Title

Towards Inclusive Low-Resource Speech Technologies: A Case Study of Educational Systems for African American English-Speaking Children

Author

Johnson, Alexander

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Towards Inclusive Low-Resource Speech Technologies: A Case Study of Educational Systems for African American English-Speaking Children

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Electrical and Computer Engineering

by

Alexander Johnson

2024

ABSTRACT OF THE DISSERTATION

Towards Inclusive Low-Resource Speech Technologies:
A Case Study of Educational Systems for African American English-Speaking Children

by

Alexander Johnson
Doctor of Philosophy in Electrical and Computer Engineering
University of California, Los Angeles, 2024
Professor Abeer A. Alwan, Chair

The potential of speech technology to improve educational outcomes has been a topic of great interest in recent years. For example, automatic speech recognition (ASR) systems could be employed to provide kindergarten-aged children with real-time feedback on their literacy and pronunciation as they practice reading aloud. Within these systems, speaker identification (SID) technology could additionally be used to identify the user's speaker characteristics in order to ensure that they receive age, language, and dialect-appropriate feedback. While these technologies are more established for well-represented groups in STEM (ie. able-bodied, adult, first-language speakers of mainstream dialects), they give much worse performance for underrepresented groups (young children, speakers of non-mainstream dialects, people with speech-related disabilities, etc.). This work focuses on improving speech technology performance for children's speech and African American English (AAE) dialect speech with the goal of creating more equitable outcomes in early education. The contributions of this work span three primary areas: 1) Dialect identification and density scoring, 2) data augmentation for speech recognition, and 3) Natural Language Processing for fair and

inclusive automatic speech assessment.

First, we create a robust system for dialect identification of African American English for both children and adult's speech. This system aims to take an input utterance from a speaker of either African American English or Mainstream American English and determine which of the two dialects the utterance belongs. The system fuses features from paralinguistics, self-supervised learning representations, automatic speech recognition system outputs, prosodic contours, and other descriptors of the speech signal in order to learn a mapping from the input acoustic information to a dialect classification decision. We further explore this architecture in automatic dialect density estimation, a task we create and develop. In dialect density scoring, we train a system to automatically predict a speaker's frequency of usage of dialect-specific patterns. This information can then be passed to a speech recognition system for more dialect-informed processing.

Second, we develop a data augmentation algorithm to improve zero-shot and few-shot speech recognition of low-resource dialects. The algorithm, named LPCAugment, deconstructs an input speech signal into a source and filter representation using linear predictive coding (LPC) analysis. The poles of the filter representation can then be perturbed independently of the source representation in order to model formant shifts that may be seen across accents and dialects. We use this perturbation method to artificially generate speech samples with shifted formant locations to serve as additional training data for a speech recognition system. This speech recognition system is then evaluated on children's speech for child speakers of a Southern California dialect and child speakers of an Atlanta, Georgia, area dialect.

Third, we explore automatic analysis and scoring of speech recognition transcripts for educational assessments. Given information about a student's spoken dialect and automatically generated transcripts of their oral response to an assessment prompt, we train a system to automatically grade the quality of the response with respect to a pre-determined criterion. This system uses language modeling and spoken information retrieval to iden-

tify key features in the spoken response and holistically decide if the response aligns with the grading criteria. Combined, the steps in this work form a framework for inclusive spoken language understanding technology that can be used to perform provide students with dialect-appropriate language training or language assessment.

The dissertation of Alexander Johnson is approved.

Lin Yang

Safiya U. Noble

Achuta Kadambi

Abeer A. Alwan, Committee Chair

University of California, Los Angeles

2024

To all the people who brought me free food while I was a broke grad student

TABLE OF CONTENTS

1 Introduction **1**

1.1 Motivation 1

1.2 Transformer-based Speech and Language Systems 2

1.3 Underrepresented Voices in Speech Technology 4

1.3.1 African American English 5

1.3.2 Children's Speech 7

1.4 Automatic Dialect Identification 8

1.5 Automatic Speech Recognition 11

1.5.1 Data Augmentation 12

1.5.2 Self-supervised Pre-training 12

1.6 Spoken Language Understanding for Education 14

1.6.1 Automatic Oral Assessment Scoring 15

1.6.2 Spoken Question Answering 16

1.7 Outline of the Dissertation 17

2 Datasets **19**

2.1 CORAAL 19

2.1.1 CORAAL QA 20

2.2 GSU Kids Speech Corpus 21

2.3 UCLA JIBO Kids speech 21

3 Dialect Identification and Dialect Density Scoring **23**

3.1 Dialect Density Estimation 23
3.1.1 Data Annotation 24
3.1.2 Methods 24
3.1.3 Feature Sets 25
3.1.4 Experiments 29
3.1.5 Results and Discussion 29
3.2 DID 33
3.2.1 Data Curation 34
3.2.2 Models 38
3.2.3 Experiments 41
3.2.4 Results and Discussion 42
3.3 Summary 45

4 Data Augmentation for Low-Resource ASR 46
4.1 Methods 47
4.1.1 The LPC Augment Algorithm 47
4.1.2 Model Training 48
4.2 Experiments and Results 50
4.2.1 Optimizing the Warping Factor 50
4.2.2 Zero Resource Scenario 51
4.2.3 Low Resource Scenario 52
4.3 Discussion 52
4.4 Summary 53

5 Fair and Inclusive Automatic Oral Assessment Scoring 57

5.1 Educational Task 57

5.1.1 TNL - Story Retelling Task 57

5.1.2 TNL - Picture Description Task 58

5.2 Experiment 59

5.2.1 Results and Discussion - TNL Story Retelling 61

5.2.2 Results and Discussion - TNL Picture Description Task 67

5.3 Summary 68

6 Inclusive Automatic Spoken Question Answering 71

6.1 Methods 72

6.1.1 Model 73

6.1.2 Experiments 74

6.2 Results 77

6.3 Discussion 77

6.4 Summary 79

7 Summary and Conclusions 83

7.1 Summary 83

7.2 Novel Contributions 85

7.3 Ethics Statement 85

7.4 Future Work 86

References 87

LIST OF FIGURES

1.1 The transformer architecture consisting of the input layer, encoder stack, and decoder stack. 4

1.2 The Wav2Vec2 architecture, demonstrating how self-supervision is used to train the encoder layer to create a robust speech representation for downstream ASR or other tasks. 14

3.1 The architecture for the fully connected (FC) network used to project the X-vectors (left) and the CNN used to project the prosodic information (right). The inputs to the CNN are the pitch (F0) and three energy contours of the utterance. The output of both networks is a vector whose elements represent the probability of the speaker belonging to each of the cities used from the CORAAL database. 28

3.2 Overview of the features used in the proposed dialect density estimation proposed framework. 30

3.3 SHAP value plot for the XGBoost model trained to predict DDMphon from the set of all features. The features are listed from top to bottom in order of significance. 34

3.4 SHAP value plot for the XGBoost model trained to predict DDMgram from the set of all features. The features are listed from top to bottom in order of significance. 35

3.5 The feature set and backend models used in the proposed dialect identification scheme. 38

4.1 An example of the LPC spectra of a child in the UCLA JIBO kid's Database pronouncing the phoneme \AA\, and the result of perturbing it with LPC Augment. In the perturbed signal, the first two formant peaks have been shifted to the left, and the third has been shifted to the right. 49

4.2 Diagram of the LPC Augment Algorithm 54

5.1 Overview of a) the string-search rubric-based approach and b) the neural linguistic feature-based approach. 62

5.2 Semantic Similarity between each student's ASR and ground truth transcript. ASR transcripts generated with Whisper, Hubert, and Hubert fine-tuned on MyST. 66

6.1 Overview of the proposed system. The long audio file, D is segmented into one minute segments, s_i. Each segment is then transcribed with ASR where the ASR system is prompted with previous context. Then both the ASR transcript from each segment and the text of an input query, q, are encoded with Sentence-BERT and scored for the likelihood that s_i answers q by the PLDA classifier. Last, the ground truth scores are used to evaluate performance. 80

LIST OF TABLES

3.1 Average dialect density by city for each of the dialect density measures shown. . 24

3.2 Pearson Correlation between actual and predicted dialect density measures for each of the three metrics: only the phonological component of the dialect density (DDMphon), only the morphosyntactic component of the dialect density measure (DDMgram), and the entire dialect density measure (DDM). The results for the model trained on six feature sets individually as well as the model trained on the combination of all of the features are shown. 31

3.3 Average Pearson Correlation between actual and predicted dialect density measures for each of the three DDMs over 200 iterations of Random Hold Out validation. 31

3.4 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. 37

3.5 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. . . . 42

3.6 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. 43

4.1 Results of the recognition experiment (in %WER) on the validation set with the proposed method for different warping factors using the transformer model. CA Val denotes the performance of the system trained with data augmentation on the speech data containing dialects found in Georgia and validated on speech containing dialects found in California. GA val similarly denotes the performance of the system trained with data augmentation on the speech data collected in California and validated on the speech data collected in Georgia. The lowest WER for each case is shown in boldface. 51

4.2 Comparison of common speech data augmentation methods with the proposed method. Each model (Transformer and HMM-DNN) is trained on either the California English training set (Train CA) or the Georgia English training set (Train GA) and then evaluated on both the California English test set (CA test) and the Georgia English test set (GA Test). Columns representing zero-resource scenarios (where the model is trained on only one dialect and tested on the other) are highlighted. The lowest word error rate for each case is shown in boldface. . 55

4.3 Results of the models trained on both Train CA and Train GA and tested on CA Test and GA Test with the proposed and other data augmentation methods. The lowest WER for each case is shown in boldface. 56

5.1 ASR Word Error Rate (WER) , Classification Accuracy (C. Acc), and classification RMSE for the fuzzy string matching approach for each system 63

5.2 The classification metrics (C. Accuracy, F1-score, and RMSE) of each of the fine-tuned language models considered when predicting scores. Numbers reported are the average of 5 trials of random hold out. 64

5.3 System performance using a backend classifier to predict assessment scores from an input concatenation of hand-crafted linguistic features and soft labels from the best-performing large language model (BERT) extracted from the best ASR transcripts (Whisper). Backend classifiers tested are: Support Vector Machines (SVM), Logistic Regression (LogReg), Random Forest (RandFor), and XGBoost. 65

5.4 Results for both the SSRM and NLF approaches across different student demographics. We present a breakdown of best performing ASR system (Whisper) word error rate, the classification C. Accuracy and RMSE of the system on the ground truth (GT) transcripts, and those metrics on the Whisper ASR transcripts for the following three demographic splits: 1) Type of Reading or Language Impairment from i) control - no impairment, ii) RD Only- student has reading disorder like dyslexia that does not occur with or as a secondary effect of a primary learning or language impairment or other condition, iii) RD + LI - A reading disorder that occurs with a primary Language impairment 2) Reading status from i) Poor - the student is evaluated to read at a level below their appropriate grade level or ii) Good - the student reads at or above their appropriate grade level, and 3) Dialect from i) African American English (AAE) or ii) Non-AAE - a mix of characteristics of General American English and Southern American English native to the Atlanta, Georgia Area. Note that the number of students in the Reading/Language Impairment and Reading Status demographic categories do not sum to the full 184. For this analysis, we excluded children with other disorders like ADHD that may complicate the test taking and children who were not able to be assessed for reading status into either the Poor or Good category. . . 69

5.5 Percent Classification Accuracy (C. Acc), Percent F1 Score, and Root Mean Square Error of each language model in predicting student scores from the input transcripts (ground truth, Whisper ASR transcript, or HuBERT asr transcripts) along with the word error rate (WER) for each. 70

6.1 Effect of the size of the Audio Segments for predictions from the PLDA model. Precision, Recall and Macro F1 statistics are calculated from predicted scores from the system. EER refers to the Equal Error Rate of the trained system . . . 77

6.2 Metrics for evaluating the quality of generated questions: Cosine distance between BERT embeddings of the generated questions and hand-written questions (Semantic Similarity), Percent Words shared between the generated questions and hand-written questions, BLEU score between the generated questions and hand-written questions, % of named entities from the ground truth transcript not included in the generated question (GT % entities included), % of named entities from the ASR transcript included in the generated question (Whisper % entities included)), Precision, Recall, and F1-score between the retrieved answer to the hand-written question and that for the generated question (Answer precision, recall, and F1-score), language model accuracy in correctly answering the question from a multiple choice set with distractors generated from the ground truth transcript (GT Distractor Acc) and distractors generated from the ASR transcript (Whisper Distractor Acc), and the Answerabile score given by Self-CheckGPT for the generated question with either the ground truth transcript or the ASR transcript given as context (GT Answerable Score and Whisper Answerable score, respectively). 81

6.3 Performance of PLDA systems trained with questions generated by different systems. Questions were generated by the respective systems from the non-prompted Whisper generated ASR transcripts. All refers to a PLDA model trained by combining the questions generated from all the individual models 82

6.4 Performance of the question answering model when the ASR transcripts from the previous N segments are used in the Whisper prompt as previous context on transcribing the current audio segment. 82

ACKNOWLEDGMENTS

This work would not have been possible without the support of several people and organizations. I would first like to thank my advisor, Prof. Abeer Alwan, and my lab mates at the UCLA Speech Processing and Auditory Perception Laboratory for guiding and assisting me throughout the PhD program. I would also like to thank our collaborators including Prof. Alison Bailey at the UCLA School of Education, Prof. Mari Ostendorf at the University of Washington, Prof. Robin Morris at Georgia State University, Prof. Julie Washington at the University of California, Irvine, and their students. This work was supported in part by the NSF, The UCLA Amazon Science Hub, and the Ralph J. Bunche Center for African American Studies at UCLA. The GSU Kids' speech data was collected with support by the Eunice Kennedy Shriver National Institute of Child Health Human Development of the NIH under Grant P01HD070837.

VITA

2018	Bachelor of Science, Electrical Engineering, Northwestern University
2018-2022	Eugene V. Cota-Robles Fellow
2019	Masters of Science, Electrical Engineering, UCLA
2018-2024	PhD Student, UCLA, Department of Electrical and Computer Engineering
2019-2023	Teaching Assistant for undergraduate and graduate classes in ECE and Neuroscience, UCLA
2019	Intern, Qualcomm
2020	Intern, 3M M*Modal
2021	Intern, Mathworks
2021	UCLA Rising to the Challenge-Graduate Summer Research Fellowship
2022	UCLA Collegium of University Teaching Fellow, Course Instructor
2022, 2023	Intern, J.P. Morgan Chase

PUBLICATIONS

N. B. Shankar, A. Johnson, C. Chance, and H. Veeramani, "CORAAL QA: A Dataset and Framework for Open Domain Spontaneous Speech Question Answering from Long Audio

Files," in ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

A. Johnson, H. Veeramani, N. B. Shankar, and A. Alwan, "An Equitable Framework for Automatically Assessing Children's Oral Narrative Language Abilities," in Interspeech 2023, 4608-4612, doi: 10.21437/Interspeech.2023-1257

A. Johnson, V. Shetty, M. Ostendorf, and A. Alwan, "Towards Automatic Dialect Identification of African American English in Adults' and Children's Speech," in ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10096614

A. Johnson, J. Washington, R. Morris, M. Ostendorf, A. Bailey, and A. Alwan, "Towards Effective Speech-based AI in the Classroom: The Case of AAE-Speaking Children" in Black in AI Workshop at NeurIPs 2023

A. Johnson, K. Everson, V. Ravi, A. Gladney, M. Ostendorf, and A. Alwan, "Automatic Dialect Density Estimation for African American English," in Interspeech 2022, 1283-1287, doi: 10.21437/Interspeech.2022-796

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," in ICASSP 2022, doi:https://doi.org/10.1109/ICASSP43922.2022.9746281, page 8577-8581

A. Johnson, A. Martin, M. Quintero, A. Bailey, and A. Alwan, "Can Social Robots Effectively Elicit Curiosity in STEM Topics from K-1 Students During Oral Assessments?" in IEEE EDUCON, 2022, pp. 1264-1268, doi: 10.1109/EDUCON52537.2022.9766662

CHAPTER 1

Introduction

1.1 Motivation

Artificial intelligence (AI) has revolutionized practices in finance, defense, and entertainment. However, the education sector has significantly lagged behind in adopting machine learning-based technologies in teaching. The deployment of AI technologies in schools could greatly alleviate labor shortages and high work loads among educators. A 2022 technical report found that the United States had a large shortage of teachers with 36,000 vacancies and 163,000 under-qualified individuals in teaching roles nationwide [1]. This shortage is even more present in specialized education roles (working in dual-language immersion programs, working with students with special needs, etc.), as illustrated by reports that 54% of speech language pathologists in schools lack the personnel to adequately perform their duties [2]. Here, voice-based AI technology, such as automatic speech recognition (ASR) or spoken language understanding (SLU) systems, could be used to lead students through educational exercises, perform oral assessments, and screen for language difficulties in situations where there is not enough staff to effectively do so for all students.

One reason for many educators' hesitancy to use AI with their students stems from AI researchers' failure to prove the efficacy and fairness of the technology. For example, work in [3] shows that many commercial ASR systems perform worse for speakers of African American English (AAE) than for speakers of Mainstream American English (MAE). In addition, studies have also shown that ASR systems designed to recognize adult speech

perform much worse for younger children [4, 5]. These deficiencies and inequities make current AI-based speech technologies unsuitable to handle the needs of classrooms consisting of diverse groups of students. In fact, the inability of many speech technologies designed for majority groups to generalize their performances to lower resource cases (eg. speakers of low resource languages and dialects or children's speech) remains a pressing problem in the speech field. This is due in large part to the fact that many data-driven speech and language systems are trained only on the most easily-accessible data, meaning that speech from under-represented groups in technology are often excluded from the training process. Solutions to this problem include the creation of new speech and language datasets which include speech from typically marginalized linguistic groups, the creation of SLU system architectures that do not rely solely on the availability of large datasets to train the model, and the adaptation of existing systems for low-resource cases. This dissertation presents a framework for more equitable training of speech and language models to perform well across speaker age, dialect, and style. We focus here on achieving fair performance in educational applications of spoken language technology for classrooms with speakers of African American English-speaking students, as AAE is a large, underrepresented dialect in the United states. The rest of this section provides background information on prior work in the fields of speech technology and linguistics on which the rest of the dissertation builds.

1.2 Transformer-based Speech and Language Systems

Many current speech and language systems utilize the transformer architecture [6]. This architecture was proposed as an efficient solution to sequence to sequence problems, or problems in which an input sequence (eg. a sequence of words in a sentence, a sequence of frames in an audio signal, or a sequence of images in a video) is mapped to an output sequence (eg. a sequence of words in a different language as in machine translation, a sequence of words corresponding to the input audio as in automatic speech recognition, or a sequence of object

labels as in video tagging). A key feature of transformers is the attention module. For each token in the input sequence, self-attention seeks to calculate three numerical representations (a key, a value, and a query) such that a mathematical combination of those representations between any two tokens will represent how related they are. For example, in the sentence "I gave his wallet to him," the words "his" and "him" reference the same object, and so the combination of their keys, values, and queries should give a high number to represent that these words are highly related. Likewise, the words "I" and "gave" are not significantly related in meaning and thus should produce a lower number when their keys, values, and queries are combined. Similar to self-attention, cross attention seeks to calculate keys, values, and queries between an input sequence and an output sequence that represent how related tokens in one sequence are to the other. For example, in a machine translation task which translates the English sentence "I need to buy fruit," to equivalent Spanish sentence "Necesito comprar fruta," the system should calculate a key, value, and query for each word so that the combination of numerical representations for the equivalent words "fruit" and "fruta" is higher than that for less related words. Transformers are composed of two portions: an encoder stack and a decoder stack. In the encoder stack, neural network layers with self-attention are used to map an input sequence to a high-level representation. Then the decoder uses cross-attention to map this high-level representation to the output sequence. A diagram of the transformer architecture is shown in Figure 1.1 Variations of this architecture have been widely successful in many speech and language tasks. For example, the transformer-based language model BERT (Bi-directional Encoder Representation from Transformers) [7] has set a benchmark in text classification, and the Generative Pretrained Transformers (GPT) [8] series of language models have set the state-of-the-art performance in generative language models. The current leading ASR models such as Wav2Vec2 [9] and Whisper [10] also heavily utilize the transformer architecture.

