- Seven members from the Records Division (DR), the team responsible for transcribing debates and producing the Assembly of the Republic's Journal (DAR);
- One member from the Committee Support Division (DAC), primarily responsible for creating meeting summaries and occasionally transcribing relevant interventions;
- One member from the Parliament TV Channel, responsible for recording and providing plenary and committee meeting audios, playing a crucial role in addressing technical aspects related to this matter;
- One member from the Division of Technological Infrastructures (DIT), the author of this work, responsible for gathering necessary information and promoting the development of a transcription solution;
- Heads of the departments interested in automatic transcription.

After unsatisfactory results with the suggested commercial transcription product, mainly due to its inadequate interface, excessive transcription time, and a high Word Error Rate (WER) of 39%, the GT-TA team decided to explore other alternatives. Between January 2022 and February 2023, the AR transcription team evaluated five other products, evaluating them based on WER, which varied significantly between 23.1 and 44.4%.

Using this work first results on the related work, the team decided to conduct initial demonstrations using the AI model Whisper and proceed with its development if the initial tests were promising in terms of WER and ease of use. During February 2023, the GT-TA team used Google Colab for transcription, taking advantage of its accessibility and processing capacity. This period served as an evaluation phase, allowing the team to evaluate transcriptions by Whisper in various scenarios and audio types. These tests were performed ah-hoc with no supervision, therefore there is no data about the number of audios evaluated. The goal was only to evaluate if the transcription was good enough to continue working with ASR Whisper, which was the case.

In April 2023, the STAAR's second iteration version was released, automating transcriptions as soon as audios became available using AR's internal resources, producing docx documents. From April to July 2023, the entire Records Division transcription team used and evaluated transcriptions generated by STAAR, processing 1915 files, equivalent to approximately five hundred hours of audio.

Following the second iteration's evaluation, new developments were implemented to address identified gaps.

In September 2023, an updated version of STAAR, corresponding to the third and final DSRM iteration, was released and evaluated by the entire Records Division transcription team. In this last iteration, from September to the end of October 2023, 486 files were transcribed, representing about 225 h of audio.

Table 6 provides key activities conducted during each iteration, highlighting the evolution from initial development and prototype testing to final refinement and innovation stages.

5.2 Evaluation

The evaluation of STAAR system allows for the measurement of the system's effectiveness and accuracy in the context of parliamentary debates. This evaluation includes technical transcription metrics, error rates (WER), as well as a more holistic assessment through the adopted methodology (DSRM).

5.2.1 Transcription metrics

To evaluate the effectiveness and precision of the Automatic Transcription System of the Assembly of the Republic (STAAR), it is necessary to quantify metrics resulting from the transcriptions performed, in terms of the volume of plenary meetings and committees processed by STAAR, as well as the associated audio hours and transcribed words.

As previously noted, the first DSRM iteration lacks statistical data, as it was limited and not actually conducted by

Table 6 Evaluation iterations

Iteration	Timeframe	Activity
1st	Dec 2022–Mar 2023	Needs identification and prototype development
2nd	Apr–Aug 2023	Optimization and automation
3rd	Sep–Oct 2023	Refinement and innovation

STAAR, but through an external platform using the same AI model, Whisper.

As assessed from Table 7, from the beginning of April to the end of October 2023, STAAR automatically transcribed over 30 days of continuous audio (724 h), which refers to 335 plenary meetings, committees, and events.

As detailed previously, in the first two iterations of STAAR, the system only transcribed audio to text. In the third iteration, it also began aligning text timestamps and performing diarization. In this last iteration, the transcription process was conducted by WhisperX, which uses the Whisper model provided by OpenAI, but with a significant reduction in GPU resource consumption and the time required for transcription.

As shown in Fig. 9, the transcription time for a 15-min fragment of a plenary meeting was reduced from an average of 3 min and 45 s (225 s) to just 33 s, which represents an efficiency increase of about 86%.

Despite the overall duration of the transcription process increasing from the second iteration (225 s) to the third (342 s), this rise in time is attributed to the additional tasks of alignment and diarization performed in the third iteration. While the alignment process is quite rapid (17 s), the diarization process is more time-consuming (292 s), accounting for approximately 85% of the total audio processing time. Although the total transcription time has increased, it remains significantly below the values considered reasonable (requirement #4). The added value provided by diarization to the final text, particularly in the context of parliamentary debates, justifies this increase in processing time.

5.2.2 Transcription evaluation

This section provides a detailed analysis of transcriptions to assess the precision of STAAR. The aim is to examine a set of transcriptions reflecting various circumstances and factors that typically influence transcription quality. The analysis includes complete audio fragments with multiple speakers as well as individual speeches.

