Steven Coats

An automatic pipeline for processing streamed content: New horizons for corpus linguistics and phonetics

Abstract: Large volumes of audio and video data are accessible through video sharing sites, streaming services, and social media platforms, but until recently, relatively little of this content has been utilized as research data for large-scale studies of grammatical or phonetic variation. This chapter discusses a notebook-based pipeline designed to analyze phonetic data from online video content, made possible by recent advances in language technology such as improvements in automatic speech recognition and forced alignment. It provides an overview of open-source frameworks for working with speech data, noting that while several tools have been developed to handle some or all of these tasks, their installation and setup may be complex and incompatibility issues may arise. Notebook-based pipelines, increasingly used in all fields of data science, offer the advantages of flexibility and adaptability. In this chapter, we introduce the Video Phonetics Pipeline (ViPP) for the extraction and analysis of audio and transcript data from video and streaming sites such as YouTube, 𝕏, TikTok, and many others, a pipeline which leverages functions from the open-source Python library yt-dlp to retrieve data, then utilizes the Montreal Forced Aligner to align audio with text. Formants are measured with Praat-Parselmouth, and packages from Python's standard library can be used for statistical analysis and visualization. The script pipeline, available as a notebook at GitHub and in a Google Colab environment, is customizable. The utility of the pipeline is demonstrated with an example: a consideration of diphthong trajectories in contemporary North American English, based on data from the Corpus of North American Spoken English (CoNASE).

Keywords: Corpus linguistics, phonetics, vowels, formants, YouTube, DASH, CoNASE, forced alignment, World Englishes

Steven Coats, University of Oulu, Finland, e-mail: steven.coats@oulu.fi

https://doi.org/10.1515/9783111434018-011

1 Introduction

The creation of speech corpora has traditionally required significant expenditure in terms of person-hours and resources, comprising collection of audio data in the form of targeted individual recordings, often in different locations, and time-consuming manual transcription of those recordings. In the past 15 years, however, it has become increasingly feasible to collect high-quality naturalistic speech data from online sources, and advances in automatic speech recognition (ASR) algorithms have greatly facilitated the preparation of orthographic transcripts, developments which are ongoing and are expected to contribute to the burgeoning field of corpus phonetics (Liberman 2019).[1]

These new perspectives make it possible to analyze linguistic variation by using automated scripting pipelines which collect, transcribe, process, and analyze speech produced in different locations, interaction contexts, or by different social groups. Compared to traditional speech corpora, much larger volumes of data can be collected, potentially enabling analysis of constructions that are rare in spoken language. Online collection and processing of speech data via pipelines is also facilitated by changes in the way people communicate: In the past 15 years, there has been an explosion in the online availability of video content, a shift which reflects the increased use of video sharing and streaming in computer-mediated communication (CMC) environments and on social media platforms. While traditional CMC formats such as mailing lists, discussion forums, and chat rooms still exist, they now typically also can include multimedia content such as embedded videos or sound files. These changes in online communication behavior are concomitant with and ultimately result from advances in the underlying communication technologies: increases in bandwidth availability and data transmission capabilities, increases in processing speed and memory which allow large files such as videos to be efficiently processed, larger storage capacities on servers, standardization of audio and video codecs, and standardization of technical protocols for video streaming under variable bandwidth conditions. Multimedia content, or simultaneous use of text, speech, and video, is the default communicative setting for CMC on popular platforms such as YouTube, Facebook, Twitch, or TikTok, whether as live streams or as recorded videos.

As of early 2024, much video content is shared using one of two transmission protocols: DASH (Dynamic Adaptive Streaming over HTTP; Sodagar 2011) and HLS (HTTP Live Streaming). These standards serve content as sequentially ordered data chunks via automatically generated URLs; the chunks are consumed by the

1 This chapter is a revised and expanded version of Coats (2023c).

end-user's browser and processed as they arrive. Depending on bandwidth availability, the protocol will serve lower-quality (i.e., requiring less data) or higher-quality (i.e., requiring more data) video and audio to the browser. Textual content such as concurrent chat interaction or comments require less bandwidth; these are also served to the end user via DASH and HLS.

From the perspective of linguistic research, the standardization of these protocols and their widespread use mean that the data is available for harvesting and analysis for anyone with an internet connection. Depending on the configuration of the processing pipeline and the individual components that are included, the researcher has access not only to audio data for phonetic analysis, but also to various ASR or manually generated transcripts, as well as to video content.[2] Transcripts can be analyzed in terms of lexis, grammar, syntax, and discourse content, for example for sociolinguistic or geolinguistic/dialectological studies, and acoustic properties of the audio can be analyzed for vowel quality and quantity, pitch, prominence, or other phonetic and prosodic phenomena. The automated analysis of video-recorded nonverbal concomitants of spoken interaction such as facial expression, gesture, kinesics, or proxemics is still in its infancy, but the relatively new field, related to *social signal processing* (Vinciarelli et al. 2009), is likely to develop rapidly in coming years.

