Evaluating Automatic Speech Recognition for Holocaust Testimonies: A Large-Scale Analysis of Whisper Performance on the Fortunoff Video Archive

William J.B. Mattingly, Christy Bailey-Tomecek

Yale University
New Haven, CT, United States
william.mattingly@yale.edu, christy.tomecek@yale.edu

Abstract

Holocaust testimonies are key primary sources documenting survivors' experiences, yet many remain inaccessible due to the labor-intensive nature of manual transcription. This paper presents a comprehensive evaluation of OpenAI's Whisper automatic speech recognition (ASR) system on 1,847 testimonies from the Fortunoff Video Archive for Holocaust Testimonies at Yale University. We assess transcription quality across multiple languages including English, French, German, Hebrew, Yiddish, Ladino, Slovak, and American Sign Language (with English voice-over), using human-reviewed captions as ground truth. Our analysis reveals a mean Word Error Rate (WER) of 15.28%, with 90.9% of testimonies achieving "Fair" or better quality (WER ≤25%). We identify systematic error patterns including challenges with disfluencies, interrupted speech, and language-specific orthographic conventions, particularly in Ladino, where Whisper's normalization to modern Spanish orthography creates systematic divergences from traditional Judeo-Spanish spelling. For Hebrew and Yiddish, we evaluate specialized models from ivrit-ai and find promising results for heritage language preservation. Our findings demonstrate that current ASR technology can substantially accelerate Holocaust testimony transcription while highlighting the need for domain-specific fine-tuning and post-processing for optimal results.

Keywords: Holocaust testimonies, automatic speech recognition, Whisper, oral history, Ladino, Yiddish, Hebrew, digital humanities

1. Introduction

The Fortunoff Video Archive for Holocaust Testimonies at Yale University holds over 4,500 testimonies comprising more than 12,000 hours of recorded interviews in 20 languages. Testimonies serve as both a research source and as a memorial to survivors and victims of the Holocaust. In both contexts, the archive has a responsibility to make testimonies accessible, and transcripts are an important piece of this.

Creating transcripts has many challenges. Historically, transcripts required cultural heritage institutions to manually transcribe audiovisual recordings with their own staff, or to contract out to vendors that utilize audio speech recognition (ASR). Manual transcription takes dozens of hours, and vendor costs, whether it involves human review, reach into the thousands of dollars (Bailey-Tomecek, 2025; Rodriguez & Brown, 2023). As a result, many institutions such as the Fortunoff Archive have historically been unable to transcribe their materials at scale. This landscape has changed with the advent of newer open-source ASR applications that use AI speech-to-text like OpenAI's Whisper.

Cultural heritage institutions have been exploring the use of Whisper because of its general accuracy and lower cost of business. (Rodriguez & Brown, 2023). However, Whisper's accuracy is dependent on features of the recording itself. (Graham & Roll, 2024; Lynema & Dunn, 2025). Holocaust testimonies have unique characteristics that can potentially impact Whisper's results.

This paper evaluates Whisper's efficacy with Holocaust testimonies from the Fortunoff Archive. We compared raw transcripts from Whisper to human-edited ground truth ones for 1,847 testimonies in eight languages. Along with evaluating Whisper's general capabilities, we examined the specific needs and challenges of Hebrew, Yiddish, and Ladino testimonies in the corpus, including using specialized models for Hebrew and Yiddish from the ivrit-ai project. Based on these results, we recommend a workflow for creating and editing ASR transcripts for Holocaust testimonies.

2. Related Work

2.1 Automatic Speech Recognition for Oral History

Complete transcription has long been a goal for many cultural heritage institutions, to provide better accessibility, as well as enhance searchability (Rao et al., 2025). Additionally, transcripts can provide opportunities for digital humanities scholars for research and analysis, such as the Fortunoff Archive's *Let Them Speak: In Search of the Drowned*[1] project and Critical Editions series.[2] Institutions have tried many vendor-based audio speech recognition (ASR) products, including 3play, Trint, Rev, Amazon Transcribe, Google speech-to-text, Kaldi,

[1] http://lts.fortunoff.library.yale.edu/

[2] https://editions.fortunoff.library.yale.edu/

Proceedings of The Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026, pages 84–92

Microsoft Stream, and Sonix (Bailey-Tomecek, 2025; Dunn et al., 2024; Lundgard, 2024; Lynema & Dunn, 2025; Myntti & Steed, 2019; Rao et al., 2025, Rodriguez & Brown, 2023). The efficacy of these services varied and were dependent on the quality of the source recordings as much as the service itself. It was also impacted by language used. (Bailey-Tomecek, 2025; Lundgard, 2024, Myntti & Steed, 2019; Rodriguez & Brown, 2023).

