JASA ARTICLE

An exploratory study on dialect density estimation for children and adult's African American English[a]

Alexander Johnson,[1,b] Natarajan Balaji Shankar,[1] Mari Ostendorf,[2] and Abeer Alwan[1]
[1]*Department of Electrical and Computer Engineering, University of California, Los Angeles, Los Angeles, California 90095, USA*
[2]*Department of Electrical and Computer Engineering, University of Washington, Seattle, Washington 98195, USA*

ABSTRACT:
This paper evaluates an innovative framework for spoken dialect density prediction on children's and adults' African American English. A speaker's dialect density is defined as the frequency with which dialect-specific language characteristics occur in their speech. Rather than treating the presence or absence of a target dialect in a user's speech as a binary decision, instead, a classifier is trained to predict the level of dialect density to provide a higher degree of specificity in downstream tasks. For this, self-supervised learning representations from HuBERT, handcrafted grammar-based features extracted from ASR transcripts, prosodic features, and other feature sets are experimented with as the input to an XGBoost classifier. Then, the classifier is trained to assign dialect density labels to short recorded utterances. High dialect density level classification accuracy is achieved for child and adult speech and demonstrated robust performance across age and regional varieties of dialect. Additionally, this work is used as a basis for analyzing which acoustic and grammatical cues affect machine perception of dialect.
 https://doi.org/10.1121/10.0025771

(Received 24 December 2023; revised 29 March 2024; accepted 2 April 2024; published online 29 April 2024)
[Editor: James F. Lynch] Pages: 2836–2848

I. INTRODUCTION

Language identification (LID) and dialect identification (DID) have become integral parts of many large spoken language systems. For example, many multilingual automatic speech recognition (ASR) systems, such as OpenAI's Whisper (Radford *et al.*, 2022) and Meta's Massively Multilingual Speech models (Pratap *et al.*, 2024), leverage large cross-lingual speech corpora for training and then perform LID during inference. Other systems, like AWS transcribe (AWS, 2023), offer DID for commercial use cases, distinguishing input speech, for example, between English dialects from the U.S., U.K., or India for better performance on regional dialects. As these models expand to support more languages and dialects, several challenges arise: First, data-driven DID methods that rely on the availability of large amounts of dialect-labeled speech may not generalize to less well-resourced dialects and variations. Second, even within a dialect, these systems are typically only trained on adult speech. Therefore, many DID systems are unable to accurately predict dialect for children's speech, making them unsuitable for speech applications in early education. Third, some speakers may use more or fewer aspects of a dialect than others (as some people are perceived to have a thicker accent than others). As such, categorizing all speakers of a dialect into the same label group regardless of the frequency of use of dialect-specific pronunciations, grammar patterns, and prosodic patterns may lead to inaccurate representations of some speakers in downstream applications.

Despite recent advances in DID systems, few works have been proposed to better explain which acoustic and linguistic cues are essential for machines to accurately predict certain dialects. Studies, such as that in Holliday (2021), attempt to better understand which acoustic and prosodic cues are used by listeners to determine a speaker's perceived ethnicity or dialect. However, it is largely unknown if machines use the same cues as humans to perform DID and, if so, to what extent they apply them. This motivates the need for further research on explainable DID systems in which the importance of different types of input cues can be further analyzed and compared to known phenomena in humans.

In this paper, we build on the dialect density estimation system originally proposed in Johnson *et al.* (2022a) to address these challenges. Particularly, we seek to better understand what acoustic cues, in addition to known morphosyntactic cues, affect machine perception of dialect. Dialect density is the frequency with which a speaker uses dialectal differences that are not present in a reference dialect (Craig and Washington, 1994; Washington and Seidenberg, 2022). Therefore, automatic dialect density estimation consists of predicting a speaker's dialect density from a short input sample of their speech. A machine can then use this estimate for better downstream model selection, tuning of decoding parameters, or data sampling techniques. The dialect density labels need not be mutually

[a]This paper is part of a special issue on Acoustic Cue-Based Perception and Production of Speech by Humans and Machines.
[b]Email: ajohnson49@g.ucla.edu

exclusive between multiple dialects and can encode dialectal aspects of grammar and pronunciation separately if desired.

This work proposes a model for African American English (AAE) dialect density estimation from short utterances on children's and adults' speech (utterances of length 30–90 s for adult speech and 2–3 min for children's speech). As education literature has demonstrated, speakers of minority dialects like AAE are often underrated in language abilities because of raters who are unfamiliar with AAE, interpreting dialectal differences as language deficiencies (Washington *et al.*, 2018). In particular, children with higher AAE dialect density have been shown to underachieve in schools that primarily teach in mainstream American English (MAE; Washington *et al.*, 2018). Therefore, DID in educational spoken language systems could be used to detect and mitigate this bias, creating a pressing use case for the dialect explored in this work. First, we train and test the proposed system on a dataset of adult's AAE. We, then, show the generality of the feature extraction and model training paradigm to children's speech by training and testing the proposed model on a corpus of spontaneous children's speech from AAE and non-AAE-speaking students from the Atlanta, GA area.

Although the phonetic and morphosyntactic dialectal features of AAE have been well-documented (Lanehart and Malik, 2015; Thomas, 2015), few studies have been performed to collect data or improve ASR system performance for the dialect, giving it status as a low-resource dialect. Notably, Koenecke *et al.* (2020) identify a performance gap between MAE and AAE for several commercial ASR systems and point to insufficiently trained acoustic models as a possible cause. They also show that commercial ASR system performance worsens as a function of increasing AAE dialect density. The model proposed in our work fuses traditional acoustic features, state-of-the-art neural network representations, and handcrafted features designed to detect documented aspects of AAE to create robust predictions of dialect density. The model combines information relating to acoustic phonetics, prosody, and morphology. We show high performance of the model for AAE-speaking children and adults, as well as offer insights on how machines can better deal with the dialectal linguistic differences present. Additionally, we show the impact of input features on the dialect density classifier to interpret how they affect the model and interact with each other. Next, we summarize the previous works related to this paper.

A. Related works

Several recent studies have offered promising DID systems for a limited number of dialects. Liao *et al.* (2023) introduce a time delay neural network, as popularized by the *X*-vector speaker embedding (Snyder *et al.*, 2018), with attention across time and frequency for classifying among a set of 16 dialects. The experiments performed in Tzudir *et al.* (2022a) also found frequency-based data augmentation to be beneficial in training a recurrent neural network to classify low-resource dialects with either speaker embeddings or a combination of Mel frequency cepstral coefficients (MFCCs) and other acoustic features. Yadavalli *et al.* (2022) designed a multitask learning framework for a conformer-based system that jointly learns to output ASR transcripts and DID labels for speech from three Telegu dialects. To overcome performance degradation caused by domain mismatch in end-to-end DID systems, Shon *et al.* (2019) create a domain-attentive fusion technique to better classify African and Arabic dialects across recording conditions and speaking styles.

Despite these advancements, several challenges remain in DID, especially for widely spoken languages such as English, which display wide variability within and across groups. For example, although many current DID systems may categorize U.S. English as distinct from British English, they do not recognize differences between MAE, AAE, Southern American English, Creole English, and other varieties. The work in Duroselle *et al.* (2021) shows that ASR systems with more knowledge of the different dialects, achieved by joint training on DID and ASR, often perform better across those dialects, implying that adding more specificity to the DID pipeline would improve the performance of downstream tasks. However, it is neither simple nor scalable to simply attempt to train current DID systems to distinguish between larger sets of dialects. First, several dialects are low-resource dialects, which means that there is not enough publicly available speech data to train large spoken language models to recognize them. Second, speech samples cannot always be categorized neatly into one dialect. Many speakers code-switch, alternating between different languages or dialects (Martin-Jones, 1995) or incorporate aspects of multiple dialects into their speech. The degree of the speaker's code-switching may depend on several factors such as the speaking style or formality of the conversation (Labov, 2006). Assigning discrete labels to samples from these speakers and forcing a model to choose a single dialect for them would likely propagate error through the system. Third, many current DID models only classify dialect from acoustic features like spectrograms or MFCCs, which mainly discern differences in pronunciation (e.g., Ali *et al.*, 2019; Lei and Hansen, 2011; Mawadda Warohma *et al.*, 2018). However, sociolinguistic variations can differ in several aspects other than just pronunciation (e.g., prosody, grammar, and diction). Previous works which have combined prosodic cues with spectral information (Tzudir *et al.*, 2022b) or attempted to classify language or dialect from grammatical features of text (Zissman and Berkling, 2001) have shown that considering other aspects of language can improve automatic DID. This is especially beneficial in DID for speakers, like children, with relatively high acoustic variability. Although children's developing vocal tracts and articulatory motor skills may cause their speech to display different acoustic properties than adults' speech (Lee *et al.*, 1999), work in Johnson *et al.* (2023a) shows that incorporating prosodic and grammar information into DID systems trained on adult's speech can make them more robust for children. Improving DID for children's speech is of particular interest in educational speech technology, as mentioned previously. Applications such as Read Along by Google (Google, 2023) use ASR and natural language processing (NLP) to recognize and provide pronunciation and literacy feedback to children as they practice reading aloud.

