Accent Classification for AAE Dialects Using Audio Feature Extraction

Amy DeMorrow, UCLA MSOL Student

Abstract

This study used an XGBoost regression model to perform accent classification on African American English (AEE) audio recordings collected in the Corpus of Regional African American Language (CORAAL). Audio features selected for model training included 20 mel frequency cepstrum coefficients, the mel spectrogram, spectral rolloffs, the complete ComParE 2016 feature set, power normalized cepstral coefficients (PNCC), and perceptual linear prediction (PLP) features.

The CORAAL dataset contained speakers from five distinct United States cities: Rochester, NY, New York, NY (lower east side of Manhattan), Washington, DC, Princeville, NC, and Valdosta, GA. Although each city contains hundreds of audio files, the number of unique speakers is extremely small. But as a study in feature performance on "clean" and "noisy" audio, this corpus is large enough to show trends in model performance between the two test sets.

Two models were trained for this study. The first model was trained on the raw audio files. The second model was trained on an augmented training set with additional time shifted, speed altered, and pitch altered audio files to help correct extreme skew in the dataset. Accuracies, feature performance, and test results in both clean and noisy data will be reported for both models. Additionally, inference testing was performed on a blind set of clean and noisy data to test generalization of the augmented model. While overall accuracy was robust on clean augmented data, noisy data and inference data performed poorly. This model proves that feature selection alone is not sufficient to create a robust AAE accent classification model.

Index Terms: Dialect Identification, African American English, limited training data, audio feature selection for noise.

1. Introduction

Dialects are variations in a language that differ by geographic region or social group and can be distinguished by a listener in terms of grammar, vocabulary, and phonology. Variations in dialect (accent) is a well-documented weakness in modeling speech for Automated Speech Recognition (ASR), with accent second only to gender in ASR performance [1]. Previous work has shown that ASR models trained in Generalized American English (GAE) demonstrate performance degradation when exposed to AAE language [2]. If the speaker can be reliably classified by dialect using a short audio clip at the beginning of an utterance, that speaker can be automatically sorted into a finely attenuated model trained specifically on that dialect. The ability to reliably classify dialect using an acoustic model may help a downstream ASR language model recognize and interpret non-standard speech patterns, which are common in AAE English and differ from Generalized American English (GAE) [3].

Recent work in this ASR category reveals continuing challenges in accent classification across a broad range of accents. Accent specific performance depends heavily on the availability of audio recordings of a specific accent, so accents with limited data suffer from error prone ASR recognition. Another challenge is the high costs associated with developing and training multiple accent specific models [4]. AAE (or AAL) is a low resource dialect currently experiencing poor performance with ASR recognition.

African American Language (AAL) is defined by the ORAAL researchers as encompassing "all varieties of language use in African American communities" [5] reflecting 'differences in age/generation, sex, gender, sexuality, social and socioeconomic class, region, education, religion, and other affiliations and identities that intersect with one's ethnicity/race and nationality" [6]. The language developed over centuries under the chattel slavery system, and a recent hypothesis called the "Substrate Hypothesis" suggests that AAL originally developed out of contact between African, Creole, and English speakers. The language continued to develop due to the Great Migration of 1910-1970, where large groups of African Americans left the southern regions of the United States and concentrated in the urban cores of large American cities like Chicago, New York, Washington, DC, Los Angeles, St. Louis, and Detroit. Once settled in these urban cores, continued segregation from adjacent communities and a strong cultural identity drove further changes to the language in subsequent generations of speakers [7].

Considering the recent common root and amount of continuing isolation imposed on this community, expecting the same regional variations seen in GAE accents in so few generations since the Great Migration for accent classification is a leap of faith. This task is further complicated by the differences in age, sex, socioeconomic class, code switching (if the researcher recording happened to be white), more recent comingling with the regional GAE accents, and the age of the recordings themselves (one older speaker in the CORAAL dataset recorded in Manhattan still sounds "southern"). The limited size of the available AAE corpuses also hinder the effectiveness of the models. But attempting accent classification on small AAE datasets could drive interest in the AAE language and push more academic resources to the AAE subset of ASR research.

