How to Use AI Transcription Services with MAXQDA

Matthew Loxton

Abstract: The advent of machine learning represents a dramatic change in the effectiveness and efficiency of voice-to-text software. The same machine-learning technology can be used to transcribe interviews. I reviewed 3 AI transcription services and used one of them throughout a current MAXQDA 2018 project.

Keywords: AI, Automatic speaker change, Computer Assisted Qualitative Analysis, functions, interview data, Machine Learning Technology, Multimedia Browser, transcribe interviews, voice-to-text

Table of Contents

1. Introduction

For some people, such as myself, MAXQDA's Transcription Mode tools are a wonderful help. The software is not, however, able to make us efficient in transcribing audio files. Even with the many functions available in MAXQDA's Multimedia Browser, such as keyboard shortcuts and automatic speaker changes, I am still far too clumsy with hitting F4, the pause button, or the various other buttons, to transcribe the audio efficiently.

My typing is too slow, my working-memory too short, and I get confused between the function keys and buttons. As a result, transcribing a 30-minute audio file takes me several hours, and by the end, I am exhausted.

2. How MAXQDA's Transcription Mode Feature Can Be Used in Conjunction with Machine Learning Voice-to-Text Services

The advent of machine learning represents a dramatic change in the effectiveness and efficiency of voice-to-text software. Where previously, voice-activated systems and voice-to-text applications were stuck at low levels of accuracy and speed, the use of machine learning has resulted in a variety of devices and services that boast of 95% and higher accuracy.

So much so, that Siri, Alexa, Cortana, and Google Home are now part of life for millions of people. Google Duplex takes this further and is almost indistinguishable from a human assistant.

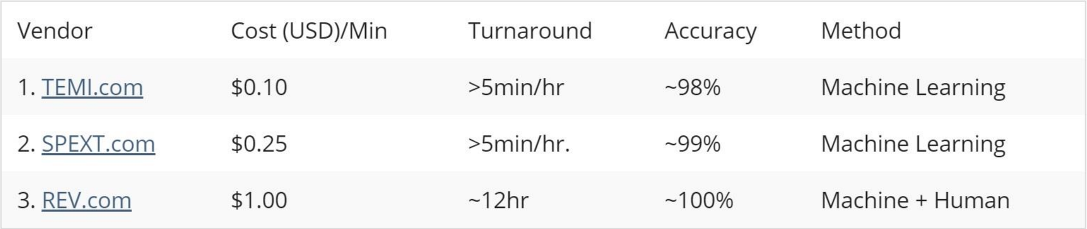
Transcribing interviews

The same machine-learning technology can be used to transcribe interviews. I reviewed three different options and services, ultimately using one of them throughout a current project.

My aim was to speed up my transcription process, and to remove the frustration that normally accompanies this step in the research process for me. I recorded nine interviews, which ranged from 18-48 minutes, and in which there was one participant, the primary interviewer, and myself.

Three voice to text services

Based on some research on popular services, the three services that I looked at were:



Vendor	Cost (USD)/Min	Turnaround	Accuracy	Method
1. TEMI.com	$0.10	>5min/hr	~98%	Machine Learning
2. SPEXT.com	$0.25	>5min/hr.	~99%	Machine Learning
3. REV.com	$1.00	~12hr	~100%	Machine + Human

I found that approximately 1-2% of the participant's text from the two Machine transcriptions needed edits, typically for similar-sounding words such as "IV" vs "IB", or "wake" vs "wait". I also needed to change the speaker names manually. I had higher error rates in the interviewer text, but this was not of significance to my needs.

The low cost and almost immediate turnaround of the machine transcriptions meant that for just over $20 in total, I could have very workable transcriptions almost ready to code within minutes for nine of the half-hour interviews.

I found that making the few edits required was far less frustrating and time-consuming than manual transcription had been in the past. This is obviously a function of my typing speed and keyboard dexterity, but may be similar for many other researchers.

