Transcription of Audio and Video with OpenAI's Whisper

By Margarete Wilkey and Geoffrey D. Wood

AI tools are new, and as the buzzword technology of our time, deserving of particularly careful consideration before implementation. Whisper AI (https://openai.com/index/whisper/), an open-source automatic speech recognition (ASR) utility, has proved a hugely popular solution for automated transcription of both audio and video. In late 2024, the Digital Initiatives team at Princeton Theological Seminary's Wright Library turned to Whisper to add transcriptions to streaming resources in our freely accessible digital library, Theological Commons (https://commons.ptsem.edu).

At the time of this writing, the Digital team comprises five full-time staff members, three of whom have been directly involved in transcription development. Team members possess a range of technical expertise, including cataloging, MARC- and non-MARC metadata, XML, JSON, XSLT, JavaScript, and web development more generally. They built Theological Commons in-house using the XML-querying language XQuery on MarkLogic Server. This combination of technical and bibliographic knowledge allows us to make informed decisions and address systemic problems readily.

In this article, we trace our institutional needs and history of transcription efforts, describe our present, unique deployment of Whisper, and offer insights and suggestions for institutions considering similar projects.

DIGITIZATION AND EARLY TRANSCRIPTION EFFORTS

Audio and video digitization began at Princeton Theological Seminary (PTSEM) in 2013, funded by a grant from the Henry Luce Foundation. Since 1940, PTSEM has maintained institutional recordings spanning from sermons to acclaimed conferences, the lion's share of which existed only on obsolete, deteriorating physical media such as reel-to-reel tapes. Digitization was performed by a specialty media restoration company, followed by audio quality and metadata checks for each sound file, which necessitated the hiring of temporary employees to perform initial review. To facilitate this work, we built a custom, user-friendly, web-based editor to enable staff of different skill levels to efficiently update the underlying XML without needing to work directly with the raw markup. This two-and-a-half-year process yielded 6,129 audio recordings and nineteen video programs from the Princeton Theological Seminary Media Archive.

Our first attempt at transcription came in 2014 when we joined a pilot project with a start-up that was working on their own speech-to-text software. While the service promised high-accuracy and time-stamped transcripts, the results were inconsistent and frequently inaccurate, especially with theological terms. For an idea of quality, "Paul" was frequently transcribed as "ball." Just over one thousand transcripts were produced using this service, but due to limited staff capacity, we did not include them in our database.

Margarete Wilkey is Metadata Librarian, Wright Library, Princeton Theological Seminary. Geoffrey D. Wood is Digital Library Technologist, Wright Library, Princeton Theological Seminary.

A more fruitful partnership began in late 2015 with a professional manual transcription service which claimed to produce "greater than 99% accuracy" and employed subject-trained specialists for transcription work. This collaboration resulted in 2,175 high quality transcripts that are publicly available in Theological Commons. High costs were prohibitive, so we only completed this single group of transcripts with our remaining grant funds.

Fast forward to August 2024 when the team began exploring new transcription solutions. Staff conducted research into commercial software services such as Trint (https://app.trint.com/) and Otter (https://otter.ai/) but none were selected due to high subscription fees, lack of local oversight, upload caps, and language restrictions. We came away from this evaluation phase with a clarified set of criteria: the solution must be free or very low-cost, ideally open source, and capable of working in tandem with our existing technologies.

EVALUATING WHISPER

Around this same time, we started to hear buzz in the library world around an open-source speech recognition system from OpenAI called Whisper. Whisper is a powerful tool that is well-documented, can automatically detect and process fifty-two different languages (obviating the need for staff with specialized language skills) and offers flexible models allowing users to choose the best accommodations between accuracy and speed.

Whisper was installed on local workstations in September 2024. The installation process required brief support from IT but was relatively painless. The team tested Whisper across a range of audio recordings, including accented lecturers, multilingual presentations, and mixed-content audio, using the various processing models to compare speed and accuracy. A pilot collection of fifty-one audio recordings from the *Black Theology and Leadership Institute* was chosen to evaluate Whisper's performance. Whisper's medium model was ultimately chosen because it offered the best balance of accuracy and efficiency.

