SpeechVis: Simplifying Speech Emotion Visualization

Luan Dopke
Arthur Accorsi
João Paulo Aires
luan.dopke@edu.pucrs.br
School of Technology
Pontifical Catholic University
of Rio Grande do Sul (PUCRS)
Porto Alegre, Brasil

Larissa Guder
larissa.guder@edu.pucrs.br
School of Technology
Pontifical Catholic University
of Rio Grande do Sul (PUCRS)
Porto Alegre, Brasil
Laboratory of Advanced Research on
Cloud Computing
Faculdade Três de Maio
Três de Maio, Brasil

Isabel Harb Manssour
Dalvan Griebler
isabel.manssour@pucrs.br
dalvan.griebler@pucrs.br
School of Technology
Pontifical Catholic University
of Rio Grande do Sul (PUCRS)
Porto Alegre, Brasil

ABSTRACT

As the amount of online content increases, analyzing and following discussions becomes harder. Relevant information, such as the main discussion topics and the emotions expressed in audio, e.g., in a podcast, requires people to watch or listen to the entire content to understand the context. However, this can take a long time, and people's interpretations of emotions can bias their understanding of them. A visual summarization of such information can help people quickly understand the audio context and analyze the content regarding speakers, their emotions, and the main topics covered. In this work, we introduce SpeechVis, a visual analytics tool that visually summarizes speech emotions from an audio source. SpeechVis extracts multiple information from the audio, such as the transcription, speakers, main topics, and emotions, to provide visualizations and statistics about the discussed topics and each speaker's emotions. We used multiple off-the-shelf machine learning models to extract audio information and developed several visual representations that aim to facilitate audio analysis. To evaluate SpeechVis, we selected two use cases and performed an analysis to demonstrate how the SpeechVis visualizations can give valuable insights and facilitate audio interpretation.

KEYWORDS

Visual Analytics, Speech Visualization, Emotion Classification, Signal Processing, Machine Learning

1 INTRODUCTION

The perception and understanding of emotions will vary according to how each person was taught to understand them. The expression also depends on each individual's idiosyncrasies. According to [30], emotion can be defined as a unified succession of feelings. Even a succession of feelings is never a permanent state. Typically, the expression of emotions happens instinctively and unconsciously, detectable through facial and bodily cues, vocal tone, pupil dilation, heart rate, and breathing [22]. Emotions may be interpreted diversely across cultures, yet cross-cultural understanding remains feasible [15].

For a computer to understand human emotions, we need to convert these emotional characteristics into a mathematical representation. This way, a machine can comprehend them. In our bibliography research, we found two main approaches to represent emotions computationally. The first one categorizes human emotions into six classes. These classes define six base emotions: anger, disgust, fear, happiness, sadness, and neutral [8]. The second approach refers to dimensionally representing emotions. Russell [24] proposed the circumplex model to represent emotions in a dimensional space, with two axes representing values for arousal (y-axis) and valence (x-axis). In the circumplex model, high valence and arousal values result in emotions characterized by happiness and excitement. Conversely, emotions such as contentment and tranquility emerge when valence is low and arousal is high. Emotions associated with low valence and arousal include sadness, fatigue, and tedium, whereas high valence and low arousal typically lead to feelings of frustration, anger, and fear. Complementing the Circumplex model, Mehrabian [17] uses a third dimension called dominance. Dominance focuses on the behavior. Lower levels reflect passivity, while higher levels indicate assertiveness.

A wide range of research studies can extract emotion from a transcribed conversation. However, when we look only at textual data, we ignore substantial information from recorded audio that can represent and distinguish emotions, such as pitch and vocal tone. Considering these aspects, it is important to analyze not only textual data but also audio recordings that can potentially reveal information about what the speakers are feeling.

Artificial intelligence, particularly machine learning, has been gaining prominence due to its effective solutions for various problems. Machine learning techniques make it possible to process audio recordings to extract valuable information related to dialogue, such as transcription, speakers, main topics, and emotions. These elements can then be analyzed and compared through an interactive visualization, providing a comprehensive understanding of the audio content.

Utilizing visualization techniques to build a visual analytics tool can significantly enhance comprehension of audio content, allowing individuals to swiftly grasp its context and analyze aspects, such as speakers, their emotions, and the primary topics discussed. A tool capable of extracting, performing analysis, and generating visualizations from audio information can help in several domains. For instance, in a call center environment, a solution like this can

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

quickly pinpoint the reasons behind a negative review, disregarding the need to listen to the entire audio to identify the root cause of the issue. This can facilitate ongoing improvements to the end-user experience. Another use case is a debate analysis, a rich source of information where individuals express a wide range of emotions and discuss diverse topics. Analyzing audio records allows a deeper understanding of the dynamics at play, including how emotions influence discourse and which topics elicit strong reactions. Moreover, by analyzing the speaker's emotions, researchers can uncover effective persuasion techniques used in debates and also identify the key issues under discussion. Also, representing visually the emotions presented in a speech can help deaf persons understand the information better.

With this objective in mind, we developed SpeechVis, a visual analytics tool designed to analyze speakers' discussions from an audio source comprehensively. SpeechVis highlights key topics, emotions, speaker behavior, and temporal patterns within the discourse, providing valuable insights into the dynamics of the conversation. We gather the data to visualize using off-the-shelf machine learning models capable of extracting information from audio recordings.

