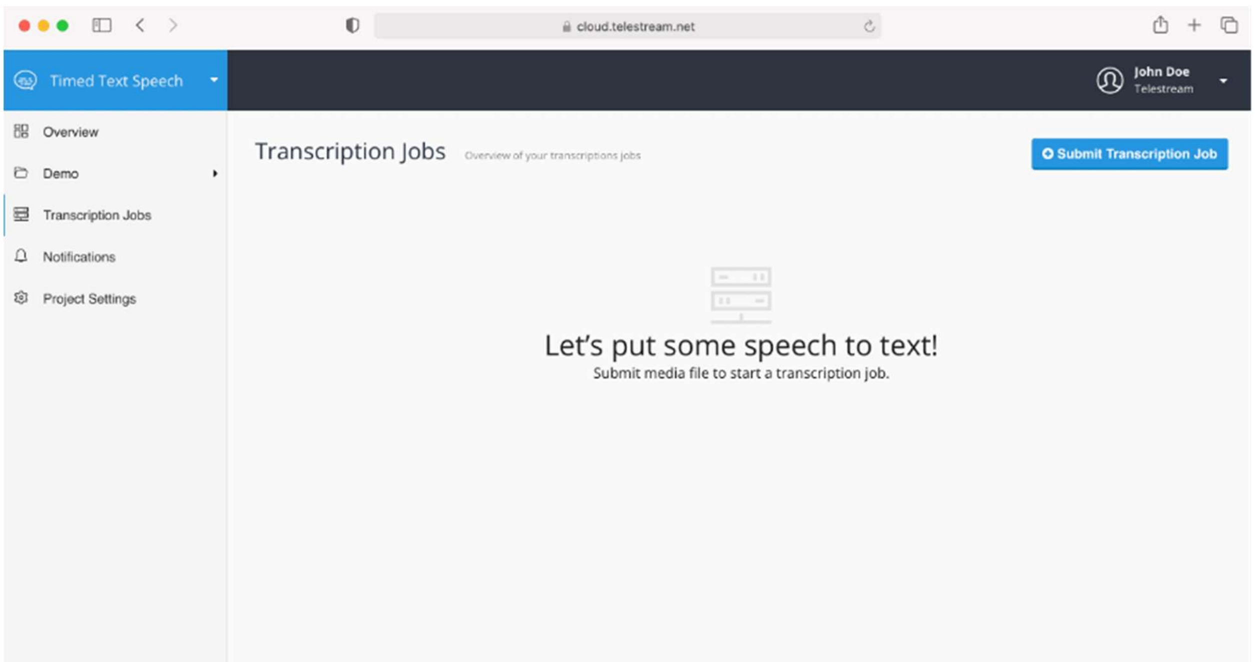
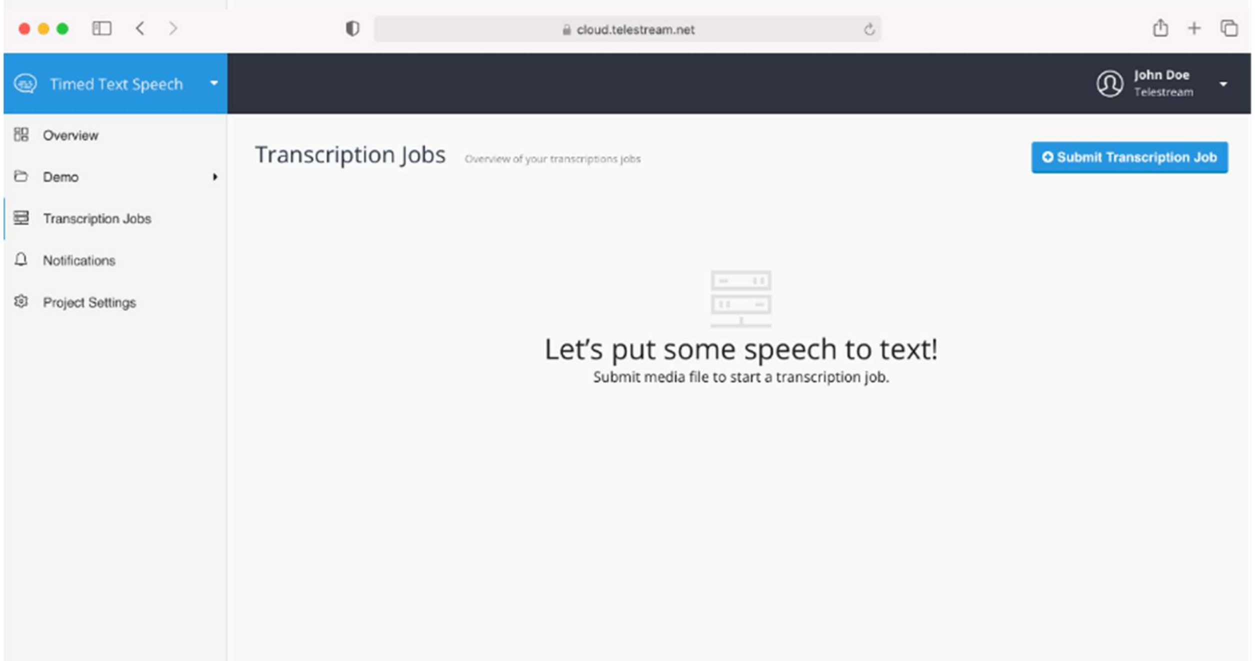
Running Transcription Jobs in the Cloud