1. Dialect density

Originally proposed in educational studies on AAE children's language usage, dialect density is a metric for measuring how much dialectal influence appears in a speaker's speech (Craig and Washington, 1994; Seymour *et al.*, 1998; Washington *et al.*, 1998). It is common to measure AAE dialect density as the percentage of words or sentences of a speaker's speech that contain well-documented AAE dialectal characteristics that are not present in MAE speakers. The language differences between MAE and AAE may cause student speakers of AAE to be observed as developing language skills incorrectly and, therefore, education researchers have found it necessary to measure one's frequency of dialect usage separately from their pronunciation abilities (Moyle *et al.*, 2014), lexical comprehension (Edwards *et al.*, 2014), and other markers of language development (Van Hofwegen and Wolfram, 2010). Drawing inspiration from these studies, we aim to enable ASR systems with similar capabilities so that they can mitigate bias that may come from dialect-specific constructions.

B. Roadmap

In Sec. II, we describe the structure of the proposed feature extraction pipeline and classification model for dialect density estimation. Then in Sec. III, we present the results of evaluating the system on adult speech from the Corpus of Regional African American Language (CORAAL) database and children's speech from the Georgia State University Kid's (GSU Kids) speech database. Section IV presents a discussion and analysis of the results. Section V provides conclusions and future work.

II. METHODS

The overall goal of this work is to train a classifier to predict the frequency and strength of a speaker's dialect usage from a short input utterance. The amount of dialect usage can be represented numerically with a dialect density measure (DDM), which gives the percentage of words in an utterance that contain a documented phonological or morphosyntactic characteristic of dialect. Here, we train a classifier to map features extracted from an utterance to the hand-labeled DDM. This section describes the datasets, feature sets, and models used in this work.

A. Datasets

This study uses adult AAE speech data from the CORAAL (Kendall and Farrington, 2021) and children's speech data from the GSU Kids speech database (data collected in Fisher *et al.*, 2019, and structured in Johnson *et al.*, 2022b). An overview of each dataset is provided. Statistics about each set and the average dialect density for the speakers are shown in Table I.

1. CORAAL

The CORAAL dataset contains recordings of interviews with AAE speakers from a variety of socioeconomic backgrounds, ages, and cities throughout the East Coast of the U.S. We use speech from five different cities in the database: Rochester, NY (ROC); Lower East Side Manhattan, NY (LES); Washington, DC (DCB); Princeville, NC (PRV); and Valdosta, GA (VLD). We avoid using recordings from the DCA (a dataset from Washington, DC, recorded two decades prior to DCB) or DTL (data recorded in Detroit, MI) datasets as these were recorded decades before the others on dissimilar devices. Preliminary experiments show that recordings from these datasets are easily distinguishable by recording device and dialect, adding confounding factors to experiments which may seek to separate recordings by regional dialectal characteristics. There was a total of 65 different speakers from across the 5 regional datasets used. The speakers ranged in age from young teens to over 90 years old. The speakers also span a range of socioeconomic groups, although this information is not available for several speakers and, thus, we do not focus on drawing conclusions from the speakers' reported socioeconomic status. From each speaker, we took 2–3 utterances, each 30–90 s in length (as performed in Koenecke *et al.*, 2020), which were annotated for dialect density. This totaled 208 utterances (about 2 h) of dialect density-labeled adult AAE speech. Despite the fact that the CORAAL dataset contains hundreds of hours of speech, the number of different speakers from whom distinct dialectal patterns can be observed is far more limited, leading to the smaller dataset used in this work. The number of utterances and speakers from each city are provided in Table I. The utterances from ROC, PRV, and DCB were selected and labeled for dialect density by Koenecke *et al.* (2020), and the utterances from VLD and LES were selected and labeled by authors of this work.[1] Note that speakers from PRV and VLD, on average, have higher dialect densities than speakers from the other cities, possibly because those southern cities have historically had larger populations of AAE speakers. The audio recordings were originally sampled at 44.1 kHz and downsampled to 16 kHz for experimentation.

TABLE I. Number of utterances, Number of speakers, and average DDM (avg. DDM) of dialect from each city for the labeled portion of the CORAAL database used and the Georgia State University Kids Speech Corpus.

City	ROC	LES	DCB	PRV	VLD	GSU Kids
Number of utterances	50	30	50	50	28	203
Number of speakers	11	10	22	10	12	203
Avg. DDM	0.047	0.042	0.088	0.194	0.141	0.040

2. GSU Kids speech database

This dataset contains audio recordings of 203 children aged 9–13 years old from the Atlanta, GA area as they perform oral assessments consisting of a picture description task. The recordings contain a mix of spontaneous and scripted speech. Each child gives one speech sample, 2–3 min in length, totaling approximately 15 h of speech. Although this leads to longer audio segments than those in the adult samples from CORAAL, we observe more similar numbers of words and number clauses between the child and adult samples of these lengths. The children's speech was transcribed by Fisher *et al.* (2019), who are experts in children's language. Authors of this paper then annotated the dialect density of each recording following the same procedure as described in Koenecke *et al.* (2020) for the CORAAL data. All of the students are from the same school district, which primarily serves children of working and lower middle-class families. We acknowledge that socioeconomic status is an important factor in acquisition of dialectal language (Craig and Washington, 1994) and control for it as best as possible with the use of this largely homogeneous dataset.

3. Dialect density labels

Each utterance was transcribed at the word level, and then any documented phonological AAE dialectal differences from MAE (i.e., differences in pronunciation) or morphosyntactic differences (i.e., differences in grammar or word choice) in the utterance were tagged as such. The DDM of each utterance is next calculated as the number of these dialectal differences divided by the number of words in the utterance (Koenecke *et al.*, 2020). For educational applications with AAE children's speech, it may also be useful to predict the child's usage of phonological dialectal patterns and morphosyntactic dialectal patterns separately. Having these two separate metrics (one corresponding to pronunciation and one corresponding to grammar) would allow spoken language systems to give dialect-appropriate feedback on a child's pronunciation, grammar, and word usage separately. To explore a classifier's ability to perform this task for children, we train the classifier to predict the total DDM, the dialect density only taking into account the phonological aspects (Phon DDM), and the dialect density only taking into account the morphosyntactic aspects (Gram DDM) for each model. Similar to the overall DDM, Phon DDM is calculated as the number of phonetic features of AAE in an utterance divided by the number of words in that utterance. We find that calculating a morphosyntactic DDM in the same way often does not produce a metric that aligns well with the raters' perception of which children are low or high density dialect speakers. Therefore, we define morphosyntactic dialect density at the utterance level as performed in Oetting and McDonald (2002). That is, we define the Gram DDM as the percentage of sentences that contain a marker of AAE grammar. Because the number of possible dialectal phonological differences is largely limited by the number of words in an utterance, and the number of dialectal morphosyntactic differences is largely limited by the number of grammar constructions (i.e., clauses), we normalize Phon DDM and Gram DDM by their respective maximum possible values. We evaluate the system performance in predicting Phon DDM and Gram DDM for only the children as the adult speech samples are too short to estimate Gram DDM. The average DDMs for each dataset are given in Table I. To format dialect density estimation as a multiclass classification problem, we then assign discrete levels to the utterances based on their DDMs: 0, dialect density of 0; 1, dialect density between 0 and 0.05; 2, dialect density between 0.05 and 0.1; 3, dialect density between 0.1 and 0.2; and 4, dialect density greater than or equal to 0.2. Each utterance together with its dialect density label then constitutes one training or testing sample. Literature shows that a dialect density greater than 0.1 (i.e., 10% of the individual's words contain a dialectal difference from the mainstream dialect) is often observed as a quite pronounced or high density dialect (Washington and Seidenberg, 2022). The number of utterances at each dialect density for each dataset is shown in Table II. We note that the majority of adult speakers from the CORAAL speakers have DDMs from level 0 to 2, and the majority of child speakers from the GSU Kids speech database have DDMs from level 0 to 1.