WORKFLOW DEVELOPMENT

Between October 2024 and March 2025, we iteratively built a workflow to integrate Whisper-generated transcripts into our existing MODS XML records. We began this process from a position of advantage, as MODS XML records were already in place for each audio file and the over two thousand human-generated transcripts already in the database provided a model for XML structure and required fields. While Whisper does not output transcripts directly in XML, we were able to select JSON as the output format, which maps cleanly to XML, simplifying the data transformation process. Drawing on our existing XML framework and technical expertise in XQuery, we wrote custom code to take Whisper's raw JSON output and transform it into XML, allowing the data to integrate seamlessly with our existing records. While having this particular technology stack already in place was a major advantage, similar results can be achieved with other tools and technical expertise depending on an institution's needs and resources.

Once the basic framework was in place, we extended the existing XML structure with additional attributes to be able to differentiate between transcripts that were "human-generated" vs. those that were "machine-generated" and flag those transcripts that had gone through the full manual review process.

To further refine Whisper's output, we wrote additional post-processing XQuery code to customize and correct the transcripts. Because Whisper does not automatically censor profanity, we implemented a dictionary-based text normalization process to identify and replace common profane words. For consistency across all transcripts and overall usability, we needed transcripts to be segmented into thirty second increments, but Whisper offers only two time-stamping options: random segment breaks or timestamps for each individual word. We instructed Whisper to timestamp each individual word and then, using XQuery, grouped words into sets of eighty (which is the approximate number of words spoken in thirty seconds), taking the timestamp of the first word as the start and the last word as the end. Additional code was then written to reformat these timestamps into HH:MM:SS format for easier readability. Finally, a post-processing XQuery script was written to identify and correct common misspellings with the goal of cutting down on manual review.

Although Whisper produces transcripts with relatively high accuracy, it is far from perfect. Even so, it supports our straightforward goals: rather than striving for publication-quality transcripts, we aim simply to make our audio and video resources not only accessible to users who are deaf or hard of hearing but also full-text searchable for all users. In this regard, ASR does for audio and video what optical character recognition does for printed text. Though we know that the degree of human review necessary to create textually perfect transcription is impossible, each transcript still needs to be reviewed manually by staff before being made public. Given the volume of transcripts, the process needs to be as streamlined and as low-tech as possible to allow multiple editors of varying technical expertise to work simultaneously. To meet this need, we developed a custom, web-based tool that allows staff to edit the transcript using a third-party application called OTranscribe. This free, online transcript editor uses a proprietary file format (.otr) designed to save the user's progress while preserving timestamps. Because of this, once editing is complete, the .otr file needs to be downloaded from OTranscribe and uploaded back into our custom transcript tool, where it is transformed back into XML using XSLT under the hood. Although the process involves several steps, the underlying logic is efficient and reliable. Two student workers were trained to carry out the manual editing process using OTranscribe and the custom web-based tool without issue.

In March 2025, we were able to expand the Digital team with the hiring of a new full-time staff member. With this new addition, the process could be horizontally scaled with four to five computers running Whisper simultaneously on a given workday. This combination of human and technological capacity enabled the Digital team to implement Whisper efficiently while maintaining oversight at every stage.

PRACTICAL INSIGHTS

In roughly eight months of steady transcription work, we generated (prior to assessment/ proofreading) approximately 2,876 transcripts. Of those, 1,435 are part of a collection of over six decades worth of lectures and sermons by a beloved minister who donated his papers and recordings to PTSEM. Known for his quick speech, repeated themes and motifs, and regular discussion of fifty or so Greek, Hebrew, and Aramaic terms, his recordings vary in quality, include small to large gaps of silence, and frequently include long sections of interstitial music.

In its large version 3 model, Whisper claims an average WER/CER (word error rate/character error rate) of between 9.3 percent and 4.1 percent for English, 13.7 percent and 10.9 percent for Greek, and 23.5 percent and 26.1 percent for Hebrew (languages that are commonly sprinkled throughout many of our recordings). By and large, we found these percentages to be accurate in the collection discussed above, even in the medium multilingual model, though performance varies with silence, interstitial

music, and multilingual content. Certain choices that Whisper makes—such as choosing whether to transcribe numbers alphabetically or numerically—seem haphazard, but not enough to challenge basic readability. Whisper does work well over longer transcripts if audio remains consistent, but the longer the recording, the more potential for drift, inconsistent capitalization or punctuation, or missing content. We regularly feed it three- to four- hour lectures with only minor inaccuracies. It also seems to work effectively with low-volume audio, often accurately transcribing audio that strains our ears to understand, though poorly recorded or distorted audio can reduce accuracy.

