A Pipeline for the Large-Scale Acoustic Analysis of Streamed Content

Steven Coats
English, Faculty of Humanities, University of Oulu, Finland
E-mail: steven.coats@oulu.fi

Abstract

Vast quantities of audio and video data are available from video sharing sites, streaming services, and social media platforms, but relatively little of this content has been utilized for acoustic, phonetic, or multimodal analysis of linguistic variation. This article describes a Python-based scripting pipeline for the extraction and analysis of audio from YouTube and other websites that use common streaming protocols. The pipeline comprises elements from the Python libraries yt-dlp and Parselmouth and uses the Montreal Forced Aligner for aligning audio with text. The scripts are customizable and suitable for the automatic extraction of video as well as audio and transcript data. An exploratory proof-of-concept analysis considers the first target of the /eɪ/ diphthong in American English: Starting from videos indexed in the Corpus of North American Spoken English, almost 9 million tokens of the segment were retrieved using the pipeline and their values in F1/F2 formant space mapped. As expected, the diphthong's first target has a more closed and back starting point for speakers in the American Southeast.

Keywords: Corpus linguistics, Phonetics, Formants, YouTube, DASH, CoNASE

1. Introduction

The study of linguistic and interactive properties of Computer-mediated communication (CMC) has historically been focused primarily on text content such as chat, instant messenger (IM) messages, or text-based posts on social media web platforms. In the past 15 years, however, continual increases in bandwidth availability and refinement of technical protocols have led to the widespread use of images, audio, and streamed video content in CMC, for example on video sharing and streaming sites or in online video meetings. Multimodality, or the concurrent use of text, speech, and video, has become central to CMC on the most widely-used video sharing and social media communication sites such as YouTube, Twitch, or TikTok.

As of 2023, most websites utilize the DASH protocol (Dynamic Adaptive Streaming over HTTP; Sodagar 2011) or the related HLS protocol (HTTP Live Streaming) to serve video, audio, and other content on the web. DASH allows the transmission of video and audio data in various formats and compression levels as well as automatic speech recognition (ASR) or manually-uploaded captioned transcripts of speech, user comments and interactions, and other types of data and metadata to the end user in a web browser.

For the researcher interested not only in text, but also in acoustic, phonetic, or gestural/kinesic properties of communication, multimodal content delivered via web streaming represents a valuable source of empirical data. This paper presents a pipeline for accessing streamed audio content on YouTube for phonetic analysis.[1] The pipeline, which is Python-based, makes use of several open-source tools, code libraries, or repositories: yt-dlp[2] for content download of audio, video, and transcript data; the Montreal Forced Aligner[3] for forced alignment of audio and transcript data; and Parselmouth-Praat[4] for identification and extraction of acoustic features of interest. The pipeline, in a Python Jupyter format, consists of modular script blocks that can be modified and adapted for specific tasks on existing datasets without needing to apply all of the steps in the pipeline. While the example provided in this paper focuses on acoustic properties of audio segments, the pipeline is also suitable for the automated download of corpora of video content.

The rest of the paper is organized as follows: Section 2 provides a brief overview of a few tools used for forced alignment and acoustic analysis of online content, including web-based services. Section 3 details the components used in the pipeline, and Section 4 demonstrates the functionality of the pipeline by providing an exploratory analysis of geographical variation for the first target of the /eɪ/ diphthong in F1/F2 formant space in North American English, starting from videos indexed in the *Corpus of North American Spoken English* (Coats 2023). The exploratory analysis reveals a pattern that corresponds to results of previous research, confirming the potential usefulness of the pipeline. Section 5 summarizes the paper and provides a brief outlook for future developments.