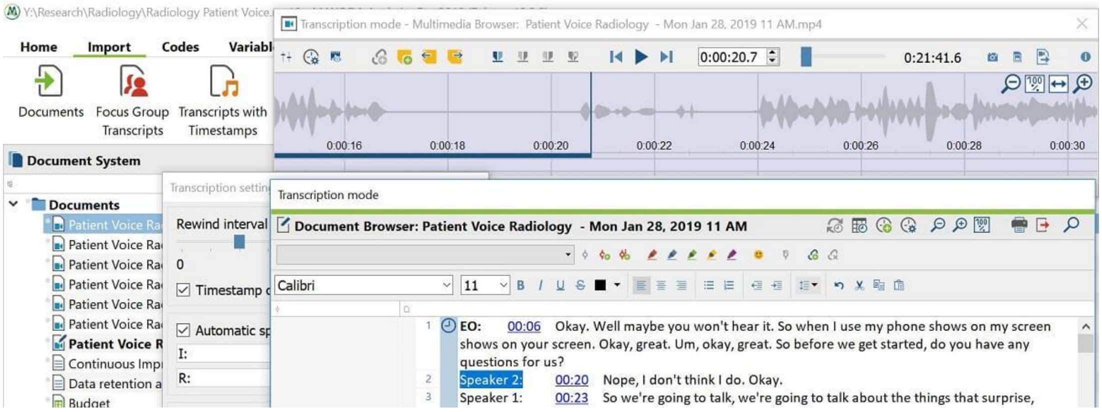
3. Recording and importing the interviews

I used Skype for Business to create the meeting requests for the interviews, provide Voice over IP (VoIP) services and dial-in numbers, and allow me to record the call as an MP4 file with good audio quality. I experimented with two alternative processes:

In the first, I used the content of an MS Word document from the transcription provider, and in the second process, I used MAXQDA's Transcripts with Timestamps function to import a SubRip (SRT) file with timestamps from the provider.

My transcription environment, captured in Figure 1, shows text from the machine learning service pasted into MAXQDA's "Document Browser" window for the selected audio file imported into the Document System (highlighted in the "Document System" window).

At this point, I am about to edit the name of the second speaker using the Automatic Speaker Change and Autotext functions:


Y:\Research\Radiology\Radiology Patient Voice...
Home
Import
Codes
Variabl
Documents
Focus Group Transcripts
Transcripts with Timestamps
Transcription mode - Multimedia Browser: Patient Voice Radiology - Mon Jan 28, 2019 11 AM.mp4
0:00:20.7
0:21:41.6
Document System
Documents
Patient Voice Ra
Continuous Impr
Data retention a
Budget
Transcription settin
Rewind interval
Timestamp
Automatic sp
Transcription mode
Document Browser: Patient Voice Radiology - Mon Jan 28, 2019 11 AM
Calibri
11

1 EO: 00:06 Okay. Well maybe you won't hear it. So when I use my phone shows on my screen shows on your screen. Okay, great. Um, okay, great. So before we get started, do you have any questions for us?
2 Speaker 2: 00:20 Nope, I don't think I do. Okay.
3 Speaker 1: 00:23 So we're going to talk, we're going to talk about the things that surprise,

Figure 1 MAXQDA's Transcription Mode Feature

4. Method 1 – Microsoft Word

My process using the MS Word file was as follows:

1. Download the transcription from the online service as an MS Word document,
2. Import the MP4 into MAXQDA,
3. Open the file for transcription,
4. Paste the content of the Word file into the Transcription Mode "Document Browser" window,
5. Run the audio in MAXQDA's Multimedia Browser, and make any edits in the text as needed,
6. Add timestamps at each speaker change or where desired.

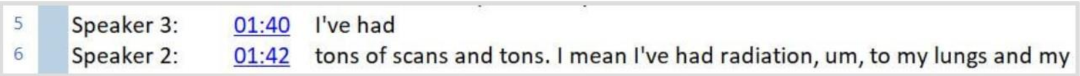
Method 1 Findings

The method was effective, and I will continue to use this in production because it has greatly reduced the time and effort to transcribe audio tracks at very low cost. There were some caveats, however.

Speaker changes

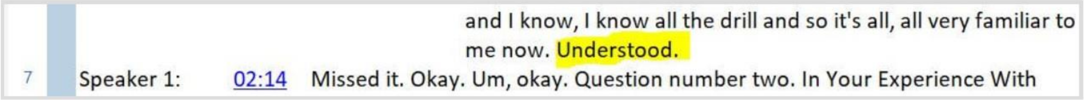
Firstly, the machine transcription did a reasonably good job of speaker changes, but sometimes it got confused. Sometimes it thought there was a different speaker, when in reality, it was the same person continuing to speak (Figure 2).