In this chapter, existing tools and approaches for working with data of this type are briefly reviewed. Although a variety of open-source tools for the management and analysis of phonetic corpora exist, a pipeline-based approach can be well suited for collection, annotation, and analysis of online speech. The Video Phonetics Pipeline (ViPP)[3] is a Python-based set of scripts that can be implemented quickly, without lengthy setup, in a Jupyter Notebook or a cloud computing environment such as Google's Colaboratory. The pipeline is designed specifically to access YouTube content, but with minor modifications can also retrieve content from other platforms by implementing widely used open-source tools and code libraries or packages. For content download of audio, video, transcript, comment, or chat data, the pipeline makes use of yt-dlp,[4] as of early 2024 the most popular Python library for harvesting YouTube content. The Montreal Forced Aligner[5] (MFA; McAuliffe et al. 2017a) is used to align transcript content with the audio signal in the down-

2 The legal contexts pertaining to copyright, fair use, and GDPR legislation are not discussed in this chapter. For a discussion of some of these issues for data collected from YouTube, please see Coats (2023b).

3 https://github.com/stcoats/phonetics_pipeline (last accessed 14 February 2025).

4 https://github.com/yt-dlp/yt-dlp (last accessed 14 February 2025).

5 https://montreal-forced-aligner.readthedocs.io (last accessed 14 February 2025).

loaded video/audio files. For the extraction of phonetic features, Python bindings for functions from the widely used Praat software (Boersma and Weenink 2023) are implemented from the Parselmouth-Praat package (Jadoul et al. 2018).[6]

The modular nature of ViPP makes it suitable for adaptation and modification for a variety of data collection and analysis tasks. For example, the script can target YouTube's own ASR captions, or manually uploaded captions. Content from platforms other than YouTube, such as videos uploaded to Twitter, Twitch, or national broadcasters such as ARD or the BBC can be retrieved. If transcripts are unavailable, the pipeline can be modified to incorporate an ASR module such as Whisper (Radford et al. 2022) or WhisperX (Bain et al. 2023). The Montreal Forced Aligner can utilize specific acoustic models, grapheme-to-phoneme models, and language models, depending on the needs of the project at hand. For phonetic analysis, Parselmouth-Praat allows virtually all of the functions in Praat to be applied. Visualization can be undertaken using widely employed packages such as Matplotlib (Hunter 2007) or Seaborn (Waskom 2021).

The remainder of the paper is organized as follows: The second section reviews some tools, architectures, and pipelines used for ASR, forced alignment, and acoustic analysis. The third section discusses the architecture of ViPP, as well as alternative implementations for specific tasks that incorporate different components. Section 4 describes a short exploratory analysis that illustrates the utility of the pipeline: the trajectory of F1 and F2 formants for the /eɪ/ diphthong is plotted for videos indexed in the *Corpus of North American Spoken English* (Coats 2023a). In the fifth section, an overview and a summary are provided and the outlook for future developments for ViPP and for similar pipelines is discussed.

2 Previous work

2.1 Software frameworks and tools

Several comprehensive free or open-source software packages for acoustic and phonetic analysis have been developed. Most tools and software for forced alignment are built on one of two frameworks: the Hidden Markov Model Toolkit (HTK, Young 1993)[7] and Kaldi (Povey et al. 2011).[8] The Penn Forced Aligner, P2FA (Yuan

6 https://github.com/YannickJadoul/Parselmouth (last accessed 14 February 2025).
7 https://htk.eng.cam.ac.uk (last accessed 14 February 2025).
8 http://kaldi-asr.org (last accessed 14 February 2025).

and Liebermann 2008), is based on HTK. It serves as the basis for forced alignment tools that have been widely used in phonetics in the last 15 years, including FAVE (Forced Alignment and Vowel Extraction, Rosenfelder et al. 2014), a Python package that, in addition to calling P2FA, can also extract vowel formant values. MAUS, or the Munich Automatic Segmentation tool (Schiel 1999), uses P2FA to align audio and text files; the web implementation WebMAUS (Kisler et al. 2017) can handle different languages and dialects of German or English by employing different underlying acoustic and grapheme-to-phoneme models. The output of MAUS and WebMAUS can be rendered as Praat TextGrid files or in other formats such as EXMARaLDA's .exb or .flk, ELAN's .eaf, .json, .xml, or .csv files.

Somewhat similar to WebMAUS, the DARLA (Dartmouth Linguistic Annotation, Reddy and Stanford 2015) framework is a website that can automatically align audio files that have been uploaded together with orthographic transcript files. The system sends user-uploaded files to the Montreal Forced Aligner for alignment and then to FAVE for vowel extraction and formant measurement; normalization and visualization (for example of vowel locations in F1/F2 formant space) are handled by the R package *vowels* (Kendall and Thomas 2010). In addition, DARLA can generate ASR transcripts from audio files by using Deepgram, a paid service.

An additional framework used for speech recognition and alignment is Julius (Lee et al. 2001; Lee and Kawahara 2009), which provides the basic underlying signal processing and acoustic modeling framework for the SPPAS software suite (Speech Phonetization Alignment and Syllabification, Bigi 2015). SPPAS can be used for alignment, annotation, and other tasks. Several other aligners are noted by Pettarin (2022).

The Language, Brain and Behaviour Corpus Analysis Tool (LaBB-CAT, Fromont and Hay 2012; Fromont 2019), developed for the Origins of New Zealand English Corpus, is a browser-based environment, implemented in Java, that powers an Apache Tomcat server and a MySQL database on a local installation. The system handles management, analysis, and visualization of audio files, transcripts, and annotations. Forced alignment can be undertaken in LaBB-CAT using a local installation of HTK and the CELEX dictionary for pronunciations (Baayen et al. 1996), as well as other pronunciation dictionaries. LaBB-CAT provides extensive search and visualization functionality, and Praat scripts can be used to analyze transcripts and audio data. Additional linguistic annotation tasks can be implemented with scripts that call third-party tools.

