ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096614

LEVERAGING MULTIPLE SOURCES IN AUTOMATIC AFRICAN AMERICAN ENGLISH DIALECT DETECTION FOR ADULTS AND CHILDREN

Alexander Johnson[1]*, Vishwas M. Shetty*[1]*, Mari Ostendorf*[2]*, and Abeer Alwan*[1]

[1]University of California, Los Angeles, Department of Electrical and Computer Engineering
[2]University of Washington, Department of Electrical and Computer Engineering

ABSTRACT

This paper[1] presents a novel system which utilizes acoustic, phonological, morphosyntactic, and prosodic information for binary automatic dialect detection of African American English. We train this system utilizing adult speech data and then evaluate on both children's and adults' speech with unmatched training and testing scenarios. The proposed system combines novel and state-of-the-art architectures, including a multi-source transformer language model pre-trained on Twitter text data and fine-tuned on ASR transcripts as well as an LSTM acoustic model trained on self-supervised learning representations, in order to learn a comprehensive view of dialect. We show robust, explainable performance across recording conditions for different features for adult speech, but fusing multiple features is important for good results on children's speech.

Index Terms— Dialect Identification, African American English, Children's Speech, Language Modeling

1. INTRODUCTION

Dialect (and accent) identification (DID) techniques have recently been subjects of great interest [1]. Successful prior work in this area includes [2] which identifies optimal time-frequency representations for a CNN-backend for English accent classification and [3] which uses a variational autoencoder to generate unsupervised learning representations for efficient Chinese and Arabic DID.

Despite these advances, DID still presents several challenges. First, as [4] notes, it is difficult to recognize accent and dialect across different speaker styles. Another challenge in DID is the multi-facetedness of the definition of dialect. A dialect is a set of pronunciations, diction, grammar, and prosodic cues that are native to the speech of a region or group [5]. Most commonly used state-of-the-art DID systems only examine a subset of these linguistic aspects, resulting in an incomplete representation of dialect. This motivates the need for DID systems that combine several sources of information to form more comprehensive classification decisions. Comprehensive language representations becomes even more important for low-data scenarios. While large end-to-end models may be able to implicitly learn characteristics of dialect in data-rich cases, they cannot avoid overfitting while training to recognize aspects of a low-resource language. Therefore, smaller models that incorporate linguistic knowledge to scaffold the training into targeted processes are preferable for low-resource DID. Of particular interest would be further studies on the role of prosody in DID. Works such as [6] show that humans use prosody when classifying dialects, and several studies (eg. [7, 8]) have found prosody useful for automatic DID, but best practices for incorporating it are not well-defined. This paper explores DID for two low-resource scenarios: 1) African American English (AAE) and 2) Children's speech. AAE is the set of regional dialects spoken by African Americans throughout the United States. It includes differences in phonology, morphology, syntax, and prosody from the Mainstream American English (MAE) dialect [9, 5]. AAE is an understudied dialect in speech processing, and there is only a small number of speakers whose labeled speech is available publicly. Automatic dialect and accent identification for children is also an area in need of further study. As children's pitches are higher than those of adults, and their pronunciations and prosody are subject to greater variability, children's speech can present challenges to systems trained only on adult's speech [10]. DID for children would be particularly useful for automatic bias mitigation in education, as child speakers of non-standard dialects are often perceived as less intelligent or possessing language impairments [11]. A few works like [12] show promise in children's accent identification by combining prosodic information with MFCCs in feature processing for a k-nearest neighbor system, but further studies are needed to show how more powerful models can take advantage of this information.

Domain adaptation has been a widely-used method of combating the difficulties seen in low-resource cases. Specifically, systems that seek to utilize large amounts of out-of-domain, high-resource data in order to learn general trends that are applicable to low-resource tasks have been successful. For example, [13] uses domain-attentive fusion to learn domain generality for unseen samples. Another example of this is shown in self-supervised learning representations, like those from Hubert [14], which have been applied to a variety of small-data tasks after being trained on large amounts of out-of-domain data [15]. In this paper, we use out-of-domain data to train a combination of neural networks, each targeted towards a different linguistic aspect of dialect, to perform low-resource DID. The novelty of our approach comprises of 1) A method which fuses different models, including multi-modal language models which use Twitter text data, in order to create a robust and interpretable DID system, 2) Performing a detailed study on which machine-readable features relate most strongly to the presence of dialect, and 3) Generalization of results across both adult's and children's speech.