The emergence of large-scale pretrained models has transformed the landscape. Whisper was trained on 680,000 hours of multilingual audio and demonstrates strong zero-shot performance across languages and domains (Radford et al., 2022). Institutions such as Emory University have reported as much as a 35% decrease in labor time for correcting Whisper output versus fully human-edited transcripts (Rao et al., 2025). For specialized applications, fine-tuned variants have shown substantial improvements. The ivrit-ai project achieved state-of-the-art results for Hebrew and Yiddish ASR through crowdsourced training data collection (Marmor et al., 2025).

However, Whisper's accuracy has been demonstrated to be dependent on the quality of the recording itself; the volume, gender, speech rate and accent of the speaker; and the presence of background noise or music (Graham & Roll, 2024; Lynema & Dunn, 2025). Depending on the type of recording, institutions like Indiana University reported word error rates between 1-8% for educational recordings, which utilize a professionally recorded English language speaker, and 20-40% for field recordings, which were done in less-than-ideal conditions with speakers with non-standard accents. (Lynema & Dunn, 2025). Similarly, Emory University reported word error rates between 5.24% for educational recordings to 24.01% for variety television shows that contain musical performances among other features. (Rao et al., 2025).

Historically, a drawback of Whisper for oral history collections such as the Fortunoff Archive is that it is not necessarily trained on corpora that have the speech qualities of oral histories (e.g. accents, disfluencies, halting speech), nor have domain-specific information (e.g. placenames in Europe that may not be commonly referred to in an English-language corpus). Researchers at University of West Bohemia Pilsen working with a similar collection of testimonies held by the USC Shoah Foundation's Visual History Archive sought to bridge that gap using a fine-tuned, monolingual Wav2Vec that trained from multiple corpora and utilized text from the Common Crawl project to build a robust vocabulary that would contain less common entities for Czech, English, and German language materials. The results were more accurate than transcripts utilizing XLS-R and Whisper models. (Lehečka et al., 2023). However, Whisper continues to improve in accuracy and is attractive to institutions due to its wide availability and ease of use.

After reviewing available literature and feedback, the majority consensus is that Whisper is accurate enough for at least topical research purposes, if not for full accessibility, and institutions consider it a realistic solution for large scale transcription of collection materials (Dunn et al., 2024; Harbert, 2025; Lundgard, 2024; Lynema & Dunn, 2025; Rao et al., 2025, Rodriguez & Brown, 2023).

2.2 Challenges in Testimony Transcription

Holocaust testimonies from the Fortunoff Archive provide significant challenges for modern ASR platforms. These include:

- **Condition of the original carriers and associated impact on digitization.** Most of the testimonies were recorded on magnetic tape formats. Magnetic tape has a limited lifespan, on average lasting 15 years. The tape degradation will impact audio quality in digitization
- **Surrounding environment during the recordings.** Many testimonies were recorded in non-studio environments, both indoors and outdoors, with background noise, and may have lacked professional audio recording equipment
- **Elderly speakers.** Most testimonies were recorded when survivors were in their 60s-90s, with age-related voice characteristics
- **Emotional speech.** Testimonies frequently include crying, pauses, and voice breaking during traumatic recollections
- **Language switching:** Survivors often mix languages, e.g. they may use the language(s) spoken at the time of the recalled event rather than the primary language of the interview
- **Non-native accents for English and Hebrew testimonies:** Many survivors speak English and Hebrew as second or third languages, with accents and pronunciations that are not standard for a native speaker
- **Disfluencies:** As testimonies are unscripted, survivors will pause, stutter, repeat words multiple times in a row, and use fillers (um, eh, uh)
- **Less well-known locations and context-specific jargon:** Names of camps, geographic locations, and institutions are less likely to be recognized by ASR. Additionally, survivors may use ghetto and camp specific jargon.