Our main contributions are:

- **Creation of a modular Machine Learning (ML) pipeline for audio processing.** We extensively analyzed off-the-shelf machine learning models helpful in extracting information for building the visualization tool. Then, we assembled an adaptable pipeline that follows a set of steps to transcribe, diarize, and classify the audio recording, and supports future model exchange to improve performance or accuracy.
- **Development of a Gantt chart for audio temporal analysis.** We designed and customized a Gantt Chart for representing the temporal dynamics of speech, enabling users to analyze the entire audio recording visually, follow who is speaking, what topics are being discussed, and which emotions are expressed throughout the recording.
- **Provision of an interactive visual analytic tool for speech analysis.** We developed a visual analytics tool that allows users to upload and analyze audio recordings by leveraging multiple coordinated views. A set of graphical representations permits exploring the emotional profile of each speaker and the main discussion topics in an audio conversation.

The next sections present the following contents: (2) Related Work, discussing existing work done in this research field; (3) Methodology, explaining the goals of our tool and describing our methodology to collect and extract our data; (4) Visualization Design, explaining the visualization tool and correlating with the goals; (5) Study Case, implementing the proposed tool in a real case scenario; (6) Evaluation using data extracted from IEMOCAP dataset; and (7) Conclusion, overviewing this work and setting new directions to follow.

2 RELATED WORK

This work combines three research fields: Speech Emotion Recognition (SER), Data Visualization, and Audio Processing. In this section, we provide an overview of the existing literature related to our work. Table 1 summarizes each work.

The E-ffective visualization proposed by Maher *et al.* [16] aims to discover what makes speeches effective. The proposed system utilizes facial, textual, and vocal data related to effectiveness and supports a rapid understanding of critical factors in inspirational speeches, such as the influence of emotions. E-ffective uses the timestamp of utterance to split the audio and applies an open-source toolbox [6] for emotional analysis to obtain each recording's valence and activation values. With the aid of these data, E-ffective creates two novel visualizations for emotions, the e-spiral and the e-script. The latter allows critical information about the text to be compared across whole speeches. e-spiral considers valence and activation values, using the direction of a spiral to show positive or negative valence and dot size to portray high or low activation, as well as colors to represent each emotion. E-ffective shows emotions per utterance and pauses in speech - since their goal is to analyze the effectiveness of speeches, it is essential information.

The PEARL visual analytic tool [33] aims to explore and examine a person's emotional style derived from this person's social media text. PEARL extracts emotions from Twitter social media to build the visualization; PEARL identifies eight emotion classes (anger, fear, anticipation, surprise, joy, sadness, trust, and disgust), but it also predicts emotion dimensions from the circumplex model (VAD): valence (degree of positiveness), arousal (degree of excitement), and dominance (degree of aggressiveness). The authors employ a lexicon-based approach to extract emotion data, the NRC lexicon [18] for emotion classes and ANEW lexicon [29] for VAD. Similar to our work, they are careful with detecting what triggers determined emotion over time; for this purpose, they summarize the tweet's content containing an emotional segment to approximate the cause or trigger related to a mood.

A lexicon-based approach is also applied on the built of EmotionWatch [13], a tool that constructs visual summaries of public emotions. This work uses the OlympLex emotion lexicon [26] to classify public event tweets into eight emotion classes used to build an emotion wheel and make a timeline for visualizing the emotion flow during the public event.

EmoCo, the visual analytics system proposed by Zeng *et al.* [32], uses facial, text, and audio data extracted from presentation videos to explore the coherence of multimodal emotions and understand emotional expressions in presentations. EmoCo uses audio features as one of the channels to identify emotions. To extract this emotional data from the audio, they split the audio into different segments and filtered out laughter. Then, they compute the Mel Frequency Cepstral Coefficient (MFCC) and feed it to a classification model [3], giving one of the seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. EmoCo has three different channels of emotion identification, and they need to display each one of them. They assign colors to different emotions and create multiple techniques to show the coherence between every channel, such as utterance clustering and a Sankey diagram. EmoCo only shows the emotion identified in each utterance. Also, they use a histogram to show the pitch, intensity, and amplitude distribution for different emotions.

Focusing on accessibility for deaf and hard-of-hearing users, Pataca *et al.* [21] proposed an emotion visualization approach for closed captions. Four different approaches were tested: (1) a conventional captioning system, (2) only prosody information, (3) dimensional

Table 1: Summary of related works

Work	Extraction	Data Type	Emotion	Objective
[16]	Toolbox	Facial, Textual, Audio	Dimensions	Supports a rapid understanding of critical factors in inspirational speeches, such as the influence of emotions
[33]	Lexicon	Textual	Dimensions, Classes	Explore and examine a person's emotional style derived from this person's social media text
[13]	Lexicon	Textual	Classes	Constructs visual summaries of public emotions
[32]	Model	Facial, Textual, Audio	Classes	Explore the coherence of multimodal emotions and understand emotional expressions in presentations
[21]	Model	Audio	Dimensions	Emotion visualization approach for closed captions
SpeechVis	Model	Audio	Dimensions, Classes	Enhance comprehension of audio content, allowing to swiftly grasp its context and analyze several aspects

emotion recognition, and (4) a combination of prosody and emotions. Considering the four approaches, according to the author's evaluation, the best performance was emotion-based. To predict valence and arousal values, a Transformer based model [28] is used. Emotion-based visualization consists of variations in font color and size. The determination of the font color is based on the valence dimension, where the scale ranges from red for negative to green for positive values. Otherwise, the font size is determined according to the arousal dimension.

SpeechVis, our Visual Analytics Tool, differentiates from the related work by proposing a pipeline of off-the-shelf machine learning models fed by a single audio file format that allows extracting all the required data for visualization from any audio recording. Moreover, the work adapts the Gantt chart to better describe the emotional characteristics of a recorded conversation and analyzes the emotions from two different approaches: emotion as dimensions and emotion as classes, performing an emotional analysis from different perspectives.