2. Project Description

For the task of acoustic accent classification on AAE, the CORAAL [8] corpus was selected for audio files, and an

XGBoost Regression Tree Model was used to assess individual feature impacts on the overall classification model. Noise files for accuracy testing were created by corrupting speakers in the CORAAL corpus using a 10dB of babble noise masker.

2.1. The CORAAL Dataset

For this study, the CORAAL data set was curated into sets of five American cities representing different regional accents in AAE:

1. Rochester, NY (ROC)
2. Lower East Side Manhattan, New York, NY (LES)
3. Washington, DC (DCB)
4. Princeville, NC (PRV)
5. Valdosta, GA (VLD)

Selected audio recordings all had lengths of greater than 10 seconds. The training dataset was highly skewed by city as well as gender, and the set contains 28 unique speakers, a low resource dataset (see Table 1).

Table 1: Baseline *CORAAL Training Set Overview*

City	% Total Audio Files	% Female	% Male	Num Females	Num Males
DCB	56%	64%	36%	5	3
ROC	15%	70%	30%	4	3
VLD	13%	55%	45%	2	3
LES	10%	58%	42%	2	2
PRV	6%	70%	30%	2	2
Total		**63%**	**37%**	**15**	**13**

With 56% of the total audio concentrated in Washington DC and 63% attributed to female speakers, a second training data set was created to help correct skew in the training data.

For the purposes of this study, four possible data augmentations were selected and assessed for possible improvements on model accuracy:

- Time Shift (1 second)
- Speed Alteration (95% of original speed)
- Pitch Alteration (Female to Male)
- Added Noise (white and pink noise)

Two attempts to add noise to the training files were made. In the first trial, half of all baseline files were augmented with white noise. In the second trial, only VLD and LES were augmented with pink noise. Neither attempt could boost noise performance without significantly degrading overall model performance (see Table 4); therefore, noise augmentation on the training set was abandoned for further study.

The ROC, VLD, LES, and PRV datasets were augmented with time shifting, speed reduction, pitch alteration as specified in the above bullet list. All files in the four underrepresented cities were augmented for time shift and speed. For altering pitch, half the female files in those four cities had their pitch lowered into the male range to boost male speaker representation. After these augmented audio files were added, the training dataset exhibited less skew (see Table 2).

Table 2: Augmented *CORAAL Training Set Overview*

City	Base Files	Time Shift Files	Speed Files	Pitch Files	New Total Files	New % of Total
DCB	2457	0	0	0	2457	28%
ROC	647	647	647	225	2166	25%
VLD	567	567	567	155	1856	21%
LES	459	459	459	132	1510	17%
PRV	242	242	242	85	812	9%
Total	**4372**	**1915**	**1915**	**597**	**8801**	

PRV and LES remain underrepresented in the augmented dataset, but the pitch alteration was a "quick and dirty" attempt to perform a complex task, and the decision was made to avoid adding additional pitch files to further balance skew in LES and VLD.

2.2. The XGBoost Model

The XGBoost (Extreme Gradient boost) is a machine learning algorithm that uses gradient boosting combined with a decision tree algorthim. This supervised algorithm attempts to accurately predict a target variable by combining the estimates of a set of simpler and weaker models. The gradient boosting operates by using regression to minimize a regularized (L1 and L2) objective function combining a convex loss and a penalty term for model complexity. The model training proceeds iteratively, adding new decision trees that predict the residuals or errors of prior trees that are then combined and pruned backwards to make a final prediction. The gradient boosting is using a gradient descent algorithm to minimize loss as trees are added. The built-in protections against overfitting combined with the lower overall time complexity make XGBoost very popular for supervised classification tasks on large datasets [9]. The other benefit of XGBoost for acoustic feature selection and evaluation is the ability for a tree model to easily show individual feature impacts on overall model performance and optimize feature selection for an acoustic model. Figure 1 shows an illustration of the XGBoost model.

