A GUIDE TO CLOx

Client Libraries Oxford

Alicia Wassink, Rob Squizzero, Campion Fellin and David Nichols

University of Washington

https://clox.ling.washington.edu/

TABLE OF CONTENTS

WHAT IS CLOx?

CLOx is a web-based service developed by the Sociolinguistics Laboratory at the University of Washington. CLOx uses Microsoft's Bing Speech Recognition API **C**lient **L**ibraries, previously known as Project **Ox**ford, to orthographically transcribe sociolinguistic or other audio-recorded interviews in a format amenable to linguistic analysis. It outputs transcriptions in a .csv format with timestamps indicating the start and end time of each turn of speech contained in an audiofile. It is estimated that CLOx enables accurate transcription of a sociolinguistic interview to be completed in one-fifth or less of the time it would take to produce a manual transcription.

Currently supported languages include Arabic (EG), Chinese (CN), English (GB), English (US), French (FR), German (DE), Italian, Japanese, Portuguese (BR), Russian, and Spanish (ES). In the event that more languages are added to the Bing Speech Recognition API in the future, CLOx should be able to integrate them.

We recommend using Chrome, Firefox, or Opera. Safari is not currently supported. We do not recommend using CLOx without high speed internet, as slow connections are susceptible to timeout errors.

HOW DO I USE CLOx?

REQUIRED SUBSCRIPTION KEY

CLOx works by directing your audiofile to Microsoft's Bing Speech Recognition server to be processed. To access this server, users need to have an access code called a "subscription key."

A free[1] subscription key may be acquired at http://azure.microsoft.com/. See the Appendix for a guide with screenshots illustrating how to obtain your key.

[1] The subscription is free for the first month, and after that provides up to 5000 calls to the server per month for free, then is billed at $4 per additional 1000 transactions. Azure allows monitoring of the number of calls. 5000 is probably more than sufficient for any small-to-medium scale transcription project. For a very large amount of transcriptions done in a short time, you may incur fees. We are not responsible for any fees incurred using the service.

PREPARING YOUR AUDIOFILE

File requirements at-a-glance:

- *.wav format*
- *mono/stereo: mono only*
- *maximum filesize: 10 MB*
- *sampling rate: 16 kHz*

CLOx operates on .wav files. Files uploaded must be mono (using one channel only), sampled at 16 kHz and under 10 MB in size, which is about 5 minutes of speech under those specifications. If your file is not formatted in this way, you may (1) use one of our automated pre-processing options below to create a set of extracted files that meet these Bing Speech requirements, or (2) pre-process it manually. **Warning:** if audio is not preprocessed, CLOx may crash without notification, produce undesirable output, or run at slower than optimal speeds.

Automated pre-processing

We offer two methods for automated preprocessing: (1) a Praat script, (2) a Python script.

A Praat script is provided with the CLOx service. This script will automatically format audiofiles to the CLOx requirements above. For example, many sociolinguists record audio in stereo at a 44.1 kHz sampling rate. Such files cannot be processed by the Bing Speech Recognition system[2]. This script does all the work necessary to

[2] Technically speaking, CLOx (and Bing, for that matter) will accept 44.1 kHz stereo audio, but Bing will automatically downsample audio to 16 kHz mono. This will cause

extract one channel, lower the sampling rate, and generate temporary audiofiles of an appropriate length to work with the Bing Speech Recognition system. To obtain the script, navigate your browser to the CLOx website and click on the download link, or download the file auto_extraction.praat from the demonstration tutorial folder. Launch Praat, then Open the script (select Praat, then Open Praat script...) and your sound file (as a Sound, not as a LongSound), then run the script. A set of temporary audiofile extracts, indexed sequentially and meeting the CLOx requirements will be saved to your computer for use with CLOx.

[[insert DN text here]]

Manual pre-processing

To prepare your audiofile yourself, you may need to convert it from stereo to mono, resample the audio signal, and segment the file into smaller extracts. The steps below describe how to do this. These steps assume you are working in Praat, but any other signal analysis software program may be used (e.g., Audacity, etc.).

