Automatic Transcription of Holocaust Testimonies in Yiddish: Orthographic Comparison and Cross-Domain Validation

Isaac L. Bleaman
University of California, Berkeley
Department of Linguistics and Center for Jewish Studies
bleaman@berkeley.edu

Abstract

The digital processing of Holocaust testimony interviews is essential for the long-term preservation and accessibility of survivors' narratives. However, automatic speech recognition (ASR) for Yiddish—the primary language of most Holocaust victims and survivors—remains underdeveloped. This paper introduces the first ASR system for European Yiddish, focused on the Northeastern ("Lithuanian") dialect and trained and evaluated on testimony interviews from the Corpus of Spoken Yiddish in Europe (42 hours of speech segments from 60 survivors). A systematic comparison of CTC-based ASR models using transcripts with different orthographic representations reveals that a Hebrew-based phonemic system with precomposed Unicode is optimal, achieving a mean word error rate (WER) of 37.96% compared to 59.40% WER for romanized Yiddish and 99.67% WER (catastrophic failure) for standard Yiddish spelled with decomposed Unicode. Cross-domain testing on Yiddish audiobooks provides additional support for a phonemic representation (27.07% WER, 6.56% CER). Together, the results suggest that automatic transcription developed from oral Holocaust testimonies can support further technological innovation in service of Yiddish-speaking communities.

Keywords: ASR, Yiddish, Holocaust testimonies, orthographic normalization

1. Introduction

Audio- and video-recorded testimony interviews are unparalleled resources for understanding the Holocaust through the firsthand accounts of survivors. The USC Shoah Foundation Visual History Archive holds tens of thousands of digitized interviews with Holocaust survivors, which were collected mostly in the 1990s in locations all around the world and delivered in dozens of languages. Automatic speech recognition (ASR) software has been developed to make the content of many of these testimonies searchable and viewable as subtitles in video players. However, these technological advances have not benefited all languages in equal measure. No such ASR capability exists for Yiddish, the primary language of most Eastern European Jewish communities destroyed in the Holocaust (Birnbaum, 2016, 42), as well as a significant number of recorded survivor testimonies. This constitutes a major barrier for the accessibility of Yiddish-language testimonies, which affects not only historians of the Holocaust but also Yiddish linguists and language learners.

Developing ASR for Yiddish-language testimonies presents both linguistic and technical challenges. Although the language continues to be used in Jewish communities around the world, most of the dialects once spoken across the pre-Holocaust European heartland are now severely endangered, and in some cases underdocumented. Furthermore, Yiddish-speaking survivors lived in highly multilingual environments, both before the Holocaust and in their post-war countries of resettlement. This makes for oral testimonies that are highly complex, in which survivors routinely engage in code-switching (with languages as diverse as Hebrew and Hungarian) as well as dialect mixing. An effective ASR system would therefore need to handle multiple Yiddish dialects and extensive language mixing.

Beyond these challenges, Yiddish presents a great deal of orthographic complexity. While it is traditionally written in a Hebrew-based orthography, Yiddish can also be transliterated into other alphabets (Latin, Cyrillic, etc.). Within Hebrew script, there is a convention in which all words are spelled phonemically unless they come from the so-called "Semitic component" (Hebrew- and Aramaic-origin words), in which case they are spelled according to the norms of those languages (Jacobs, 2005, 48). Additionally, numerous orthographies (not all standardized) have been in use in different times and places, and in today's digital texts, there are also competing Unicode normalization forms. The interaction between orthographic representations and modern neural ASR architectures remains unexplored for Yiddish, yet this choice has significant downstream consequences for model training and performance.

