ASHA American Speech-Language-Hearing Association

PERSPECTIVES

SIG 19

Research Article

Demographic and Acoustic Factors Related to Automatic Speech Recognition Inaccuracies for Child African American English Speakers

Brittany N. Fletcher,[a,b] Wei-Wen Hsu,[c] Vesna D. Novak,[d] Mary E. Wilkens,[a] Amy W. Hobek,[a] Amy S. Pratt,[a] Michelle Leon,[e] Kimmerly Harrell,[f] and Victoria S. McKenna[g]

[a]Department of Communication Sciences and Disorders, University of Cincinnati, OH [b]Department of Speech, Language, and Hearing Sciences, Boston University, MA [c]Department of Biostatistics, Health Informatics and Data Sciences, University of Cincinnati, OH [d]Department of Electrical Engineering and Computer Science, University of Cincinnati, OH [e]Department of Speech, Language, and Hearing Sciences, Mount St. Joseph University, Cincinnati, OH [f]Department of Speech-Language Pathology, Albizu University, Miami, FL [g]School of Communication Sciences and Disorders, University of Central Florida, Orlando

ARTICLE INFO

Article History:
Received February 26, 2025
Revision received June 13, 2025
Accepted July 10, 2025

Editor-in-Chief: Kendrea L. (Focht) Garand
Editor: Jason A. Whitfield

https://doi.org/10.1044/2025_PERSP-25-00052

ABSTRACT

Purpose: This study investigated the relationship between acoustic measures and Google's Speech-to-Text inaccuracies in recognizing speech of children ages 4–9 years who speak African American English (AAE).
Method: Audio recordings were collected from 11 AAE-speaking children with speech stimuli targeting final plosive variations observed within the AAE dialect. Dialectal density was measured using the Diagnostic Evaluation of Language Variation Language Screener. Recordings were transcribed using Google's Speech-to-Text application (Google Voice), and inaccuracies were determined through comparison to researcher-extracted transcriptions. Acoustic measures from vowels preceding final plosives (including vowel duration, fundamental frequency, average first formant) were extracted using Praat and a custom MATLAB algorithm. Individual mixed-effects logistic regression models were conducted to analyze the relationships between acoustic measures and transcription accuracy (accurate vs. inaccurate) for voiced and voiceless plosives separately.
Results: There were no significant differences between inaccuracy rates for voiced and voiceless plosive productions, nor were acoustic measures predictive of automatic speech recognition inaccuracy. However, age and dialect density were significantly related to voiceless plosive accuracy.
Conclusions: The complexities of voice, motor, and articulatory development within children can be characterized by acoustic measures. These measures inform acoustic algorithms created for speech technology. Research on acoustic measures in young child AAE speech, with considerations for dialect variability and age, will enhance speech recognition technology and clinical best practices.
Supplemental Material: https://doi.org/10.23641/asha.30150178

Over the past decade, artificial intelligence (AI) systems have been increasingly utilized by children within their homes, communities, and educational systems through tools such as speech-to-text, text-to-speech, and chatbots. Within the educational system, speech-to-text, or voice recognition tools, provide increased accessibility to technology and information for children with learning disabilities, physical disabilities, and vision impairments (National Center for

Correspondence to Brittany N. Fletcher: brittany.fletcher@ucf.edu.
Disclosure: *Brittany Fletcher serves as an editorial board member for Perspectives of the ASHA Special Interest Groups. Brittany Fletcher is a member of the National Black Association for Speech-Language and Hearing. The authors have declared that no other competing financial or nonfinancial interests existed at the time of publication.*

Technology Innovation, 2010). Additionally, advancements within the communication sciences and disorders discipline are opening doors for use of AI as a tool for speech and language assessments (Albudoor & Peña, 2022) and treatment (e.g., Say It Labs, 2023). Despite these advancements, speech-to-text technology, such as Google Home, have been shown to have inherent biases. That is, biases exist for adults from underrepresented communities, such as the African American English (AAE)–speaking community (Koenecke et al., 2020).

Approximately 85 million users utilize Google Home products with an expected increase of approximately 92 million users by 2025 (Keating, 2023). Due to its popularity and accessibility within the home, Google was the speech-to-text tool of focus for this exploratory study. As AI increasingly becomes an accessible tool within the home and for clinicians and educators to utilize, we must gain a deeper understanding of what measures impact speech recognition inaccuracies for nonmainstream American English (non-MAE) speakers to ensure access and equity for underrepresented populations.

AI Systems for Speech

AI speech technology systems involve a complex process to perceive the speakers' commands, interpret the command, and provide an output via information or action. The first step of the process, perceiving the command, is completed through automatic speech recognition (ASR). ASR is a multistep process composed of (a) extracting speech features from the speakers' productions; (b) developing acoustic, pronunciation, and language models; and (c) building machine-learning algorithms to characterize and classify speech and linguistic elements (Du, 2023). Previous studies have found that discrepancies in ASR transcription accuracies for non-MAE speakers are due to the acoustic models, not the language models (Koenecke et al., 2020).

The classic machine learning algorithm for ASR was the hidden Markov model (HMM; Deng & Li, 2013), a type of probabilistic network that associates a sequence of acoustic patterns with specific phones. Essentially, it converts a speech signal to a sequence of shorter signals that are approximately stationary processes, extracts acoustic features from each shorter signal, and then finds the phoneme sequence, depending on the dialect or language being trained, that is most likely given this sequence of acoustic features (Fischer et al., 2006; Hunt, 2006; Waserblat & Eilam, 2014). Spectral information, such as vowel formant frequencies and duration (O'Shaughnessy, 2024), are key features extracted from the acoustic models. These same acoustic features also contribute to the speech perception and the differentiation between dialects (Holt, 2018). To identify the most probable phoneme sequence, the model must be trained with labeled training data—a database of speech signals annotated with corresponding phoneme sequences. Once trained with such data, the model can pick among the possible phoneme sequences in the data set in order to recognize new speech signals. However, while HMMs were the dominant approach for ASR for decades, they have significant well-known limitations (Deng & Li, 2013), and modern approaches propose several alternatives.

One common modern alternative to HMMs is the deep neural network (Nassif et al., 2019). A classic artificial neural network is a model that consists of connected nodes ("neurons"), with each node receiving information from previous nodes, processing it, and sending results to later nodes in the model. Such nodes are aggregated into consecutive layers, and a deep neural network is one with a large number of layers, allowing it to model complex patterns. Early examples of neural networks were feedforward networks (i.e., data can only flow from early layers to later layers), but modern neural networks for ASR are usually recurrent networks (i.e., data can also flow from later layers back to early ones). Like an HMM, a deep neural network is most commonly trained with a database of speech signals annotated with corresponding phoneme sequences, and outputs the most probable phoneme sequence for a new speech signal. Unlike an HMM, a deep neural network makes fewer assumptions about the underlying data and can thus utilize the "raw" acoustic waveform rather than relying on manually crafted acoustic features such as formant frequencies (Radha et al., 2024). Such deep networks can even be used to perform end-to-end ASR: a form of ASR where the different models (acoustic, pronunciation, language) are all learned together (unlike HMMs, which require separate components for each model). However, since such a deep neural network is very complex, it can need a very large amount of data to be properly trained.