3 SYSTEM DESIGN

After defining our research topic, we thoroughly reviewed current literature to identify gaps and off-the-shelf machine learning models that extract the needed data to build SpeechVis. This process enabled us to identify datasets for our study and choose the models for our pipeline. Additionally, the literature review aided in establishing clear goals to drive our research and develop a strong, logical, and coherent visual analytics tool. We identified four distinct goals (G):

- **G1: Identify and display the behavior of multiple speakers.** It is necessary to identify how many people speak in the input audio. Every speaker will exhibit different emotions, we need to be able to assign them to each one and display them.
- **G2: Identify and display temporal characteristics of the conversation.** A conversation has a natural flow. An emotion might manifest when a specific topic is brought up or after someone speaks. We want to showcase these characteristics.
- **G3: Identify and display the main topics of the conversation.** Conversations may have multiple topics. We want to display them so it is easier to see what the speakers are saying without having to listen to it.
- **G4: Connect topics with speaker emotions.** Some topics are more sensitive and delicate, and can express different emotions among speakers. We aim to detect how each speaker deals with each topic.

The literature demonstrates that establishing a validation dataset is important to verify the selected models. We used IEMOCAP [7] to validate our visualization. It is a publicly available dataset comprising approximately 12 hours of audio data labeled by multiple annotators. Each speaker's dialogue is accompanied by transcriptions and timestamps. IEMOCAP uses two approaches for emotional classification: discrete categorical-based annotations and dimensional attribute-based annotations. The emotional categories include anger, happiness, excitement, sadness, frustration, fear, surprise, other, and neutral states.

We must collect data in multiple forms to reach our goals, such as the number of speakers, speech duration, conversation content, and emotions. We will present the necessary data processing, library, and pre-trained models in the next sections. The input data of SpeechVis is an audio file in *.wav* format. Each subsequent step is necessary to extract information from it. These pipeline steps are explained in Figure 1. All data generated by the pipeline model's outputs and pre-processing steps are saved in a Comma-separated values (CSV) file. The visualization tool received this CSV file and the audio file as data input to construct the charts.

3.1 Transcription and Diarization

An audio file is capable of containing multiple sound sources, and some of them are not of interest to our application, such as laughing,

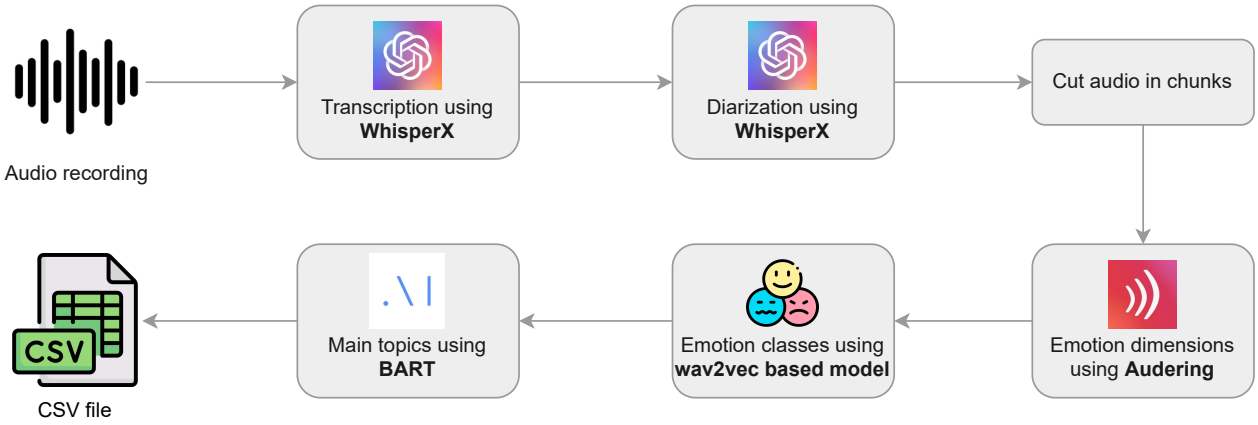

Figure 1: Proposed pipeline for audio processing.

clapping, or coughing. Since we want to estimate an emotion within a person's voice, we can't take into consideration such "noises", as it will affect the outcome. In order to avoid that, we need to create smaller audio files containing only one utterance. It is achieved by using the transcription model, which gives us the beginning and end of each utterance within a speech; it is a very important process because the smaller audio clips are used in the remaining models. Also, a diarization task is used to differentiate each speaker in a conversation. Diarization is the ability to know who said what in a given utterance. We use *WhisperX* [2] for transcription and diarization, which uses a voice activity detection algorithm to identify regions of the audio that contain speech. It also gives an accurate transcription using the original Open AI's Whisper and a phoneme model to force align the data. WhisperX uses *Pyannote.Audio* [5] to give an estimated number of speakers (up to six), as well as which speaker spoke which utterance. We chose WhisperX because of its easy use and fast processing, due to parallel implementation.

3.2 Emotion Recognition

In our approach, we employ different machine learning models for each type of emotion classification. Therefore, two different models are necessary for our application: one for discrete and one for dimensional emotion prediction. The data input for these models is the audio file generated from the beginning and end of the utterance (data taken from WhisperX's segmentation). This way, every utterance has a predicted emotion and dimensional values.

3.2.1 Dimensional-based Emotion. Dimensional-based emotion classification assigns values to valence and activation variables. The valence dimension measures how positive or negative someone feels; negative values represent negative emotions, and positive values represent positive emotions. The activation dimension measures whether an emotion makes someone take action or not, from active to passive [23]. Anger has a low valence and high activation, and relaxation has a high valence and low activation. The model used for the prediction of valence and arousal values for each file is *w2v2-L-robust-12*[27], which is a pre-trained wav2vec 2.0 model fine-tuned on a speech emotional dataset [28]. Wav2vec 2.0 is an automated speech recognition (ASR) model that generates representations directly from raw speech audio, significantly reducing the need for annotated data [1]. The model predicted values are normalized between *[-1;1]* for better fitting in Russell's Circumplex Model.