This paper presents a proof-of-concept ASR system for Northeastern Yiddish, also known as *Litvish* 'Lithuanian' Yiddish, a cluster of dialects spoken across a territory that includes present-day Lithuania, Latvia, Belarus, northeastern Poland, and northern and eastern Ukraine (Weinreich, 1963, 337; Jacobs, 2005, 65). Northeastern Yiddish was chosen because standard Yiddish

Proceedings of The Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026, pages 20–28
11 May 2026.

spelling (and standard romanization) is more-or-less phonemically transparent for this dialect. Using audio recordings and transcripts from the Corpus of Spoken Yiddish in Europe (CSYE; Bleaman and Nove, 2025), we investigate which orthographic representations enable effective ASR training. More specifically, this work makes the following contributions:

1. We systematically compare three orthographies for representing Yiddish speech during ASR training and evaluation: romanization (ROM); a standardized Hebrew-based script, with phonemic spellings of all words in decomposed Unicode (STD); and a Hebrew-based representation of Yiddish phonemes in precomposed Unicode (PHON). Our experiments demonstrate that the PHON system—which can readily be back-transformed into a more human-readable standard Yiddish spelling—achieves a much lower word error rate (WER) than the ROM system.

2. We identify a critical incompatibility between STD, which uses Hebrew letters and combining diacritics, and Connectionist Temporal Classification (CTC)-based ASR training. The STD approach fails catastrophically, producing unintelligible output. Given that decomposed Unicode is standard for digital Yiddish text today, this finding has immediate implications for corpus development, and it also extends to other languages that use combining characters to capture phonemic distinctions.

3. We validate the robustness of an ASR system trained on Holocaust testimony interviews through a cross-domain evaluation of Yiddish audiobooks, using a dataset compiled for text-to-speech (TTS) applications (Webber et al., 2022; Bleaman et al., 2023). Despite being trained on spontaneous speech, our ASR system generalizes to read speech as indicated by WER.

4. We provide resources to support future work: a trained ASR model, orthographic preprocessing utilities, and an interactive demo.

The results of this project demonstrate that effective ASR for Yiddish-language testimonies is achievable but highly dependent on specific design choices, which have consequences for other technologies developed for Yiddish-speaking communities. The remainder of this paper is organized as follows: Section 2 describes the testimonies and datasets used for model training and evaluation. Section 3 details the orthographic representations chosen for training. Section 4 presents the model architecture, training configuration, and evaluation methodology. Section 5 reports results and discusses findings and implications. Section 6 concludes with limitations and future directions. Brief statements about data availability and ethical considerations are provided before the reference lists.

2. Data

2.1. The Speech Corpus

The primary data for this project come from the Corpus of Spoken Yiddish in Europe (CSYE; Bleaman and Nove, 2025), a collection of manually transcribed Holocaust survivor testimonies in Yiddish sourced from the USC Shoah Foundation Visual History Archive (VHA). At the time of model training, the corpus contained interviews with 60 speakers of Northeastern Yiddish. Like other testimonies in the VHA, these interviews in Yiddish were conducted by trained volunteers in locations all around the world, and generally proceed chronologically as survivors recount their personal and family histories before, during, and after World War II. In addition to the complexities outlined above related to dialect and language mixing, the conversational nature of these interviews—including overlapping speech between survivor and interviewer, disfluencies including filled pauses, and moments of emotional intensity—presents important ASR challenges that are not typical for read or scripted speech.

The CSYE includes downloadable audio files extracted from VHA video files (digitized video cassettes). These were converted to 16kHz mono WAV format. Transcripts in the CSYE are annotated as *reviewed* or *unreviewed*, reflecting whether or not the transcript for a particular video cassette was reviewed by a member of the CSYE team other than the original transcriber. We included both reviewed and unreviewed segments in our training and testing to maximize coverage.

Aside from speaker diarization, which is the result of a machine learning algorithm and manual correction, all of the transcripts in the CSYE were produced by hand by Yiddish-speaking team members trained in the CSYE transcription conventions. The survivor and interviewer are transcribed on separate time-aligned text tiers in ELAN files (Max Planck Institute for Psycholinguistics, 2021). Transcription conventions are based on standard YIVO transliteration, an orthographic representation in a Latin character set widely used in the Yiddish scholarly community (YIVO, 1999; Bleaman, 2019). CSYE conventions instruct transcribers to faithfully transcribe dialectal vocabulary items, but not to modify spellings to reflect historical sound changes that predictably differenti-

ate the dialects. For example, the written form ⟨beygl⟩ 'bagel(s)' can represent either /beɪgl/ in Northeastern and Southeastern Yiddish or /baɪgl/ in Central Yiddish.[1] Because the corpus was designed (in part) to support research in sociophonetics, the transcripts include faithful representations of partial words, filled pauses, and other disfluencies—elements that are often omitted from ASR output. More information on CSYE transcription methodology is documented in Bleaman and Nove (2025).

