Using collection context to reduce errors in automated transcription of historical audio

Internet History Initiative · practice note · 2026-07-23

Background

We recently accessioned the Internet Talk Radio collection (Carl Malamud, 1993–1996; 567 audio items mirrored from archive.org). To make the audio searchable and citable we need transcripts. We ran a single-episode pilot to test a transcription workflow before processing the full collection, using the first “Geek of the Week” episode: an interview with Marshall T. Rose recorded March 31, 1993 (48 minutes).

The error pattern

Automated speech recognition handles ordinary speech in this material well, but systematically mis-renders the specialized vocabulary of the period: acronyms, protocol names, product names, and personal names. The model substitutes phonetically plausible guesses. From the raw output of this episode:

- “ISO-DE” for ISODE (a software package the guest authored)
- “the IEPF” for the IETF (six occurrences)
- “X400” for X.400 (six occurrences)
- “Marty Shostal” for Marty Schoffstall
- “Keith McLaughery” for Keith McCloghrie

Name errors are the most serious for our purposes, since names are access points. These errors are not correctable by a general-purpose model working alone: nothing in the audio distinguishes “McLaughery” from “McCloghrie” without outside knowledge of who worked with the guest.

Method

The workflow has two stages. Both use information already held in the collection.

Stage 1: speech recognition with a vocabulary prompt. We transcribed the audio with Whisper (large-v3 model, one GPU). Whisper accepts a short free-text prompt that biases its vocabulary. We filled it with the episode metadata (show, guest, interviewer, date) and roughly forty period terms. This reduces, but does not eliminate, the errors above. The output — 741 timestamped segments — is saved unchanged as the raw transcript.

Stage 2: constrained correction using a brief derived from the collection. We then passed the transcript in chunks to a language model (Claude Sonnet) with a correction brief and narrow instructions: correct only mis-transcribed technical terms, acronyms, and proper nouns that the brief supports; do not rewrite, paraphrase, or improve the speech; do not alter timestamps.

The brief (about 1,100 characters) was assembled from two collection sources:

1. *The guest's authority record* from our catalog, which lists his projects, publications, employers, and collaborators. This is what allows “Shostal” to be corrected to Schoffstall (a colleague at the guest's employer, PSI) and “McLaughery” to McCloghrie (a co-author of the SNMP standards).
2. *Vocabulary mined from contemporaneous holdings.* We extracted the most frequent capitalized terms and acronyms from the subject lines of our 1992–1994 mailing-list corpora (tcp-ip, com-priv). This produces the working vocabulary of the period as its participants actually wrote it, with no manual glossary-building.

The corrected output is saved as a second file alongside the raw one. Both are kept permanently: the raw transcript is the preservation copy; the corrected transcript is the access copy; a line-by-line comparison of the two shows every change made.

Results

The correction stage changed 41 of 741 segments. Examples, with timestamps into the recording:

Time	Raw output	Corrected
0:56	the author of ISO-DE	the author of ISODE
6:02	Marty Shostal of PSI	Marty Schoffstall of PSI
7:33	X400 (and 5 further occurrences)	X.400
17:15	a call from the IEPF (and 5 further)	a call from the IETF
17:42	Keith McLaughery	Keith McCloghrie

We reviewed the full diff. All 41 changes were within the instructed scope (terms, acronyms, proper nouns); none rewrote the wording of the speech.

Processing cost for the episode: about 9 minutes of GPU time for transcription and 8 language-model calls for correction. Extrapolated to the full 567-episode collection: roughly 10 GPU-hours and a few thousand model calls, which is within our normal batch budgets.

Limitations

- The correction stage can only fix what the brief covers. A name absent from the catalog and the mailing-list corpora will stay as the model heard it. We do not ask the model to guess beyond the brief.
- Errors outside the vocabulary class (mis-heard ordinary words, speaker attribution, overlapping speech) are untouched by this workflow.
- This is a one-episode pilot. Error rates will vary with audio quality and speaker; we will re-check a sample as the batch runs.
- The corrected transcript remains a machine product and is labeled as such in our provenance fields, at the same “receipt” tier we use for other machine-generated finding aids. It is not a human-verified transcript.

Adapting this elsewhere

The workflow assumes nothing specific to internet history. It requires: (a) an ASR system that accepts a vocabulary prompt; (b) a language model that will follow narrow correction instructions; and (c) collection materials contemporaneous with the recordings — correspondence, minutes, newsletters, an authority file — from which to build the brief. For archives that hold such materials, the marginal effort per recording is small: the brief for this episode was assembled automatically from existing records.

Contact: the IHI reading room (history.cooperate.social). Tooling: Whisper large-v3 with initial_prompt; Claude Sonnet for correction; brief assembly scripted against our catalog API and message index.