3.2.2 Categorical Based Emotion. For categorical emotion recognition, we used a model based on *wav2vec 2.0*, but trained on a different dataset [12]. It gives a score for different emotion classes: anger, happiness, sadness, disgust, and fear. Since people do not always express emotions, we add a neutral emotion when the model confidence is below 75%. To reach this threshold value, we performed a manual analysis comparing the model's output with the annotated data from the IEMOCAP dataset.

3.3 Natural Language Processing

Aiming to extract the main topics of each utterance, we assume that every topic is a noun. We performed a series of steps to classify the main topics. First, we used the generated transcription from the audio with WhisperX to perform a part-of-speech (POS) task. We chose *SpaCy* for this task, a widely adopted open-source library for NLP applications [11] due to its ease of integration and fast processing on local environments. The POS tagging enable acquires every noun from each sentence. If the sentence only has one noun, we use that noun as the main topic of that sentence. In the cases where the sentence had more than one noun, we use *Bart* [14], a zero-shot-text-classification model with these nouns as candidate labels and the sentence as a sequence to classify the main topic.

3.4 Loud Voice Detection

In order to evaluate our system in unlabeled audio, we decided to use moments when someone speaks loudly to show which utterance was said with a higher activation [31], providing possible regions of interest in the audio. Louder voices have a high sound intensity [10]. The intensity of the audio is taken from the amplitude of the sound wave. The amplitude of the sound wave of each audio track is calculated from *Librosa*'s Root Mean Square (RMS) amplitude function, and its maximum value is stored in an array. The 4^{th} percentile of the amplitude array is used to detect at which utterance

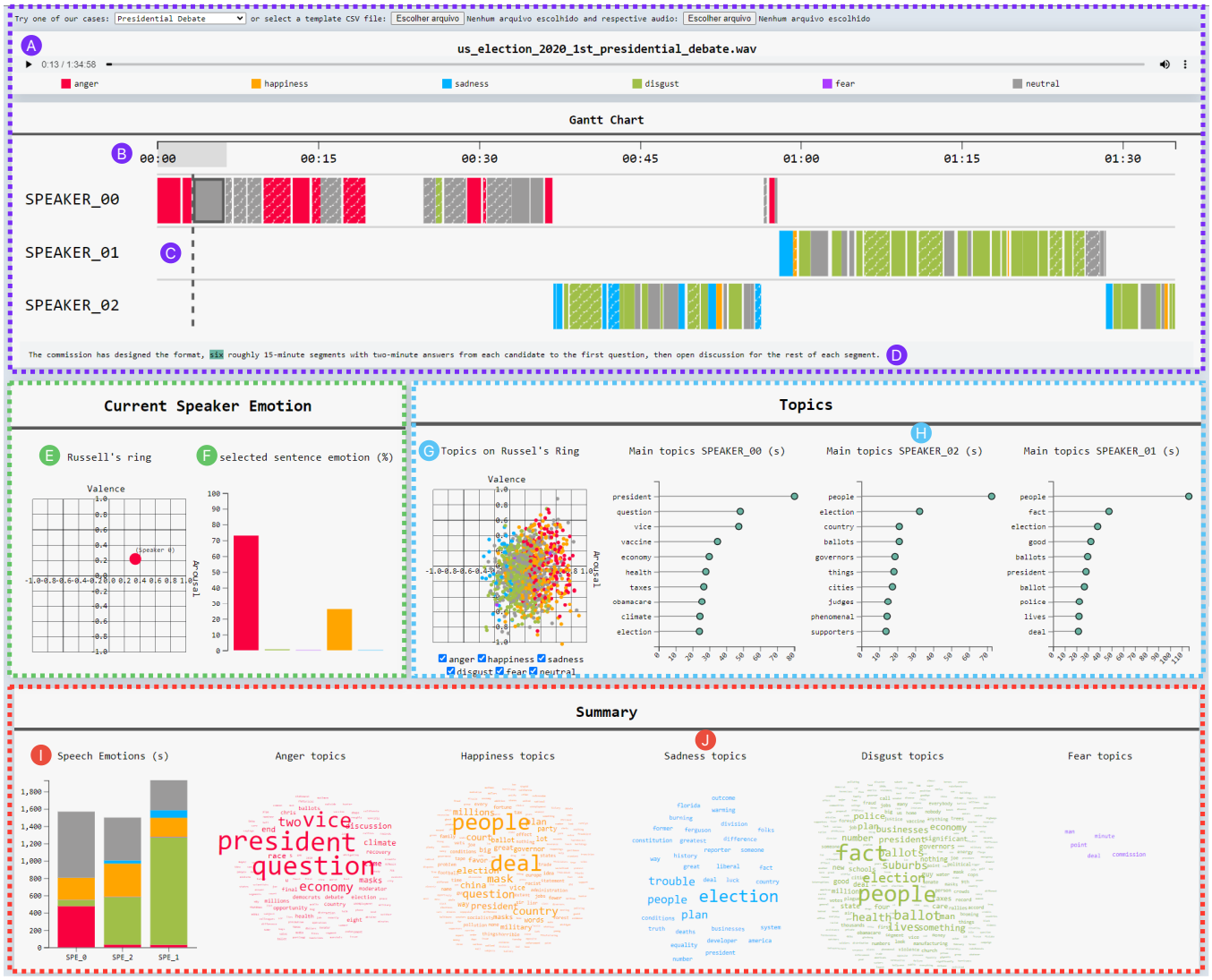

Figure 2: SpeechVis Overview.

there was a louder voice than usual. Since every RMS amplitude used in the process is from an audio segment, the 4th percentile array doesn't include moments when no one is speaking or when there is noise from non-vocal sources in the audio.