Step A: Converting from Stereo to Mono

To extract one channel in Praat:

1. *Open your audiofile as a "Sound" object (not as a "LongSound").*
2. *Select the file in the object list.*
3. *Select "Convert -" then "Extract one channel..." and select the channel you wish to extract.*
4. *Save your new audiofile (in .wav format).*

CLOx to run more slowly without any accompanying improvement in output quality. In order to avoid timeout errors, CLOx will not accept files larger than 10 MB.

5. *If your new (mono) file exceeds 5 minutes, either submit it to the preprocessor as in step 1 above for segmenting and resampling (recommended) or follow the steps below for Resampling (B) and Segmenting (C).*
6. *If your new (mono) file does not exceed 5 minutes, follow step B. below and proceed to "Transcribing with CLOx."*

Step B: Resampling

Your .wav file must be sampled at 16,000 Hz. Resampling may be done in Praat by selecting the mono file and clicking Convert – Resample… and entering 16000 in the New Sampling Frequency (Hz) box. The precision does not need adjusting.

Step C: Segmenting

If you would like to manually segment your audio into extracts under 5 minutes in length or if you have multiple files that you would like included in the same transcript:

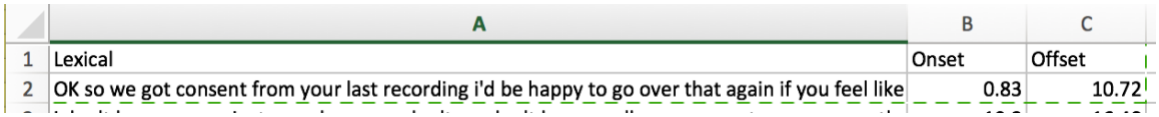
1. *Make sure that each extracted audiofile name ends with _"startTime".wav, where "startTime" is the beginning time of the file in milliseconds. For example, if you are using 5-minute (300000 millisecond) extracts, the first .wav should be named soundName_0.wav, the second soundName_300000.wav, the third soundName_600000, and so on. (This is done to ensure that timestamps are represented accurately in the output .csv file.)*
2. *Create a unique local directory to hold the extracted .wav files that you want included in the same transcription.*
3. *Sort the extracted files by name, so the first file listed in the directory is soundName_0.wav, the second is soundName_300000.wav, etc.*

TRANSCRIBING WITH CLOx

CLOx works with an open connection to a Bing Speech Recognition server. For best results, make sure that the screensaver and sleep (sometimes called energy saver) system settings of your computer are turned off while CLOx runs. You may, alternatively, set both to start only after an extended period of time (2 hours should be more than sufficient for even a lengthy CLOx session on an older computer). This ensures that you will not experience disconnection from the server due to periods of inactivity. You should also avoid navigating your internet browser away from the CLOx webpage while CLOx is running.

1. *Enter your subscription key in the "Subscription" field.*
2. *Select the language of your interview/audio file.*
3. *Enter a name for your output file.*
4. *Click "Select Files and Start."*
5. *A dialog box will appear. Navigate to the folder containing the preprocessed files. Select desired files using shift+click, ctrl+click, or cmd+click. Files should be sorted by name in ascending order. Press enter or click ok.*
6. *The status box will update to display CLOx's progress. It shows which audiofile is being accessed, and indicates that text is being added to the output file row by row. As a rule of thumb, allow up to one minute of transcription time for every one minute of speech in your audio file(s).*
7. *When complete, "Done" appears in the status box. The comma-separated output file will be downloaded to the folder you specified for output. This file may be opened in Microsoft Excel, or any standard text editor.*

8. *The output file contains 3 columns, labelled "Lexical," "Onset" and "Offset." an example is shown below:*



	A	B	C
1	Lexical	Onset	Offset
2	OK so we got consent from your last recording i'd be happy to go over that again if you feel like	0.83	10.72

The first column, "lexical," contains the speech recognizer output. The second and third, "onset" and "offset," respectively, contain the beginning- and ending-times of each "turn," determined by Bing's algorithm as a run of speech ending with a period of silence above an arbitrary threshold, or when Bing detects a change in speaker vocal quality. Each row indicates a "turn." CLOx appends all output to a single .csv file. So, regardless of whether a single audiofile or multiple audiofiles were selected, all output is concatenated into a single output file. This saves the user from having to concatenate multiple transcripts associated with a single interview session.