The Emu speech corpus database system (Cassidy and Harrington 1996) was developed to organize and provide query functionality to recorded speech data with multiple levels of annotation. Emu has been refined and developed over the years, resulting in an R-based tool suite comprising several libraries (Winkelmann et al. 2017) as well as a web application for visualization, annotation, and analysis,

the EMU-WebApp;[9] collectively, these comprise the EMU-SDMS (Speech Database Management System). While EMU-SDMS is suitable for a range of visualization and analyzation tasks, it is not designed for retrieval of online video or audio content, ASR, or forced alignment.

The PolyglotDB system (McAuliffe et al. 2017b) is a database for corpus-phonetic management, written mostly in Python, which enables a variety of analysis tasks from data with various input formats. The related Integrated Speech Corpus Analysis (ISCAN) platform (McAuliffe et al. 2019), similar in some ways to EMU-SDMS, provides extensive functionality for visualization and phonetic analysis. ISCAN, available in a dockerized container from source files hosted on GitHub, creates a browser-based interface in which queries and functions from PolyglotDB are automated for ease of use.[10] PolyglotDB and ISCAN notably include functions for formant extraction which automatically discard formant tracking errors, as described in Mielke et al. (2019). While PolyglotDB and ISCAN provide extensive functionality, setup may be complicated due to many possible dependency and installation issues that can arise, and the tools are not designed for the purposes of online content harvesting, ASR, or forced alignment.

Additional tool suites that allow organization, transcription, search functionality, visualization, and analysis of speech corpora include EXMaRALDA (Schmidt and Wörner 2014) and ELAN (Wittenburg et al. 2006), developed specifically for annotation and analysis of video data. *Visible Vowels* (Heeringa and Van de Velde 2018) is a site built using Shiny in R that can perform various types of analysis and visualization of vowels for files containing speaker, vowel, timing, duration, and formant information that have been uploaded in Excel format.[11]

For high-quality audio, accurate transcripts, and well-resourced languages, HTK- and Kaldi-based aligners can produce alignments that are generally comparable in quality to those produced by human annotators. DARLA, which uses MFA for alignment, and FAVE, which uses P2FA, both generate accurate alignments for regional British English speech (MacKenzie and Turton 2020), despite the acoustic models and phonemic representation dictionaries not having been trained on those specific varieties. Similarly, MFA can generate accurate alignments of Australian English speech, even when using the default American English language models and phoneme-grapheme dictionaries (Gonzalez et al. 2020).

9 https://ips-lmu.github.io/EMU-webApp (last accessed 14 February 2025).

10 As of early 2024, the dockerfile and requirements.txt files for ISCAN need manual editing in order to be launchable and the resulting docker environment may generate errors due to package inconsistencies.

11 https://www.visiblevowels.org/ (last accessed 14 February 2025).

2.2 Pipeline approaches

Convergence of tools has resulted in the development of similar approaches, often making use of core functionalities of HTK- or Kaldi-based aligners and Praat (Boersma and Weenink 2023) for acoustic analysis. Specifically for YouTube, the PEASYV tool (Phonetic Extraction and Alignment of Subtitled YouTube Videos; Méli and Ballier 2023; Méli et al. 2023)[12] utilizes yt-dlp-based data collection, then alignment with P2FA and SPPAS; acoustic analysis is conducted with Praat scripts. Ahn et al. (2023) used a pipeline comprising Praat and Python scripts to identify outlier values in vowel formant measurements values for several speech corpora. A number of projects have developed and documented automated pipeline approaches for the acoustic analysis of World Englishes (e.g., Fuchs 2023; Meer 2020; Meer et al. 2021).

Recent approaches have also incorporated the general-purpose speech recognition model Whisper (Radford et al. 2022) into speech processing pipelines for linguistic analysis. Whisper can generate high-quality ASR transcripts in multiple languages, for example on the multilingual Fleurs dataset (Conneau et al. 2022). As of early 2024, transcriptions generated by Whisper contain timestamps indicating the start and end of utterance chunks of variable length, ranging from one to twenty or more words; word timestamps can also be generated. Although the transcription accuracy of Whisper ASR is high, especially for the large models, the word timing information can be inaccurate and is not immediately suitable for further phonetic processing tasks such as forced alignment. WhisperX (Bain et al. 2023) is a set of tools and a Python package that generates word-level alignment and speaker diarization from Whisper output. The package harnesses other open-source models and repositories such as Wav2Vec2 (Baevski et al. 2020) for word and phone alignment and Pyannote.audio for speaker diarization (Bredin 2023; Plaquet and Bredin 2023). Likewise, these packages build upon algorithms, models, and training data sets that have been made available to the research community at large, such as the AVA-AVD dataset (Xu et al. 2022). Pipelines for automatic analysis of pause and lexical stress have also been developed, incorporating Whisper, WhisperX, Pyannote, MFA, and other tools (e.g., Coulange et al. 2023).[13] As of 2024, the use of WhisperX in phonetics research is ongoing in several projects. WhisperX can easily be integrated into notebook-based pipelines such as ViPP.

The possibilities offered by new tools and models have been embraced by researchers, but audio from online sources such as videos may be vulnerable to

12 https://adrienmeli.xyz/peasyv.html (last accessed 14 February 2025).

