Software for Automatic Speech Recognition via Whisper models applied to Oral History interviews in the Portuguese language

Edgleide de Oliveira Clemente da Silva[1], Fernando Rezende Zagatti[1,2,3,4], Filipe Loyola Lopes[1,2,3], Anderson Dias Duarte[1], Rodrigo Bonacin[1,2,3], Angela Maria Alves[1]

[1]Center for Information Technology Renato Archer, Campinas, SP, Brazil
[2]Centro Universitário Max Planck (UniMAX), Indaiatuba, SP, Brasil
[3]Centro Universitário de Jaguariúna (UniFAJ), Jaguariúna, SP, Brasil
[4]Federal University of São Carlos, São Carlos, SP, Brazil

Correspondence: fzagatti@cti.gov.br

Abstract

This paper presents Ethos AT, a desktop software for automatic transcription that uses OpenAI Whisper models, enabling local processing and ensuring data privacy and accessibility for users who are not necessarily programming experts, such as oral history researchers. A comparative analysis of six Whisper models (*small*, *medium*, *large*, *large-v2*, *large-v3*, and *turbo*) was conducted to analyze performance in terms of transcription accuracy, error types, and processing time. Results indicate that larger models achieve higher lexical accuracy, while smaller ones provide faster execution with acceptable quality for general use; the *turbo* model showed an effective balance between accuracy and speed. Overall, Ethos AT offers a secure, efficient, and user-friendly solution for academic and institutional contexts.

1 Introduction

In recent years, the incorporation of Artificial Intelligence (AI) technologies into the field of Oral History has transformed the way that interviews are processed (Gref, 2022; Stoykova et al., 2024). Among these innovations, automatic transcription approaches stands out for its ability to convert speech into text quickly and with relative accuracy, contributing to the optimization of the work of researchers and institutions dealing with large volumes of audio records.

According to the literature, transcription involves transforming orality into written language; however, there is a risk that automated transcription may erase the nuances of voice, such as hesitations, emphases, and emotions, which are common in spoken language (Severino, 2016). In this way, automatic technologies should be used as support tools, with the user being responsible for ensuring the integrity and coherence between what was said and what will be recorded (Shopes, 2002).

The automation of this process represents an advance compared to manual transcription, traditionally recognized as a costly stage due to the time and mental fatigue spent by human agents (Haberl et al., 2024). Tools available on the market (such as Clipto[1], Descript[2], and Otter.ai[3]) allow hours of recordings to be transcribed quickly, while also offering complementary features such as speaker diarization (Kanda et al., 2022), timestamping (Shi et al., 2022), and text export in different formats. These features enhance data analysis and reuse, making the material more accessible and integrable into digital repositories and databases.

Beyond operational efficiency, automatic transcription also contributes to the democratization of access to oral content. By providing textual versions of interviews or other types of audio, researchers can make collections more inclusive, benefiting people with hearing impairments or communities that profit from automatic translations (Pragt et al., 2022; Fatehifar et al., 2025). Thus, technology is a way to expand oral memory and strengthen its circulation in the digital environment.

However, the use of these tools still faces relevant technical and ethical challenges. The accuracy of transcriptions depends on audio quality, recording environment, and particularities of speech (such as accents, pauses, overlaps, and regional expressions). In addition, the models may generate semantic errors, misinterpretations, and omissions of important nuances in speech. Finally, there are concerns about data privacy and confidentiality, since many services perform processing in the cloud, out-

[1]https://www.clipto.com/
[2]https://www.descript.com/transcription
[3]https://otter.ai/

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 302–310

side the researcher's direct control.

The role of the researcher is not replaced by technology, but rather redefined. Automatic transcription should be understood as an initial stage, to be complemented by human review and curation. Thus, we developed a transcription software named **Ethos AT - Audio Transcriber**[4] designed to facilitate access to these technologies for people who are not programming experts, or even experts who want something quick and easy to access. This technology aims to enable the processing of long audios while ensuring privacy in its application.

The structure of this paper is organized as follows: Section II presents the main related works, highlighting approaches and tools in the context of automatic transcription. Section III describes the developed software, detailing its architecture, functionalities, and distinguishing features compared to existing solutions. Then, Section IV presents the evaluation process. Section V discusses aspects related to data security and privacy handling, emphasizing the importance of local processing and ethical data management. The corresponding results and analyses are discussed in Section VI. Finally, Section VII presents the conclusions and proposes directions for future work.