2. METHODS

In this paper, we create a system for binary classification of utterances as containing AAE or not containing AAE speech. An utterance is defined as belonging to the AAE class if it contains at least one phonological or morphosyntactic marker of AAE as transcribed by expert transcribers. The following section describes the datasets and models used in the experiments.

[1]This work is supported in part by the NSF.

2.1. Datasets

The focus of this work is on dialect detection given spontaneous speech, particularly adult and children's AAE speech. There is no dataset available for this task, so we build on multiple datasets, as described below. The AAE data used in this work reflects southern variants, due to the availability of such data for children's speech.

2.1.1. Speech Datasets

A particular challenge in this work is learning dialect representations that are robust to recording conditions, speaker style, and speaker traits (eg. age, gender, et.). We select these datasets for their coverage of a wide range of these scenarios. All speech data are resampled to 16kHz for experimentation. The utterances used are each approximately 5-15sec in length.

CORAAL. The Corpus of Regional African American Language (CORAAL), [16] is a speech dataset containing recordings of oral interviews with speakers of AAE. These recordings contain spontaneous speech in a variety of different recording conditions. In this work, we utilized the recordings of speakers from Princeville, NC, Valdosta, GA, and Washington DC, as speakers from those cities had the highest average dialect density or frequency of use of dialectal characteristics [17, 8]. From these speakers, we selected utterances that contained at least five spoken words, as denoted by the ground truth transcripts, and were free of non-speech sounds. This resulted in a speaker-independent training and test set totalling approximately 20 hours and 2 hours of speech respectively.

Librispeech. In order to show how availabe large out-of-domain datasets can be used for training, we use the popular Librispeech corpus [18] to train models to learn the negative class (samples that contain only MAE and no AAE). We randomly selected utterances from train-clean-100 dataset to create a training set and utterances from the dev-clean set to create a validation set. These speaker-independent data splits were created to contain the same number of utterances as those from CORAAL.

SITW. The Speakers in the Wild Challenge (SITW) dataset [19] contains recordings of conversational speech in various recording environments, primarily involving MAE speakers. We randomly selected a subset of the same number of utterances as that of the CORAAL test set. This subset is used only for testing and serves as a reference for spontaneous, non-dialect speech in background noise.

GSU Kids: The Georgia State University Kids' Speech Dataset [2] (GSU Kids) [20] is a speech dataset of approximately 200 children aged 8-13 from the Atlanta, GA area. The children were recorded in a noisy classroom environment as they performed educational assessments in story-telling and picture-description tasks. The children's speech was annotated by the authors for aspects of AAE dialect, and the dataset was subsequently divided into AAE-dialect and non-AAE dialect speaking children. In this work, a subset of approximately 800 utterances totalling about 3 hours was randomly selected for use such that approximately half of the utterances contained AAE speech. In order to determine which children in the dataset spoke AAE, the dataset was annotated for dialect tokens that are widely accepted to be common markers of AAE as in [17].

The speaking styles and train/test usage of different data sets are summarized in Table 1. We use "non-AAE" instead of MAE for the Kid's speech, since it is mostly a southern dialect. The adult corpora may also contain dialects that are not MAE, but the data are dominated by the MAE dialect.

Source	Dialect	Style	# speakers Train/Test		avg # test utt./spkr
CORAAL	AAE	spon/noisy	61	11	72
Libri	MAE	read/clean	251	40	20
SITW	MAE	spon/noisy	–	119	6
GSU Kids	AAE	spon/noisy	–	117	3
GSU Kids	non-AAE	spon/noisy	–	76	4

Table 1. Summary of characteristics and usage of speech datasets. We show the number of speakers used in training and testing to highlight the low-resource problem caused by the lack of available training data from AAE speakers. The datasets with no entry in the "Train" column were used only for testing. We also include the average number of utterances per speaker in each test set. There are approximatley 8000 utterances in each training set, 800 utterances in the CORAAL, Librispeech, and SITW test sets, and approximately 400 utterances in the GSU AAE and GSU non-AAE test sets.