2.2. Data Preprocessing

We segmented the audio and transcripts into short phrase-level chunks, based on the segmentation already present in the CSYE. Only speech from the survivors, not the interviewers, was included in our dataset. Because the current project was envisioned to be a proof-of-concept for a Yiddish ASR system, we applied filtering to remove the following speech segments:

- **Overlapping speech:** Segments produced by the survivor that overlapped with segments produced by the interviewer
- **Unclear or misheard words:** Segments containing *UNK* (convention for words that were unintelligible to the transcriber) or angle brackets (convention for uncertain transcriptions)
- **Partial words:** Segments containing any word strings ending in a hyphen (convention for partial words)
- **Fillers:** Segments containing one or more predefined filled pauses (*uh*, *uhm*, *ehm*, etc.)
- **Borrowings:** Segments containing words marked as borrowings (those with 2+ adjacent uppercase letters)
- **Short segments:** Segments shorter than 0.5 seconds

Of the 114,092 total speech segments from Northeastern Yiddish-speaking survivors, 63,346 segments (55.5%) remained after these filtering steps. This corresponds to 42.21 total hours of isolated speech segments.

We then applied several orthography-specific text preprocessing steps. These included whitespace normalization, punctuation removal, replacing remaining hyphens (those used in compounds) with spaces, and various orthography-specific character transformations, which are detailed in Section 3.

[1] In this paper, angle brackets are used to represent orthographic forms. Slashes represent phonemic forms.

Finally, we partitioned speakers (*not* segments or tape transcripts) into a training set (70%: 42 speakers), a validation set (10%: 6 speakers), and a test set (20%: 12 speakers) using a fixed random seed. This ensures that test speakers are completely unseen during model training and can be used to evaluate how well the ASR system generalizes to new voices. A fixed random seed ensures that the same speaker partition is used across all three orthographic representations for a fair comparison.

2.3. Cross-Domain Corpus

For cross-domain validation of an ASR system trained on spontaneous conversational speech, we used the Reading Electronic Yiddish Documents (REYD) corpus of audiobook narrations, which was assembled for a project to create a text-to-speech (TTS) dataset and system for Yiddish (Webber et al., 2022; Bleaman et al., 2023). The dataset consists of short audio segments matched with text files from readings of Yiddish literature, which were taken from two different public repositories: the Yiddish Book Center's Sami Rohr Library of Recorded Yiddish Books, originally recorded in the 1980s and 1990s in Montreal, and the "World of Yiddish" webpage, recorded in the early 2000s at the University of Haifa. From the dataset, we used recordings from the speakers labeled *lit1* and *lit2* (two Northeastern Yiddish-speaking narrators: Sara Blacher-Retter and Leib Rubinov) and the set of utterances labeled as *yivo_respelled* (those written in a Hebrew-based script, with all words spelled phonemically). The entire subcorpus for these two narrators was used for cross-domain testing. For each orthographic model, we transformed the REYD reference texts using the same preprocessing pipeline applied to the CSYE data.

Table 1 provides summary statistics for all datasets used in this study.

3. Orthographic Representations

As mentioned above, Yiddish can be written using multiple orthographic systems, and even within a single system, users can apply various encoding and normalization choices. For ASR development, the choice of orthographic representation constrains the model's output vocabulary and plays an important role in training. In this project, we systematically compare three approaches, which involve two different scripts (Latin-based vs. Hebrew-based) and Unicode normalization strategies, as well as other Yiddish-specific choices that are elaborated on in this section.