2 Related work

Automatic Speech Recognition (ASR), also known as Speech-to-Text (STT), has two major lines of study that are related to this present paper: (i) the technological and methodological advances in ASR for Portuguese, and (ii) the application of these technologies in contexts of interviews, oral history, or sound archives.

Gris et al. (2022) presented the NURC/SP corpus, compiled by the Urban Educated Linguistic Norm (NURC) Project, consisting of 375 inquiries taken in São Paulo (approximately 334 hours of audio), in which 328 inquiries had no transcription and 47 had no alignment between audio and transcription. Evaluating four open-source ASR models, the study found that the CORAA ASR model (Candido Junior et al., 2023) achieved the best average performance and selected to automatically transcribe the remaining 284 hours of audio.

For other approaches, Gris et al. (2023) presents an evaluation of the OpenAI Whisper ASR model for the Portuguese language, focusing on speech audios and videos extracted from the life stories of the Museu da Pessoa (MuPe). The main objective was to assess Whisper's ability to generate transcriptions with punctuation and capitalization, a functionality absent in the output of most ASR systems. The results indicate that Whisper achieves good performance, although they conclude that certain punctuation still requires improvement.

Medeiros et al. (2023) propose a transfer learning approach based on a model previously optimized for the English language, using the NVIDIA NeMo framework, to adapt ASR to European Portuguese (EP). Since EP has fewer resources available in the literature, this work investigates domain adaptation using the SpeechDat dataset (only EP data) and the Multilingual LibriSpeech (mixed languages, portuguese are mostly Brazilian).

Nascimento et al. (2024) introduced the Automatic Transcription System for the Assembly of the Republic of Portugal (STAAR), a tool that uses Whisper ASR models and was developed to identify nuances of the Portuguese language and parliamentary procedures. The system demonstrated a low transcription error rate, ranging between 1.7% and 11.3% depending on the context and speech style. STAAR reduced the time required to produce the official journal of the Assembly of the Republic and enabled the transcription of committee meetings that were previously undocumented.

Our work distinguishes itself by developing a user-oriented desktop application that integrates pre-trained ASR models into an accessible, privacy-preserving, and locally executed environment. The proposed tool enables automatic transcription of audio or video files for non-specialized users. Furthermore, we conducted an analysis of text transcription in long interviews, reinforcing the usability of Whisper models for the Portuguese language.

3 Ethos AT software

Although there are several ASR approaches in the literature, are scarce studies that present a software application aimed at non-specialist users, combining usability, local processing, and support for long audio and video files. In this work, we propose the development of a desktop application designed to import audio or video files and perform transcription through pre-trained AI models. The main goal of the development was to offer an efficient and secure solution capable of performing transcriptions locally, without the need for connection to external servers. Figure 1 shows the program screen.

[4] https://ethos-at.github.io/

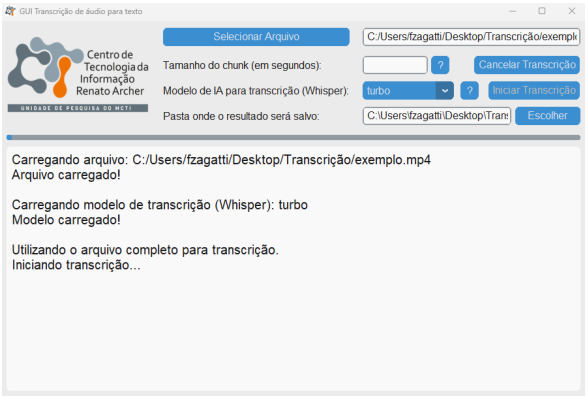

Figure 1: Ethos AT software screen

The software was developed in Python and integrates the OpenAI Whisper library (Radford et al., 2023), recognized for its multi-tasking format and multilingual transcription. Unlike commercial cloud-based solutions, processing occurs entirely on the user's machine, ensuring greater control over sensitive data and suitability for research contexts that require confidentiality, such as those in Oral History. Figure 2 shows a simplified software pipeline, while Algorithm 1 presents the steps for audio transcription.