2.1.2. Text Data

In order to train language models for dialect detection, we utilize two large corpora of Twitter text data. All Twitter text is preprocessed to match wav2vec ASR transcript format. The data is lowercased, and we remove hashtags, mentions, and punctuation (excluding periods and apostrophes). While primarily adult twitter data may be less applicable for training models for children's speech, the volume and availability of the data makes it an interesting use case.

TwitterAAE [21] is a dataset of over one million tweets that were automatically found to have a high probability of being authored by a speaker of AAE. Through training a probabilistic model that took into account the geographic location of the tweeter, the N-gram probability of the words used in the tweets, the grammatical structure of the tweet as identified by an automatic part-of-speech tagger, and the presence of AAE syntax, these tweets were found to display many common aspects of AAE.

The **Sentiment 140** dataset [22] is a database of 1.6 million tweets on various subjects labeled with the corresponding user sentiment of the message. In this work, we use this dataset as a reference set of non-AAE text of the same format as text of Twitter AAE.

2.2. Models

We train several models, each using one of three different architectures (CNN, LSTM, or BERT-style masked language model), to learn different aspects of dialect from different linguistically-focused features of the data. The goal of the model training is binary classification of the input data as containing or not-containing AAE speech.

2.2.1. CNN

We use a modified version of the Convolutional Neural Network from [2] to map acoustic and prosodic features to dialect. The CNN layers had kernel sizes of 4x4 with: kernel strides of 1, 16 output channels in the first layer, and 32 output channels in the second layer. The convolutional layers were followed by max pooling and then two fully connected layers that mapped to the final output decision. While [2] found that the spectrogram was the best feature for DID, [12] saw more success using MFCCs. We evaluate the performance of both of these features for child and adult DID. We extract

[2]The GSU data was collected with support by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the NIH under Grant P01HD070837.

the spectrogram with a window size of 10ms and window shift of 5ms. For the MFCCs, we extract the 20dim-feature along with the first and second derivatives. We additionally use prosodic features as described in [23, 8]. These include the F0 contour extracted with Praat [24], the energy contour of the signal, the energy contour of the signal lowpass filtered at 1kHz, and the energy contour of the signal highpass filtered at 1kHz. We perform DID both using the prosody features alone and in concatenation with the best from the MFCC and spectrogram features in the CNN.

2.2.2. LSTM

We employ the popular self-supervised learning representations extracted by Hubert [14] in this task. The Hubert hidden layer outputs are input into a one-layer 128-dim Long Short Term Memory (LSTM) layer and then two fully connected layers with sizes of 64 and 1 to make the binary dialect classification decision.

2.2.3. Language Models

One prominent difference between AAE and MAE is the pronunciation of certain words in given contexts. For example, Southern AAE may include reductions of word final consonant clusters (e.g. pronouncing "band" as "ban") and a raising of the /IH/ vowel (e.g. pronouncing "kill" as 'keel') [5]. Character-level ASR systems may capture these pronunciation differences. We use a Wav2Vec2.0 model [25] trained on the Switchboard Telephone Corpus [26] to generate ASR transcripts for the speech data. We evaluate the performance of the ASR system and find it consistent with previously reported results on AAE and non-AAE speech for the given cases [17]. Using the ASR transcripts as input, we apply a character-level BERT-style transformer language model (LM) [27], pre-trained using a masked language model (MLM) objective and finetuned to distinguish between the AAE and MAE text in a binary classification task. The use of the LM allows us to take advantage of large language models that benefit from large amounts of text data and utilize the abundant text data on Twitter. We explore two LM configurations, both building on a pretrained small BERT model,[3] with the CLS token embedding input to a single fully connected layer used to decide whether or not the speech contains AAE dialect. One model simply trains this classifier with a cross-entropy (CE) objective using the two sources of Twitter data, also updating weights of the BERT model. For the second model, we further pretrain the model with the MLM objective on the Twitter data, followed by additional pretraining on the Librispeech and CORAAL ASR transcripts. We then train the last classification layer with the LM weights frozen using CE with the CORAAL-Libri transcripts, and finally further fine-tune the full model for a few iterations with CE on the ASR transcripts.