Dataset	Speakers	Hours	Segments	Domain
CSYE training	42	30.83	46,432	Testimony (conversational speech)
CSYE validation	6	2.80	3,803	Testimony (conversational speech)
CSYE test	12	8.58	13,111	Testimony (conversational speech)
REYD	2	5.32	3,632*	Audiobooks (read speech)

Table 1: Corpus statistics after preprocessing and data partitioning. All CSYE orthographic representations use identical speaker partitions. *REYD segment counts vary somewhat by orthography, due to the respelling of reference texts and filtering rules.

3.1. Romanization (ROM)

The CSYE transcripts are originally produced in romanized Yiddish adapted from YIVO conventions for transliteration, and this orthographic representation serves as the baseline for our experimentation. After the preprocessing and filtering steps outlined above, we convert the remaining transcribed segments to lowercase. The resulting vocabulary contains 24 characters: 21 letters (<a>–<z> excluding <c j q w x>), word boundary marker, and standard special tokens for padding and unknown characters.

While the ROM system is phonemically transparent, it is not a one-to-one mapping of grapheme to phoneme; many Latin letter combinations correspond to a single phoneme. For example, <tog> 'day' corresponds to /tog/, but diphthongs and certain consonants are represented by multiple characters each, e.g., <boykh> 'stomach' corresponds to /boɪ͡χ/.

3.2. Standard Hebrew-Based with Decomposed Unicode (STD)

To create Hebrew-script representations, we automatically respelled the original romanized transcripts using the `detransliterate()` function from the `yiddish` library (Bleaman, 2024). This uses rule-based pattern matching to convert standard YIVO transliteration into the Hebrew-alphabetic script, without correcting the spelling of words of Semitic origin (i.e., these are spelled phonemically). This output is then fed into the `replace_with_decomposed()` function to represent vowel marks and other *nekudes* ("pointing") as combining diacritics with preceding letter graphemes. Further, the argument `vov_yud=True` is specified to produce a few Yiddish-specific ligatures (<וו וי יי>). With the exception of the phonemic spelling of Semitic-origin words, the output of all of these steps reflects how standard Yiddish is typically encoded in most digital documents today.

While a Hebrew-based orthography addresses *some* of the many-to-one character-to-phoneme mappings of the ROM system—e.g., the consonant /χ/ corresponds to <kh> in ROM but to the singleton grapheme *khof* <כ> in STD—the use of combining diacritics means that an even larger number of sounds are represented by multiple Unicode characters. For example, the consonant /f/ becomes <פֿ>, a two-character sequence consisting of a plain *fey* <פ> followed by the *rofe* diacritic. Additionally, a silent *alef* <א> is required in many vowel-initial words, and five consonant phonemes have special letter forms (distinct Unicode allographs) when they appear in word-final position. For example, the word /oɪ͡χ/ 'also' (romanized as <oykh>) is represented in STD as <אויך>, which begins with a silent *alef* and ends with the word-final allograph of *khof* <כ>.

The vocabulary contains 36 characters: Hebrew base letters (excluding <ח> and <ת>, which only appear in Semitic spellings), the combining diacritics used in standard Yiddish (those seen here: <אַ אָ יִ פּ פֿ>), word-final allographs (<ך ם ן ף ץ>), Yiddish ligatures <וו וי יי>, word boundary, and standard special tokens.

3.3. Phonemic Hebrew with Precomposed Unicode (PHON)

In anticipation of the problems that might arise from the use of decomposed Unicode with a CTC-based ASR system, we created a Hebrew-based phonemic representation in precomposed Unicode characters. This differs from STD in the following ways:

- All letters with combining diacritics, e.g., <אַ> (U+05D0 for *alef* plus U+05B7 for *pasekh*), are replaced with precomposed equivalents from the "Alphabetic Presentation Forms" Unicode block of ligatures, e.g., <אַ> (U+FB2E).
- Silent *alef* letters are removed throughout.
- Word-final allographs are replaced with their nonfinal forms.
- The consonant /j/ and the vowel /i/, which are both (usually) represented in STD by the letter *yud* <י>, are distinguished in PHON: <י> for the consonant and <יִ> (U+FB1D) for the vowel.

- The vowel /u/, which is represented in STD either as ‹ו› or as ‹וּ› depending on context, is consistently represented in PHON as ‹וּ› (U+FB35).

