Link to GitHub

Introducing LingTube: An open-source toolkit for linguistic analysis of YouTube data

Lauretta S. P. Cheng[1] & Mathew A. Kramer[1]
[1]Department of Linguistics, University of Michigan, Ann Arbor, MI, USA

Link to Poster

Key Points

- **LingTube** is a suite of tools for automating the downloading and processing of captioned YouTube audio for textual and/or phonetic analysis.
- We present ongoing exploratory work using LingTube to investigate phonetic features in the English of Asian North American (ANA) speakers.

YouTube Background

Vast, yet relatively untapped source of publicly-available linguistic data (Schneider, 2016)

- **Advantages:**
  - 'Naturalistic', (auto-)captioned speech data representing various contexts
  - Large user base → improved access to lesser-studied language varieties/communities
- **Disadvantages:**
  - Not straightforward to download and process raw data
  - Manual work needed to screen usable videos/speech (BG music, noise, audio quality, etc.)
  - Minimal, inferred or unavailable speaker background information

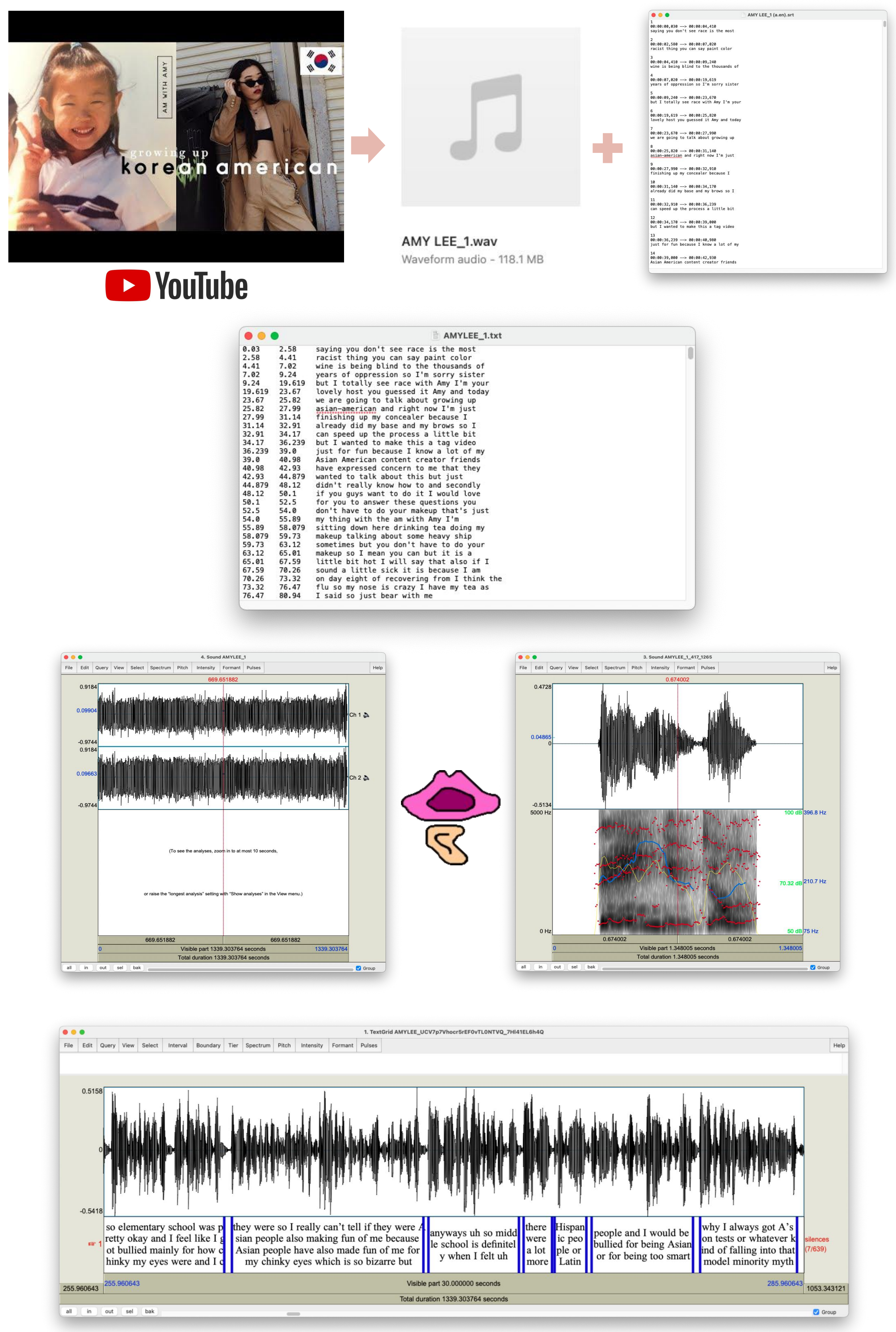
Some previous uses of YouTube:

- *Lee, 2017*: Style-shifting of one speaker across contexts
- *Coats, 2020*: Articulation rate across US regions
- *Kramer, 2021*: Cross-linguistic patterns of dependency length minimization

LingTube Pipeline: https://github.com/Narquelion/LingTube

Base: Download & pre-process captions	
1. Scrape channels	Scrape video URLs and about page information from channel(s)
2. Scrape videos	Download video audio and captions
3. Clean captions	Convert SRT files to tidy TXT files with columns; automatic cleaning of common issues
4. Correct captions	Open caption text file, YouTube video, and GUI for user to manually correct captions

YouSpeak: Process audio and transcript for forced alignment	
1. Convert audio	Converts MP4 files to mono WAV
2. Chunk audio	Create textgrid with utterances separated by breath breaks and saves audio chunks
3. Validate chunks	Allow user to listen to audio chunks, validate transcript, and mark as usable/not
4. Create textgrids	Add validated text to textgrid; optionally copy files to a folder for forced alignment



YouTube ANA English Analysis

The corpus

- 66 YouTubers: 46 ANA-identified, 3 multiracial ANA, 17 non-ANA
  - All presented as women and were East or Southeast Asian
- Used video(s) discussing (ANA) identity; otherwise, personal Q&A or conversational speech
- Demographic/language info coded from videos/"About" page

Steps

1. *Identified video URLs via "Growing Up Asian American tag", etc.*
2. **Scraped channel info, video captions, and audio**
3. **Cleaned captions; RAs manually corrected auto-transcription**
4. **Converted audio to mono WAV and chunked into utterances**
5. **RAs coded audio usability and validated transcript alignment**
6. **Created TextGrids with time-aligned transcribed utterances**
7. *Forced-aligned using Montreal Forced Aligner* (McAuliffe et al., 2017)
8. *Extracted vowel duration and formants (25%, midpoint, 75%)*

Analysis

- Assessed phonetic features linked to ANA identity (Bauman, 2016)
  - Back-vowel retraction
  - Monophthongization
  - "Syllable-timed rhythm"
- Performed hierarchical clustering using Ward's method on 5 measures per speaker
  - Two clusters but no pattern based on (non-)ANA identification
- Fig 1 shows that ANA vs. non-ANA speakers are distributed across each measure's range