3. Dataset

3.1 The Fortunoff Video Archive

The Fortunoff Video Archive for Holocaust Testimonies holds over 4,500 testimonies recorded between 1979-2024. The testimonies were recorded in 20 languages throughout North America, South America, Europe, and Israel. Testimonies range from 30 minutes to over 40 hours in length, with an average length of one and a half hours. Interviews were unscripted and unguided by the interviewers except for occasional, clarifying questions based on the survivor's discussion.

For this study, we analyzed a subset of 1,847 testimonies for which both Whisper-generated transcripts and human-reviewed captions were available. The distribution by primary language is shown in Table 1.

Language	Code	Count	Percentage
English	eng	1,699	91.9%
Unknown	--	71	3.8%
German	ger	35	1.9%
French	fre	25	1.4%
Hebrew	heb	5	0.3%
Ladino	lad	5	0.3%
Yiddish	yid	4	0.2%
American Sign Language	sgn	2	0.1%
Slovak	slo	1	0.05%
Total	--	1,847	100%

Table 1: Testimony distribution by language

3.2 Ground Truth Captions

Most ground truth captions were created by several vendor products and later corrected by archive staff. Vendors either used a hybrid of ASR with human editors employed by the vendor or purely ASR. These include 3play for English (hybrid), Trint for English, German, French, and Slovak (ASR only), Verbit for Hebrew (hybrid), and Sonix for all languages except Yiddish and Ladino. (ASR only). Yiddish and Ladino captions were created manually by scholars associated with archive projects. Yiddish transcripts were created in the Elan editor, which includes time synchronization while Ladino transcripts were transcribed without time synchronization and later aligned using Sonix's forced alignment tool with the Spanish model.

The archive's style guide includes the following:

- Diarization, either with the speaker names, or general labels like "interviewer" or "subject"
- Verbatim transcription, including disfluencies (e.g. "uh", "um"). Grammar and phrasing are not normalized or corrected as the archive prioritizes fidelity to the speaker. Mispronunciations are not documented in the text
- Notation of interrupted speech or false starts with em-dashes (e.g., "I was—")
- Bracketed annotations for non-verbal sounds using all-caps (e.g. [LAUGHS], [CRYING])
- Bracketed annotations for unclear or inaudible speech using all-caps (e.g. [INAUDIBLE])
- Bracketed annotations for words and phrases that the transcriptionist thinks are being used, but not completely certain of, using question marks directly after the opening bracket and directly before the closing bracket (e.g. "[? He said ?]")

Ground truth captions may have errors that contribute to false error rates. These include:

- Time misalignment in the vendor's editing platform. Most of the editing platforms used by the archive's vendors do not show any sort of timestamping beyond line break/paragraph levels. If more than a couple of words need to be edited, it can accidentally misalign individual syllables or even pin an entire sentence to one timestamp
- Time misalignment by manual transcriptionists. While Elan provides time-synchronization, some amount of misalignment still can happen.
- Choices by human editors. Some editors corrected grammar when asked not to, or did not transcribe sections of audio, e.g. conversation with crew members. These were corrected by archive staff whenever found, but may have been missed in other transcripts

Ladino transcripts present additional challenges as ground truth captions due to the language's Romanized orthography not being fully standardized. (Kohen & Kohen-Gordon, 2000) Choices by one transcriptionist may not be the same as another.

3.3 ASR System Configuration

We employed Whisper using the faster-whisper implementation with the large-v3-turbo model variant for optimal speed-accuracy tradeoffs. For Hebrew and Yiddish testimonies, we utilized specialized models from the ivrit-ai project:

- Hebrew: ivrit-ai/whisper-large-v3-turbo-ct2

- Yiddish: ivrit-ai/yi-whisper-large-v3-turbo-ct2

These models were fine-tuned on crowdsourced Hebrew and Yiddish audio data, achieving substantially lower word error rates than the base Whisper model on these languages.

As Ladino lacks a dedicated model, but shares a strong common root with Old Castilian Spanish, we utilized Whisper's Spanish language model. This choice was influenced by past success with forced alignment of Ladino transcripts with testimony audio using Sonix's Spanish model (Bailey-Tomecek, 2025).