The result of these modifications is a system of one-to-one mappings between Unicode characters and phonemes. While this differs from the way Yiddish is customarily represented in digital text, it can readily be converted back to a familiar, human-readable format by performing "roundtrip" transliteration, i.e., by first applying the `yiddish` library's `transliterate()` function and then the `detransliterate()` function.

The vocabulary for PHON contains 30 characters: several Hebrew base letters (‹ב ג ד ה ז ט י כ ל מ נ ס ע צ ק ר ש›), Yiddish ligatures (‹וו וי יי›), Hebrew letters precomposed with diacritics (‹אַ אָ אּ בּ וּ יִ פּ פֿ›), word boundary, and standard special tokens.

A comparison of orthographic representations for an example Yiddish utterance from the CSYE is provided in Table 2.

4. Experimental Setup

4.1. Model Architecture

We built upon Wav2Vec-BERT 2.0 (Seamless Communication et al., 2023), a large-scale self-supervised speech encoder developed by Meta, which was pretrained on over 4.5 million hours of unlabeled multilingual audio. The model uses convolutional acoustic feature extraction followed by a conformer-based encoder, producing high-dimensional speech embeddings that generalize across languages and downstream tasks. To adapt the model to Yiddish, we followed the fine-tuning procedure of Lacombe (2024), which covers dataset preparation, CTC-based vocabulary construction, processor setup, and CTC-based fine-tuning, with a convolutional adapter layer added.

The base model contains approximately 600 million parameters. We initialized from the publicly available *facebook/w2v-bert-2.0* checkpoint and fine-tuned all parameters on the CSYE dataset described above. The CTC training objective assumes a monotonic alignment between acoustic frames and output characters, a constraint that becomes critical for evaluating different orthographic representations (see Section 5).

4.2. Training Configuration

All models were trained with identical hyperparameters (Table 3), adapted from Lacombe (2024) and held constant across all three orthographies. No data augmentation (e.g., noise injection, speed perturbation) was applied during fine-tuning. We fine-tuned for up to 10 epochs with early stopping based on validation WER. Models were evaluated every 300 steps, with training stopping if no improvement occurred for 3 consecutive evaluations. Training used mixed-precision (FP16) and gradient checkpointing to reduce memory consumption. Total training time ranged from 1.3 to 1.7 hours on a Tesla T4 GPU.

4.3. Evaluation Metrics

We report word error rate (WER) and character error rate (CER) on the held-out test set. WER measures the edit distance between predicted and reference word sequences normalized by the number of reference words:

$$\text{WER} = \frac{\text{Substitutions} + \text{Deletions} + \text{Insertions}}{\text{Total Reference Words}} \tag{1}$$

CER applies the same metric at the character level. Unlike WER, which disproportionately penalizes misplaced whitespace, CER provides a more forgiving measure of accuracy by treating word-boundary errors as single-character edits rather than complete word-level mismatches. While WER is the standard metric, and the one used for evaluation during training, we report both in Section 5.

Initial experiments with a single random seed (42) across all three orthographies identified PHON as best-performing. We therefore trained PHON with four additional random seeds (43–46) to quantify initialization variance and establish statistical confidence. ROM (the baseline orthography) and STD (which failed to produce intelligible output; see below) were trained just once.

Therefore, for the PHON models, we report mean ± standard deviation across the five random seeds, but report a single value for ROM and STD. Cross-domain evaluation using the REYD audiobook corpus reports the same metrics, where the reference texts undergo the same orthographic transformations defined above in Section 3.

5. Results and Discussion

5.1. In-Domain Performance (CSYE)

Table 4 presents our main results on the CSYE test set. The phonemic Hebrew representation (PHON) achieved the best performance with a mean 37.96% WER, representing a 36.1% relative improvement over the romanization baseline (59.40% WER). This substantial improvement is consistent with the hypothesis that one-to-one mapping between phonemes and graphemes optimizes CTC training for Yiddish ASR.