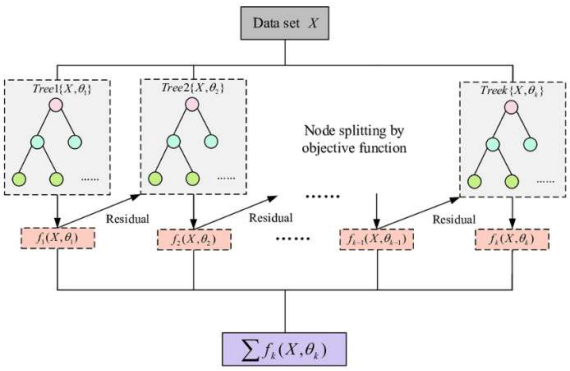

Figure 1: *Diagram of an XGBoost Model [10]*

2.3. Acoustic Features

For the purposes of this model, acoustic features known to be robust to noise and cited as useful in other accent classification studies were chosen to provide inputs to the XGBoost model. Frequency domain features tend to perform best, but some time domain features were retained in the model after they were shown to boost overall model accuracy in clean data. One attempt was made to give the ComParE 2016 feature set a "haircut" by discarding all features below an impact threshold of 0.0001, but since accuracy performance degraded, the full set of ComParE features were retained in the final model. Features shown to provide the highest impacts to model performance were:

- MFCCs (mel frequency cepstrum coefficients)
- Mel Spectrograph
- Rolloffs
- ComParE 2016 feature set
- PLP (perceptual linear prediction)
- PNCC (power normalized cepstral coefficients)

The PLP features were retained in the baseline model, whereas the PNCC features were retained in the augmented data model. PLP might have been more useful than PNCC in the augmented dataset; however, computational resource constraints precluded an assessment of PLP features with that dataset. Relative feature impacts in the models will be shown in the Results section of this study.

3. Related work

Because a successful ASR model relies on an acoustic model as well as a language model for grammar and vocabulary, the rise of neural networks for accent classification have accelerated the work in this space. Early work in this category included a DNN generalized model paired with an accent specific acoustic top layer, which showed significant improvement over a baseline cross entropy model [11]. A later improvement to this model structure was made by adding i-vectors to the model [12].

More recent work involves supplementing models with a standalone accent classifier which feeds into this DNN ASR model. Jain et al. incorporated a standalone classifier that input features as a separate input into the model and found that combining accent embedding input into the multi-task learning model performed better than relying on a single output layer for all accents [13].

Moving beyond a basic DNN model to more advanced models has yielded some accuracy gains using bias adaptation in the first layer of an LSTM paired with large datasets of training audio eliminated the need for complex adaptation techniques [14]. The use of LSTM architecture was further refined by Yoo et al. using Feature-wise Linear Modulation (FiLM) to apply feature wise affine transformations between dialect conditioned LSTM layers [15]. These advanced techniques are well beyond the scope of this study but show that advances in machine learning could one day capture the essence of the AAE language with accuracy to rival GAE.

4. Results

4.1. The Baseline CORAAL Dataset

Early experiments with features available in the librosa library in Python using a full suite of available features for exploration included MFCCs, Mel Spectrogram, Chromas, Zeros, Spectral Centroids, Spectral Rolloffs, and Spectral Bandwidths. Adding MFCCs beyond the first 13 seemed to increase model accuracy, so 20 MFCCs were chosen for the final baseline.

After evaluating feature importance to the model, a "Librosa" baseline feature set were down selected to include MFCC=20, the mel spectrogram, and rolloffs (highlighted in yellow). The rolloffs are not highly ranked, but were retained due to literature studies citing their importance to accent recognition (see Figure 2). The results of these early experiments are shown in Table 3 below.

Table 3: Baseline *CORAAL Training Set: Early Feature Evaluation*

Features	Train Acc	Test Clean Acc	Test Noise Acc
MFCC=13 (baseline)	*0.979*	*0.796*	*0.582*
MFCC=13, Spectrogram, Chromas. Zeros. Centroids, Rolloffs, Bandwidths	0.996	0.808	0.608
MFCC=16, Spectrogram, Rolloffs	1.0	0.817	0.646
MFCC=20, Spectrogram, Rolloffs ("Librosa")	1.0	0.857	0.634

