Emotions In Oral History Interviews: A Multimodal Approach to Holocaust Testimonies

Nele Mantaj, Vaibhav Agarwal, Inés Matres
University of Trier, Technical University of Munich, University of Helsinki
s2nemant@uni-trier.de, vaibhav.agarwal@tum.de, ines.matres@helsinki.fi

Abstract

Video interviews with Holocaust survivors and witnesses comprise, to date, the most globally distributed and comprehensive oral history documentation. As survivors among us disappear, these sources are increasingly important to understand the impact of the Holocaust and mechanisms to overcome the trauma experienced. While historians often rely on written transcripts, these omit emotional nuances conveyed through audiovisual cues such as facial expressions, pauses, and eye movements. This article outlines the resources, data-preparation steps, and analytical methods used during a 10-day Digital Humanities Hackathon project to examine emotions in Holocaust testimonies, incorporating video, audio, and text. The group aimed to determine whether audiovisual signals offer meaningful emotional or sentimental information beyond transcripts. To achieve this, the group worked with a sample of 10 interviews facilitated by the US Holocaust Memorial Museum (USHMM); which were separated into video, audio, and textual components for machine processing and realigned side-by-side for analysis. This resulting "cookbook" lays out a workflow, resources, and practical entry points for preparing oral history interviews for multimodal emotion and sentiment annotation, or to aid the detection of emotionally significant moments for deeper examination.

Keywords: Holocaust, Oral History interviews, Multimodal analysis, Emotion analysis

1. Background

Testimonies are a crucial source for understanding the Holocaust, providing first-hand accounts and personal narratives from survivors and witnesses. These narratives offer insight into individual experiences; however, much of the existing analysis has focused primarily on the accounts of events from these testimonies (Waxman, 2012). In order to leverage the large amounts of oral testimony in diverse oral history archives, machine learning (ML) methods based on transcripts have proven accurate in identifying topics (Ifergan et al., 2024), named entities, such as places or events (Anuradha Nanomi Arachchige et al., 2023), or relationships in Holocaust and other oral history testimony (Anuradha et al., 2023; Laato et al., 2025). Our approach is inspired by the turn in historical research that considers emotion as an important lens to examine past phenomena that helps bridge the gap between personal and collective; and between experience and expression (Eustace et al., 2012). While emotional expressions within these accounts have been explored through qualitative methods, a large-scale approach to emotion with oral testimony on the Holocaust has been minimal and only has aimed to improve Automatic Speech Recognition (ASR) methods (Bukreeva et al., 2023).

The main goal of this paper is to improve the preconditions for a more holistic emotional analysis of Holocaust testimony at scale, by laying out the processes involved in tackling multimodal information contained in interviews. After a detailed description of the data used, firstly, we present a workflow for splitting the interview into three components making each signal (transcript, audio and video) fit for machine-processing while interoperable to be analysed side-by-side (section 3). Secondly, we show what insights about emotion and sentiment can be exacted with aid of digital methods from the video and audio in addition to the transcript analysis (section 4); Finally, in sections 4 and 5 we discuss limitations of computational models and vistas for developing ASR and computer vision to support this approach.

2. Hackathon setting and data used

The setting of this project was an academic Hackathon in the space of 10 days in Spring 2025 at the University of Helsinki. The group formed by nine MA students from language studies, history, data and computer science and four instructors [1] gained access to a sample of 100 testimonies of Holocaust survivors and witnesses from the United States Holocaust Memorial Museum [2]. The data selection was done by researchers from the CLARIN Network [3] specialists in corpus linguistics. The dataset was balanced in terms of gender, distribution of witnesses and survivors; and one particular criteria was its language diversity, including 30 interviews in Czech, 23 in Polish, 20 in English, 15 in Dutch and 12 in French (Anuradha et al., 2026).

Although the dataset was not created attending to

[1] seeacknowledgements
[2] https://www.ushmm.org/
[3] https://www.clarin.eu/

Proceedings of The Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026, pages 66–73
11 May 2026.

their emotional content, highly emotional moments were identified by all students in preparation for the Hackathon. This consisted in viewing at least two interviews and suggesting topics to explore. Early on and after consulting with oral history experts among the instructors, the emotional content and heterogenous expressiveness of interviewees were the most remarkable for students and became the target of our project.