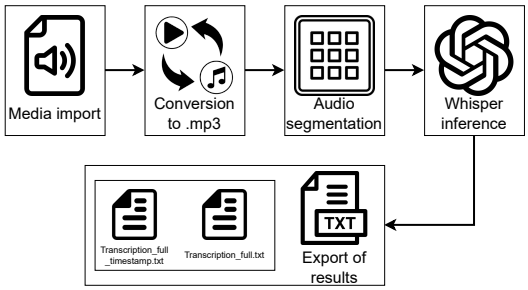

Figure 2: Pipeline of the Ethos AT software for ASR, from audio import to text export

Algorithm 1: Pseudocode of the Ethos AT's internal pipeline

Data: Audio A, Whisper model W, Chunk size C
Result: Result in text s
load the file in A;
if *A is not .mp3* **then**
  transform A into *.mp3* file;
end
if *(C is > 0) and (C < A size)* **then**
  split A into C sized parts;
end
load whisper model W;
apply W in A;
save the result in s;
return s;

As can be observed, the system architecture was designed in a modular way, comprising four main components: (i) media import and conversion module; (ii) segmentation module; (iii) Whisper model inference module; and (iv) results export module. This separation facilitates maintenance and future expansion of the software, allowing the integration of new functionalities such as automatic translation or speaker diarization. In addition, local processing reinforces the data privacy and integrity, essential in research projects involving sensitive material.

The system accepts audio and video files in various formats, namely: *.mp4, *.mkv, *.avi, *.mov, *.mp3, and *.wav. When the file is not in the *.mp3 format, the application automatically performs the conversion, using FFmpeg[5] as an auxiliary tool. Next, the system evaluates the segmentation parameter defined by the user, called *chunk size*, which represents the duration, in seconds, of each portion into which the audio can be divided. If the value provided is greater than the total duration of the audio, or if no *chunk size* is specified, the material is processed entirely in a single stage. This division allows the user to obtain preliminary transcription results, eliminating the need to wait for the entire audio transcription before performing other tasks.

Subsequently, the Whisper model is loaded according to the user's choice, who can select different variants – *tiny*, *base*, *small*, *medium*, *large*, *large-v2*, *large-v3*, and *turbo* – depending on the availability of computational resources and the desired accuracy. The model is applied to each audio segment (or to the complete file), producing two outputs: `transcription_full.txt`, which contains the plain text, and `transcription_full_timestamp.txt`, which associates each sentence with its corresponding time interval.

In summary, the proposed software stands out by combining the technical capability of AI-based transcription models with the practicality required by non-programmer researchers. The local execution, the ability to transcribe long files, and the generation of results with time markers make the tool an accessible alternative to support transcription work in academic and institutional contexts.

3.1 Whisper models

The transcription software utilizes Whisper models, an open-source ASR proposed by OpenAI in 2022,

[5] https://www.ffmpeg.org/

trained on approximately 680,000 hours of multilingual and multitasking audio data collected from various domains and languages (Radford et al., 2023). Whisper uses a Transformer encoder-decoder architecture (Vaswani et al., 2017), inspired by machine translation models. The encoder component converts the audio spectrogram into a high-dimensional vector representation, while the decoder generates the corresponding text, considering the acoustic and linguistic context. Figure 3 demonstrates the text transcription operation.

The Whisper models used by the application differ mainly in the number of parameters, inference speed, and transcription performance. Each version represents a trade-off between computational cost and transcription quality. Table 1 presents a comparative summary of the available models, allowing the user to select the one that best meets their needs for accuracy and execution time.

Model	Size	Required VRAM	Relative Speed	Parameters
tiny	72.1 MB	∼1 GB	∼10x	39 M
base	139 MB	∼1 GB	∼7x	74 M
small	461 MB	∼2 GB	∼4x	244 M
medium	1.42 GB	∼5 GB	∼2x	769 M
large	2.88 GB	∼10 GB	∼1x	1550 M
large-v2	2.87 GB	∼10 GB	∼1x	1550 M
large-v3	2.88 GB	∼10 GB	∼1x	1550 M
turbo	1.51 GB	∼6 GB	∼8x	809 M

Table 1: Features of the Whisper models available for automatic transcription, adapted from OpenAI Whisper's GitHub[6]

In this context, the smaller versions (*tiny*, *base*, and *small*) offer faster processing speeds, making them suitable for devices with lower hardware capabilities or for quick transcriptions. The larger variants (*medium*, *large*, *large-v2*, and *large-v3*) achieve better performance in terms of accuracy, especially with longer audio files or those with greater acoustic complexity. The *turbo* model, inspired by Distil-Whisper (Gandhi et al., 2023), represents an optimization focused on accelerating inference while maintaining quality levels comparable to the larger models.