B. Features

We extract several feature sets that relate to documented aspects of AAE dialect and then train a backend classifier to predict the dialect density level of a given utterance. Section II B 1 describes the five proposed feature sets and backend model.

1. Grammatical features

AAE has different grammar than MAE. For example, AAE constructions may contain verb conjugations, collocations, or word usages that are not observed in MAE.

TABLE II. Number of utterances in each DDM bin for the CORAAL and GSU Kids speech database datasets.

Label	0	1	2	3	4
Bounds	DDM = 0	$0 < \text{DDM} \leq 0.05$	$0.05 < \text{DDM} \leq 0.1$	$0.1 < \text{DDM} \leq 0.2$	$0.2 < \text{DDM}$
CORAAL DDM	28	49	51	54	26
GSU Kids DDM	68	80	31	17	7
GSU Kids Phon DDM	95	82	16	8	2
GSU Kids Gram DDM	84	14	12	43	50

Motivated by the desire to capture these grammatical aspects of AAE, which have been well-documented in linguistic studies (Lanehart and Malik, 2015), we create a handcrafted feature set composed of the following values to detect the most commonly noticed of these differences (as determined in Craig *et al.*, 2003) in ASR transcripts of spoken AAE.

First, we use the ASR system, HuBERT-base (Hsu *et al.*, 2021), to automatically transcribe each utterance. As an ultimate goal of this work is to perform transcription and dialect density estimation with as little work required from the teacher as possible, we use the ASR transcripts as is without human corrections. Our previous work in Johnson *et al.* (2023a) showed that HuBERT achieved lower average word error rate (WER) than Wav2Vec2 (Baevski *et al.*, 2020) but worse performance than Whisper (Radford *et al.*, 2022). However, Whisper's language modeling often forced the output to align with a language pattern similar to that observed in training, removing AAE constructions from the transcripts. For example, Whisper may interpret the utterance, "We *wasn't* doin' nothin,'" as "We weren't doing nothing," which does not represent the dialectal grammar and pronunciation differences present in the speech sample. Whereas HuBERT gave a higher average WER, we noticed that it represented these differences more faithfully for the higher dialect density speakers. From the HuBERT ASR transcripts, next, we calculated the following quantities intended to capture commonly recognized grammatical traits of AAE:

- *GPT2 sentence perplexity*: We calculated the perplexity of the ASR transcript under the GPT2 language model (Radford *et al.*, 2019). This gives the average negative log likelihood of a sequence of words occurring in their given order (i.e., a value inversely proportional to the likelihood of the sentence being spoken). As GPT2 is likely trained on primarily MAE text, we hypothesize that AAE constructions and ASR errors caused by dialectal differences will give higher perplexity to ASR transcripts from higher density speakers;
- *habitual or future "be" perplexity*: AAE grammar constructions may contain an unconjugated instance "to be," as in "They be crazy out there." We calculate the ratio of perplexity of the original sentence with the perplexity of the sentence, replacing the verb "be" with the contraction of "are" or "is" (e.g., "They're crazy out there."). We similarly calculate the perplexity for the use of "future be" (e.g., "He be here tomorrow.");
- *completetive "done"* (e.g., "They done finished it."): We calculate the ratio of the perplexity of the original utterance to the perplexity of the utterance with the word "done" removed. This value will return "1" if the word "done" does not appear in the ASR transcript, and we choose a backend classifier that can ignore this or other values if they are not informative;
- *simple past "had"* (e.g., "She had went inside" to express the simple past, "She went inside."): We compute the ratio of the perplexity of the original utterance to the perplexity of the utterance with the word "had" removed;
- *subject-verb agreement*: Like the habitual "be," AAE has several grammar constructions that contain subject-verb combinations that do not follow the typical subject-verb agreement patterns of MAE. For example, AAE constructions can include double marking of number and tense (e.g., "he wants to hit*s* them" or "they both fell*ed*"), generalization of "is" and "was" to plural and second person (e.g., "They was from Los Angeles."), and use of a verb stem as past tense (e.g., "They come here yesterday."). To capture these, we use the SpaCy Python library (Honnibal *et al.*, 2020) to automatically apply parts of speech and dependency taggings to the input utterances and then return a binary decision on whether or not a mismatched subject-verb pair (i.e., a plural subject and singular verb or vice versa) was detected. We also apply direct string matching to detect common subject-verb pairs with irregular verbs (e.g., "they was" or "we is");
- *consecutive nouns*: Some AAE constructions, such as absence of possessive "s" (e.g., "That's John house."), absence of plural "s" (e.g., "It's two inch long."), and use of appositive or pleonastic pronouns (e.g., "That girl, she likes chocolate."), can be detected by the presence of consecutive nouns. We use SpaCy part of speech tagging to tag nouns in the ASR transcript and return a binary decision for whether or not consecutive nouns (not including possessives or proper noun phrases) were detected;
- *"ain't" as a preverbal negator*: We return a binary decision on whether or not the word "ain't" is detected in the utterance through string matching on the ASR transcripts;
- *negative concord*: AAE grammar constructions may include double negatives or negative concord (e.g., "They ain't done nothing to nobody."). We use SpaCy part of speech and dependency tagging to automatically detect whether or not a negative verb with a negative object appears in the transcript to return a binary decision for this;
- *existential "it" and "got"*: AAE speakers may use an existential "it" or "got" in place of reference words (e.g., "*it was* a ton of people" or "They got a ton of people." instead of "*there were* a ton of people."). We calculate the ratio of the sentence perplexity of the utterance with the perplexity of the utterance replacing phrases with existential "it" or "got" with the corresponding MAE phrase (e.g., replacing "It was" or "They got" with "There were");
- *indefinite article*: AAE may include invariant use of the indefinite article regardless of the starting sound of the following noun (e.g., saying "a airplane"). We use string matching to determine the presence of the article "a" followed by a word starting with a vowel and return a binary decision for this;
- *irregular participle*: AAE may include using regular verb forms for irregular participles (e.g., "a broke down car" instead of "a broken down car"). We use SpaCy part of speech tagging to identify verbs that modify nouns and

are not in participle form and then return a binary decision for the detection of these; and

- *zero preposition*: Some prepositions are variably included in AAE. Notably, the preposition "of" is often omitted in constructions with the preposition "out" (e.g., "She came out the car."). We use SpaCy part of speech tagging to identify the presence of prepositions after the word "out" and return a binary decision for the detection of these.