While emotional content could be found in any interview, we soon remarked that ways of expressing and verbalizing emotions were highly dependent on the individual rather than being defined by gender, language spoken, or type of testimony. This allowed some freedom when sampling this huge dataset. A final selection of 10 videos was done (see table 1) with the main requirement that they contained full-transcripts and videos. Additionally, these interviews lasted approximately one hour, because video models took substantial time in processing the material. Balance in gender was maintained and the language diversity was respected in selecting an equal number of interviews in English and Polish, as the group included native speakers of both languages to ensure verification of results. A witness bias (7 of the 10) is due to survivor testimonies being in average longer, but for the purpose of this study, this had not much impact, as both long and short interviews could contain strong emotional content, and each individual is unique in expressing or containing their emotions.

No.	Type	Born	Lang.	Gender
1	Survivor	1933, Germany	EN	Female
2	Witness	1915, Unknown	EN	Male
3	Survivor	1934, today Czech Rep.	EN	Female
4	Survivor	1929, Poland	EN	Male
5	Witness	Unknown, United States	EN	Male
6	Witness	1915, today Ukraine	PL	Male
7	Witness	1917, Poland	PL	Female
8	Witness	1914, Poland	PL	Female
9	Witness	1931, Poland	PL	Female
10	Witness	1917, Poland	PL	Male

Note: EN = English, PL = Polish

Table 1: Subset of interviews used in the hackathon

3. Workflow

In this section we explain the steps taken to extract the three signals from the interviews in independent layers, and the processes to transform these into machine-readable formats and applying models and results. In doing this, dependencies emerged and some of these processes were done concomitantly, hence while an illustration of the workflow and dependencies is shown in Figure 1 (next page), for clarity we describe them separately.

In addition to more established sentiment and Emotion analysis applied to text (transcripts), paralinguistic features such as speech patterns, silences, changes in voice, hesitations, or gestures, provide cues for emotional intensity and variation in oral history interviews. To account for granularity for these features we choose to test specialised models which prioritized accuracy and input data type over generalisable output produced using LLMs. Adopting this approach also shed light on the pitfalls of the current state of the art models, which are discussed in section 5.

It is also important to note that a full analytical pipeline should aim to model the entire spectrum of paralinguistic features central to oral testimonies; however, given the setting and time constraints associated to a Hackathon project, modeling these features turned out to be infeasible. While it was possible to successfully detect silences and perform diarization to ensure coherent audio utterances and corresponding text transcripts, the pipeline had to rely on computational models which captured these features in hidden layers (for example, using the Wav2vec 2.0 model based on (Wagner et al., 2023) for valence and arousal detection), instead of an analysis catered towards the paralinguistic features, which would have added further nuance to the results from the modalities.

3.1. Text

The original 100 piece dataset included heterogeneous text material in addition to the video recordings. These could be transcriptions in PDF format in their original language, a few had additional translations in English, and some had summaries instead of transcripts. This heterogeneity posed several technical challenges: transcripts were stored in non-machine-readable formats, varied greatly in structure, language, and completeness, and some interviews lacked transcripts entirely. Since the interviews were conducted in different languages, this required language-specific processing pipelines and models.

Existing PDF transcripts were first converted into plain TXT files to enable natural language processing. Due to inconsistent formatting, automatic methods were insufficient to reliably identify speakers or dialogue turns. Therefore, the text was automatically segmented into utterances, defined as uninterrupted sequences of speech by a single speaker. Assigning each utterance to a speaker (interviewer or interviewee) required manual annotation. Finally, the annotated transcripts were converted into structured JSON files. Each JSON object corresponded to a single utterance and included metadata such as speaker identification and interview ID, making the data directly compatible with machine learning models.