According to a benchmark conducted by OpenAI, the *turbo* model shows a performance level comparable to *large-v2*, although it exhibits higher degradation in certain languages, such as Thai. Figure 4 presents the performance of the model in the top-10 transcription tasks using the Common Voice 15 dataset[7], based on both Word Error Rate (WER) and Character Error Rate (CER) metrics.

In our developed application, the user can select the desired model according to their performance and accuracy needs. This flexibility seeks to balance computational efficiency and transcription fidelity, allowing the software to be used in both technical experimentation contexts and practical research applications. Furthermore, by performing processing entirely locally, the use of Whisper models ensures that user data is not transferred to or stored on external servers, respecting information security principles and research ethics.

4 Methodology and evaluation process

The experiments were conducted on a computer with an Intel Core i7-1165G7 processor, 16 GB of RAM, and an NVIDIA GeForce MX330 graphics card with 2 GB of GDDR5 memory, an intermediate configuration that represents a realistic scenario for researchers and professionals seeking to perform local transcriptions efficiently, without the need for high-performance infrastructure.

4.1 Interview audios

Evaluations involve interviews with different levels of complexity (such as variations in accent, speech overlap, audio quality, and the use of technical terms), allowing a critical analysis of the Whisper's ability to recognize and transcribe the fundamental elements of oral narratives, as detailed below.

- **Processing speed:** For long files such as Oral History interviews, the reduction of transcription time has a direct impact on productivity. This enables researchers to handle larger volumes of material in less time, without compromising the quality of the transcription process.
- **Varied audio quality:** Recordings with different levels of background noise, volume, and clarity were used to evaluate the Whisper's ability to maintain transcription accuracy under suboptimal conditions.
- **Regional accents:** Interviews with speakers from different regions of Brazil to verify the capacity to recognize linguistic variations and informal vocabulary. Regional speech patterns contributes to the authenticity and interpretive value of the transcribed material.

[6] https://github.com/openai/whisper

[7] https://huggingface.co/datasets/fsicoli/common_voice_15_0

[8] https://github.com/openai/whisper/discussions/2363

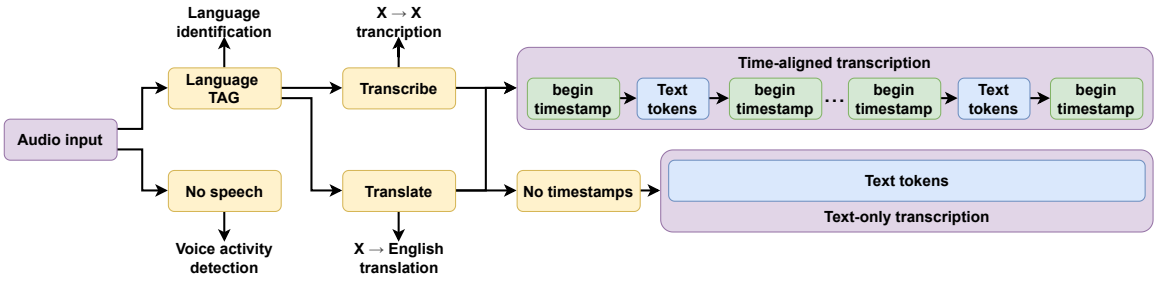

Figure 3: Text transcription (or translation) operation for new audios using Whisper models (Radford et al., 2023)

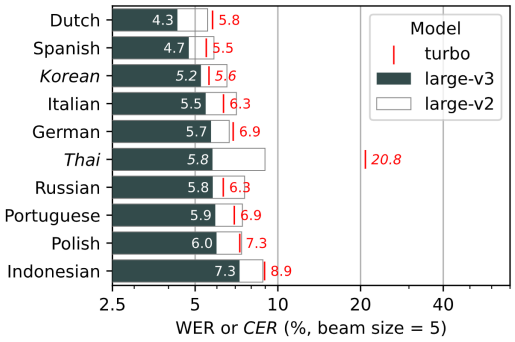

Figure 4: Comparison of transcripts in different languages, considering the metrics Word Error Rate (WER) and Character Error Rate (CER), from *turbo* model discussion on GitHub[8]

- **Presence of multiple speakers:** Audio recordings feature interaction between the interviewer and the interviewee, resulting in segments with overlapping voices and rapid alternation between speech turns. This evaluates dialog dynamics, ensuring that speaker alternation does not compromise transcription accuracy or temporal alignment.