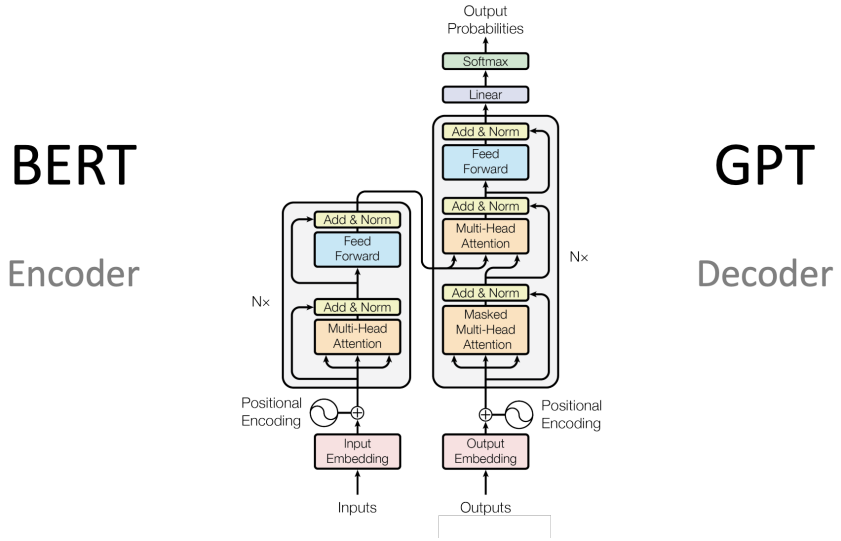

Figure 1.1: The transformer architecture consisting of the input layer, encoder stack, and decoder stack.

1.3 Underrepresented Voices in Speech Technology

Transformer-based speech systems often use a large number of trainable parameters (e.g 1.6 billion parameters in Whisper-Large [10]). In order to train a large neural network, the system in turn requires large amounts of labeled training data. While this requirement does not typically present challenges to well-resourced linguistic groups with a large digital footprint, such large amounts of training data are not always available for speakers of under-resourced languages dialects or speakers whose speech characteristics are not well-represented online (young children, speakers with speech related-disabilities, etc.). This means that researchers will not be able to train these systems on large amounts of data from speech of people of these minority groups in the way that they can for majority groups, and the system will consequently give lower performance for the minority groups. Therefore, further work is needed to bridge the performance gap between majority and minority users of these data-driven speech technologies. In the scope of inclusive educational technology, this paper focuses on speech-language systems for two groups of underrepresented speakers in technology: Speakers of

the African American English dialect and child speakers.

1.3.1 African American English

AAE is one of the most studied varieties of American English for both child and adult speakers [11, 12]. While AAE can display several regional and generational differences, many characteristics of AAE are common across most or all of the variants. Many scholars point to the origin and evolution of AAE over time as the reason for the shared traits between these variants [13]. The first large populations of Black people in the United States were enslaved people who were brought to the US South in the 1600's. There, AAE and White Southern American English (SAE) grew out of British colonial dialects and heavily influenced each other as they developed. While origins of specific AAE variants are often subjects of debate, many scholars agree that several defining features of AAE were created during this time and persisted through the Great Migration of African Americans throughout other regions of the US [13]. Despite the historical and linguistic work done to better understand AAE, the dialect is still understudied in the area of Spoken Language Systems. AAE impacts all domains of American English (AE), but most significantly presents differences in rules for production of the morphology , syntax, phonology, and prosody from those of Mainstream American English (MAE).

AAE Phonology: A widely recognizable feature of AAE is it's collection of phonological differences, often expressed as differences in pronunciation of specific sounds or words, from MAE. For example, AAE speakers can display word-initial labiodentalization of dental fricatives (e.g. pronuncing "this" as "dis"), word final g-dropping (e.g. pronouncing "nothing" as "nothin"), and word final r-lessnes (e.g pronouncing "four" as "fou") [14]. While many of these phonological patterns are not strictly unique to AAE (e.g. g-dropping has become common in SAE and other dialects), they are more likely to appear in many AAE speakers than in MAE speakers. Some phonological features of AAE also vary with region, such as a regional vowel shift or lacktherof in a region where other speakers display a vowel

shift [15].

AAE Morphosyntax: Morphosyntax, which encompasses features of grammar, word choice, and word usage, is perhaps the most widely used studied aspect of the AAE dialect. Grammatical features of AAE such as zero copula (e.g stating "they are rich" as "they rich"), negative concord (e.g. "They ain't never got no money"), and preterite "had" (e.g. stating "she had went to the store" to express the simple past "she went to the store") are well-documented, and the evolution of their usage since the early 1900's has been a popular subject of study [16]. Interestingly, it has also been shown that the number of many morphosyntactic AAE features used by AAE-speaking children declines as the children get older [17]. This may be due to increased exposure to MAE with age or as a result of how the US educational system structures its lesson instruction.

AAE Prosody: Prosody, which describes usage of pitch, intonation and rhythm in speech, is perhaps one of the most difficult aspects of speech to document. Prosodic features can occur both within words and across longer segments, change based on a speaker's style or intended subject of emphasis. Due to the complex nature of prosodic patterns, a trained linguist is often needed to annotate the speech sample accurately. In general, linguistic work in prosody has progressed less quickly than that in other areas such as phonology and syntax. In studying AAE, this is no exception. Many scholars agree that AAE has unique prosodic features that are distinct from other dialects [18]. However, it is difficult to document exact rises and falls in pitch or segment-level intonations that would make an utterance sound like an AAE construction. The relatively small body of well-agreed upon work on AAE prosody makes the field an area in need of future analysis.

AAE Dialect Density AAE occurs on a continuum of low to high density usage. Children who are high density users of AAE tend to be those who are growing up in poverty [19, 20]. Factors such as isolation and widespread school segregation likely influence the density of language variation in these low-income speakers [21]. In addition, children in the Southern United States have been documented to use particularly high levels of dialect

overall, as regional variation is also prevalent, and combines with AAE to result in oral language that differs significantly from the language of print [22, 23].

Speakers of African American English often face bias, being perceived as less educated or professional than speakers of MAE. For example, typically-developing child speakers of AAE are often under-rated in language exams and put in special needs classes at significantly higher rates than their MAE-speaking counterparts [24, 19]. AAE speakers with more exposure to dialects outside of their own often learn to code-switch or translanguage, meaning that they incorporate varying amounts of AAE and MAE dialectal characteristics into their speech depending on the situation [25]. Low-income speakers typically do not get the same number and quality of opportunities to hear and learn to use different types language, either inside or outside of school, meaning that they often do not learn to code switch [21]. These children in particular are at high risk for receiving assessment scores that do not reflect their actual abilities and, subsequently, inadequate education.

Dialect density, or the frequency of one's use of dialect-specific linguistic patterns, is a commonly used term in discussing the effects of dialect in the classroom. The dialect dialect density measure (DDM) of an utterance can be calculated as the number of dialect-specific phonological and morpho-syntactic tokens in an utterance divided by the number of words in that utterance. Of particular interest to the field of educational speech technology is the fact that speakers who speak with higher AAE dialect density (i.e. their utterances have a relatively high average DDM) have been shown to face more educational disparities than their counterparts who speak with lower dialect density [19].

1.3.2 Children's Speech

There are several well-known differences between children's speech and adults' speech that cause performance discrepancies when neural network-based ASR systems trained only on adults' speech are tested on children's speech [26]. First, young children have yet to master the more complex articulatory gestures needed to produce conventional or adult-like English

speech sounds [27]. This means that children's speech often contains several production errors (substituted speech sounds for example) and variability in how they produce speech sounds. In fact, there is higher inter- and intra-speaker variability in child speech when compared to adults [28]. Second, the frequency range of children's voices is much higher than that of adults, making them less compatible with systems trained on adult's speech [28]. Third, ASR systems trained to recognize the words of an adult's vocabulary will likely have bias towards interpreting the child's words as ones more commonly used by adults [29]. Therefore, in order to more effectively create an ASR system for children, that system should be trained using child-specific speech data. While some children's speech corpora do exist, they are not nearly as plentiful as those for adults. In addition, the available children's speech datasets often do not contain dialectal or sociolinguistic information on the participants, making it difficult to ensure fair performance across diverse child speakers.

1.4 Automatic Dialect Identification

Language identification (LID) and dialect identification (DID) are the processes of automatically identifying a speaker's spoken language and dialect from a short input utterance. LID and DID identification have become integral parts of many large spoken language systems. For example, many multilingual automatic speech recognition (ASR) systems like OpenAI's Whisper [10] and Meta's Massively Multilingual Speech models [30] leverage large cross-lingual speech corpora for training and then perform LID during inference. Other systems like AWS transcribe [31] offer DID for commercial use cases, distinguishing input speech, for example, between English variants from the US, UK, 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 how 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 frequency of use of dialect-specific pronunciations, grammar patterns, and prosodic patterns may lead to inaccurate representations of some speakers in downstream applications.

Several recent studies have offered promising DID systems for a limited number of dialects. [32] introduces a time delay neural network, as popularized by the X-vector speaker embedding [33], with attention across both time and frequency for classifying between a set of 16 dialects. The experiments performed in [34] additionally 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. The authors of [35] designed a multi-task learning framework for a conformer-based system that jointly learns to output ASR transcripts and DID labels for speech from three Telegu dialects. In order to overcome performance degradation caused by domain mismatch in end-to-end DID systems, [36] creates 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 both within and across groups. For example, while many current paradigms may categorize US English as distinct from British English, they do not recognize differences between Mainstream American English (MAE), African American English (AAE), Southern American English, Creole English, and other variants. The work in [37] 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, meaning 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 [38], or incorporate aspects of multiple dialects into their speech. 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 Mel frequency cepstral coefficients which mainly discern differences in pronunciation (e.g. [39, 40, 41]). However, dialects are a multi-faceted aspect of language which can differ in prosody, grammar, and diction in addition to pronunciation. Previous works which have combined prosodic cues with spectral information [42], or that have attempted to classify language or dialect from grammatical features of text [43] have shown that considering other aspects of language can improve automatic DID. This is especially beneficial in DID for speakers with relatively high acoustic variability like children. Although children's developing vocal tracts and articulatory motor skills may cause their speech to display different acoustic properties than adults [44], work in [45] shows that incorporating prosodic and grammar information into DID systems trained on adults speech can make them more robust for children.

Improving DID for children's speech is of particular interest in educational speech technology. Applications like Read Along by Google [46] use ASR and natural language processing (NLP) to recognize and provide pronunciation and literacy feedback to children as they practice reading aloud. As education literature has demonstrated, speakers of minority dialects like AAE are often underrated in language abilities due to raters who are unfamiliar with AAE interpreting dialectal differences as language deficiencies [19]. In particular, children with higher AAE dialect density have been shown to underachieve in schools that primarily

teach in MAE [19]. Children's DID in educational spoken language systems could be used to detect and mitigate this bias.

1.5 Automatic Speech Recognition

Automatic Speech Recognition (ASR) is the task of generating a text transcript of the spoken words contained within an audio recording. Many traditional ASR systems operate in the following way[47]: First, a training corpus of matched audio data and corresponding text transcripts is compiled. From each audio file, a frame-level acoustic representation of speech is extracted. These representations may be calculated deterministically, as with Mel Frequency Cepstral Coefficients [48], or learned by the input layers of a neural network. These frame-level features are then passed to an acoustic model which aims to predict which speech sound, if any, was spoken in each frame of the input audio data [49]. As the output of the acoustic model is a sequence of speech sounds that has the same length as the number of input audio frames, a lexical model is then tasked with mapping this sequence of speech sounds to the intended words that most likely produced them. Finally, a language model incorporates grammatical and semantic knowledge to determine the most likely sentence spoken by the speaker that would produce the given output of the lexical model [50]. During training, the ASR system is trained to learn the optimal parameters that will produce the closest transcript to the ground truth transcript. Many current end-to-end models seek to combine the acoustic model, lexical model, and language model into one step which is learned by a large neural network. These end-to-end models have been shown to achieve high performance when trained on large datasets whose speaker distribution matches that of the evaluation set. For example, OpenAI's Whisper [10] achieves high state-of-the-art after training on 680,000 hours of audio data scraped from the internet. However, these end-to-end systems often experience large degradations in output quality when tested on out-of-domain data. That is, these systems perform worse for speakers whose linguistic

group or speech patterns were not well-represented in the training data. Therefore, Whisper and other widely-used ASR systems typically show less efficacy when transcribing speech from young children, speakers of African American English, and speakers with other accents and dialects. In order to overcome these and other effects of domain mismatch, researchers have proposed several training strategies such as data augmentation and self-supervised pre-training.

1.5.1 Data Augmentation

Many ASR systems have been shown to perform better when trained with more audio data. Therefore, researchers have sought out methods to cheaply generate more training data from the existing datasets. Data augmentation methods seek to create artificial training data containing deviations from the original samples in order to make the model less sensitive to expected variations. For example, vocal tract length perturbation (VTLP) [51] applies a piece-wise mapping to the frequency axis of input spectral features of a speech signal in order to simulate speech having come from a speaker with a different vocal tract length. Speed perturbation [52] speeds up or slows down portions of an audio signal to simulate having speech of different speaking rates. SpecAugment [53] masks out and re-scales portions of an input spectrogram to simulate speech time-frequency components that are missing or changed from the original recording. By training a speech recognition system to process the augmented speech, the system learns to transcribe the speech in an invariant manner with respect to target characteristics such as speaking rate or vocal tract length.

1.5.2 Self-supervised Pre-training

Thus far, we have described ASR systems as trained with supervised learning. Supervised learning means that the system is given audio data and corresponding human-labeled text transcripts at the time of training and then tasked with finding an optimal mapping from

the audio data to the transcripts. This process requires that all audio recordings used have corresponding human-written labels in the form of transcripts. However, there is much more audio data available than just what has been curated by transcribers. For example, websites like YouTube contain many hours of audio recordings that do not contain human-written text transcripts. This begs the question of whether or not non-labeled data can be leveraged for training an ASR system. This task of leveraging unlabeled data is commonly referred to as unsupervised learning if no label is used during training, or self-supervised learning if a training task in which the system learns to predict an attribute of the data itself is created. One of the most popular architectures for unsupervised or self-supervised learning is Meta's Wav2Vec2 system [9]. Before performing supervised training on the matched audio-text pairs in the training data, Wav2Vec2 first attempts to learn information about the structure of the audio data through a self-supervised pre-training step. Ideally, after the system completes the pre-training step, it will learn a better neural network weight initialization from which to start supervised training. At the input, Wav2Vec2 uses a convolutional neural network to extract acoustic features from the speech signal. Some of these features are intentionally masked out or removed from the system. These features are then fed to a BERT (Bidirectional Encoder Representation from Transformers) [7]. The BERT encoder is then tasked with learning a feature representation from which the missing information from the masked out frames can be interpolated. After the system has been trained to do this, the output features of the BERT encoder can be readily applied to a downstream speech recognition model or another task. A diagram of the Wav2Vec architecture is shown in Figure 1.2. This self-supervised pre-training task has been shown to significantly improve model performance. Other architectures like HuBERT [54] have improved on Wav2Vec2's design for increased accuracy. HuBERT uses much of the same architecture as Wav2Vec2. However, one notable difference is that the system is also tasked with learning to cluster similar acoustic features in order to discover hidden units which may correspond to sounds or characteristics of the language. Notably, the features output by the BERT encoder in

HuBERT have been shown to be useful for a variety of speech tasks outside of ASR such as emotion recognition, keyword spotting, and automatic speaker verification [55].

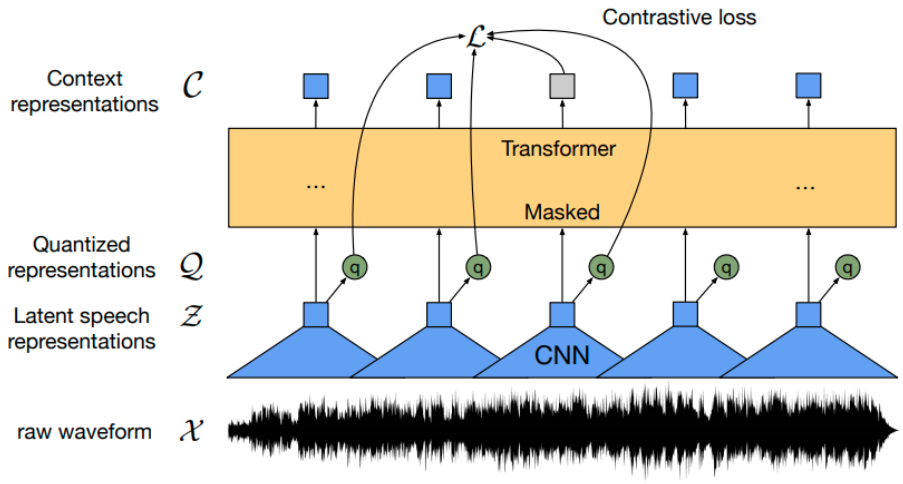

Figure 1.2: The Wav2Vec2 architecture, demonstrating how self-supervision is used to train the encoder layer to create a robust speech representation for downstream ASR or other tasks.

1.6 Spoken Language Understanding for Education

After transcribing a student's spoken response to a question, we may then want to offer feedback on completeness, complexity, or overall quality of their answer. For this, we design natural language processing (NLP) pipelines to extract information from or categorize the ASR transcripts of student responses. Given a text representation of the student's answer, we can build on common methods in NLP for education to adapt them for spontaneous speech and speech diverse dialects. Here, we build on methods in automatic assessment scoring and question answering.

1.6.1 Automatic Oral Assessment Scoring

Automatic Oral Assessment scoring seeks to train a machine to give a score to a student's verbal response to a prompt such that the score has high agreement with that assigned by a human rater. In recent years, great strides have been made to automate spoken language assessments (SLAs) that measure fluency and goodness of pronunciation. For example, [56] explores multitask learning as an approach to overcoming the problem of limited data in automatic oral English proficiency SLAs for Mandarin speakers. In addition, [57] compares the performance of Wav2Vec2.0 [58] and Kaldi TDNN-based [59] grapheme embeddings as features for evaluating children's phonological working memory for nonwords. Similarly, the authors of [60] use hidden states from Wav2Vec2.0 [58] to predict mispronunciations and abnormalities in children's speech. Such methods that take advantage of large pre-trained automatic speech recognition (ASR) systems seem particularly promising given the recent advancements in training strategies for architectures like HuBERT [54], WavLM [61], and Whisper [10]. However, challenges remain in automatic SLA, especially for children. Children's developing language skills and growing speech articulators cause their speech to be highly variable [44], which in turn creates challenges in recognition and assessment [62, 63].

In order to assess language abilities such as grammar usage, coherence, and reasoning, NLP systems that infer over longer contexts are necessary. This task has been explored in tasks such as automatic essay scoring. Studies in essay scoring have used natural language understanding (NLU) to score written essays for narrative language proficiency [64, 65]. Notably, [66] combines hand-crafted linguistic features which capture advanced semantics with soft label predictions from the language model, RoBERTa [67], in a hybrid model which achieves state-of-the-art-performance readability score classification (i.e. classifying the complexity and depth of an essay).

However, further work is needed to adapt these state-of-the-art essay scoring systems to spoken language.

1.6.2 Spoken Question Answering

In addition to scoring the overall quality of an oral response, a rater may also want to extract specific pieces of information from the answer in order to give targeted feedback on the completeness and correctness of certain sections of the response. For example, if a student is asked to verbally describe a person's appearance, the rater may consider the description incomplete (i.e. deserving of a lower score) if it does not mention the person's hair color. Using automatic question answering, an NLU system could query the student's response for any mention of hair color to determine the student's score in that area. Recent advancements in BERT [7] and GPT [68]-based language models have revolutionized performance in question answering and information retrieval tasks on text. Now, a desirable outcome is to replicate the performance of these systems in the speech domain. That is, given a set of audio recordings and a user's input query for information, we seek to return audio recordings or spans that are relevant to the query. Successful architectures for this task typically take one of two frameworks: a cascade system or an end-to-end model. A cascade system first uses automatic speech recognition (ASR) to transcribe a spoken document and then passes that transcript to a downstream language model for text-based question answering. End-to-end systems seek to bypass the need for transcription and answer a question directly from audio features. Notable cascade models include [69] which introduces a self-supervised dialogue learning framework from conversational question answering and [70] which proposes a unified pipeline for multiple spoken language understanding tasks. End-to-end spoken question answering models of interest include SpeechBERT [71], which jointly encodes audio and text information for downstream spoken question answering, GhostT5 [72] which extracts and passes a lightweight speech feature representation to a pre-trained language model to answer questions from speech without the need for complete automatic speech recognition (ASR) transcription, and [73] which implements a dual attention mechanism for smoother incorporation of both text and audio. While end-to-end models show promise in eliminating errors propagated by ASR systems [74], cascade models are able to leverage large language models

trained on massive amounts of text data for open domain question-answering. Currently, these cascade models may be especially preferable in low-resource applications for which there does not exist enough in-domain data to effectively train an end-to-end model from scratch. End-to-end systems may match or surpass the performance of cascade models as more labeled datasets for spoken question answering become available.