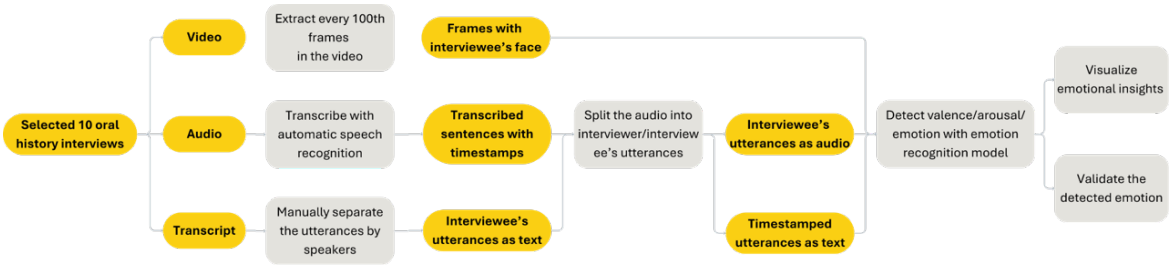

Figure 1: Proposed workflow for multimodal emotion analysis of oral history interviews

For interviews without transcripts, automated speech-to-text transcription was tested. A Whisper-based large language model fine-tuned for French speech recognition was used and produced acceptable transcription quality for French interviews, although manual correction and annotation were still required. In contrast, automated transcription for Polish interviews yielded poor results. This contributed to the decision to adjust the dataset, ultimately relying exclusively on existing Polish and English transcripts, annotated as described above.

All sentiment, emotion, and affective analyses were performed using transformer-based language models, which represent the current standard in natural language processing. Transformers model contextual relationships between all tokens in a text using self-attention mechanisms, enabling them to capture nuanced semantic and emotional information beyond keyword-based approaches. The models used in this workflow were retrieved via the Hugging Face platform.

Sentiment analysis aimed to classify text segments according to their overall polarity (positive, negative, or neutral). Transformer-based sentiment models derived sentiment from contextual embeddings, making them robust to linguistic phenomena such as negation. For English data, the RoBERTa-based model *cardiffnlp/twitter-roberta-base-sentiment*[4] was used, which output predictions for positive, neutral, and negative sentiment. For Polish data, the GPT-2–based model *nie3e/sentiment-polish-gpt2-large*[5] was applied. In addition to positive, negative, and neutral labels, this model output an "ambiguous" category, which was excluded from further analysis to avoid uncertain classifications.

The Emotion analysis expanded sentiment analysis by identifying specific emotional categories. For English transcripts, the model *j-hartmann/emotion-english-roberta-large*[6] was used, targeting seven emotions (anger, disgust, fear, joy, neutral, sadness, and surprise). For Polish data, the model *hplisiecki/polemo_intensity*[7] was applied, which predicted intensity scores for six emotions (happiness, sadness, anger, disgust, fear and pride) but lacked a neutral label. These models output continuous scores rather than single categorical labels.

In addition to categorical emotions, affective states were modelled along the continuous dimensions of valence and arousal. For English data, valence was estimated using the transformer-based model *chrlukas/stories-emotion-c0*[8]. For Polish data, the emotion detection model also provided valence and arousal estimates. Since model outputs differed in scale, all valence scores were linearly transformed to a standardized range between -1 and 1 to ensure comparability.

3.2. Audio

The task of analyzing the interviewees' speech to track emotional changes over time involved further pre-processing. This required segmenting the speech into coherent parts that contained relevant information about the emotions portrayed in the speech. The pre-processing had to take into consideration the current speaker, the flow of the speech and possibly the topics handled.

In natural speech, emotional states are dynamic, i.e., they fluctuate over time. This is important to note when looking at oral testimonies due to their long duration. When survivors or witnesses recall distinct events, their emotional expression varies across different segments. To capture these nuances, it was necessary to split the continuous audio into coherent segments. Following this, we segmented the speech into utterances, defined as uninterrupted chains of speech that follow the speaker's natural flow. This ensures input for the emotion detection models remaining contextually consistent.

[4] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

[5] https://huggingface.co/nie3e/sentiment-polish-gpt2-large

[6] https://huggingface.co/j-hartmann/emotion-english-roberta-large

[7] https://doi.org/10.1007/s11135-025-02116-8

[8] https://huggingface.co/chrlukas/stories-emotion-c0

Furthermore, we were only interested in the interviewee's emotions, so we didn't need to predict the interviewer's emotional expression. This was possible through diarization, which was done using the powerset multi-class segmentation code available in the open-source *pyannote.audio*[9] library to detect the timestamps when the speaker changes in an audio file (Plaquet and Bredin, 2023; Bredin, 2023).