Per Google's report, the vision of Google Cloud's speech-to-text application (Zhang et al., 2023) approach is to create a universal ASR model that covers all the spoken languages in the world, and its performance evaluation suggests that it outperforms other "universal" ASR models such as Whisper, a well-known open-source machine learning model (Radford et al., 2022). Nonetheless, studies show that both Chirp (used by Google Cloud) and other ASR models such as Whisper underperform on speech of various marginalized groups such as d/Deaf speakers (Zhao et al., 2025).

Discrepancies in AI for AAE Speakers

AAE is a dialect of American English spoken by many African Americans or people who live within the community where African Americans currently or have

historically lived (Charity Hudley et al., 2018). AAE speakers have experienced disadvantages related to educational and societal resources and opportunities for decades due to systemic racism and bias (Bailey et al., 2017). Even recent advancements in commonly used speech technology continue to have inherent biases that may compound the already existing disadvantages for AAE speakers due to the lack of representative AAE speech within the corpuses of training data (Kendall & Farrington, 2018).

Mengesha et al. (2021) studied adult AAE speakers' experiences with the use of common speech-to-text technology (e.g., Google). It was found that they presented with feelings of discrimination when the ASR system did not understand their speech. In addition to the common errors made by the system, the participants felt they were forced to alter their speech more to the likeness of MAE in order to have more success with system comprehension. As users, they felt a lack of representation in that the system was not initially developed to understand their speech.

Johnson et al. (2023) presented a novel framework to the persistent issue of nonmainstream English dialectal bias within ASR, which is a first step to improving algorithmic approaches for nonmainstream dialects. They developed a model that estimated dialectal density (the percentage of words within short utterances that contain nonmainstream dialectal productions) with the goal of being able to use a dialect density measure to improve ASR. Using the Corpus of Regional African American Language database of adult AAE speakers (Kendall & Farrington, 2024), they investigated several demographic (e.g., geographic location) and acoustic factors (e.g., consonant clusters, voicing, prosody) to characterize dialect density. They used a Rasta-styled filtered auditory spectrum (Hermansky & Morgan, 1994) for speech processing, which captured context-dependent psychoacoustic information useful to ASR. Some of the more impactful features determined to predict AAE phonological tokens within the model included mean rising slope (prosody/intonation measure), final consonant deletions, cluster reductions, and speaker-specific information within geographic regions. Therefore, final consonant features (a focus of the present work) were among some of the key features in dialect density determination within the model. In the future, this model could determine dialect density to then adjust an ASR model's speech identification based on the dialect used and its degree (or density) of use.

AAE Speech and Language Characteristics

AAE is composed of various phonological and morphosyntactic features that make it a distinct dialect (Hamilton et al., 2018; Holt, 2018; Stockman, 2010; Thomas, 2007). The focus of this study solely includes phonological features, due to its impact, more so than morphosyntactic features, on ASR accuracy (Koenecke et al., 2020). There are a variety of phonological features observed in AAE; features such as consonant cluster reduction, syllable addition or deletion, monopthongization (Craig et al., 2003), and voiceless productions of voiced final plosives are of the most common (Thompson et al., 2004). For example, production of a final voiced plosive (i.e., /kæb/) may also be produced as a voiceless plosive (e.g., /kæp/), produced without a plosive (e.g., /kæ/), or produced with glottalization of the voiced plosive (e.g., /kæʔ/). This phonological pattern will be referred to as *voiced final plosive variation* throughout this study, to use noncomparative descriptions of AAE speech (Hamilton et al., 2018).

Within the dialect of AAE, variation can be due to regionality, environmental setting, and other cultural and linguistic influences (Llamas & Watt, 2010). Dialect density provides a way to measure the degree to which a speaker uses a dialect and captures the variation of AAE spoken even within a concentrated regional area. The quantification of dialect use in children typically involves measuring the number of tokens or types of dialect features present in a language sample relative to the total number of utterances in that sample. The lower the density measure, the fewer AAE phonological and morphological patterns the child used within the target utterances. The higher the dialect density measure, the more AAE patterns used. When considering the lower ASR accuracies for adult AAE speakers, dialect density provides information on the degree of dialect variation that impacts the inaccuracies (Johnson et al., 2023; Radha et al., 2024).

Research indicates that dialect density is not static but evolves significantly with age and environmental exposure (Craig et al., 2004; Ivy, & Masterson, 2011; Thompson et al., 2004; Washington et al., 2018). A major influence on this change is exposure to MAE in school settings (Terry et al., 2012). Longitudinal studies reveal a consistent pattern: With each additional year of formal schooling, children tend to exhibit a decrease in dialect density in both oral and written language, relative to their dialect use at the time of school entry (Craig et al., 2009; Puranik et al., 2020; Washington et al., 2018). Therefore, dialect density characterizes the degree of AAE use within a given environment, age, and stimuli.

Acoustic Measurements

Acoustic cues such as vowel spectral measures and voice onset time (VOT) can both be used to quantify perception of voiced final plosive variation across AAE speakers. However, due to the final positioning of the voiced plosive, VOT is often too difficult to visually determine and measure. The silent interval of voiced plosives

tend to be short, and the burst tends to be reduced in amplitude and duration making VOT difficult to identify (Abramson & Whalen, 2017). Vowel duration of the vowel preceding the plosive, as well as average fundamental frequency (*F*0) and the average first formant (*F*1), instead, have been used to provide information to the listener on the intent of the speaker and help to distinguish between voiced and voiceless plosives (Benki, 2001; Lisker & Abramson, 1964; Stevens & Klatt, 1974).

Previous research has established a longer vowel duration for voiced final plosive productions compared to voiceless productions, even when the voiced plosive produced by the speaker is perceived as voiceless by the listener (Snow, 1997). For AAE speakers, there is greater variation in final plosive productions compared to MAE, which may be reflected within vowel duration, formant, and frequency measures. As most ASR systems have utilized mainly MAE speech training data, there remains to be explored if the same acoustic measures related to listener perception in variation of AAE final plosives also informs ASR transcription accuracy.

Discrepancies in AI for Child Speakers With Nonmainstream English Dialects

Classrooms and clinical caseloads are composed of a diverse range of multilingual, multidialectal, and multicultural children. However, phonological, morphological, and anatomical differences in child speech, in comparison to adult speech, has created great challenges in the effective use of speech technology by children (Bhardwaj et al., 2022; Yeung & Alwan, 2018). In addition to dialectal variations occurring across the continuum of AAE, there are various characteristics of child speech that impact the accuracy of speech technology for children. Similar to the situation of reduced accuracy for adult AAE speakers, the initial speech recognition technology was not trained using representative data for children (Du, 2023). Therefore, differences in acoustic patterns between children and adults—such as children having higher formant frequencies and less control over prosodic features—lead to reduced accuracies by the system when transcribing speech based on child acoustic speech patterns (Bhardwaj et al., 2022; Sunil et al., 2016). Hence, there is a need to establish the accuracy of speech technology used not only by young children who speak MAE but also for children who speak a nonmainstream dialect, such as AAE.