13 https://gricad-gitlab.univ-grenoble-alpes.fr/lidilem/plspp (last accessed 14 February 2025).

measurement errors. Formant frequencies can be affected by the acoustic properties of the recording space, and algorithms may have difficulties reliably detecting formants at low and high frequencies (Aalto et al. 2018). The suitability of audio data collected under highly variable recording conditions has been investigated in several recent studies. Freeman and de Decker (2021a) compared the audio quality of vowels and nasals from recordings made on smartphones, tablet devices, and laptops with recordings made on professional equipment in a studio environment. Recordings from personal devices were mostly able to recapitulate the major divisions of the vowel space "relatively faithfully". Similarly, in Freeman and de Decker (2021b), audio from video conferencing platforms was found to be mostly suitable for sociophonetic analysis, albeit with the caveat that measurement points for low back vowels exhibited considerable variability. Conklin (2023) compared vowel reduction in lossless recordings undertaken in a controlled studio environment with lossless recordings from smartphones and lossy recordings from laptops made via a website interface. She found that the different recording setups and audio compression settings result in values that are generally reliable for coarse comparisons but are not suitable for fine comparisons requiring precision.

Overall, a wide variety of tools for the collection, processing, alignment, and analysis of speech have been developed, and continued advancements in the application of neural networks and large acoustic and language models have resulted in new possibilities for phonetic research. Still, in some cases, existing tools are difficult to install and setup due to dependency incompatibilities, or are not well suited for collection of online data. The next section describes a notebook-based pipeline that can be used out-of-the-box.

3 ViPP Notebook

The Video Phonetics Pipeline (ViPP) was developed as a Jupyter Notebook that offers the analyst a fast means of retrieving, accessing and analyzing audio from online video without requiring extensive installation of software or tools. The pipeline is hosted on GitHub and designed to run on Google's Colab service or comparable cloud computing environments. Its main functionality comprises use of the open-source Python libraries yt-dlp and Praat-Parselmouth (Jadoul et al. 2018); alignment is achieved with a temporary local installation of the Montreal Forced Aligner in a Miniconda environment.

ViPP retrieves YouTube ASR transcripts and audio with functions from yt-dlp, a fork of the popular YouTube-DL library in Python. The pipeline's default settings retrieve, for a given video, the highest-quality audio available and convert it to .wav

format, if necessary, using ffmpeg. Transcripts are converted from .vtt files to one of two formats: a string representing the orthographic transcription, or a string in which each word token has attached timing information, which can be used if the pipeline is modified to target utterances or sequences. By specifying the language of transcripts to be targeted, the script can be used with videos in languages for which YouTube provides ASR captions: as of early 2024, English, Dutch, French, German, Italian, Japanese, Korean, Portuguese, Russian, and Spanish. If desired, part-of-speech annotation can be implemented using models from SpaCy.[14]

The pipeline scripts install the Montreal Forced Aligner in a local Miniconda environment and retrieve a pronunciation dictionary and an acoustic model for English.[15] Calling the aligner will analyze and align the converted ASR transcript with the .wav file, outputting files in Praat's .TextGrid format.

Textgrid files, together with the corresponding .wav files, can then be used to examine acoustic properties of speech segments by using functions from Praat-Parselmouth. The default code in ViPP measures F1 and F2 formant values, but other acoustic properties can also be measured with minor changes to the code. ViPP's formant extraction approach uses the default Praat parameter values to retrieve F1 and F2 at a monophthong's durational midpoint, as determined by the Montreal Forced Aligner. Formant values can then be plotted in F1/F2 space for a single or for multiple videos using Mahalanobis distance to exclude outliers. For analyses of format trajectories, multiple measurement points can be used.

4 Example: Diphthong trajectory in F1/F2 space

Recent studies have sought to characterize vowel quality in terms of dynamic trajectories, rather than as single measurement points for monophthongs or onset/target measurement points for diphthongs (see, e.g., Fox and Jacewicz 2009; Sóskuthy et al. 2019; Renwick and Stanley 2020). The Video Phonetics Pipeline can be used to quickly assemble data for comparison from YouTube videos, then visualize and assess diphthong trajectories for locations or social groups.

As an example, Figure 1 shows the trajectories of 92 tokens of /eɪ/ extracted from YouTube videos uploaded to the channel of the city of California City, California. These tokens, which were filtered on the basis of having at least 5 measure-

14 For example, the en_core_web_sm model (https://spacy.io/usage/models, last accessed 14 February 2025).

15 Available models are described and can be downloaded from https://mfa-models.readthedocs.io/en/latest/ (last accessed 14 February 2025).

ment points as well as monotonic decreases in F1 values and increases in F2 values, are represented by dashed lines, with circles showing the values at individual measurement points, which are evenly distributed throughout the duration of the phone. The black line shows the mean diphthong trajectory for the 92 tokens. The exploratory visualization implies a bimodal distribution for the values which may correspond to the sex of the speakers in the sampled videos. Such analyses can serve as the starting point for comparisons of diphthong trajectories for different social groups or in different locations.