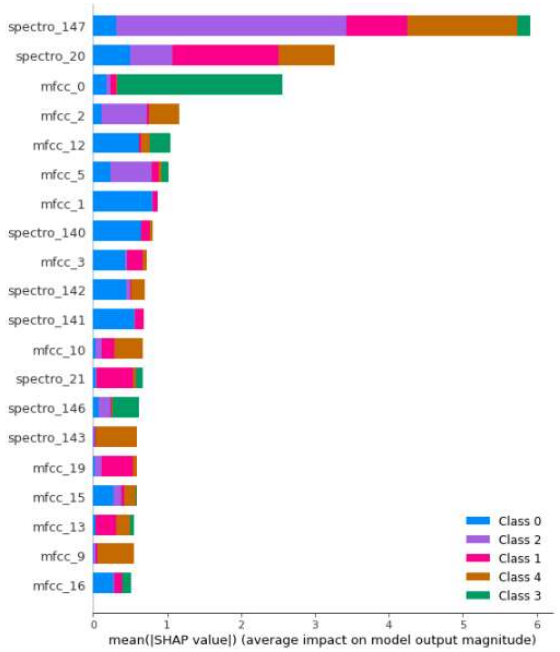

Figure 2: *SHAP diagram ranking feature importance for "Librosa" baseline feature set*

Moving beyond the librosa library, the ComParE 2016 feature set (available from openSMILE) and PNCC and PLP features (available from the spafe library) were extracted and evaluated alongside the "Librosa" feature set. The results on model performance are shown in Table 4.

Table 4: Baseline *CORAAL Training Set: Early Feature Evaluation*

Features	Train Acc	Test Clean Acc	Test Noise Acc
Librosa (baseline)	*1.0*	*0.857*	*0.634*
ComParE	1.0	0.897	0.602
Librosa + ComParE	1.0	0.875	0.671
Librosa + ComParE + PNCC	1.0	0.826	0.674
Librosa + ComParE + PNCC + PLP	1.0	0.837	0.674
Librosa + PNCC + PLP	1.0	0.785	0.654
ComParE + PNCC + PLP	1.0	0.857	0.654
ComParE + PNCC	1.0	0.843	0.651
ComParE + PLP	1.0	0.906	0.605
Librosa + ComParE + PLP	1.0	0.886	0.669

Various tradeoffs between clean and noisy test set performance are observed in these combinations of feature sets. The "Librosa + ComParE + PLP" feature set was chosen as the baseline for the baseline due to the best overall balance between clean and noisy data performance. The feature performance impacts of this feature set are show in Figure 3. This SHAP analysis shows clear impacts due to ComParE features. PLP features do not rank in the Top 20, but overall accuracy gains prove their value in the baseline feature set.

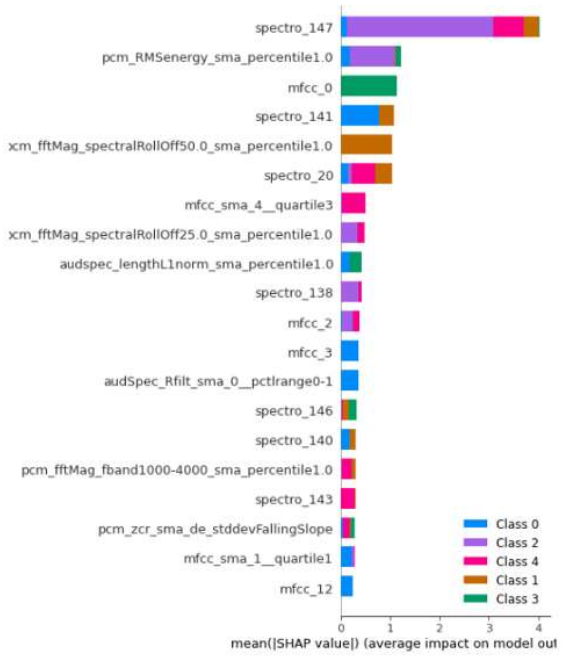

Figure 3: *SHAP diagram ranking feature importance for "Librosa + ComParE + PLP" baseline feature set*

4.2. The Augmented CORAAL Dataset

Due to the observed skew in the baseline dataset, and obvious performance degradations observed in the underrepresented cities (ROC, LES, PRV, VLD), augmentation of the dataset was performed to help balance the audio input data. The techniques and amounts of skew correction were discussed in Section 2.1 and shown in Table 2. The results on experiments with various augmentation techniques are shown in Table 5.

