Cross-Study Reliability of Neural Network Auto-Coders for Rhoticity

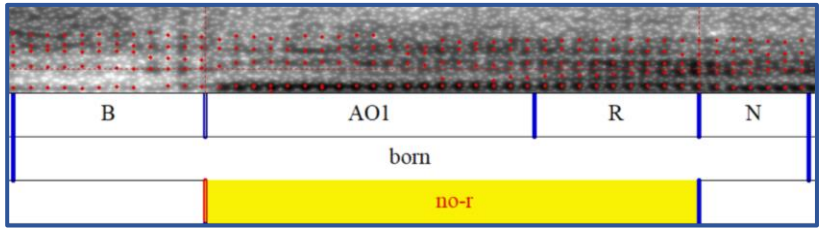
Brandon Prickett (bprickett@umass.edu),[1] Sarah Gupta (sarahguptaa@gmail.com),[2] Monica Nesbitt (nesbittm@indiana.edu),[3]
Joe Pater (pater@umass.edu),[1] and James N. Stanford (James.N.Stanford@Dartmouth.edu)[4]

1. Introduction

- Auto-coders help with the time-consuming and inconsistent nature of manual transcription.
- Many sociophonetic features can be auto-coded with a high degree of reliability and replicability (Labov et al. 2013; Sonderegger 2015; Villarreal et al. 2020).
  - A reliable auto-coded approach to (r) is lacking.
  - Though there has been progress (McLarty et al. 2019; Gupta and DiPadova 2019).
- In this study, we used a neural-network auto-coder with Mel Frequency Cepstral Coefficients (MFCCs) as input, following prior work by Gupta and DiPadova (2019).
  - We then examine whether, when trained on white New England speakers, the auto-coder will generalize to Black speakers from the same region.

2. Methods

- Training data (Gupta and DiPadova 2019) included word lists, sentences, and reading passages read aloud by 208 white New England speakers (WhNE).
  - V+/r/ sequences were extracted from recordings and coded for (r) by 2 judges.
  - Each datum was divided into 100 timepoints.
  - For each timepoint, 12 MFCCs were measured.



- We also tested the auto-coder on a similarly transcribed New England dataset (Nesbitt & Watts 2022) that includes 59 African American and Caribbean American speakers (AA&CA) from Boston.
- All tests used data withheld from training and metrics borrowed from the machine learning literature:
  - **Precision** = $\frac{True\ Positives}{True\ \&\ False\ Positives}$ **Recall** = $\frac{True\ Positives}{True\ Positivs\ \&\ False\ Negatives}$
  - **F1 scores** (harmonic mean of precision and recall) and **AUC** (area under the curve of a true positive vs. false positive plot at various cutoff values).

3. Results

- The results from our auto-coder, compared with past results:

	Accuracy	Precision	Recall	F1	AUC
Past work (WhNE)	.811	.829	.830	.829	.892
Our results (WhNE)	.796	.811	.841	.826	.833
Our results (AA&CA)	.793	.878	.889	.884	.473

- A logistic regression (speaker=random) found no significant effects of speakers' age/ethnicity/gender or sub/urban status on r-fulness in AA&CA.
- We ran a model with the same predictors on the auto-coder's output and found a significant effect of age ($z=5.286$, $p<.001$) on its accuracy (older=better).
- When trained on both WhNE and AA&CA, the model did even better on withheld data (acc=.91, prec=.97, recall=.92, F1=.94, AUC=.97).

4. Discussion

- We replicated Gupta and DiPadova's (2019) results showing that a neural network can auto-code rhoticity with relatively high accuracy and showed this approach may generalize well across diverse studies.
- Future work should explore other novel datasets with this approach and see whether other factors might affect performance (we've been pursuing this and found that generalizing to some datasets *does* challenge the auto-coder).

Select References

Gupta & DiPadova (2019). Deep learning and sociophonetics: Automatic coding of rhoticity using neural networks. • Nesbitt & Watts (2022). Socially distanced but virtually connected: pandemic fieldwork with Black Bostonians. • Villarreal, Hay, & Watson (2020). From categories to gradience: Auto-coding sociophonetic variation with random forests.

Author affiliations: *(1) University of Massachusetts Amherst, (2) AWS AI Labs (work unrelated to the author's position at Amazon), (3) Indiana University, (4) Dartmouth College.*