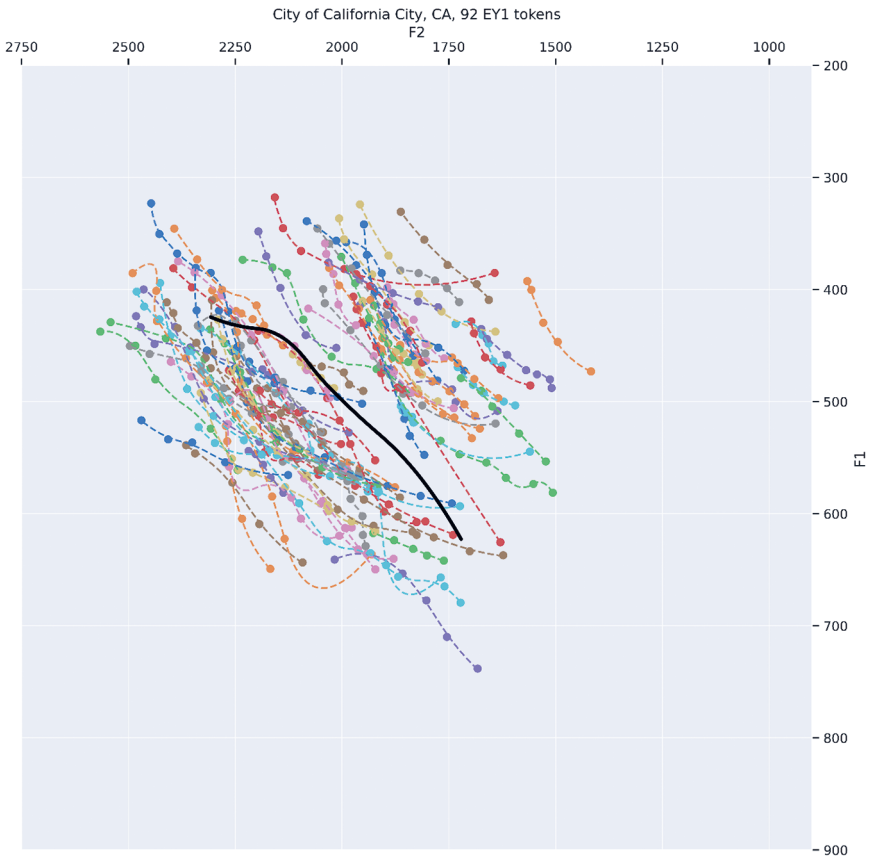

Figure 1: /eɪ/ trajectories from the YouTube channel of California City, California.

With minor modifications and the use of functions from interactive visualization libraries in Python such as Bokeh (Bokeh Development Team 2024), the ViPP can

also render interactive visualizations that play the audio for a token upon a mouse click or rollover.[16]

The pipeline can be used to extract formants from large numbers of videos in order to (for example) gauge regional variation in vowel quality. Figure 2 shows the values of a spatial autocorrelation statistic, the Getis-Ord G_i^* (Getis and Ord 1992; Ord and Getis 1995), for F2 values of the onset of the /eɪ/ diphthong, based on millions of vowel tokens from videos uploaded by American local government YouTube channels (see Coats 2023a). Each point on the map represents a location in which at least 100 tokens were sampled; a 20-nearest-neighbors binary spatial weights matrix was used to calculate the statistic on the basis of the mean formant value at each location. As can be seen in Figure 2, /eɪ/ onsets are more back in the American Southeast, and more front in the upper Midwest, Canada, and Southern California, a pattern which corresponds to intuitions about American dialects as well as quantitative findings (e.g., Labov et al. 2006: 94; Grieve et al. 2013: 49).

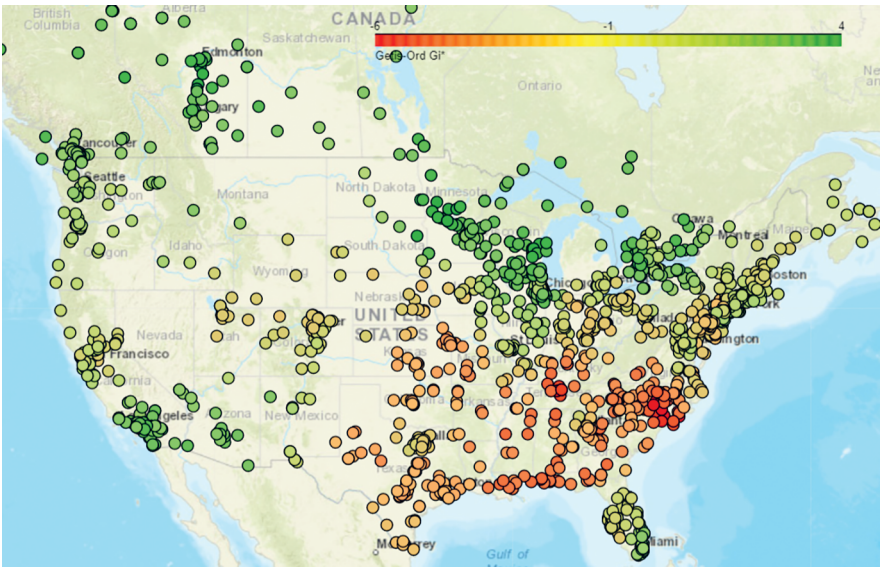

Figure 2: Getis-Ord Gi* values for F2, onset of /eɪ/ diphthong (8,788,999 tokens).

16 An example, from the YouTube channel of a town in Tennessee, can be found at https://cc.oulu.fi/~scoats/example_Gallatin_all.html (last accessed 14 February 2025).