Despite the achievements presented by the aforementioned studies, several challenges remain in creating robust spoken question answering and information retrieval systems. First, much of the work done in spoken question answering is evaluated on datasets such as the Spoken SQuAD dataset [75] or Spoken CoQA dataset [73]. These datasets often only contain spoken questions and contexts that were either generated using text-to-speech or read from a script created from an existing text question answering dataset. This means that further work may be necessary to create spoken language understanding systems that are robust to the disfluencies and lack of proper logical organization often found in spontaneous speech [76]. Second, many of these works format the problem of spoken question answering as finding an answer from a short context (e.g. a one minute audio recording). Many contexts (e.g. a lecture, an instructional video, or a meeting recording) may be significantly longer, and it is non-trivial to scale a model trained for short contexts to infer answers from a longer context. Last, further work is needed to ensure that these systems are robust to differences in dialect, accent, speaking style, and regional diction or other out of vocabulary words. This may be especially true for cascade systems employing pre-trained models that were trained, for example, on only one dialect.

1.7 Outline of the Dissertation

The rest of this dissertation is organized as follows:

Chapter 2 describes the primary databases used in this work

Chapter 3 describes the work done in African-American English Dialect Identification

and Dialect Density scoring. We frame this work in the context of performing linguistic evaluation for more linguistically-informed downstream tasks such as speech recognition and spoken language understanding

Chapter 4 outlines a method created for improving automatic speech recognition for child speakers of low-resource dialects such as African American English. This method uses data augmentation to produce additional training data with targeted characteristics of

Chapter 5 details frameworks for spoken language understanding from speech recognition transcripts of diverse children's speech. These frameworks use state-of-the-art natural language processing algorithms and large language models to automatically score children's oral language exam responses and retrieve specific pieces of information from oral responses for use in educational technology.

Chapter 6 describes methods for spoken language understanding and spoken question answering from long audio files which contain speech from speakers of African American English.

Chapter 7 is a conclusion of the dissertation which offers a summary of key findings and suggestions for future work

CHAPTER 2

Datasets

This dissertation primarily uses data from three datasets. The Corpus of Regional African American Language [77] was used to perform experiments in dialect identification and dialect density estimation on adult African American English speech signals. We additionally created a spoken question answering dataset from CORAAL, CORAAL QA, for use in a spoken question answering task on dialectal speech. The Georgia State University Kids Speech Corpus (GSU Kids Corpus) [78, 79] was used in both children's AAE DID experiments and automatic oral assessment scoring. Last, the UCLA JIBO Kids speech database [78, 80] was collected and used as non-AAE speech in cross-dialect children's speech recognition experiments.

2.1 CORAAL

The Corpus of Regional African American Language database contains spoken interactions between an interviewer and an interviewee who speaks a regional variant of AAE. The set of speakers range in age from under 15 to over 90 and contain roughly equal numbers of male and female identifying participants. The interviewees were asked to describe their daily lives, experiences, and opinions on their communities as well as given space to discuss other topics of interest to them. The entire CORAAL database contains over 200 hours of speech that are divided into 8 components, where each component is a set of speakers from a particular city and time of recording 8. We used the following numbers of speakers with regional dialects from the following five US cities: 22 speakers from Washington DC (DCB), 10 speakers from

Princeville, NC (PRV), 11 speakers from Rochester, NY (ROC), 10 speakers from Lower East Side Manhattan, NY (LES), and 12 speakers from Valdosta, GA (VLD). We chose to use these components, or splits of the dataset, because of their use in prior work [3] and their coverage of different regional speaking styles. These five splits contain 143 audio files total, of which 34 are under 15min in length, 39 are between 15min and 45min in length, and 70 are greater than 45min in length. This totals over 100 hours of spontaneous AAE speech. For each speaker, several utterances with good audio quality ranging from 5sec to 1min in length were selected, and their dialect densities were scored by hand as ground truths. The dialect densities of the speakers in DCB, PRV, and ROC were scored by the authors of [3] while the dialect densities of the speakers in LES and VLD were scored by the authors of this dissertation. This results in a total of approximately 3 hours of dialect density-scored utterances from 65 speakers.

2.1.1 CORAAL QA

To assess performance in the spoken question answering task, we introduce the CORAAL QA dataset [1]. This dataset consists of hand-labeled answer question-answer pairs created from speech contained in the LES, ROC, DCB, PRV, and DCB splits of CORAAL (same as listed in the previous section). From each interview recording, we created a set of questions using the following criteria: 1) The question can be factually and objectively answered by information contained in a continuous time span of the audio file that is 45sec or less in length, 2) The answer to the question is given only once in the audio file, and 3) the answer to the question is not common knowledge and must be answered through extraction from the given audio file. The question answer pairs are given in the format: "query: answer_start_span, answer_end_span" where the answer starting and ending span are given in seconds (e.g. "Who is the speaker's favorite basketball player? : 831.25, 842.76" where the numbers after the colon indicate the start and stop time in the audio file where the speaker gives the answer

[1] data available at https://github.com/christinachance/CORAAL-QA/

to the question).

2.2 GSU Kids Speech Corpus

This dataset contains recordings of approximately 200 students between the ages of 8 and 12 years old from the Atlanta, Georgia area. The data was originally collected in [79] for educational studies. As part of that work, metadata about the student's reading ability and presence of any language impairments was recorded. We later annotated a portion of the dataset for speech tasks. The children in the dataset were recorded while performing educational exercises in reading, language, and pronunciation with a facilitator. First, the students were administered a portion of the GFTA [81] sounds in words exercise in which they were recorded stating phonemically diverse words in isolation. The students were then administered two portions of the Test of Narrative Language (TNL) [82], a story retelling task and a set of picture description tasks, which assessed their oral narrative language abilities. Each child was recorded in 4 sessions each lasting about 2 to 10 minutes. The students were also given additional tasks including sentence formulation and non-word repetition. The entire dataset contains approximately 100 hours of labeled and unlabeled speech data. All children recruited to the study lived in the Atlanta Georgia Area and were native English speakers. The audio was recorded by a computer microphone with a sampling rate of 44.1kHz.

2.3 UCLA JIBO Kids speech

This dataset contains recordings of approximately 130 children between the ages of 4 and 7 years old, the critical age range for early acquisition of literacy. The children were recorded while they performed educational exercises in reading and pronunciation (eg. picture-naming tasks). Each child was recorded in 3 sessions each lasting about 15 minutes. The entire dataset contains approximately 90 hours of labeled audio. The child speakers in the dataset

conversed with the social robot, Jibo [2], following a protocol created by experts in early childhood education [83]. A facilitator was also present at each session and intervened verbally if the child had difficulty interacting with the social robot. Each child sat approximately two feet away from the robot with a microphone placed equidistantly between them. The children then were administered a portion of the Goldman Fristoe Test of Articulation-3 (GFTA3) [81] as well as exercises in counting and spelling. All children recruited to the study lived in Southern California and were proficient in English. Many of these children spoke second languages at home. The audio was recorded by a Logitech C920 Webcam microphone with a sampling rate of 48kHz.

[2] "Jibo Robot - He can't wait to meet you," Boston, MA, 2017. [Online]. Available: https://www.jibo.com

CHAPTER 3

Dialect Identification and Dialect Density Scoring

In this chapter, we introduce novel systems to perform AAE dialect identification and dialect density scoring from short utterances. Recall from Chapter 1 that dialect density is defined as the proportion of a speaker's speech that contains dialect-specific phonological, morphosyntactic, or prosodic cues. DID may be used to automatically provide dialect information to downstream tasks such as ASR or NLU for more dialect-informed processing. Given that an utterance was detected as containing characteristics of AAE, we may want to further estimate the dialect density measure of the utterance in order to processes low and high density utterances differently. This task of dialect density estimation could also be especially useful for data mining in building dialect-specific text-to-speech systems or for linguistic cataloging of a dialect.

3.1 Dialect Density Estimation

This section focuses on performing dialect density estimation for AAE adult speech. The focus of this work is to assess the feasibility of estimating dialect density in a low-resource scenario such as AAE. In order to overcome the lack of data needed to train large language models to perform such tasks, we incorporate linguistic knowledge to attempt to targetedly extract features corresponding to well-documented aspects of AAE.

3.1.1 Data Annotation

This work uses 208 utterances from the CORAAL database. The utterances ranged in length from 15sec to one minute in length. For each utterance, the number of phonological aspects of AAE and the number of morphosyntactic aspects of AAE were counted and divided by the number of words in the utterance to calculate that utterance's DDM. We calculate one DDM that only takes into account the phonological aspects of dialect (DDMphon), one DDM that only takes into account the morphosyntactic or grammatical aspects of dialect (DDMgram), and one DDM that takes into account both (DDM). The utterances from the PRV, ROC, and DCB sets of the CORAAL database were selected and annotated for dialect density by the authors of [3]. The utterances from the LES and VLD sets of CORAAL were annotated for DDM by the authors of this work. Only well-documented phonological and morphosyntactic markers of AAE were counted as linguistic aspects of the dialect. The average DDM for each city in the CORAAL dataset is shown in Table 3.1.

	DDMphon	DDMgram	DDM
DCB	0.083	0.004	0.088
ROC	0.041	0.006	0.047
PRV	0.166	0.028	0.194
LES	0.018	0.025	0.042
VLD	0.122	0.029	0.141

Table 3.1: Average dialect density by city for each of the dialect density measures shown.

3.1.2 Methods

For each DDM-labeled utterance, we extract a feature set which is hypothesized to correlate strongly with a particular aspeect of AAE dialect. We then train a backend classifier to map the input features to a continuous dialect density prediction.

3.1.3 Feature Sets

From each utterance, we extracted the following six feature sets:

Wav2Vec2.0 Transcripts: For the first three feature sets, we generated ASR transcripts using a pretrained Wav2Vec2.0 model [9] trained on the 960hr LibriSpeech database [84]. While these transcripts contained errors and misrepresented the out-of vocabulary (OOV) words, we implicitly attempted to utilize consistent errors and accurate portions of the transcripts to identify useful phonetic and grammatical information.

1. ASR Output Character Combination Frequency: The frequency of each sequence of two characters (bigram) in the transcript was counted and used as a feature. The Wav2Vec2.0 model can output 31 different characters leading this feature to be a 961 x 1 vector which can be thought of as the flattened 31 x 31 matrix in which the element in row i and column j is the number of times character i was followed directly by character j in the generated transcript for the given utterance. We hypothesize that this feature will capture consonant clusters that commonly occur in a particular dialect.

2. ASR Output Character Duration: From the output logits of the Wav2Vec2.0 model, the average duration of each output character was computed. We hypothesize that this feature will be useful in determining which sounds are more or less frequently spoken or stressed by speakers of a particular dialect.

3. ASR Output Language Modeling: In addition to the previously mentioned features, we were interested in how neural language modeling techniques could be applied to automatically generated transcripts of speech in order to predict dialect density. We noticed that, of the most commonly noted features of AAE [21, 85], language differences relating to the tense, collocation, and negation of verbs (eg. absence of copula, negative concord, generalization of "is" and "was" to use with plural and second person subjects, etc.) were especially prevalent. This led us to pay particular attention to verbs. First, the verbs in each utterance were found using a pre-trained FLAIR part-of-speech tagging model

[86]. We then used the Fisher corpus [87], consisting largely of MAE conversations, to train word- and character-based LSTM language models, which provide probability distributions over the next word or character in an utterance given the history. To measure mismatch of verbs in the MAE training data and AAE testing data, we then extracted the verb OOV rate (using the word-based model vocabulary) and the average verb surprisal [88] (using the character-based model) for each utterance, where the surprisal of the i-th word ($S(w_i)$) is calculated from the letter LM as the negative log probability of the i-th word occuring in sequence after words $w_1, ..., w_{i-1}$. We also calculate the overall utterance perplexity from the character-based model (char_ppl), the average surprisal for all words, and the ratio of average verb surprisal to average overall surprisal. Since the LM is trained on MAE, word choices more characteristic of AAE will have high surprisal.

4. ComParE16 Features: The widely used ComParE16 features [89] were extracted from the audio segments using the OpenSmile toolkit [90]. This set includes pitch, energy, spectral, cepstral coefficients (MFCCs) and voicing related frame-level features which are referred to as low-level descriptors (LLDs). It also includes the zero crossing rate, jitter, shimmer, the harmonic-to-noise ratio (HNR), spectral harmonicity and psychoacoustic spectral sharpness. In total, this feature set contains 6373 features resulting from the computation of various statistics, polynomial regression coefficients, and transformations calculated over the low-level descriptor contours.

5. X-Vector: The popular X-vector was incorporated to capture speaker-specific information [33]. These 512-dimensional neural network-generated embeddings contain speaker-specific information that may relate to dialect. As described earlier, the 512-dimensional vectors were projected into 5-dimensional feature vectors using the fully connected network shown in Figure 3.1. This network achieved a validation accuracy of 72.6%.

6. Prosodic Embedding: Inspired by [91], four pitch and energy features were extracted across time from the utterances: F0 (extracted with Praat [92]), the total energy in the frame, the energy in the spectrum below 1kHz, and the energy in the spectrum above

1kHz. These features were then normalized and used as the input to a CNN (as shown in Figure 1) that was trained to predict the region of origin of the speaker. This forces the CNN to classify region specific information from only the prosodic information contained in the speaker's changes in pitch and energy. This CNN achieved a validation accuracy of 70.7%. The output probability vector was then used as the final prosodic embedding.

Weak Supervision: To create the X-vector and Prosodic Embedding Features, we employed a weakly-supervised learning technique. We noticed that the five cities used from the CORAAL database have widely varying average dialect densities, with the averages from PRV and VLD being much higher than those from ROC and LES, and with the DCB average in between. Therefore, we believed that an utterance's city of origin could serve as a weak label in a preliminary step before dialect density estimation. We gathered the set of utterances from the entirety of the 200hr CORAAL database from the five cities of interest that matched the following criteria: 1) Contained at least 10 words to have enough speech to estimate dialect density, 2) Contained no interruptions from the interviewer 3) Were not contained in the set of dialect density-scored utterances. We then used shallow neural networks to map larger input feature vectors into 5-dimensional vectors for which the ith element represents the probability that the utterance was spoken by a speaker from the ith city in the database. This step is intended to project larger sources of information into smaller features vectors which contain only relevant dialect information. The idea is that training a model to classify diverse utterances by region would prompt it to learn region-specific information such as dialectal traits without the need to label the dialect density of all of the utterances in the training set. The output 5-dimensional vector is then used as the representative feature. This framework is depicted in Figure 3.1

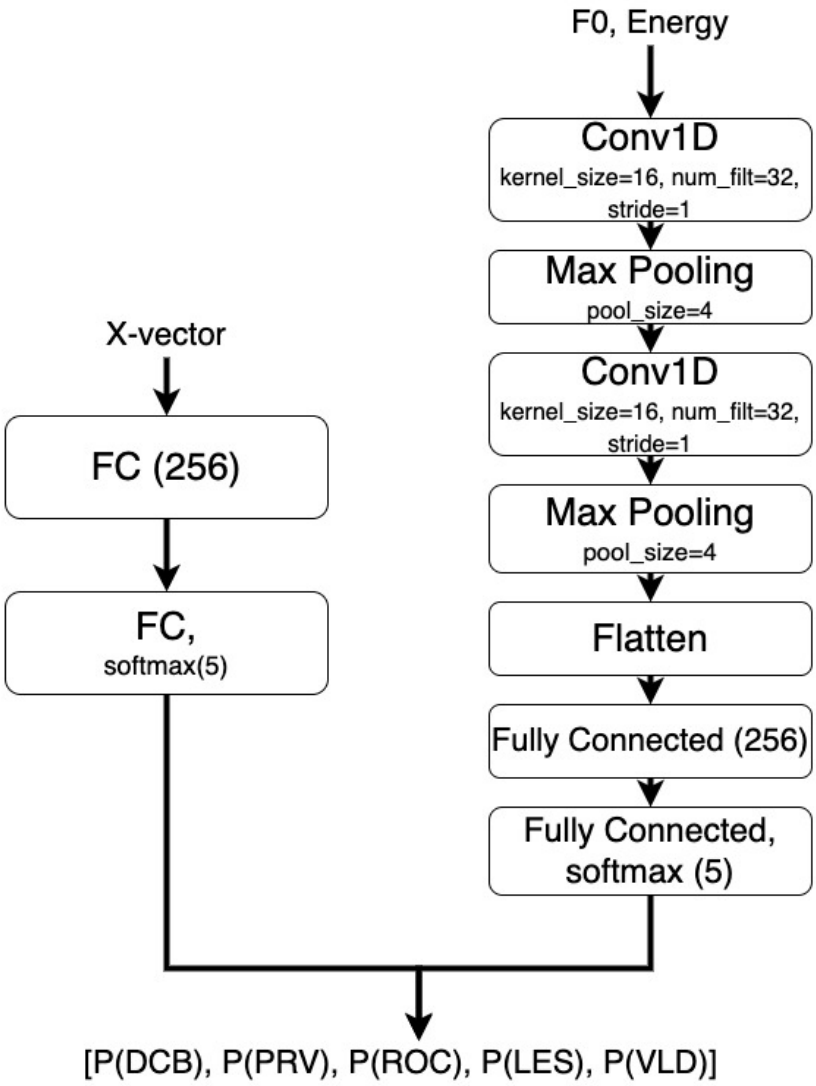

Figure 3.1: The architecture for the fully connected (FC) network used to project the X-vectors (left) and the CNN used to project the prosodic information (right). The inputs to the CNN are the pitch (F0) and three energy contours of the utterance. The output of both networks is a vector whose elements represent the probability of the speaker belonging to each of the cities used from the CORAAL database.

3.1.4 Experiments

First, one distinct XGBoost model [93] was trained for each of the six feature sets. This boosted decision tree model has the advantage of allowing us to easily measure the impact of the input features on the output value for explainability. Each of the six models was trained to predict dialect density scores from one of the given input feature sets. Then the correlation between the predicted dialect density labels and actual dialect density labels was calculated. We chose correlation as the performance metric because human-performed dialect density assessments are subject to possibly high inter-rater variability within the ranges of their scores [94], and so evaluation methods that rely heavily on the absolute value of the dialect density may be subject to measurement noise. However, raters do tend to assign higher or lower scores to the same speakers, and so we expect correlation between predicted and ground truth scores to be meaningful. As some features may only correlate with phonological aspects or only correlate with morphosyntactic aspects of dialect density, we train each model to predict each of the three types of dialect density scores (DDMphon, DDMgram, and DDM). Finally, we used the set of all features as the input to the XGBoost model, as shown in Figure 3.2. As the ComParE16 feature set was large, only the most impactful 10 ComParE16 features were used in the combined feature set.