4. Methodology

4.1 Evaluation Metrics

We compute the standard Word Error Rate (WER) metric using Python's difflib.SequenceMatcher. For each testimony, we:

1. Parse both ASR output and ground truth VTT files
2. Clean ground truth text by removing speaker labels and technical annotations
3. Normalize both texts to lowercase with punctuation removed
4. Normalize number representations (e.g., "fourteen" → "14")
5. Filter common disfluencies (e.g., "um") that may differ by transcription convention
6. Calculate WER as (substitutions + deletions + insertions) / total_reference_words

WER is the standard metric in speech recognition evaluation, where values below 20% are generally considered good for conversational speech. Note that WER can exceed 100% when the output contains significantly more words than the reference (due to insertions or hallucinations).

4.2 Word-Level Error Analysis

We categorize word-level errors into three types:

- Missed words: Present in ground truth but absent from ASR output
- Extra words: Present in ASR output but absent from ground truth (potential hallucinations)
- Replaced words: Words that differ between transcripts

We further annotate errors using spaCy's part-of-speech tagger to identify patterns across word categories (nouns, verbs, proper nouns, function words, etc.).

5. Results

5.1 Overall Performance

Across 1,847 analyzed testimony segments, we observed the following aggregate statistics:

Metric	Value
Mean WER	15.28%
Median WER	13.81%
Standard Deviation	10.92%
Minimum	0.00%
Maximum	191.04%
Total Words	12,070,412
Total Errors	1,917,171

Table 2: Overall WER Statistics

These results indicate that Whisper achieves approximately 85% word-level accuracy on Holocaust testimonies, a figure that compares favorably with human inter-annotator agreement on complex transcription tasks. The maximum WER exceeding 100% indicates cases where ASR output contained significantly more content than the reference, typically due to language mismatches or audio-transcript misalignment (see below).

5.2 Quality Distribution

We categorized testimonies into quality tiers based on WER. These categories are based on the complexity of the problem space.

Quality Tier	WER Range	Count	Percentage
Excellent	0-15%	1,074	58.1%
Good	15-20%	427	23.1%
Fair	20-25%	179	9.7%
Poor	25-30%	82	4.4%
Very Poor	>30%	85	4.6%

Table 3: Quality Distribution

Notably, 90.9% of testimonies fall within the "Fair" quality tier or better (WER ≤25%), suggesting that Whisper outputs can serve as useful first drafts for human review. The majority (58.1%) of testimonies achieve "Excellent" quality with WER ≤15%.

5.3 Performance by Language

Language significantly impacts ASR performance:

Language	Count	Mean WER	Min	Max
Sign Language (voice-over)	2	4.12%	1.87%	6.37%
Slovak	1	11.28 %	11.28 %	11.28 %
Hebrew	5	13.12 %	9.01%	19.26 %
Unknown	71	13.15 %	1.74%	61.69 %
English	1,699	15.44 %	0.00%	191.04%
French	25	16.02 %	6.92%	27.71 %
German	35	20.42 %	10.96 %	44.70 %
Ladino	5	80.51 %	73.49 %	102.86%
Yiddish	4	111.71%	103.06%	132.02%

Table 4: Performance by Primary Language

Sign language testimonies with voice-over achieve the best results (4.12% WER), as these contain professionally narrated English translations. Hebrew testimonies using the ivrit-ai fine-tuned model perform well (13.12% WER). English testimonies, comprising the majority of the corpus, achieve strong performance at 15.44% mean WER.

German testimonies exhibit higher WER (20.42%), potentially due to:

- Historical German variants and dialectal features
- Code-switching between German narrative and English interviewer questions

Heritage languages present significant challenges. Ladino testimonies show 80.51% WER due to Whisper's orthographic normalization to modern Spanish (discussed in Section 6.4).

5.4 High-WER Case Analysis

Extreme WER values (>80%) typically indicate systemic issues rather than poor ASR performance:

- Script mismatch: Hebrew-script languages (Hebrew, Yiddish) compared against Latin-script ASR output, or vice versa
- Language mismatch: Testimonies conducted in unmodeled languages (Ladino) where Whisper defaults to the closest supported language
- Orthographic conventions: Ladino testimonies where Whisper outputs modern Spanish orthography rather than traditional Judeo-Spanish spelling
- File mismatches: In some cases, ground truth captions corresponded to different testimony segments than the ASR output

When excluding heritage language testimonies (Ladino and Yiddish) where specialized models are needed, mean WER drops to approximately 13.36%.