Adaptation of neural network models have been applied in attempts to resolve the dilemma of multilingual children utilizing models trained mostly by adult MAE speakers (Lo et al., 2020; Matassoni et al., 2018). Radha et al. (2024) developed goals similar to Johnson et al. (2023) to help improve ASR transcription accuracy of native and nonnative British English preschoolers' speech. To solve the issue of lower transcription accuracies for child speech due to anatomical and linguistic development variation, they utilized pitch harmonic estimation, mel frequency cepstral coefficients (MFCCs), and the *k*-nearest neighbor approach to achieve accent classification. The authors supported similar findings from Krishna et al. (2020), Salau et al. (2020), and Yarra and Ghosh (2018) that frequency, formant, and MFCC metrics showed promise as acoustic targets for refining ASR algorithms and should be included in future investigations. Recent developments in research-based ASR systems trained and specific to children have reduced the inaccuracies to less than 10% (Attia et al., 2024; Yeung & Alwan, 2018). However, advancements related to additional factors influencing accuracy (e.g., age and dialect) still need to be explored. These questions also remain for the more publicly accessible speech recognition tools, including but not limited to Google Voice.

Purpose

With the increased accessibility and opportunities for use of speech technology tools by children, it is critical to investigate the factors, such as acoustic parameters, age, and dialect density related to the transcription accuracies for AAE speakers. This will further contribute to the knowledge of potential bias within AI and how ASR may contribute to educational inequities among children who speak nonmainstream English dialects, including AAE. Considering acoustic models heavily impact ASR's capability to accurately transcribe child AAE speech (Lin et al., 2009), we aimed to understand the relationship between specific acoustic features (e.g., formant frequencies) and ASR's inaccuracy. We specifically investigated the phonological pattern characteristic of AAE, voiced final plosive variations, and its relationship with Google speech-to-text transcription inaccuracies. The following study aims and hypotheses were addressed:

ASR Inaccuracy in Final Voiced Versus Voiceless Plosives

Aim 1: Determine differences in speech-to-text inaccuracy between final position voiced and voiceless plosive productions within child AAE speech, considering age and dialect density as covariates.

Hypothesis 1: AAE is a dialect that has a characteristic of final plosive voicing variation, in which voiced plosives can also be produced as voiceless or glottalized. Therefore, we expected greater ASR transcription inaccuracies for voiced compared to voiceless plosives within child AAE speech. Furthermore, we expect age and dialect density to influence inaccuracies, with younger age and higher dialect density associated with greater inaccuracies in Google speech-to-text.

Acoustic Correlates of ASR Inaccuracy in Final Plosive Productions

Aim 2: Investigate the relationship between acoustic measures (vowel duration, mean *F*0, and *F*1) and ASR inaccuracy in final plosive productions, with additional consideration of covariates of dialectal density and age for child AAE speakers.

Hypothesis 2: Previous literature indicates that acoustic characteristics of vowels preceding plosives vary depending on whether the plosive is voiced or voiceless (Lisker, 1986; Snow, 1997). In our study, we expected voiced plosives that have acoustic characteristics of longer vowel durations, lower mean *F*0, and lower *F*1 formants, to be associated with lower degrees of inaccuracy, as these are the acoustic cues that assist a listener in distinguishing speaker intent. Conversely, we would expect shorter vowel durations and higher mean *F*0 and *F*1 to be observed as the vowel characteristics of voiceless plosives and reduced inaccuracy. Furthermore, we anticipated that younger age and higher dialectal density would further contribute to the greater ASR inaccuracy. This is because ASR systems are typically trained with adult MAE speech patterns and may struggle to accurately transcribe speech that includes a child's speech features and a greater density of dialect-specific features.

Method

The study protocol received approval from the University of Cincinnati Institutional Review Board (IRB No. 2022-1099). Prior to children participating in the study, parental written consent and child written or verbal assent were obtained.

Participants

The participant pool consisted of 11 AAE child speakers (nine girls and two boys; $M = 6.1$ years, range: 4.0–9.3 years) from the northern Kentucky region. Participants met the following inclusion criteria: (a) between the ages of 4;0 and 9;11 (years;months), (b) passed a hearing and speech screening appropriate for age, and (c) native speakers of American English. Exclusion criteria included children who met any of the following conditions: (a) had an Individualized Education Program targeting speech goals, (b) formally diagnosed speech or language disorder, (c) parent-reported hearing impairment or neurological issue, (d) unwell or experienced illness within the week prior to data collection, or (e) were multilingual or did not primarily speak AAE.

Dialect classification was determined based on parent-reported information. Prior to parents completing demographic information, the researcher, who identifies as an AAE speaker, provided an in-depth definition, explanation, and examples of "What is AAE?". First, the researcher asked the question, "How do you identify with Black culture?" and then proceeded to ask, "Do you believe Black culture can be expressed through language and how we talk?". The researcher then prefaced the conversation with the history of the misidentification of AAE based on the term *Ebonics*, how people within the Black community view AAE with a negative connation, followed by a brief history of the development of AAE into what is used today. Syntactic and phonological examples were provided. Parents were then asked, "Do you use some of these patterns (or others) at home, church, and/or with friends but tend to switch your speech when you are at work or in non-predominantly Black spaces?". AAE use was then confirmed in the household by asking, "Do you express your Black culture through your language and how you speak at home and in your communities?". See Supplemental Material S1 for dialect classification protocol.

The Diagnostic Evaluation of Language Variation (DELV) Screener (Seymour et al., 2003) is a standardized language screener that identifies any use of a nonmainstream English dialect. AAE-speaking children were the main population the DELV was normed on. Still, it should be noted that the DELV Screener has been observed to have inconsistencies in accurately evaluating language abilities of young nonmainstream dialect speakers, especially AAE (Marencin et al., 2024; Moland & Oetting, 2021). Therefore, the DELV Screener was used to describe the AAE speakers in the study, but participants were not excluded solely based on their language scores. Instead, dialectal density and variation were calculated using information from the DELV Screener.