Special focus was given to different speech styles, including both pre-written and read or impromptu speeches, and various frequently transcribed parliamentary contexts such as plenary sessions and committee meetings. The debate environment, ranging from calm discussions without interruptions to more intense debates with background noise and multiple interruptions, was also considered.

Furthermore, to assess the effect of interruptions, speaker changes, background noise, and side comments on transcription quality, fragments from various meetings and debates were analyzed, including heated debates with several interruptions and background noise.

Table 7 Transcription metrics

Iteration	Engine	Timeframe	Proceedings	Meetings	Hours	Words
2nd	Whisper	Apr–Aug 2023	Plenary	50	180	1,495,826
			Committee	122	318	2,537,262
			Events	2	1	11,000
3rd	WhisperX	Sep–Oct 2023	Plenary	22	79	675,894
			Committee	139	146	1,053,725
Total				335	724	5,773,707

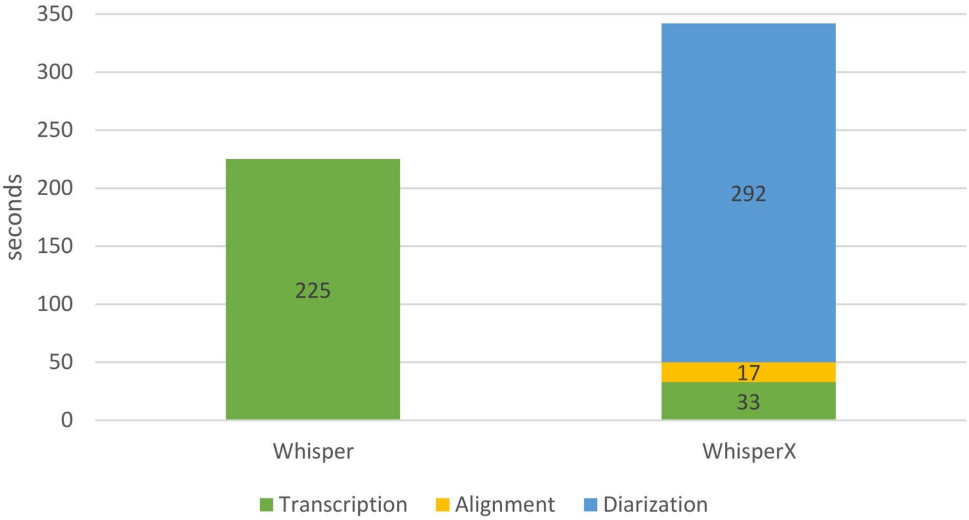


Fig. 9 Comparison between Whisper and WhisperX processing time for a 15-min audio

Three read speeches/interventions in plenary and one testimony read at an Inquiry Committee, five impromptu interventions in plenary, and two short impromptu interventions in a parliamentary committee were analyzed to evaluate the influence of each speaker/speech on the quality of the automatic transcriptions.

The goal was to understand how individual speaker characteristics, such as pronunciation, rhythm, intonation, and speech clarity, affect transcription quality. The analysis included speakers from different regions of Portugal and even a case with a possible speech disorder.

In total, eleven transcriptions were analyzed, amounting to just over two hours. Despite their relatively short duration, these transcriptions are considered to provide a representative view of the daily challenges faced by the team transcribing parliamentary debates. A summary of this analysis can be found in Table 8.

5.2.3 Word error rate analysis

In the context of the analyzed transcriptions, although their number and duration are limited (suggesting the need for more extensive analyses and cautious interpretation of results), it is evident that STAAR's Word Error Rate (WER) is, on average, significantly lower than initially anticipated and below AR's acceptable minimum (15%). In ideal situations, such as read interventions in plenary sessions, the WER can be as low as 4.3%, a value lower than even what the creator of the Whisper model used by STAAR, indicated for Portuguese transcriptions, as illustrated in Fig. 10.