5 Discussion

Many open-source libraries, frameworks, and tools have been developed to facilitate corpus creation and phonetic analysis, but the tools themselves, as well as internet data transmission protocols, programming languages, and operating systems, are in a constant state of flux. Not all tools and frameworks are robust to changes in underlying operating system architecture or package dependencies. Open-source software may stop working for any number of reasons, but common causes include incompatible dependencies (i.e., the software requires a newer or older version of a package than what is installed), syntax changes in the underlying programming language that may introduce conflicts (e.g., Python 3.11 instead of 3.4, or Python 3 instead of 2), non-portability of code to different operating systems, or incompatibility of code in different OS environments of the same OS due to different availability of packages (for example, Ubuntu vs. Red Hat Linux). Developers of operating systems, programming languages, and libraries/packages, as well as authors of software tools for linguistic analysis, do their best to ensure compatibility when new versions are introduced, but some problems are inevitable. Open-source software, including linguistic software, may not be actively maintained. The team needed to maintain and support the software may have run out of funding. Team members may have moved on to different institutions or to non-academic jobs and no longer have the time to maintain an older software package. Institutions such as universities and libraries may no longer be able to provide server space to host necessary parts of the infrastructure. Many other possibilities are conceivable.

Notebook-based approaches such as ViPP can address some of these issues:

- Notebooks are portable and are relatively easy to implement under different operating systems and Python/R versions.
- Notebooks may not require lengthy and time-consuming installation and configuration of complex underlying dependencies such as database or web server software.
- Many users may already be familiar with Python and R and thus be able to follow and modify code cells.
- Data collection and analysis tasks which are divided into modular code blocks, implemented as notebook cells, are easier to customize and modify in the case of problems, compared to stand-alone programs run from the command line or in a custom interface.
- Notebooks designed to run in cloud-based environments may be less subject to dependency or incompatibility issues, compared to more static scripts and tools. A notebook can be designed to install and use software packages and library

versions which are mutually compatible in the local operating environment, for example. Colab automatically uses a recent, stable version of Linux and a recent Python kernel, and the most widely used packages are automatically installed in the environment.
– Using a notebook in a cloud-based environment generally does not require administrator knowledge (or system privileges), and data collection, analysis, and visualization can be done almost immediately.

The use of notebooks is not without its own set of problems, which may include missing documentation for code in cells, lack of modularity for scripts, or unclear/incompatible dependency declarations, among others (Pimentel et al. 2021). In addition, notebook setups may offer only limited functionality compared to dedicated software platforms. ViPP, for example, does not implement syllabification of input data. Depending on the local settings, a notebook may not be suitable for long-running tasks or for processing large amounts of data. Google's default access to the Colab service has limitations on runtime, processor, and memory availability. In addition, the processing of data on commercial platforms such as Colab may introduce privacy and copyright issues that need to be carefully considered before research is undertaken.

Nevertheless, despite these limitations, notebook-based data collection, processing, and analysis approaches may offer an expedient means to quickly retrieve and analyze linguistic data. Especially for YouTube content, ViPP provides a framework which can be implemented immediately, allowing the analyst to focus on linguistic phenomena, rather than troubleshooting the installation of open-source phonetic analysis software.

6 Summary and outlook

Software and tools for linguistic and phonetic analysis change and evolve rapidly. For some data collection and analysis tasks, a notebook-based approach may be suitable. The Video Phonetic Pipeline is a Python notebook that incorporates functionality from yt-dlp, the Montreal Forced Aligner, and Parselmouth-Praat to harvest transcript and audio data from YouTube videos. With minor modifications, the pipeline can be adapted to collect data from other platforms. ViPP can be used for creation of small, specialized corpora from YouTube content as well as for larger corpora of YouTube transcripts and audio (Coats 2023c, 2024). The pipeline, and the notebook-based approach in general, represent a framework for the creation, processing, and analysis of online data for a diverse range of content

types which is compatible with the general trend towards use of cloud-based services and tools for data analysis, rather than processing with software installations on local machines.

As notebooks are by design customizable, recent AI models such as Whisper or WhisperX for automated ASR transcript generation and diarization can be incorporated into the pipeline. Additional tools for specific speech processing tasks can be included, for example with models from Hugging Face. From the perspective of linguistic analysis, research involving the correlation of speech content or acoustic quality with automatically annotated facial expression, gestures, proxemics, or kinesics remain a relatively under-researched domain. Because data harvested from video platforms is fundamentally open, and considering the "generalizability of the body activity cues across datasets" (Beyan et al. 2023: 16), one future perspective may be to modify ViPP to incorporate sophisticated large models for tasks such as automated analysis of video content, including movement, gesture, or facial expression.

While these perspectives are expected to materialize in the future, the capabilities of ViPP, and the versatility of notebook approaches in general, offer practical utility for linguistic data collection and analysis tasks such as creation of transcript corpora and phonetic analysis of vowel space. In this context, ViPP shows potential for researchers aiming to quickly access interesting, new, or notable linguistic phenomena in the ever-growing universe of online content.

References

Ahn, Emily P., Gina-Anne Levow, Richard A. Wright & Eleanor Chodroff. 2023. An outlier analysis of vowel formants from a corpus phonetics pipeline. In *Proceedings of Interspeech 2023*.

Aalto, Daniel, Jarmo Malinen & Matti T. Vainio. 2018. Formants. In *Oxford research encyclopedia of linguistics*. Oxford: Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.419.

Baayen, R. Harald, Richard Piepenbrock & Leon Gulikers. 1996. *The CELEX lexical database* (cd-rom).

Baevski, Alexei, Henry Zhou, Abdelrahman Mohamed & Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv*:2006.11477 [cs.CL]. https://doi.org/10.48550/arXiv.2006.11477.