It is important to note that the interviews used in this evaluation contain sensitive data from the Center for Information Technolgy Renato Archer (CTI Renato Archer), and therefore cannot be publicly disclosed due to confidentiality and data protection agreements. Nonetheless, the developed software is not limited to this specific context. Its modular and adaptable design allows it to be applied to a wide range of audio materials, including lectures, meetings, oral history archives, podcasts, and other research or institutional recordings that require accurate and secure transcription.

4.2 Evaluation by human specialist

The evaluation of the software involves the participation of a human specialist in Oral History, intending to verify the automatically generated transcriptions and the usability of the tool. The validation process includes a comparison between automatic transcriptions and manually revised versions, allowing measurement of fidelity to the original content and identification of the most frequent types of errors, such as spelling of proper names, technical terms, and common words.

In addition to linguistic accuracy, the evaluation includes usability aspects of the interface, considering the clarity of the available functions, the ease of switching between different transcription models, and the export options for the results. This stage seeks to understand the researcher's experience during the use of the tool, ensuring that the software is accessible and intuitive to users without experience in data science or programming.

5 Data security and information privacy

Information security is grounded on the CIA triad (Goodman and Rowland, 2021): Confidentiality, Integrity, and Availability, which serve as references for the creation and development of data protection policies in digital environments. Confidentiality ensures that information is accessed only by those with proper authorization, while integrity guarantees that data has not been modified. Availability means that data will remain accessible over time.

Beyond these basic information security issues, there is an additional concern in works like this, because they contain data and information derived from the lived history of a community and must comply with legal requirements, especially the Brazilian General Data Protection Law (LGPD). In this sense, local data processing – as implemented in this system – ensures that all storage and analysis occurs in a controlled environment. By preventing sensitive information from being transmitted to external servers or the cloud, this approach mitigates risks of interception, unauthorized changes, and improper exposure of personal data

while ensuring that data handling respects the requirements of purpose and necessity established by LGPD. Furthermore, converting voice into text in the transcription process enhances anonymization, making it unfeasible to identify the speaker.

Therefore, by opting for local storage, this work demonstrates a strong commitment to data protection and compliance with current legislation, thus reinforcing the integrity of the research and trust within the scientific community.

6 Result analysis

A comparative analysis was developed to evaluate the performance of different artificial intelligence models for ASR, including the exploration of proper names, acronyms, and technical vocabulary. A total of 67 words and expressions were evaluated in six different models: *small*, *medium*, *large*, *large-v2*, *large-v3* and *turbo* (Radford et al., 2023). Note that the *tiny* and *base* models were excluded from the evaluation, as they proved to be the least efficient models during preliminary screening. The selected 67 words correspond to some items that exhibited recognition issues in at least one of the tested models, ensuring that the analysis focused on challenging and representative cases for ASR evaluation. This analysis aimed to guide the selection of the most suitable model for the transcription task.

The difference between the transcriptions of the six models involves a lack of precision in proper names (people and places), institutional acronyms, technical and academic terms, and common words. Table 2 shows a summary of these errors. The complete table is available at Zenodo[9].

The confusion between "CTI" and "CPI" demonstrates the weakness in recognizing technical and institutional vocabulary, that can compromises textual reliability in specific contexts. Technical terms were also distorted, such as "intangível" (*intangible*) becoming "intergível" (no translation).

The *turbo* model demonstrates greater consistency and accuracy, suggesting superior performance. The *large*, *large-v2* and *large-v3* models also show satisfactory results, with a high number of correct recognitions. On the other hand, the *small* and *medium* models have a high error rate, frequently replacing or distorting terms. Table 3 shows a summary of these errors.

It was observed that the *small* model, although faster, presented some inconsistencies in the identification of words that are easily understood in the audio by humans. Word substitution errors, omissions, and difficulties with proper names were more frequent in this model. The *medium* model halves the errors of the *small* model, but still presents a considerable number of incorrect words.