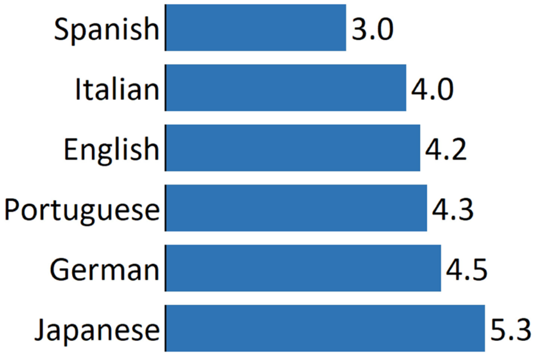


Fig. 10 Whisper's "large-v2" WER for several languages, in particular Portuguese (Radford et al., 2022)

Careful analysis of the various transcriptions show that diverse factors influence the transcription quality, including different speech styles (pre-written/read and impromptu) and parliamentary contexts (plenary sessions, committee meetings). The study also considers the debate environment, varying from calm to intense with background noise and interruptions. The major key findings are:

- Prepared/read speeches generally have higher accuracy than impromptu speeches;
- Plenary interventions are transcribed more precisely than committee meetings;
- The lowest WER observed was 1.7% in a read intervention from the tribune (T3), followed by 2% in a speech by the President of the AR in a solemn session (T1);
- The highest WER (11.3%) was in a question-and-answer session in an inquiry committee (T9);
- The read testimony in T7 had the lowest WER (6.1%) among all committee transcriptions;
- The nature of the intervention (read or impromptu) significantly impacts transcription quality;

Table 8 Transcript analysis with calculated word error rate (WER)

#	Proceedings	Type of speech	Pronunciation	Type of debate	WER (%)	
					Manual	Without disfluencies
T1	Plenary	Read / Written	North	No noise or interruptions	2.0	
T2	Plenary	Impromptu	Center	Noisy and with interruptions	9.7	4.5
T3	Plenary	Read / Written	Center	No noise or interruptions	1.7	
T4	Plenary	Impromptu	Center	Noisy and with interruptions	8.2	
T5	Plenary	Read / Written	Madeira	No noise or interruptions	8.1	
T6	Plenary	Impromptu	Center / North	Noisy and with interruptions	5.8	1.8
T7	Committee	Written statement	Dyslalia	Noisy	6.1	5.4
T8	Committee	Written statement	Dyslalia	Noisy and with interruptions	7.3	
T9	Committee	Question/Answer	Dyslalia	Noisy and with interruptions	11.3	
T10	Committee	Impromptu	Center	Noisy	5.2	
T11	Committee	Impromptu	Center	Noisy	4.0	1.5

- Environmental and procedural factors also play a role; speeches from the podium face fewer interruptions and less background noise, benefiting transcription accuracy;
- Individual speech characteristics (pronunciation, diction, prosody) have a lesser impact on transcription quality compared to other factors;
- Interruptions, speaker changes, background noise, and side comments in debates showed limited errors, indicating a need to fine-tune technical parameters for speech detection (VAD);
- Whisper's omissions of repetitions resemble to what a human editor would do, which facilitate transcription work by avoiding unnecessary corrections;

Based on the collected data it is possible to determine the factors that influence automatic transcription quality in parliamentary debates, as shown on Table 9.

Despite the limited size of the sample and the brief period for analysis, the diversity of speeches reviewed allows for a confident conclusion that STAAR's efficiency and accuracy rate surpass not only all previously tested software and solutions but also the initial predictions of the GT-TA regarding the capabilities of an automatic transcription system. In the best possible conditions, the calculated WER ranged between 1.7 and 2%, while in the worst conditions evaluated, the WER did not exceed 11.3%, staying below the 15% initially set as the maximum acceptable limit (requirement #14).

5.2.4 DSRM evaluation

This section outlines the Evaluation stage of the DSRM (Peffers et al., 2007), which aims to obtain a clear understanding of the system's performance in a real-world context and identify areas of success and potential improvements, and with it align the development with the work objectives.

Following the demonstration period of the automatic transcription prototype using Whisper on Google Cloud (first iteration), users recognized the system's utility but identified various areas needing improvement:

- Complexity of the process: Although the transcription was automatic, it was perceived as complex and time-consuming, potentially compromising the desired efficiency;
- Time required for transcription: In a few occurrences, the cloud system took up to ten minutes to transcribe a fifteen-minute audio and frequently failed, requiring the process to be repeated;
- Difficulty in file handling: The need to individually upload each audio and then download the corresponding text document made the process very impractical, especially when dealing with a large volume of audios;
- Inaccuracies in transcription: In its initial form, the system was not adapted to the specific terminology and jargon of the parliamentary context, leading to transcription errors that required additional effort from the reviewers;
- Writing Rules: Various writing and formatting rules were not applied by the transcription, necessitating manual review to ensure compliance with established standards.