2. Previous Work

Phonetic analysis of speech audio requires a transcribed text and a forced alignment of the transcript with the speech signal, permitting the acoustic analysis of words, phonemes, and other segments. Several tools for forced alignment have been built on the Hidden Markov Model Toolkit (HTK, Young 1993)[5] and Kaldi (Povey et al. 2011)[6]: The Penn Forced Aligner (Yuan & Liebermann 2008) and the MAUS aligner (Schiel 1999), for example, are built on HTK, while the Montreal Forced Aligner (McAuliffe et al. 2017) builds upon Kaldi. Other forced alignment tools include Julius (Lee et al. 2009),[7] and SPPAS,[8] developed for French on the basis of Julius but capable of aligning

[1] https://github.com/stcoats/phonetics_pipeline
[2] https://github.com/yt-dlp/yt-dlp
[3] https://montreal-forced-aligner.readthedocs.io
[4] https://github.com/YannickJadoul/Parselmouth
[5] https://htk.eng.cam.ac.uk
[6] http://kaldi-asr.org
[7] http://julius.osdn.jp/en_index.php
[8] https://sppas.org

additional languages (Bigi 2015).

Composite tool suites and web-based speech processing platforms have incorporated these aligners into their functionality, making it easier to process audio recordings without having to install and configure the software locally. FAVE-Extract (Forced Alignment and Vowel Extract, Rosenfelder et al. 2011), for example, uses the Penn Forced Aligner, while WebMAUS (Kisler et al. 2017) and DARLA (Dartmouth Linguistic Annotation, Reddy & Stanford 2015), which use MAUS and MFA, respectively, are websites that allow users to upload audio files and transcripts for forced alignment. A recent option in DARLA allows users to generate ASR transcripts from audio files by sending them to Deepgram, a paid service that hosts large neural network speech-to-text models.

Studies have shown that the Penn Forced Aligner and the Montreal Forced Aligner can produce results comparable to those of human annotators. MacKenzie and Turton (2020), for example, used FAVE and DARLA to align samples of speech from six regional British English varieties. Comparing them with alignments produced by human annotators, they found that DARLA performed slightly better than FAVE, but that both tools perform well and produce alignments comparable to those created by human annotators. They remark that "the fact that they have been provided with phonological systems that differ – sometimes rather radically – from the systems they have been trained on has not hindered their performance" (2020: 9), and conclude "our analysis has shown impressive performances from both DARLA and FAVE, and we have full confidence in recommending that researchers who work on non-American and non-standard varieties of English use these tools for forced alignment" (2020: 11). Similarly, Gonzalez et al. (2020) found that the Montreal Forced Aligner generated accurate alignments for recordings of Australian English, even when using an American English model.

Once the audio and transcript have been aligned, acoustic analysis can be undertaken with Praat (Boersma & Weenink 2023) or other software, for example to investigate vowel quality and quantity, pitch, timing and prosody, or other features.

For YouTube, the PEASYV tool (Phonetic Extraction and Alignment of Subtitled YouTube Videos, Méli 2023) provides for individual videos functionality similar to that of the pipeline described in this paper. PEASYV makes use of yt-dlp and aligns transcripts with the Penn Forced Aligner and SPPAS. Source code for the tool, however, is not available, as of mid-2023. Notable is also youglish.com, a service through which users can search YouTube ASR transcripts for specific utterances; links to the utterance in YouTube videos are returned.[9]

4. Pipeline components

The pipeline has been provided as a Jupyter Notebook hosted on GitHub which can be run on the Google Colab service. Due to restrictions on user accounts imposed by the database underlying the Montreal Forced Aligner, using the pipeline on a local or cloud machine may be more efficient than Colab for extensive data collection.

4.1 Yt-dlp

Yt-dlp is a fork of YouTube-DL, an open-source library for accessing YouTube or other streamed content. The fork provides some additional functionality, compared to the original library, and can be used to retrieve content not only from YouTube, but from many websites that stream using DASH or HLS protocols, including broadcasters, social media, and content sharing websites.