The DELV Screener was used to identify AAE features based on phonological (e.g., consonant cluster reduction), morphosyntactic (e.g., invariant "be," zero copula), and syntactic patterns (e.g., multiple negation). Features were cross-referenced with culturally and linguistically representative norms for AAE (Charity, 2008; Green, 2010; Hendricks & Adlof, 2017). To differentiate AAE from maturational variations, responses were compared to age-specific language developmental norms (Crowe & McLeod, 2020a; Wyatt, 1995). Patterns consistent with AAE were those that aligned with cultural and linguistic norms rather than transient developmental errors. For instance, zero copula (e.g., "She tall") was classified as AAE when consistently used by children older than age 3 years, as transient omissions are expected at younger ages. Dialect density numerical and categorized scores were computed from the proportion of identified AAE features across 15 DELV tasks targeting phonological ("I

see her brushing *teeth*"), morphosyntactic ("The boy has little kites, but the girl *have big kites*"), and syntactic dimensions ("The boy always wash the plates, but the girl always *wash the dishes*"). A score of "0%" represented complete use of MAE within speech, and "100%" represented complete use of AAE during speech. The AAE-speaking children in our study scored in a range of 6%–100% for dialect density with an average of 57%. Furthermore, degrees of variation such as "strong variation from MAE," "some variation from MAE," or "no variation from MAE" were assigned to all participants based on the Degree of Language Variation portion of the assessment. It is important to note that older age categories, approximately age 6 years and greater, have a lower dialect density percentage threshold to be categorized as "some variation" versus a "strong variation." See Table 1 for demographic information.

Screening and Background Information

Pure-tone hearing screenings were performed to verify that hearing was within normal limits, defined as 25 dB HL (Ohio Department of Health, 2022) at frequencies of 500, 1000, 2000, and 4000 Hz. Screening to verify typical speech development included the Diagnostic Evaluation of Articulation and Phonology (DEAP) Screener (Dodd et al., 2006) for all children. The DEAP Screener identified any articulation or phonological disorders. Children's data were not included if they had any articulation or phonological patterns inappropriate for the child's age. Parents completed a questionnaire, supplying demographic, speech, language, and cognitive information. A total of 13 AAE children were screened for inclusion. Two children did not pass the articulation and phonology screener; therefore, only 11 met the inclusion criteria.

Data Collection

Data were collected in a quiet location in one of three different environments: a University of Cincinnati Speech and Hearing Clinic treatment room, the child's home environment, or in a quiet office at the child's school or day care. Each environment met criterion for minimal background noise measured via a sound pressure level (SPL) meter (< 50 dB SPL; Patel et al., 2018) with a range of environmental noise of 35.3–47.4 dB SPL.

To collect data, targeted phonological patterns of AAE minimal pairs were created by the first author. The minimal pairs task targeted words using a carrier phrase (see Table 2 for speech stimuli). Children produced a total of 28 targeted voiced final consonant productions at phrase level from a minimal pair's stimuli. The targeted minimal pairs varied on final consonant voicing, for example, "I see a cob" and "I see a cop." These were specifically chosen due to the known characteristics of final voiced plosive variation in AAE. Children were provided with a picture and asked to use the carrier phrase. Modeling was provided if the child could not read the carrier phrase or was not familiar with the target picture. The researcher eliciting the speech stimuli is a bidialectal speaker of MAE and AAE and therefore provided a range of productions of both dialects. Nonimmediate repetition was used when the researcher provided the model of the entire phrase followed by a prompt of "Now can you say that?" to elicit the correct target phrase. This approach is supported by findings that phonetic imitation is selective and influenced by both phonetic and social factors (Trudgill, 2008), highlighting the importance of considering researcher culture and dialect when eliciting naturalistic speech patterns. Nonimmediate repetition cueing was used for six out of the 11 children, where a range from < 1% to 86% of stimuli required cueing.

Table 1. Descriptive statistics of race, ethnicity, age, sex, and dialect density of all participants.

Participant	Race	Ethnicity	Age (years;months)	Sex	Dialect variation from MAE	Dialect density (%)
P_01	Black	Non-Hispanic	8;0	M	None	6
P_02	Black	Non-Hispanic	5;2	F	Strong	87
P_03	Black	Non-Hispanic	9;4	F	Strong	33
P_04	Black	African American	5;8	F	Strong	80
P_05	Black	African American	4;0	M	Strong	60
P_06	Black	African American	5;7	F	Strong	87
P_07	Black and White	African American	6;10	F	Strong	67
P_08	Black and White	African American	4;9	F	Some	47
P_09	Black	African American	7;8	F	None	13
P_10	Black	African American	5;2	F	Strong	100
P_11	Black	African American	5;1	F	Some	47

Note. Dialect density categorization and percentages were scored using the DELV screener, which varied also considering age. MAE = mainstream American English; M = male; F = female.

Table 2. Stimuli list of targeting final plosives within phrases.

Target phrase	Voicing	Vowel	Place of articulation
I see a **bag**	voiced	/æ/	velar
I see her **back**	voiceless	/æ/	velar
I see a **cab**	voiced	/æ/	bilabial
I see a **cap**	voiceless	/æ/	bilabial
I see a **cob**	voiced	/ɑ/	bilabial
I see a **cop**	voiceless	/ɑ/	bilabial
I see a **cub**	voiced	/ʌ/	bilabial
I see a **cup**	voiceless	/ʌ/	bilabial
I see a **dog**	voiced	/ɑ/	velar
I see a **dock**	voiceless	/ɑ/	velar
The cat **hid**	voiced	/ɪ/	alveolar
The girl **hit**	voiceless	/ɪ/	alveolar
I see a **log**	voiced	/ɑ/	velar
I see a **lock**	voiceless	/ɑ/	velar
The man is **mad**	voiced	/æ/	alveolar
I see a **mat**	voiceless	/æ/	alveolar
I see a **mob**	voiced	/ɑ/	bilabial
I see a **mop**	voiceless	/ɑ/	bilabial
I see a **pig**	voiced	/ɪ/	velar
I see the girl **pick**	voiceless	/ɪ/	velar
I see a **pod**	voiced	/ɑ/	alveolar
I see a **pot**	voiceless	/ ɑ/	alveolar
I see a **robe**	voiced	/oʊ/	bilabial
I see a **rope**	voiceless	/oʊ/	bilabial
I see them **tug**	voiced	/ʌ/	velar
I see him **tuck**	voiceless	/ʌ/	velar
What does the boy **ride**?	voiced	/aɪ/	alveolar
What does the girl **write**?	voiceless	/aɪ/	alveolar

Speech productions were recorded using a wireless or wired cardioid microphone placed on the right side of a vest placed on the child. Microphones were an average of 14.6-cm distance from the lips. The wireless battery pack was placed inside the secure vest worn by the child to minimize noise in the signal caused by interference with the microphone cord. Recordings were acquired at a 44.1-kHz sampling rate and 16-bit quantization (Kehoe & Philippart de Foy, 2023), adhering to standards for high-quality acoustic recordings. Audio files were exported as .wav files, and signal amplitudes were normalized prior to any data analysis to maintain amplitude consistency across recordings when presented to the speech-to-text software.