To find the optimal way to split the audio into utterances, we had to select what level of intensity is the cut-off point for a "silence". Some interviews had very long periods of silence that we had to investigate by listening to the recording, while in others, the silence was a part of the natural flow and added value. Some interviews also contained quiet periods where the interview team changed the recording tape, but it could have been possible that some silent periods were marked erroneously. The silences were detected based on Rudolfbyker's code to *split wav*[10] files by silence which takes into account how quiet must the audio be, as well as, for how long must the silence last before noting the cut.

It is also important to note that during diarization, numerous splits were created, which were partly caused by the various non-speaking voices and noises in the interviews. However, given the rich pre-existing transcripts with timestamps and the audio outputs from whisper with timestamps, we were able to overlay the silences detected with the nearest gap in the audio and get utterances that were coherent to the natural speech flow, and aligned with the transcriptions.

The output of the diarization was used together with the utterances to generate the relevant audio files for input in the emotion modeling. Since finally a subset of 10 interviews was selected for qualitative interpretation, all audio and text utterances were checked and corrected manually. The manual check served as an informal verification of the utterances and was performed by two researchers at the hackathon by cross referencing the automated speaker turns and utterances against the original recordings and generated text transcripts. In the few instances where the diarization failed to detect the transition or, if an utterance could be split again, the timestamps and transcript breaks were adjusted manually to ensure perfect alignment. Given the sample size, this manual process was feasible; however, for larger sample sizes in the future, Human-LLM assisted verification could be used.

The emotion modeling for audio was accomplished with large pre-trained models, mainly provided by the Hugging Face model library. For speech emotion recognition (SER) modeling, we used the Wav2vec 2.0 model as implemented and evaluated by Wagner et al. (Wagner et al., 2023). Wav2vec 2.0 is a neural network model relying heavily on transformer architectures with 12 transformer layers. For our analysis, we decided to use the wav2vec model for both English and Polish interviews as the authors of the paper claim the validity of the model for languages other than English. This also kept our results invariant to possible bias introduced by using different models for English and Polish. Training a specific model for Polish analysis was attempted but we observed there to be a lack of labeled Polish data for valence and arousal detection from audio, effectively making the task infeasible given the project's time constraints.

3.3. Visual

The video analysis task was to determine the emotions expressed by the survivors and witnesses based on their facial expressions. Rather than replacing traditional text analysis, this visual approach serves as a tool to identify specific 'emotional spikes' that call for deeper qualitative investigation by the researcher. We do not argue for automating the process of understanding oral testimonies, rather to facilitate the researchers to gain insights into potential emotional moments.

In the video analysis pipeline, the videos were analysed frame by frame and not by segmenting them into utterances. While it was necessary for the text and audio inputs to be split into utterances to maintain contextual consistency, facial expression analysis benefited from the higher granularity provided by the individual frames in the video. This is because, firstly, emotions change within seconds; we are interested in these changes or spikes, which would otherwise be lost if averaged across an utterance. Secondly, most model input requirements detect the emotional state based on a single snapshot (i.e., one frame), unlike the text and audio models which require a temporal contextual window. Finally, emotional facial expressions do not necessarily stop when the speaker stops talking. This approach therefore, allows the detection of emotional shifts that may occur across utterances, within a single utterance or during periods of silence.

To process the videos, a first challenge was that they had different frame-per-second rates (30-40 FPS), while it would have been more accurate to extract one per second, we extracted one frame every 100th, which is equivalent to having one snapshot every 2.5-3.3 seconds. In a more dynamic type of video a tighter extraction rate would be advised, however in these videos the interviewee is conven-

[9] https://github.com/pyannote/pyannote-audio

[10] https://gist.github.com/rudolfbyker/8fc0d99ecadad0204813d97fee2c6c06

tionally situated at the centre of the frame and is recorded from a fixed angle. The Emonet [11] emotional detection model, was applied to identify the emotions expressed, and to quantify valence (degree of positivity or negativity) and arousal (level of emotional intensity or excitement). Emonet is a convolutional neural network (CNN) optimized for estimating valence and arousal levels from faces in naturalistic conditions (Toisoul et al., 2021). The model claims to estimate the valence and arousal in a given image with a small margin of error. The model was run on all the extracted frames of the selected interviews to generate frame level predictions for categorical emotion labels (e.g., happy, excited, anger, fear etc.) and valence and arousal scores. In order to maintain feasible processing times, these processes were assisted using a High Performance Computing environment.