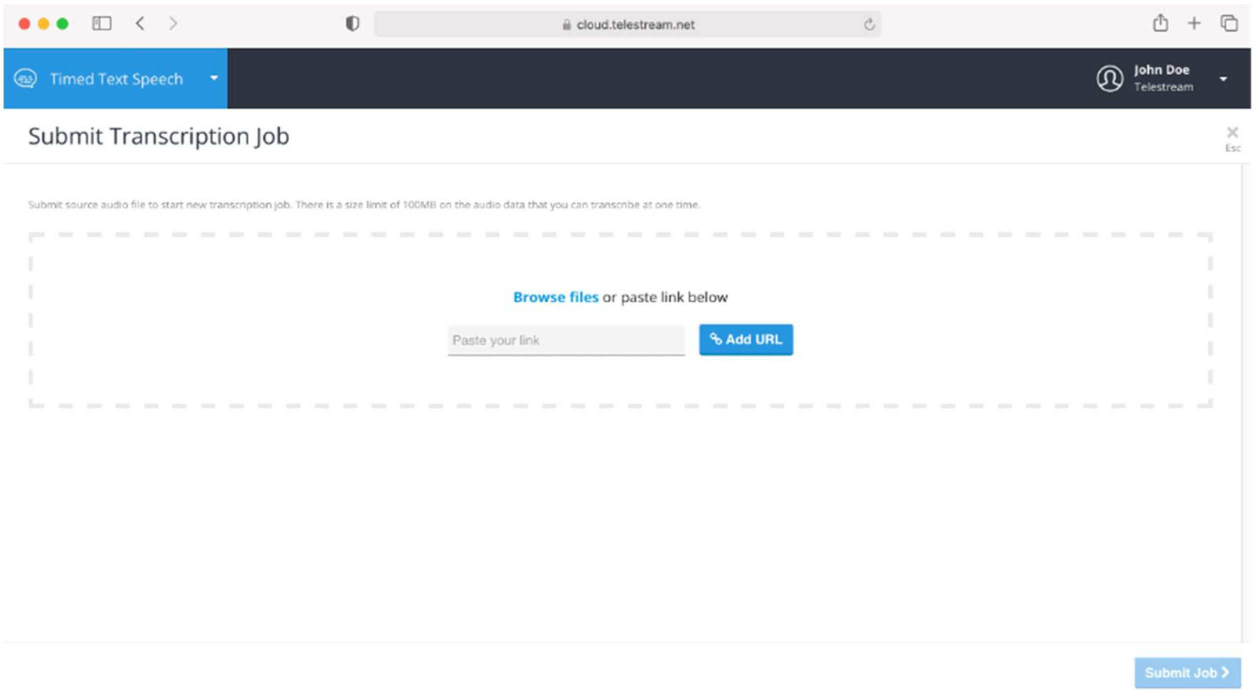
Jobs can be submitted to Telestream Cloud Timed Text Speech in a few ways - API, from CaptionMaker and MacCaption, or in this case through web console.