WHAT DO I DO WITH MY OUTPUT?

While the Bing Speech API service provides highly accurate transcriptions, it is far from flawless. In addition to occasional transcription errors of words or clauses, CLOx is currently unable to reliably separate speakers on a recording and may have difficulty accurately transcribing overlapping or otherwise obscured speech. Therefore, it is important to check and manually correct your transcriptions. We recommend importing your CLOx output to a software application such as ELAN, that allows auditing of your audio alongside your CLOx output. Here's how:

Manual Correction in ELAN

CLOx transcripts are designed for easy importing in ELAN. Follow these steps:

A. Before opening ELAN, open the .csv in Excel. Select the "Onset" and "Offset" columns, right-click, select "Number," ensure the number of decimal places is set to 2 and click OK. Then save the file. If Excel asks you if you want to convert the file format, say no. This resolves the first known issue, below.

B. In ELAN, instead of creating a new project, select File → Import → CSV / Tab-delimited Text File…

    1. Select the "Lexical" column as "Annotation"
    2. Select the "Onset" column as "Begin Time"
    3. Select the "Offset" column as "End Time"
    4. Specify first row of data: 2

C. Now add your audio file by selecting Edit → Linked Files…

    1. On the Linked Media Files tab (selected by default), click Add…, select the file containing the entire recording of your transcription, and click Apply.

You should see the waveform of the audio file appear above the transcription in the main ELAN window, and it should be properly aligned with your transcription.

KNOWN ISSUES

?? I don't have an API/Subscription Key

It is not possible to get an API key at this time. Microsoft stopped issuing these for the Bing Speech API as of 10/24/2018 as they have started to transition over to their new unified Speech API. The way they have structured this transition requires us to rewrite a large portion of CLOx from scratch in order to be compatible with the new Speech API, so it is not possible to use a subscription key for the new

Speech API. In the meantime, we are able to supply a limited number of users with an API key – please contact cloxhelp@uw.edu for details.

?? My connection terminated prematurely:

If you experience connection issues mid-transcription, you will need to refresh the page and start again from the beginning. You may wish to run transcriptions in segments if this error persists. For best results, the screensaver and sleep (sometimes called energy saver) system settings should be turned off while CLOx runs, and you should avoid navigating the browser away from the page while CLOx is running.

?? My onset/offset times are slightly off:

This is a bug that sometimes occurs when a new audio segment is transcribed. If you notice all segments after a timestamp are slightly off in the same direction, you can fix this in ELAN's annotation mode.

- *Zoom in to measure how much the segment is misaligned.*
- *Place the cursor to the left of the first misaligned segment.*
- *From the annotation menu at the top, select "Shift →" "Annotations on All Tiers, Right of Crosshair…"*
- *If the annotations are misaligned to the left (the start times are early), enter the time you found in step a above. If they are misaligned to the right, put a negative sign (-) and then enter the time.*

?? CLOx returned a blank .csv transcript:

There are two common reasons for this:

1. *There is a problem with your audio file. Ensure that it is preprocessed correctly, and then listen to the preprocessed audio files to ensure they contain the correct audio.*

2. *There is a problem with your subscription key. You may have entered it wrong, or, on rare occasions, it may have been invalidated by Microsoft. To remedy this, regenerate your key as shown in Step 10 of the Appendix. If you have not upgraded your account after your free trial period, your key will not work until you have upgraded. You may need to regenerate your key after upgrading.*

CONTACT US

Questions? Issues? Contact Rob Squizzero at cloxhelp@uw.edu.

Thanks for using CLOx!

SECURITY AND PRIVACY OF DATA

It is expected that many CLOx users are working with audio data that may contain subjects' personal or identifying information and may be subject to scrutiny by Institutional Review Boards (IRBs). To that end, we recognize that data security is of paramount importance and would like to explain the precautions we have taken.

Audio files and generated transcriptions never pass through or are intercepted by the CLOx server. CLOx, as a service, merely facilitates transfer between a user's computer and Microsoft's speech-to-text servers. Microsoft's policies prohibit their storage of any audio recordings or transcripts generated for any purpose.