Grammatical features are another defining aspect of AAE. For example, AAE can include a dropping of auxiliary verbs (e.g. "he gone" instead of "he is gone") or a deletion of the infinitive marker "to" (e.g. "it's your turn go" instead of "it's your turn to go"). In order to capture these differences, we applied the automatic part-of-speech (POS) tagger from the Python SpaCy library to the Twitter text data and the ASR transcripts. For example, the POS tagger may take the transcript, "who all goin" as input and produce the output sequence of the same length, "PRON DET VERB." Anecdotally, we find that even when the ASR system spells words differently than in the standard English dictionary, these words are often tagged as the correct part of speech (e.g. tagging "goin" as a verb here). We then learn a token-level transformer language model using MLM pretraining on the Twitter data to predict dialect as MAE or AAE from the sequence of POS tags, similarly to the character LM.

[3]https://tfhub.dev/google/collections/bert/1

2.3. Experiments

Using the features listed above, we train the CNN, LSTM, and Bert MLM to perform AAE DID. All systems are trained with the CORAAL training set as the positive class and the Librispeech training set as the negative class. The language models are additionally pre-trained on the Twitter text data. Although the positive samples come entirely from one dataset and the negative samples come entirely from another, we chose training datasets that are each compilations of various recordings from across different speakers, years, locations, and recording devices, meaning that there will not likely be spurious channel effects or recording conditions that can help distinguish recordings of the same database. We evaluate the performance training on CORAAL (noisy, spontaneous) and Librispeech (clean, read) in two cases: Resolving AAE-speech in CORAAL from the non-AAE speech in SITW (noisy, spontaneous) and Resolving the AAE-speech from the non-AAE speech in the GSU Kids' speech database (noisy, spontaneous). This will show the robustness of the systems to different speaking styles and recording conditions. We additionally show the performance of score-level fusion of the best models. The model output scores are added and then the new detection threshold is taken to be the median score of the test set. This method of fusion allows us to fuse the scores in the case when we do not have enough data to create a separate validation set to train a fusion model. We choose the median confidence score as the threshold because we know in advance that the test sets are balanced in the number of utterances in each class. In a real scenario, the demographics of a group of users would likely be known, and the threshold could be chosen to match those demographics (eg. if the system were used in an area where approximately two-thirds of the population spoke AAE then the threshold could be set at the 33rd percentile value of the output scores if it could not be found through validation). In order to show the performance of the fused models without respect to threshold, we also calculate their Area under the ROC Curve (AUC) values.

3. RESULTS AND DISCUSSION

Table 2 shows the performance of the individual models trained on a particular feature or a concatenation of 2 features. Each row shows the input features to the model, the model backend, the target linguistic correlate of the model, and the accuracy and F1 score of that model for the validation set and two test sets. Table 3 shows the Accuracy, F1 score, and AUC for the models. In Table 2, the models are trained with a detection threshold of 0.5. The fused models in Table 3 use the median value of the testset as the detection threshold. Therefore, we recalculate the performance of the individual models with the median threshold for inclusion in Table 3 in order to show the effects of thresholding and fusion separately.

We observe that several of the individual models, including those trained on the spectrogram, MFCC, and prosody features perform significantly worse for the children's speech test set than for the adult speech test set. This may be an indication that these models overfit to the acoustic features or speaker style of the adult speech. The largest drop is for prosody features; it may be that prosody is less reliable for children because of the high F0 and disfluencies and/or because of greater variability, The model trained on the concatenated spectrogram and prosody features performs better than the models trained on either feature individually in nearly

Feature	Backend	Linguistic Correlate	Validation Set (CORAAL AAE vs. Librispeech MAE)		CORAAL AAE vs. SITW MAE		GSU AAE vs. GSU non-AAE	
			Acc.	F1	Acc.	F1	Acc.	F1
1. Spectrogram	CNN	Acoustic	91.1	92.2	72.9	76.5	55.3	54.2
2. MFCC	CNN	Acoustic	73.8	83.5	60.5	69.8	55.7	58.3
3. Prosody feat	CNN	Prosody	90.8	91.2	83.3	80.1	52.4	52.9
4. concat(Spec.,Pros)	CNN	Acoust, Pros.	91.8	92.9	88.9	88.9	58.2	55.6
5. Hubert feat	LSTM	Acoustic	78.1	87.7	71.1	82.9	64.8	74.3
6. Char-level text pre-train Twitter	MLM	Phonology	82.6	79.9	66.9	56.8	51.5	58.9
7. Char-level text finetune CORAAL-Libri	MLM	Phonology	91.0	89.3	88.2	81.4	62.7	71.2
8. POS-token pre-train Twitter	MLM	Grammar	69.2	60.7	67.5	60.1	46.8	61.4
9. POS-token finetune CORAAL-Libri	MLM	Grammar	84.8	77.4	87.1	77.5	55.2	68.4

