Is the Future Almost Here?
Large-Scale Completely Automated Vowel Extraction of Free Speech Recordings

Sravana K. Reddy (Dartmouth College) and James N. Stanford (Dartmouth College)

Automatic Speech Recognition (ASR) is reaching farther into everyday life through applications like Apple's Siri. Likewise, sociolinguists have been considering new technologies for vowel formant extraction, e.g., semi-automated alignment/extraction techniques like Penn Aligner (Yuan & Liberman 2008; Evanini et al. 2009) and Forced Alignment Vowel Extraction (FAVE) (Rosenfelder et al. 2011). With humans transcribing recordings into sentences, these semi-automated methods produce effective results (Labov et al. 2013; Evanini et al. 2009). But sociolinguistics may be on the brink of another transformational technology: large-scale, completely automated vowel extraction without any need for human transcription. It would then be possible to quickly extract vowels from virtually limitless hours of recordings, such as YouTube, publicly available audio/video archives, and even live-streaming video. How far away is this transformational moment? In the present study, we apply state-of-the-art ASR to a real-world sociolinguistic dataset as a feasibility test. Our results show that meaningful sociolinguistic results are possible, although a number of ASR challenges remain.

Methods: Unlike other ASR applications where accurate word-recognition is the primary goal, sociophonetic vowel research typically focuses on a narrower objective: extracting a representative vowel-space for each speaker. For this reason, we believe that *completely automated vowel extraction* (CAVE) is becoming feasible for sociolinguistic research.

We trained a Hidden Markov Model-based speech recognizer (Jelinek et al. 1975) on publicly available corpora using the CMU Sphinx toolkit. This recognizer represents the state-of-the-art research standard in ASR. We then examined the U.S. Southern Vowel Shift (SVS) in the Switchboard corpus of phone conversations (Godfrey & Holliman 1993), randomly selecting 10 Southerners (5 women/5 men) and 10 Northerners (5 women/5 men). We used CAVE to automatically transcribe these recordings and extract F1/F2 from 143,266 stressed vowel tokens (15+ hours of conversation, averaging 7,163 tokens/speaker), normalizing with Lobanov (Kendall & Thomas 2010). As a control, we used Switchboard's manual transcriptions and extracted the formants with FAVE.

Results: Comparisons of individual speakers in FAVE and CAVE shows inconsistency at the level of individual vowel tokens, thus highlighting the need for improved ASR in the future. Even so, both methods produced comparable sociolinguistic analyses of Southern features, suggesting that CAVE can already provide usable results for certain research questions.

Our analysis shows that both CAVE and FAVE revealed clear north/south contrasts in the tense/lax shifts of BAIT/BET (EY/EH) and BEAT/BIT (IY/IH). For both methods, these shifts appear in the expected SVS directions (EY and IY are lowered/backed, EH and IH are raised/fronted). Both CAVE and FAVE also show Southern fronting of AW, UW, and OW. The north/south contrast in the EY/EH shift was significant for both methods ($p<0.002$, Repeated-Measures ANOVA, using Euclidean distances between EY/EH), but the contrast in IY/IH was only significant in FAVE ($p=0.011$), not CAVE ($p=0.284$). In prior SVS work, EY/EH is typically more advanced than IY/IH (Kendall & Fridland 2012), and both methods showed this effect as well.

As ASR improves, completely automated methods are likely to become reliable enough for fast, accurate analyses of vast amounts of data.

[500 words]

Selected References

Evanini, Keelan, Stephen Isard, and Mark Liberman (2009). Automatic formant extraction for sociolinguistic analysis of large corpora. *Proceedings of Interspeech*.

Godfrey, John and Edward Holliman (1993). Switchboard-1 Release 2 LDC97S62 [Corpus]. Philadelphia: Linguistic Data Consortium.

Jelinek, Fred, Lalit Bahl and Robert Mercer (1975). Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory* IT 21:250–256.

Kendall, Tyler and Valerie Fridland (2012). Variation in perception and production of mid front vowels in the U.S. Southern Vowel Shift. *Journal of Phonetics* 40:289-306.

Kendall, Tyler and Erik R. Thomas (2010). Vowels: vowel manipulation, normalization, and plotting in R. R package, version 1.1.
Available from http://cran.r-project.org/we/packages/vowels/

Labov, William, Sharon Ash and Charles Boberg (2006). *The Atlas of North American English* (ANAE). Berlin: Mouton.

Labov,William, Ingrid Rosenfelder and Josef Fruehwald (2013). One hundred years of sound change in Philadelphia: Linear incrementation, reversal and reanalysis. *Language* 89(1):30–65.

Rosenfelder, Ingrid, Josef Fruehwald, Keelan Evanini and Jiahong Yuan (2011). FAVE (Forced Alignment and Vowel Extraction) Program Suite [Computer Program].

Yuan, Jiahong and Mark Liberman (2008). Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics '08*.