2. HuBERT self-supervised learning representations (SSLR)

As noted in Yang *et al.* (2021), the HuBERT SSLR have proven to be useful for a variety of speech tasks. Here, we apply them to train a classifier to predict dialect density. For each utterance, first, we extract the hidden state from the last layer of HuBERT. Then, wedivide the $1024 \times N$ output SSLR (where N is the number of 20 ms frames in the audio signal) into segments of five frames (corresponding to 100 ms of the audio signal). These 100 ms segments are compiled with a sliding window with a shift of 20 ms, which means that there is overlap between adjacent segments. We compute the average of each 5-frame segment and use these 1024×1 vectors to train a K-nearest neighbor (Knn) classifier to predict the dialect density level of a new input averaged segment of HuBERT SSLR during inference. These 1024×1 vectors are extracted from all 100 ms frames of every training utterance and given the dialect density label of the utterance from which they came for training the Knn classifier. Tuning on the validation set showed that the best K for the Knn classifier was 90. After training the Knn classifier on the frames of the training set, next, we similarly extracted the HuBERT SSLR from the test set, averaged over each 100 ms segment, computed the Knn prediction for each segment, and computed the percentage of frames assigned to each of the five dialect density levels. The soft label 5×1 vector, containing the percentages of frames at each dialect density level, is then used as an input to the backend classifier for final dialect density level prediction. As HuBERT is trained with an unsupervised clustering step, we hypothesize that its SSLR will be useful in a downstream dialect-related task using clustering.

3. ASR phoneme-level features

For this feature set, we use the Wav2Vec2-Phoneme model first (Xu *et al.*, 2022) to transcribe each utterance at a phoneme level. Wav2Vec2-Phoneme has a total of 391 different possible phoneme outputs. Validation on the CORAAL adult speech training set, which holds out CORAAL DCB as the test set, showed that only 38 of these phonemes were present in the dataset and, therefore, we restricted the output of the system to only consider those 38 for all experiments. Then, we compute the frequency of each phoneme and bigram frequency of each phoneme pair normalized by the number of phonemes in the utterance. This created a 38-dim feature vector for the unigram phoneme frequency and a 1444-dim feature vector (i.e., 38^2) for the bigram phoneme frequency counts. We note that the majority of entries in the bigram feature vector were zero as many phonemes would not typically occur next to each other in a given order. A vector containing these counts for each phoneme or phoneme combination is then used as an input to the backend classifier.

4. OpenSmile features

The OpenSmile feature set (Eyben *et al.*, 2010), which extracts paralinguistic features relating to speaker pitch, voice quality, spectral shape, MFCCs, and other factors has proven to be effective in low-resource DID in multiple studies (Johnson *et al.*, 2022a; Tzudir *et al.*, 2022b). Here, we investigate the performance Geneva minimalistic acoustic parameter set (GeMAPS; Eyben *et al.*, 2016) feature set of the OpenSmile features in dialect density classification. We elect to use the smaller GeMAPS v01a feature set instead of the larger ComparE 2016 feature set (62 vs 6373 features, respectively) as we wish to use the feature set primarily to investigate the prosodic information contained in the utterance, which can be achieved through the use of the low level descriptors (LLDs) and their statistical functionals available in GeMAPS. Although the DDMs used in this paper are calculated without respect to prosodic markers, previous work shows that prosodic markers of dialect often cooccur with phonological and grammatical markers of dialect and are used by human listeners to discern dialect, as shown with AAE in Holliday (2021). Whereas the LLDs of the GeMAPS set are available in the ComparE set as well, the large number of features contained in the overall feature set compared to the size of the available dataset might cause the classifier to overfit and, thus, we opt against using the full ComparE 2016 feature set.

5. X-vector speaker embeddings

Originally proposed as a feature for speaker identification, X-vectors are the output of a later hidden layer of a time delay neural network trained for speaker discrimination (Snyder *et al.*, 2018). These features have proven to be useful in DID (Johnson *et al.*, 2022a; Liao *et al.*, 2023). Here, we use them as a feature to train the backend system to learn dialect density. From each utterance, we extract the 512-dimensional X-vector using the Kaldi toolkit (Povey *et al.*, 2011). We also perform a comparison of these embeddings with the more recent ECAPA-TDNN X-vectors (Desplanques *et al.*, 2020).

C. Model

After extracting features, we use an XGBoost model (Chen and Guestrin, 2016) to map the input features to a discrete dialect density level. XGBoost is an ensemble method which iteratively trains decision trees to perform classification, adding new trees to the ensemble to compensate for the errors of the previous tree in each iteration. These models perform well in classification tasks that rely on fusing information from different feature sets and have proven useful in dialect density estimation in our previous work (Johnson *et al.*, 2022a). These models also offer much more explainability than deep neural networks as the impact of each feature used in decision can be explored through SHAP value analysis (Lundberg and Lee, 2017). That is, we can calculate a measure of feature importance for each input feature in the five-class dialect density level classification problem.

D. Prediction tasks

We perform three sets of experiments to validate our proposed system.

Task 1: Individual feature performance. We use the features described in Sec. II C as the input to the XGBoost model with the goal of predicting the speaker's DDM from one of five discrete levels. We, first, test the performance of each feature individually in predicting the overall DDM for adults and children and the Phon DDM and Gram DDM for children.

Task 2: Combined feature performance. Given the performance of the individual features in predicting the DDM classes, we next use a concatenation of the features in the model to perform the five-class dialect density level classification.

Task 3: Binary thresholding. We acknowledge that choosing boundaries for each dialect density level requires domain knowledge which may not exist for every dialect or accent. Therefore, the multi-class classification method that we present is less reproducible for some low-resource dialects. As an alternative, we also perform the experiment as a binary classification task. In this experiment, we choose a threshold and train the classifier to predict whether or not the DDM for each test sample is less than or equal to that threshold. Then, we shift the threshold across the range of DDMs for the test set.

For the adult speech, where data is labeled for different dialect regions, we consider two train/test configurations: cross-region and multi-region. In the *cross-region* case, we train the system on four regions and test on the held-out region, rotating over all regions. This scenario is designed to show the performance of the system with no training data from the same region as the test set. In the *multi-region* case, we randomly hold out 20% of the full CORAAL data set for testing and train on the remaining data, repeating the experiment five times and reporting the average performance. Because the children's speech data all comes from a single region, we perform a fivefold validation experiment and present the average results.

E. Comparison with our previous work

We make several modifications to our previous framework for dialect density estimation in Johnson *et al.* (2022a) in accordance with new developments in speech and language processing. First, we previously noted that sentence perplexity calculated with a long short term memory–based language model was an effective feature in estimating dialect density. With the increasing effectiveness of GPT-based language models, we instead try a perplexity feature calculated with the most recent open-source GPT model (GPT-2 at time of writing). We also implement more granular handcrafted features to target specific grammatical patterns that may affect perplexity for greater interpretability. Next, we add SSLR from HuBERT here as they have recently been shown to be effective in a variety of speech tasks (Yang *et al.*, 2021). In addition, our previous work with the OpenSmile feature set of over 6000 features showed that several of the most impactful features from the set related to voice quality and prosody. We opt to use the more compact GeMAPS feature set from OpenSmile because it contains features relating to the most useful features of our previous work and reduces the chance of overfitting. We, again, use the *X*-vector speaker embedding in this work. Previously, we trained a neural network to predict a speaker's regional accent (using the speaker's city of origin as a label) from the input *X*-vectors extracted from non-dialect density-labeled speech CORAAL. The output softmax probability from that system was then used as a feature in dialect density estimation. We have since found that some region's recordings in CORAAL are highly separable by recording quality and channel effects and, hence, we use instead the raw *X*-vector as a feature here. Last, we used correlation between the sets of predicted and actual DDM labels in our previous work. In this work, we format the problem as a classification problem for greater interpretability of the machine performance on individual samples.

III. RESULTS

Because this is the first reported effort on automatic dialect density prediction, the results for all three tasks are reported in comparison to the accuracy associated with predicting the most frequent class in the training data, i.e., the prediction based only on class priors. The *training prior* condition represents an uninformed baseline; model accuracy below this baseline reflects over-fitting. A low training prior result indicates train/test mismatch in the class distributions for the cross-region scenario.