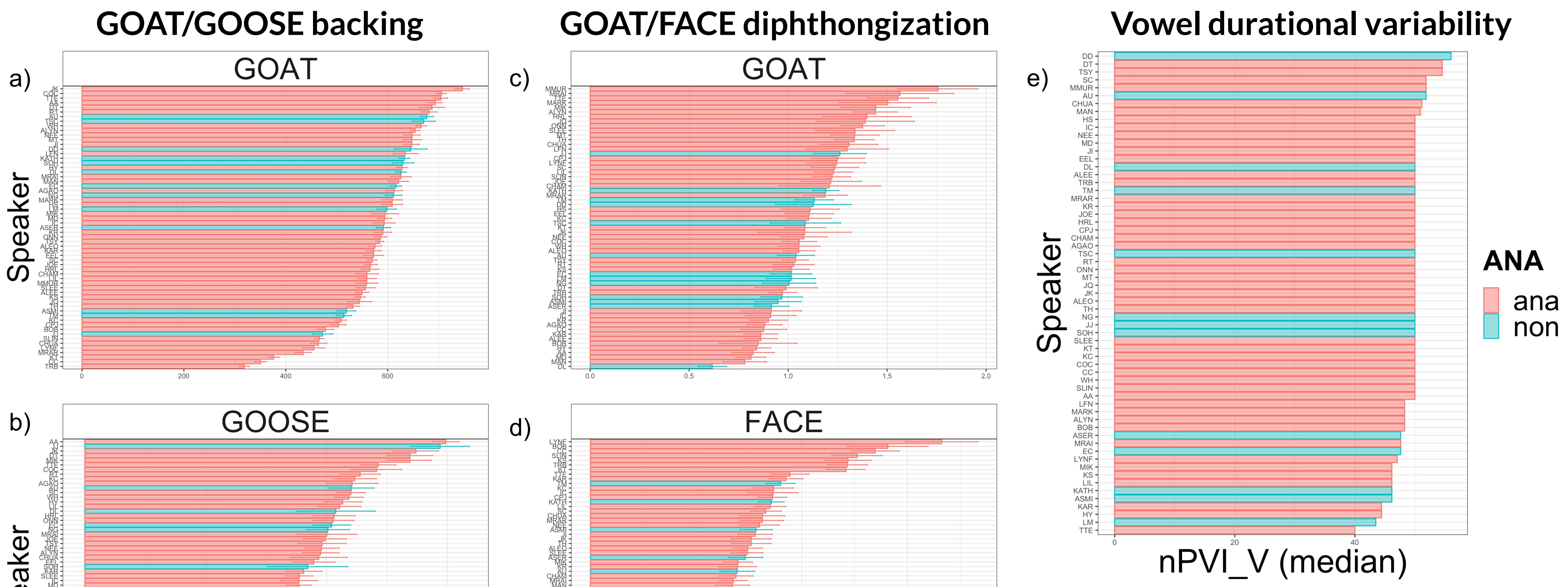


Fig 1. Values per measure plotted by speaker, sorted from highest to lowest. Color indicates ANA (red) and non-ANA (blue) identity.

Measures: Midpoint F2 difference from /i/ for (a) GOAT and (b) GOOSE. Norm. F1-F2 Euclidean distance between 25% and 75% for (c) GOAT and (d) FACE. (e) Normalized Pairwise Variability Index (Grabe & Low, 2002)

Limitations & Future Steps

- LingTube **only thoroughly tested on MacOS and English speech**, but easily extendible
- Still **requires somewhat time-intensive semi-manual work**, but **planned upgrades** will reduce this
  1. Music/noise detection and removal
  2. Word-level auto-alignment (prior to phoneme-level forced-alignment)
- YouTube data **may not be suitable for some questions**
  - E.g. Hard to get ethnic and/or regional identity info
- Both **auto-captioning and forced alignment may struggle with non-standardized American English varieties** introducing bias into data
  - Current automatic vowel data needs correction
- **Next steps**:
  - Continued analysis with hand-corrected vowels, additional measures, consideration of more speaker information (e.g. region, age, etc.)
  - Possible extension to morphosyntactic analysis

References Bauman, C. (2016). Speaking of Sisterhood: A Sociolinguistic Study of an Asian American Sorority. Coats, S. (2020). Articulation Rate in American English in a Corpus of YouTube Videos. Grabe, E., & Low, E. L. (2002). Durational variability in speech and the Rhythm Class Hypothesis. Kramer, A. (2021). Dependency Lengths in Speech and Writing: A Cross-Linguistic Comparison via YouDePP, a Pipeline for Scraping and Parsing YouTube Captions. Lee, S. (2017). Style-Shifting in Vlogging: An Acoustic Analysis of "YouTube Voice." McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. Schneider, E. W. (2016). World Englishes on YouTube: Treasure trove or nightmare?