4 VISUALIZATION DESIGN

A detailed overview of SpeechVis is available in Figure 2 and is next described. Our visualization tool stands out with its four unique views, each designed to play a specific role in presenting data and understanding emotional information within audio recordings. The *Gantt Chart*, the first view at the top of Figure 2, illustrates the flow of emotions and temporal characteristics throughout the discourse. The *Current Speaker Emotion* view below provides a more precise analysis of emotions within the currently selected utterance. Next to it, the *Topics* view attempts to correlate emotional sentiment with discussed topics, highlighting the most prevalent subjects. Lastly, the *Summary* view provides a comprehensive overview, summarizing the emotions and topics discussed throughout the entire recording.

4.1 Gantt Chart and Current Speaker Emotion views

SpeechVis allows users to upload their own CSV data and audio files, enabling them to analyze their use cases and samples. The tool is responsive, adapting to various screen sizes. We used the open-source D3.js library [4] to create the visualization tool, utilizing only HTML, CSS, and JavaScript.

The main View of SpeechVis enables an analysis of the temporal and emotional characteristics of the recordings; the first component (A) in Figure 2 allows the user to select one of our use cases

or select his/her own CSV data and respective audio recordings. It also has a sound player for the audio and a legend displaying the emotions and their colors used throughout the tool. The tool allows the user to navigate through the conversation using the top component (B), which synchronizes with component (A), allowing the user to select a specific period and navigate along the audio duration. The Gantt Chart (C) follows the component (B) and has the number of lines equal to the number of speakers identified from the recording; here, we can check the period that a speaker talked and their current emotion, and if an utterance is characterized as loud, this information has represented with the hatched pattern. A tooltip is shown with the start and end times of the utterance when we hover the cursor over the component. Component (D) shows a recording transcription, which is changed according to the current audio time and highlights the main utterance topic.

For the Gantt chart, we use Coordinated Multiple Views (CMV) interaction technique [20], in which selections made on this chart directly impact the other charts. Furthermore, we chose to use more traditional charts, such as word clouds, bar and lollipop charts, as they are more appropriate considering the possible users' different experiences, expectations, and visualization literacy [9].

The green highlight in Figure 2 shows the Current Speaker Emotion view. It displays two charts: Russell's Ring (E) and Selected Sentence Emotion (F). These charts are intricately linked with the Gantt view because the graph data is about the current moment selected on the Gantt chart. Russell's Ring shows the emotion according to the dimension of emotion theory, with arousal on the x-axis and valence on the y-axis; the color corresponds to the emotion class with higher confidence provided by the model. On Selected Sentence Emotion, we have the percentage of the emotions of the current utterance selected and rated by the class model.

Both views are essential to accomplishing the (G1) and (G2) goals. The Gantt chart view allows one to identify the number of speakers in the recording, the recording duration, the emotions that these speakers feel during the conversation flow, and if a part of the recording is louder than the average, accomplishing the (G1) goal. We reached the (G2) goal when we analyzed the entire recording displayed by the Gantt View and could detect an emotional change in the flow, and brought up the topic that changed this conversation behavior.

4.2 Topics View

Topics, the next view, can be visualized in the blue highlight in Figure 2, which aims to raise the main topics from the record. The chart Topics on Russel's Ring (G), organizes all the main topics extracted from the utterances into a dimensional vector, with valence and arousal, the same properties that the Chart Russel's Ring on View Current Speaker Emotion. Each dot on this chart is a topic, and its colors refer to the dominant emotion class. We can filter this chart data through the checkboxes for emotion above it and zoom in using the scroll. Also, this chart is very useful for directly comparing how each ML model deals with emotions and getting an overview of the mood of the conversation.

The other charts, Main Topics (H), are built dynamically and displayed according to the speaker's count on the conversation recording, generating one chart for each speaker. Each speaker has their own chart and demonstrates the ten most frequent topics identified in the speech. We use a Lollipop chart to visualize the information. When we click on one of these topics, a menu called About appears. It contains every occurrence in which the topic was mentioned in the conversation. This menu allows the user to navigate precisely where the speakers have commented on the selected topics. We can visualize the menu in Figure 3.

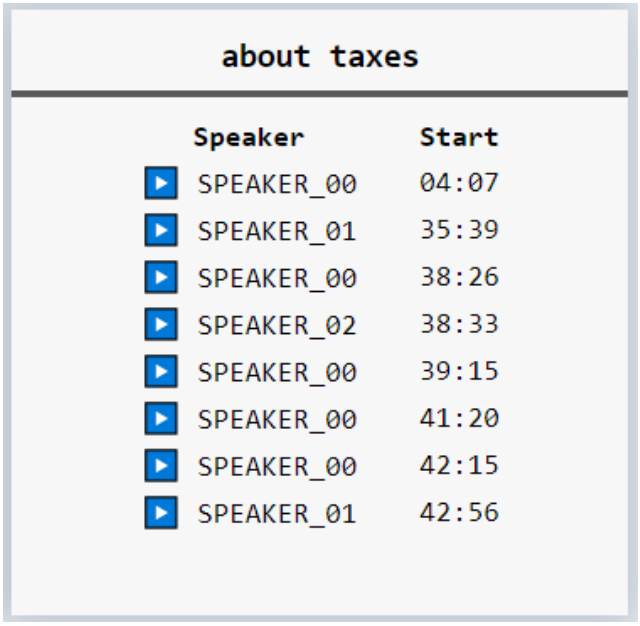

Figure 3: Menu about raised when a topic was clicked.