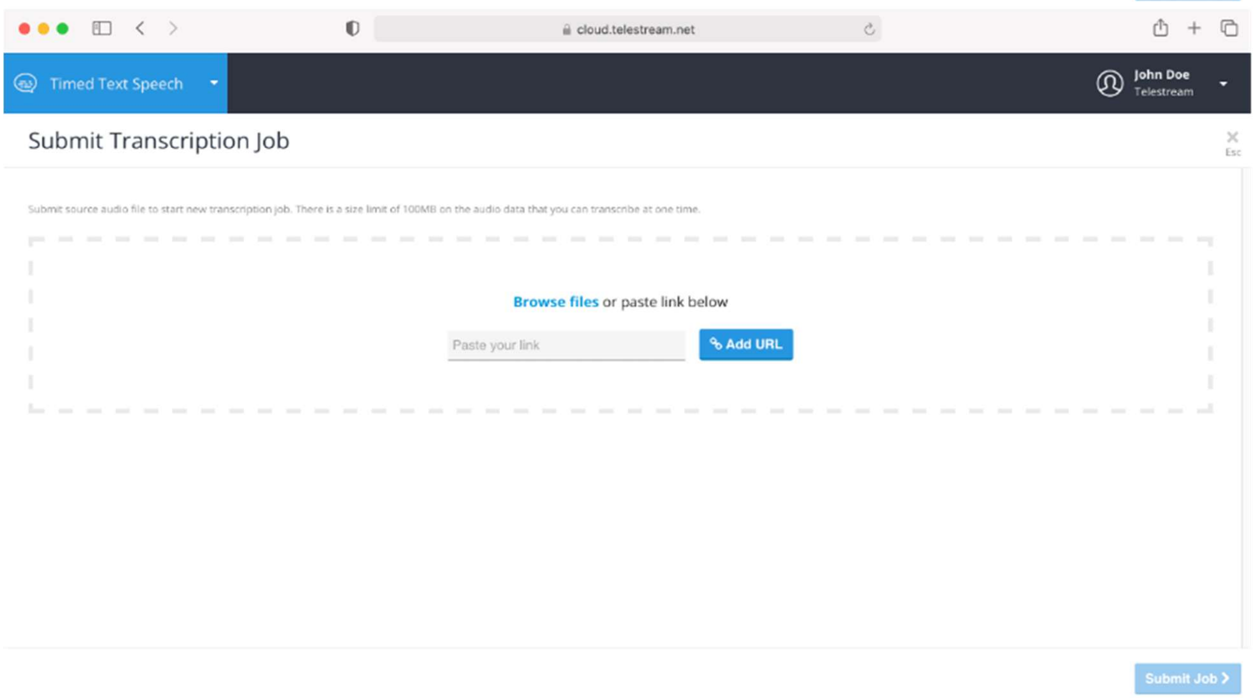
Once logged in, select the Project that you'd like to use and you'll be looking at the Transcription Jobs list. This is where you can keep track of all jobs that have been processed or are currently in progress.





Click the Submit Transcription Job button to select files for processing. You can either drag & drop source files from your local disc or paste the URL to your media file. There is a size limit of 100MB on the audio data that you can transcribe at one time. If you upload a media file with both video and audio we will extract the audio track before the transcription process.