The evaluation of the outcome of the second iteration was conducted by the same team, GT-TA, who found that many of the previously identified issues had been resolved. However, to fully meet the identified requirements (Table 3), further improvements were necessary to address the following issues:

- Temporal imprecision: The timestamps associated with the transcribed phrases were not accurate, potentially compromising the fidelity of the transcription in relation to the original audio;
- Text format: The lack of line breaks in the transcribed text made it difficult to read and navigate the document, making the review and editing process more challenging;
- Absence of transcription at the beginning/end of some interventions: It was observed that sometimes the system failed to transcribe segments close to periods of silence in the audio, often corresponding to speaker change.

The third iteration marked the final phase of development and optimization of the system. Building on continuous feedback from the GT-TA and with the lessons learned from previous iterations, the development focused on

Table 9 Factors that influence STAAR's automatic transcription and its impact on WER

Factor	Description	Impact on WER
Nature of the intervention	Whether the intervention is pre-written/read or impromptu	High
Debate environment	Level of noise and number of interruptions during speech	High
Change of Speaker	Frequency of changing speakers or interjections	Medium
Omission of disfluencies	Whether the system deliberately omits disfluencies	Medium
Speech characteristics	Pronunciation, diction, rhythm, etc	Low
Regimental aspects	Nature of the intervention, like requests for clarification	Low

enhancing existing functionalities and introducing new features to better meet user needs. This phase saw significant improvements in transcription accuracy, the precision of timestamps per word, identification of text related to each speaker, text formatting, and fine-tuning of speech detection (VAD) parameters. Additionally, new functionalities were implemented, such as the ability to customize substitution dictionaries, and STAAR began transcribing several committee meetings that were previously outside its scope of coverage.

The continuous involvement of the GT-TA throughout all stages of the DSRM (Peffers et al., 2007) ensured that the developed system was aligned with the expectations and requirements of the end-users. Table 10 presents the evaluation conducted by the GT-TA, based on the requirements (Table 3) that were set, categorized by criteria such as efficiency, usability, quality, and security, among others. This evaluation was conducted after using each updated version of the system resulting from the development iterations, which allowed them to assess the results produced by STAAR.

According to the proposed methodology, the defined evaluation scale is:

- Not Achieved (N);
- Partially Achieved (P);
- Largely Achieved (L);
- Fully Achieved (F).

5.3 User feedback

The success of any technological system depends not only on its functionality or technical efficiency but also on its acceptance and perceived utility by end-users. Therefore, user feedback is fundamental for evaluating a solution's effectiveness, identifying areas for improvement, and ensuring the system meets needs and expectations.

For that purpose, a survey was conducted to gain a holistic view of user interaction with the system and its impact on the parliamentary workflow, particularly regarding parliamentary debate transcription. The survey was chosen for its ability to systematically reach a large number of users, allowing for the collection of both quantitative and qualitative data.

The first section of the survey focuses on participants characterization, understanding the demographic and professional profile of users through aspects like age, department affiliation, and transcription experience. This data is crucial for contextualizing responses and identifying specific patterns or trends among user groups. The second section centers on direct experiences with STAAR, inviting users to share their opinions on usability, efficiency, accuracy, and other relevant system aspects, in order to understand user satisfaction, identify strengths, and identify improvement areas.

A 5-point Likert scale (Likert et al., 1934) was used for the responses to the survey, with verbal descriptions for extremes, such as "Yes, very much" and "Not at all." This scale allowed participants to indicate their level of agreement

Table 10 Evaluation of requirements per DSRM iteration

Criteria	#	Requirement	1st	2nd	3rd
Efficiency	1	Automatically transcribe speech to text	P	L	F
	2	Produce legible and comprehensible transcripts	L	L	F
	4	Process audios quickly once available	P	F	F
Usability and compatibility	3	Allow editing and formatting of text directly in Word	N	F	F
	5	Be compatible with existing equipment	F	F	F
	6	Allow use of hardware for audio control per transcription	N	F	F
	7	Identify speaker change and associate transcribed text	N	N	F
Flexibility	8	Obtain audios from various locations	N	F	F
	9	Transcribe manually introduced audios	F	F	F
	12	Be expandable to other parliamentary contexts	N	L	F
Linguistic specificity	10	Consider the particularities of the Portuguese language	F	F	F
	13	Allow the creation and customization of dictionaries	N	L	F
Quality	11	Allow evaluations and tests	F	F	F
	14	Have a WER below 15%	F	F	F
	18	Handle background noise	L	L	F
Security and privacy	15	Process audios within the AR infrastructure	N	F	F
	16	Allow remote work capability	F	F	F
	17	Provide distinct levels of access permissions	N	F	F

or disagreement with the questions on an ordinal scale of 1 to 5, facilitating data analysis and result comparison.

