Automatic Transcription Systems: A Game Changer for Court Hearings

Alan Lyra[1], Carlos Eduardo Barbosa[1,2], Herbert Salazar[1], Matheus Argôlo[1], Yuri Lima[1], Rebeca Motta[1] and Jano Moreira de Souza[1]
[1]*Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil*
[2]*Centro de Análises de Sistemas Navais (CASNAV), Marinha do Brasil, Rio de Janeiro, Brazil*

Keywords: Automated Transcription, Court Hearings, Computing, Legal Documentation, Decision-Making Support.

Abstract: Lawsuits typically require a long time for resolution, and many court hearings may occur during the trial process. Legally, both parties must transcribe them and be open to the public if desired by the court. In court proceedings, a transcript is a record of all judges' decisions, the spoken arguments by the lawyers, and the depositions of the defense and witnesses. The scenario in Brazil is that for a long time, this process was manual, with a person responsible for the typing transcription. Today, with the electronic process, the court does not provide typed transcriptions anymore, but instead, the audio or video recordings of the hearings. In our work, we developed an automatic transcription solution for court hearings to obtain the best possible transcription considering current technologies' limitations, recording quality, participants' diction, and commonly used jargon in the legal sphere. With this work, we expect to ease this burdensome task with technical support and have a direct contribution to the legal environment.

1 INTRODUCTION

In the legal sphere, several testimonies and hearings resolve lawsuits. Historically, specialized individuals manually transcribed, making the process expensive and time-consuming. Technology can have many applications in court proceedings and facilitate a comprehensive end-to-end process, whether generating audiovisual media records or even processing information – transcribing (Nadaraj & Odayappan, 2020).

Organizing and managing these digital records can potentially contribute to court proceedings since this type of media can deliver better results, with the possibility of capturing subtle insights that cannot be written on paper, such people reactions.

We believe that transcribing hearing content allows easier assimilation of all procedural content by those who come to work with the proceeding. This approach does not exclude audiovisual recording but can highlight significant sections in procedural documents, including citations that can be easily referenced and classified (Gomes & Furtado, 2017).

Courts have applied new technologies and services to manage information and records of vast legal content. Services such as e-Litigation ("Integrated Electronic Litigation System," 2020) and e-Discovery ("Electronic Discovery," 2020) are examples that contribute to this approach. With the COVID-19 pandemic, the digital court is being experimented with and evolving daily to a more digital style (Nadaraj & Odayappan, 2020).

This work aims to organize the data generated from court hearings through a knowledge management process and transcribe the media using an automatic transcription system. We also try to refine this automated process to reach the highest level of accuracy through collaboration mechanisms.

This transcript, associated with the trial process and the corresponding media, will help significantly unfold a court lawsuit. In the current scenario – in Brazil's legal sphere – reports are generated after court hearings for decision-making and motions. However, reports and minutes – the common strategies to keep the memory of a court hearing – are formal and often limited, mostly when the institution handles extensive multi-stakeholder legal processes.

According to Chiu et al. (2001), the minutes of a trial process are summaries of it, constituting a part of the procedural memory. Soon after a hearing, looking at the minutes to review and act based on that

Lyra, A., Barbosa, C., Salazar, H., Argôlo, M., Lima, Y., Motta, R. and Moreira de Souza, J.
Automatic Transcription Systems: A Game Changer for Court Hearings.
DOI: 10.5220/0012891400003838

In *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2024) - Volume 3: KMIS*, pages 163-174
ISBN: 978-989-758-716-0; ISSN: 2184-3228

information is often helpful. Even during a hearing, it may be useful to refer to something from an earlier point, for example, by asking a question about a previous event in the legal process. For this reason, we argue that proposing a system that presents information from the legal process with mechanisms that add value to the decision-making can contribute to advancements in the legal system.

One of the challenges faced in the solution is adapting the existing Portuguese language technologies in a world where English has more significant scientific potential and technology options. The second challenge is regarding transcription correctness. Regardless of the efficiency, errors will appear due to the recording quality, participants' diction, and even jargon commonly used in the legal field. Another crucial challenge for developing the transcription system is voice identification through participants' voice biometry, which enables the automatic creation of a dialog in text format from the hearing inputs.

2 LITERATURE

This section overviews the Brazilian legal process and how the resulting media are processed. We also discuss the essential elements of knowledge management.

2.1 Legal Process in Brazil

Understanding how the trial process and hearings work, with motions, judges, and different courts, is not simple for those not from the legal field. The step by step of the legal process can vary according to the matter involved (civil law, criminal law, tax, etc.). Generally, a legal process consists of a request from the author to resolve a conflict. From this, the judge will determine the presentation of reasons and the production of evidence and will decide to recognize the right of one of the parties.

In the Brazilian legal process, the court communicates all steps to the parties through publications in the *Diário Oficial* (the official journal of the Brazilian Government). Nowadays, lawyers and stakeholders can monitor these proceedings online, facilitating quicker and more efficient processing. Additionally, legal software like ProJuris (2024), available on the market, can optimize the court's progress.

The court hearing is a unique and significant event that defines the fate of a lawsuit. It is a complex, agile, and dynamic event and usually results in the lawyer's only chance to demonstrate the Constitution (Brasil, 1988), the impediment, modification, or extinction of a right.

Figures 1, 2, and 3 present a simplified business process model (BPM) (OMG, 2011) of the Brazilian Legal Process, showing the generic steps from beginning to end.

Court hearings are the primary audio/video media producer during the legal process steps. The proposed system aims to use that media to automatically transcribe its contents to support the law worker, enabling them to perform faster and more efficient work. Analyzing the BPM, we highlight that court hearings often occur more than once in a lawsuit. According to Rio de Janeiro's General Public Defender, Dr. André Luis Machado de Castro, and the General Public Subdefender, Dr. Denis de Oliveira Praça (personal communication, Ago 07, 2018), analyzing the transcription of court hearings offers a straightforward and fast way to make their decisions, presenting a significant improvement in the working dynamics in the legal environment.