The yt-dlp component of the pipeline extracts ASR transcripts for video(s) of interest; these are tokenized and then converted to either a format in which the transcript is rendered as a standard text or a format in which each word token has timing information appended in the form `word_1.00`, where the numerical value indicates the time offset in seconds from the start of the corresponding video. SpaCy can be used in the pipeline for part-of-speech tagging.[10] The script works "out of the box" for any of the languages for which YouTube provides ASR captions.[11]

Texts prepared with word timing information in this manner can then be used to extract audio or video content from the corresponding videos, again using yt-dlp. With regular expressions, specific lexical items, word sequences, speech acts, or exchanges can be targeted for audio or video extraction. The pipeline script uses the timing information to retrieve the corresponding audio segment and transcript fragment for a variable-length "window" around the targeted word sequence: for example, if the regular expression targets the sequence "need to", the window can be set to capture (three words) + "need to" + (three words), resulting in hits such as "then if we need to ask about the".[12]

4.2 Montreal Forced Aligner

The extracted text fragment and its corresponding audio segment are aligned with the Montreal Forced Aligner, using an acoustic model trained on the librispeech dataset (Panayotov et al. 2015). The output is Praat TextGrid files which contain the exact start and end times for the words and phones within the corresponding audio; phones are represented with the ARPA dictionary (Gorman & Howell 2011).

4.3 Parselmouth (Praat)

Parselmouth (Jadoul et al. 2018) is a Python port of functions from Praat. In the exploratory analysis in Section

[9] Youglish uses YouTube's API and automatically-generated metadata to associate individual videos with English varieties (American, British, Australian, etc.). The service provides access to the videos at YouTube's website, but audio and video content are not available for download and further processing such as forced alignment without using additional tools.

[10] With the en_core_web_sm model (https://spacy.io/usage/models).

[11] As of mid-2023, English, Dutch, French, German, Italian, Japanese, Korean, Portuguese, Russian, and Spanish.

[12] Aligning shorter segments prevents ASR or other errors from causing cascading alignment errors in the entire video. A window length from approximately 7–20 words was found to be effective.

5, Parselmouth is used to measure formant frequencies, but the software can be used to investigate other acoustic phenomena pertaining to the speech signal as well, such as pitch, intensity, timing phenomena, stress, or intonation. An advantage of using Parselmouth, compared to standalone Praat, is Python integration: while shell scripts can be used to pass data from Python to Praat, the process can be cumbersome, and integration of Praat functions, via Parselmouth, into common Python development environments such as Jupyter can facilitate analysis and visualization workflows.

5. /eɪ/ Nuclei in North America

This section describes an exploratory analysis of regional phonetic variation undertaken using the pipeline.

The Corpus of North American Spoken English (Coats 2023), a 1.3-billion-word corpus of geolocated YouTube ASR transcripts, was used as a starting point for extraction of /eɪ/ diphthongs. A regex script targeted monosyllabic words in CoNASE containing /eɪ/ and extracted a seven-word span of transcript and audio from the corresponding videos. These alignments were used to extract F1 and F2 formant values at nine measurement points during vowel duration for the monophthongs and diphthongs of American English.

Figure 1 demonstrates the results for videos from a single YouTube channel, that of the municipality of Hendersonville, Tennessee. The figure shows the trajectories in formant space for /eɪ/, as well as the diphthongs /aʊ/ and /oʊ/, for 10,745 vowel tokens extracted from 133 videos. Each circle represents a single measurement in F1/F2 space. The size of circles shows the number of measurements at the corresponding duration quantile. The mean trajectories of the diphthongs correspond to line segments joining the centers of the individual measurement points for that diphthong.