Google Forms (Google Forms & | Google Workspace, 2023) was used for collecting user feedback on STAAR, chosen for its widespread accessibility, flexibility in question creation, integrated analysis capabilities, and confidentiality assurance. An email with the survey link was sent on November 1st, 2023, explaining the survey's purpose, and ensuring that responses were anonymous.

A week after the email, twenty-two responses were received, nineteen from the Records Division (DR) and three from the Committee Support Division (DAC. Notably, all AR employees dedicated to parliamentary debate transcription responded, providing a clear and comprehensive view of user perception and experience using STAAR.

Age-wise, participants varied significantly, from 30 to over 60 years, reflecting different life stages. In terms of professional experience, there was an even distribution between employees with less than nine years and those with over ten years of experience, highlighting AR's ongoing efforts in staff renewal and the presence of highly experienced professionals in parliamentary debate transcription. Figure 11 shows the distribution of participants by age and current professional experience related to transcription.

Regarding STAAR usage frequency, the results show that the system is a regular part of participants routines, with 45.5% using it daily and 40.9% on a weekly basis.

Statistical information, like the mean, median, minimum, maximum values, and standard deviation of responses, were structured to provide statistical insights. Table 11 shows the descriptive statistics of the survey responses.

Overall, lower standard deviations in categories like ease of use, transcription needs, and positive impact suggest a consistently favorable perception of STAAR in these aspects. On the other hand, greater variation in topics like text formatting and transcription speed responses indicate areas where user experiences vary more and may need improvements or adjustments.

The global analysis of responses indicates an incredibly positive reception of STAAR. All participants rated the ease of use of transcriptions as "Easy" (40.9%) or "Very easy" (59.1%), and the speed of transcription availability was predominantly evaluated as "Very fast" (54.5%). In terms of accuracy, the transcriptions were mostly considered "High" (54.5%) or "Acceptable" (40.9%), suggesting satisfactory system performance with room for improvement.

The text separation by speaker functionality, possible through diarization, was well-received, with the vast

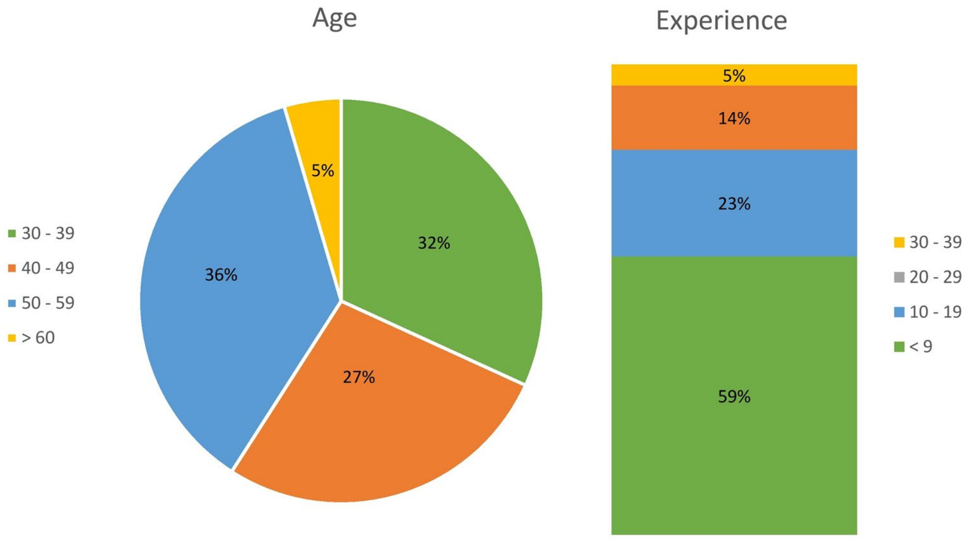


Fig. 11 Participants distribution by age and experience

Table 11 Statistics of survey responses

Topic	Mean	Median	Minimum	Maximum	Deviation
Ease of use	4.6	5	4	5	0.5
Speed of transcription	4.4	5	3	5	0.7
Accuracy	3.6	4	3	5	0.6
Text separation by speaker	4.5	5	3	5	0.7
Text formatting	4.3	4	2	5	0.8
Responds to transcription needs	4.3	4	4	5	0.5
Positive impact on job	4.6	5	4	5	0.5

majority finding it useful (90.9%). Text formatting and automatic correction identification were seen as facilitators (86.4%), improving the review, and editing experience of automatic transcriptions.