3.1.5 Results and Discussion

Table 3.2 gives the Pearson Correlation of the predicted dialect density measure with the ground truth labels for the test set for an XGBoost model trained on the listed feature sets. We also include the SHAP value plots [95] which give the relative importance of each feature to the model during prediction. Figures 3.3 and 3.4 give the SHAP value plots for the models trained on all features for predicting DDMphon and DDMgram, respectively. As the DDMphon term dominiates the total dialect density measure, the SHAP value plot for DDM is nearly identical to that of DDMphon. In these plots, the Wav2Vec2.0 Char

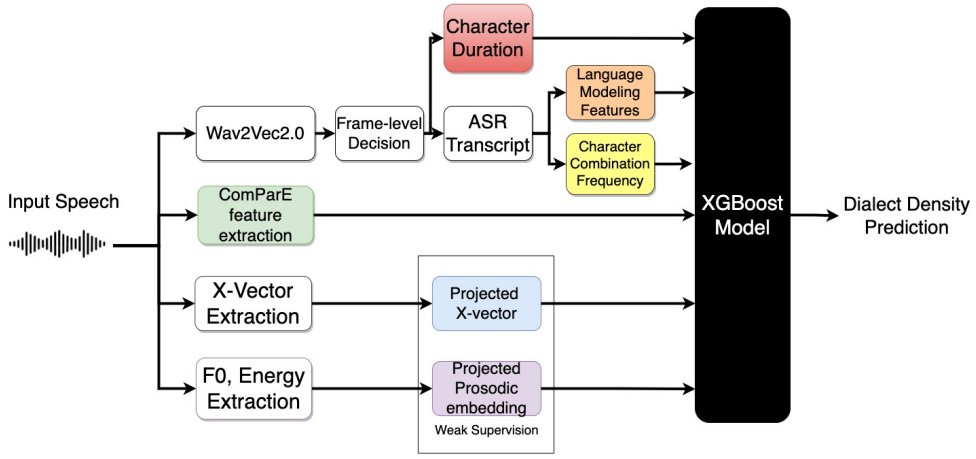

Figure 3.2: Overview of the features used in the proposed dialect density estimation proposed framework.

Comb features are listed as *char*1_*char*2 (e.g. N_space is the frequency of the "N" character being followed by a space character), and the Wav2Vec2.0 Char Dur features are listed as the character whose duration was used as the input feature from the letters A-Z, period, apostrophe, space, or silence (sil) characters.

In order to demonstrate the reliability of our results, we also perform random hold out on the highest performing features. Here, we randomly select speaker-independent train and test split (80% train, 20% test) from the data 200 times and report the average scores over all runs in Table 3.3.

Looking at the individual features, we note that the Wav2Vec2.0 Character Combinations and Wav2Vec2.0 LM features were especially effective in estimating dialect density. Many of the character combinations appear to relate to word initial and word final sounds (eg. N_space (frequency of an "N" followed by a space character in the ASR transcripts), F_space (frequency of an "F" followed by a space), and space_U (frequency of a space followed by a "U")). This is in line with observations that AAE includes dropping of word final nasals and glides and simplification of word initial and word final consonant clusters. Character perplexity (char_ppl) from the language modeling features was the most impactful feature in

Correlation	DDMphon	DDMgram	DDM
Wav2Vec2.0 Char Dur.	0.382	-0.013	0.359
Wav2Vec2.0 Char Comb	0.303	0.124	0.503
Wav2Vec2.0 LM	0.520	0.108	0.637
X-vector	0.404	0.392	0.369
ComParE	0.102	0.189	0.443
Prosody	0.029	0.376	0.008
All features	0.552	0.430	0.718

Table 3.2: Pearson Correlation between actual and predicted dialect density measures for each of the three metrics: only the phonological component of the dialect density (DDMphon), only the morphosyntactic component of the dialect density measure (DDMgram), and the entire dialect density measure (DDM). The results for the model trained on six feature sets individually as well as the model trained on the combination of all of the features are shown.

Correlation	DDMphon	DDMgram	DDM
Wav2Vec2.0 Char Comb	0.339	0.126	0.495
Wav2Vec2.0 LM	0.502	0.173	0.629
All features	0.569	0.385	0.678

Table 3.3: Average Pearson Correlation between actual and predicted dialect density measures for each of the three DDMs over 200 iterations of Random Hold Out validation.

estimating all three DDM scores. This feature is particularly useful in providing an objective distance metric between the MAE of the Fisher Corpus and the ASR transcripts of the target dialect speech which, unlike WER, does not require ground truth transcripts or suffer as heavily in the presence of OOV words. The features derived through weakly supervised embedding (projected X-vector and Prosody embedding) have the most significant correlation with DDMgram. This may indicate that learning grammar from audio files or imperfect transcripts requires larger amounts of data which our method of weak supervision allows us to utilize. In general, the ComParE features using Auditory Rasta filtering proved to be most useful. The RASTA-style filtered auditory spectrum is inspired by psychoacoustics and has been shown to capture context-dependent information useful in ASR [96].

As Figure 3.3 shows, the combination of the five most impactful features in predicting DDMphon was: character perplexity (char_ppl), mean rising slope of the Rasta-filtered auditory spectrum, the frequency of an "N" character followed by a space character in the ASR transcripts (N_space), the standard deviation of distances between peaks in the Rasta-filtered auditory spectrum, and the PRV component of the projected X-vector. As Figure 3.4 shows, the five most impactful features in estimating DDMgram are character perplexity, duration of sounds predicted to be silence or unintelligible by Wav2Vec2.0 (sil), the PRV component of the prosody embedding, the ROC component of the projected X-vector, and the frequency of an "F" followed by an "A" in the ASR transcripts (F_A). The frequency of F_A as a feature may due to a formant shift of the vowel following "F" in several words such as "fell" or "fire" as is seen in some dialects of the US South. We note that the features taken from the Wav2Vec2.0 output are most useful in predicting phonological aspects of AAE dialect density. The character perplexity in particular has a remarkably high correlation with the DDMphon. A high character perplexity is indicative of the presence of character strings that would be unlikely to occur in the MAE sentences of the Fisher corpus. Therefore, this feature proves effective in separating MAE and AAE utterances by the pronunciation perceived by the Wav2Vec2.0 model. The character duration and frequency of combinations

of characters output by the Wav2Vec2.0 model similarly have high correlations with the dialectal phonological differences. The prosody feature by far has the highest correlation with DDMgram. This may indicate that grammatical differences between AAE and MAE often co-occur with prosodic differences. We note that the utterances from different cities of the CORAAL database contain largely disparate numbers of AAE grammatical features, with the utterances from PRV and LES containing several and the utterances from DCB and ROC containing relatively few of these. This may make our method of training prosody embeddings with the utterances' city of origin as target particularly effective in identifying the expected amount of grammatical features. As there are many more phonological AAE tokens than morphosyntactic AAE tokens in the dataset (ie. each spoken clause likely contains only one verb phrase whose grammar structure can be modified but several words whose pronunciation can be changed), the phonological features dominate the total dialect density measure (DDM). As a result, the features that are most useful in predicting DDMphon (eg. character perplexity and character combination frequency) are also more useful in predicting DDM. The ComParE16 features are significantly better at predicting the total dialect density measure than either DDMphon or DDMgram alone. This set contains a large number (approximately 6400) of features, and we note that the XGBoost model utilizes different features from this set for predicting each dialect density measure. For predicting DDMphon, the spectral feature, 'RASTA-style filtered auditory spectrum,' and the OpenSmile feature, 'ZCR' (zero-crossing rate), are the two features with the most impact. These may correlate to dialectal aspects of pronunciation like vowel shifts and durations.

3.2 DID

The previous section describes methods for AAE dialect density estimation for adults given that the utterance is known to be from a potential AAE speaker. In this section, we expand the framework to perform dialect identification for both children and adults whose dialect

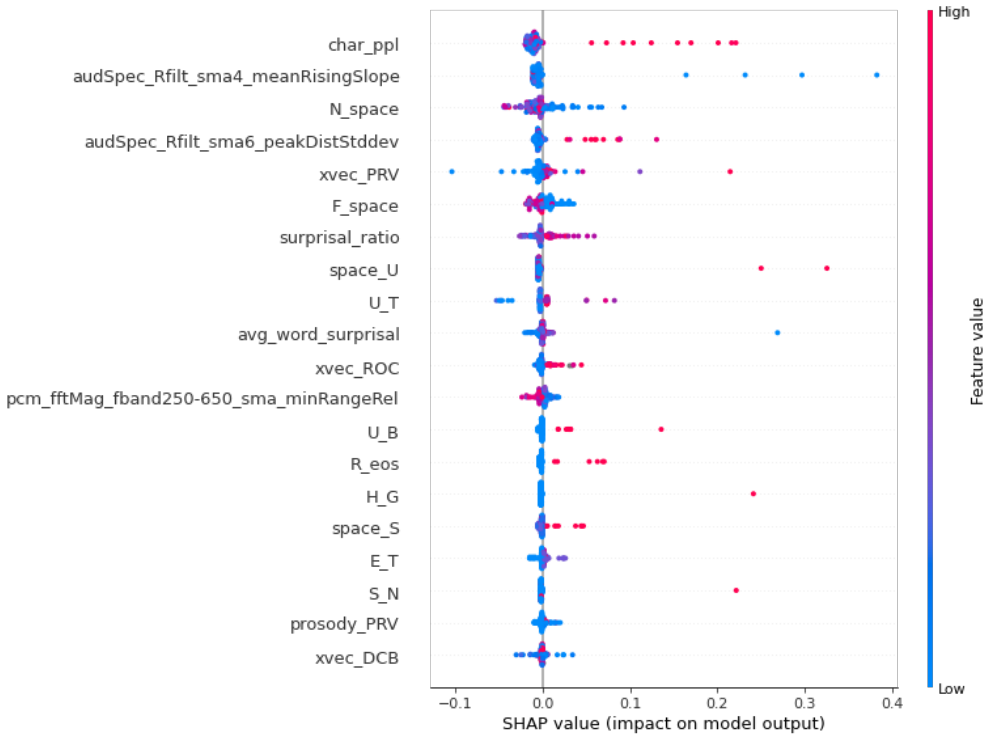

Figure 3.3: SHAP value plot for the XGBoost model trained to predict DDMphon from the set of all features. The features are listed from top to bottom in order of significance.

is not previously known. That is, we seek to determine whether a person is a speaker of AAE or MAE from only a short utterance of their speech. In addition, we seek to create a framework that can be applied regardless of speaker age (adult or child) or speaker style (e.g. spontaneous or read speech).

3.2.1 Data Curation

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.

A particular challenge in this work is learning dialect representations that are robust to

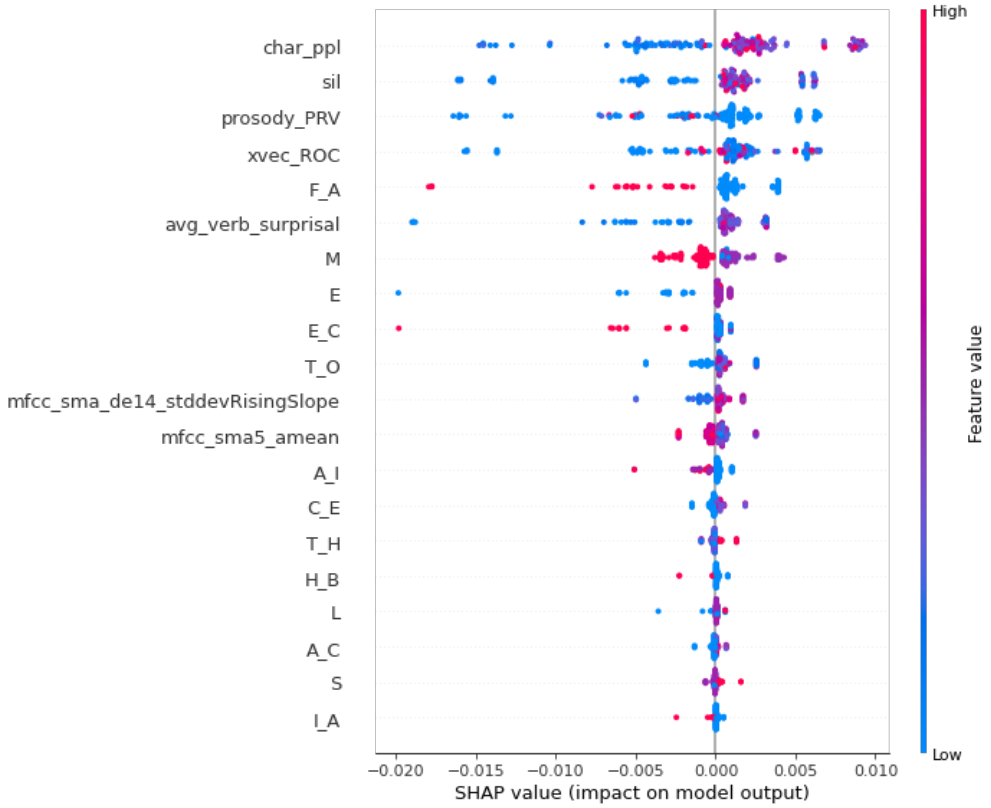

Figure 3.4: SHAP value plot for the XGBoost model trained to predict DDMgram from the set of all features. The features are listed from top to bottom in order of significance.

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. To train the system to perform DID for AAE adult speech, we utilized the recordings of speakers from the Princeville, NC, Valdosta, GA, and Washington DC, sets of CORAAL. The speakers from these sets as had the highest average dialect density, or frequency of use of dialectal characteristics [3, 97], making them more apt to use for DID. 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 [98] 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 [99] 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 [1] (GSU Kids) [100] 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 [3].

The speaking styles and train/test usage of different data sets are summarized in Table 3.4. 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

[1]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.

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 3.4: 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.

dominated by the MAE dialect.

3.2.1.1 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 Wav2Vec2.0 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 [101] 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 [102] 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.

3.2.2 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. An overview of the models used is shown in Figure 3.5.

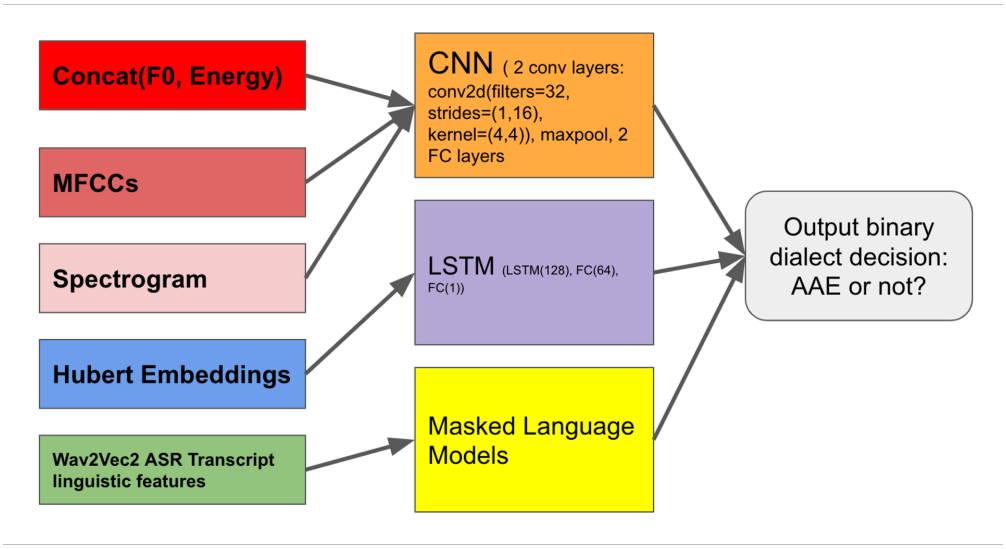

Figure 3.5: The feature set and backend models used in the proposed dialect identification scheme.

3.2.2.1 CNN

We use a modified version of the Convolutional Neural Network from [103] 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 [103] found that the spectrogram was the best feature for DID, [104] saw more success using MFCCs. We evaluate the performance of both of these features for child and adult DID. We extract 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 [105, 97]. These include the F0 contour extracted with Praat [106], 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.

3.2.2.2 LSTM

We employ the popular self-supervised learning representations extracted by Hubert [54] 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.

3.2.2.3 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’) [107]. Character-level ASR systems may capture these pronunciation differ-

ences. We use a Wav2Vec2.0 model [108] trained on the Switchboard Telephone Corpus [109] 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 [3]. Using the ASR transcripts as input, we apply a character-level BERT-style transformer language model (LM) [110], 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,[2] 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

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

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.

3.2.3 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: 1) Resolving AAE-speech in CORAAL from the non-AAE speech in SITW (noisy, spontaneous) and 2) 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

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 3.5: 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.

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.

3.2.4 Results and Discussion

Table 3.5 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	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 3.6: 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.

model for the validation set and two test sets. Table 3.6 shows the Accuracy, F1 score, and AUC for the models. In Table 3.5, the models are trained with a detection threshold of 0.5. The fused models in Table 3.6 use the median value of the test set as the detection threshold. Therefore, we recalculate the performance of the individual models with the median threshold for inclusion in Table 3.6 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 bet-

ter than the models trained on either feature individually in nearly 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 3.6 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 3.5 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.

3.3 Summary

This section describes novel frameworks for both AAE dialect density estimation and AAE dialect identification. We introduce a custom feature set for predicting the extent of a speaker's dialect usage for AAE-speaking adults from the CORAAL database. We then extend that work, incorporating self-supervised learning representations and pre-training on larger amounts of data, to perform dialect identification for both adult and children's speech.

This work was presented in part at Interspeech 2022 [97] and ICASSP 2023 [45].

CHAPTER 4

Data Augmentation for Low-Resource ASR

After the stage of automatic dialect identification or dialect density classification decisions for a speech utterance, we may then desire to select a dialect-specific down-stream language tasks. For example, we may train ASR models for specific dialects, regional variants of the dialect, age groups of speakers of the dialect. However, there is often not enough training data to train a large language model with such specificity of the speaker's sociolinguistic identity. The research question then becomes: How can we train ASR systems to perform dialect or age-specific recognition with limited training data?

In this chapter, we investigate methods for data augmentation for low-resource ASR applications such as ASR for AAE-speaking children. In other words, we seek to create artificial training data that can be used to train an ASR system to learn to be more invariant to characteristics of AAE and children's speech when transcribing audio data. We hypothesize that an LPC-based method of estimating the formant frequencies (i.e. poles) of the signal, shifting the poles of the LPC filter, and reconstructing the signal with the perturbed filter coefficients can model formant shifts found in southern AAE variants in available non-AAE training data. Training on this perturbed data would then teach the ASR system to recognize speech even with these formant shifts. We evaluate the performance of this method, named LPC Augment, in both low-resource and zero-resource training scenarios for Georgia AAE and California English-speaking children.

4.1 Methods

This work uses speech from the UCLA JIBO Kids' Speech Database and the GSU Kids' Speech Database as described in Chapter 2. In this chapter, we seek to evaluate how well data augmentation algorithms can transfer aspects of dialect to the newly generated training data. To test this, we attempt to use LPC Augment to add characteristics of Georgia dialect (we henceforth refer to the combination of Southern American English and African American English represented in the dataset as "Georgia English") vowel shifts to the California English contained in the UCLA JIBO Kids's Speech database in order to train a system to recognize the US southern dialect with little to no prior in-domain training data. Conversely, we also attempt to warp the formant locations of the Georgia English contained in the GSU Kids' Speech Database to the expected frequency locations of California English to train the systems to recognize California English with little or no in-domain training data. We use recordings of children performing the Goldman-Fristoe Test of Articulation from both datasets. This creates a matched-vocabulary set of words being spoken in isolation by child speakers of either Georgia English or California English. There are approximately 5 hours of speech available for either dialect, making this a low-resource training case.

4.1.1 The LPC Augment Algorithm

In this section, we propose a novel data augmentation technique. We use the notation from the linear prediction equation $s[n] - \sum_{k=1}^{P} a_k s[n-k] = e[n]$ where $s[n]$ is the windowed frame of the signal and $e[n]$ is the residual, the prediction order, P, is estimated as $P = 2F_{max}$ (in kHz) $+2$ where F_{max} is one half the sampling frequency.

The algorithm is then given as follows assuming an all pole model. For each frame:

1: Compute the LPC coefficients, $a_1, a_2, ...a_P$ of the windowed signal.

2: Compute the residual $e[n]$ as the result of passing $s[n]$ through the filter $A(z) = 1 - \sum_{k=1}^{P} a_k z^{-k}$

3: Solve for the complex conjugate roots, r_k, of the prediction filter polynomial $A(z)$

4: Compute the magnitude and phase of each root r_k

5: Multiply the phase of each complex conjugate pair of roots by a warping factor $w_k \forall k = 1, 2, ...P$ where the warping factor values are chosen from a random uniform distribution $\in [x, y]$ once for each utterance and held constant across all frames of the utterance.

6: Recombine the magnitude with phase of each pole, creating the warped polynomial roots $\hat{r}_k = |r_k| * e^{j(w_k * \angle r_k)}$. The warping does not affect the magnitude in order to ensure filter stability.