Orthography	Alphabet	Example
Original (from the CSYE)	Latin	mayn familye-nomen iz Dimantshteyn
ROM (romanized)	Latin	mayn familye nomen iz dimantshteyn
STD (standard Hebrew-based)	Hebrew, decomposed	מײַן פֿאַמיליע נאָמען איז דימאַנטשטיין
PHON (phonemic Hebrew-based)	Hebrew, precomposed	מײַנ פֿאַמיִליע נאָמע֫נ יִז דִימאַנטשטיינ

Table 2: An utterance from the Corpus of Spoken Yiddish in Europe, as represented in the training data for each orthographic system after preprocessing. The utterance comes from the testimony of Holocaust survivor Aizik Dimantstein (1996).

Parameter	Value
Base model	w2v-BERT 2.0
Optimizer	AdamW
Effective batch size	32
Learning rate	$5x10^{-5}$
LR scheduler	Cosine
Warmup steps	1,000
Max epochs	10
Evaluation frequency	Every 300 steps
Early stopping patience	3 evaluations
Precision	FP16

Table 3: Training hyperparameters. All models use identical speaker-based data splits. Five PHON models were trained with different random seeds; ROM and STD each trained with a single random seed.

Model	WER (%)	CER (%)
ROM	59.40	18.73
STD	99.67	93.21
PHON	37.96 ± 0.77	13.39 ± 0.46

Table 4: Test set performance on CSYE (13,111 segments from 12 unseen speakers). PHON results show mean and standard deviation across five random seeds.

The five PHON models showed consistent performance across random seeds, with WER ranging from 37.22% (seed 44) to 39.14% (seed 45). The small standard deviation indicates that our results are robust to initialization variance. All five seeds achieve substantial improvements over the ROM baseline.

The ROM model established a baseline to demonstrate that ASR for Yiddish testimonies is feasible even with the default Latin-script representations from the CSYE. However, the Hebrew orthography with decomposed Unicode (STD) catastrophically failed, with 99.67% WER and 93.21% CER—producing text largely consisting of isolated diacritics, repeated characters, and empty strings.

This failure presumably stems from the use of a decomposed Unicode character set for Yiddish. CTC assumes a roughly monotonic alignment between acoustic frames and output tokens, but decomposition often splits single phonemes like /a/ into multiple code points (a letter plus a combining diacritic). In such cases, the model must predict multiple sequential tokens for essentially the same acoustic span. Other factors, such as the use of silent *alef*s or multiple allographs to represent the same phoneme (e.g., word-initial and -medial ‹כ› vs. word-final ‹ך›), could increase variability in the target sequence, but they do not fundamentally contradict the temporal alignment assumptions in the same way as decomposition.

Model	WER (%)	CER (%)
ROM	51.84	10.53
STD	98.98	84.17
PHON	27.07 ± 2.99	6.56 ± 0.69

Table 5: Cross-domain performance on REYD audiobooks (2 speakers, 5.32 hours). PHON results show mean ± standard deviation across five random seeds.

5.2. Cross-Domain Performance (REYD)

Table 5 summarizes the results of applying the ASR models trained on transcribed Holocaust testimonies to the REYD audiobook corpus, and Table 6 provides example ASR outputs. Remarkably, the PHON models achieve better performance on the REYD dataset (27.07% WER) than on the CSYE test set (37.96% WER), demonstrating robust cross-domain generalization. This improvement likely reflects the slower, more careful speaking style of audiobook narration compared to the spontaneous speech of testimony interviews. Another relevant factor may be the reduced background noise variability in the audiobooks, which were recorded in a studio environment rather than in the speakers' homes. In any event, the cross-domain improvement suggests that the model has learned generalizable acoustic phonetic patterns of Northeastern Yiddish rather than testimony-specific characteristics.

The ROM baseline also improved on REYD (51.84% WER vs. 59.40% on the CSYE), while STD's failure remained consistent across domains (98.98% WER on REYD and 99.67% on CSYE). For example, the STD model produced 537 completely empty predictions out of 3,632 REYD utter-

ances, further confirming that its failure is systematic rather than dataset-specific.

The strong cross-domain performance suggests that PHON-based models can be applied to diverse Yiddish audio collections beyond Holocaust testimonies. The WER on REYD approaches the performance levels where ASR can become practically useful for searching and information extraction—all the more so if the output is manually corrected.