6. Error Analysis

6.1 Disfluencies Handling

The most significant source of WER stems from differential handling of disfluencies. Ground truth captions meticulously transcribe verbal fillers following archival guidelines, while Whisper tends to omit or normalize them.

Disfluency	Missed (GT > ASR)	Extra (ASR > GT)	Net Differnce
uh,	4,978	311	-4,667
um,	1,118	0	-1,118
uh--	1,029	0	-1,029
yeah.	583	459	-124
mm-hmm.	499	370	-129
eh,	397	0	-397
ok.	370	0	-370
mhm.	360	0	-360
um--	296	0	-296
mm-hm.	286	0	-286

Table 5: Disfluency Discrepancies (Top 10)

The pattern is consistent: Whisper systematically under-represents verbal fillers compared to verbatim ground truth transcription. This is by design, as Whisper is optimized for readability rather than forensic transcription. For Holocaust testimony applications where preserving speech patterns may be analytically important, post-processing to restore disfluencies may be desirable.

Further work still needs to be done in the identification of disfluency in non-English

languages. These affect the presented WER, but for downstream applications, we map these to a single standardized form. This allows us to represent these sounds in a unified way.

6.2 Interrupted Speech

Ground truth captions use em-dashes to indicate interrupted or self-corrected speech (e.g., "I was—I mean, we were"). These interrupted forms are heavily represented in missed word statistics:

Pattern	Missed Count
the--	2,485
I--	1,636
a--	1,410
to--	1,127
in--	791
was--	730
and--	692

Table 6: Interrupted Speech Markers

Whisper typically completes or omits interrupted words rather than preserving the interruption marker. This represents a fundamental difference in transcription philosophy—verbatim vs. normalized—rather than an ASR error per se.

6.3 Function Word Variation

High-frequency function words dominate both missed and extra word categories:

Word	Missed	Extra	Net
the	5,427	4,597	-830
and	4,019	3,238	-781
a	3,290	3,298	+8
I	3,364	2,870	-494
to	2,795	2,525	-270

Table 7: Functional Word Errors

The near-balance of missed and extra function words, combined with the high "replaced word" count (1.3M total), suggests that many apparent errors are actually timing/segmentation differences. When text from adjacent segments is considered, the effective error rate decreases substantially.

6.4 Ladino Orthographic Normalization

The most linguistically interesting error pattern emerges in Ladino (Judeo-Spanish) testimonies. Whisper, trained primarily on modern Spanish, systematically normalizes Ladino orthography:

Ladino (Ground Truth)	Whisper (Modern Spanish)	Pattern
katorze	14 / catorce	k → c
anyos	años	ny → ñ
kuando	cuando	k → c
deportasyon	deportación	syon → cíon
famiya	familia	y → i
ermana	hermana	∅ → h
klasika	clásica	k → c
djudya	judía	dj → j
ke	que	k → qu
kozas	cosas	k → c, z → s
skola	escuela	sk → esc
avlava	hablaba	v → b, ∅ → h
munchos	muchos	n deleted
djidyos	judíos	dj → j

Table 8: Ladino Orthographic Mappings

These are not ASR errors in the conventional sense as Whisper correctly recognizes the spoken words but outputs them in modern Spanish orthography rather than romanized Ladino spelling. This creates high WER despite semantic accuracy.

The implications are significant for heritage language preservation:

- Phonetic accuracy: The underlying speech is correctly recognized
- Orthographic loss: Distinctive Ladino spelling conventions are erased
- Cultural significance: Ladino orthography carries historical and cultural meaning that normalization destroys

For Ladino testimony transcription, we recommend:

- Using the ASR output as a phonetic guide
- Post-processing with Ladino-specific spelling rules
- Training specialized Ladino ASR models on available corpora

6.5 Hebrew and Yiddish Performance

For Hebrew testimonies, we utilized the ivrit-ai whisper-large-v3-turbo-ct2 model, which was fine-tuned on crowdsourced Hebrew audio. Qualitative analysis shows strong performance:

Ground Truth: אני נולדתי ביוגוסלביה בסובוטיצה, שזו עיר גבול יוגוסלביה-הונגריה.