As more robust models were applied (*large*, *large-v2*, *large-v3*, and *turbo*), there was an improvement in transcription accuracy, with a greater ability to recognize complex terms, natural speech pauses, and more appropriate segmentations. Curiously, despite achieving very similar results, the *turbo* model performed slightly better than the *large-v3* model, which differed from what was reported in OpenAI's research for Portuguese (Figure 4). This evaluation shows that the more advanced models offer transcriptions that are closer to the original speech, although they require greater processing capacity and execution time. Table 4 presents a comparison of the processing times.

The execution time analysis highlights a clear trade-off between transcription accuracy and processing efficiency. As shown, the *small* and *medium* models deliver results significantly faster, with the *small* model completing the transcription in half the time of the original audio for interviews #1 and #2. However, this speed comes at the cost of higher error rates. Conversely, the *large* and *large-v3* models require substantially longer execution times, exceeding three and four hours for some audios. The *turbo* model, designed for efficiency, demonstrated a promising balance between precision and speed, maintaining accuracy levels close to *large-v3* while reducing processing time by more than half. It was also observed that interviews #2 and #4 presented a longer processing time than the other instances when using the *large* model. This longer processing time may be linked to factors that can affect computational cost, such as speech clarity, speech speed, vocabulary used, and even the computer's usage time and temperature. Despite the processing time, it is worth highlighting the software's advantages: it is free, prioritizes transcription quality, and prioritizes information security, since the process occurs locally without sending data to the cloud.

Recent literature on qualitative research methods reports that manual transcription is both time-consuming and costly (Battaglia, 2024), with estimates indicating that approximately five hours are required to transcribe a single hour of interview

[9]https://doi.org/10.5281/zenodo.17583668

Word type	Examples (Correct - Incorrect)	Models with errors
Proper names	**"Angela"** - "Jolene", "Lajla", "Juliette"	Small and Medium
	"Edgleide" - "Aglídea", "Eliglory", "Glíria", "Glídia", "Glíria", "Edgluide"	All models
	"Romildo" - "Romulo"	Small
Acronyms	**"MicroMed"** - "micromédio", "micromédia"	All models
	"CTI" - "CPI"	Large, Large-v2 and Large-v3
	"MAST" - "MASH"	Large
Technical terms	**"automação"** - "Timentação", "Burjultron", "abertura", "abro os"	All models
	"intangível" - "intergível", "inteligente"	Medium and Large-v3
	"tripartite" - "de partite"	Small
Common words	**"boiadeira"** - "banhadeira", "brasileira"	Small, Medium and Large-v2
	"optei" - "obtei", "obtenho"	Medium and Large-v2
	"Trajetória" - "tranjetória", "transição"	Small and Large
Foreignism	**"cluster"** - "clustering"	Small
	"commodities" - "comótito", "comódito"	All models
	"BigTechs" - "big-tex", "big tecs"	Small and Large

Table 2: Examples of transcription variations observed among Whisper models

Model	Accuracy	Observations	Num. of errors
Small	Low	Many incorrect names, random substitutions, severe distortions in vocabularies.	60
Medium	Reasonable	Retains some errors, especially in proper names, technical words and some common words.	31
Large	Good	Corrects several words, but still fails with proper names and acronyms.	22
Large-v2	High	Some failures in acronyms and names, but with progress compared to previous models. There are still distortions in technical terms.	21
Large-v3	High	Few errors, stable performance, but some failures in acronyms and technical terms persist.	21
Turbo	High	Shows the best results, with minor errors and contextual consistency. However, reproduces some errors already improved in previous models.	20

Table 3: Transcription comparison among different models

Summary		Transcription models					
Interview	**Audio time**	**Small**	**Medium**	**Large**	**Large-v2**	**Large-v3**	**Turbo**
#1	34m55s	14min	33min	1h13min	1h03min	2h	58min
#2	1h05m25s	37min	1h13min	4h18min	2h35min	3h30min	1h02min
#3	1h22m19s	1h09min	2h24min	3h28min	2h48min	3h32min	2h32min
#4	1h13m31s	1h39min	1h21min	4h46min	2h19min	3h10min	1h18min
#5	1h05m51s	33min	1h18min	3h31min	3h14min	3h20min	1h15min

Table 4: Interview processing times for different Whisper models

recording (Bokhove and Downey, 2018). Our case studies have shown that the total time spent transcribing and reviewing each interview takes at least three days. Ethos AT streamlines this transcription process, allowing researchers to spend more effort on other activities, such as text revision.