In terms of meeting transcription needs, all participants felt the system fully (31.8%) or partially (68.2%) met their needs. STAAR's implementation was seen as having a positive impact on all participant's roles, with the majority expressing "Very" positive (59.1%), indicating significant improvement over previous transcription methods.

Based on the responses from the survey, Fig. 12 presents a radar chart that shows a comparative perspective on various dimensions evaluated by STAAR users.

The participants feedback reflects a high overall satisfaction with STAAR, highlighting its significance in transforming and optimizing the transcription process in the parliamentary context. With its speed, accuracy, and innovative features, STAAR not only meets the needs of its users but also establishes itself as a valuable tool in the modernization and efficiency of parliamentary documentation.

5.4 Limitations

Throughout the development and implementation of the Automatic Transcription System for the Assembly of the Republic (STAAR), various challenges were encountered that required an adaptive approach and innovative solutions suitable for the parliamentary context.

Technically, integrating and optimizing the innovative Whisper model for parliamentary use presented significant challenges due to limited documentation and case studies available at the time of its development. Overcoming these challenges involved extensive IT specialist involvement and continuous testing.

Operationally, transitioning from a manual to an automatic system was challenging, requiring new work methods. Initially, the Automatic Transcription Group (GT-TA) often requested the replication of existing work methods rather than adopting innovative approaches. User training and technical support were crucial in overcoming resistance to change and ensuring system acceptance and adoption.

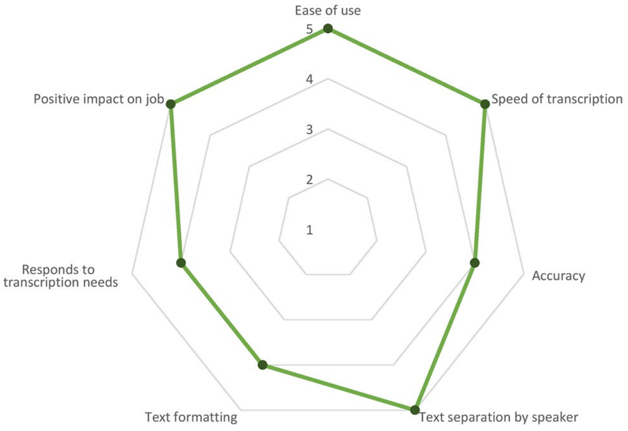


Fig. 12 Radar chart with survey response medians

Contextually, the parliamentary environment's unique jargon and the rigid format of the Assembly of the Republic journal (DAR) posed distinct challenges. The system needed customization to recognize and accurately transcribe specific parliamentary language, and adjustments were made to align the transcribed text with DAR standards and formats. Collaboration with the Assembly's transcription team was key to adapting the system to the specific needs of the parliamentary environment.

Despite these challenges, the right combination of technical expertise and a deep understanding of the parliamentary context by GT-TA enabled the successful creation of a robust and effective solution for transcribing parliamentary debates.

6 Conclusions and future work

The development of the Automatic Transcription System for the Assembly of the Republic (STAAR) represents a significant advancement in the transcription of parliamentary debates, closely aligning with the objectives set at the beginning of this work.

The first goal, focused on identifying and analyzing automatic speech recognition technologies and their applicability in transcribing parliamentary debates, was achieved through an investigation of the state of the art in automatic transcription, not only through literature review but also based on the experience of other Parliaments. This detailed analysis led to the selection of the Whisper model, which stood out for its ease of use and suitability for the needs of the parliamentary environment.

By implementing advanced Artificial Intelligence technologies like the Whisper model, it was possible to automate and optimize a process that traditionally relied on intensive manual efforts and involved significant physical strain. The modular architecture of STAAR, that comprehends stages of audio collection, audio processing, text treatment, and storage, proved effective for the specific needs of the parliamentary environment in terms of audio-to-text transcription. The ability to adapt to the particularities of parliamentary language and specific jargon, as well as integration with existing systems, highlighted the system's flexibility and robustness, fully achieving the second goal of this work.

The third and final objective, implementing speaker change recognition technologies to improve the transcription of parliamentary debates, was successfully achieved through diarization methods in STAAR. This feature enriched the transcriptions and facilitated reading and review through clear visual segmentation of the text by speaker, improving the accuracy and utility of the automatic transcriptions.

The overall results obtained with STAAR exceeded expectations not only for being extremely fast in producing transcriptions but also because the error rate (WER) of these was much lower than what was initially considered acceptable. The incorporation of advanced features, such as automatic text correction via specialized dictionaries, identification of speaker alternation, and appropriate text formatting, resulted in documents that closely resemble those previously done manually in the context of the AR. For the transcription team, this system represented a profound transformation in how their work is conducted, completely eliminating the need for manual transcription of parliamentary debates, although the task of revision became more demanding due to the need to identify and correct potential transcription errors.