Table 2. The results of binary classification for each model using 0.5 as the detection threshold. For each model, we present the targeted linguistic correlate of dialect (Acoustic Phonetics (Acoustic), Phonology, Morphology/Syntax (Grammar), or Prosody (Pros)) and the Accuracy (Acc.) and F1 score (as calculated by Python SKlearn). Twitter refers to both TwitterAAE and Sentiment140 text data.

Model	CORAAL AAE vs. SITW MAE			GSU AAE vs. GSU non-AAE		
	Acc.	F1	AUC	Acc.	F1	AUC
4.	90.0	89.6	90.2	55.4	55.4	55.6
5.	76.9	84.4	75.4	65.6	76.2	57.3
7.	88.6	85.4	77.3	61.8	70.2	62.3
4 + 5	88.6	89.8	78.1	67.3	72.5	65.5
4 + 7	86.1	86.3	77.3	61	68.2	62.6
5 + 7	89.2	83.4	79.4	68.6	74.4	69.2
4 + 5 + 7	89.5	86.8	81.1	70.7	77.6	70.4

Table 3. The results of binary classification for the individual and fused models when the threshold is taken as the median output score. We also report the AUC values as threshold-invariant metrics.

all cases, showing that these features may provide complementary dialect information. This model (4) does better than any other individual models for the CORAAL vs. SITW test set, suggesting that the combination of spectrogram and prosody made the model more invariant to the change in speaker style between the training and test case. However, this model still does not generalize well to the children's test set. Although the model trained on Hubert self-supervised learning representations performs worse for the validation set than the other acoustic features, it appears to generalize much better to the children's speech. This may be because the wide range of speaker variability seen by Hubert during pre-training has allowed it to learn more robust representations of higher-pitched voices and disfluent speech as seen in children. Both language models see a significant improvement after being fine-tuned on data from the ASR transcripts. The character-level MLM trained directly on the transcripts seems to learn information about AAE pronunciations from the Twitter and ASR transcript data that meaningfully translate to other datasets. The grammar-based MLM trained on POS tags does more poorly. This may be due to tagging errors or indicate that dialect-specific grammatical patterns are not consistent enough across age and geographic region to be useful for classification. Table 3 shows that fused models improve performance over individual models for the children's data, but give no significant benefit for the adult test set. The model trained on Hubert features seems most important to obtaining good results on the kids' speech, as the fused model without it does less well for the GSU test set. The fusion of the models trained on concatenated spectrogram and prosody features, the Hubert features, and the language modeling representations gives the best results for children, with statistically significantly higher accuracy and F1 scores than any other model.

The table also shows that use of the median threshold with the individual models improves performance for the adult test set compared to the 0.5 threshold, especially for the Hubert features. This may suggest that the detection threshold should be shifted with a shift in domain, and further studies are needed to create thresholding strategies that do not require large amounts of in-domain development data for low-resource cases. For the children's speech case, only the model (5) sees an improvement from the change in threshold. Comparing the individual models in Table 3 to the fused model, we see that the model (4 + 5 + 7) still shows significantly better performance for the children's speech and is not significantly worse than any model for the adult speech. Note that this model also has the highest AUC for the children's case and good AUC for the adult's case. This indicates that fusion may be a promising method of capturing dialectal differences in children's speech.

4. CONCLUSIONS AND FUTURE WORK

This study introduces a framework for DID of AAE dialect by drawing linguistic information from several features and training strategies which deliberately target multiple aspects of dialect. This is a particularly difficult task, as many of the distinguishing features of AAE are underrepresented in speech datasets for adults, let alone children. We show that compensating for this by both incorporating Twitter text data into a Bert MLM and by fusing different features, including self-supervised learning representations and prosodic features, yields a promising method for advancing low-resource DID, especially for children's speech. Future work includes analyzing specific aspects of children's speech, especially prosodic speech patterns, which cause confusions in DID.

5. REFERENCES

[1] A. Ali et al., "The mgb-5 challenge: Recognition and dialect identification of dialectal arabic speech," in *IEEE ASRU*, 2019, pp. 1026–1033.