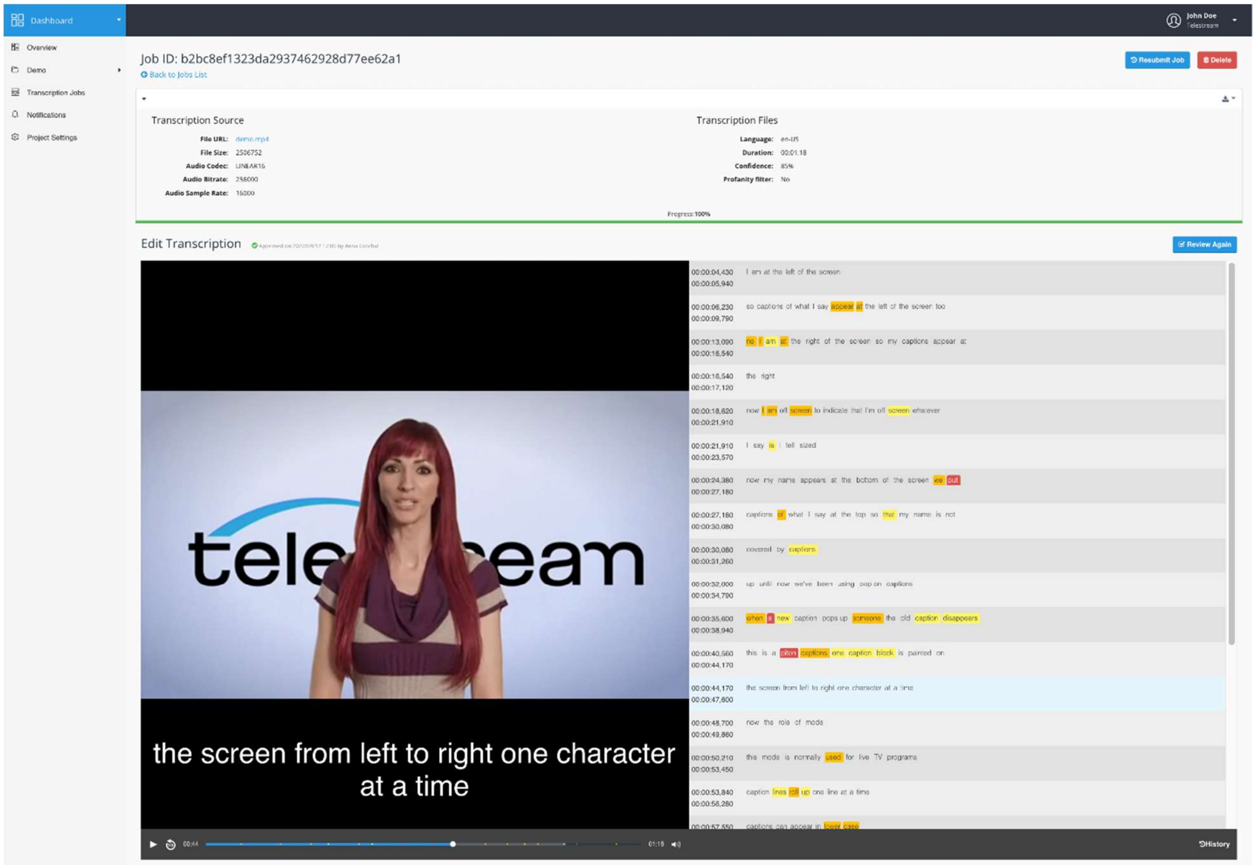


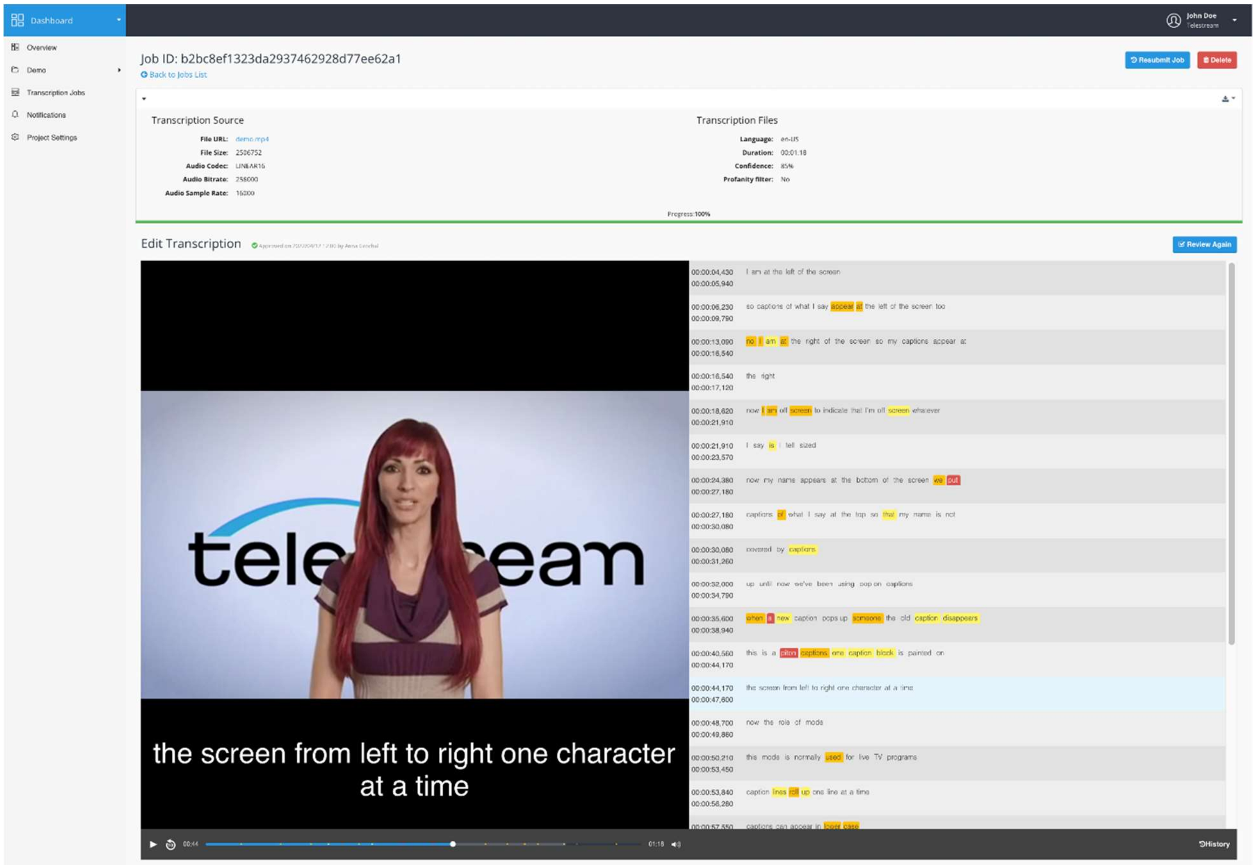


When ready, click the Submit Job button to start the upload and transcription process. You can follow the general progress in the jobs list. You can also click the job in-progress from the list to see a more detailed view. When the transcription is finished it's time to move on to the final stage - review and editing.

Transcription Review and Editing

While Timed Text Speech transcription accuracy is usually very high you may want to review the final result anyway. That's why we added the ability to generate a proxy file (either audio or video) which makes the review process even easier.