This View is fundamental to accomplishing (G3) and (G4) goals related to topics and emotions. The Main Topics speakers chart can easily display the most talked-about subjects and identify them in the Gantt chart, achieving the (G3) goal. The topics on Russell's Ring chart can summarize the emotion of the main topics, and the About menu can identify how all the speakers address the same subject, thus achieving the (G4) goal as well.

4.3 Summary View

The last View, called Summary, can be seen in red highlight in Figure 2. This View summarizes all the audio recording data. The first chart, Speech Emotion (I), quantifies how long each speaker feels each emotion; we can discover the predominant emotion for each speaker here. Additionally, a word cloud is provided for each emotion (J), showcasing the main topics discussed while speakers are experiencing that particular emotion, correlating topics with emotions.

This view supports other views by complementing and providing an overview of the recording data and speakers. It aids the goals (G1) through the chart summarizing speakers' information and emotions, and (G4) through the word clouds that show a correlation between topics and emotions.

5 STUDY CASE

Aiming to analyze SpeechVis in different scenarios, we applied the proposed pipeline and extract the data from two used cases: one is a recording of a married couple talk available in IEMOCAP dataset (Ses01M_script02_2.wav file), another is a real case scenario, the first United States of America 2020 presidential debate between Donald Trump and Joe Biden hosted in Cleveland, OH, mediated by Chris Wallace and extracted from Cable-Satellite Public Affairs Network (C-SPAN) YouTube Channel.

5.1 Presidential Debate

On a quick overview in the SpeechVis tool, available in Figure 2, we can easily detect that the debate contains three speakers; one conducts its speech aloud and neutrally, as extracted by analyzing the (C) and (I) charts and the main topics of this speaker are: "president", "question", and "vice", following popular electoral subjects, such as "vaccines", the "economy", and "health", according to (H) charts. This speaker is a debate mediator who constantly questions the other two speakers. The second speaker (SPEAKER_01) spoke much more than the other two and, in its discourse, had a prevalence of disgusting emotions, according to the Speech Emotion chart (I). The presence of disgust emotion in this discourse can be related to socio-moral disgust [25], indicating that the speaker is possibly presenting social or moral boundaries that appear to be violated, centering on human violations of the autonomy and dignity of others. In deep discourse analysis using the Gantt chart, we can check the preoccupation with health, racism, and lives that can be related.

As reported by the (I) chart, the SPEAKER_02 talks less than your opponent, and its discourse is predominantly neutral, happy, and disgusting. The main topics chart (H) shows that subjects discussed in the debate are "COVID-19 pandemic management", "ballots", "vaccines", "climate", "public administration", "police violence", and "public health". The overall topic of emotions during the debate stays with positive arousal (G), which means a heated debate.

The model may classify certain statements made by debate mediators as angry. However, upon closer examination, they are often spoken loudly and transparently, which comes across as neutral. This behavior suggests that the model used to determine the speaker's emotion has learned to associate a loud voice with anger.

5.2 Couple Talk

The SpeechVis tool permits easy detection of the conversation flow between the two speakers, as we can visualize in Figure 4. SPEAKER_01 starts with a positive and neutral tone. After some heated and aggressive talk from SPEAKER_00, SPEAKER_01's tone turns into sadness and disgust. SPEAKER_00 responds with predominantly disgusting emotion, and SPEAKER_01 becomes sad and transitions into a loud and heated talk, indicated by the hashes on (B) point. SPEAKER_01 then displays feelings of disgust and sadness. Also, upon analyzing the emotional flow, we can determine that the question "Are you on your period?" triggers an angry response from SPEAKER_00, as visible on point (A).

Upon analyzing the (C) chart, we discerned that SPEAKER_01 is more vocal and predominantly expresses a range of poignant emotions—sadness, neutrality, and disgust—in their discourse. On the other hand, SPEAKER_00, while speaking less, conveys a more contrasting mix of happiness and anger.

The recording period was only 10 minutes, which may not have been sufficient to gather a comprehensive dataset for emotions, word clouds, and topic analysis. To focus on the discussion of the insights we can extract from analyzing this case study, we have decided to present only the views that contributed most significantly to our debate.

6 EVALUATION

In order to analyze the efficiency of the models used for our visualization, we compared data from IEMOCAP's dataset and data extracted directly from the audio file. We analyzed 110s-long audio recordings from the improvisation section of IEMOCAP. It is a challenging task to compare them since the diarization model and IEMOCAP transcription segment the sentences among speakers differently. The dataset contains 30 sentences, while the diarization model results in 58. Our solution to this problem is to use the transcribed text. Since we have the information of who spoke which sentence, it is possible to evaluate diarization and transcription simultaneously through Jiwer's Match Error Rate [19]. We also compared speech time (ST) by comparing sentence times from each speaker, presented in Table 2. The models substantially detect the minor speech time of the female (F) voice. However, the MER metric is also lower in the female case.

Table 2: Model Evaluation

Speaker	ST from Dataset	ST from Models	MER
M	59s	32s	17.21%
F	69s	62s	26.23%

For emotion analysis, the difference in segmentation creates a more significant problem; to bypass this problem, we use the segmentation pattern from IEMOCAP and feed these sentences to the classification model. The emotion classes in IEMOCAP are not the same as the ones predicted with the model, making it impossible to evaluate them. Valence and arousal values predicted by the model, on the other hand, were both normalized between -1 and 1, matching with IEMOCAP data. We calculated the valence and activation of these model errors using the Mean Squared Error (MSE) and Root Mean Square Error (RMSE), presented in Table 3.