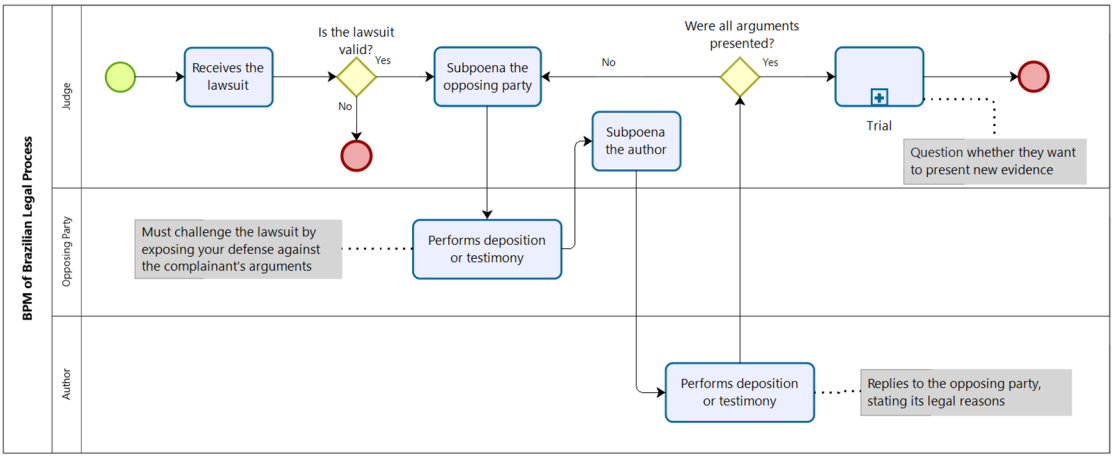

Figure 1: BPM of Brazilian Legal Process.

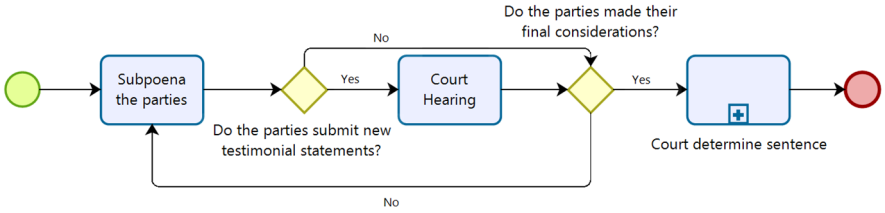

Figure 2: BPM of Brazilian Generic Subpoena Subprocess.

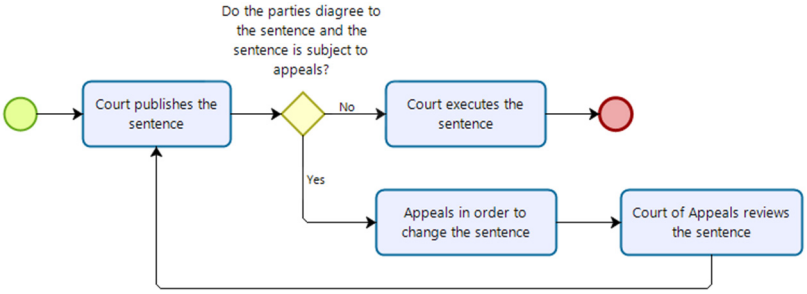

Figure 3: BPM of Brazilian Generic Verdict Subprocess.

Courts typically store media files from hearings in a digital lawsuit process. However, many legacy lawsuits still run on paper, with court hearing data stored in CD/DVD media. The most common type of lawsuit in Brazil that records court hearings is criminal. Civil cases rarely have their court hearings recorded.

2.2 Knowledge Management

Knowledge Management (KM) is a broad research topic, and as a discipline, it has several definitions, perceptions, and approaches. We understand KM as retrieving and organizing an organization's data and information to make it accessible to support decision-making and achieve strategic objectives through the discovered knowledge. The interest in and use of KM are increasingly expanding as more and more data are being produced. The KM is an area that permeates the companies' administration, information systems, management, and data science, with many research and practical and actual case initiatives for KM solutions (Patel, 2012).

Knowledge is a justified belief that enhances an entity's ability for effective action (Huber, 1991; Nonaka, 1994). Alavi and Leidner (2001) propose six different perspectives for observing knowledge: 1) Data and Information, 2) A state of mind, 3) An object, 4) A process, 5) A condition to have access to information, and 6) A capacity.

The first perspective of Data and Information shows only facts, raw numbers, processed information, and interpreted data. The second perspective of knowledge as a state of mind focuses on enabling individuals to expand their knowledge and apply it to the organization's needs. The third view defines knowledge as an object and postulates that knowledge can be seen as something to be stored and manipulated. Alternatively, knowledge can be seen as a simultaneous process of knowing and acting (Carlsson et al., 1996; McQueen, 1998; Zack, 1998), focusing on applying knowledge (Zack, 1998). The fifth view of knowledge is a condition of access to information (McQueen, 1998). According to this view, organizations must organize knowledge to facilitate access and retrieval of content, extending the knowledge base. Finally, in the latter perspective, knowledge can be viewed as a capacity that can influence future actions (Carlsson et al., 1996).

Along with understanding how we can observe knowledge, it is also paramount to have alternatives for managing it. Regardless of the definition, KM presents some core activities and central factors that structure any KM model (Stollenwerk, 2001). The generic KM process has the following activities (Alegbeleye, 2010; Dhamdhere, 2015; Mutula & Mooko, 2008):

- **Identification:** this activity focuses on strategic issues, such as identifying which competencies are relevant to the organizational context.
- **Capture:** in this activity, the objective is to acquire knowledge, skills, and experiences necessary to create and maintain the core competencies and areas of knowledge selected and mapped.

- **Selection and Validation:** this activity aims to filter knowledge, evaluate its quality, and synthesize it for future application.
- **Organization and Storage:** the goal is to ensure fast, easy, and correct knowledge recovery through effective storage systems.
- **Sharing:** this activity aims to ensure information and knowledge access to a more significant number of people that otherwise would remain restricted to a small group of individuals. Also, knowledge distribution refers to implementing a mechanism capable of automatically disseminating knowledge to various stakeholders to share new knowledge with those who need it quickly.
- **Application:** here, the objective is to put into practice the knowledge disseminated by the previous process, recording lessons learned from the use of knowledge, the benefits, and the challenges to be overcome.
- **Creation:** creating new knowledge involves learning, externalization, lessons learned, creative thinking, research, experimentation, discovery, and innovation. Many organizational activities can contribute to the creation of new knowledge. According to Nonaka (1994), knowledge creation is related to continuously transforming and adapting different types of expertise, such as practice and interactions.