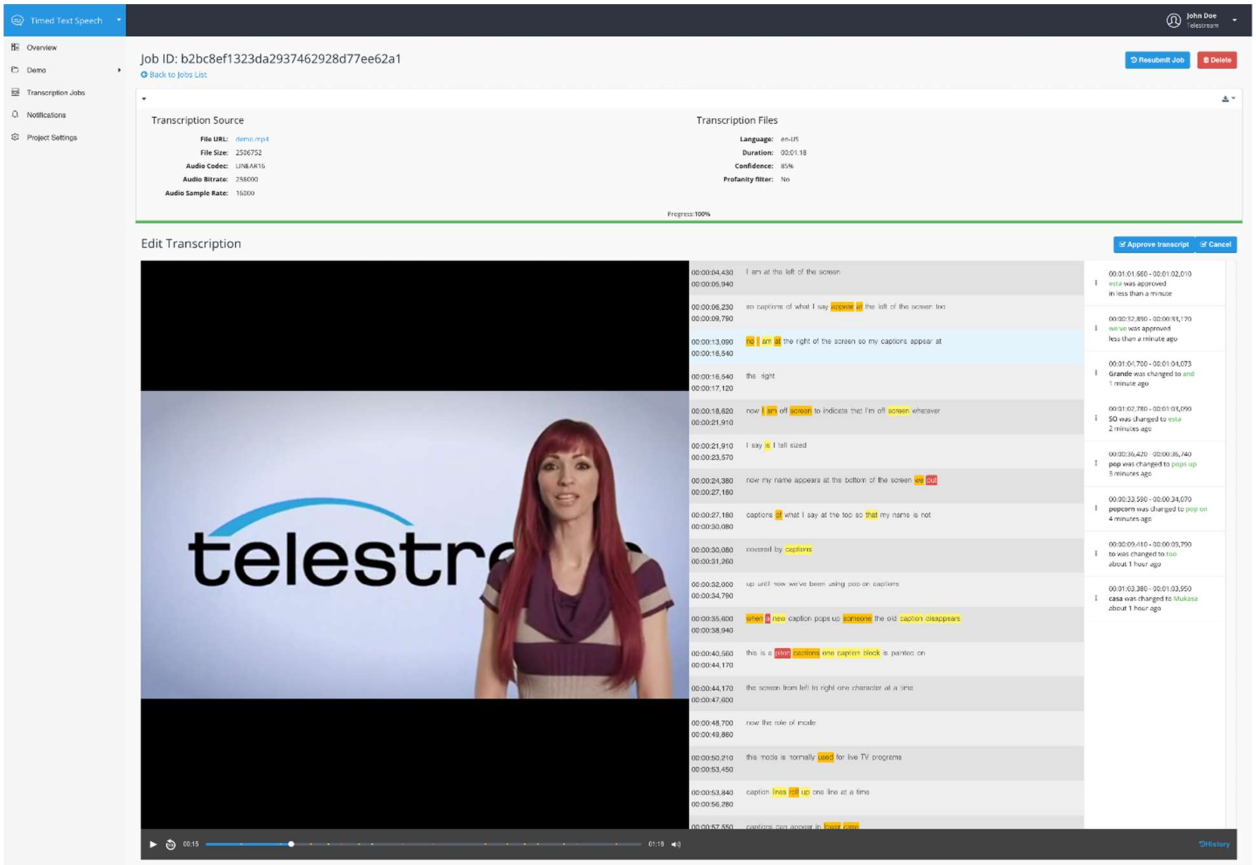

Dashboard
John Doe
Telestream
Overview
Demo
Transcription Jobs
Notifications
Project Settings
Job ID: b2bc8ef1323da2937462928d77ee62a1
Back to Jobs List
Resubmit Job
Delete
Transcription Source
File URL: demo.mp4
File Size: 2506752
Audio Codec: LINEAR16
Audio Bitrate: 256000
Audio Sample Rate: 16000
Transcription Files
Language: en-US
Duration: 00:01:18
Confidence: 85%
Profanity filter: No
Progress:100%
Edit Transcription
Review Again
00:00:04,430 00:00:05,940 I am at the left of the screen
00:00:06,230 00:00:09,790 so captions of what I say appear at the left of the screen too
00:00:13,090 00:00:16,540 no I am at the right of the screen so my captions appear at
00:00:16,540 00:00:17,120 the right
00:00:18,620 00:00:21,910 now I am off screen to indicate that I'm off screen whatever
00:00:21,910 00:00:23,570 I say is I tell sized
00:00:24,380 00:00:27,180 now my name appears at the bottom of the screen we put
00:00:27,180 00:00:30,080 captions of what I say at the top so that my name is not
00:00:30,080 00:00:31,260 covered by captions
00:00:32,000 00:00:34,790 up until now we've been using pop on captions
00:00:35,600 00:00:38,940 when a new caption pops up someone the old caption disappears
00:00:40,560 00:00:44,170 this is a piton captions one caption block is painted on
00:00:44,170 00:00:47,600 the screen from left to right one character at a time
00:00:48,700 00:00:49,860 now the role of mode
00:00:50,210 00:00:53,450 this mode is normally used for live TV programs
00:00:53,840 00:00:56,280 caption lines roll up one line at a time
00:00:57,550 captions can appear in lower case
the screen from left to right one character at a time
00:44
01:18
History