In this context, developing academic and institutional projects requires high writing accuracy, so errors in names, acronyms, and technical terms compromise the software's reliability. In Oral History interviews, the use of the *turbo* or *large-v3* models is justified by their superior lexical fidelity, although it should be noted that these models sometimes lose precision in identifying certain words that were correctly recognized by smaller models. On the other hand, if the task requires greater speed and can tolerate a higher error rate, smaller models such as *small* or *medium* are more suitable.

In addition to evaluating transcription accuracy, the graphical interface (Figure 1) was developed to provide a simple and intuitive experience, allowing users with no programming knowledge to perform automatic transcription tasks efficiently. The interface includes buttons to select the audio or video file, define the *chunk size*, choose the desired Whisper model, and indicate the output directory

to save the result. To assist the user, there is a dynamic log panel that displays each processing stage, from file loading to the transcription itself. This transparency helps the user understand what is happening in real time, minimizing uncertainty during longer transcription tasks. Additionally, the interface features a *Cancel Transcription* button that allows the user to interrupt the process.

The software also provides contextual help buttons (**"?"**) next to key parameters, giving short explanations about each setting. This feature reduces the learning curve and reinforces accessibility for users outside the field of computer science. The design emphasizes a balance between visual simplicity and functional completeness, following principles of user-centered design.

During the interface testing phase, an occasional error message ("[WinError 5] Access denied: 'chunks"') was observed on some computers, especially those with restricted permissions or synchronized folders (e.g., OneDrive directories). This issue is related to file access authorization during the creation of temporary audio segments for transcription. Although it did not compromise the overall functionality of the system, the event highlights the importance of verifying write permissions in the selected output directories. Future versions of the application will include an automatic verification step to prevent this problem and improve user feedback when such access restrictions occur.

In summary, the system demonstrated good usability and operational robustness during testing. Users were able to complete the entire transcription workflow, from file selection to the generation of the final text, without requiring technical support. The straightforward structure of the interface makes the software suitable for research environments and academic transcription activities, aligning with the objective of democratizing access to speech recognition technologies. However, an occasional access error was observed on some computers; although this behavior did not affect the overall system performance, it revealed the need for improvements in permission management and user feedback in future versions.

7 Conclusion and future work

This study presented a software named Ethos AT and a comparative analysis of different Whisper models applied to the automatic transcription of Oral History interviews. The results demonstrate that, although smaller models (*small* and *medium*) offer faster processing times, their accuracy is significantly affected, particularly in the recognition of proper names, acronyms, and technical vocabulary. Larger models (*large*, *large-v2*, and *large-v3*) provide more faithful transcriptions, capturing complex expressions and nuances of spontaneous speech, but at the cost of higher computational demand and longer execution time.

In contrast, the *turbo* model reduced inference time with accuracy comparable to the larger versions. These findings suggest that recent optimizations represent a promising direction for a balance between performance and efficiency in ASR applications. However, variations in performance depending on language and acoustic conditions highlight the need for continuous evaluation and adaptation of models to specific linguistic contexts.

Beyond quantitative measures, this work reinforces the role of automatic transcription in democratizing access to oral content. The integration of these tools in research environments can accelerate documentation workflows and broaden access to speech-based resources for diverse communities. However, human review remains necessary to verify and correct erroneously transcribed terms in all models used, demonstrating the need for human agents for final evaluation.

As future work, we intend to implement hybrid approaches that combine Whisper models with specialized institutional glossaries, improving the recognition of acronyms, foreign words, and technical terms. Additionally, we plan to integrate speaker diarization, enabling the automatic distinction of different speakers within the same recording and improving the structure and usability of the transcriptions, particularly in multi-participant interviews. And finally, we aim to improve the graphical interface and correct errors, aiming for greater compatibility between operating systems and stability in execution.

Acknowledgments

The authors gratefully acknowledge the financial and research support provided by Centro Universitário Max Planck (UniMAX), Centro Universitário de Jaguariúna (UniFAJ), the Center for Information Technology (CTI) Renato Archer, and the Brazilian Ministry of Science, Technology and Innovation (MCTI). Their support was essential for the development and execution of this research.