Data Processing

Speech targets were first orthographically transcribed using an established transcription protocol. The first author completed all AAE transcriptions using their dialectal and transcription expertise to provide consistency across all participants. Speech recordings analyzed did not include omissions or substitutions of target plosives within phrases; therefore, all final plosives used in analysis were accurate MAE and/or AAE productions. The entire phrase was then played aloud from a speaker (JBL Control 1 Pro) placed 30 cm from a computer that had Google's speech-to-text software to provide ecological validity for the distance spoken when children utilize speech-to-text systems. All phrases were presented in a sound booth to minimize any additional background noise. Google speech-to-text transcripts were then compared to the researcher's transcription to determine if the whole word of each target minimal pair (e.g., cob, cop) was correctly or incorrectly transcribed by ASR. Correctly transcribed targets were coded as "0," and incorrectly transcribed targets were coded as "1" for further analysis. See Table 3 for a description of the types of ASR transcription errors that occurred.

Acoustic Measurements

Semi-automated extraction with MATLAB. A semi-automated algorithm (Heller Murray, 2024; Story & Bunton, 2016) in MATLAB (Version R2023b) was utilized to extract average vowel *F*1. In brief, the algorithm automatically identified *F*1 values from the vowels extracted from the minimal pair target words and provided a spectrogram for the researcher to

Table 3. Automatic speech recognition (ASR) transcription error categorization.

ASR transcription error categorization	%
Final consonant substitution: nonminimal pair contrast[a]	23.30
Final consonant substitution: minimal pair contrast[b]	12.30
Vowel substitution	8.09
No transcription produced for target word[c]	6.15
Vowel and final consonant substitution	5.18
Initial and final consonant substitution	2.91
Initial consonant substitution	2.59
Substitution of all phonemes within target word[d]	2.59
Final consonant addition	1.62

[a]Final consonant substitution: nonminimal pair contrast: Errors were transcribed using phonemes other than the voiced/voiceless counterpart. For example, a target of "cab" was transcribed as "cat." [b]Final consonant substitution: minimal pair contrast: Errors were transcribed as the voiced/voiceless counterpart of the target word. For example, "cab" was transcribed as "cap." [c]No transcription produced for target word: Errors were observed as no transcription provided by the ASR tool. [d]All phonemes substituted within target word: Errors were transcribed with no phonemes matching the target word. For example, a target of "cab" was transcribed as "dog."

review with formants identified. The researcher was then able to confirm or deny accurate identification of *F*1 by the algorithm, using visual inspection and comparison to established *F*1 values in the literature (Chung, 2020; Holt, 2018; Lulich & Charles, 2020). The *F*1 average reported was over each vowel.

When the algorithmic data of *F*1 were out of bounds of the established norms for child speech (Vorperian & Kent, 2007), we utilized Praat for *F*1 extractions (please see further details below). Due to known difficulties in tracking formants in child speech (Derdemezis et al., 2016; Vallabha & Tuller, 2002), of the 1,630 possible data points, 8% (131 points) of data were missing due to the inconsistencies of the semiautomated algorithm and/or Praat difficulty identifying formants (Schiel & Zitzelsberger, 2018).

Manual extraction with Pratt. We used Praat software (Version 6.13.17; Boersma & Weenink, 2018) to extract frequency, durational, and formant measures from the vowel in the target final plosive minimal pair stimuli produced by the child AAE speakers. An undergraduate research assistant completed training in the processes related to frequency, formant, and durational measures analyzed in this study. Interrater reliability was accomplished for all researchers prior to beginning analyses. Training reliability was established as an intraclass correlation coefficient (ICC) of > .80. After experimental data extraction, interrater reliability was also completed for a sample of 10% (n = 2) of participants for all acoustic measures (i.e., mean *F*0, *F*1 average, vowel duration), in which the researchers were blinded to each other's data extractions. The interrater reliability revealed good-to-excellent reliability (Koo & Li, 2016) for mean *F*0 (ICC = .96), *F*1 average (ICC = .91), and vowel duration (ICC = .91) across the two samples.

Experimental data extractions were completed as follows. First, Praat software parameters were established to extract acoustic measures. The pitch range was set between 90 and 500 Hz, the window length was set to 0.025 s, and the formant maximum was set to 8000 Hz. Measures of mean *F*0, *F*1 average, and vowel duration were extracted manually. Measures of mean *F*0 (Hz) and *F*1 (Hz) were extracted from the vowel steady state, which was identified as the central, most stable portion of the vowel (0.050–0.100 s). Vowel duration, measured in seconds, was defined as the timing between the commencement of periodicity in the waveform and a substantial decrease in amplitude or the transition to the following plosive (Hillenbrand et al., 1995).

Statistical Analysis

Statistical analyses were completed in SAS statistical software (Version 9.4). Descriptive statistics were reported to summarize the acoustic information for voiced and voiceless plosives. To address Research Aim 1, which sought to compare ASR inaccuracies between voiced and voiceless plosive productions, a mixed-effects logistic regression model (GLIMMIX procedure in SAS) that can accommodate the underlying associations induced by repeated measurements collected from the same subjects was created with plosive voicing (voiced, voiceless), and covariates of age (ranging from 4.0 to 9.4 years old), sex (male, female), and dialect density (%). Interaction terms were examined in each model. The outcome variable of the model was the binary inaccuracy (yes/no) for each word. Significance was $p < .05$ for the model. Then, separate receiver operator characteristic (ROC) curves were calculated, including 95% confidence intervals (CIs), to report which factors can predict ASR inaccuracies based on voiced and voiceless plosive data.

To address Research Aim 2, we investigated the relationships between three acoustic measures (vowel duration, mean *F*0, and *F*1 average) and ASR accuracies by using separate mixed-effects logistic regression models for voiced and voiceless plosives. All other predictors remained the same as the model in Aim 1, including age, sex, and dialect density. Interaction terms were examined in each model. ASR inaccuracy was once again defined as the binary (yes/no) outcome measure. Separate ROC curves were generated for voiced and voiceless stimuli.

Results

Voiced Versus Voiceless Plosives, Research Aim 1

In the mixed-effects logistic regression model, there was no significant association between ASR inaccuracy and plosive voicing (p = .77), nor inaccuracy and sex (p =

.16). Interaction terms were not significant and so are not included in the tables or discussed further. ASR transcription of voiced plosives was 41% inaccurate compared to 43% inaccurate in transcribing voiceless plosives. Dialect density was not significant but approached significance (p = .07), with an increasing likelihood of inaccuracy as dialect density increased. Age was a significant predictor of inaccuracy (p < .001). As age decreased, inaccuracies increased. Further analysis of ROC curves for voiced and voiceless plosives separately, after adjusting for age, sex, and dialect density, revealed areas under the curve (AUCs) of .68 (CI [.59, .77]) and 0.75 (CI [.67, .83]), respectively. See Table 4 for additional details regarding the logistic regression model.