Table 5: Augmentation Experiments on *CORAAL Training Set:*

Features	Train Acc	Test Clean Acc	Test Noise Acc
Librosa (baseline)	*1.0*	*0.857*	*0.634*
Librosa + Shift	1.0	0.881	0.682
Librosa + Shift + Speed	1.0	0.908	0.663
Librosa + Shift + Speed + Pitch	1.0	0.902	0.723
Librosa + Shift + Speed + Pitch + White Noise	1.0	0.832	0.744
Librosa + Shift + Speed + Pitch + Pink Noise (VLD, LES Only)	1.0	0.892	0.735

Adding noise to the files improves model performance in noise but pays a penalty in overall model performance. Due to the degradations in clean accuracy, using noise as an augmentation technique was abandoned. Instead, adding time shift, speed, and pitch augmentations were chosen for the baseline augmentation techniques. All further evaluations with feature performance use shift, speed, and pitch augmented files to evaluate model performance.

Using this augmented dataset, model performance with various feature sets was carried out with the results shown in Table 6. GeMAPS features from openSMILE were also attempted but discarded due to poor performance.

Table 6: Augmented *CORAAL Training Set: Effects of Feature Sets on Model Performance*

Features	Train Acc	Test Clean Acc	Test Noise Acc
Librosa (baseline)	*1.0*	*0.902*	*0.723*
ComParE	1.0	0.957	0.594
GeMAPS	1.0	0.852	0.334
Librosa + ComParE	1.0	0.987	0.671
Librosa + ComParE + GeMAPS	1.0	0.980	0.663
Librosa + ComParE + PNCC	1.0	0.987	0.666
ComParE + PNCC	1.0	0.940	0.643
Librosa + PNCC	1.0	0.872	0.775

Although the overall accuracies seem to indicate choosing "Librosa + ComParE" as the baseline is the best choice, better misclassification patterns seem in the confusion matrices led the author to believe that choosing "Librosa + ComParE + PNCC" as the baseline feature set could help boost performance in an blind inference dataset.

The author would have rather had a "Librosa + ComParE + PLP" feature set for the baseline feature set (as was chosen in Section 4.1), but computational resource constraints precluded that evaluation.

Inference testing was accomplished on the augmented and baseline datasets with the results shown in Table 7.

Table 7: *CORAAL Training Set: Inference Testing With "Unseen" Speakers from the Five Cities on Baseline and Augmented Datasets*

Dataset	Features	Test Clean Acc	Test Noise Acc
Augmented	Librosa + ComParE	0.590	0.234
Augmented	Librosa + ComParE + PNCC	0.599	0.281
Baseline	Librosa + ComParE + PLP	0.873	0.766
Baseline	Librosa + ComParE + PNCC + PLP	0.837	0.760
Baseline	Librosa + ComParE + PNCC	0.837	0.766
Baseline	Librosa + ComParE	0.873	0.764

These results indicate poor performance on inference testing in the augmented dataset, demonstrating that either a highly constrained set of unique speakers is overfitting the model or errors in the augmentation techniques negatively impacted inference. Based on the extremely low accuracy values, augmentation implementation errors are the likely culprit.

As for the baseline dataset, inference testing yields the best overall accuracy on both clean data and noisy data using the "Librosa + ComParE + PLP" feature set. When compared to the original test results from Table 4, the accuracy drop in the clean data is mild, whereas the model performs better in noise (see Table 8).

Table 8: *CORAAL Training Set: Inference Vs Training Testing on Baseline Dataset*

Dataset	Features	Test Clean Acc	Test Noise Acc
Baseline Testing	Librosa + ComParE + PLP	0.886	0.669
Inference Testing	Librosa + ComParE + PLP	0.873	0.766

Looking at the confusion matrices of the inference test set, the misclassification errors for noise are clustering in LES and VLD (see Figures 4 and 5), which is similar to misclassifications seen in the original noise test set. This result infers that audio quality in these cities is a variable. In the clean data set, the misclassification in ROC might be attributed to some unique characteristic of the unseen speakers, especially considering that classification was boosted in noise.