This was not a major hurdle, and the changeovers were reasonably obvious, and mistakes were easily corrected. Sometimes it changed over to a new speaker a little late, and switched speakers several words into the next speaker's text. In Figure 3 the highlighted text belongs to Speaker 1, not to the previous speaker. This too was reasonably easy to spot, and to edit.



5 Speaker 3: 01:40 I've had
6 Speaker 2: 01:42 tons of scans and tons. I mean I've had radiation, um, to my lungs and my

Figure 2 Same Speaker



and I know, I know all the drill and so it's all, all very familiar to me now. Understood.
7 Speaker 1: 02:14 Missed it. Okay. Um, okay. Question number two. In Your Experience With

Figure 3 Different Speaker

Multiple speakers

Secondly, it did not like multiple people talking at the same time. When there were multiple interleaved or simultaneous speakers, it tended to think it was the same speaker, and also sometimes lost several words of each speaker.

This issue can be seen in the first paragraph in Figure 1 (above) where the participant and interviewer start the interview talking about whether the VoIP system had announced that it was recording. Both speakers were captured as a single person, and their phrases intermingled.

This was harder to identify, and to edit but tended to happen at specific points in the interview, such as when there was a transition in topic. However, there were at least one or two occurrences in each interview, especially when something startling or funny was said, and multiple people interjected.

For example, when one participant commented that they had been left alone in the MRI room and staff had left for lunch, both interviewers interjected with comments, laughter, whistles, etc. The machine transcription missed some of the words, and mingled them together in small phrases. This was less easy to edit, but not any worse than was the case for me doing it manually.

Lower volumes

Lastly, it often ignored audio that was significantly lower volume than the rest (it may have calculated an average volume and then filtered out quieter sounds as noise). One recurring situation brought this to light. At the end of each participant track, the interviewer typically thanked them in a much quieter voice and then used normal volume to initiate the next question or comment.

In many cases, these “asides” were not transcribed at all. This was not a problem for my situation, but it may be a significant issue if the participant varies in volume.

5. Method 2 – SRT File

My method for the SRT file used MAXQDA’s Import Transcripts with Timestamps function:

1. Download the SRT file,
2. Import the SRT into MAXQDA using the Import Transcripts with Timestamps function,
3. Point to the associated MP4 file in the resulting dialogue box,
4. Open the file for transcription,
5. Run the audio in MAXQDA’s Multimedia Browser,
6. Make edits in the text as needed.

Method 2 Findings

The accuracy rate was the same, and the same cost and logistics considerations applied. The caveats were also similar, and although importing the SRT had an advantage in not requiring me to add timestamps manually, it resulted in a highly fragmented text layout inherited from the SRT file structure.

While the audio and text tracked well together, viewing the text as a vertical list of fragments was difficult to read, and coding was made significantly more difficult. As a result, I did not continue with the use of the SRT file method after the initial test.

6. Conclusion

The low cost and short turnaround time make machine-learning transcription services worthwhile to some researchers whose keyboard dexterity makes manual transcription tedious. The services do not offer perfect transcription, however, and each method to use them comes with some caveats.

Using the MS Word import was found to be preferable to using the SRT method, and what was lost in the more comprehensive timestamping was gained in the greater readability of the resulting text of the Word import.

My overall finding was that both TEMI and SPEXT were good enough to continue to use, but that if high fidelity was needed and cost and time were less of an issue, the human-machine combination would be an attractive option.

Editor´s Note

Matthew Loxton is a Principal Analyst at Whitney, Bradley, and Brown Inc. focused on healthcare improvement, serves on the board of directors of the Blue Faery Liver Cancer Association, and holds a master's degree in KM from the University of Canberra. Matthew is the founder of the Monitoring & Evaluation, Quality Assurance, and Process Improvement (MEQAPI) organization, and regularly blogs for Physician's Weekly. Matthew is active on social media related to healthcare improvement and hosts the weekly #MEQAPI chat.