With this, the implementation of STAAR not only improved the efficiency of transcriptions and reduced the time needed to produce the first internal version of the Diary of the AR but also provided significant resource savings, allowing transcription professionals to focus on more complex and less routine tasks. Additionally, due to the positive results achieved by STAAR, there was a rapid request for the system to be used in transcribing parliamentary committee meetings, a task previously not performed due to human resource limitations, thus expanding the scope and detail in documenting parliamentary activities.

In conclusion, STAAR not only addressed the identified problems and requirements, fulfilling all the objectives of this work, but also added significant value to the AR positioning it at the forefront of technological innovation in the context of transcribing parliamentary debates.

6.1 Future work

Despite the significant advancements achieved with STAAR in the transcription of parliamentary debates, the dynamic nature of technology and the evolving needs of the Assembly of the Republic suggest there is always room for improvement and expansion. Reflecting on the work done and anticipating future needs, several opportunities stand out, aiming to optimize STAAR and expand its applicability and ongoing relevance:

- Conduct further testing in future research to include long-term usage data and a detailed analysis of different speech styles and environmental conditions;
- Optimization of the ASR model: With the continuous advancement of AI technology, there is room to further refine the model used, making it more precise and adapted to the nuances of parliamentary debates;
- Creation of a speaker corpus: This innovation would enable STAAR to accurately identify the author of each parliamentary intervention. By automatically recognizing and assigning the name and party of the speaker to the transcription, it would eliminate a manual task currently performed by the transcription team, further optimizing the process, and providing other potential applications;
- Reduction of overall transcription duration: Given that the diarization stage represents about 85% of the time needed for the transcription of audio, optimizing this would significantly reduce the overall duration of the process, thus increasing the system's efficiency and speeding up the availability of transcriptions for parliamentary use;
- Probabilistic contextual replacements: A significant evolution for STAAR would be the introduction of a text replacement mechanism based on approximation, going beyond simple word or expression matching, and incorporating a contextual analysis of the transcription text. Through probabilistic approaches, the system could identify, and correct errors based on matching thresholds, ensuring more precise corrections adapted to the context in which words or expressions appear;
- Adoption of a structured transcription format: Transitioning from a traditional format like docx to a more structured and interoperable format would enable segmented queries by speaker, theme, or any other relevant criteria, potentially enabling more agile and personalized handling of transcriptions. This change would benefit not only Members of Parliament but also make parliamentary debates more accessible and understandable to the general public;
- Creation of automatic summaries: Implementing advanced automatic summarization algorithms could transform extensive textual records into clear and relevant syntheses, enhancing the efficiency of work, for example, in Parliamentary Committees;
- Integration with other platforms: STAAR's integration with other digital platforms would broaden and simplify access to parliamentary debates.

Funding Open access funding provided by FCT|FCCN (b-on).

References

14:00-17:00 ISO/IEC 15504-2. (2003). Retrieved 23 Oct 2023, from https://www.iso.org/standard/37458.html

Alumaë, T., Tilk, O., & Ullah, A. (2018). Advanced rich transcription system for Estonian speech. *Frontiers in Artificial Intelligence and Applications, 307*, 8.

Baevski, A., Zhou, H., Mohamed, A., Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *34th Conference on neural information processing systems (NeurIPS 2020)*, (Vol. 2020), Vancouver, Canada.

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). WhisperX: Time-accurate speech transcription of long-form audio. https://doi.org/10.48550/arXiv.2303.00747

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., & Gill, M.-P. (2019). Pyannote.Audio: Neural building blocks for speaker diarization.

Campr, P., Kunešová, M., Vaněk, J., Čech, J., Psutka, J. (2014). Audio-video speaker diarization for unsupervised speaker and face model creation. In *Proceedings of the lecturer notes on computer sciences* (Vol. 8655 LNAI, pp. 465–472). Springer.