Bain, Max, Jaesung Huh, Tengda Han & Andrew Zisserman. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. In *Proceedings of Interspeech 2023*, 4489–4493. https://doi.org/10.21437/Interspeech.2023-78.

Beyan, Cigdem, Alessandro Vinciarelli & Alessio Del Bue. 2023. Co-located human-human interaction analysis using nonverbal cues: A survey. *ACM Computing Surveys* 56 (5). 1–41.

Bigi, Brigitte. 2015. SPPAS – Multi-lingual approaches to the automatic annotation of speech. *The Phonetician – International Society of Phonetic Sciences* 111. 54–69.

Boersma, Paul & David Weenink. 2023. *Praat: doing phonetics by computer* [Computer program]. Version 6.3.09. http://www.praat.org (last accessed 14 February 2025).

Bokeh Development Team. 2024. *Bokeh: Python library for interactive visualization*. http://bokeh.org (last accessed 14 February 2025).

Bredin, Hervé. 2023. Pyannote.audio 2.1 speaker diarization pipeline: Principle, benchmark and recipe. In *Proceedings of Interspeech 2023*, 1983–1987. https://doi.org/10.21437/Interspeech.2023-105.

Cassidy, Steve & Jonathan Harrington. 1996. Emu: An enhanced hierarchical speech data management system. In *Proceedings of the Sixth Australian International Conference on Speech Science and Technology*, 361–366.

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

Coats, Steven. 2023b. A new corpus of geolocated ASR transcripts from Germany. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-023-09686-9.

Coats, Steven. 2023c. A pipeline for the large-scale acoustic analysis of streamed content. In Louis Cotgrove, Laura Herzberg, Harald Lüngen, and Ines Pisetta (eds.), *Proceedings of the 10th International Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora 2023)*, 51–54. Mannheim: Leibniz-Institut für Deutsche Sprache. https://doi.org/10.14618/1z5k-pb25.

Coats, Steven. 2024. CoANZSE Audio: Creation of an online corpus for linguistic and phonetic analysis of Australian and New Zealand Englishes. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti & Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 3407–3412.

Conklin, Jenna. 2023. Examining recording quality from two methods of remote data collection in a study of vowel reduction. *Laboratory Phonology* 14 (1). https://doi.org/10.16995/labphon.10544.

Conneau, Alexis, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera & Ankur Bapna. 2022. FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech. *arXiv*:2205.12446 [cs.CL]. https://doi.org/10.48550/arXiv.2205.12446.

Coulange, Sylvain, Tsuneo Kato, Solange Rossato & Monica Masperi. 2023. *Comprehensibility diagnosis of spontaneous L2 English: Automated analysis of pausing and lexical stress patterns*. Paper presented at the workshop Tools in L2 research, November 2023, Zurich, Switzerland. http://i3l.univ-grenoble-alpes.fr/~coulangs/languages2023/CoulangeAl2023_Zurich.pdf (last accessed 14 February 2025).

Fox, Robert Allen & Ewa Jacewicz. 2009. Cross-dialectal variation in formant dynamics of American English vowels. *Journal of the Acoustical Society of America* 126 (5). 2603–2618.

Freeman, Valerie & Paul de Decker. 2021a. Remote sociophonetic data collection: Vowels and nasalization from self-recordings on personal devices. *Language and Linguistics Compass* 15. https://doi.org/10.1111/lnc3.12435.

Freeman, Valerie & Paul de Decker. 2021b. Remote sociophonetic data collection: Vowels and nasalization over video conferencing apps. *The Journal of the Acoustical Society of America* 149 (2). 1211–1223. https://doi.org/10.1121/10.0003529.

Fromont, Robert. 2019. Forced alignment of different language varieties using LaBB-CAT. In *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)*, 1327–1331.

Fromont, Robert A., & Jennifer Hay. 2012. LaBB-CAT: An annotation store. In *Proceedings of the Australasian Language Technology Association Workshop 2012*, 113–117.

Fuchs, Robert. 2023. Analysing the speech rhythm of New Englishes: A guide to researchers and a case study on Pakistani, Philippine, Nigerian and British English. In Guyanne Wilson & Michael Westphal (eds.), *New Englishes, New Methods*, 132–155. Amsterdam: Benjamins.

Getis, Arthur and Ord, J. Keith. 1992. The analysis of spatial association by use of distance statistics, *Geographical Analysis* 24 (7). 189–206.

Gonzalez, Simon, James Grama & Catherine E. Travis. 2020. Comparing the performance of forced aligners used in sociophonetic research. *Linguistics Vanguard* 5. https://doi.org/10.1515/lingvan-2019-0058.

Grieve, Jack, Dirk Speelman & Dirk Geeraerts. 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.

Heeringa, Wilbert & Hans Van de Velde. 2018. Visible Vowels: A tool for the visualization of vowel variation. In *Proceedings of the CLARIN Annual Conference 2018, 8 - 10 October, Pisa, Italy*, 120–123. CLARIN ERIC.

Hunter, John D. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 9 (3). 90–95.

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

Kendall, Tyler & Erik R. Thomas. 2010. *Vowels: Vowel manipulation, normalization, and plotting in R.* R package. https://cran.r-project.org/web/packages/vowels/index.html (last accessed 14 February 2025).

Kisler, Thomas, Uwe Reichel & Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45. 326–347.

Labov, William, Sharon Ash & Charles Boberg. 2006. *The atlas of North American English*. Berlin: Mouton de Gruyter.

