5aSC27. Acoustic cues to linguistic profiling? Machine learning of phonetic features of African American English

Meisam K. Arjmandi, Laura Dilley, Suzanne Wagner

Michigan State University, Department of Communicative Sciences and Disorders, East Lansing, MI

BACKGROUND

- Listeners can rapidly draw inferences about the likely background of a speaker – including their dialect and racial background – within milliseconds of hearing their voice (Munson, 2007; Lattner & Friederici, 2003; Scharinger et al., 2011).
- The accuracy of perceptual recognition of dialect is better than chance (Purnell et al., 1999).
- African American English (AAE) is a dialect spoken by many of the approximately 45 million African Americans.
- Acoustic cues carried by the word "*hello*" over a phone call were enough to identify speaker's racial background (Purnell et al., 1999; Scharinger et al., 2011)
  - This dialect identification has led to subsequent discrimination in housing (Purnell et al., 1999).
- It is still not clear which acoustic cues and phonetic contexts facilitate this rapid inference about dialect and racial background.
- Speech is the outcome of a dynamic interaction between vocal folds vibratory patterns and patterns of articulatory states and movement in the vocal tract.
- Dialect modulates phonatory and articulatory patterns during speech, leading to distinct cross-dialectal acoustic representations (Fox & Jacewicz, 2009).
- Formant dynamic information is informative for separation of AAE from Standard American English (SAE) dialect (Arjmandi et al., 2017), but the degree of contribution of other acoustic dimensions has not yet been investigated.

RESEARCH QUESTIONS:

- **What are the acoustic dimensions relevant to the glottal source and/or the vocal tract which are most informative for AAE versus SAE dialect separation?**
- **How does the degree of informativity of these acoustic cues for dialect differentiation vary across different phonological contexts?**

METHODS

MATERIALS:

- Six female speakers, all from Lansing, Michigan, participated in an sociolinguistic interview.
  - 3 AAE speakers and 3 SAE speakers
- Tokens of vowels conditioned on certain phonological contexts were identified.
- Closed syllables with a sonorant coda (/l/, /r/, /n/, or /m/) or non-sonorant coda, from specific lexical items, to control for coarticulation (Table 1).
  - Target stretches of speech consisted of vowel (V) or vowel-consonant (VC) sequences (Total analyzed speech = 183.3 secs (100.1 sec AAE & 83.4 sec SAE)
- Sonorant sounds (*e.g.*, V and VC) carry substantial acoustic cues relevant to dialect identification (Jacewicz & Fox, 2013).

ACOUSTIC MEASURES:

- Four general categories of acoustic features were calculated to characterize acoustic variations in multiple dimensions with respect to their informativeness in AAE vs. SAE dialect separation.
  1) **Speech-based Features (Glottal Source + Vocal Tract):** Measures that reflect the behavior of both glottal source and vocal tract.
     - **H1-H2, H1-A1, H1-A1, H1-A2, A1-A3, H1-A3:** These measures were calculated by amplitudes of the 1st and 2nd harmonics (cf. H1, H2) relative to each other and to the amplitude of the 1st, 2nd, and 3rd formants (cf. A1, A2, A3)
     - **Spectral Slope (SS):** Reflects the rate of decline in spectral amplitudes.
  2) **Vocal Tract Features:** Measures that represent the natural resonances of the vocal cavity.
     - **F1, F2, & F3:** The 1st, 2nd, and 3rd formant frequencies
  3) **Voice Quality (VQ) Measures**: Measures that reflect the quality of voice during V or VC pronunciation.
     - **Jitter & Shimmer:** The average absolute difference between consecutive periods (jitter) and amplitude (shimmer), normalized by average period and average amplitude.