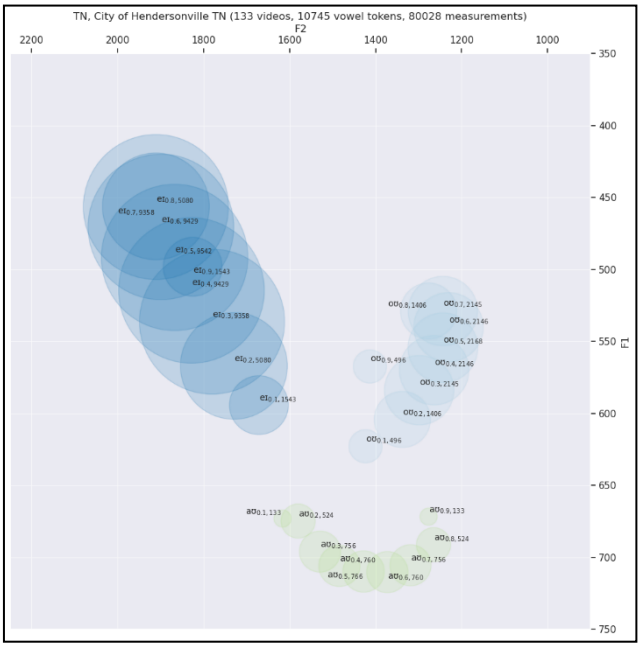

Figure 1: /eɪ/, /aʊ/, and /oʊ/ diphthongs for Hendersonville, TN

This level of detail allows the analyst to consider characteristic qualities of vowels in different regions or locations.

Integration of the pipeline into Python development environments makes it possible to create interactive visualizations as well. Figure 2 is a screenshot of an interactive visualization of a sample of /eɪ/ diphthongs from another Tennessee locality, the town of Gallatin. [13] Diphthong trajectories for individual tokens are represented as lines; the circles on each line mark the measurements at the corresponding quantile. Users interacting with the plot can click on a line to hear the diphthong; the plot can be used to demonstrate relative closedness and backness of /eɪ/ for many speakers from this locality (and elsewhere in the American Upper South).

From a broader geographical perspective, the formant extraction procedure can provide an overview of variation in the phonemic inventory of American English. Figure 3 shows the Getis-Ord Gi* value for the F2 value of the first target of the /eɪ/ diphthong, based on almost 9 million vowel tokens. As can be seen, the diphthong nucleus is somewhat more back in the American Southeast, but more front in the upper Midwest, Canada, and Southern California. This pattern largely corresponds to our knowledge of the distribution of formant values for this diphthong (e.g., Labov et al. 2006: 94; Grieve et al. 2013: 49), providing a preliminary confirmation of the validity of the phonetic extraction pipeline.

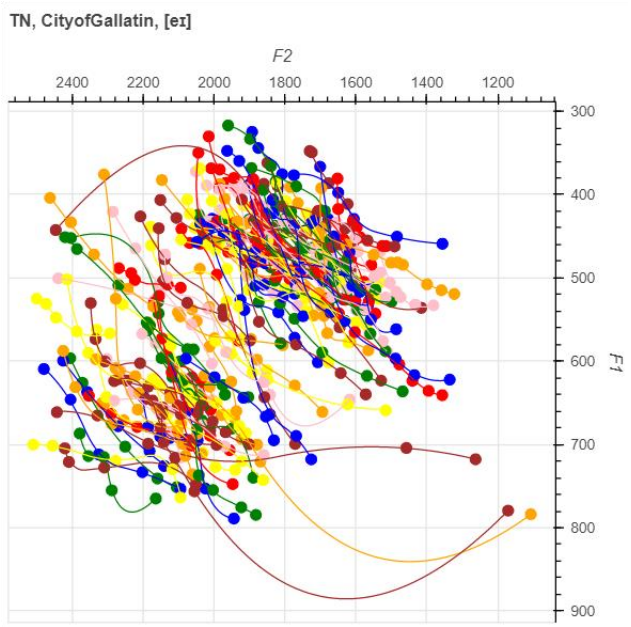

Figure 2: Screenshot of interactive /eɪ/ formant tracks for Gallatin, TN

[13] https://cc.oulu.fi/~scoats/example_Gallatin_all.html

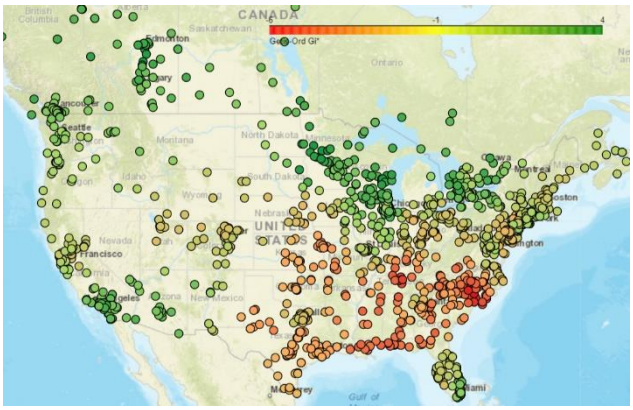

Figure 3: Getis-Ord Gi* values for F2 nucleus of /eɪ/ diphthong (8,788,999 tokens)