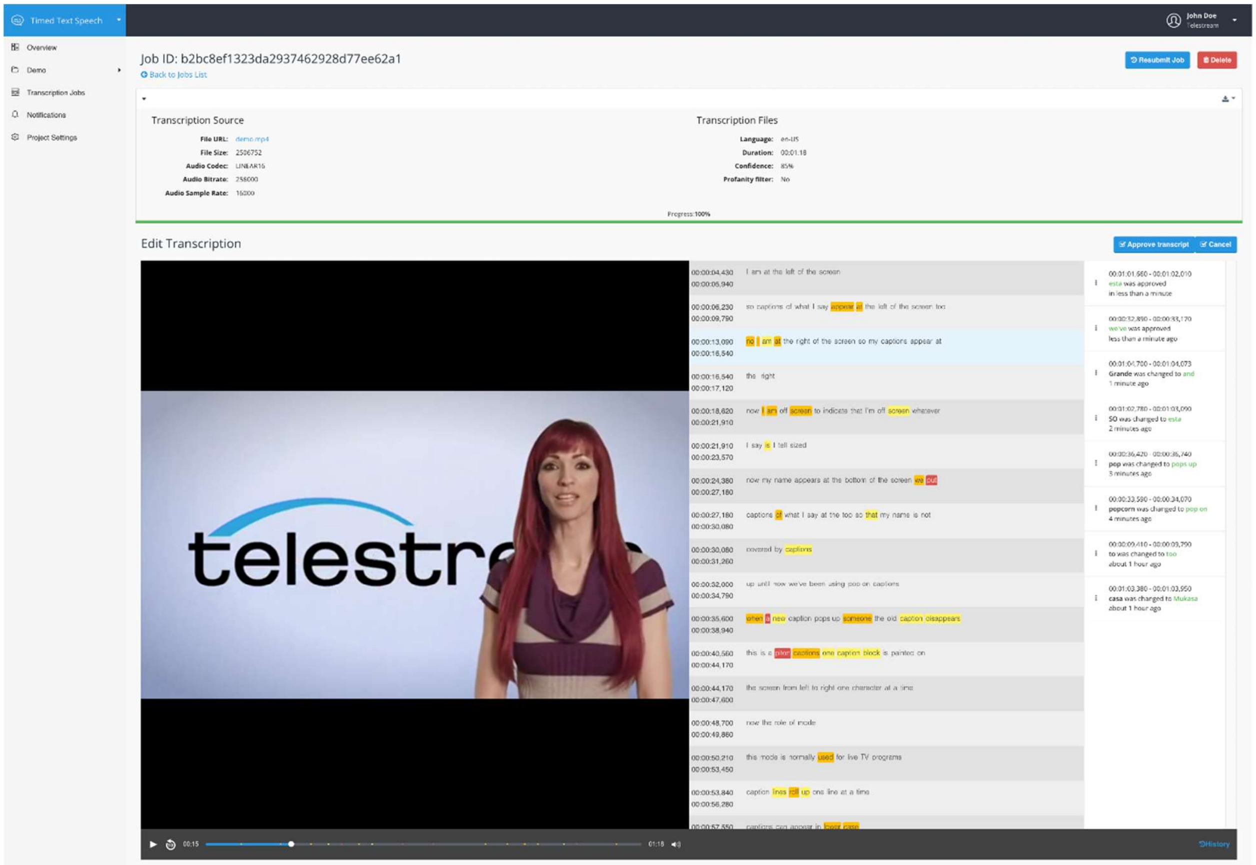


The media file timeline has markings to show parts of the transcription which may need your special attention. This usually happens when the transcription model finds alternative words with similar confidence levels, or if the transcription confidence falls below a certain level.

The timeline is aligned with the transcription editor below, so any time you click on it you will be taken to the relevant place on the video timeline. Text highlighted in yellow, red, or orange means that the speech-to-text engine does not have high confidence that it located the correct word. This is where you need to step-in and either confirm it's the right word, correct it by choosing one of the available options, or enter the correct word manually. You can also edit whole lines at a time instead of single words.

There are also formatting options which allow you to split or merge lines to match the media file. All changes are visible on the fly so you have instant feedback on your actions. If at any time you need to review your changes there is history view available.


Timed Text Speech
John Doe
Overview
Demo
Transcription Jobs
Notifications
Project Settings
Job ID: b2bc8ef1323da2937462928d77ee62a1
Back to Jobs List
Resubmit Job
Delete
Transcription Source
File URL: demo.mp4
File Size: 2506752
Audio Codec: LINEAR16
Audio Bitrate: 256000
Audio Sample Rate: 16000
Transcription Files
Language: en-US
Duration: 00:01:18
Confidence: 85%
Profanity filter: No
Progress: 100%
Edit Transcription
Approve transcript
Cancel
telestre
History



As soon as you're happy with the final result you can download the transcription file in one of the available formats - TXT, JSON, CSV or SRT and use it in your project.

Building Effective Custom Vocabulary

Speech-to-text engines are trained on a vast collection of sample recordings and texts. This means they perform well when your source is similar to "average" speech – i.e. typical conversations on common topics and using colloquial vocabulary and phrases that you would commonly find in your language. They will not perform as well if your source contains a lot of unique words or specialized terminology or phrases that the engine has not encountered before.

Using custom vocabulary allows you to inform the speech engine about unique words and phrases that are likely to occur in your source audio, so that it is more likely (but not guaranteed) to recognize them correctly.

Ideally, you should include things like proper nouns, acronyms, specialized terms, and short phrases which frequently occur in your audio but which are not part of typical every day conversations.

The downside of using custom vocabulary is that the words and phrases you include will be prioritized over more commonly spoken words and phrases. Using a large custom vocabulary with too many unnecessary words or phrases can actually decrease the accuracy of the results.

Following a few tips listed below will improve transcription accuracy.

Do Include

- Proper nouns (names of people, places, companies, etc.) – especially if they are from a different language, non-dictionary words, or words with an unusual spelling
- Acronyms (company names, abbreviations, etc.)
- Short phrases (less than 100 characters) which are unique to your source and are repeated often (e.g. a catch phrase often spoken by a character, or industry terminology)

Do not Include

- Words or phrases that are unlikely to occur in your source file
- Phrases longer than 100 characters
- Long phrases, sentences, or paragraphs that occur only once in your source
- Long lists of unrelated or unlikely words or phrases

Format

The custom vocabulary should be specified as a comma separated list of words or short phrases.
Longer phrases can be placed on separate lines (line delimited).

Limits

Please note these limits which are enforced by the API engine. Attempting to use a vocabulary file over the limits could result in the job failing.
These limits are so high that if you are approaching them, you should review the recommendations above.