6. Conclusion, Limitations, and Future Directions

Holocaust testimony archives hold thousands of hours of interviews in Yiddish, yet these recordings remain largely unsearchable and inaccessible for large-scale analysis. We address this barrier by developing the first automatic speech recognition system for Northeastern Yiddish, trained on transcribed survivor testimonies from the Corpus of Spoken Yiddish in Europe.

Our phonemic Hebrew orthography (PHON) achieves a mean WER of 37.96% on conversational testimony speech, a large improvement over a romanized baseline. Critically, we identified that decomposed Unicode—commonly used in digital Yiddish text—fails in CTC-based ASR, which underscores the importance of normalization for both corpus creation and downstream applications. This finding extends beyond Yiddish to other languages that use combining diacritics for phonemic distinctions reflected in spelling.

Several important limitations should be noted. Our data preprocessing removed a large portion of the original transcribed speech segments through aggressive filtering for speaker overlap, disfluencies, borrowings, and code-switches. While this filtering was done to maximize the success of ASR training, it means our models are optimized for clean, single-speaker utterances rather than naturalistic testimony speech. Interviews with Holocaust survivors frequently contain overlapping speech between the survivor and the interviewer, false starts and hesitations, and language mixing; it is not yet known how an ASR system that includes such segments would perform in training or evaluation, and thus whether it would be suitable for production deployment in archival settings.

Additionally, we trained exclusively on Northeastern Yiddish, represented by 60 of the currently available 158 speakers in the CSYE. A natural extension of the current project would strive for inclusion of all three broad dialects of Eastern Yiddish: Northeastern ("Lithuanian"), Central ("Polish"), and Southeastern ("Ukrainian"). Improved dialect coverage would likely make this ASR system much more useful across archives of Yiddish-language testimonies, and other collections of Yiddish speech. One potential future direction would be to combine dialect identification with dialect-specific ASR: an initial classifier would first identify the speaker's regional variety and then route the audio to a specialized model trained on that dialect. This pipeline would provide a unified interface for a "whole language" ASR model for Yiddish.

Several other technical improvements could enhance performance. Integrating language models trained on text corpora—a standard approach in production systems—could substantially improve ASR performance for Yiddish. Even for the best phonemic models tested on clean audiobook data, 27.07% WER remains too high for production transcription without a significant amount of correction. Repositories such as the Yiddish Book Center's digital library (Yiddish Book Center, 2022) provide large-scale text data suitable for training language models aimed at correcting phonetically plausible but lexically invalid outputs. Additionally, data augmentation techniques such as noise injection could improve robustness to different datasets and recording conditions.

7. Acknowledgments

I gratefully acknowledge the USC Shoah Foundation – The Institute for Visual History and Education for its support of this project.

Thank you to Jacob J. Webber for sharing a tutorial that helped me debug my training code, and to Elan Rosenfeld for comments on the paper.

This material is based upon work supported by the National Science Foundation under Award No. BCS-2142797.

8. Data Availability and Ethical Considerations

The transcripts and audio recordings from the Corpus of Spoken Yiddish in Europe (CSYE) are available to the public online. Users of the corpus must abide by the USC Shoah Foundation Terms of Use as well as the CSYE Terms of Use, both of which are provided in the CSYE User Guide. The dataset compiled for the Reading Electronic Yiddish Documents (REYD) text-to-speech project, a subset of which was used for cross-domain testing, is also available online. See the Language Resource References section below for more details.

A phonemic ASR model is released at https://huggingface.co/ibleaman/w2v-bert-2.0-yiddish-northeastern for non-commercial research and educational purposes only. The model card also includes links to notebooks containing orthographic preprocessing functions and an interactive ASR demo.