Based on the objectives of this research area, the understanding and use of KM can support knowledge treatment in several places. In this work, we used these principles to guide this research related to the hearings of legal processes.

2.3 Related Work

Similar work was proposed by Huet (2006) when he created the Transcript Coding Scheme (TCS), which was a tool used to encode transcripts of corporate meetings – meetings – through the analysis of audio recordings, which remains the same.

Huet (2006) states that the data collected and explained in the transcription scheme offer the reader a vast amount of information, some of which may be considered critical and not captured using traditional reading techniques. However, processing and interpreting these highly detailed records require significant time and effort. Experiments have shown that accurately transcribing and encoding 30 minutes of recording takes approximately 10 hours.

The scheme proposed by Huet, to be better understood, can be divided into two parts:

- **Transcriber:** This step explicitly shows the transcription basics, stating the user identification, the speech transcription, and the timestamp when the speech occurred.
- **Coding Scheme:** This step provides information that aims to complement the transcript to offer more details to facilitate the decision-making process: type of speech; the purpose of speech; type of information provided; the argument; type of argument; participant's knowledge area; description; and origin of speech.

Another work concerning transcriptions is by Williams et al. (2011), which proposes crowdsourcing techniques in complex transcriptions that allow explicit exchanges between precision, recall, and cost. These techniques include automatic transcription, incremental collaboration redundancy, and a regression model to predict the transcription's reliability. The main idea is to use the crowd to assist in the final result of a transcription performed by an automatic transcription system (Parent & Eskenazi, 2010) based on requests for Amazon's MTurk (Crowston, 2012). The techniques of redundancy and regression are defined by refining the process to validate a collaboration through the redundancy of correct answers through algorithms.

3 THE PROPOSED KM PROCESS

The developed KM model manages the information associated with the court hearing and produces a faster and better understanding for law workers. Analyzing the KM activities related to lawsuit trials, we understand that knowledge creation occurs during court hearings (opening statements, witness examination, closing arguments, and jury verdicts). We propose a Knowledge Management Model for Court Hearings, inspired by the study of Motta et al. (2022), which is focused on identifying, acquiring, and storing knowledge, as presented in Figure 4.

We detail each step in the proposed process:

- **Hearing Preparation:** The court hearing occurs when there is a need to collect information about the lawsuit from the parties or witnesses. The judge decides when and who will attend the court hearing.
- **Hearing Conduction:** This is a hearing instance. In a lawsuit, several hearings are made and recorded. This work relies on hearing the recording.

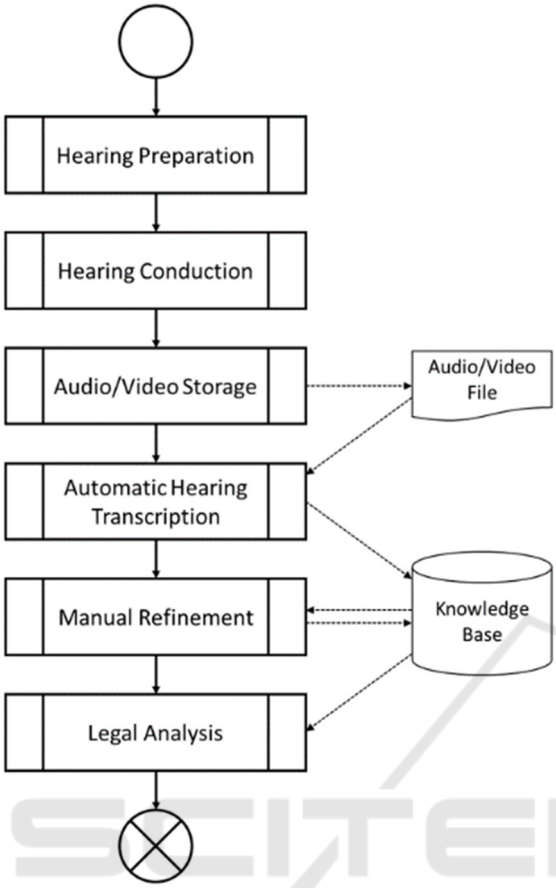

Figure 4: Knowledge Management Process for Court Hearings. Adapted from Motta et al. (2022).

- **Audio/Video Storage:** the default way to capture the hearing is to record audio or video media. Such an approach is simple to implement and frees the members of the accuracy of the stored hearing data – audio/video media is precise to solve any doubt about a hearing.
- **Automatic Hearing Transcription:** In this step, the system provides an automatized transcription of the hearing and determines the speaker of each sentence from the hearing audio/video. All knowledge will be stored in a shared database, allowing legal memory (processes, hearings, transcripts, and the source of the significant decisions).
- **Manual Refinement:** after the transcript generation, the result becomes available to be evaluated and fixed by any member of the judicial apparatus or the population in general through collaborative tools. This step fixes errors due to poor audio quality, poor diction, or legal jargon.
- **Legal Analysis:** based on the transcription, independent of the manual refinement, the law worker (judge, prosecutor, or lawyer, for example) may proceed in the hearing analysis using the text instead of the audio/video. Using specific keywords, they can find particular moments from the hearing without watching the whole media, in many cases, more than once. Such ability enables faster analysis of the court hearings. Besides, since the transcription is temporally associated with the media, it may be seen as a proxy to the media, allowing searching a video for the keyword spoken, for example. Although the transcription may have errors, the source (audio/video) is accurate.

4 THE PROPOSED TRANSCRIPTION SYSTEM

The developed system aims to contribute to the knowledge management of hearings by their automatic transcription. This solution enables a party to visualize the legal process and its transcription, locate other related audiences, and conduct a textual search of the available hearings. The system seeks to facilitate the work of the legal professional who will analyze the process, providing features like:

- Display of the media corresponding to the audience or hearing;
- Automatic transcription of this media along with the participant identification;
- Caption inserted in the media for direct accompaniment in the video;
- Search for words associated with the legal process; and
- Edit the transcript manually.

The system automatically generates the transcription scheme at the end of a hearing, providing the transcription and identifying the participants. The system also associates the data with the legal process, participant role (i.e., defense), and trial results to enrich the hearing. The hearing process is recorded in a media file (audio or video) during its execution and saved in a database. The system automatically generates and indexes a transcript for each individually submitted hearing.