Lee, Akinobu, Tatsuya Kawahara & Kiyohiro Shikano. 2001. Julius—an open source real-time large vocabulary recognition engine. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, 1691–1694.

Lee, Akinobu & Tatsuya Kawahara. 2009. Recent development of open-source speech recognition engine Julius. In *Proceedings of APSIPA ASC 2009*, 131–137.

Liberman, Mark Y. 2019. Corpus phonetics. *Annual Review of Linguistics* 5. 91–107. https://doi.org/10.1146/annurev-linguistics-011516-033830.

MacKenzie, Laurel & Danielle Turton. 2020. Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard* 6. https://doi.org/10.1515/lingvan-2018-0061.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner & Morgan Sonderegger. 2017a. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of the 18th Conference of the International Speech Communication Association*.

McAuliffe, Michael, Elias Stengel-Eskin, Michaela Socolof & Morgan Sonderegger. 2017b. Polyglot and speech corpus tools: A system for representing, integrating, and querying speech corpora. In *Proceedings of Interspeech 2017*, 3887–3891.

McAuliffe, Michael, Arlie Coles, Michael Goodale, Sarah Mihuc, Michael Wagner, Jane Stuart-Smith & Morgan Sonderegger. 2019. ISCAN: A system for integrated phonetic analyses across speech corpora. In *Proceedings of Interspeech 2019*, 1322–1326.

Meer, Philipp. 2020. Automatic alignment for New Englishes: Applying state-of-the-art aligners to Trinidadian English. *The Journal of the Acoustical Society of America* 147 (4). 2283–2294. https://doi.org/10.1121/10.0001069.

Meer, Philipp, Thorsten Brato & José A. Matute Flores. 2021. Extending automatic vowel formant extraction to New Englishes: A comparison of different methods. *English World-Wide* 42 (1). 54–84. https://doi.org/10.1075/eww.00060.mee.

Méli, Adrien & Nicolas Ballier. 2023. PEASYV: A procedure to obtain phonetic data from subtitled videos. In *Proceedings of the International Congress of Phonetic Sciences 2023*, 3211–3215.

Méli, Adrien, Steven Coats & Nicolas Ballier. 2023. Methods for phonetic scraping of Youtube videos. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, 244–249.

Mielke, Jeff, Erik R. Thomas, Josef Fruehwald, Michael McAuliffe, Morgan Sonderegger, Jane Stuart-Smith & Robin Dodsworth. 2019. Age vectors vs. axes of intraspeaker variation in vowel formants measured automatically from several English speech corpora. In *Proceedings of the International Congress of Phonetic Sciences 2019*, 1258–1262.

Ord, J. Keith & Arthur Getis. 1995. Local spatial autocorrelation statistics: Distributional issues and application. *Geographical Analysis* 27 (4). 286–306.

Pettarin, Alberto. 2022. *Forced-alignment-tools*. https://github.com/pettarin/forced-alignment-tools (last accessed 14 February 2025).

Pimentel, João Felipe, Leonardo Murta, Vanessa Braganholo & Juliana Freire. 2021. Understanding and improving the quality and reproducibility of Jupyter notebooks. *Empirical Software Engineering* 26. https://doi.org/10.1007/s10664-021-09961-9.

Plaquet, Alexis & Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proceedings of Interspeech 2023*, 3222–3226. https://doi.org/10.21437/Interspeech.2023-205.

Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer & Karel Vesely. 2011. The Kaldi speech recognition toolkit. In *Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey & Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv*:2212.04356 [eess.AS]. https://doi.org/10.48550/arXiv.2212.04356.

Reddy, Sravana & James Stanford. 2015. A web application for automated dialect analysis. In *Proceedings of NAACL-HLT 2015.*

Renwick, Margaret E. L. & Joseph A. Stanley. 2020. Modeling dynamic trajectories of front vowels in the American South. *The Journal of the Acoustical Society of America* 147 (1). 579–595. https://doi.org/10.1121/10.0000549.

Rosenfelder, Ingrid Josef Fruehwald, Keelan Evanini, Scott Seyfarth, Kyle Gorman, Hilary Prichard & Jiahong Yuan. 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.

Schmidt, Thomas & Kai Wörner. 2014. EXMARaLDA. In Jacques Durand, Ulrike Gut & Gjert Kristoffersen (eds.), *The Oxford handbook of corpus phonology*, 402–419. Oxford: Oxford University Press.

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

Sóskuthy, Márton, Jennifer Hay & James Brand. 2019. Horizontal diphthong shift in New Zealand English. In *Proceedings of the 19th International Congress of Phonetic Sciences*, 597–601.

Vinciarelli, Alessandro, Maja Pantic & Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and Vision Computing* 27 (12). 1743–1759.

Waskom, Michael L. 2021. Seaborn: Statistical data visualization. *Journal of Open Source Software* 6, 60, 3021. https://doi.org/10.21105/joss.03021.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann & Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, 1556–1559.

Xu, Eric Zhongcong, Zeyang Song, Satoshi Tsutsui, Chao Feng, Mang Ye & Mike Zheng Shou. 2022. AVA-AVD: Audio-visual speaker diarization in the wild. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*, 3838–3847. New York: Association for Computing Machinery. https://doi.org/10.1145/3503161.3548027.

Young, Steve J. 1993. *The HTK hidden Markov model toolkit: Design and philosophy*. Cambridge: Cambridge University.

Yuan, Jiahong & Mark Liberman. 2008. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America* 123 (5). 3878.