To aim at a correct identification of key moments based on the emotion expression of the interviewees, it was important to minimize instances where the interviewer's faces might have been captured. While the camera was focused on the interviewee for the majority of the interview, we included only those frames that detected one face. However, it is still possible that there could be some noise in the dataset that detected the interviewer rather than the interviewee. This resulting dataset contains the emotional trajectory of the testimonies based on the videos. Differences in emotional display across gender and survivor/witness testimonies were analysed using R [12]. By identifying frames with high emotional expression, we can identify specific narratives in the testimony where the visual data provides an unique layer of context, complementing the text and audio workflows.

4. Excerpts from the Analysis

To illustrate the main results from the emotion analysis, we extract in this section key moments of two interviews from our dataset (Anuradha et al., 2026). These moments show agreement between two signals (Figure 2) and disagreement (Figure 3) . For layout purposes, the figures are displayed in the following page.

First we highlight an excerpt of the testimony given by Judith Balassa Zucker, born circa 1934 in Czechoslovakia (Figure 2) [13]. Zucker survived the Holocaust in her hometown of Krupina, hiding in the mountains with her family and other Jews towards the end of World War II. In the moment shown she recalls the arrival of German soldiers in later stages of the war. The emotional analysis identifies fear, something that a human reader could agree with and would most likely tag the same. Incorporating the emotional analysis overtime for the video, we see an overall agreement (while the frames give more frequent results than the longer utterances in the timeline).

The next excerpt (Figure 3) shows a disagreement between transcript and audiovisual analysis. A peak was detected by the audio and visual models, both concern valence that refers to the positivity of the emotion. In this interview Josefa Anasiewicz, born in 1914, gives testimony of a mass-shooting in her village [14]. When the head-shot from the moment we see the peak we see a smile, according to the transcript, she recalls Jewish Easter holidays and making bread in the immediacy of sad memories about a fire in her house. While the smile validates the result, neither the sentiment or emotional analysis from the transcript identified the fast emotional change.

5. Limitations of data and models

There were methodological limitations arising from the data, the hackathon setting and the applied models. Our goal in acknowledging them in detail is to signpost vistas for improving other-than-English and audiovisual models.

Concerning the transcript material was highly heterogeneous, requiring extensive pre-processing and manual standardization. Furthermore, some transformer-based models imposed input length constraints (approximately 500 tokens), which made processing the data at the level of individual utterances necessary. While the selected interviews contained responses short enough to be analysed without further segmentation, longer responses would require splitting, potentially fragmenting semantic and emotional context. While it was not among our goals to make comparisons, the use of different language-specific models for English and Polish sentiment analysis of transcripts limits direct comparability, as the models differ in architecture, label sets, and training data. Finally, as all models were pre-trained and not fine-tuned on interview-specific data, their predictions may not fully capture the nuances. This limitation is further illustrated by the Polish emotion model used in this study, which was reported by its authors to have suffered from corrupted weights in an earlier version, leading to largely random predictions. Although this issue was later corrected, it highlights the broader risk of relying on pre-trained models whose internal limitations or instabilities may not be immediately apparent. If one wishes for comparability across