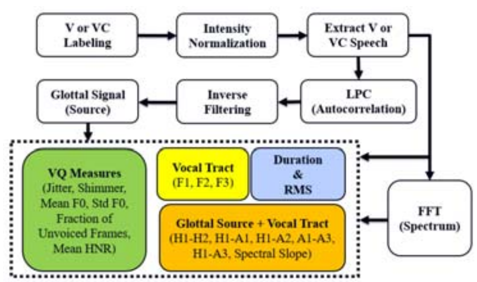
ACOUSTIC MEASURES:

3) **Voice Quality (VQ) Measures (cont.):**
   - **Mean F_0 & STD F_0:** Mean and standard deviation of F_0
   - **Fraction of Unvoiced Frames (FOUF):** Fraction of locally unvoiced frames
   - **Mean Harmonic-to-Noise Ratio (HNR):** It reflects the degree of periodicity.
4) **Duration & RMS:** Measures to characterize energy and linguistic stress. Duration is used as a physical correlate of linguistic stress (Fry, 1955), and RMS characterizes the amount of energy in the voice.

ANALYSES:

1) **Feature Evaluation:** The informativeness of these acoustic features were individually evaluated to identify their informativeness across these four categories in separation of AAE versus SAE dialect contrast.
   - Principal component analysis (PCA) is conducted to understand the most optimum new dimension which explains major sources of variability in the data.
   - The *Mahalanobis distance* (Theodoridis & Koutroumbas, 2011) was used as a non-probabilistic measure to rank the acoustic features. It evaluates the distance of a feature in a multi-dimensional space from the mean of the class.
2) **Machine Learning of Sonorant Speech**: A support vector machine (SVM) was trained on the feature space to identify how much the acoustic dimensions formed by these features are informative.

Figure 1. Schematic of the method implemented to extract the acoustic features of the four categories from V and VC speech sounds. These acoustic measures were calculated from the original signal or the glottal source signal or FFT spectrum.



RESULTS

- The average and standard deviation of the fraction of variance explained by principal component analysis (PCA) across 17 phonological contexts suggest that only 3 principal components (PCs) are enough to represent the variability in acoustic feature space.

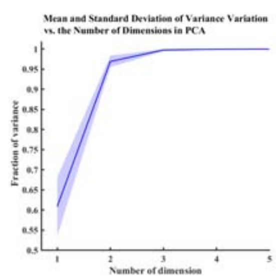
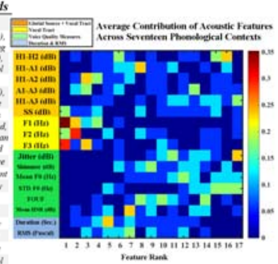


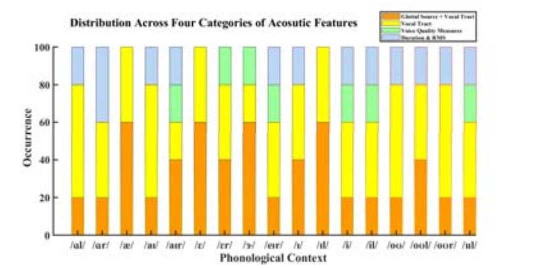
Table 1. The distribution of acoustic features, which are individually ranked by *Mahalanobis distance* measure, is shown for each phonological context. The first five acoustic features with more information in dialect separation is listed.

Context	*1*	*2*	*3*	*4*	*5*	*Words*
/ɑl/	Spectral Slope	F2	F3	F1	RMS	*all*
/ɑr/	F2	Spectral Slope	Duration	F1	RMS	*Bark(s), Barking*
/æ/	F1	F2	Spectral Slope	H1-A3	H1-A1	*cap(s), capital*
/aɪ/	F3	H1-A2	RMS	F1	F2	*My*
/aɪr/	Spectral Slope	F3	RMS	H1-A2	STD(F0)	*Fire(d), entire*
/ɛ/	F3	F1	H1-A2	Spectral Slope	H1-H2	*guess*
/ɛr/	F3	HNR	F1	Spectral Slope	H1-A3	*Married, American*
/ɝ/	Spectral Slope	H1-H2	F1	H1-A2	HNR	*Heard*
/eɪr/	F3	F1	A1-A3	RMS	STD	*They're*
/ɪ/	F1	A1-A3	F3	H1-H2	Duration	*Different*
/ɪl/	F2	A1-A3	F3	H1-A1	H1-A2	*Really*
/i/	F1	F2	A1-A3	HNR	Duration	*Me*
/il/	F2	F1	A1-A3	Duration	HNR	*Feel*
/oʊ/	F1	F2	H1-A2	F3	Duration	*Know*
/oʊl/	F1	F2	H1-A2	Duration	A1-A3	*Told*
/oʊr/	F1	Duration	F3	H1-A2	F2	*More*
/ul/	F3	F1	H1-A2	Unvoiced Frames	Duration	*School*