7: Determine the new prediction polynomial, $\hat{A}(z)$ as the polynomial whose roots are the warped prediction filter roots, $\hat{r}_k$

8: Create the perturbed output frame by passing the residual $e[n]$ through the filter with transfer function $1/\hat{A}(z)$

An example of the spectrum of signal perturbed with LPC Augment is shown in Figure 4.1. A block diagram of the algorithm is shown in Figure 4.2.

4.1.2 Model Training

In this section, we introduce the experimental setup including the data and models. For the data, both the Californa English speech data from the UCLA JIBO kids' dataset and Georgia English from the GSU Kids' Dataset are split into training, validation, and test sets with a ratio of 7:1:2 with no overlap between speakers. As a result, approximately 3.5 hours of data are used for training. We verify the proposed method across both a hybrid HMM-DNN model and an end-to-end attention-based model to demonstrate that LPC Augment is not model-dependent.

The hybrid system is built based on Kaldi [59] and Pykaldi2 [111], where Pykaldi2 is used for acoustic model training. Specifically, a four-layer bidirectional long short term memory (BLSTM) model with 512 nodes in each direction is used. The input is 80 dimensional log

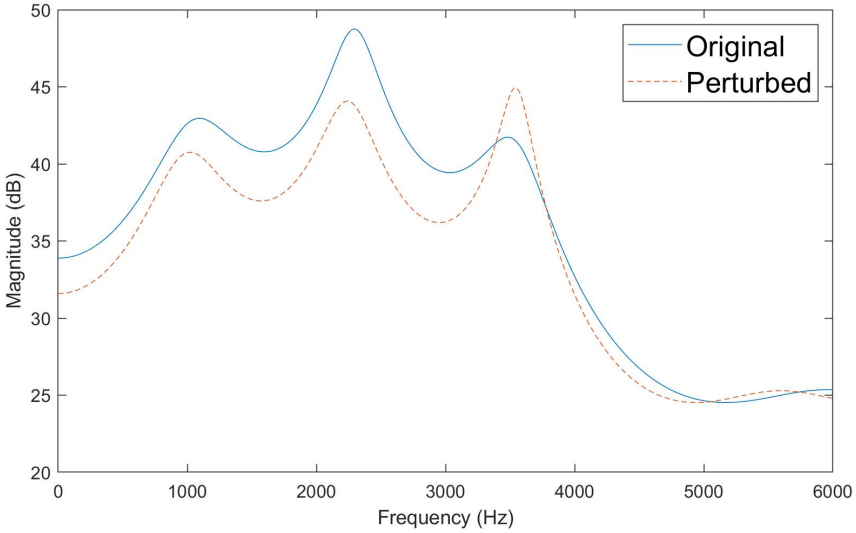

Figure 4.1: An example of the LPC spectra of a child in the UCLA JIBO kid's Database pronouncing the phoneme \AA\, and the result of perturbing it with LPC Augment. In the perturbed signal, the first two formant peaks have been shifted to the left, and the third has been shifted to the right.

filter-bank energy extracted with a frame length of 25ms and a frame shift of 10ms. We also use a frame skipping strategy [112] by concatenating two adjacent frames and then skipping the frame by a ratio of 2 to accelerate the model training. The acoustic model then outputs an approximately 3488 dimensional vector representing the senone probabilities for each frame, which are then decoded using a pre-constructed WFST graph in Kaldi. Prior to the acoustic model training for kids' data, HMM-GMM and BLSTM models are trained using the Librispeech clean 100 hours data [84], as the model for the forced alignment acquisition and the start point for kids' acoustic model training, respectively.

The end-to-end model is the sequence-to-sequence (S2S) Speech Transformer as proposed in [113]. The model input is the spectrogram calculated with a frame size of 25ms and a frame shift of 10ms. The input is then passed to the network which consists of a series of three convolutional layers each with a receptive field of size 11, and an encoder and decoder

block both composed of six stacked multi-head attention units and fully connected layers with residual connections. The output then contains 31 classes: 26 lowercase letters, apostrophe, period, space, noise marker, and end-of-sequence tokens.

We use the proposed data augmentation scheme to train the model and evaluate the performance across training and testing conditions. Data augmentation was performed using MATLAB's LPC coefficient algorithm. Each utterance is windowed using a 20ms long Hamming window before being passed into the augmentation algorithm. The length of the Hamming window was determined empirically in pilot experiments. The perturbed audio sample is then input into the neural network. We first optimize the range of the warping factors w_k using the validation set. The training set size was increased by 3x with the proposed method. Preliminary experiments showed that augmenting the training set size to 5x yielded no additional increase in performance. We then compare the performance of the optimized proposed method with that of other common data augmentation methods in recognizing children's speech of both the in-domain dialect and an out-of-domain dialect.

We first train the models on speech from one of the English dialects and test on the other dialect for the zero resource scenario. We then train a model jointly on both children's datasets for the low-resource scenario.

4.2 Experiments and Results

4.2.1 Optimizing the Warping Factor

In order to optimize the range of warping factors, w_k, used in the proposed method, we first train the transformer model on the California English training set and evaluate it on both the California English and Georgia English validation set (CA val and GA val respectively). We use the training set containing California English because it is considered a widely-spoken American dialect. Adapting the California English training set to the Georgia English validation set then represents adapting from a more standard dialect to the less standard

Warping Factor	CA val	GA val
$w \in [0.8, 1.0]$	16.41	63.44
$w \in [1.0, 1.2]$	15.97	69.16
$w \in [0.8, 1.2]$	14.63	**51.63**
$w \in [0.9, 1.1]$	14.86	55.93
$w \in [0.7, 1.3]$	**13.94**	58.85

Table 4.1: Results of the recognition experiment (in %WER) on the validation set with the proposed method for different warping factors using the transformer model. CA Val denotes the performance of the system trained with data augmentation on the speech data containing dialects found in Georgia and validated on speech containing dialects found in California. GA val similarly denotes the performance of the system trained with data augmentation on the speech data collected in California and validated on the speech data collected in Georgia. The lowest WER for each case is shown in boldface.

dialect as in low-resource scenarios. Table 4.1 shows the performance in percent word error rate (%WER) of the proposed algorithm for warping factors within the indicated range.

4.2.2 Zero Resource Scenario

Here, we are primarily concerned with achieving the best result on the out-of-domain data (GA Test), and so we continue with the warping factor chosen in the range $[0.8, 1.2]$. Zero resource scenarios occur when the model is trained on only one dialect and tested on another. We proceed to compare the performance of the proposed method (abbreviated LPC Aug) in zero resource dialect children's ASR with three of the most commonly used data augmentation algorithms: VTLP, Speed Perturbation (Speed Pert.), and Spec Augment (SpecAug) (described in Chapter 1). We also combine the more successful data augmentation methods to determine their cumulative effects. The results of both the transformer and hybrid model

are shown in Table 4.2. The proposed method, LPC Augment, achieves a statistically significant ($p < 0.05$) reduction in WER over the baseline ("No Aug") for mismatched dialect cases. The lowest WER when training on one dialect and testing on the other is achieved when LPC Augment is used in conjunction with SpecAugment. In testing and training on the same dialect, the lowest WER is achieved by using SpecAugment alone in three out of four cases.

4.2.3 Low Resource Scenario

We train the models on data from both the CA dataset and the GA dataset in order to create an ASR system that performs well over multiple dialects and ages. This represents the low-resource case, as the training sets from both dialects are small. We show the results in %WER in Table 4.3. Note that the high baseline WERs observed in Tables 4.2 and 4.3 have been observed in previous low-resource accented children's ASR tasks in other languages as well [114].

4.3 Discussion

We observe that LPC Augment creates a significant reduction in WER for zero resource dialect children's ASR as compared to the other frequency-based data augmentation method, VTLP. This is likely due to the algorithm changing formant locations independently of each other rather than according to a predefined warping function. It appears that LPC Augment is complementary to SpecAugment, as they can typically be used together to give better performance than either alone or compared to other data augmentation methods. In the zero-resource scenario in Table 4.2, the HMM-DNN ASR system results in a much higher reduction in WER for the CA test set than for the GA test set. This may be a result of pre-training the model on Librispeech 100-clean which contains speech of a dialect more similar to California English. Further work is necessary to determine how the pre-training dataset

biases the model towards better performance for a given dialect. We also notice in Table 4.3 (low-resource case) that the transformer model typically benefits more (% improvement over the baseline) from data augmentation than the HMM-DNN system. The transformer's implicit language modeling may allow it to better learn relevant groupings of characters and hence may have a bigger advantage for the children's small vocabulary task. In the low-resource task, we observe that LPC Augment used simultaneously with SpecAugment and Speed Perturbation appears to give improved performance across dialects. We conclude that LPC Augment shows promise in creating robust low and zero-resource dialect ASR systems.

4.4 Summary

This chapter introduces the LPCAugment technique for training a cross-dialect children's speech recognition system. By decoupling the linear predictive coefficient representation of the vocal tract filter effects from the speech source signal, perturbing the LPC coefficients, and then reconstructing the speech signal, the LPCAugment attempts to model dialectal formant shifts in the augmented training data. We apply this technique to few and zero-shot children's speech recognition tasks and achieve state-of-the-art performance.

This work was presented in part at ICASSP 2022 [100].

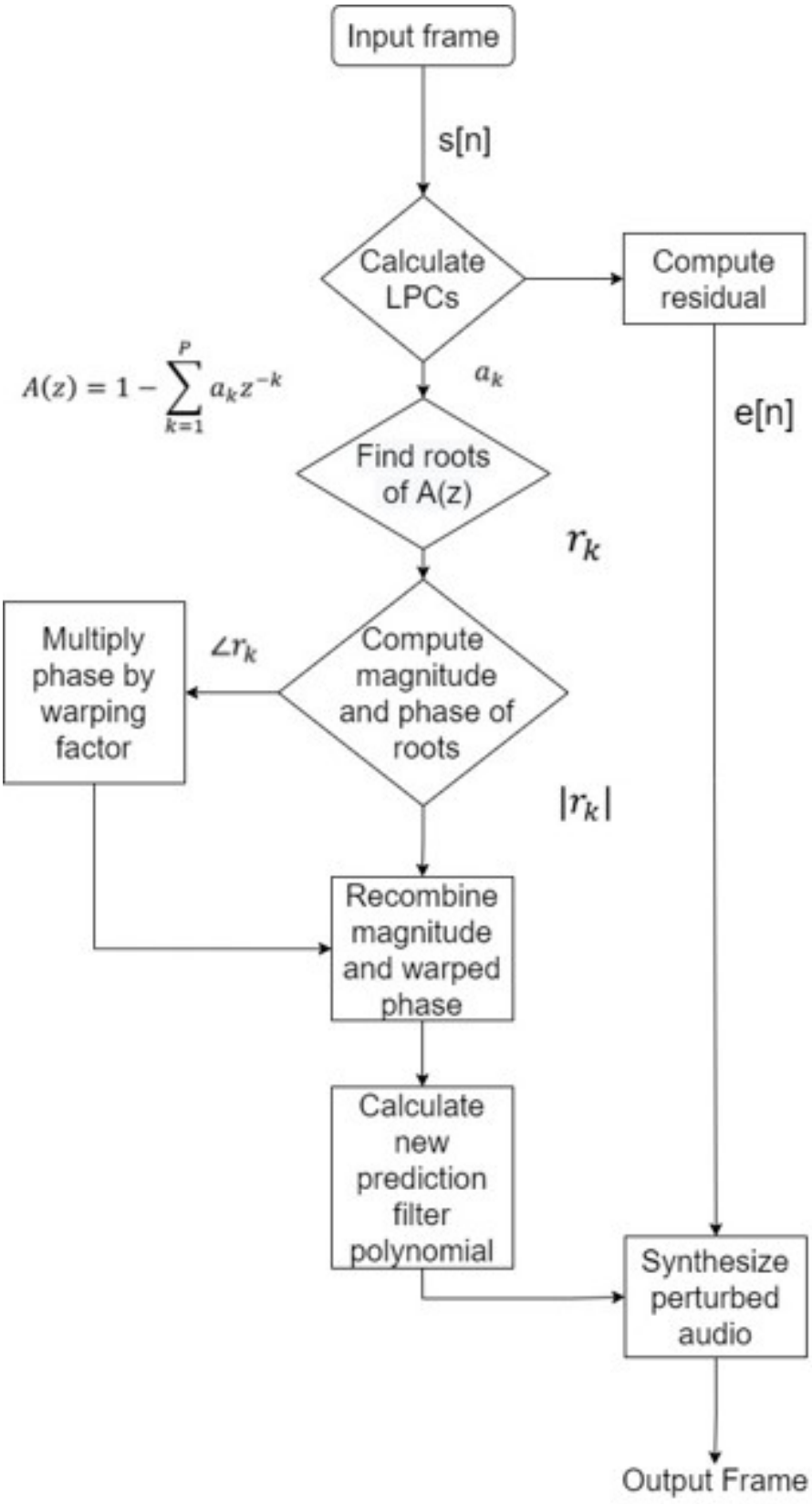

Figure 4.2: Diagram of the LPC Augment Algorithm

	Train CA				Train GA			
%WER	Transformer		HMM-DNN		Transformer		HMM-DNN	
Test Set	CA test	GA test	CA test	GA test	CA test	GA test	CA test	GA test
No Aug	18.34	70.00	16.39	76.29	56.56	24.01	92.26	37.74
VTLP [51]	20.07	71.46	15.85	77.83	65.41	25.10	91.26	37.29
Speed Pert. [52]	26.39	69.58	14.57	76.74	63.12	27.82	90.44	38.82
SpecAug [53]	**17.85**	62.84	13.93	76.47	54.84	**22.64**	88.71	**34.84**
LPC Aug	19.49	62.70	14.30	76.74	51.76	24.79	81.33	38.73
Speed Pert. + SpecAug	21.32	68.63	14.30	76.92	63.85	22.95	90.16	37.56
Speed Pert. + LPC Aug	23.86	71.88	13.75	77.29	55.68	23.54	83.15	36.83
SpecAug + LPC Aug	18.61	**59.80**	**13.30**	**75.02**	**51.13**	22.90	**75.41**	35.48

Table 4.2: Comparison of common speech data augmentation methods with the proposed method. Each model (Transformer and HMM-DNN) is trained on either the California English training set (Train CA) or the Georgia English training set (Train GA) and then evaluated on both the California English test set (CA test) and the Georgia English test set (GA Test). Columns representing zero-resource scenarios (where the model is trained on only one dialect and tested on the other) are highlighted. The lowest word error rate for each case is shown in boldface.

	Transformer		HMM-DNN	
%WER	CA test	GA test	CA test	GA test
No Aug	26.12	21.18	16.94	37.83
VTLP	21.47	15.84	17.49	36.83
Speed Pert.	19.68	14.57	15.76	37.92
SpecAug	19.23	13.80	15.03	**35.20**
LPC Aug	19.76	14.39	14.66	38.46
Speed Pert. + SpecAug	**18.72**	13.76	14.39	36.74
Speed Pert. + LPC Aug	19.01	13.52	14.30	37.47
SpecAug + LPC Aug	18.91	**13.34**	**13.84**	35.29

Table 4.3: Results of the models trained on both Train CA and Train GA and tested on CA Test and GA Test with the proposed and other data augmentation methods. The lowest WER for each case is shown in boldface.

CHAPTER 5

Fair and Inclusive Automatic Oral Assessment Scoring

This chapter is concerned with producing natural language processing algorithms for spontaneous speech that are robust to disfluencies and patterns not found in written text. We seek to apply these algorithms to the educational domain for tasks such as automatically providing feedback on oral exercises and categorizing student responses. To accomplish this task, we train a model to automatically grade recordings of oral responses from children taking portions of the Test of Narrative Language (TNL). We do so with the goal of designing the system to be inclusive and robust to speech differences found across diverse dialects and language abilities.

5.1 Educational Task

In this work, we use recordings from the GSU Kids' Speech Corpus. This corpus contains approximately 200 recordings of children in 3rd to 8th grade performing two types of tasks from the TNL: A story retelling task and a picture description task.

5.1.1 TNL - Story Retelling Task

The TNL Story Retelling assessment is a task in which students are read a story by the test administerer. The students are then asked to retell the story and are graded on their ability to use the set of pre-determined test keywords from the original story-telling. These keywords contain story elements (e.g. character names, locations, times, important objects,

and action verbs) that must be retold in the same verb tense and order to receive credit in the test scoring. For example, if a test item contained the sentence, "**Tim eats** his lunch while **Matt** plays **football**" where the bolded words are the scored keywords, the child will receive points for two of the four keywords if they retell it as "**Tim** played **football** while **Matt ate** lunch," as the word order or tense of the other two keywords are incorrect. Each child's assessment was administered and audio recorded by a trained member of the project staff according to the TNL standardization manual protocols. The recordings were then independently scored by a speech-language pathologist and a second trained speech-language staff member. If disagreements occurred in scoring, the two scorers reviewed the audio and discussed differences to come to a consensus. Each child's score was an integer value between 0 and the total number of test keywords. Although not necessary to the TNL protocol or the training procedure below, the project team additionally transcribed ground truth transcriptions for each audio recording. The dataset additionally contains demographic metadata on the students in the following categories: 1) the presence of reading/language impairment, 2) the student's reading ability (good or poor) as rated by a team of experienced teachers and learning specialists as a selection criterion for the study, and 3) the speaker's dialect (either African American English (AAE) or Southern American English) as labeled by the authors according to the procedure in [115].

5.1.2 TNL - Picture Description Task

Picture description tasks are often used to elicit spontaneous speech from children. Students are shown a picture with multiple characters or elements relating to a story plot. Images are generally chosen by experts in education to be straightforward to describe and contain enough content for the child to give a lengthy answer. The students are then asked to tell a story about the picture. Students are graded based on completeness of the description, coherence of the story, proper use of grammar, and other aspects relating to narrative language ability. In the GSU Kids' Speech Corpus data used in this work, children were shown an image

from the TNL containing a character and several elements to describe. The students were then asked to tell a story about the image, making their story as complete as possible. Each child's response to the prompt was recorded, and each child, on average, took about 3 minutes to complete their story. Then, specialists in children's language education graded the assessment as described in [79].

5.2 Experiment

Given the recordings and grades from the students' TNL story retelling and picture description task responses, we then sought to train a model to automatically map an ASR transcript of the response to an assessment score. We first map the test scores to discrete labels in order to formulate the assessment scoring problem as a multi-class classification task. This is intended to reduce over-fitting to negligibly small differences between scores. We sort the scores into five equally spaced histogram bins and then assign a class to each sample based on its bin, resulting in a five-class classification task. We predict the class automatically from ASR-generated transcripts. We consider ASR transcripts from Whisper, Wav2Vec2.0, and HuBERT. A 4-gram KenLM language model [116] trained on the LibriSpeech Corpus [84] is applied to the the output transcripts of each ASR system, and we report scoring accuracy both with and without the effects of the external language model. We also fine-tune HuBERT on the MyST Database [117] to explore possible improvements from training on additional children's speech data. To establish a baseline performance for the task, we first implement a deterministic string-search based assessment scoring method on the student response ASR transcript. We then propose a Neural Linguistic Feature-based approach to improve over the deterministic approach by training a system to more flexibly take dialectal speech differences into account and learn despite ASR errors.

String-Search Rubric Matching-based scoring (SSRM) The TNL provides a comprehensive scoring rubric which assigns points to each keyword given. As a preliminary

approach (shown in Figure 5.1, left), we apply fuzzy string matching with a similarity ratio of 85% to the ASR transcripts to identify close matches to the specified keywords. This method then awards points to a student's predicted score if a word whose Levenshtein edit distance with a test keyword is less than or equal to 15% of the word length appears in the ASR transcript. We then present the accuracy and root mean square error (RMSE) of this method for each ASR systems used.