One of our most pleasant discoveries is that Whisper does not require cutting-edge hardware to run locally. Exact hardware requirements are flexible, and OpenAI's GitHub gives the necessary minimum RAM per model. Our initial work was on three relatively new Mac laptops running Apple M1 through M3 chips, 16 GB RAM, and Sequoia 15; eventually we also incorporated two largely unused student-worker PCs running Intel Core i5 chips, 8 GB RAM, and Windows 10 Pro. Whisper runs equally reliably, albeit at different speeds, on both groups of machines, allowing us to run an average of ten transcripts a day on each Mac and five a day on each of the older PCs. While this system makes the quantity of output unpredictable based on competing computing needs, it has allowed us to generate a sizable number of transcripts of equivalent quality. Of note, we have only worked with the original Whisper model, but there are other efficient, compressed models available like Turbo, which works well on English-only recordings, and Hugging Face's Lite-Whisper, which may work better in scenarios with limited staff and computing resources.

However, these strengths are offset by notable limitations. While the 2,786 transcripts mentioned earlier look very nice, especially considering that Whisper runs in the background after only a few minutes of daily setup, there are practical caveats. For one, this does entail machines running nearly twenty-four hours a day, seven days a week. We have found that heat does build up more than average in some computers (and laptops in particular) running round the clock; in hopes of being better stewards of our institution-supplied computers, we chose to purchase cooling pads which run external fans under the machines.

Total transcription failures are a small, but persistent problem. In our sample set, fewer than 4 percent have come back empty or so near to empty as to be unusable. There are instances where Whisper will "hallucinate" or randomly generate text that is not heard in the recording. Other times, Whisper will fail to generate any text at all when a recording begins with an extended period of non-speech, whether it be silence, music, or ambient noise. When Whisper is unable to recover, it may output an entire file of only punctuation (typically dots) or, interestingly, Whisper may default to a form of unintelligible Welsh if it is not able to identify the language at the start of a recording. Errors like these can sometimes be corrected by specifying the language of the audio in the initial Whisper instructions or by trimming non-speech audio from the start of a recording, but there are still some instances where transcription will continue to fail regardless.

Another related problem is that Whisper can begin to lose its capitalization and punctuation over time. Whisper operates in thirty second "chunks" and mistakes in one chunk tend to carry over to the following chunks. We have not discovered any specific cause for this error, but once one thirty-second chunk has dropped either its capitalization or punctuation, all following chunks will follow suit. Moreover, the one absence—capitalization or punctuation—seems to encourage the other, so in most cases the transcript will lose both.

Some challenges arise not from Whisper itself, but from the distinctive qualities of our recordings. Interstitial music, common in worship services and holiday recordings, often causes Whisper to either mis-transcribe the sung words or ignore any spoken words during the remainder of that thirty

second chunk, leading to elisions of as many as several sentences in a transcript. On rare occasions, Whisper will default to songs on which it was originally trained; "The Star-Spangled Banner" is a favorite replacement. Our solution is to replace the sung text with a simple bracketed placeholder "[music]" during the final proofreading phase. Proper names have also proven to be frequently misspelled, especially depending on the accent of the speaker. Examples within our specific collection include "T Lick" for Thielicke, "Tyre Tyre" for Thyatira, and "Bucks the Hooty" for Buxtehude. While these misspellings are usually innocuous, you may need to keep an eye out for more problematic ones: "Presbyterian" often turns into "Predatorian" in our transcripts, which could cause concern for our users.

Whisper claims that users can use prompting as a workaround to help guide the model and reduce misspellings. We've experimented with using initial_prompt in our Whisper command line instructions by specifying exemplary text to "teach" Whisper the pattern of text we are expecting, hoping to pre-correct frequent misspellings, and create predictive patterns for punctuation and sentence structure, but our testing of this technique yielded inconsistent results at best.

REFLECTIONS

In aggregate, we have been very satisfied with Whisper's performance. It has allowed us to affordably and quickly step up our transcription process in a way nothing before it has. But in the 2,876 transcripts we've generated so far, we've encountered several notable pitfalls: a percentage of transcripts will fail to generate for different reasons; additional post-processing scripts must be folded in afterward; and every produced transcript will require manual proofreading. Whether in the future even more reliable and widely trained versions of Whisper, or all-in-one transcription utilities using Whisper as their base, will become available is unknown, but Whisper is to date the most valuable and cost-saving ASR utility we have seen, and worth the consideration of most libraries.