In investigating which transcription method to use, we considered the following factors: accuracy, cost, and support for the Portuguese language. We initially evaluated two open-source transcription software – Kaldi (Povey et al., 2011) and

Pocketsphinx (Huggins-Daines et al., 2006) – but they only support English. Next, we assessed the Google Cloud Speech-to-Text API (Google, 2020a), which supports many languages – including Portuguese. The Google API presented a lower error rate in our tests, and we selected it for our solution. We use Java (Arnold et al., 2005) and FFmpeg (Tomar, 2006) technologies to complement the Google API to implement the full solution.

4.1 Diarization

We implemented the speaker identification algorithm. Firstly, we developed a speaker diarization system based on a convolutional neural network capable of identifying speech in a speaker's frontal video without using the associated audio wave for use cases in which the latter is either low quality, noisy, or outright missing. For this purpose, we extracted facial landmarks from each video frame using a facial identifier present in the Dlib (King, 2009) library based on the Histogram of Oriented Gradients (HOG). We fed them into the neural network using Tensorflow (Abadi et al., 2016), as presented in Figure 5.

With the onset of the COVID-19 pandemic and societal restrictions such as the mandatory use of face masks, the technique developed to detect facial landmarks became unfeasible, rendering diarization impractical. Considering this, we shifted our approach by creating an audio-based diarization model.

Speaker diarization involves identifying different speakers in multimedia content to temporally separate them, defining who spoke when, and producing a script. Traditionally, audio analysis has tackled this problem by extracting features as I-vectors (Dehak et al., 2011) and subsequently clustering them. Typically, these audio processing algorithms rely on prior knowledge of the number of speakers involved in the audio.

Google researchers proposed the utilization of a Long Short-Term Memory (LSTM) for creating feature vectors, termed D-vectors (Wang et al., 2017). They applied a clustering technique called Spectral Clustering, which involves constructing an affinity matrix among the examples and applying specific refinement operations to consider the temporal locality of the data, aiming to smooth, normalize, and remove noise from the data. It enabled a dynamic identification of the number of speakers in the audio.

We trained the model using video datasets from public hearings. We extracted and converted the audio signals into 25-millisecond frames with a 10-millisecond step. During the testing phase, we used the Voice Activity Detection (VAD) algorithm to identify spoken segments in the audio. Extracting audio signals is the first step in preparing them for the I-vector learning algorithm due to the complexity of raw signals. Initially, filter banks of the signals are computed to isolate different frequency components. We chose to extract Log Mel Frequency Energy filter banks, which involve performing a Fourier Transform, mapping the spectrum to the Mel scale, and other post-processing steps, as shown in Figure 6.

We implemented two models for feature extraction: Gaussian Mixture Model (GMM) and LSTM. Traditional models for I-vector extraction typically use dimensionality reduction of the GMM super-vector.

The final step of our implemented model involves clustering the feature vectors (I-vectors or D-vectors) extracted in the previous step. The chosen clustering models were: Spectral Clustering (Wang et al., 2017), Density-based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996), Agglomerative Clustering (Hastie et al., 2009), and Self Organizing Maps (SOM) (Kohonen, 1990) followed by K-Means (MacQueen, 1967). We used different libraries for each clustering model. Figure 7 illustrates the performed activities. We fine-tuned each clustering algorithm.

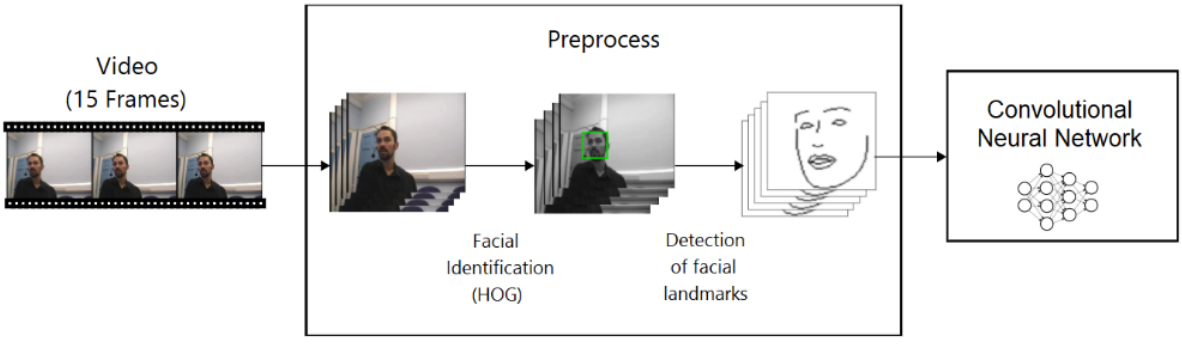

Figure 5: The sequence of data loading and model training.