Acoustic Associations With ASR Inaccuracy by Voicing, Research Aim 2

Summary statistics on voiced and voiceless plosive acoustics can be found in Table 5. As expected, voiced plosives had longer vowel durations (0.265 s vs. 0.175 s) as well as lower mean $F0$ (227 Hz vs. 251 Hz) and lower $F1$ formants (959.5 vs. 1013.3), compared to voiceless plosive productions. However, the mixed-effects logistic regression models showed no association between ASR inaccuracy and these acoustic measures (mean $F0$, average $F1$, vowel duration). A trend in longer vowel durations during accurate ASR transcription is observed in Figure 1. Sex was not significant for either voiced or voiceless plosive models, and no interaction effects were significant either. For the voiced plosive model, dialect density was not a significant predictor (p = .77), and the factor of age only approached significance (p = .06). The AUC was .77 (CI [.69, 85]). However, both dialect density (p = .03) and age (p = .001) were significant predictors in the voiceless plosive model, with an AUC of .70 (CI [.61, 80]). Younger age and lower dialect density were associated with greater inaccuracies. Figure 2 visualizes the trends in age and accuracy for voiced plosives (accurate M = 6.78 years; inaccurate M = 5.75 years) and for voiceless plosives (accurate M = 6.62 years; inaccurate M = 5.80 years). See Table 6 for additional details regarding the logistic regression model.

Discussion

This study investigated final plosive variations and the relationship between acoustic measures and speech recognition inaccuracy for child speakers of AAE using Google's ASR technology, as well as factors (e.g., age, dialect density), which may impact acoustic measures and ASR transcription accuracies. We found no significant differences in ASR inaccuracy between voiced and voiceless plosives in our sample, nor were there significant relationships between the targeted acoustic measures of AAE final plosives and the ASR inaccuracy; however, age and dialect density were significantly related to ASR inaccuracy in voiceless plosives. Because these factors were not significant within ASR accuracy of voiced plosives, which display acoustic characteristics of AAE, the transcription inaccuracies within this study may be more so related to age than dialect density.

A Comparison of Voiced Versus Voiceless Plosives

Contrary to our hypothesis, the comparison of voiced and voiceless plosives did not reveal any significant differences, with inaccuracy rates of 41% versus 43%, respectively. This was unexpected because we expected a difference between the voiced and voiceless counterparts due to previous evidence on final plosive variation in

Table 4. Statistical results for Aim 1 mixed-effects logistic regressions.

Variable		ASR inaccuracy			ASR inaccuracy (voiced plosives)			ASR inaccuracy (voiceless plosives)		
		Estimate	*SE*	*p* value	Estimate	*SE*	*p* value	Estimate	*SE*	*p* value
Intercept		4.658	1.353	.011*	4.592	1.904	.047*	4.649	1.409	.013*
Age (yrs)		–0.641	0.170	< .001**	–0.652	0.241	.008**	–0.617	0.178	< .001**
Gender	Female	0.801	0.571	.162	0.814	0.812	.318	0.827	0.576	.153
	Male (reference)	—	—	—	—	—	—	—	—	—
Plosive voicing	Voiced	0.074	0.253	.771	—	—	—	—	—	—
	Voiceless (reference)	—	—	—	—	—	—	—	—	—
Dialect density		–0.017	0.009	.067	–0.014	0.013	.295	–0.020	0.010	.043*
AUC (95% CI)		.713	[.654, .772]		.748	[.669, .827]		.679	[.592, .766]	

Note. Em dashes mark cells where calculations did not occur. ASR = automatic speech recognition; yrs = years; AUC = area under the curve; CI = confidence interval.

*p < .05. **p < .01.

Table 5. Descriptive statistics of voiced and voiceless plosive acoustics.

Acoustic measure	Plosive voicing	*N*	*M*	*SD*	Minimum	Q1	*Mdn*	Q3	Maximum
Mean *F*0 (Hz)	Voiced	137	227.69	49.64	151.03	193.85	217.24	254.35	441.15
	Voiceless	124	251.11	58.48	156.94	205.29	238.39	280.35	444.63
*F*1 (Hz)	Voiced	147	959.5	293.9	374.9	727.6	963.5	1,176.0	1,704.4
	Voiceless	145	1,013.3	321.0	411.8	767.9	1,039.0	1,187.2	2,233.5
Vowel duration (s)	Voiced	141	0.26522	0.06601	0.11000	0.21650	0.25600	0.30880	0.47000
	Voiceless	138	0.17518	0.04967	0.07000	0.13500	0.17395	0.21003	0.33100

Note. *F*0 = fundamental frequency; *F*1 = first formant; Q1 = first quartile; Q3 = third quartile.

AAE speakers, in which voiced final plosives can perceptually manifest as voiceless (Lisker, 1986). Therefore, we hypothesized the final voiced plosive variations would lead to greater ASR inaccuracies compared to voiceless; however, our findings did not support this.

A trend in higher dialect density was observed for inaccurate ASR transcriptions (accurate = 53% dialect density rate, inaccurate = 61% dialect density rate); this approached significance, however, was not a statistically significant covariate in the model (p = .07). Given this information, it seems that dialect was not a factor in the accuracy of the ASR system. The description of ASR error types previously stated in Table 3 supports these results, as most errors were due to final consonant substitutions of nonminimal pairs (23.3%) compared to minimal pair errors (12%). The minimal pair errors reflect ASRs' inconsistency in recognizing the voiced plosive variation patterns of AAE. However, since most inaccuracies were related to nonminimal pairs, this can be considered less related to dialect, and more so age. The target plosives were substituted by phonemes that were not consistently of the same place and manner. For example, "cab" was transcribed as "cat" across eight speakers (M_{age} = 6.2 years, *Mdn* = 5.7 years, range: 4.0–9.4 years), which does not represent a phonological process observed within child development (Cohen & Anderson, 2011). These findings support that the improvements in Google ASR systems within adult speakers of nonmainstream dialects and languages (Zhang et al., 2023) may have expanded to children, even though previous studies have expressed concerns over the functioning of ASR for AAE speakers (Koenecke et al., 2020; Mengesha et al., 2021).

Our findings are positive in that they support that Google may have improved algorithmic accuracy, related to dialect. Still, the current inaccuracies hovered around 40%, markedly higher than previous reports of adult AAE speech at 30% (Koenecke et al., 2020). Our model did find a significant age effect with younger age of the child being related to higher inaccuracies. This result is consistent with previous studies (Johnson et al., 2023) on ASR

Figure 1. Interval plot of vowel duration by automatic speech recognition (ASR) transcription inaccuracy.