[11] https://github.com/face-analysis/emonet

[12] https://www.r-project.org/

[13] https://collections.ushmm.org/search/catalog/irn511823

[14] https://collections.ushmm.org/search/catalog/irn507914

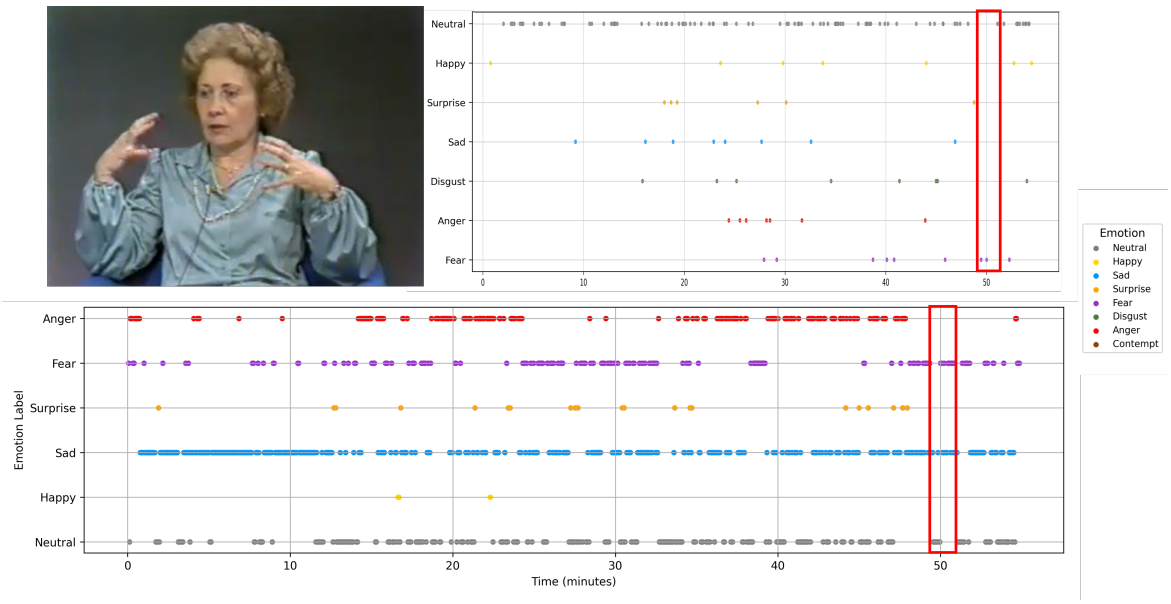

Figure 2: Emotional analysis over time from interview with Judith Zucker (emotion lables for transcript above, for video below): "About 10 o'clock, we got the news that Germans are coming. It was so cold [...] that you wouldn't send a dog out there. The wailing winds, it was unbelievable cold. **We hear the Germans are coming. We got to go.** So we go."