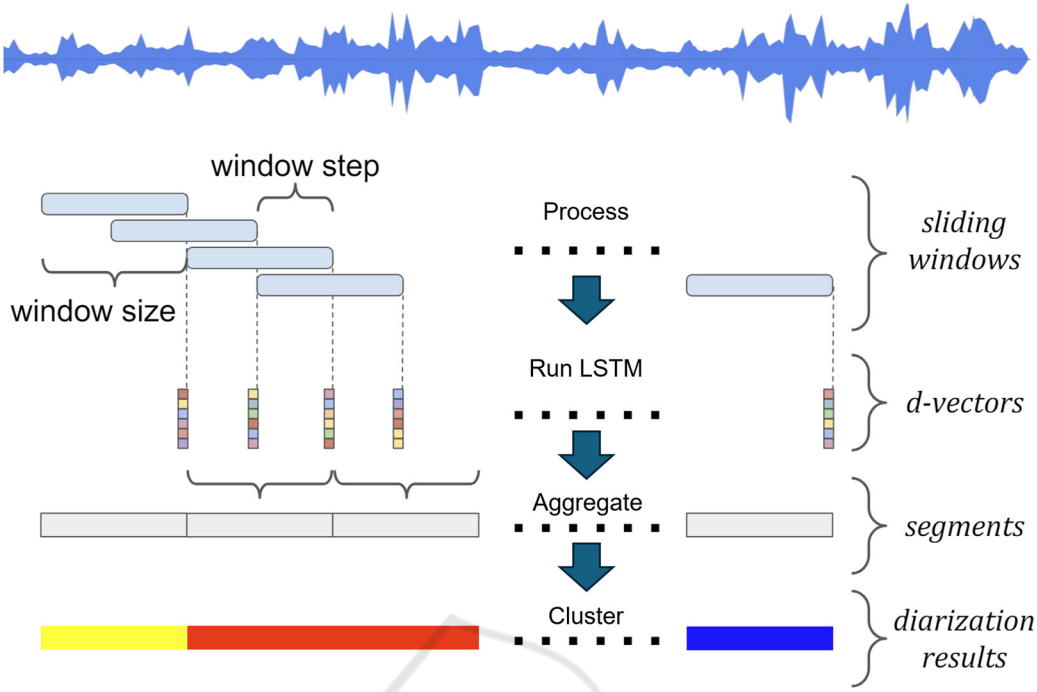

Figure 6: Flowchart of d-vector based diarization system, adapted from Wang et al. (2017).

The fine-tune process included:

- **Spectral clustering:** The percentile *p* for the Row-wise Thresholding and the *sigma* value for the Gaussian Blur refinement operation.
- **DBSCAN:** The *epsilon* represents the maximum distance between two samples for one to be considered in the other's neighborhood and the minimum number of samples in a sample to be considered a core point.
- **Agglomerative Clustering:** The criteria for cluster union and the metric used to compute this union.
- **SOM + K-Means:** Parameters such as algorithm initialization, training type, neighborhood type, and grid map format.

4.2 The Transcriptor Architecture

The architecture of the transcription system is shown in Figure 8, presenting the technologies used, interacting with the Database Management System (DBMS), and the system interface. The proposed transcription system continuously monitors the transcription, besides being responsible for collecting the metadata generated.

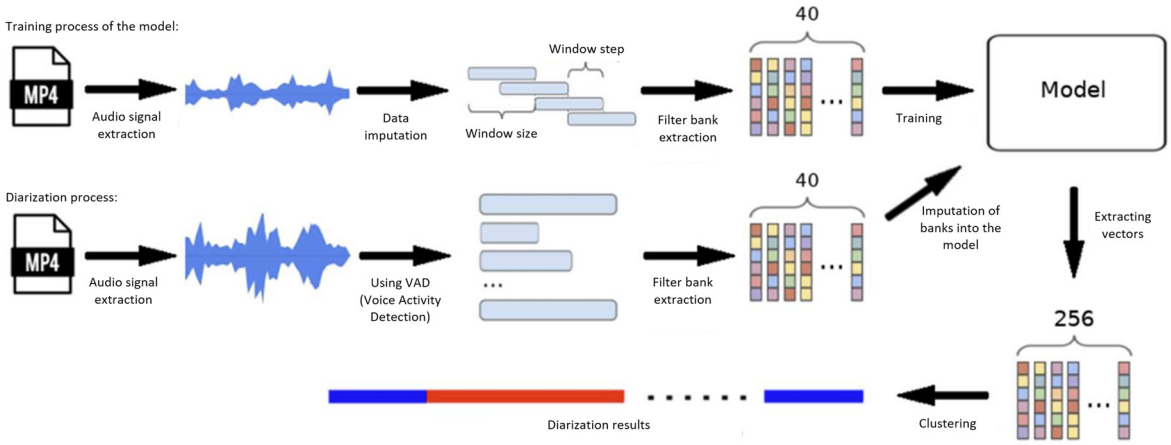

Figure 7: Flowchart of all steps of diarization system development.

The transcription begins when a user uploads new hearing audio/video to the system. The Transcriptor System detects the new file, extracts the audio, and uploads it to Google Cloud Storage. It then uses the Google Speech-to-Text API to transcribe the file and process the transcription results, storing them in the Transcriptor Data database.

Next, the Transcriptor System applies the Speaker Diarization Algorithm to identify the speaker for each transcribed word, storing the resulting metadata in the Transcriptor Data database. This database indexes the text and includes metadata such as the hearing identifier, the timestamp of each word, and the confidence level of the transcription.

The Transcriptor Interface provides user-friendly access to the transcriptions. Users can view the media (audio or video) with the transcription as subtitles, read the full transcription of the hearing, and perform searches within the text.

We highlight that Google's cloud for transcription through Google Speech API requires uploading audio files that are longer than one minute, which differs from the other two transcription methods, which are processed locally.

5 THE TRANSCRIPTOR IN USE

Developed as a legal system module in Brazil, the system stores digital lawsuit processes, stakeholders, and resolutions. Users can upload hearings in video format to obtain transcriptions. Users manually insert the description, date, participants, and media files on the upload screen. If users upload multiple media files for the same hearing, they can combine them into a single video. Users can submit videos from their computer or the Brazilian Judicial Court system.

After the file is uploaded, the transcription processes start automatically, running in the background, and take a few minutes. Depending on the video's size, the user can monitor the transcription status (in progress, error, or complete). The user can view the transcription after the processing step completion, as seen in Figure 9 (left).

The Visualize Hearing screen has the transcript of the selected hearing and the corresponding media side-by-side. The system presents the transcription as a caption attached to the video. The user can click on the word and be directed to the exact moment of the video when it was said. Also, by selecting editing mode, the alternative words defined in the transcript are displayed next to the option to correct the word when clicking on a word. The user can improve the transcription by selecting a present alternative or editing it as flowing text. The system also interpolates the metadata associated with each new word to compose a new transcription.

The user can improve speaker identification and add bookmarks to the transcript. Both features are accessible by selecting a word or sentence and the corresponding checkbox, as shown in Figure 9 (right).