Table 3: Model Evaluation

Dimension	MSE	RMSE
Valence	0.124	0.353
Activation	0.062	0.249
Dominance	0.094	0.307

To evaluate cases from data that are not from datasets and have no annotated data, we compared the values of valence and activation from the audio to the values extracted from text using the NRC-lexicon [18], a well-known text emotion estimator. We applied our visualization data in the 2020 U.S. first presidential debate [1].

[1] https://www.youtube.com/watch?v=wW1lY5jFNcQ

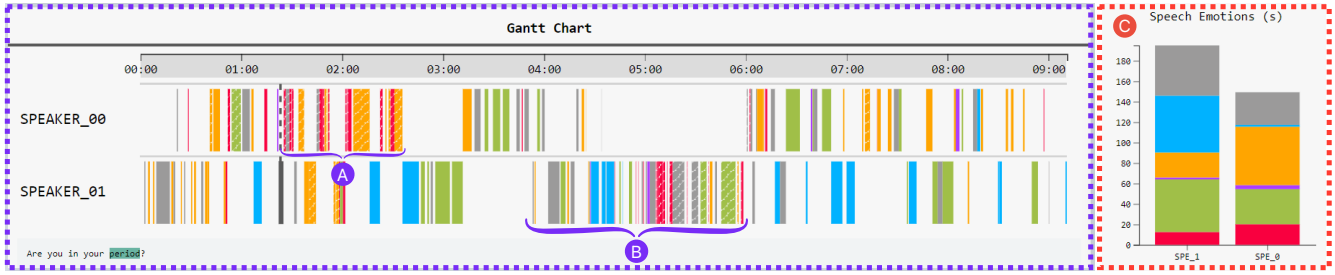

Figure 4: SpeechVis elements from the Couple Talk study case overview.

Since the audio duration of a debate is 1h30min long, it takes a long time to process and get information from it. We limited the duration to the first 40 minutes, around half the debate's duration. The processing time needed to extract information from the audio was 51 minutes, which is longer than the audio recording. The MSE and RMSE calculated metrics can be compared among valence, activation, and dominance values, as shown in Table 4.

Table 4: Model Evaluation

Dimension	MSE	RMSE
Valence	0.738	0.859
Activation	0.145	0.380
Dominance	0.096	0.309

7 CONCLUSIONS

The implemented visual analytics tool, SpeechVis, successfully provided solutions to the initial questions raised. SpeechVis interactive visualizations can identify and display the behavior of multiple speakers, the temporal characteristics of the conversation, and the main topics, connecting the topics with the speaker's emotions. It is available online[2], and any user can upload their audio file to analyze it. Moreover, we extracted the required data from audio recordings using off-the-shelf machine learning models. We evaluated this by comparing it with the widely recognized IEMOCAP dataset as a baseline. The successful examination of an entire presidential debate case further proves that our tool can extract valuable information from conversation recordings. Also, the SpeechVis tool enables other researchers to analyze your data, potentially leading to new insights from different audio recordings.

It is important to note that there are some limitations in the study that should be considered. For instance, inaccurate analyses may occur due to the machine learning models used in the work. Although these models effectively extract emotional information from audio, there may be cases where the analyses could be more precise, as the models are subject to limitations and uncertainties inherent to the learning process. Conversely, the SpeechVis is a modular tool that allows overcoming this drawback by swapping the models selected in the pipeline and substituting them with more accurate machine learning models.

[2]https://gmap.pucrs.br/speechvis/

Additionally, it is essential to highlight that the visualization requires preprocessing of the audio before being presented, which currently does not allow real-time audio analysis using the current machine learning techniques selected in the study.

Future work may focus on enhancing machine learning models to improve the accuracy of analyses, taking into account different accents, regional culture, and background noise. Further, developing faster and parallel audio preprocessing techniques to extract our required data can allow real-time analysis. These potential improvements could significantly enhance SpeechVis capabilities and usability and mitigate current limitations.

ACKNOWLEDGMENTS

This work was partially supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, FAPERGS 09/2023 PqG (Nº 24/2551-0001400-4), CNPq Research Program (Nº311012/2025-6 and Nº 303208/2023-6), and PUCRS institutional program to incentivize Stricto Sensu graduate studies - PRO-Stricto.

REFERENCES

[1] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv:2006.11477 [cs.CL]

[2] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. arXiv:2303.00747 [cs.SD]

[3] Pablo Barros and Stefan Wermter. 2016. Developing crossmodal expression recognition based on a deep neural model. *Adaptive Behavior* 24, 5 (Oct. 2016), 373–396. https://doi.org/10.1177/1059712316664017

[4] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D3: Data-Driven Documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)* (2011). http://vis.stanford.edu/papers/d3

[5] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2019. pyannote.audio: neural building blocks for speaker diarization. arXiv:1911.01255 [eess.AS]

[6] Paul Buitelaar, Ian D. Wood, Sapna Negi, Mihael Arcan, John P. McCrae, Andrejs Abele, Cecile Robin, Vladimir Andryushechkin, Housam Ziad, Hesam Sagha, Maximilian Schmitt, Bjorn W. Schuller, J. Fernando Sanchez-Rada, Carlos A. Iglesias, Carlos Navarro, Andreas Giefer, Nicolaus Heise, Vincenzo Masucci, Francesco A. Danza, Ciro Caterino, Pavel Smrz, Michal Hradis, Filip Povolny, Marek Klimes, Pavel Matejka, and Giovanni Tummarello. 2018. MixedEmotions: An Open-Source Toolbox for Multimodal Emotion Analysis. *IEEE Transactions on Multimedia* 20, 9 (Sept. 2018), 2454–2465. https://doi.org/10.1109/tmm.2018.2798287

[7] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation* 42, 4 (Dec. 2008), 335–359.