Whisper: אני נולדתי ביוגוסלביה, בסובוטיצה, שזו עיר גבול, יוגוסלביה-הונגריה.

The ivrit-ai model correctly handles:

- Hebrew script and diacritics
- Place names (סובוטיצה / Subotica)
- Historical dates in Hebrew
- Natural speech patterns

For Yiddish, the ivrit-ai yi-whisper-large-v3-turbo model, trained on Yiddish audio data, shows promising semantic results. However, upon review by a Yiddish transcriptionist, it is clear that the ivrit-ai corpus is built on modern spelling and pronunciation conventions that do not always align with YIVO-standard conventions that more closely matches period Yiddish (B. Sadock, personal communication, March 17, 2026). The WER may improve if the raw output is converted to be more in line with YIVO standards. We are exploring using Gemini 3.1 Pro to resolve this challenge.

7. Discussion

7.1 Implications for Testimony Transcription

Our findings suggest that Whisper provides a strong foundation for accelerating Holocaust testimony transcription. With 90.9% of testimonies achieving ≤25% WER and 58.1% achieving excellent quality (≤15% WER), ASR outputs are in line with other institutions that reported time savings using Whisper and can serve as usable first drafts (Rao et al., 2025). However, several caveats apply:

- Verbatim requirements: If verbatim transcription is required (preserving all disfluencies and interruptions), Whisper outputs require substantial post-editing
- Heritage languages: Ladino, Yiddish, and other heritage languages need specialized models or post-processing
- Quality control: Approximately 5% of testimonies with very poor performance (WER >30%) require human transcription or specialized handling

7.2 Recommendations for Implementation

Based on our analysis, we recommend the following pipeline for Holocaust testimony ASR:

- Language detection: Automatically identify primary language(s) in each testimony
- Model selection: Use language-specific fine-tuned models where available (ivrit-ai for Hebrew/Yiddish)
- Post-processing: Apply domain-specific corrections:
    - Restore common disfluencies based on audio analysis
    - Apply heritage language orthographic conventions
    - Standardize Holocaust-specific terminology
- Quality scoring: Compute confidence scores to prioritize human review for low-confidence segments
- Human review: Focus human effort on proper nouns, dates, and low-confidence passages

7.3 Limiting Factors on Study

This study has several limitations:

- Language coverage: Our quantitative analysis focuses primarily on English testimonies; other languages require further study
- Ground truth variation: Transcription guidelines may have evolved over time, introducing inconsistencies in ground truth
- Model versions: Whisper continues to improve; newer versions may show different error patterns
- Semantic accuracy: Our WER metric captures word-level accuracy but not semantic accuracy—correctly transcribed content in different words would register as errors
- Hallucinations: Hallucinations certainly appear as a result of this process. We are currently developing solutions to identify and flag these hallucinations, including running the model over the same video multiple times to identify disfluency

7.4 Future Work

Several directions merit further investigation:

- Fine-tuning on Holocaust testimonies: Training domain-specific models on available transcribed testimonies
- Ladino ASR development: Creating the first dedicated Ladino speech recognition model
- Named entity recognition: Developing specialized NER models for Holocaust-related names, places, and terminology
- Multimodal analysis: Incorporating visual information from video testimonies to improve transcription accuracy

8. Conclusion

This paper presents the first large-scale evaluation of automatic speech recognition on Holocaust testimonies, analyzing 1,847 testimony segments from the Fortunoff Video Archive across 11 languages. We find that OpenAI's Whisper achieves approximately 84% word-level accuracy (16.3% mean WER), with 90.9% of testimonies achieving WER ≤25% and suitable for human review with ASR assistance.

Our error analysis reveals that apparent errors often reflect differences in transcription philosophy rather than ASR failures—particularly

for disfluencies and interrupted speech that Whisper normalizes but archival guidelines preserve. For heritage languages like Ladino, Whisper's orthographic normalization to modern standards presents both an opportunity (phonetic accuracy) and a challenge (loss of traditional spelling). Languages without dedicated model support (Yiddish without specialized models) require alternative approaches.