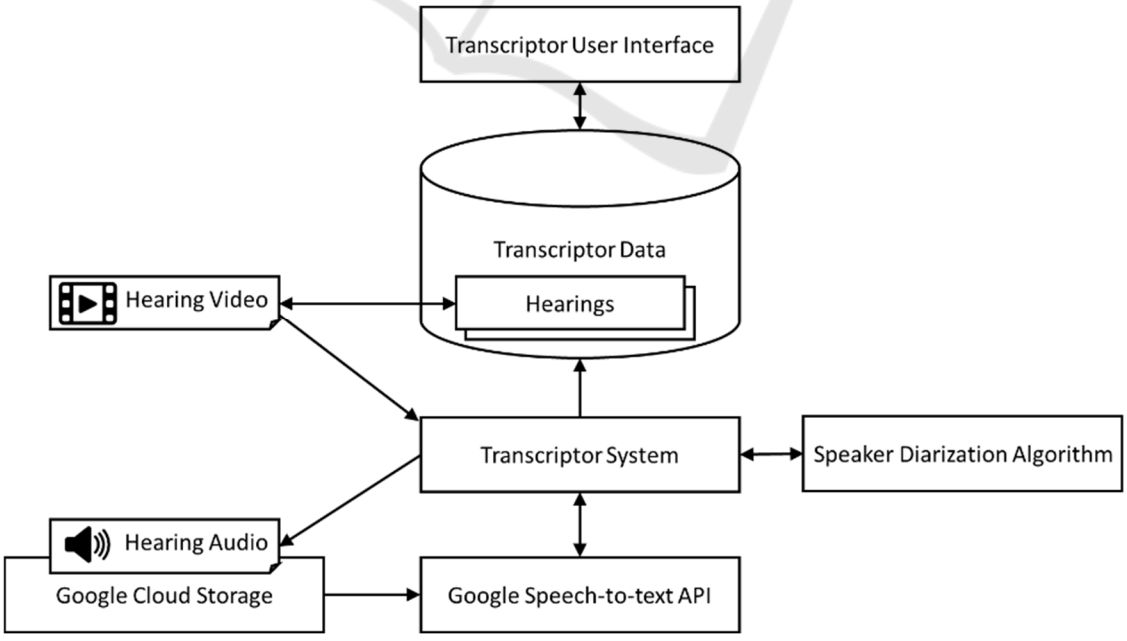

Figure 8: Transcription system of court hearings architecture.

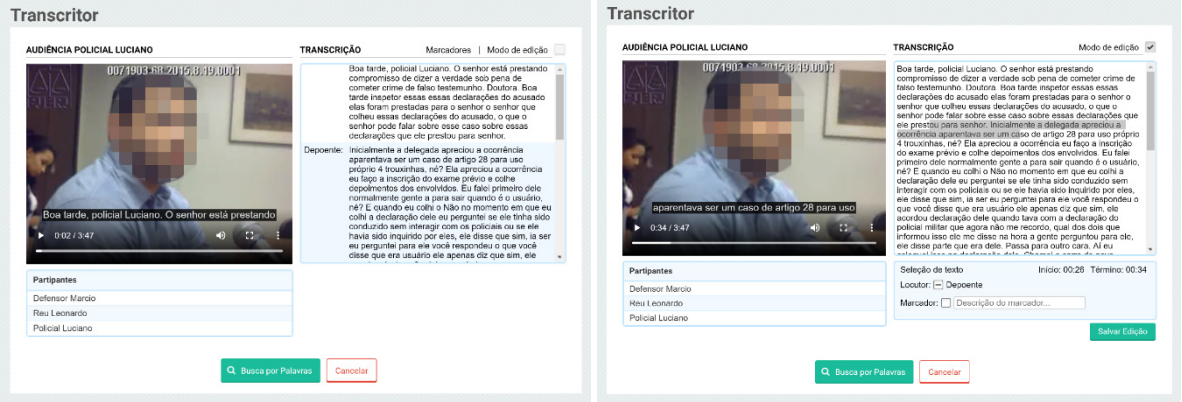

Figure 9: The visualization (left) and the editing mode (right) of a public Hearing transcribed.

Another feature available in the system is the word search. With it, the user can perform a textual search for any word present in the hearings transcribed by the system. The search can be general or filtered by a trial process.

The result lists the hearings in order, including the context in which they appear—predecessor and successor words—and the minute when the word was said in the media. The user can click on the "View full transcript" link to return to the main screen, and the system will present the complete transcript of the audience in question from the beginning. The user can also click the "View transcript at the specified time of the word" link to return to the main screen. The system will present the media and the complete transcript of the audience, starting a few seconds before the word searched is spoken. If the user searches for an alternative word, the result will display the description of an alternative word and the word transcribed by the transcription system. This functionality can facilitate the search and understanding of a hearing, for example, enabling a user to return quickly in the specific minute a sentence was given.

6 EXPERIMENTS AND RESULTS

Although the evaluation included three transcription methods, the implementation used the Google Cloud Speech-to-Text API (Google, 2020a) to provide a lower error rate and meet the Portuguese language without needing modifications. Moreover, the speech recognition technology provided by Google is continuously improving. The word recognition error rate in the company's last disclosure was 4.9% (Pichai, 2017).

Word recognition supported by Google machine learning achieved 95% accuracy in May 2017 for English. This rate also limits human accuracy (Meeker, 2017).

In tests performed using Google's API, the average confidence rate observed was ~ 0.92 for audios extracted from real audiences with good recording quality. The API itself provides this confidence rate for each transcription excerpt performed and assigned to each word corresponding to the excerpt.

The API provides more than one transcription alternative for certain words. Thus, corrections can be made based on the user's feedback through the system interface, using the alternatives provided or manual insertions. A list of other options must be stored in the DBMS with the corresponding section's indexation to return in a search for words or when requested in the system interface. Thus, the analysis of all knowledge obtained in the hearings can return the desired term, even if the transcription mechanism has not effectively transcribed that term. The beta version of the API also offers automatic scoring features (Google, 2020b) to improve the transcription result.

In the data transcription, the first issue was the language: Portuguese. Most algorithms are fine-tuned for English usage, and Portuguese models have demonstrated lower precision. We solved that issue by changing our transcription method to the Google Cloud Speech-to-Text API. Even so, Google's API presents issues: besides being a paid product, it demands uploading files in more than one minute to Google Cloud Storage – which brings the discussion about data security and privacy. We experimented with dividing the files into one-minute chunks to keep files locally. However, the splitting algorithm cut the audio abruptly, often cutting words, which impacted

the transcription quality. Developing a smarter audio splitter was out of the question due to its complexity.

In speaker identification, before we develop our diarization model, we test the beta functionality of diarization of Google's Speak-to-Text API (Google, 2020a), which labels different voices identified in the audios by IDs. Voice training or prior registration is unnecessary since it does not recognize the person but changes the speaker or returns to the same. However, this functionality of Google's API needs to estimate how many people speak in a hearing to process the audio correctly. Using the incorrect value leads to errors in speaker identification. We do not have such data to provide to the algorithm.