But regardless of these promising inference test results, additional speakers are sorely needed to improve overall model accuracy within these five regional accent classifications.

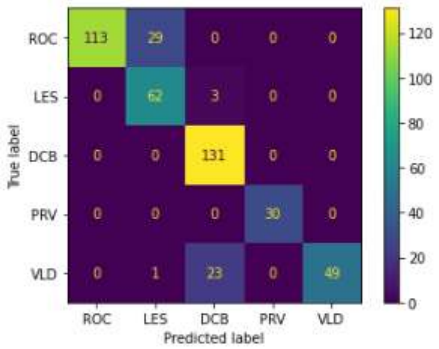

Figure 4: *Confusion matrix for inference test set on clean data using "Librosa + ComParE + PLP" baseline feature set*

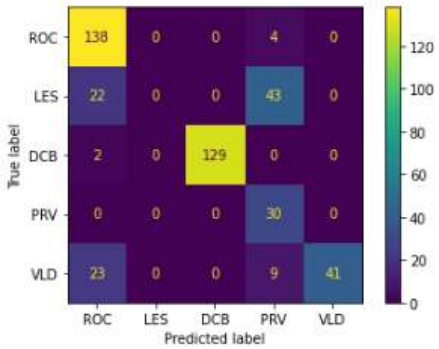

Figure 5: *Confusion matrix for inference test set on noisy data using "Librosa + ComParE + PLP" baseline feature set*

5. Discussion

Surprisingly high accuracies can be achieved on clean audio data with a suite of both noise resistant and less noise robust features. Based on the pattern of results in this study, evidence points to including noise resistant features in the feature set, but not relying exclusively on those features for model training. The evidence is seen in noise accuracies increasing when using the Librosa, PNCC, and PLP features together without ComParE 2016 – the noise and clean test data accuracy results begin to converge, but at the expense of overall model accuracy. This effect lends credence to handling noise elsewhere in the ASR model and avoiding adding noise to audio files. Regardless, the wide variety of noise to consider and train into models is cost prohibitive.

The poor results on the augmented dataset were surprising but appear to indicate errors in the augmentation implementation rather than overfitting in the model. The results in the baseline dataset were more aligned with expected performance, with the effects of "unseen" speakers accounting for the variations in performance.

As discussed earlier, dialect depends heavily on various variables such as age, gender, and socioeconomic class and this dataset contained all three variables. These subsets are further

complicated by the effects of the Great Migration (older speakers were likely born in the south) and the effects of cultural pride on younger speakers. Several speakers remain difficult to classify regardless of features selected. Listening to their audio files provides an explanation. These speakers tended to be older with a more pronounced "southern" accent, or the audio files presented as poorer quality than the DCB dataset, rendering some of the speech unintelligible to the human ear.

Without the ability to analyze the inference dataset, it is impossible to separate the effects of speaker demographics and audio quality from the effects of feature selection in the model. But based on the high variations in speakers paired with the low number of overall unique speakers, the performance of these features is quite robust.

Additional speakers would boost acoustic model accuracy, and perhaps multiple models are needed to cluster speakers born in Southern regions from younger speakers born and raised in these cities. Other models based on gender and socioeconomic class are also needed, based on the author's years spent attending the Boys and Girls Clubs (amongst a 90% African American cohort) and various residences and travels around the United States.

For future work on this topic, adding the phone level aligned text tags to incorporate a language model into the overall ASR model could further improve accuracy and hopefully help address the variations in accent seen in this cohort. Incorporating differences in grammar usage between AAE and GAE would improve the model because once these words are tagged, they can be isolated and studied for differences in accent between regions. The author knows the grammar and vocabulary differences exist due to hearing them firsthand, and a very well cited Wikipedia article agrees with this anecdotal assessment [16]. Beyond adding a robust language model to the ASR system, once feature selections were finalized on inference test sets, neural network models could be incorporated to fine tune overall model performance. Other backend features such as accent adaptation and perhaps deviations from expensive noise training such as "utterance level noise vectors" [17] are also needed but are beyond the scope of this study.