Neural Linguistic Feature-based Scoring (NLF)

A weakness of the SSRM scoring is that it only considers whether or not a close match to a keyword appears in the transcripts. It does not consider, for example, whether words were used in reference to the right characters or appeared in the correct order. These tasks require a neural network-based approach to capture finer scoring details. For this, we split the samples from the TNL into a 45-15-40, train-val-test split. We arrived at this split by starting with a 70-10-20 train-val-test split and reducing the amount of data in the training set until performance significantly worsened. This was done to best simulate the low-resource data scenario found in many children's speech responses. With this data, we employ methods used in readability assessment from [66]. We first apply transfer learning to large language models to generate soft label features from the transcripts for downstream scoring. Here, we experiment with BERT [118], RoBERTa, BART [119], and XLNET [120] using Huggingface. A 5-dim fully-connected layer is appended to the output of the large language model, and then, we fine-tuned this extended model on a combination the WeeBit Corpus [121] and the training set of the TNL data for more task-specific text-scoring. A parameter grid search over the validation set using the AdamW Optimizer found a linear learning schedule (beginning with a learning rate of 2e-5 and a weighted decay of 0.01 after 10% of the total training steps were used in warmup) and a batch size of 8 as the best system hyperparameters. The other model hyperparameters were not changed from their original implementations. The WeeBit Corpus contains short news article-style texts used for children's reading comprehension tasks. These texts are each labeled with an integer

difficulty rating between 1 and 5 with difficulty 1 meant for children ages 7-8 and difficulty 5 meant for children ages 14-16. By having the network simultaneously learn to both predict difficulty levels of children's texts and scores for narrative language proficiency on spoken transcripts, we create a multitask learning framework in which the machine must learn to combine both knowledge of age-appropriate reading texts (WeeBit) and knowledge of oral language abilities (TNL). As education literature shows that children's reading proficiency and comprehension abilities are directly correlated with their oral language proficiency [122], we use this strategy to make the machine use the same weights to jointly predict both tasks. We report the accuracy of this system in predicting the TNL scores of the test set. Next, we extract the subset of the 255 hand-crafted linguistic features found optimal for the WeeBit corpus in [66] and try additional features from that study in the "Discourse", "Semantic", and "Traditional" categories which capture several measures used to score essay quality (eg. ease of identifying main topics, density of predicted entities, lexical difficulty of words used) as proposed by [66]. Finally, we concatenate the hand-crafted features with the soft-labels given by the large language model for input to a backend classifier trained to perform score prediction (as shown in Figure 5.1, right). We experiment with logistic regression, support vector machines, Random Forest, and XGBoost [93] and report the accuracy of each system. As no training is necessary for the SSRM scoring, we report the accuracy over the entire set of speakers. For the NLF scoring, we report metrics as averaged over 5 train-test splits. To ensure that the model performs fairly for diverse students, we also report test accuracy for the student demographic categories (reading or language disability, reading status, and dialect spoken).

5.2.1 Results and Discussion - TNL Story Retelling

Table 5.1 shows the results of the SSRM method for the different ASR transcripts in comparison to the performance on the ground truth transcripts. We report ASR WERs and the 5-class classification accuracy and RMSE. In addition to lowest overall WER, Whisper also

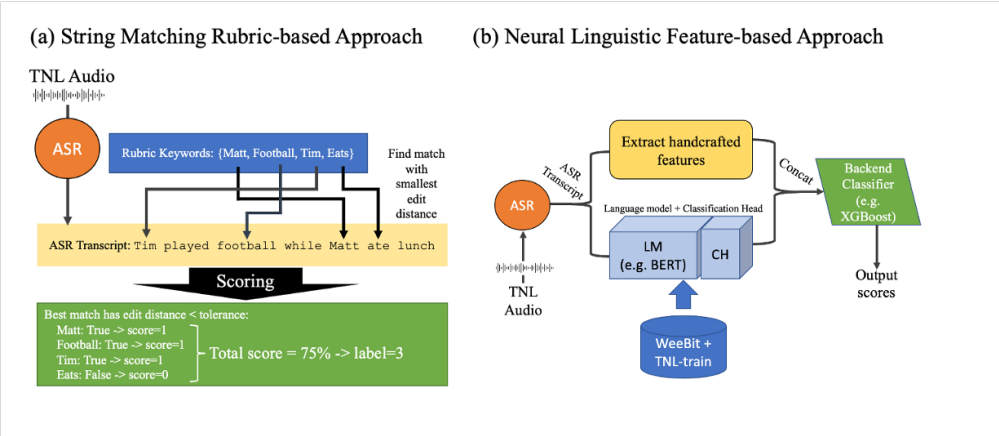

Figure 5.1: Overview of a) the string-search rubric-based approach and b) the neural linguistic feature-based approach.

had 93.6% precision and 93.7% recall in correctly detecting and transcribing the test keywords with a detection threshold of 85% string matching. Table 5.2 shows the performance of the NLF method while only using the language model with a final classification layer (no linguistic features) after being fine-tuned on the WeeBit corpus and training set of the TNL transcripts. To better understand the effects of different steps in this training pipeline, we perform an ablation study in which we remove stages from the training. Fine-tuning the best performing language model on only the TNL with no WeeBit text data gives a maximum classification accuracy of approximately 60%. Likewise, fine-tuning this language model on only the WeeBit text without using the TNL transcripts gives a maximum classification accuracy of about 58%. Next, Table 5.3 shows the performance of backend classifiers using a concatenation of the soft-labels from the best system in Table 5.2 and hand-crafted linguistic features. Table 5.2 shows that BERT performs equally well with either the Whisper ASR transcripts or the Hubert-finetuned on MyST transcripts. We proceed with the Whisper transcripts because of their higher semantic similarity with the groundtruth transcripts (depicted in Figure 5.2). Finally, we divide the student samples into the three demographic groups listed in the previous section and show the performance of the best system for each group in Table 5.4.

Table 5.1: ASR Word Error Rate (WER) , Classification Accuracy (C. Acc), and classification RMSE for the fuzzy string matching approach for each system

Transcripts	WER	C. Acc	RMSE
Ground Truth	-	88.0%	0.120
Wav2Vec2	37.0%	61.4%	0.402
HuBert	46.7%	62.0%	0.413
HuBert Finetuned	42.5%	64.1%	0.407
HuBert XL	43.9%	73.3%	0.282
HuBert XL w/ 4-gram LM	38.9%	76.0%	0.282
Whisper Large	26.8%	86.4%	0.136
Whisper Large w/ 4-gram LM	**26.3%**	**87.0%**	**0.130**

A comparison of the string-matching (SSRM) approach and neural (NLF) approach shows that the machine learning method far outperforms the rubric-based baseline. The proposed system (BERT soft labels + hand-crafted linguistic features + XGBoost) achieves a classification accuracy of 98.5% using the Whisper ASR transcripts. In comparison, the rubric-based approach achieves only about 87% classification accuracy and sees marginal improvement even with the ground truth transcripts. Since the rubric-based method only considers whether or not test keywords appeared in the story and not whether they're used coherently, the performance difference between the two approaches suggests that the proposed system is able to capture grammar and logic rules used in scoring the assessments that cannot be assessed with a simple fuzzy string search. The ablation study shows a significant degradation in performance of the proposed approach when either the test set or the added WeeBit set is removed from the fine-tuning process. This further suggests that the language model only performs well given a sufficient amount of in-domain data but can make use

Table 5.2: The classification metrics (C. Accuracy, F1-score, and RMSE) of each of the fine-tuned language models considered when predicting scores. Numbers reported are the average of 5 trials of random hold out.

	BERT			ROBERTA			BART			XLNET		
Transcript	C. Acc	F1	RMSE	C. Acc	F1	RMSE	C. Acc	F1	RMSE	C. Acc	F1	RMSE
Ground Truth	96%	0.95	0.04	91%	0.90	0.15	80%	0.78	0.40	84%	0.83	0.16
Whisper	**96**%	**0.95**	**0.04**	90%	0.89	0.16	78%	0.77	0.43	82%	0.82	0.18
HuBert Large	95%	0.94	0.04	86%	0.85	0.27	75%	0.60	0.50	80%	0.79	0.24
HuBert Base	89%	0.88	0.11	83%	0.83	0.16	71%	0.69	0.29	80%	0.77	0.35
HuBert Base Ft	**96**%	**0.95**	**0.04**	93%	0.93	0.09	82%	0.82	0.18	82%	0.81	0.19
Wav2Vec2	85%	0.83	0.15	87%	0.84	0.25	70%	0.70	0.40	85%	0.83	0.30

of the correlation between reading proficiency measures (with WeeBit readability scores) and oral proficiency measures (with the TNL training set) in order to learn relationships in children's language well. Given that we only use 45% of the 184 samples in training, this approach appears to successfully deal with the low-resource data problem in children's language assessments. The demographic splits in Table 5.4 imply that our method performs fairly across language ability (or presence of disability), reading ability, and dialect. Across the Reading/Language Impairment demographics, the NLF method matches or outperforms the rubric-based approach in all cases. We note, however, that the rubric-based approach performs more fairly across these categories, with scores from the ASR transcript varying by less than 3% absolute value from the control students (no impairment) to the students with both a reading disorder and language impairment. The NLF method achieves nearly perfect accuracy for the control case. However, this system performs worse for students with impairments who may make non-standard language errors not present in the WeeBit training set. The complex nature and differing severities of these language disorders creates high variability in the narrative language abilities of the students in these groups. This suggests

Table 5.3: System performance using a backend classifier to predict assessment scores from an input concatenation of hand-crafted linguistic features and soft labels from the best-performing large language model (BERT) extracted from the best ASR transcripts (Whisper). Backend classifiers tested are: Support Vector Machines (SVM), Logistic Regression (LogReg), Random Forest (RandFor), and XGBoost.

BERT Soft Labels + Linguistic Features				
Transcripts	Classifier	C. Acc	F1 Score	RMSE
Ground Truth	SVM	96.9%	0.96	0.034
	LogReg	97.0%	0.96	0.032
	RandFor	98.5%	0.97	0.029
	XGBoost	99.2%	0.99	0.025
Whisper	SVM	96.5%	0.95	0.038
	LogReg	96.0%	0.96	0.039
	RandFor	97.6%	0.97	0.032
	XGBoost	**98.5**%	**0.98**	**0.030**

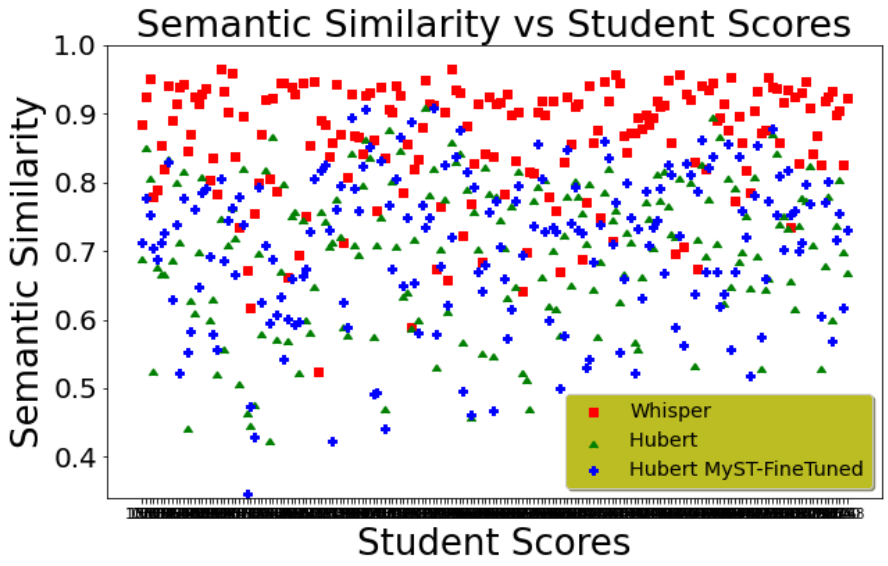

Figure 5.2: Semantic Similarity between each student's ASR and ground truth transcript. ASR transcripts generated with Whisper, Hubert, and Hubert fine-tuned on MyST.

the need for additional data or data augmentation strategies to model disordered children's language in order to improve performance more fairly across these demographics. We observe a similar trend in the poor vs good reading status demographics. However, the NLF approach performs better than the rubric-based approach across both the "good" and "poor" reading students. The almost 6% drop in performance of this system from the ground truth transcripts to the Whisper transcripts for the Poor reading status group may mean that further ASR improvements are needed for the speech differences that these children display. We observe relatively unbiased performance across dialect for the proposed system. While the work in [115] demonstrates that many commercially available ASR systems give worse performance for US East-coast AAE speakers than California General American English, little has been done to compare ASR performance for AAE vs other American dialects like Georgia's Southern American English. The comparably high WERs for these two Georgia dialects shown in Table 5.4 demonstrate that further work is needed in understanding and improving ASR for multiple types of children's regional dialectal speech.

5.2.2 Results and Discussion - TNL Picture Description Task

Given the success of the Neural Linguistic Feature Approach in automatically scoring the story retelling responses, we adapt it for use in scoring the picture description task responses. To generate the language model soft labels used in the NLF approach, we experimented with the four language models: 1) BERT [118], which performed best in the story retelling task scoring, 2) ALBERT [123], which has shown good performance in the sentence reordering pre-training task and may consequently have the ability to deal with misordered phrases that can appear in spontaneous speech 3) DistilBERT [124] which is trained by distilling the BERT model into a smaller, more computationally efficient model, and XLNET [120], whose autoregressive structure has proven advantageous over BERT in several text classification tasks. 75% of the GSU Kids Speech was used for training and the other 25% for testing. We performed a 4-fold split to ensure that all data was used in testing and report the average performance over all folds. In addition to the GSU Kids dataset, we also jointly train the system with text from the VHED dataset [125] to augment the size of training corpus. This dataset contains image captions corresponding to sequences of images which form short stories. Each set of captions is also labeled by human annotators with an average quality ranking of the overall story on a scale of 1 to 5. The VHED dataset is a composite of multiple audio-visual story-telling datasets with a wide representation of story tasks. Our intention in using it here is to implicitly teach the model to score story quality in addition to being trained on in-domain data from the GSU Kids training set. We use 80% of the VHED dataset (roughly 10,000 text samples) in training the language model to simultaneously predict cumulative scores of the GSU Kids story-telling assessment and average quality rankings of the VHED text samples. We show the classification accuracy to demonstrate overall system performance, F1-score to demonstrate fair performance across imbalanced classes, and root mean square error to show the magnitude of the machine's errors in Table 5.5.

Our results show that the proposed multitask training scheme used with BERT achieves

high accuracy in predicting the overall scores in the GSU Kids story-telling samples. The next sentence prediction training objective and larger parameter size of BERT may contribute to its improved performance over the other models. We note that ALBERT, the model with the fewest parameters, is least robust to an increase in WER in the input transcripts. While transcriptions for children's speech generated from ASR systems still contain several errors, we demonstrate that the usage of language models helps extract high level linguistic features in spite of the high WER of these systems. In the near future, we will train the system to predict individual score components in order to return a detailed score report for each system. For example, the annotators have marked whether or not the child included character names in their story, if they have described key parts of the scene in the picture, and if they keep the same verb tense throughout their story. The promising results in Table 5.5 imply that, given the annotator labels, the system can be trained to predict these characteristics of the student response individually. Teachers can use use such results to understand which areas a student needs to improve in.

5.3 Summary

This section details our frameworks for automatic oral language assessment scoring. We demonstrate the performance of the systems for both a story-retelling task and a picture description task. Using state-of-the-art ASR and downstream BERT-based classification, the systems perform well in categorizing the participants' performance across tasks. We also show that the system performs relatively fairly across the participants' ability levels and spoken dialect.

Demographic			String-Search Rubric (SSRM) Approach				Neural Linguistic Feature (NLF) Approach			
Reading/Language Impairment	WER	# of students	GT RMSE	GT C. Acc	Whisper RMSE	Whisper C. Acc	GT RMSE	GT C. Acc	Whisper RMSE	Whisper C. Acc
Control	21.1%	32	0.125	87.5%	0.125	87.5%	0.020	99.9%	0.025	99.9%
RD only	26.8%	60	0.116	88.3%	0.133	86.6%	0.061	87.5%	0.064	87.5%
RD+LI	34.5%	27	0.111	88.8%	0.148	85.0%	0.110	87.8%	0.130	85.0%
Reading Status	WER	# of students	GT RMSE	GT C. Acc	Whisper RMSE	Whisper C. Acc	GT RMSE	GT C. Acc	Whisper RMSE	Whisper C. Acc
Poor	27.0%	142	0.119	88.0%	0.140	85.9%	0.022	95.0%	0.033	89.5%
Good	21.1%	32	0.125	87.5%	0.125	87.5%	0.086	97.5%	0.094	96.9%
Dialect	WER	# of students	GT RMSE	GT C. Acc	Whisper RMSE	Whisper C. Acc	GT RMSE	GT C. Acc	Whisper RMSE	Whisper C. Acc
AAE	26.8%	116	0.119	87.9%	0.155	84.4%	0.062	94.0%	0.061	94.1%
Non-AAE	25.4%	68	0.108	88.2%	0.102	89.7%	0.013	99.2%	0.013	99.2%

Table 5.4: Results for both the SSRM and NLF approaches across different student demographics. We present a breakdown of best performing ASR system (Whisper) word error rate, the classification C. Accuracy and RMSE of the system on the ground truth (GT) transcripts, and those metrics on the Whisper ASR transcripts for the following three demographic splits: 1) Type of Reading or Language Impairment from i) control - no impairment, ii) RD Only- student has reading disorder like dyslexia that does not occur with or as a secondary effect of a primary learning or language impairment or other condition, iii) RD + LI - A reading disorder that occurs with a primary Language impairment 2) Reading status from i) Poor - the student is evaluated to read at a level below their appropriate grade level or ii) Good - the student reads at or above their appropriate grade level, and 3) Dialect from i) African American English (AAE) or ii) Non-AAE - a mix of characteristics of General American English and Southern American English native to the Atlanta, Georgia Area. Note that the number of students in the Reading/Language Impairment and Reading Status demographic categories do not sum to the full 184. For this analysis, we excluded children with other disorders like ADHD that may complicate the test taking and children who were not able to be assessed for reading status into either the Poor or Good category.

Model (Size)		BERT (110M)			ALBERT (11M)			DistilBERT (66M)			XLNET (110M)		
Metric	%WER	C. Acc	F1	RMSE	C. Acc	F1	RMSE	C. Acc	F1	RMSE	C. Acc	F1	RMSE
Groundtruth	-	98.0	97.5	0.06	95.5	93.0	0.09	84.0	83.0	0.3	92.0	90.0	0.072
Whisper-Large	22.4	96.5	95.0	0.067	95.5	93.0	0.1	84.0	82.5	0.44	91.3	91.0	0.12
HuBERT-Large	33.5	96.0	96.0	0.12	87.5	85.0	0.22	83.0	83.0	0.27	91.0	90.0	0.16

Table 5.5: Percent Classification Accuracy (C. Acc), Percent F1 Score, and Root Mean Square Error of each language model in predicting student scores from the input transcripts (ground truth, Whisper ASR transcript, or HuBERT asr transcripts) along with the word error rate (WER) for each.

CHAPTER 6

Inclusive Automatic Spoken Question Answering

In the previous chapter, we introduced speech-robust NLP methods for cumulative oral assessment scoring. However, an educator assessing a student's oral response may want to provide more detailed feedback to a, such as which target words from a story retelling task were missed or whether or not a student correctly described a character's appearance in a picture description task. In order to provide this more fine-grain feedback, a system would need to retrieve target pieces of information from the input context. This information retrieval task has been explored in spoken question answering, as in [69, 70, 71, 72]. However, several challenges remain in creating robust spoken question answering and information retrieval systems. First, much of the work done in spoken question answering is evaluated on datasets such as the Spoken SQuAD dataset [75] or Spoken CoQA dataset [73]. These datasets often only contain spoken questions and contexts that were either generated using text-to-speech or read from a script created from an existing text question answering dataset. This means that further work may be necessary to create spoken language understanding systems that are robust to the disfluencies and lack of proper logical organization often found in spontaneous speech [76]. Second, many of these works format the problem of spoken question answering as finding an answer from a short context (e.g. a one minute audio recording). Many contexts (e.g. a lecture, an instructional video, or a meeting recording) may be significantly longer, and it is non-trivial to scale a model trained for short contexts to infer answers from a longer context. Last, further work is needed to ensure that these systems are robust to differences in dialect, accent, speaking style, and regional diction or other out of vocabulary words. This may be especially true for cascade systems employing

pre-trained models that were trained on only one dialect or speaking style.