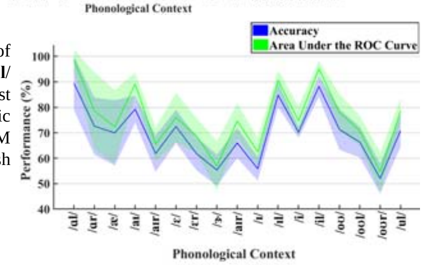


RESULTS

- The results from ranking the acoustic features based on their informativeness in AAE-vs-SAE dialect separation suggest that the main contributions come from **speech-based features** and **vocal tract** features.



- Sonorant contexts of **/ɑl/**, **/ɪl/**, and **/il/** provide the most informative acoustic cues for the SVM classifier to distinguish AAE from SAE.



CONCLUSIONS

- The results from this study suggest that rapid recognition of AAE dialect from SAE dialect is facilitated through interaction of acoustic features representing both phonatory behaviors and articulatory gestures.
- Formants in V and VC provide substantial acoustic cues for recognition of AAE from SAE.
- Investigating the acoustic cues from continuous speech, including obstruents, rather than merely sonorant regions, can be planned for future studies.
- These findings also suggest that auditory perceptual categorization of AAE from SAE occurs through the interaction of multiple acoustic cues in a multidimensional acoustic space. Listeners dynamically adjust their cue weighting mechanisms with respect to these dimensions to retrieve dialect-related information.

References

[1] Lattner, S., & Friederici, A. D. (2003). Talker's voice and gender stereotype in human auditory sentence processing–evidence from event-related brain potentials. *Neuroscience Letters*, 339(3), 191-194.

[2] Scharinger, M., Monahan, P. J., & Idsardi, W. J. (2011). You had me at "Hello": Rapid extraction of dialect information from spoken words. *Neuroimage*, 56(4), 2329-2338.

[3] Munson, B. (2007). The acoustic correlates of perceived masculinity, perceived femininity, and perceived sexual orientation. *Language and Speech*, 50(1), 125-142.

[4] Purnell, T., Idsardi, W., & Baugh, J. (1999). Perceptual and phonetic experiments on American English dialect identification. *Journal of Language and Social Psychology*, 18(1), 10-30.

[6] Fox, R. A., & Jacewicz, E. (2009). Cross-dialectal variation in formant dynamics of American English vowels. *The Journal of the Acoustical Society of America*, *126*(5), 2603-2618.

[7] Arjmandi, M. K., Dilley, L., & Ireland, Z. (2017). Applying pattern recognition to formant trajectories: A useful tool for understanding African American English dialect variation. *The Journal of the Acoustical Society of America*, *141*(5), 3980-3980.

[9] Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *The Journal of the Acoustical Society of America*, *27*(4), 765-768.

[11] Theodoridis, S., & Koutroumbas, K. (2001). Pattern recognition and neural networks. In *Machine Learning and Its Applications* (pp. 169-195). Springer Berlin Heidelberg.

Acknowledgements

We gratefully acknowledge the support provided by the Diversity Research Network (DRN), the Charles Strosacker Foundation, and the Stockman Fund provided by Drs. George and Ida Stockman for backing of this study.