[8] Tim Dalgleish and Mick Power (Eds.). 1999. *Handbook of Cognition and Emotion*. John Wiley & Sons, Chichester, England. https://doi.org/10.1002/0470013494

[9] Catherine D'Ignazio. 2017. Creative data literacy: Bridging the gap between the data-haves and data-have nots. *Information Design Journal* 23, 1 (2017), 6–18.

[10] Christian Hildebrand, Fotis Efthymiou, Francesc Busquet, William H. Hampton, Donna L. Hoffman, and Thomas P. Novak. 2020. Voice analytics in business research: Conceptual foundations, acoustic feature extraction, and applications. *Journal of Business Research* 121 (2020), 364–374. https://doi.org/10.1016/j.jbusres.2020.09.020

[11] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. *spaCy: Industrial-strength Natural Language Processing in Python*. 10.5281/zenodo.1212303

[12] Harshit Katyal. 2024. harshit345/xlsr-wav2vec-speech-emotion-recognition. https://huggingface.co/harshit345/xlsr-wav2vec-speech-emotion-recognition

[13] Renato Kempter, Valentina Sintsova, Claudiu Musat, and Pearl Pu. 2014. EmotionWatch: Visualizing Fine-Grained Emotions in Event-Related Tweets. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014), 236–245. https://doi.org/10.1609/icwsm.v8i1.14556

[14] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv:1910.13461 [cs.CL]

[15] Kristina Loderer, Kornelia Gentsch, Melissa C. Duffy, Mingjing Zhu, Xiyao Xie, Jason A. Chavarría, Elisabeth Vogl, Cristina Soriano, Klaus R. Scherer, and Reinhard Pekrun. 2020. Are concepts of achievement-related emotions universal across cultures? A semantic profiling approach. *Cognition and Emotion* 34, 7 (April 2020), 1480–1488. https://doi.org/10.1080/02699931.2020.1748577

[16] Kevin Maher, Zeyuan Huang, Jiancheng Song, Xiaoming Deng, Yu-Kun Lai, Cuixia Ma, Hao Wang, Yong-Jin Liu, and Hongan Wang. 2022. E-ffective: A Visual Analytic System for Exploring the Emotion and Effectiveness of Inspirational Speeches. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (Jan. 2022), 508–517. https://doi.org/10.1109/tvcg.2021.3114789

[17] Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament. *Current Psychology* 14 (Dec. 1996), 261–292. https://doi.org/10.1007/BF02686918

[18] Saif Mohammad and Peter Turney. 2010. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Diana Inkpen and Carlo Strapparava (Eds.). Association for Computational Linguistics, Los Angeles, CA, 26–34. https://aclanthology.org/W10-0204

[19] Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *Interspeech*. ISCA, 2765–2768. https://doi.org/10.21437/Interspeech.2004-668

[20] Chris North and Ben Shneiderman. 2000. Snap-together visualization: a user interface for coordinating visualizations via relational schemata. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (Palermo, Italy) *(AVI '00)*. Association for Computing Machinery, New York, NY, USA, 128–135. https://doi.org/10.1145/345513.345282

[21] Caluã de Lacerda Pataca, Matthew Watkins, Roshan Peiris, Sooyeon Lee, and Matt Huenerfauth. 2023. Visualization of Speech Prosody and Emotion in Captions: Accessibility for Deaf and Hard-of-Hearing Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 831, 15 pages. https://doi.org/10.1145/3544548.3581511

[22] Rosalind W Picard. 1997. *Affective Computing*. MIT Press, Cambridge, MA.

[23] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98–125. https://doi.org/10.1016/j.inffus.2017.02.003

[24] James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (Dec. 1980), 1161–1178. https://doi.org/10.1037/h0077714

[25] Jane Simpson, Sarah Carter, Susan H. Anthony, and Paul G. Overton. 2006. Is Disgust a Homogeneous Emotion? *Motivation and Emotion* 30, 1 (March 2006), 31–41. https://doi.org/10.1007/s11031-006-9005-1

[26] Valentina Sintsova, Claudiu Musat, and Pearl Pu. 2013. Fine-Grained Emotion Recognition in Olympic Tweets Based on Human Computation. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Alexandra Balahur, Erik van der Goot, and Andres Montoyo (Eds.). Association for Computational Linguistics, Atlanta, Georgia, 12–20. https://aclanthology.org/W13-1603

[27] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. 2022. Model for Dimensional Speech Emotion Recognition based on Wav2vec 2.0. https://doi.org/10.5281/zenodo.6221127

[28] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. 2023. Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 9 (2023), 10745–10759.

[29] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13, 915 English lemmas. *Behavior Research Methods* 45, 4 (Feb. 2013), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

[30] W.M. Wundt and C.H. Judd. 1897. *Outlines of Psychology*. W. Engelmann. https://psychclassics.yorku.ca/Wundt/Outlines/sec13.htm

[31] Irena Yanushevskaya, Christer Gobl, and Ailbhe Ní Chasaide. 2013. Voice quality in affect cueing: does loudness matter? *Frontiers in Psychology* 4 (2013). https://doi.org/10.3389/fpsyg.2013.00335

[32] Haipeng Zeng, Xingbo Wang, Aoyu Wu, Yong Wang, Quan Li, Alex Endert, and Huamin Qu. 2020. EmoCo: Visual Analysis of Emotion Coherence in Presentation Videos. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 927–937. https://doi.org/10.1109/TVCG.2019.2934656

[33] Jian Zhao, Liang Gou, Fei Wang, and Michelle Zhou. 2014. PEARL: An interactive visual analytic tool for understanding personal emotion style derived from social media. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 203–212. https://doi.org/10.1109/vast.2014.7042496