We research other frameworks for speaker identification, such as MARF (Mokhov, 2008) and Microsoft Azure's Speaker Identification API (Microsoft Azure, 2020), which are based on neural networks to identify voice patterns. For this reason, these frameworks require a pre-trained set of possible speakers, and such algorithms associate the identified voice with the closest registered one. These algorithms cannot handle unknown voices; thus, every speaker must register in the system, and their voice must be trained previously.

In diarization, we focused our analysis on the vectors created by the extraction techniques and the performance of the implemented approaches using the diarization metric.

Initially, a dimensionality reduction was performed on the vectors generated by both chosen approaches to only two principal components, enabling their visualization in two dimensions to show the distance between each vector visually. Figure 10 (left) shows a close cluster in the D-vectors generated, making efficient clustering unfeasible. Conversely, the I-vectors – Figure 10 (right) – are more evenly distributed and farther apart. We highlight that the graphs shown in Figure 10 have different scales.

The Diarization Error Rate (DER) represents the fraction of total diarization time incorrectly attributed, whether due to False Alarms, Misses, Speaker Overlaps, or Confusion (Ryant et al., 2019).

$$DER=\frac{\text{False Alarm} + \text{Miss} + \text{Overlap} + \text{Confusion}}{\text{Time}} \quad (1)$$

Table 1 illustrates the performance of each implemented approach on the provided dataset. We also compare it with the previously obtained result using only speaker video images.

Table 1: Performance of each approach on the dataset.

Model	GMM	LSTM
Speaker video	32.5%	32.5%
Spectral Clustering	**21.79%**	80.90%
DBSCAN	57.24%	72.51%
Agglomerative	28.49%	84.47%
SOM + K-Means	65.39%	**67.85%**

As observed in Table 1, significant results were not obtained with the LSTM, suggesting that the model may not be learning the D-vectors correctly. As a neural network, LSTM is a complex model and more challenging to train, which may explain the difficulty in achieving significant results. On the other hand, the GMM generated I-vectors with substantial differences from each other, enabling efficient clustering and achieving a DER of 21.79% using Spectral Clustering. This result represents a 32.95% improvement over the previously designed model using only video images.

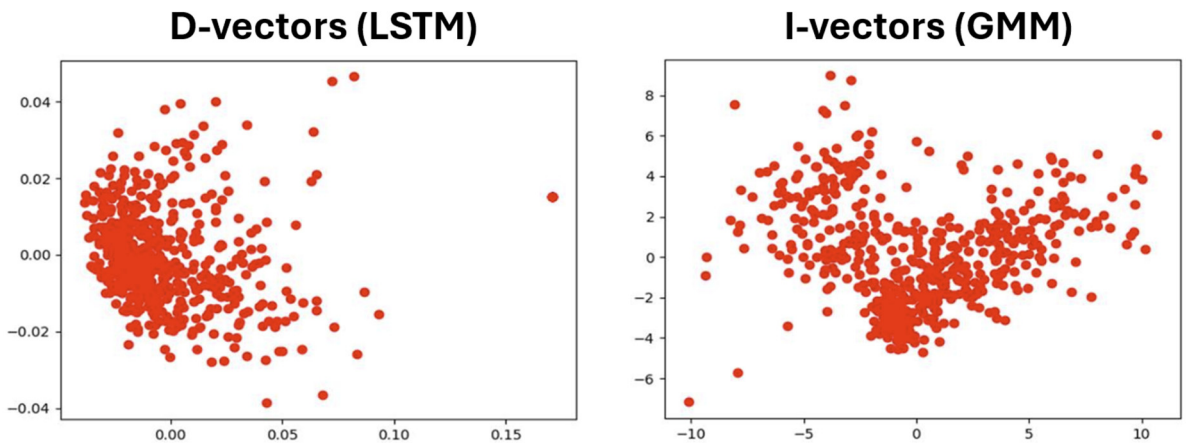

Figure 10: Dimensionality reduction of the D-vectors (LSTM) (left) and I-vectors (GMM) (right).

7 CONCLUSIONS

Legal proceedings involve conducting several testimonies and hearings. When performed, the transcriptions of these media are part of an expensive and slow process done manually by specialized people. Thus, few processes have their media records transcribed. In this work, an automatic transcription system was proposed, capable of organizing the media content generated in legal depositions/hearings based on a knowledge management model. The system allows collaboration to refine the automated process to achieve higher accuracy.

The system underwent an evaluation by legal members and proved to be quite helpful for its proposed use. It is currently in use in the Defense Office of Rio de Janeiro, with almost one hundred hours of court hearings transcribed. It can significantly contribute to the progress of legal processes and their decision-making. From a technical point of view, the transcripts had an average hit rate of 92% for audio extracted from real audiences with good sound quality. An automatic transcription system significantly contributes to the legal scope, streamlines the decision-making process, and maintains a legal memory that can be analyzed more quickly and efficiently later.

The limitations of this work are due to the existing technology for automatic transcriptions, which, for Portuguese, is still limited, in addition to features in Beta versions, such as automatic punctuation and announcer diarization. The latter contributes significantly to distinguishing interleaved speeches but still does not solve the problem of complete identification of the user, which implies voice training previously registered in the system to work for any user present in the hearings. This question is not viable for the type and quantity of users who participate in Brazilian legal processes. The quality of the generated media files also significantly impacts the result of the transcription.

In conclusion, the advancement of technology, such as better recording media quality and the Google API improvements for Portuguese, will enable a higher success rate in the future and the implementation of features still in beta. We highlight that this system must stay consistent and integrated into the legal system to effectively use and provide the desired legal memory easily accessible to law workers.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv:1603.04467 [Cs]*. http://arxiv.org/abs/1603.04467

Alavi, M., & Leidner, D. E. (2001). Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. *MIS Quarterly*, *25*(1), 107–136. https://doi.org/10.2307/3250961

Alegbeleye, B. (2010). Old wine in new bottle: A critical analysis of the relationship between knowledge and library and information science. *Nigeria Library Association National Conference and AGM*, 18–23.

Arnold, K., Gosling, J., & Holmes, D. (2005). *The Java programming language*. Addison Wesley Professional.