Table III shows the five-class dialect density level classification accuracy for task 1, where an XGBoost model is trained separately on each of the specific feature types described in Sec. II. We show the performance of the models trained separately for adults (cross-region and multi-region scenarios) and children. The DCB set is used in this exploratory work for the cross-region scenario because it has the median dialect density of the CORAAL database. With the exception of the ECAPA-TDNN *X*-vector, all features provide benefit over the uninformed training prior baseline for the adult conditions. For children, as discussed further in Sec. IV, grammar features are only informative for the Gram-DDM score, and most of the acoustic features are uninformative for the Gram-DDM score. The experiments showed that the Wav2Vec2-Phoneme Bigram features and ECAPA-TDNN *X*-vector feature perform substantially worse than their related counterparts, the

TABLE III. DDM classification accuracy of XGBoost classifier trained on each individual feature set (task 1) for adults (cross-region DCB and multi-region) and children with results for the training prior maximum (Tr-prior) for reference. The accuracy is shown with the overall dialect density for adults and children. In addition, for children, results are given for the dialect density taking into account only phonological characteristics of dialect (Phon DDM) and the dialect density taking into account only grammatical characteristics of dialect (Gram DDM).

Feature set	Feature dimension	CORAAL adults (test DCB, train other)	CORAAL adults (20% RHO)	GSU Kids (fivefold validation)		
Metric		Overall DDM	Overall DDM	Overall DDM	Phon DDM	Gram DDM
Grammar feature	13	32.0%	47.6%	37.7%	25.1%	56.4%
HuBERT SSLR-knn	5	40%	45.2%	56.2%	51.3%	35.8%
Wav2Vec2-Phoneme unigram	38	44.0%	52.4%	52.1%	54.5%	51.3%
Wav2Vec2-Phoneme bigram	1444	40.0%	51.1%	48.9%	46.6%	36.4%
OpenSmile GeMAPS	62	36.0%	41.7%	48.2%	56.4%	43.6%
Kaldi X-vector	512	34.0%	48.2%	44.0%	53.6%	33.3%
ECAPA-TDNN X-vector	128	16.0%	37.8%	42.8%	55.2%	29.2%
Tr-prior		24.0%	26.0%	39.4%	43.2%	41.3%

Wav2Vec2-Phoneme unigram feature and Kaldi *X*-vector feature, respectively. Therefore, these features are dropped in subsequent experiments.

Table IV shows the classification accuracy for task 2, where we concatenate the features and train a single model to perform the dialect density level estimation. The average cross-region (avg CR) and average random hold out (RHO) results are not directly comparable because of random sampling, but the performance difference is substantial in that it is roughly double the standard deviation of the RHO results. In all cases, the model substantially outperforms the uninformed training prior baseline and the results for all individual features as expected.

Figure 1 presents the result of the binary DDM classification experiment (task 3) for the different regions of the adult speech (cross-region) and children's speech. For each test set, we compute the accuracy of the system in predicting whether or not the speaker of a given sample had a DDM above a series of different thresholds. The corresponding plots show the difference in model prediction accuracy relative to the uninformed training prior baseline. Small values (positive or negative) indicate that performance is not significantly different from the training prior, i.e., the features are not informative, which will be the case for thresholds where one class has few examples. Larger negative values reflect overtraining, which is generally associated with a mismatch in the binary class distribution between training and testing.

IV. DISCUSSION

In this section, we analyze the experimental results. The Knn-generated soft labels, using the frames of the HuBERT SSLR and the Wav2Vec2-Phoneme unigram model, classify dialect density level best for the children's and adults' speech. It is worth noting that there are typically more phonological than morphosyntactic aspects of AAE dialects in a speaker's speech because a sentence can have several words containing pronunciation differences but will often only have one subject-verb structure that can be modified. Therefore, the overall DDM is often dominated by the Phon DDM term, and features that capture acoustic differences in pronunciation like the HuBERT SSLR and the Wav2Vec2-Phoneme outputs appear best for predicting the overall DDM. However, these features do not appear to capture grammatical features of AAE dialect well. The handcrafted grammar features and *X*-vector features perform best for predicting DDM-gram for the adults. Although the *X*-vector features are derived for speaker identification and not semantic tasks, the TDNN used to extract the feature pools information over several time windows, capturing segment-level information. This segment-level information is likely more useful in categorizing a speaker's likelihood of speaking with a morphosyntactic dialectal difference than features that operate at the frame-level only (e.g., Wav2Vec2 or HuBERT features). Although the handcrafted grammatical features perform well for adult speech, their performance degrades for the children's speech. This is likely the result

TABLE IV. Performance of the XGBoost model trained on the combined feature set (task 2; excluding the Wav2Vec2-Phoneme bigram and ECAPA-TDNN *X*-vector features). For reference, we show performance associated with the training prior maximum (Tr-prior) for each test set. Results are reported for the overall DDM score for cross-region (CR) and multi-region conditions for the adult AAE speech in CORAAL. For children's speech, cross-validation results are reported for DDM, Phon DDM, and Gram DDM.

	CORAAL adults							GSU Kids fivefold validation		
	Overall DDM							DDM type		
Model	ROC	LES	DCB	PRV	VLD	Average CR	Average RHO	Overall	Phon	Gram
Tr-prior	18.0%	5.0%	24.0%	2.0%	28.6%	15.5%	26.0%	39.4%	43.2%	41.3%
XGboost	46.0%	56.7%	48.0%	48.0%	32.9%	46.3%	60.1%	73.8%	61.2%	59.0%

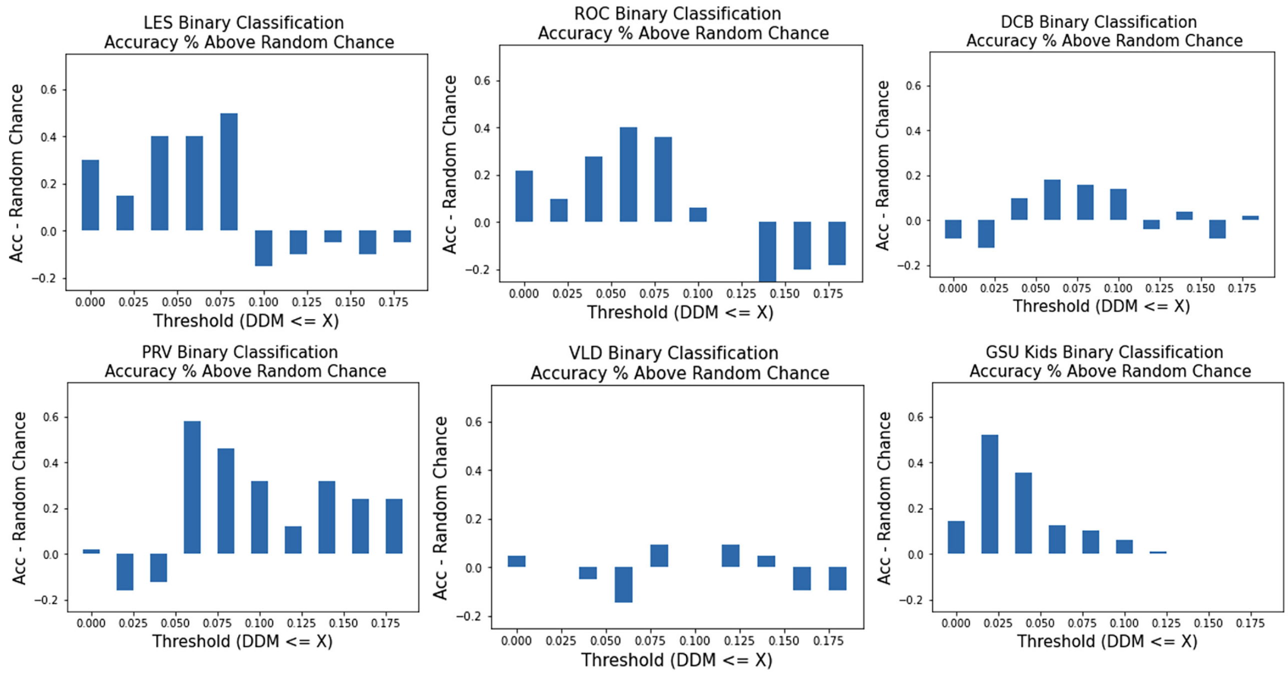

FIG. 1. (Color online) Performance of the task 3 binary model in predicting whether or not a speech sample displayed an overall DDM higher than a given threshold. Each plot shows the difference in the classification accuracy relative to performance associated with the training prior decision vs the DDM threshold. The plots for adults correspond to the five cross-region systems, and the plot for GSU Kids speech database shows the fivefold CV average for overall DDM.

of the higher number of ASR transcription errors in children's speech, which may prevent the downstream NLP algorithm from accurately matching grammatical patterns. The *X*-vector feature, again, performs well for classifying the number of morphosyntactic dialectal differences in the children. We also note that the OpenSmile prosodic features are useful for this as some grammatical patterns may typically co-occur with specific intonation or changes in pitch, making the prosody a good indicator of grammatical differences. Whereas the adults in the study each display one of five different regional dialects, all of the children are from the same school district, making them more likely to share prosodic and dialectal grammar patterns that generalize better across the training and testing sets.