In this work, we aim to advance methods for spoken question answering from long contexts on spontaneous speech. We introduce the CORAAL Question Answering (CORAAL QA) dataset which is composed of hand-labeled question-answer-span pairs about information present in long audio files (typically 30min-1hr) from the Corpus of Regional African American Language [77]. Next, we train a model to rank the relevance of segments of ASR transcripts of a long audio file to an input query and return the most likely span to contain a corresponding answer. Finally, we leverage large language models to generate new questions for data augmentation and further process the returned outputs to improve performance.

6.1 Methods

This work uses spoken question answer pairs from the CORAAL_QA dataset. We designed this dataset with the intention of creating a spoken question answering dataset benchmark dataset with 1) speakers of diverse regional dialects such as AAE, 2) spontaneous speech, and 3) long audio files whose context far exceeds just the span of the answer segment. We format the question answering from long audio files as the following information retrieval task: an audio file, D, is composed of several short segments, $D = \{s_1, s_2, s_3, ..., s_n\}$. The user inputs a query, q, whose answer, a, can be found in one or more consecutive segments, i.e. $a = \{s_i, ..., s_{i+k}\}$. Given the input query, q and audio file, D, which may be up to an hour or more in length, we then seek to return the set of answer segment, a. We accomplish this by assigning a score to each segment in D based on its likelihood of answering q and return the segments with the highest scores. For simplicity, we do not consider queries that can not be answered by any segment in the audio file.

6.1.1 Model

An overview of the framework is given in Figure 6.1. The input audio file is first divided into short segments with an overlap between them. To identify the ideal segment size and overlap, we validate over several choices (shown in Table 6.2) and arrive at an ideal segment size of 60 sec and overlap size of 20 sec. The input audio segments are then transcribed using the Whisper-Large [10] ASR model. Prior work shows that Whisper achieves state-of-the-art performance for the African American English contained in the CORAAL database (16.2% WER) [10, 126]. From the ASR transcript generated from each audio segment, we then use Sentence-BERT [127] to compute a sentence embedding. BERT-based sentence embeddings have been shown to be useful for separating information by topic relevance in text information retrieval tasks [128], and so we seek to apply those benefits in the spoken domain. Inspired by the popular speaker embedding approach of [33], we train a Probabilistic Linear Discriminant Analysis (PLDA) [129] model to score the relevance between the 384-dim BERT sentence embeddings from an input query and the BERT sentence embeddings from the segment-level ASR transcripts. During training, embeddings of text questions from the training set and ASR transcripts from the corresponding answer segments are labeled as coming from the same distribution. During inference, we then use embeddings of target questions as the enrollment set and embeddings of segment-level ASR transcripts as the test set. We then evaluate the system performance by the equal error rate (EER) as well as the precision, recall, and F1-score in correctly retrieving the relevant audio segments. For calculating precision and recall, the PLDA scores for each segment are converted to binary detection decisions through thresholding at the equal error rate. We then compare these scores to the ground truth segment-level labels.

6.1.2 Experiments

We use the VLD split of CORAAL as testing data and the other splits of CORAAL as training and validation splits. This creates an approximately 80%, 10%, 10% split. Splitting data in this way ensures that no speaker appears in more than one split, and that the regional diction and dialect from the test set has not been previously seen by the model. We first establish the performance of the baseline model with this test-train split in Table 1, validating over different input audio segment lengths. Then, we experiment with two methods designed to improve the model performance: Data augmentation and Prompt Engineering. Inference for all models utilized for these tasks is conducted on a single Nvidia A6000 GPU.

6.1.2.1 Data Augmentation with Question Generation

In order to improve model performance, we investigate using large language models to generate more diverse training data. In addition to the hand-written questions of the training set, we also use question generation with DeBERTa [130], ChatGPT (gpt-3.5-turbo) [8], and Llama 2-7b [131] to generate additional training questions. Each language model is given the Whisper ASR transcript from each 60sec segment of each audio file in the training set. Then, using each segment-level transcript as context, the language models are prompted to generate a question with an extractive answer. This question and corresponding time frame from the audio are then used as additional training data. In order to evaluate the quality of the generated questions, we generate questions from the same context as the hand-written questions and compare them with the following metrics: **Semantic Similarity**: the cosine distance between the BERT sentence embeddings of the hand-written question and the generated question. **Percent Words Shared**: The number of words that the generated and hand-written questions have in common after lemmatization and removal of stop words divided by the number of words in the hand-written question. **BLEU Score**: As the BLEU score is a commonly accepted metric for the quality of a machine-generated sentence with

respect to a human-written sentence, we report the BLEU of the generated question with respect to the hand-written question. **Percent entities included**: We report the number of named entities that appear in the generated question divided by the number of entities that appeared in the context from which the question was generated. We perform this both using the ground truth transcript and ASR transcript as context. This metric gives a measure of the language model's ability to ask questions about specific names, dates, locations, and other named entities as well as the ASR system's ability to correctly transcribe these entities before they are passed to the language model as context. **Answer Precision, Recall, and F1 score**: We first ask a RoBERTa model optimised on the SQuAD dataset [132] to extractively answer both the hand-written and the generated question from the ASR transcript. We then score the precision, recall, and F1-score of the retrieved answer of the generated question with respect to that of the hand-written question using the SQuaD evaluation script [133]. This gives a measure of similarity between the answers to the generated questions and the answers to the hand-written question. **Distractor Accuracy**: For each question generated by each language model, we utilise the MQAG framework [134] to generate a correct answer to the question as well as three incorrect distractor answers. We then ask ChatGPT to answer the four-choice multiple response question from the created answers and report the accuracy. We perform this with distractors both generated from the ground truth transcript and from the ASR transcript. **Answerable score**: We use SelfCheckGPT [135] to derive a score for how answerable a given generated question is given the context. This will ideally return a low score if the generated question contains several errors or does not correspond to the given input context. We perform this using both the ground truth and ASR transcripts as input context. These metrics are shown in Table 2. The performance of the data augmentation experiments is shown in Table 3.

For question generation with Llama 2 and ChatGPT, we elect to feed the models with the following prompt:

You are a Teacher. Your task is to setup a question for an upcoming quiz. The question should be simple in nature. Restrict the question to the context information provided
TRANSCRIPT FROM AUDIO SEGMENT

In total we generated 7347 questions each using ChatGPT and DeBERTa, and 7230 questions from Llama 2, resulting in a combined total of 21924 augmented questions. The discrepancy in the number of questions generated from Llama 2 arises due to safeguards placed in the model that lead to a refusal to generate questions from certain segments covering sensitive topics.

For predicting the right answer to a question from a set of distractors, we feed ChatGPT with the following prompt:

You are a Student. Your task is to select the correct answer in a test. You are provided with some context information, a question and some multiple choice options. Answer the question only with the context information provided.Return only the correct option.
TRANSCRIPT FROM AUDIO SEGMENT
Question is below
GENERATED QUESTION
Options are below
A: **OPTION 1** *B:* **OPTION 2** *C:* **OPTION 3** *D:* **OPTION 4**

6.1.2.2 Whisper Prompting

Whisper is powerful in its ability to provide previous context to the decoder in order to improve transcription, and this has led to significant improvements in word error rate for zero-shot spoken language tasks [136]. In this work, we apply Whisper's prompting to take advantage of the temporal relationship of audio segments upon being input into the classifier model. We try prompting Whisper with the concatenation of ASR transcripts from the last N segments upon transcribing the current segment. This is intended to ensure that segments

Chunk Size \(Overlap)	Precision ↑	Recall ↑	Macro F1 ↑	EER ↓
15s \(5s)	0.666	0.567	0.59	0.244
30s \(10s)	0.748	0.622	0.654	0.203
60s \(20s)	0.712	0.631	0.655	0.181

Table 6.1: Effect of the size of the Audio Segments for predictions from the PLDA model. Precision, Recall and Macro F1 statistics are calculated from predicted scores from the system. EER refers to the Equal Error Rate of the trained system

are transcribed with previous knowledge of the conversation and that important information is preserved and consistently transcribed over time. We experiment with using the last N segments as context in the prompt for $N = 1, 2$, and 3 (shown in Table 4).

6.2 Results

Table 6.2 shows the performance of the model with input segments of varying length and overlap. We determine that the question answering model performed best (in terms of both F1 score and EER) with input segments that were 60 seconds long with 20 seconds of overlap between adjacent segments, and so we keep these parameters throughout the rest of this paper. Table 6.2 gives the evaluation metrics for the questions generated by DeBERTa, ChatGPT, and Llama 2 with respect to the hand-written questions. Table **??** reports the performance of the question answering model using generated questions from each language model as training data. Table 6.2 reports the performance of the question answering model when using N segments of previous context as the prompt to Whisper in the current step.

6.3 Discussion

As all the language models used in this work are trained on text data, their metrics in generating data for a spontaneous speech task are expected to be lower than if evaluated

on written words. However, ChatGPT seems to be more robust to the differences in spoken speech vs written text style than DeBERTa and Llama 2. We observe from Table 2 that ChatGPT often produces questions most in line with the hand-written questions, having the highest semantic similarity, number of shared words, and BLEU score with the human-generated samples. ChatGPT's questions also had higher distractor accuracy, though we acknowledge that there may be bias in evaluating its performance, as we predict distractor scores through a secondary prompt to ChatGPT itself. The questions generated by DeBERTa give a significantly lower answerable score than those from either ChatGPT or Llama 2. However, when asked to answer both its own generated question and a hand-written question derived from the same audio segment, DeBERTa gives higher answer precision, recall, and F1-score than Llama 2. However, the number of named entities that Llama 2 included in its questions generated from the both the ground truth transcript and the ASR transcript was significantly higher than that from DeBERTa and not significantly different from that from ChatGPT.

Although using generated questions from ChatGPT in training improved the performance (in terms of precision, recall, and F1 score) over the baseline without data augmentation, we see that additional training data generated by DeBERTa and Llama 2 was also beneficial. This may imply that using artificially generated questions in the question-answering system is beneficial regardless of the generative model used, although some models may produce more human-like or diverse sets of training samples. It also appears that the semantic similarity and Whisper answerable score metrics of the generated questions correlate relatively well to the precision, recall, and F1 score of the question answering model trained on that synthetic data. These metrics may be useful for data mining or automatic quality analysis of generated data in the future. Next, the combination of questions generated by DeBERTa, ChatGPT, and Llama 2 together in data augmentation gives a larger benefit than the use of generated questions from any one model. This result may mean that having a combination of questions from different models, (i.e. a more diverse augmented training set) is more important than

the quality of questions generated by any one model.

The model appears to benefit from the additional context in the prompt. The system shows the highest F1 score when using $N = 3$ segments of previous context, implying that the model performs better when given more context. However, the increase in performance from $N = 2$ to $N = 3$ is marginal, and the benefits gained would likely be outweighed by the additional memory cost if significantly more audio segments were used.

6.4 Summary

This chapter shows our work in automatic spoken question answering on the CORAAL QA database introduced in here. We expand on existing ASR and NLP techniques to extract semantic information from spontaneously spoken African American English speech. The cascaded system of Whisper ASR, SentenceBERT neural embeddings, and PLDA scoring achieves state-of-the-art precision and recall when identifying an answer span from a long audio segment.

Part of this work was accepted to ICASSP 2023.

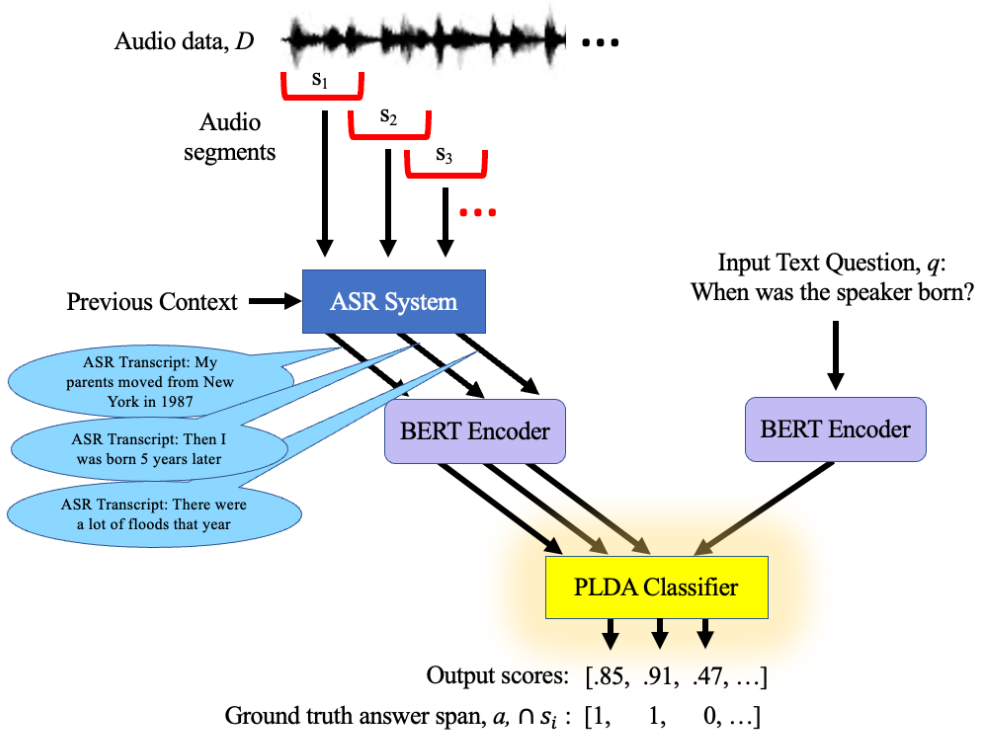

Figure 6.1: Overview of the proposed system. The long audio file, D is segmented into one minute segments, s_i. Each segment is then transcribed with ASR where the ASR system is prompted with previous context. Then both the ASR transcript from each segment and the text of an input query, q, are encoded with Sentence-BERT and scored for the likelihood that s_i answers q by the PLDA classifier. Last, the ground truth scores are used to evaluate performance.

Model	Semantic Similarity ↑	Percent Words Shared ↑	BLEU ↑	GT % Entities Included ↑	Whisper % Entities Included ↑	Answer (precision/ recall/ f1) ↑	GT Distractor Acc ↑	Whisper Distracto Acc ↑
DeBERTa	0.3914	28.28	0.063	20.07	14.52	37.1 /35.4/34.2	72.97	71.86
ChatGPT	0.5670	43.17	0.065	43.19	20.49	41.3/ 40.8/ 39.2	73.43	72.97
Llama 2	0.4751	38.85	0.054	39.61	28.42	30.0/ 29.9/ 28.4	65.28	64.46

Table 6.2: Metrics for evaluating the quality of generated questions: Cosine distance between BERT embeddings of the generated questions and hand-written questions (Semantic Similarity), Percent Words shared between the generated questions and hand-written questions, BLEU score between the generated questions and hand-written questions, % of named entities from the ground truth transcript not included in the generated question (GT % entities included), % of named entities from the ASR transcript included in the generated question (Whisper % entities included)), Precision, Recall, and F1-score between the retrieved answer to the hand-written question and that for the generated question (Answer precision, recall, and F1-score), language model accuracy in correctly answering the question from a multiple choice set with distractors generated from the ground truth transcript (GT Distractor Acc) and distractors generated from the ASR transcript (Whisper Distractor Acc), and the Answerabile score given by SelfCheckGPT for the generated question with either the ground truth transcript or the ASR transcript given as context (GT Answerable Score and Whisper Answerable score, respectively).

Model	**Precision ↑**	**Recall ↑**	**Macro F1 ↑**	**EER ↓**
Deberta	0.732	0.64	0.667	0.183
ChatGPT	0.76	0.66	0.688	0.175
Llama 2	0.748	0.654	0.681	0.164
All	0.765	0.668	0.697	0.166

Table 6.3: Performance of PLDA systems trained with questions generated by different systems. Questions were generated by the respective systems from the non-prompted Whisper generated ASR transcripts. All refers to a PLDA model trained by combining the questions generated from all the individual models

#Seg	**Precision ↑**	**Recall ↑**	**Macro F1 ↑**	**EER ↓**
$N = 1$	0.728	0.641	0.680	0.175
$N = 2$	0.759	0.658	0.688	0.174
$N = 3$	0.761	0.663	0.689	0.175

Table 6.4: Performance of the question answering model when the ASR transcripts from the previous N segments are used in the Whisper prompt as previous context on transcribing the current audio segment.

CHAPTER 7

Summary and Conclusions

This dissertation examined novel methods for fair speech recognition and understanding systems with applications towards educational technologies. We primarily focus on developing speech technology for speakers of African American English because of the pressing social need to improve educational outcomes for such groups that are underrepresented in the technology sector. The chapters of this work can be taken as the steps in a pipeline which 1) performs dialect analysis on an input utterance to better inform downstream processing, 2) perform dialect or age-specific speech recognition on the input utterance, and 3) perform spoken language understanding on the ASR transcripts of the utterance for use in educational settings.

7.1 Summary

Chapter 2 provided a summary of the main databases used in this work including CORAAL, the CORAAL QA Spoken Question Answering dataset, the UCLA JIBO Kids' Speech Corpus, and the GSU Kids' Speech Corpus. Notably, with the exception of the publicly available CORAAL dataset, the other datasets mentioned were largely collected or compiled by the authors of this work and their collaborators. Other publicly available databases used in this work include the Librispeech database [84] and the Speakers in the Wild database [99].

Chapter 3 examined how dialect identification and dialect density estimation could be performed on speech from a low-resource dialect by incorporating linguistic knowledge. We

used ASR, acoustic, and language models to extract features relating to documented linguistic phenomenon in AAE such as formant shifts, alternate pronunciations, grammar constructions, and prosodic patterns. We then achieved state of the art performance in using these features to distinguish the speaker's dialect. We also defined and set the benchmark for the task of dialect density estimation where we correlated the features to the speaker's frequecny of dialect usage.

Chapter 4 proposed the novel ASR data augmentation method, LPC Augment, to synthesize training data for low-resource dialect ASR. The LPC augment algorithm used LPC analysis to decouple the speech source signal from the vocal tract filter, perturb the filter pole locations to better match those that might be representative of formants in the target low-resource dialect, and reconstructed a speech signal with those formant locations. We saw marked improvement over baseline data augmentation methods in training either a transformer-based or a hybrid HMM-DNN system with this augmented data.

Chapter 5 proposed a novel framework for automatically scoring children's oral narrative language abilities. By training a system which predicted student oral assessment scores from a combination of input BERT soft labels with hand-crafted linguistic features extracted from the ASR transcripts, we were able to show high performance in the education task. Furthermore, we showed relatively robust performance across the speakers' dialects, reading ability, and presence of language impairment, indicating a high potential for fair and inclusive speech technology in the space.

Chapter 6 explained the construction of the new CORAAL QA database and describes a PLDA-based spoken question answering system for answering questions from dialectal spontaneous speech in long audio files. By using large language models to generate data for data augmentation and experimenting with the ASR input context, we improved over the baseline to create a high-performing spoken question answering system for the open-domain spontaneous speech task.

7.2 Novel Contributions

Some of the novel contributions of this dissertation work include:

- The creation and establishing of a benchmark performance for the automatic dialect density estimation task
- The proposal of an AAE dialect identification system for both children and adults
- The LPCAugment data augmentation method for low-resource children's dialectal speech
- State-of-the-art performance in children's oral narrative language assessment evaluation
- The creation of the CORAAL QA database for open-domain spontaneous speech question answering from long audio files

7.3 Ethics Statement

The work presented in this dissertation is intended to be used only in service of educators, students, and stakeholders in education. It is not intended to be used to discriminate, violate privacy, or otherwise cause harm. We acknowledge that the proposed systems are trained with sensitive data such as voice bio-metric markers, demographic information, and educational testing scores which could be used to uncover protected characteristics of participants. Malicious users may attempt to use this information to attack or discriminate against vulnerable community members. To prevent this, we propose the following recommendations: 1) This technology is to be used in conjunction with human evaluation and should not replace human intervention entirely. For example, the machine learning-based system may provide an initial evaluation of a student's oral language assessment abilities, and then a human language specialist would verify the results before further action is decided. 2) All

users of the system should be made fully aware of any risks involved. The system will not be employed for any user before they have given informed consent, 3) The system should not be used outside of educational purposes without the consent of all users and stakeholders, and 4) As little sensitive data as possible should be retained after training and inferring with the system.