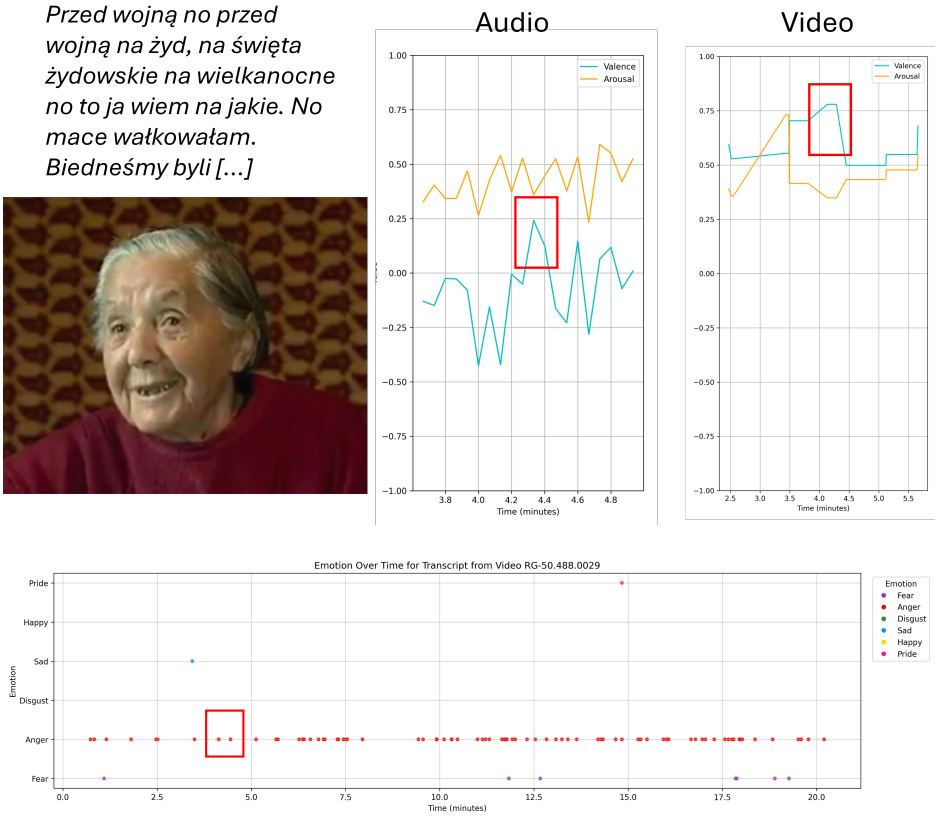

Figure 3: Emotional analysis from interview with Josefa Anasiewicz. Above the valence and arousal peak in the audio and video signals. Below, the emotional analysis based on the transcript, the positive sentiment is not detected. Translation: "Before the war for Jew [sic] for Jewish holidays, for Easter [sic], what do I know which one; I used to roll matzoh [bread]. We were poor [...]"

languages, it would be necessary to build and train a model of their own.

Concerning the audio workflow, a challenge of this approach is the limited reliability of the utterances. As an automatic process, the splitting of the speech might introduce errors. An additional concern remains whether the silences in the speech (context window around silences, breathing patterns, changes in voice etc.) are meaningful for analysis. On the other hand, even if the silent parts of speech contain relevant information about the emotions portrayed, the modeling we used for emotion detection would likely lack the capability for emotion recognition for silent audio. It is also important to note that the pre-trained models are mostly trained in English and with samples from younger adults. This introduces risks of age bias and unavailability of models relevant to different languages. Finally, as noted in the workflow, the diarization and silence detection models were sensitive to the acoustic environment of the interviews. For example, issues such as tape changes, background noise, or long, meaningful silences required manual oversight to ensure that segments were not erroneously discarded.

During the visual analysis, the primary limitation was the age bias in emotion detection. Pretrained facial emotional recognition models, including Emonet, are predominantly trained on datasets of younger individuals. This bias in training data is especially visible in expressions erroneously classified as sadness or anger. Regardless of the speaker's true emotion, changes in physiological features such as drooping eyelids and marionette lines led to these errors. In addition, the environment of the video recordings themselves introduce noise; for instance, camera zooms, lighting changes, changes of video tapes etc., which affect the model's ability to map facial landmarks coherently. Similarly, during instances where the interviewer's face may appear in the frame, the model will predict emotional labels for the interviewer, requiring data cleaning.

Finally, when looking at the textual, audio and visual models, which were developed on different training datasets and architectures, an unbiased comparison between these modalities is impossible. In addition, pre-trained models on the English language performed better when compared to other languages such as Polish or French, highlighting the need for improvement in language-specific or multilingual emotion detection models. It is clear that there is a need to test and develop multilingual and multimodal models, specifically fine-tuned for different use cases, where the training inputs and requirements more closely match the characteristics of Holocaust testimonies.

6. Conclusion, applications and vistas for research

An important motivating factor for focusing on emotions considering their verbal and non-verbal expressions, was tied to a fundamental value of having recorded and preserved Holocaust testimony as full-length video interviews: a great deal of emotional communication occurs in the audiovisual dimensions of testimony. The public online catalogue of USHMM alone contains over 14.000 digitized recorded interviews, of which less than a tenth include a transcript. While improvements in ASR, ML and NLP are enabling to ever more accurately transform and translate speech-to-text, the analysis of Holocaust testimony still relies on textual sources. The main contribution of this recipe book is to offer a proof of concept for multimodal large-scale analysis incorporating rich non-verbal emotional information left out of transcripts. Developing multimodal models, or replicating this in a more long-term project that allows to fine-tunine those tried during this project, can help identifying in one quick glance emotionally charged moments from long recordings, and down the line generating labels or metadata to enrich transcripts.