Original utterance:	זי הייסט: "די רוסישע רעוואָלוציע אין קייזערלעכן הויף"
Translation:	It [the new play] is called: “The Russian Revolution in the Imperial Court”
ROM	
Reference:	zi heyst di rusishe revolutsye in keyzerlekhn hoyf
Prediction:	zi heystdi rusisherevolutsiyein keyzerlekh un hoyf
STD	
Reference:	זי הייסט די רוסישע רעוואָלוציע אין קייזערלעכן הויף
Prediction:	ע א
PHON	
Reference:	זִי הייסט דִי רוּסִישע רעוואָלוּציע יִן קייזערלעכן הויפֿ
PHON (seed 42)	
Prediction:	זִי הייסט דִי רוּסִישע רעוואָלוּציע יִן קייזערלעכן האָויפֿ
PHON (seed 43)	
Prediction:	זִי הייסט דִי רוּסִישע רעוואָלוּצע יִן קייזערלעכן הויפֿ
PHON (seed 44)	
Prediction:	זִי הייסט דִי רוּסִישע רעוואָלוּציע יִן קייזערלעכן הויפֿ
PHON (seed 45)	
Prediction:	זיי הייסטדִי רוּסִישע רעוואָלוּציע יִן קייזערלעכ ן הויפֿ
PHON (seed 46)	
Prediction:	זִי הייסט דִי רוּסִישע רעוואָלוּציע יִן קייזערלעכן הויפֿ

Table 6: Predictions for an example utterance from the REYD test set across all models and seeds. ROM primarily shows word boundary errors, STD produces unintelligible output, and PHON models achieve accurate transcription for seeds 44 and 46 with others showing minor errors.

The survivor testimonies used in this project were sourced from the USC Shoah Foundation VHA and licensed for inclusion in the CSYE. Due to the sensitive nature of testimony data, all CSYE transcripts are produced by hand with utmost care to ensure the texts are accurate and reliable. Users of the ASR model should be aware of its performance limitations and verify the accuracy of all generated transcripts against the original audio.

9. Bibliographical References

References

Solomon A. Birnbaum. 2016. *Yiddish: A Survey and a Grammar*, 2nd edition. University of Toronto Press, Toronto. Originally published in 1979.

Isaac L. Bleaman. 2019. Guidelines for Yiddish in bibliographies: A supplement to YIVO transliteration. *In geveb*.

Isaac L. Bleaman. 2024. `yiddish`: A Python library for processing Yiddish text.

Aizik Dimantstein. 1996. Interview 20327. *USC Shoah Foundation Visual History Archive*. USC Shoah Foundation. Accessed April 1, 2026.

Neil G. Jacobs. 2005. *Yiddish: A Linguistic Introduction*. Cambridge University Press, Cambridge.

Yoach Lacombe. 2024. Fine-tune w2v2-BERT for low-resource ASR with Transformers. Hugging Face, Community Blog & Articles.

Max Planck Institute for Psycholinguistics. 2021. ELAN [computer program].

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual expressive and streaming speech translation. arXiv.

Uriel Weinreich. 1963. Four riddles in bilingual dialectology. In *American Contributions to the Fifth International Congress of Slavists, Sofia, September 1963*, volume 1: *Linguistic Contributions*, pages 335–359. Mouton, The Hague.

Yiddish Book Center. 2022. Steven Spielberg digital Yiddish library.

YIVO (Yidisher visnshaftlekher institut). 1999. *Der eynheytlekher yidisher oysleyg [The Standardized Yiddish Orthography]*, 6th edition. YIVO Institute for Jewish Research and League for Yiddish, New York.

10. Language Resource References

Isaac L. Bleaman and Chaya R. Nove. 2025. The Corpus of Spoken Yiddish in Europe: Goals, methods, and applications. *Language Documentation & Conservation*, 19:142–157. Corpus available online: https://www.yiddishcorpus.org/csye.

Isaac L. Bleaman, Jacob J. Webber, and Samuel K. Lo. 2023. Speech synthesis in the “mother tongue”: Designing, training, and evaluating a text-to-speech system for Yiddish. *Journal of Jewish Languages*, 11(1):15–43. Dataset available at: https://github.com/REYD-TTS.

Jacob J. Webber, Samuel K. Lo, and Isaac L. Bleaman. 2022. REYD – The first Yiddish text-to-speech dataset and system. In *Proceedings of Interspeech 2022*. Dataset available at: https://github.com/REYD-TTS.