- Number of phrases: 5000
- Characters per phrase: 100
- Total characters: 100,000

- # Getting the Best Results from Auto-Transcription
-
- **Using Telestream Cloud for auto-transcription of media files is a great way to get started on a caption or subtitle project.**
- All submissions will result in an automatically generated transcript timed to match your media file. The timed transcript can be reviewed and edited on the fly in the Telestream Cloud console or directly populated into your MacCaption or CaptionMaker project. The more accurate the results the less clean up and editing is needed to make your transcript perfect. Below are some best practices and tips that can help increase the accuracy of the auto generated transcript from Timed Text Speech.
- **Isolating the Spoken Word**
  For original content or media with multichannel audio, the dialogue-only track can be isolated to eliminate noise, music, and sound effects. This isolation can be done with any video editing software or audio production tool. By submitting audio with only spoken words to the Timed Text Speech engine, accuracy of the auto-generated transcripts can be greatly increased.
- To accomplish this a video editor can open their project in Adobe Premiere or Avid and silence all audio tracks which do not contain dialogue. Next, they can simply export an audio-only file. Timed Text Speech can handle audio files such as .mp3, aiff, and wav.
- In some cases your media file such a QuickTime or .MXF may contain multiple audio tracks. This could be a 5.1 mix, or isolated tracks for archival or transcoding purposes. Within MacCaption or CaptionMaker users have the option to select any of the audio tracks within the video file before submitting their project to Timed Text Speech. By default the software will submit tracks 1 and 2. If there is different audio track in the file which contains the spoken words to transcribe, the software can be configured to submit the alternate audio track instead. This means that the spoken-word-only track will be processed and the results will in turn be much more accurate.
- **Training the Speech Engine**
  In many cases media files which require transcription may contain names, phrases, and acronyms that are not common. The speech engine may consistently get these wrong, causing users to manually correct the results again and again. To remedy this, Telestream Cloud’s console offers a way

to train the Timed Text Speech engine by uploading a corpus text file, or by manually entering these uncommon words.

- This file is a simple plain text .txt document that contains a list of names and phrases that are used in a project. Users can upload this .txt document to any specific project that may require training to increase accuracy. We recommend that the .txt document contain a list of phrases on each line instead of only individual words.
- For example an **effective** corpus text document would like like this:
- John Galveston
  CEO of the Company
  Working with CDN providers
  Transcoding and captioning solutions
  MacCaption Software
- An example of a corpus text file that is **not effective** looks like this:
- John
  Galveston
  CDN
  Providers
  Transcoding
  Captioning
  MacCaption
  Software
- By using phrases the speech engine knows what to expect and which other words are typically used with the new vocabulary. This added context means that results when using the custom vocabulary will greatly increase in accuracy.
- In some cases, creating a corpus text file for training is very easy and takes very little time. Some users simply repurpose the old transcripts or caption files from the same TV program or Project. For example, if a broadcaster needs to create a transcript for season 3 of a TV program, they can open the caption files from season 1 and 2 using MacCaption or CaptionMaker and export a corpus file which can be used for training Timed Text Speech. These 2 previous seasons contain the names and phrases that would greatly increase the vocabulary.
- Another way that users can leverage the vocabulary training of Timed Text Speech is when a rough transcript is already available of the media file prior to submission to Telestream Cloud. This rough transcript will also contain the names and key phrases for the project. Timed Text Speech would then automatically time the rough transcript and fill in the text that is missing.

- **Content Types Best Suited for Automatic Speech Recognition (ASR)**
  The type of video content plays an important role in the level of accuracy achieved using auto-transcription software. For example, a news show with a professional announcer and clear studio audio will have great accuracy compared to a video shot outdoors in a noisy environment on a mobile phone. In addition, loud music, singing, and shouting will also bring down the level of accuracy. There are also cases where speakers may change their voice to provide dialogue for children's programming or for dramatic effect. This means that a speech engine that is designed and trained for standard voices may not be able to understand these voice tones. Generally speaking, projects with clear studio quality recordings, minimal background noise, and a professional speaker will always result in the best accuracy.
- **Creating a Proxy**
  Professional video companies generally work with high quality video master files called mezzanine files. These files are used the same way tape masters were used in the old days. The original video must be uncompressed or high bitrate when submitted for processing. This is not the case for Timed Text Speech workflows. Because Telestream Cloud requires only the audio content for auto-transcription, users can submit a low bitrate MP4, or just the audio file. As long as the audio quality is good you will achieve high quality results, the video quality or resolution does not affect your results.
- **Voiceover for the Purpose of ASR**
  For video editing workflows, it's quite common for editors to do a rough voiceover when editing prior to bringing in voice talent to the studio to record the final audio. This also provides an opportunity for video editors to re-speak any portions of the video project that do not have clear audio. This rough voice over can then be exported from the video editing system and submitted to Timed Text Speech for processing. This means that results will have a greater accuracy than the original audio.