[2] Mariia Lesnichaia, Veranika Mikhailava, Natalia Bogach, Iurii Lezhenin, John Blake, and Evgeny Pyshkin, "Classification of Accented English Using CNN Model Trained on Amplitude Mel-Spectrograms," in *Proc. Interspeech 2022*, 2022, pp. 3669–3673.

[3] Q. Zhang and J. H. L. Hansen, "Language/dialect recognition based on unsupervised deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 5, pp. 873–882, 2018.

[4] G. Liu, Y. Lei, and J. H. L. Hansen, "Dialect identification: Impact of differences between read versus spontaneous speech," in *2010 18th European Signal Processing Conference*, 2010, pp. 2003–2006.

[5] S. Lanehart and A. M. Malik, "Language Use in African American Communities: An Introduction," in *The Oxford Handbook of African American language*, J. Bloomquist, L. J. Green, and S. L. Lanehart, Eds. Oxford University Press, Oxford, 2015.

[6] N. R. Holliday and Z. S. Jaggers, "Influence of suprasegmental features on perceived ethnicity of american politicians," in *Proc. 18th Int. Congress Phonetic Sci.*, 2015.

[7] J. Lee, K. Kim, and M. Chung, "Korean dialect identification based on intonation modeling," in *2021 24th Conference of the Oriental COCOSDA International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, 2021, pp. 168–173.

[8] A. Johnson, K. Everson, V. Ravi, A. Gladney, M. Ostendorf, and A. Alwan, "Automatic dialect density estimation for african american english," in *Interspeech*, 2022.

[9] E. R. Thomas, "Prosodic Features of African American English," in *The Oxford Handbook of African American language*, J. Bloomquist, L. J. Green, and S. L. Lanehart, Eds. Oxford University Press, Oxford, 2015.

[10] Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan, "Analysis of children's speech: duration, pitch and formants," in *Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, 1997, pp. 473–476.

[11] J. A. Washington and H. K. Craig, "Dialectal Forms during Discourse of Poor, Urban, African American Preschoolers," *J Speech Hear Res.*, pp. 816–23, 1994.

[12] Kodali Radha, Mohan Bansal, and Shaik Mulla Shabber, "Accent classification of native and non-native children using harmonic pitch," in *2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP)*, 2022, pp. 1–6.

[13] S. Shon, A. Ali, and J. Glass, "Domain attentive fusion for end-to-end dialect identification with unknown target domain," in *ICASSP*, 2019, pp. 5951–5955.

[14] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *CoRR*, vol. abs/2106.07447, 2021.

[15] Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee, "SUPERB: speech processing universal performance benchmark," *Interspeech*, vol. abs/2105.01051, 2021.

[16] T. Kendall and C. Farrington, "The Corpus of Regional African American Language. Version 2021.07.," 2021.

[17] A. Koenecke et al., "Racial Disparities in Automated Speech Recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR Corpus based on Public Domain Audio Books," in *ICASSP*, 2015, pp. 5206–5210.

[19] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The Speakers in the Wild (SITW) Speaker Recognition Database," in *Proc. Interspeech 2016*, 2016, pp. 818–822.

[20] A. Johnson, R. Fan, R. Morris, and A. Alwan, "LPC AUGMENT: An LPC-Based ASR Data Augmentation Algorithm for Low and Zero-Resource Children's Dialects," *ICASSP*, 2022.

[21] S. Blodgett, L. Green, and B. O'Connor, "Demographic dialectal variation in social media: A case study of African-American English," in *EMNLP*, Austin, Texas, Nov. 2016, pp. 1119–1130, Association for Computational Linguistics.

[22] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *Stanford CS224N Project Report*, 01 2009.

[23] T. Tran et al., "Parsing speech: a neural approach to integrating lexical and acoustic-prosodic information," in *Proc. NAACL*, 2018, pp. 69–81.

[24] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.13)," 2009.

[25] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A Framework for Self-supervised Learning of Speech Representations," in *NeurIPS*, 2020.

[26] J. J. Godfrey and E. Holliman, "Switchboard-1 release 2," *Linguistic Data Consortium*, vol. LDC97S62, 1993, Web Download.

[27] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *ArXiv*, vol. abs/1810.04805, 2019.