In the combined model, we dropped the worse performing Wav2Vec-Phoneme bigram and *X*-vector features. Although studies have shown that bigram features typically outperform unigram features, the smaller size of the data used in this work may be insufficient to adequately train a model using bigram features, which are much higher dimension than the unigram features. We also examine the performance of the speaker embeddings in this task. The 512-dimensional Kaldi *X*-vectors outperformed the 128-dimensional ECAPA-TDNN *X*-vectors. This may indicate that the more compressed ECAPA-TDNN *X*-vectors contain only more identity focused information, whereas the larger Kaldi *X*-vector feature retains more information on dialect.

The model trained on the combined feature set outperforms all models trained on individual feature sets for CORAAL DCB and performs well across the other test sets. We note that the model trained on the other four sets and tested on CORAAL VLD has the lowest dialect density level prediction performance. This is likely due to the fact that the speakers from Valdosta display some aspects of southern American dialect that are not observed in the other datasets and, thus, are difficult for the model to learn. Particularly, AAE speakers from North Carolina and Georgia have been shown to exhibit vowel shifts more in line with those observed in Southern American English, whereas AAE speakers from Washington, DC and New York often display vowel shifts that are more unique to AAE (Thomas, 2001; Yaeger-Dror and Thomas, 2010). The model performed best for CORAAL LES out of the adult datasets as the LES dialect has been influenced by speakers of several other regions, and training on speech from other areas will likely generalize better to speakers from there than to a more isolated area. The work in Koenecke *et al.* (2020) also demonstrates that commonly used ASR systems have shown better ASR WER for the northern AAE dialects than the southern AAE dialects and, therefore, a higher number of ASR errors in the VLD Wav2Vec2-Phoneme features and grammatical feature input transcripts may have caused worse prediction accuracy.

Figure 2 shows the bee swarm plot depicting which features were most used in predicting dialect density level for the adult's speech when testing on CORAAL DCB. Each line shows how separable each utterance was by the feature displayed on the left. We note that the Knn soft labels generated from the HuBERT SSLR were most often used in the classification (where knn_0, knn_1, knn_2, knn_3, and knn_4 denote the Knn soft labels for dialect density level in ascending order). The unsupervised pre-training on a large amount of

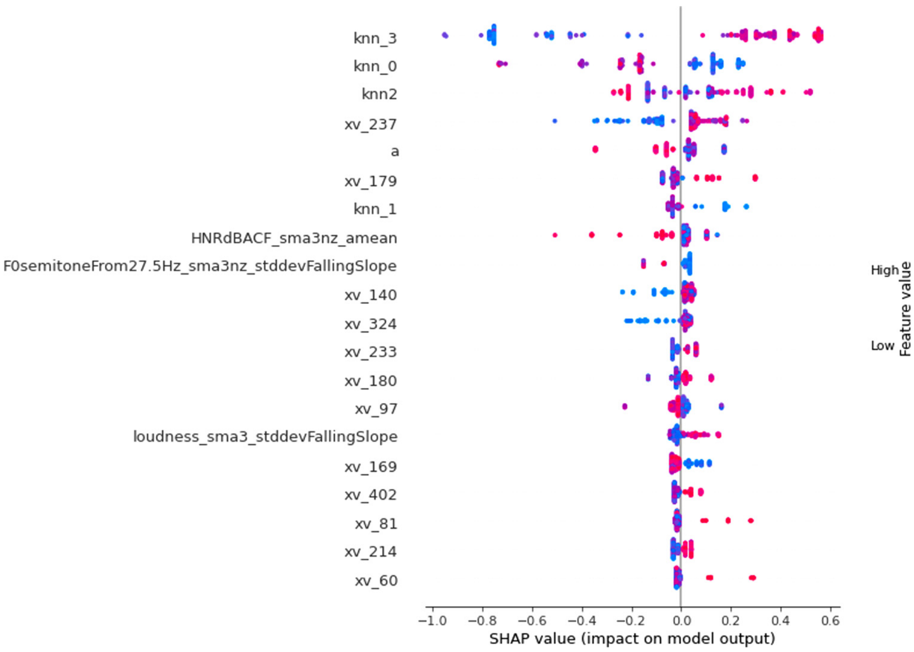

FIG. 2. (Color online) Bee swarm plot depicting the relative impact of features in the ensemble classifier for the adult speech from CORAAL.

data appears to have made the HuBERT SSLR especially potent in capturing small acoustic differences relating to pronunciation and regional phonological varieties. Then, we see that several components of the X-vector feature (e.g., *xvec*_237) were effective in distinguishing dialect density level, capturing shared traits across speakers at the segment level. From the Wav2Vec2-Phoneme model, it appears that a higher number of detections of the vowel \a\ (shown as "a" in the bee swarm plot) correlated with a lower predicted dialect density. This is consistent with documented phenomena in which vowel formant frequencies shift between MAE and southern American dialects, including varieties of AAE, which may result in alternate pronunciations of some vowel sounds and cause the model trained on MAE to recognize them as other sounds (Johnson *et al.*, 2022b; Lanehart and Malik, 2015). Last, several of the OpenSmile features, such as the harmonic-to-noise ratio (HNR), autocorrelation function, standard deviation of the $F0$ semitone, and standard deviation of the slope of the loudness, were also often used by the decision trees of the ensemble classifier. HNR has been shown to be useful in distinguishing several speaker characteristics such as age and speaking style (Ferrand, 2002) while changes in $F0$ and loudness over time may be indicative of the presence of dialect-specific prosodic patterns.

The model trained on the combined features performs well for the children's speech (Fig. 3). The model achieves over 70% classification accuracy in the five-class dialect density level prediction task. One reason why this model performs better for the children's speech than for the adults' speech may be that the children in the test and train splits are from the same geographic area, whereas the adult models are trained on speech from other regional AAE variants. Another reason is that although the recordings of the children performing the picture description educational assessment are unscripted, the children are all performing the same assessment and are likely to share some of the same vocabulary and grammar while performing it. This may make it easier for the model to analyze shared traits across content that are not available across the completely spontaneous interview speech in CORAAL. However, the high variability in children's speech still presents challenges for the model.

It can be observed that the overall DDM prediction accuracy (DDM acc.) is sometimes lower than the DDM considering only phonological differences (Phon DDM acc.)