ChatGPT. Retrieved 5 Nov 2023, from https://chat.openai.com

DALL·E 3. Retrieved 5 Nov 2023, from https://openai.com/dall-e-3

de Lima, T. A., & Da Costa-Abreu, M. (2020). A survey on automatic speech recognition systems for Portuguese language and its variations. *Computer Speech and Language, 62*, 101055. https://doi.org/10.1016/j.csl.2019.101055

de Vos, H., & Verberne, S. (2023). *Political corpus creation through automatic speech recognition on EU debates*. https://doi.org/10.48550/arXiv.2304.08137

De Wet, F., Badenhorst, J., & Modipa, T. (2016). Developing speech resources from parliamentary data for South African English. *Procedia Computer Science, 81*, 45–52.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. https://doi.org/10.48550/arXiv.1810.04805

Diáz-Munió, C. V. G., Silvestre-Cerdà, J. A., Jorge, J., Giménez, A., Iranzo-Sánchez, J., Baquero-Arnal, P., Roselló, N., Pérez-González-de-Martos, A., Civera, J., Sanchis, A. et al. (2021). Europarl-ASR: A large corpus of parliamentary debates for streaming ASR benchmarking and speech data filtering/verbatimization. In *Proceedings of the annual conference on international speech communication association* (Vol. 6, pp. 4371–4375). International Speech Communication Association (Interspeech)

El Emam, K. (1998, Nov). The internal consistency of the ISO/IEC 15504 software process capability scale. In *Proceedings of the proceedings fifth international software metrics symposium*. Metrics (Cat. No.98TB100262) (pp. 72–81).

Google Forms | Google Workspace. Retrieved 13 June 2023, from https://www.google.com/forms/about/

Google Colaboratory. Retrieved 19 Oct 2023, from https://colab.research.google.com/

HTK Speech Recognition Toolkit. Retrieved 29 Oct 2023, from https://htk.eng.cam.ac.uk/

Introducing Whisper. Retrieved 25 Sept 2023, from https://openai.com/research/whisper

Kaldi: Kaldi. Retrieved 29 Oct 2023, from https://kaldi-asr.org/doc/index.html

Kawahara, T. (2018, Feb) Automatic meeting transcription system for the Japanese Parliament (diet). In *Proceedings of the Asia-Pacific signal on information processing association annual summit conference (APSIPA ASC),* (Vol. 2018, pp. 1006–1010). Institute of Electrical and Electronics Engineers Inc.

LIBE | Committees | European Parliament. Retrieved 5 Dec 2023, from https://www.europarl.europa.eu/committees/en/libe/home/highlights

Likert, R., Roslow, S., & Murphy, G. (1934). A simple and reliable method of scoring the thurstone attitude scales. *Journal of Social Psychology, 5*, 228–238. https://doi.org/10.1080/00224545.1934.9919450

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. https://doi.org/10.48550/arXiv.1907.11692

LIUM SpkDiarization. Retrieved 29 Oct 2023, from https://projets-lium.univ-lemans.fr/spkdiarization/

Ma, R., Gales, M. J. F., Knill, K. M., & Qian, M. (2023). N-Best T5: Robust ASR error correction using multiple input hypotheses and constrained decoding space. *INTERSPEECH, 2023*, 3267–3271. https://doi.org/10.21437/Interspeech.2023-1616

Mansikkaniemi, A., Smit, P., & Kurimo, M. (2017, Aug). Automatic construction of the Finnish Parliament speech corpus. In Lacerda F., Strombergsson S., Wlodarczak M., Heldner M., Gustafson J., & House D. (Eds.), *Proceedings of the annual conference on international speech communication association* (Vol. 2017, pp. 3762–3766). International Speech Communication Association (Interspeech).

Multilingualism in the European Parliament. Retrieved 5 Dec 2023, from https://www.europarl.europa.eu/about-parliament/en/organisation-and-rules/multilingualism

OpenAI GPT-4 technical report. (2023). https://doi.org/10.48550/arXiv.2303.08774

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*, 45–77. https://doi.org/10.2753/MIS0742-1222240302

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision*. https://doi.org/10.48550/ARXIV.2212.04356

Scopus—Document search. Retrieved 21 May 2023, from https://www.scopus.com/search/form.uri?display=basic#basic

The Official Home of the Python Programming Language. Website Python.Org. Retrieved 25 Sept 2023, from https://www.python.org/

Web of Science Core Collection—Document search. Retrieved 21 May 2023, from https://www.webofscience.com/wos/woscc/basic-search

WER | calculate the word error rate with our tool. Retrieved 21 Oct 2023, from https://www.amberscript.com/en/wer-tool/

Whisper/Approach.Png at Main Openai/Whisper. Retrieved 27 Sept 2023, from https://github.com/openai/whisper/blob/main/approach.png

IEEE Xplore—Document search. Retrieved 21 May 2023, from https://ieeexplore.ieee.org/Xplore/home.jsp

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). *A survey of large language models*. https://doi.org/10.48550/arXiv.2303.18223

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.