7.4 Future Work

A question that remains unexplored in some of this work is how well these models generalize to other low resource cases. For example, it would be interesting to examine how well the data augmentation methods we propose for ASR and spoken question answering on African American English speech transfer to other low resource dialects, accents, languages, and speech from speakers with speech-related disabilities like dysarthria. It also remains unseen how well the proposed spoken language understanding systems generalize to applications with specialized vocabulary or pronunciations, such as education for medical, legal, or business fields. A future step is to evaluate these methods on other low resource domains.

In addition, scalability is another important consideration for low-resource speech systems. For example, the fact that a system performed relatively well with a small neural network or when trained on a small number of hours of data does not guarantee that performance will scale linearly with the number of parameters used or amount of training data provided. We hypothesize that our proposed low-resource systems would perform significantly better if given significantly more data or scaled to the size of current large language models. However, the extent of this effect has yet to be shown.

REFERENCES

[1] Tuan D Nguyen, Chanh B Lam, and Paul Bruno, "Is there a national teacher shortage? a systematic examination of reports of teacher shortages in the united states," *Annenberg Institute at Brown University*, 2022.

[2] Kelly Farquharson, Michelle Therrien, Andrea Barton-Hulsey, and Ann F. Brandt, "How to recruit, support, and retain speech-language pathologists in public schools," *Journal of School Leadership*, vol. 32, no. 3, pp. 225–245, 2022.

[3] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel, "Racial Disparities in Automated Speech Recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[4] Satwik Dutta, Sarah Anne Tao, Jacob C Reyna, Rebecca Elizabeth Hacker, Dwight W Irvin, Jay F Buzhardt, and John HL Hansen, "Challenges remain in building asr for spontaneous preschool children speech in naturalistic educational environments," *ISCA INTERSPEECH-2022*, 2022.

[5] Gary Yeung and Abeer Alwan, "On the difficulties of automatic speech recognition for kindergarten-aged children," 09 2018, pp. 1661–1665.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019, pp. 4171–4186, Association for Computational Linguistics.

[8] OpenAI, "Gpt-4 technical report," 2023.

[9] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 12449–12460, Curran Associates, Inc.

[10] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[11] Sonja Lanehart, Ayesha M Malik, and SL Lanehart, "Language use in african american communities," in *The Oxford handbook of African American Language*, pp. 1–19. Oxford University Press, 2015.

[12] Janneke Van Hofwegen, "The development of african american english through childhood and adolescence," *Oxford handbook of African American language*, vol. 454, pp. 474, 2015.

[13] Donald Winford, "85The Origins of African American Vernacular English: Beginnings," in *The Oxford Handbook of African American Language*. Oxford University Press, 07 2015.

[14] Erik R. Thomas and Guy Bailey, "403Segmental Phonology of African American English," in *The Oxford Handbook of African American Language*. Oxford University Press, 07 2015.

[15] William A. Kretzschmar, "219African American Voices in Atlanta," in *The Oxford Handbook of African American Language*. Oxford University Press, 07 2015.

[16] Lisa J. Green and Walter Sistrunk, "355Syntax and Semantics in African American English," in *The Oxford Handbook of African American Language*. Oxford University Press, 07 2015.

[17] Janneke Van Hofwegen, "454The Development of African American English through Childhood and Adolescence," in *The Oxford Handbook of African American Language*. Oxford University Press, 07 2015.

[18] Erik R. Thomas, "420Prosodic Features of African American English," in *The Oxford Handbook of African American Language*. Oxford University Press, 07 2015.

[19] Julie A Washington, Lee Branum-Martin, Congying Sun, and Ryan Lee-James, "The impact of dialect density on the growth of language and reading in african american children," *Language, speech, and hearing services in schools*, vol. 49, no. 2, pp. 232–247, 2018.

[20] Janet L McDonald and Janna B Oetting, "Nonword repetition across two dialects of english: Effects of specific language impairment and nonmainstream form density," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 5, pp. 1381–1391, 2019.

[21] Julie A Washington and Holly K Craig, "Dialectal forms during discourse of poor, urban, african american preschoolers," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 4, pp. 816–823, 1994.

[22] Janna B Oetting and Sonja Pruitt, "Southern african-american english use across groups," *Journal of Multilingual Communication Disorders*, vol. 3, no. 2, pp. 136–144, 2005.

[23] Janna B Oetting, "Some similarities and differences between african american english and southern white english in children," 2015.

[24] Salikoko S Mufwene, John R Rickford, Guy Bailey, and John Baugh, *African-American English: structure, history, and use*, Routledge, 2021.

[25] Angela Creese and Adrian Blackledge, "Translanguaging and identity in educational settings," *Annual review of applied linguistics*, vol. 35, pp. 20–35, 2015.

[26] S. Das, D. Nix, and M. Picheny, "Improvements in children's speech recognition performance," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, 1998, vol. 1, pp. 433–436 vol.1.

[27] Laura L Koenig, Jorge C Lucero, and Elizabeth Perlman, "Speech production variability in fricatives of children and adults: Results of functional data analysis," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3158–3170, 2008.

[28] 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.

[29] Alexandros Potamianos, Shrikanth Narayanan, and Sungbok Lee, "Automatic speech recognition for children," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[30] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli, "Scaling speech technology to 1,000+ languages," *arXiv*, 2023.

[31] AWS, "What is amazon transcribe?," *Amazon Web Services*, 2023.

[32] Chao Liao, Jinwen Huang, Huan Yuan, Peng Yao, Jianchao Tan, Dawei Zhang, Feng Deng, Xiaorui Wang, and Chengru Song, "Dynamic tf-tdnn: Dynamic time delay neural network based on temporal-frequency attention for dialect recognition," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[33] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[34] Moakala Tzudir, Shikha Baghel, Priyankoo Sarmah, and S. R. Mahadeva Prasanna, "Under-resourced dialect identification in Ao using source information," *The Journal of the Acoustical Society of America*, vol. 152, no. 3, pp. 1755–1766, 09 2022.

[35] Aditya Yadavalli, Ganesh Mirishkar, and Anil Kumar Vuppala, "Multi-task end-to-end model for telugu dialect and speech recognition," in *Proc. Interspeech*, 2022, pp. 1387–1391.

[36] Suwon Shon, Ahmed Ali, and James Glass, "Domain attentive fusion for end-to-end dialect identification with unknown target domain," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5951–5955.

[37] Raphaël Duroselle, Md. Sahidullah, Denis Jouvet, and Irina Illina, "Modeling and Training Strategies for Language Recognition Systems," in *Proc. Interspeech 2021*, 2021, pp. 1494–1498.

[38] Marilyn Martin-Jones, "Code-switching in the classroom: Two decades of research," *One speaker, two languages: Cross-disciplinary perspectives on code-switching*, pp. 90–111, 1995.

[39] Ayu Mawadda Warohma, Puspa Kurniasari, Suci Dwijayanti, Irmawan, and Bhakti Yudho Suprapto, "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*, 2018, pp. 522–527.

[40] Ahmed Ali, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James Glass, Steve Renals, and Khalid Choukri, "The mgb-5 challenge: Recognition and dialect identification of dialectal arabic speech," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 1026–1033.

[41] Yun Lei and John H. L. Hansen, "Dialect classification via text-independent training and testing for arabic, spanish, and chinese," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 85–96, 2011.

[42] Moakala Tzudir, Priyankoo Sarmah, and S Prasanna, "Prosodic information in dialect identification of a tonal language: The case of ao," in *Interspeech*, 09 2022, pp. 2238–2242.

[43] Marc A Zissman and Kay M Berkling, "Automatic language identification," *Speech Communication*, vol. 35, no. 1, pp. 115–124, 2001, MIST.

[44] Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[45] Alexander Johnson, Vishwas M. Shetty, Mari Ostendorf, and Abeer Alwan, "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)*, 2023, pp. 1–5.

[46] Google, "Read along by google," Accessed May 2023.

[47] LR Rabiner, B-H Juang, and C-H Lee, "An overview of automatic speech recognition," *Automatic speech and speaker recognition: advanced topics*, pp. 1–30, 1996.

[48] Bruce P Bogert, "The quefrency alanysis of time series for echoes: Cepstrum, pseudoautocovariance, cross-cepstrum and saphe cracking," in *Proc. Symposium Time Series Analysis, 1963*, 1963, pp. 209–243.

[49] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[50] Frederick Jelinek, Bernard Merialdo, Salim Roukos, and Martin Strauss, "A dynamic language model for speech recognition," in *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*, 1991.

[51] Navdeep Jaitly and Geoffrey E Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, 2013, vol. 117, p. 21.

[52] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[53] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[54] 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," *IEEE/ACM TASLP*, vol. 29, pp. 3451–3460, oct 2021.

[55] 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.

[56] Jeremy Heng Meng Wong, Huayun Zhang, and Nancy Chen, "Variations of multi-task learning for spoken language assessment," in *Proc. Interspeech 2022*, 2022, pp. 4456–4460.

[57] Ilja Baumann, Dominik Wagner, Sebastian Bayerl, and Tobias Bocklet, "Nonwords pronunciation classification in language development tests for preschool children," *Proc. Interspeech 2022*, 2022.

[58] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 12449–12460, Curran Associates, Inc.

[59] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number CONF.

[60] Yaroslav Getman, Ragheb Al-Ghezi, Katja Voskoboinik, Tamás Grósz, Mikko Kurimo, Giampiero Salvi, Torbjørn Svendsen, and Sofia Strömbergsson, "wav2vec2-based Speech Rating System for Children with Speech Sound Disorder," in *Proc. Interspeech 2022*, 2022, pp. 3618–3622.

[61] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[62] Satwik Dutta, Sarah Anne Tao, Jacob C. Reyna, Rebecca Elizabeth Hacker, Dwight W. Irvin, Jay F. Buzhardt, and John H.L. Hansen, "Challenges remain in Building ASR for Spontaneous Preschool Children Speech in Naturalistic Educational Environments," in *Proc. Interspeech 2022*, 2022, pp. 4322–4326.

[63] Gary Yeung and Abeer Alwan, "On the difficulties of automatic speech recognition for kindergarten-aged children," *Interspeech 2018*, 2018.

[64] Dadi Ramesh and Suresh Kumar Sanampudi, "An automated essay scoring systems: a systematic literature review," *Artificial Intelligence Review*, vol. 55, no. 3, pp. 2495–2527, 2022.

[65] Takumi Shibata and Masaki Uto, "Analytic automated essay scoring based on deep neural networks integrating multidimensional item response theory," in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 2917–2926.

[66] Bruce W. Lee, Yoo Sung Jang, and Jason Lee, "Pushing on text readability assessment: A transformer meets handcrafted linguistic features," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic, Nov. 2021, pp. 10669–10686, Association for Computational Linguistics.

[67] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[68] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al., "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, pp. 9, 2019.

[69] Nuo Chen, Chenyu You, and Yuexian Zou, "Self-Supervised Dialogue Learning for Spoken Conversational Question Answering," in *Proc. Interspeech 2021*, 2021, pp. 231–235.

[70] Suwon Shon, Siddhant Arora, Chyi-Jiunn Lin, Ankita Pasad, Felix Wu, Roshan S Sharma, Wei-Lun Wu, Hung-yi Lee, Karen Livescu, and Shinji Watanabe, "SLUE phase-2: A benchmark suite of diverse spoken language understanding tasks," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, July 2023, pp. 8906–8937, Association for Computational Linguistics.

[71] Yung-Sung Chuang, Chi-Liang Liu, Hung yi Lee, and Lin shan Lee, "SpeechBERT: An Audio-and-Text Jointly Learned Language Model for End-to-End Spoken Question Answering," in *Proc. Interspeech 2020*, 2020, pp. 4168–4172.

[72] Xuxin Cheng, Zhihong Zhu, Ziyu Yao, Hongxiang Li, Yaowei Li, and Yuexian Zou, "GhostT5: Generate More Features with Cheap Operations to Improve Textless Spoken Question Answering," in *Proc. INTERSPEECH 2023*, 2023, pp. 1134–1138.

[73] Chenyu You, Nuo Chen, Fenglin Liu, Shen Ge, Xian Wu, and Yuexian Zou, "End-to-end spoken conversational question answering: Task, dataset and model," in *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, United States, July 2022, pp. 1219–1232, Association for Computational Linguistics.

[74] Guan-Ting Lin, Yung-Sung Chuang, Ho-Lam Chung, Shu wen Yang, Hsuan-Jui Chen, Shuyan Annie Dong, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Lin shan Lee, "DUAL: Discrete Spoken Unit Adaptive Learning for Textless Spoken Question Answering," in *Proc. Interspeech 2022*, 2022, pp. 5165–5169.

[75] Chia-Hsuan Lee, Szu-Lin Wu, Chi-Liang Liu, and Hung yi Lee, "Spoken SQuAD: A Study of Mitigating the Impact of Speech Recognition Errors on Listening Comprehension," in *Proc. Interspeech 2018*, 2018, pp. 3459–3463.

[76] S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori, "Speech-to-text and speech-to-speech summarization of spontaneous speech," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 4, pp. 401–408, 2004.

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

[78] Alexander Johnson, Ruchao Fan, Robin Morris, and Abeer Alwan, "Lpc augment: an lpc-based asr data augmentation algorithm for low and zero-resource children's dialects," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8577–8581.

[79] Evelyn L Fisher, Andrea Barton-Hulsey, Casy Walters, Rose A Sevcik, and Robin Morris, "Executive functioning and narrative language in children with dyslexia," *American journal of speech-language pathology*, vol. 28, no. 3, pp. 1127–1138, 2019.

[80] Alexander Johnson, Alejandra Martin, Marlen Quintero, Alison Bailey, and Abeer Alwan, "Can social robots effectively elicit curiosity in stem topics from k-1 students during oral assessments?," in *2022 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 2022, pp. 1264–1268.

[81] R Goldman, "Goldman-fristoe test of articulation–third edition (gfta-3)," *Circle Pines, MN: AGS*, 2015.

[82] Ronald Bradley Gillam and Nils A Pearson, *Test of narrative language*, Pro-ed, 2017.

[83] Gary Yeung, Alison L. Bailey, Amber Afshan, Morgan Tinkler, Marlen Q. Pérez, Alejandra Martin, Anahit A. Pogossian, Samuel Spaulding, Hae Won Park, Manushaqe Muco, Abeer Alwan, and Cynthia Breazeal, "A robotic interface for the administration of language, literacy, and speech pathology assessments for children," in *Proc. 8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)*, 2019, pp. 41–42.

[84] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[85] Cynthia Puranik, Lee Branum-Martin, and Julie A Washington, "The relation between dialect density and the codevelopment of writing and reading in african american children," *Child Development*, vol. 91, no. 4, pp. e866–e882, 2020.

[86] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf, "Flair: An easy-to-use framework for state-of-the-art nlp," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, 2019, pp. 54–59.

[87] Christopher Cieri, David Miller, and Kevin Walker, "The fisher corpus: A resource for the next generations of speech-to-text.," in *LREC*, 2004, vol. 4, pp. 69–71.

[88] John Hale, "A probabilistic earley parser as a psycholinguistic model," in *Second meeting of the north american chapter of the association for computational linguistics*, 2001.

[89] Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), Vols 1-5*. ISCA, 2016, vol. 8, pp. 2001–2005.

[90] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[91] Trang Tran, Shubham Toshniwal, Mohit Bansal, Kevin Gimpel, Karen Livescu, and Mari Ostendorf, "Parsing speech: a neural approach to integrating lexical and acoustic-prosodic information," *Proc. Conf. North American Chapter Assoc. for Computational Linguistics (NAACL)*, 2018.

[92] Paul Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, pp. 341–345, 2014.

[93] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al., "Xgboost: extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[94] Julie A Washington and Mark S Seidenberg, "Language and dialect of african american children," in *Handbook of Literacy in Diglossia and in Dialectal Contexts: Psycholinguistic, Neurolinguistic, and Educational Perspectives*, pp. 11–32. Springer, 2022.

[95] Sergiu Hart, "Shapley value," in *Game theory*, pp. 210–216. Springer, 1989.

[96] Hynek Hermansky and Nelson Morgan, "Rasta processing of speech," *IEEE transactions on speech and audio processing*, vol. 2, no. 4, pp. 578–589, 1994.

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

[98] 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.

[99] 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.

[100] 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.

[101] 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.

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

[103] 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.

[104] 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.

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

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

[107] 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.

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

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

[110] 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.

[111] Dogan Can, Victor R Martinez, Pavlos Papadopoulos, and Shrikanth S Narayanan, "Pykaldi: A python wrapper for kaldi," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5889–5893.

[112] Yajie Miao, Jinyu Li, Yongqiang Wang, Shi-Xiong Zhang, and Yifan Gong, "Simplifying long short-term memory acoustic models for fast training and decoding," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2284–2288.

[113] Linhao Dong, Shuang Xu, and Bo Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5884–5888.

[114] Roberto Gretter, Marco Matassoni, Daniele Falavigna, A Misra, Chee Wee Leong, Katherine Knill, and Linlin Wang, "Etlt 2021: Shared task on automatic speech recognition for non-native children's speech," in *Proceedings of the Annual Conference of the International Speech Communication Association*. ISCA, 2021, pp. 3845–3849.

[115] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[116] Kenneth Heafield, "KenLM: Faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, July 2011, pp. 187–197, Association for Computational Linguistics.

[117] Wayne Ward, Ron Cole, Daniel Bolaños, Cindy Buchenroth-Martin, Edward Svirsky, and Tim Weston, "My science tutor: A conversational multimedia virtual tutor.," *Journal of Educational Psychology*, vol. 105, no. 4, pp. 1115, 2013.

[118] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio, Eds. 2019, pp. 4171–4186, Association for Computational Linguistics.

[119] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Annual Meeting of the Association for Computational Linguistics*, 2019.

[120] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. 2019, vol. 32, Curran Associates, Inc.

[121] Sowmya Vajjala and Detmar Meurers, "On improving the accuracy of readability classification using insights from second language acquisition," in *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, USA, 2012, NAACL HLT '12, p. 163–173, Association for Computational Linguistics.

[122] Falk Huettig and Martin J. Pickering, "Literacy advantages beyond reading: Prediction of spoken language," *Trends in Cognitive Sciences*, vol. 23, no. 6, pp. 464–475, 2019.

[123] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. 2020, OpenReview.net.

[124] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf, "Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter," *CoRR*, vol. abs/1910.01108, 2019.

[125] Chi-Yang Hsu et al., "Learning to rank visual stories from human ranking data," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 6365–6378.

[126] Alexander Johnson, Hariram Veeramani, Balaji Natarajan, and Abeer Alwan, "An equitable framework for automatically assessing children's oral narrative language abilities," *Proc. Interspeech*, 2023.

[127] Nils Reimers and Iryna Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Nov. 2019, pp. 3982–3992, Association for Computational Linguistics.

[128] Min Pan, Junmei Wang, Jimmy X. Huang, Angela J. Huang, Qi Chen, and Jinguang Chen, "A probabilistic framework for integrating sentence-level semantics via bert into pseudo-relevance feedback," *Information Processing & Management*, vol. 59, no. 1, pp. 102734, 2022.

[129] Sergey Ioffe, "Probabilistic linear discriminant analysis," in *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part IV 9*. Springer, 2006, pp. 531–542.

[130] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen, "Deberta: Decoding-enhanced bert with disentangled attention," in *International Conference on Learning Representations*, 2021.

[131] Hugo Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023.

[132] Xiang Pan, Alex Sheng, David Shimshoni, Aditya Singhal, Sara Rosenthal, and Avirup Sil, "Task transfer and domain adaptation for zero-shot question answering," *CoRR*, vol. abs/2206.06705, 2022.

[133] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, Nov. 2016, pp. 2383–2392, Association for Computational Linguistics.

[134] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales, "MQAG: multiple-choice question answering and generation for assessing information consistency in summarization," *CoRR*, vol. abs/2301.12307, 2023.

[135] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales, "Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models," 2023.

[136] Puyuan Peng, Brian Yan, Shinji Watanabe, and David Harwath, "Prompting the Hidden Talent of Web-Scale Speech Models for Zero-Shot Task Generalization," in *Proc. INTERSPEECH 2023*, 2023, pp. 396–400.