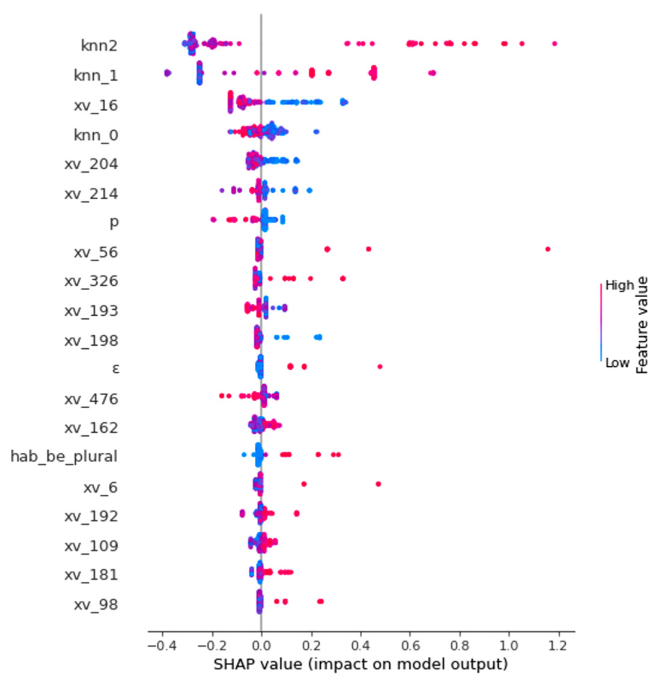

FIG. 3. (Color online) Bee swarm plot depicting the relative impact of features in the ensemble classifier for the child speech from the GSU Kids' speech database.

or the DDM considering only grammatical differences (Gram DDM acc.). This may indicate that for some regional variations of AAE or age-specific ways of speaking, the grammatical and phonological language characteristics are not strongly correlated. It then becomes a more complex problem for the classifier to jointly identify the presence of both of these types of linguistic tokens in an input utterance. As a result, the performance of the joint prediction may be worse than the individual phonological or grammatical DDM prediction. For these cases, we may either explore increasing the complexity of the classifier or simply creating a weighted sum of the individual predictions in the future.

The model trained to perform binary classification with thresholded DDMs as opposed to multi-class classification generally performed well across the choice of threshold for the test sets. For all CORAAL test sets except VLD, the greatest benefit of the classifier over the training prior for a DDM threshold is 0.05 and 0.1. This is likely because most regions had a larger number of samples in this range and, hence, the classifier was able to better learn to distinguish DDMs from the input data. For the children's speech, we see that the classifier accuracy is mostly above training prior baseline when the DDM threshold is in the lower range. This follows logically as many of the children's speech samples have lower dialect density and, therefore, the classification problem becomes easier as the threshold rises to the point where most samples in the test and training sets will have lower DDMs than the threshold. Therefore, the training prior baseline performance will be much higher at the higher DDM thresholds for this case. Overall, these experiments give insight on which DDM thresholds the classifier performs best at given the variation across regions and currently available training data. From this, we observe a tendency for the classifier to become more accurate as more data in the target dialect density range is added, pointing to a possibility for the classifier to become much stronger with additional training data.

As expected, the multi-region scenario outperforms the cross-region scenario (Table IV) because of the reduced mismatch in the train/test distributions. We realize that the recordings taken from the same region (i.e., recorded by the same interviewer) may share some recording conditions or channel effects that can be used to form spurious correlations with the speaker's dialect density. However, the Wav2Vec-Phoneme and grammar features do not pass information on the background conditions or channel effects to the backend classifier. Therefore, background effects that may be common across multiple recordings of the same region cannot be used to indirectly learn dialect density classification. Because these features perform similarly to those that are more subject to background noise and channel effects (e.g., OpenSmile GeMAPS features and Kaldi *X*-vector feature), we believe that feature correlations with characteristics of the audio that are not directly related to speaker and dialectal qualities are minimal.

We note a few comparisons with our prior work in Johnson *et al*. (2023b). Previously, we found character-level perplexity of the transcripts to be a useful feature in dialect density estimation. However, in this work, we do not see the word level perplexity from GPT2 used by the classifier as often as several of the other features available. As Holtzman *et al*. (2019) points out, large language models like GPT2 may learn a bias for longer sentences and repetitive grammar structures when training on large text corpora, which means that their prediction of likelihood of words occurring in a sequence does not generalize well to spontaneous spoken speech. Our previous work also found the frequency of several sounds in the transcripts to be good indicators of the dialect spoken. This is especially true for vowels that may undergo a formant shift or consonants that are more often dropped or de-emphasized in different dialects. We noticed a similar trend for a few vowels and consonants for the adults' and children's speech. The addition of the Knn soft labels in this work seems to improve performance over our previous results, and the features are relatively robust for adult and child speech.

V. CONCLUSIONS

This work shows promising progress in automatically detecting dialect density levels of speakers across age and regional dialect. Given the limited size of the datasets, we achieve reasonably high dialect density level classification accuracy over the adults' speech (often ranging from 10% to 40% above the uninformed max training prior baseline) and over 70% accuracy for children. We demonstrate the utility of HuBERT self-supervised representations, prosodic features from OpenSmile, handcrafted grammatical features, speaker embeddings, and phoneme-level transcripts in the prediction task. The feature sets provided may be adapted for use in several other language and DID tasks, and the framework presented offers explainability for which speech features capture dialectal differences that are useful for automatic classification. We anticipate that additional training data would lead to improved results with high enough fidelity for real-time classroom use. This study also highlights the degree of dialectal speaker variability within and across regions and how spoken language systems should be adapted to handle them. Our future work includes using dialect density predictions in downstream tasks such as bias mitigation in language technology, fair educational speech technologies that provide dialect-appropriate automatic feedback to spoken responses in oral assessments, and applying this framework to other dialects.

ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation (NSF). The GSU data were collected with support by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health (NIH) under Grant No. P01HD070837.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

https://doi.org/10.1121/10.0025771

DATA AVAILABILITY

The data that support the findings of this study are openly available in the CORAAL database at https://doi.org/10.7264/1ad5-6t35. The GSU Kid's database is available upon reasonable request from the authors of Fisher *et al.* (2019) at GSU. These data are not publicly available due to privacy concerns for the sensitive population that it was sampled from (i.e., children).

[1]Data are available at https://drive.google.com/drive/folders/1g4ypxQB_fYaOCuMXW_vAUtWLhqFZ-h1f?usp=sharing,codetobereleaseduponacceptanceofthispaper (Last viewed December 20, 2023).

Ali, A., Shon, S., Samih, Y., Mubarak, H., Abdelali, A., Glass, J., Renals, S., and Choukri, K. (**2019**). "The MGB-5 challenge: Recognition and dialect identification of dialectal Arabic speech," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Singapore (14–18 December) (IEEE, New York), pp. 1026–1033.

AWS (**2023**). "What is Amazon transcribe?," Amazon Web Services, available at https://docs.aws.amazon.com/transcribe/latest/dg/what-is.html (Last viewed December 20, 2023).

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (**2020**). "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Curran Associates, Inc., San Diego, CA), pp. 12449–12460, Vol. 33, available at https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf (Last viewed December 20, 2023).

Chen, T., and Guestrin, C., (**2016**). "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

Craig, H. K., Thompson, C. A., Washington, J. A., and Potter, S. L. (**2003**). "Phonological features of child African American English," Am. Speech Hear. Assoc. **46**(3), 623–635.

Craig, H. K., and Washington, J. A. (**1994**). "The complex syntax skills of poor, urban, African-American preschoolers at school entry," Lang. Speech Hear. Serv. Sch. **25**(3), 181–190.

Desplanques, B., Thienpondt, J., and Demuynck, K. (**2020**). "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech 2020*, pp. 3830–3834.

Duroselle, R., Sahidullah, M., Jouvet, D., and Illina, I. (**2021**). "Modeling and training strategies for language recognition systems," in *Proceedings of Interspeech 2021*, pp. 1494–1498.

Edwards, J., Gross, M., Chen, J., MacDonald, M. C., Kaplan, D., Brown, M., and Seidenberg, M. S. (**2014**). "Dialect awareness and lexical comprehension of mainstream American English in African American English–speaking children," J. Speech. Lang. Hear. Res. **57**(5), 1883–1895.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., and Truong, K. P. (**2016**). "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," IEEE Trans. Affective Comput. **7**(2), 190–202.

Eyben, F., Wöllmer, M., and Schuller, B. (**2010**). "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th Association for Computing Machinery International Conference on Multimedia*, pp. 1459–1462.

Ferrand, C. T. (**2002**). "Harmonics-to-noise ratio: An index of vocal aging," J. Voice **16**(4), 480–487.

Fisher, E. L., Barton-Hulsey, A., Walters, C., Sevcik, R. A., and Morris, R. (**2019**). "Executive functioning and narrative language in children with dyslexia," Am. J. Speech. Lang. Pathol. **28**(3), 1127–1138.