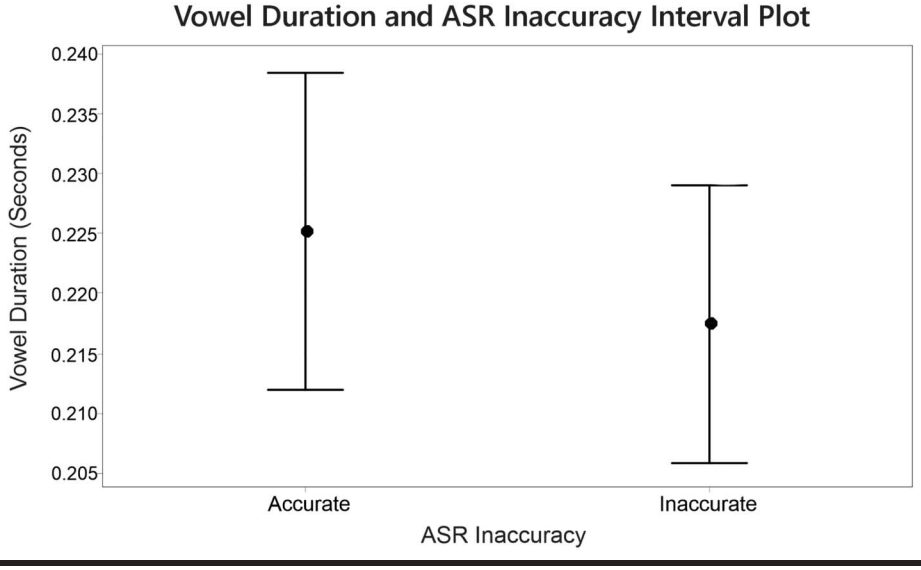
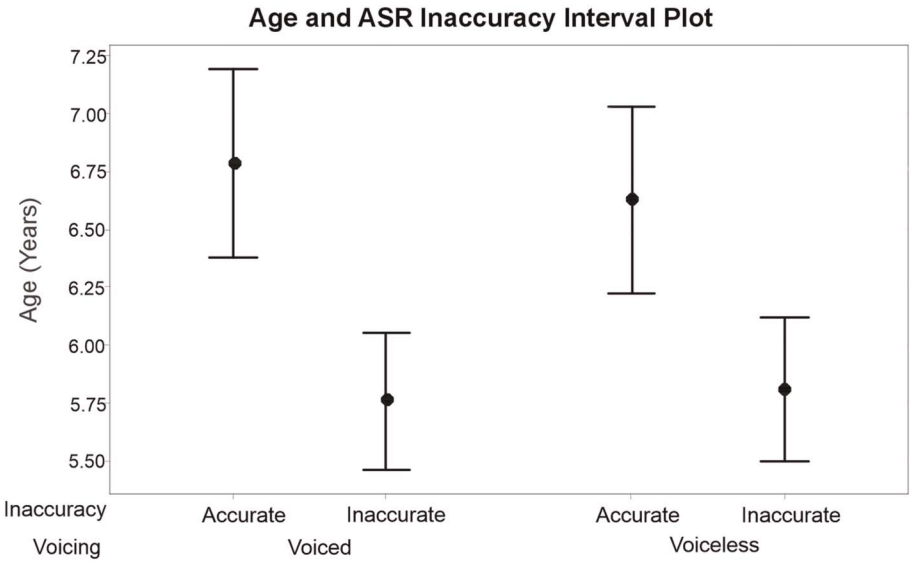


Figure 2. Interval plot of mean age by automatic speech recognition (ASR) transcription inaccuracy.



exhibiting difficulties with child speech. Therefore, as dialect variation has become a focus within ASR improvement, age more so appears to remain a limitation for the Google speech-to-text system.

Acoustic Measurements and Voiced/Voiceless Plosive Comparisons

Using acoustic measures related to listener perception of AAE voiced final plosives variations (Benki, 2001; Lisker, 1986), our second aim explored whether these same measures inform ASR transcription accuracy of AAE final plosives. Unexpectedly, the acoustic measures in this study (vowel duration, *F*0, and *F*1 average) did not have a statistically significant relationship with ASR transcription accuracies. Likely, other features of speech (e.g., consonant variability, intonation patterns, speech rate, or phonological processes) perhaps provided more relevant information for the ASR system (Johnson et al., 2023; Radha et al., 2022). Furthermore, ASR systems often rely on contextual and linguistic information in addition to acoustic features to reduce inaccuracies. The system's reliance on a broader linguistic context might compensate for the influence of the information, or lack thereof, that individual acoustic measures provide. In the present study, a simple carrier phrase was used consistently across all

Table 6. Statistical results for Aim 2 mixed-effects logistic regressions.

Variable		ASR inaccuracy			ASR inaccuracy (voiced plosives)			ASR inaccuracy (voiceless plosives)		
		Estimate	*SE*	*p* value	Estimate	*SE*	*p* value	Estimate	*SE*	*p* value
Intercept		3.411	1.624	.074	1.961	2.675	.487	4.534	1.868	.046*
Age (yrs)		–0.607	0.180	< .001**	–0.529	0.284	.065	–0.678	0.208	.002**
Gender:	Female	0.748	0.612	.223	0.762	1.001	.448	1.037	0.672	.126
	Male (reference)	—	—	—	—	—	—	—	—	—
Dialect density		–0.015	0.010	.137	–0.005	0.016	.772	–0.026	0.012	.038*
Mean *F*0		0.002	0.003	.531	0.005	0.005	.312	0.001	0.003	.771
Average *F*1		0.0005	0.0005	.303	0.0007	0.0007	.293	0.0003	0.0007	.617
Vowel duration		0.0134	1.954	.995	–2.451	3.246	.452	–0.201	3.961	.960
AUC (95% CI)		.721	[.658, .784]		.768	[.689, .847]		.703	[.609, .798]	

Note. Em dashes mark cells where calculations did not occur. ASR = automatic speech recognition; yrs = years; *F*0 = fundamental frequency; *F*1 = first formant; AUC = area under the curve; CI = confidence interval.

*$p < .05$. **$p < .01$.

stimuli ("I see a _____") in order to reduce linguistic prediction patterns and force a reliance on acoustic features of the minimal pair word. Google's ASR algorithm was able to utilize other acoustic information of the preceding vowels and plosives to make correct distinctions between voiced/voiceless and predict the target word at similar rates.

Dialect Density and Age as Factors for ASR Inaccuracy in Voiceless Plosives

Investigation of additional factors that may have impacted transcription accuracies revealed that dialect density and age were significant predictors *only for voiceless plosives*. Within the context of this study, age and dialect density did not impact ASR inaccuracy specifically in tokens that reflect the phonological patterns of AAE (i.e., voiced plosives).

Although we saw a trend for higher dialect density percentages during inaccurate ASR transcriptions (accurate = 53%, inaccurate = 61%) in the large model that evaluated voiced versus voiceless plosives, the relationship appeared to change within the analysis of voiceless plosives. The analysis revealed lower dialect density was associated with higher ASR inaccuracy, a result in complete opposition to our prediction.