Another aim of this project, was to test a multimodal and partly multilingual approach to Holocaust Testimony. The time invested in preparation of data did not allow to produce a reliable study, but we succeeded in identifying the suitability of models or cues, such as valence and arousal in video and audio being valid indicators of emotionally charged moments. Furthermore we could recognise in models important flaws, such as the deficiency of audio models in other-than-English languages, or the bias of audio and visual models trained with samples of younger population and interpreting overly negative facial expressions of ageing population. Hence, we were able to sign-post vistas to further develop audio-visual models. This points at a related contribution, the preparation of the used dataset which contains rich multilingual interviews that with further refinement can become a benchmark on which existing or new models could be fine tuned or trained (Anuradha et al., 2026).

In zooming into in the workflow that we followed, one last contribution is the itemization of steps that need to be made in order to prepare non-machine readable transcripts and audiovisual recordings for machine-aided analysis. The recipes in this paper can be guide archives that hold oral history interviews and researchers working with audiovisual material, to turn this rich but heterogeneous data into machine readable datasets and benchmarks that allow further development of emotion detection models in other-than-English and ageing populations. Finally, we hope this inspires newcomers to

emotional approaches to Holocaust testimonies or digital humanities researchers to pay attention to the rich emotional information contained in audiovisual recordings.

7. Acknowledgements

We want to acknowledge the valuable work of other students of the Oral History group at the DHH2025: Visa Alamännistö, Haruka Buss, Joonatan Huang, Xiaoyue Wang, Rahel Albicker, Muhammad Hassan Qadeer Butt and Ellen Yang; as well as other instructors: Saara Kekki, Yu Wu and particularly Edyta Gawron. We also want to thank the US Holocaust Memorial Museum, Isuri Anuradha and Martin Wayne who facilitated an unrestricted access to the interviews.

8. Bibliographical References

Isuri Anuradha, Le An Ha, Ruslan Mitkov, and Vinita Nahar. 2023. Evaluating of Large Language Models in Relationship Extraction from Unstructured Data: Empirical Study from Holocaust Testimonies. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 117–123, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Isuri Anuradha Nanomi Arachchige, Le Ha, Ruslan Mitkov, and Johannes-Dieter Steinert. 2023. Enhancing Named Entity Recognition for Holocaust Testimonies through Pseudo Labelling and Transformer-based Models. In *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*, HIP '23, pages 85–90, New York, NY, USA. Association for Computing Machinery.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

Liudmila Bukreeva, Daria Guseva, Mikhail Dolgushin, Vera Evdokimova, and Vasilisa Obotnina. 2023. Emotional Speech Recognition of Holocaust Survivors with Deep Neural Network Models for Russian Language. In Alexey Karpov, K. Samudravijaya, K. T. Deepak, Rajesh M. Hegde, Shyam S. Agrawal, and S. R. Mahadeva Prasanna, editors, *Speech and Computer*, volume 14338, pages 68–76. Springer Nature Switzerland, Cham. Series Title: Lecture Notes in Computer Science.

Nicole Eustace, Eugenia Lean, Julie Livingston, Jan Plamper, William M. Reddy, and Barbara H. Rosenwein. 2012. *AHR* Conversation: The Historical Study of Emotions. *The American Historical Review*, 117(5):1487–1531.

Maxim Ifergan, Renana Keydar, Omri Abend, and Amit Pinchevski. 2024. Identifying Narrative Patterns and Outliers in Holocaust Testimonies Using Topic Modeling. ArXiv:2405.02650 [cs].

Joonatan Laato, Jenna Kanerva, John Loehr, Virpi Lummaa, and Filip Ginter. 2025. Extracting Social Connections from Finnish Karelian Refugee Interviews Using LLMs. ArXiv:2502.13566 [cs].

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. 2021. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*.

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):10645–10659.

Zoë Waxman. 2012. Chapter 7. Transcending History? Methodological Problems in Holocaust Testimony. In Dan Stone, editor, *The Holocaust and Historical Methodology*, pages 143–157. Berghahn Books.

9. Data Sources

Isuri Anuradha, Inés Matres, and Yu Wu. 2026. Multilingual dataset of interviews with survivors and witnesses of the holocaust https://doi.org/10.5281/zenodo.18701345.