References

Fabio Battaglia. 2024. From "listen and repeat" to "listen and revise": how to transcribe interviews offline quickly and for free using voice recognition software. *International Journal of Qualitative Methods*, 23:16094069241247473.

Christian Bokhove and Christopher Downey. 2018. Automated generation of 'good enough'transcripts as a first step to transcription of audio-recorded data. *Methodological innovations*, 11(2):2059799118790743.

Arnaldo Candido Junior, Edresson Casanova, Anderson Soares, Frederico Santos de Oliveira, Lucas Oliveira, Ricardo Corso Fernandes Junior, Daniel Peixoto Pinto da Silva, Fernando Gorgulho Fayet, Bruno Baldissera Carlotto, Lucas Rafael Stefanel Gris, and 1 others. 2023. Coraa asr: a large corpus of spontaneous and prepared speech manually validated for speech recognition in brazilian portuguese. *Language Resources and Evaluation*, 57(3):1139–1171.

Mohsen Fatehifar, Josef Schlittenlacher, Ibrahim Almufarrij, David Wong, Tim Cootes, and Kevin J Munro. 2025. Applications of automatic speech recognition and text-to-speech technologies for hearing assessment: a scoping review. *International Journal of Audiology*, 64(6):537–548.

Sanchit Gandhi, Patrick Von Platen, and Alexander M Rush. 2023. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv preprint arXiv:2311.00430*.

Howard B. Goodman and Pam Rowland. 2021. Deficiencies of compliancy for data and storage. In *National Cyber Summit (NCS) Research Track 2020*, pages 170–192, Cham. Springer International Publishing.

Michael Gref. 2022. *Robust Speech Recognition via Adaptation for German Oral History Interviews*. Ph.D. thesis, Universitäts-und Landesbibliothek Bonn.

Lucas Rafael Stefanel Gris, Arnaldo Candido Junior, Vinícius G dos Santos, Bruno A Papa Dias, Marli Quadros Leite, Flaviane Romani Fernandes Svartman, and Sandra Aluísio. 2022. Bringing nurc/sp to digital life: the role of open-source automatic speech recognition models. *arXiv preprint arXiv:2210.07852*.

Lucas Rafael Stefanel Gris, Ricardo Marcacini, Arnaldo Candido Junior, Edresson Casanova, Anderson Soares, and Sandra Maria Aluísio. 2023. Evaluating openai's whisper asr for punctuation prediction and topic modeling of life histories of the museum of the person. *arXiv preprint arXiv:2305.14580*.

Armin Haberl, Jürgen Fleiß, Dominik Kowald, and Stefan Thalmann. 2024. Take the atrain. introducing an interface for the accessible transcription of interviews. *Journal of Behavioral and Experimental Finance*, 41:100891.

Naoyuki Kanda, Xiong Xiao, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka. 2022. Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed asr. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8082–8086. IEEE.

Eduardo Medeiros, Leonel Corado, Luís Rato, Paulo Quaresma, and Pedro Salgueiro. 2023. Domain adaptation speech-to-text for low-resource european portuguese using deep learning. *Future Internet*, 15(5):159.

Pedro Nascimento, João C Ferreira, and Fernando Batista. 2024. Automatic transcription system for parliamentary debates in the context of assembly of the republic of portugal. *International Journal of Speech Technology*, 27(3):613–635.

Leontien Pragt, Peter van Hengel, Dagmar Grob, and Jan-Willem A Wasmann. 2022. Preliminary evaluation of automated speech recognition apps for the hearing impaired and deaf. *Frontiers in Digital Health*, 4:806076.

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

Antonio Joaquim Severino. 2016. História oral como arte da escuta, de alessandro portelli. *EccoS–Revista Científica*, (41):238–243.

Xian Shi, Yanni Chen, Shiliang Zhang, and Zhijie Yan. 2022. Achieving timestamp prediction while recognizing with non-autoregressive end-to-end asr model. In *National Conference on Man-Machine Speech Communication*, pages 89–100. Springer.

Linda Shopes. 2002. Making sense of oral history. *History matters: The US survey course on the web*, 6.

Radina Stoykova, Kyle Porter, and Thomas Beka. 2024. The ai act in a law enforcement context: The case of automatic speech recognition for transcribing investigative interviews. *Forensic Science International: Synergy*, 9:100563.

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