Specialized models, such as those from the ivrit-ai project for Hebrew and Yiddish, demonstrate that targeted fine-tuning can substantially improve performance on underrepresented languages. We encourage the development of similar resources for Ladino and other Holocaust-relevant languages.

As archives worldwide work to make Holocaust testimonies more accessible, ASR technology offers a crucial tool for scaling transcription efforts. Our findings provide a foundation for implementing ASR pipelines while highlighting the need for domain expertise, quality control, and respect for the linguistic heritage embedded in survivors' own words.

9. Acknowledgments

We thank the Fortunoff Video Archive for Holocaust Testimonies at Yale University for providing access to testimony recordings and transcriptions. We acknowledge the ivrit-ai project for their work on Hebrew and Yiddish ASR models. We are grateful to the transcriptionists and archivists whose careful work created the ground truth captions that made this evaluation possible.

10. Bibliographical References

Bailey-Tomecek, C. (2025, October 28). *Comparing Sonix and aTrain for transcribing French, German, Hebrew, and Russian Testimonies* [Presentation] AI for Libraries, Archives, and Museums Speech-to-Text Working Group.

Dunn, J., Lynema, E., Wheeler, B., Cameron, J. (2024, October 18). *Whisper applied to digitized historical audiovisual materials* [Conference presentation]. Fantastic Futures 2024.

Graham, C., & Roll, N. (2024). Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2). https://doi.org/10.1121/10.0024876

Harbert, R.H. (2025 March 10). *Whisper at High Volumes: How to transcribe an Archive* [Conference presentation]. Code4Lib 2025, Princeton, NJ, United States. https://www.youtube.com/live/A3I4OSTOQzo?si=PecwW9RMjqMPBPA3&t=6021

Kohen, E., & Kohen-Gordon, D. (2000). *Ladino-English, English-Ladino: Concise Encyclopedic Dictionary (Judeo-Spanish)*. Hippocrene.

Lehečka, J., Švec, J., Psutka, J.V., Ircing, P. (2023). Transformer-based Speech Recognition Models for Oral History Archives in English, German, and Czech. *Proceedings of Interspeech 2023*, 201-205. https://doi.org/10.21437/Interspeech.2023-872

Lynema, E. & Dunn, J. (2025, March 10). *Whisper speech-to-text for digitized historical audiovisual materials* [Conference presentation]. Code4Lib 2025, Princeton, NJ, United States. https://www.youtube.com/live/A3I4OSTOQzo?si=rJiq5HxG6wkZyxMW&t=5194

Lundgard, A. (2024, June). *Automatic speech recognition tools for audiovisual media at Stanford Libraries*. [Presentation] AI for Libraries, Archives, and Museums Speech-to-Text Working Group.

Marmor, Y., Lifshitz, Y., Snapir, Y., & Misgav, K. (2025). Building an Accurate Open-Source Hebrew ASR System through Crowdsourcing. *Proceedings of Interspeech 2025*, 723-727. https://doi.org/10.21437/Interspeech.2025

Myntti, J. & Steed, M.R. (2019). Audiovisual accessibility: evaluating workflows for transcribing and captioning digital archive content. *Journal of Digital Media Management*, 8(3), 264-274.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision*. https://doi.org/10.48550/arxiv.2212.04356.

Rao, N., O'Riodan, S., & Coulis, J. (2025). AI and labor: Captioning audiovisual content with Whisper. *IFLA Journal*, 51(3), 803-813. https://doi.org/10.1177/03400352241310534

Rodriguez, D., & Brown, B.J. (2023). Comparative analysis of automated speech recognition technologies for enhanced audiovisual accessibility. *Code4Lib Journal* 58. https://journal.code4lib.org/articles/17820

Appendix

Software configuration:

- ASR Engine: faster-whisper 1.0.0
- Base Model: openai/whisper-large-v3-turbo
- Hebrew Model: ivrit-ai/whisper-large-v3-turbo-ct2
- Yiddish Model: ivrit-ai/yi-whisper-large-v3-turbo-ct2
- Text Processing: spaCy 3.7 (en_core_web_sm)

- Evaluation Framework: Custom Python implementation using difflib

The Fortunoff Video Archive testimonies are available through Yale University Library with appropriate access permissions. Information about accessing the collection is available on their website: https://fortunoff.aviaryplatform.com/.