Google (**2023**). "Read along by google," available at https://play.google.com/store/apps/details?id=com.google.android.apps.seekh&hl=en_US&gl=US (Last viewed December 20, 2023).

Holliday, N. R. (**2021**). "Perception in black and white: Effects of intonational variables and filtering conditions on sociolinguistic judgments with implications for ASR," Front. Artif. Intell. **4**, 642783.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (**2019**). "The curious case of neural text degeneration," in *International Conference on Learning Representations*.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (**2020**). "spaCy: Industrial-strength natural language processing in Python."

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (**2021**). "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 3451–3460.

Johnson, A., Everson, K., Ravi, V., Gladney, A., Ostendorf, M., and Alwan, A. (**2022a**). "Automatic dialect density estimation for African American English," in *Proceedings of Interspeech 2022*, pp. 1283–1287.

Johnson, A., Fan, R., Morris, R., and Alwan, A. (**2022b**). "LPC augment: An LPC-based ASR data augmentation algorithm for low and zero-resource children's dialects," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore (23–27 May) (IEEE, New York), pp. 8577–8581.

Johnson, A., Shetty, V. M., Ostendorf, M., and Alwan, A. (**2023a**). "Leveraging multiple sources in automatic African American English dialect detection for adults and children," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece (4–10 June) (IEEE, New York), pp. 1–5.

Johnson, A., Veeramani, H., Natarajan, B., and Alwan, A. (**2023b**). "An equitable framework for automatically assessing children's oral narrative language abilities," *Proceedings of Interspeech*.

Kendall, T., and Farrington, C. (**2021**). "The Corpus of Regional African American Language. version 2021.07," available at http://oraal.uoregon.edu/coraal (Last viewed December 20, 2023).

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., and Goel, S. (**2020**). "Racial disparities in automated speech recognition," Proc. Natl. Acad. Sci. U.S.A. **117**(14), 7684–7689.

Labov, W. (**2006**). *The Social Stratification of English in New York City* (Cambridge University Press, Cambridge, UK).

Lanehart, S., and Malik, A. M. (**2015**). "Language use in African American communities: An introduction," in *The Oxford Handbook of African American Language*, edited by J. Bloomquist, L. J. Green, and S. L. Lanehart (Oxford University Press, Oxford, UK).

Lee, S., Potamianos, A., and Narayanan, S. (**1999**). "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," J. Acoust. Soc. Am. **105**(3), 1455–1468.

Lei, Y., and Hansen, J. H. L. (**2011**). "Dialect classification via text-independent training and testing for Arabic, Spanish, and Chinese," IEEE Trans. Audio. Speech Lang. Process. **19**(1), 85–96.

Liao, C., Huang, J., Yuan, H., Yao, P., Tan, J., Zhang, D., Deng, F., Wang, X., and Song, C. (**2023**). "Dynamic TF-TDNN: Dynamic time delay neural network based on temporal-frequency attention for dialect recognition," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes Island, Greece (4–10 June) (IEEE, New York), pp. 1–5.

Lundberg, S. M., and Lee, S.-I. (**2017**). "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., Long Beach, CA), Vol. 30, pp. 4765–4774, available at http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf (Last viewed December 20, 2023).

Martin-Jones, M. (**1995**). "Code-switching in the classroom: Two decades of research," in *One Speaker, Two Languages: Cross-Disciplinary Perspectives on Code-Switching* (Cambridge University Press, Cambridge, UK), pp. 90–111.

Mawadda Warohma, A., Kurniasari, P., Dwijayanti, S., Irmawan, Y., and Suprapto, B. (**2018**). "Identification of regional dialects using mel frequency cepstral coefficients (mfccs) and neural network," in *2018 International Seminar on Application for Technology of Information and Communication*, Semarang, Indonesia (21–22 Septmeber) (IEEE, New York), pp. 522–527.

Moyle, M. J., Heilmann, J. J., and Finneran, D. A. (**2014**). "The role of dialect density in nonword repetition performance: An examination with at-risk African American preschool children," Clin. Linguist. Phonet. **28**(9), 682–696.

Oetting, J. B., and McDonald, J. L. (**2002**). "Methods for characterizing participants' nonmainstream dialect use in child language research," J. Speech. Lang. Hear. Res. **45**(3), 505–518.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P *et al.*, (**2011**). "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* (IEEE Signal Processing Society, Piscataway, NJ).

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., and Auli, M. (**2024**). "Scaling speech technology to 1,000+ languages," J. Mach. Learn. Res. **25**(97), 1–52, available at http://jmlr.org/papers/v25/23-1318.html.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (**2022**). "Robust speech recognition via large-scale weak supervision," arXiv:2212.04356.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (**2019**). "Language models are unsupervised multitask learners," OpenAI blog **1**(8), 9.

Seymour, H. N., Bland-Stewart, L., and Green, L. J. (**1998**). "Difference versus deficit in child African American English," Lang. Speech Hear. Serv. Sch. **29**(2), 96–108.

Shon, S., Ali, A., and Glass, J. (**2019**). "Domain attentive fusion for end-to-end dialect identification with unknown target domain," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*, Brighton, UK (12–17 May) (IEEE, New York), pp. 5951–5955.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (**2018**). "*X*-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada (15–20 April) (IEEE, New York), pp. 5329–5333.

Thomas, E. R. (**2001**). *An Acoustic Analysis of Vowel Variation in New World English* (Duke University Press, Durham, NC), Vol. 85.

Thomas, E. R. (**2015**). "Prosodic features of African American English," in *The Oxford Handbook of African American Language*, edited by J. Bloomquist, L. J. Green, and S. L. Lanehart (Oxford University Press, Oxford, UK).

Tzudir, M., Baghel, S., Sarmah, P., and Prasanna, S. R. M. (**2022a**). "Under-resourced dialect identification in Ao using source information," J. Acoust. Soc. Am. **152**(3), 1755–1766.

Tzudir, M., Sarmah, P., and Prasanna, S. (**2022b**). "Prosodic information in dialect identification of a tonal language: The case of Ao," in *Interspeech*, pp. 2238–2242.

Van Hofwegen, J., and Wolfram, W. (**2010**). "Coming of age in African American English: A longitudinal study," J. Sociolinguist. **14**(4), 427–455.

Washington, J. A., Branum-Martin, L., Sun, C., and Lee-James, R. (**2018**). "The impact of dialect density on the growth of language and reading in African American children," Lang. Speech Hear. Serv. Sch. **49**(2), 232–247.

Washington, J. A., Craig, H. K., and Kushmaul, A. J. (**1998**). "Variable use of African American English across two language sampling contexts," J. Speech. Lang. Hear. Res. **41**(5), 1115–1124.

Washington, J. A., and Seidenberg, M. S. (**2022**). "Language and dialect of African American children," in *Handbook of Literacy in Diglossia and in Dialectal Contexts: Psycholinguistic, Neurolinguistic, and Educational Perspectives* (Springer, New York), pp. 11–32.

Xu, Q., Baevski, A., and Auli, M. (**2022**). "Simple and effective zero-shot cross-lingual phoneme recognition," *Proc. Interspeech*.

Yadavalli, A., Mirishkar, G., and Vuppala, A. K. (**2022**). "Multi-task end-to-end model for Telugu dialect and speech recognition," in *Proc. Interspeech*, pp. 1387–1391.

Yaeger-Dror, M., and Thomas, E. R. (**2010**). *African American English Speakers and Their Participation in Local Sound Changes: A Comparative Study* (Duke University Press, Durham, NC).

Yang, S., Chi, P., Chuang, Y., Lai, C. J., Lakhotia, K., Lin, Y. Y., Liu, A. T., Shi, J., Chang, X., Lin, G., Huang, T., Tseng, W., Lee, K., Liu, D., Huang, Z., Dong, S., Li, S., Watanabe, S., Mohamed, A., and Lee, H. (**2021**). "SUPERB: Speech processing universal performance benchmark," arXiv:2105.01051.

Zissman, M. A., and Berkling, K. M. (**2001**). "Automatic language identification," Speech Commun. **35**(1), 115–124.