6. Summary and Outlook

The acoustic analysis pipeline utilizes components from yt-dlp, the Montreal Forced Aligner, and Parselmouth-Praat, and can be used to harvest transcript and acoustic data from YouTube. Content from other websites that utilize the common streaming protocols can also be harvested, including video data. The pipeline can be used to create custom corpora for acoustic and multimodal analysis, or can serve as the starting point for acoustic analyses of large existing corpora of YouTube transcripts, such as CoNASE or CoBISE (Coats 2023, 2022). The pipeline represents a potentially useful framework for the creation of corpora and the acoustic analysis of naturalistic speech from a range of geographical contexts, content types, and pragmatic situations.

7. References

Bigi, B. (2015). SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech. *The Phonetician - International Society of Phonetic Sciences* 111, 54–69.

Boersma, P. & Weenink, D. (2023). *Praat: doing phonetics by computer* [Computer program]. Version 6.3.09. http://www.praat.org

Coats, S. (2023). Dialect corpora from YouTube. In Beatrix Busse, Nina Dumrukcic, and Ingo Kleiber (eds.), *Language and linguistics in a complex world*, 79–102. Berlin: de Gruyter. https://doi.org/10.1515/9783111017433-005.

Coats, S. (2022). The Corpus of British Isles Spoken English (CoBISE): A new resource of contemporary British and Irish speech. In Karl Berglund, Matti La Mela, and Inge Zwart (eds.), *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference, Uppsala, Sweden, March 15–18, 2022*, 187–194. Aachen, Germany: CEUR. http://ceur-ws.org/Vol-3232/paper15.pdf

Gonzalez, S., Grama, J. & Travis, C. (2020). Comparing the performance of forced aligners used in sociophonetic research. *Linguistics Vanguard* 5. https://doi.org/10.1515/lingvan-2019-0058

Gorman, K. & Howell, J. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics* 39(3), 192–193.

Grieve, J., Speelman, D. & Geeraerts, D. (2013). A multivariate spatial analysis of vowel formants in American English. *Journal of Linguistic Geography* 1, 31–51. https://doi.org/10.1017/jlg.2013.3

Jadoul, Y., Thompson, B. & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71, 1–15. https://doi.org/10.1016/j.wocn.2018.07.001

Kisler, T., Reichel, U. D. & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language* 45, 326–347.

Labov, W., Ash, S. & Boberg, C. (2006), *The Atlas of North American English*. Berlin: Mouton de Gruyter.

Lee, A. & Kawahara, T. (2009). Recent development of open-source speech recognition engine Julius. In *Proceedings of APSIPA ASC 2009*, pp. 131–137.

MacKenzie, L. & Turton, D. (2020). Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard* 6, no. s1. https://doi.org/10.1515/lingvan-2018-0061

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of the 18th Conference of the International Speech Communication Association*.

Méli, A. (2023). *PEASYV: Phonetic Extraction and Alignment of Subtitled YouTube Videos*. https://adrienmeli.xyz/peasyv.html

Panayotov, V., Chen, G., Povey, D. & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G. & Vesely, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Hilton Waikoloa Village, Big Island, Hawaii, US*. IEEE Signal Processing Society.

Reddy, S. & Stanford, J. (2015). A Web Application for Automated Dialect Analysis. In *Proceedings of NAACL-HLT* 2015.

Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H. & Yuan, J. (2014). *FAVE (Forced Alignment and Vowel Extraction) Program Suite v1.2.2* https://doi.org/10.5281/zenodo.22281

Schiel, Florian. (1999). Automatic phonetic transcription of non-prompted speech. In *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS)*, 607–610.

Sodagar, I. (2011). The mpeg-dash standard for multimedia streaming over the internet. *IEEE multimedia*, 18(4), 62–67.

Young, S. J. (1994). *The HTK hidden Markov model toolkit: Design and philosophy*.

Yuan, J. & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics '08*.