6. Conclusions

In this study, the impacts of various acoustic features were evaluated on the CORAAL corpus of AAE speakers using an XGBoost powered regional dialect classification model. The results of these experiments revealed that a suite of features including:

- MFCCs (the first 20)
- The Mel Spectrogram
- Rolloffs
- The complete ComParE 2016 feature set
- PLP

yielded the highest accuracies in both clean and noisy data, with the expected performance degradation in noisy data still evident but well mitigated.

7. References

[1] C. Huang, T. Chen, S. Li, E. Chang, and J. L. Zhou, "Analysis of speaker variability," in Proc. EuroSpeech, Aalborg, Denmark, Sep. 2001, vol. 2, pp. 1377–1380.

[2] Koenecke, A Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition." Proceedings of the National Academy of Sciences Apr 2020, 117 (14) 7684-7689; DOI: 10.1073/pnas.1915768117.

[3] Hamilton, Megan-Brette, "An Informed Lens on African American English," in The ASHA Leader, Jan-Feb 2020, vol. 25, issue 1. https://doi.org/10.1044/leader.FTR1.25012020.46.

[4] Hinsvark, Arthur, Delworth, Natalie, Del Rio, Miguel, et al. "Accented Speech Recognition: A Survey," arXiv:2104.10747 [cs.CL]. https://doi.org/10.48550/arXiv.2104.10747

[5] ORAAL, "What is AAL and who speaks it?," *Online Resources for African American Language*, https://oraal.uoregon.edu/AAL/What-is-AAL.

[6] Lanehart, Sonja L., and Ayesha M. Malik. 2015. "Language Use in African American Communities: An Introduction." In The Oxford Handbook of African American Language, edited by Sonja L. Lanehart, 1–22. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199795390.013.62.

[7] ORAAL, "How did AAL develop?," *Online Resources for African American Language*, https://oraal.uoregon.edu/AAL/Development.

[8] Kendall, Tyler and Charlie Farrington. 2021. *The Corpus of Regional African American Language.* Version 2021.07. Eugene, OR: The Online Resources for African American Language Project. [https://doi.org/10.7264/1ad5-6t35].

[9] Pedamkar, Priya. EDUCBA. https://www.educba.com/xgboost-algorithm/

[10] Guo, Rui, Zhao, Zhiqian, et al, "Degradation state recognition of piston pump based on ICEEMDAN and XGBoost," in Applied Sciences, September 2020. DOI: 10.3390/app10186593.

[11] Huang, Y., Yu, D., Liu, C., Gong, Y. (2014) Multi-accent deep neural network acoustic model with accent-specific top layer using the KLD-regularized model adaptation. Proc. Interspeech 2014, 2977-2981, doi: 10.21437/Interspeech.2014-497.

[12] M. Chen, Z. Yang, J. Liang, Y. Li, and W. Liu, "Improving deep neural networks based multi-accent mandarin speech recognition using i-vectors and accent-specific top layer," in Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[13] A. Jain, M. Upreti, and P. Jyothi, "Improved accented speech recognition using accent embeddings and multi-task learning." in Interspeech, 2018, pp. 2454–2458.

[14] M. Grace, M. Bastani, and E. Weinstein, "Occam's adaptation: A comparison of interpolation of bases adaptation methods for multi-dialect acoustic modeling with lstms," in 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018, pp. 174– 181.

[15] S. Yoo, I. Song, and Y. Bengio, "A highly adaptive acoustic model for accurate multi-dialect speech recognition," in ICASSP 2019- 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 5716–5720.

[16] "African-American Vernacular English," Wikipedia. https://en.wikipedia.org/wiki/African-American_Vernacular_English

[17] Desh Raj, Jesus Villalba, Daniel Povey, Sanjeev Khudanpur, "Frustratingly Easy Noise-aware Training of Acoustic Models," arXiv:2011.02090 [eess.AS], https://doi.org/10.48550/arXiv.2011.02090