Brasil, S. F. do. (1988). *Constituição da República Federativa do Brasil*. Senado Federal, Centro Gráfico.

Carlsson, S., Sawy, O. E., Eriksson, I., & Raven, A. (1996). *Gaining Competitive Advantage Through Shared Knowledge Creation: In Search of a New Design Theory for Strategic Information Systems.* 1067–1076.

Chiu, P., Boreczky, J., Girgensohn, A., & Kimber, D. (2001). *LiteMinutes: An Internet-based system for multimedia meeting minutes*. 140–149. https://doi.org/10.1145/371920.371971

Crowston, K. (2012). Amazon Mechanical Turk: A Research Tool for Organizations and Information Systems Scholars. In A. Bhattacherjee & B. Fitzgerald (Eds.), *Shaping the Future of ICT Research. Methods and Approaches* (Vol. 389, pp. 210–221). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-35142-6_14

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788–798. https://doi.org/10.1109/TASL.2010.2064307

Dhamdhere, S. N. (2015). Knowledge Management Strategies and Process in Traditional Colleges: A Study. *International Journal of Information Library and Society*, *4*(1), Article 1.

Electronic discovery. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Electronic_discovery&oldid=969855680

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Knowledge Discovery and Data Mining*. https://api.semanticscholar.org/CorpusID:355163

Gomes, T. C., & Furtado, B. S. (2017, July 22). *A Justiça deve transcrever audiência de caso complexo*. Consultor Jurídico. http://www.conjur.com.br/2017-jul-22/opiniao-justica-transcrever-audiencia-complexo

Google. (2020a). *Cloud Speech-to-Text Documentation*. Cloud Speech-to-Text API. https://cloud.google.com/speech-to-text/docs/

Google. (2020b). *Getting automatic punctuation*. Cloud Speech-to-Text API. https://cloud.google.com/speech-to-text/docs/automatic-punctuation

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The elements of statistical learning: Data mining, inference, and prediction. In *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed, pp. 520–528). Springer.

Huber, G. P. (1991). Organizational Learning: The Contributing Processes and the Literatures. *Organization Science*, *2*(1), 88–115. https://doi.org/10.1287/orsc.2.1.88

Huet, G. (2006). *Design transaction monitoring: Understanding design reviews for extended knowledge capture* [PhD Thesis]. University of Bath.

Huggins-Daines, D., Kumar, M., Chan, A., Black, A. W., Ravishankar, M., & Rudnicky, A. I. (2006). Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices. *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, *1*, I-185-I–188. https://doi.org/10.1109/ICASSP.2006.1659988

Integrated Electronic Litigation System. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Integrated_Electronic_Litigation_System&oldid=935231346

King, D. E. (2009). Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research*, *10*, 1755–1758.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480. https://doi.org/10.1109/5.58325

MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations*. https://api.semanticscholar.org/CorpusID:6278891

McQueen, R. (1998). *Four views of knowledge and knowledge management.* 609-11.76.

Meeker, M. (2017, May 31). *Internet Trends 2017 Report*. 22nd edition of the Internet Trends Report at the Code Conference, Rancho Palos Verdes, California. https://www.slideshare.net/kleinerperkins/internet-trends-2017-report

Microsoft Azure. (2020). *Speaker Recognition | Microsoft Azure*. https://azure.microsoft.com/en-us/services/cognitive-services/speaker-recognition/

Mokhov, S. A. (2008). Introducing MARF: A Modular Audio Recognition Framework and its Applications for Scientific and Software Engineering Research. In T. Sobh (Ed.), *Advances in Computer and Information Sciences and Engineering* (pp. 473–478). Springer Netherlands. https://doi.org/10.1007/978-1-4020-8741-7_84

Motta, R., Barbosa, C. E., Lyra, A., Oliveira, J., Zimbrão, G., & De Souza, J. M. (2022). Extracting Knowledge from and for Meetings. *2022 12th International Conference on Software Technology and Engineering (ICSTE)*, 82–90. https://doi.org/10.1109/ICSTE57415.2022.00019

Mutula, S., & Mooko, N. (2008). *Knowldge Management IN Information and Knowldge Management in the digital age:concepts, Technologies and African Perspectives*.

Nadaraj, P., & Odayappan, S. (2020). *Digital Courts: Are We Really Availing Infinite Possibilities Of Technology?* https://www.outlookindia.com/website/story/opinion-digital-courts-are-we-really-availing-infinite-possibilities-of-technology/351800

Nonaka, I. (1994). A Dynamic Theory of Organizational Knowledge Creation. *Organization Science*, *5*(1), 14–37.

OMG. (2011). *Business Process Model and Notation (BPMN), Version 2.0* (p. 538) [Technical report]. Object Management Group.

Parent, G., & Eskenazi, M. (2010). Toward better crowdsourced transcription: Transcription of a year of the Let's Go Bus Information System data. *2010 IEEE Spoken Language Technology Workshop*, 312–317. https://doi.org/10.1109/SLT.2010.5700870

Patel, D. S. (2012). *Management Discovery Of Knowledge Management In*.

Pichai, S. (2017). *Google I/O*. Google's Annual Developer Conference, Montain View, CA.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motl´ıcek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. http://kaldi-asr.org/doc/

ProJuris. (2024). *ProJuris- Software para Escritórios de Advocacia de alto desempenho*. ProJuris. https://www.projuris.com.br

Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., & Liberman, M. (2019). *The Second DIHARD Diarization Challenge: Dataset, task, and baselines*. https://doi.org/10.48550/ARXIV.1906.07839

Stollenwerk, M. de F. L. (2001). Gestão do conhecimento: Conceitos e modelos. In K. Tarapanoff, *(Org.) Inteligência organizacional e competitiva* (pp. 143–163).

Tomar, S. (2006). Converting video formats with FFmpeg. *Linux Journal*, *2006*(146), 10.

Wang, Q., Downey, C., Wan, L., Mansfield, P. A., & Moreno, I. L. (2017). *Speaker Diarization with LSTM*. https://doi.org/10.48550/ARXIV.1710.10468

Williams, J. D., Melamed, I. D., Alonso, T., Hollister, B., & Wilpon, J. (2011). *Crowd-sourcing for difficult transcription of speech*. 535–540. http://ieeexplore.ieee.org/document/6163988/

Zack, M. (1998, September). *An architecture for managing explicated knowledge*.