Dialect Density as a Factor of ASR Inaccuracy

The phonological pattern of voiced final plosive variation involves the production of a final voiced plosive (i.e., /kæb/), which may be produced as voiced (i.e., /kæb/), voiceless (e.g., /kæp/), no plosive produced (e.g., /kæ/), or glottalized production of the voiced plosive (e.g., /kæʔ/). Children who speak AAE are often bidialectal and communicate using AAE and MAE based on the environment. Therefore, code mixing—or the interchangeable use of at least two dialects or languages during speech depending on the environment (Kim, 2006)—is common for AAE speakers. Based on the translanguaging theoretical framework (García & Lin, 2016), we acknowledge there is a continuum of AAE use ranging from "no use" to "complete use" of AAE phonological and other linguistic features (i.e., morphosyntatic, semantic, and prosodic), where in the case of these participants there is a continuous interplay between the use of AAE and MAE during conversation and tasks. AAE speakers can present with a great deal of variability in the use of AAE and MAE productions due to how each speaker uses this continuum across different environments (Giles & Hewstone, 1982). One reason for less impact of dialect density on ASR transcription in child AAE speakers with higher dialectal density may be due to the similarity in the acoustical production of voiceless plosives between AAE and MAE speech, enabling accurate ASR discrimination based on the more commonly trained MAE speech patterns.

Google Cloud's speech-to-text application programming interface specifically utilizes Chirp as its ASR model (Zhang et al., 2023). It aims to address the challenge of limited training data by first being pretrained on a large unlabeled multilingual data set (12+ million hours of YouTube-based audio in 300+ languages, with no text attached) and then optionally fine-tuned with a smaller data set of labeled multilingual data (i.e., audio with corresponding text) and an additional text-only corpus. ASR recognition is then completed using end-to-end approaches based on various types of recurrent neural networks, for example, Connectionist Temporal Classification (Graves et al., 2006) and Listen, Attend and Spell (Chan et al., 2016; Chiu et al., 2018) transducers, with the most effective algorithm depending on the task. Therefore, another reasoning for dialect density not impacting ASR transcription accuracy of voiceless plosives could be that Google's speech-to-text technology may have algorithmic properties that effectively discern voiceless plosives regardless of dialect density, due to Google's broad multilingual training goals.

Age as a Factor of ASR Inaccuracy

Age-related speech developmental changes should be considered for the optimization of ASR systems utilized for child speech. Compared to adult speakers, children's speech is characterized by higher variability across acoustic measures due to ongoing motor and linguistic development (Kent et al., 2010). This variability stems from the early-stage development of the anatomical structures involved in speech production and the gradual acquisition of motor control (Kent, 1976; Lee et al., 1997). As children grow older, their articulatory systems develop anatomically and motorically, allowing them to gain greater precision and consistency in producing phonological and intonational features (Bhardwaj et al., 2022). Previous literature supports children achieve the phonetic inventory similar to adults at approximately 6 years (Crowe & McLeod, 2020a); however, vocal development (with regard to *F*0 changes and formant shifts) and a decrease in speech disfluencies both continue into the teenage years with the onset and process of puberty (Bhardwaj et al., 2022; Evans et al., 2008; Gelfer & Denor, 2014). Because the most popular and accessible ASR systems are solely or mostly trained on adult speech, word error rates have historically been two to five times greater with child speech compared to adults (Potamianos & Narayanan, 2003). Differences in these features across development, as compared to adults, may contribute to the poor accuracy of ASR systems (Bhardwaj et al., 2022; Kumar et al.,

2020). That is, a child's speech is more variable (Levy & Hanulíková, 2019) as they establish adultlike speech patterns and furthermore have different acoustic features (i.e., higher *F*0, higher formant values) when compared to adults (Gerosa et al., 2007; Potamianos et al., 1997).

Our findings reflect the impact of the anatomic and motoric maturity in ASR transcription accuracy, specifically in the context of voiceless plosives. One theory may lie in the motor development of vocal fold abduction and oral articulation coordination timing required for voiceless plosive production. Regarding final plosives, children can continue voicing from the preceding vowel to slightly after the plosive articulatory closure to produce a voiced plosive (Higgs & Hodson, 1978), but the temporal coordination between phonation, abduction, and articulation can be more difficult for voiceless plosive production for younger children. Voiceless plosive production is characterized by the abduction of the vocal folds to release air into the glottis to the targeted constriction point. In the context of final plosives, precise timing and placement of the articulators are required to transition from vocal fold adduction of the preceding voiced vowel to the vocal fold abduction for production of the final voiceless phoneme (Dixit, 1989; Grigos et al., 2005; Löfqvist, 1980; Modolo et al., 2010; Munhall et al., 1994). These speech, motor, and vocal skills are acquired across child development, which do not reach full maturation until approximately 12 years (Kent, 1976; Kewley-Port & Preston, 1974). There is a discrepancy in acquisition development specifically related to plosives, where Crowe and McLeod (2020) states this occurs before the age of approximately 4 years, and other studies identify a higher mean age of approximately 6 years (Gilbert & Purves, 1977; Kewley-Port & Preston, 1974; Zlatin & Koenigsknecht, 1976). Studies such as Wilpon and Jacobsen (1996) have found greater accuracies of ASR that use acoustic models that were trained with and performed on specific age ranges. Our data support an inaccuracy shift between the ages of 5 and 6 years with inaccurate voiceless plosive productions an average of 5.8 years and accurate productions of 6.7 years. Therefore, further investigation into the transition from inaccurate to accurate ASR transcription in a more focused age cohort (i.e., children aged 4–7 years) may be beneficial to fully tease out the importance of maturation, motor control, and critical age ranges related to ASR inaccuracy.

ASR Inaccuracy With Voiced Plosives

The lack of significance for these factors in voiced plosives, which inherently display acoustic characteristics of AAE (e.g., glottalizing final stops or devoicing patterns), suggests transcription inaccuracies observed in this study may be more closely tied to age-related articulatory or acoustic variability than dialect density. This finding implies that the developmental stage of the speaker plays a more prominent role in ASR performance when it comes to voiced plosives, potentially because ASR models may be less sensitive to age-related speech variability compared to dialectal nuances. Alternatively, the acoustic cues associated with voiced plosives in AAE might be better captured or interpreted by the ASR system, thereby reducing the impact of dialect density on transcription accuracy related to this specific phonological pattern.

Acoustic features, which characterize the production of voiced final plosive variations in AAE, are established within the literature (Lee et al., 2022; Lisker, 1986). We investigated the impact of three of those features, including mean *F*0, *F*1, and the duration of the preceding vowel. However, there is an additional measure to consider: the *F*1 transition. The acoustic characterization of a final voiced plosive variation based on *F*1 transition includes a downward trajectory (curved decrease) in the *F*1 transition between the end of the preceding vowel to the end of voicing. Such specific and consistent measures may be collected within the spectral features of ASR algorithm during training. Furthermore, previous research has supported using more specific frequency information (i.e., harmonics, mel cepstral coefficients) to improve algorithmic accuracy (Krishna et al., 2020; Radha et al., 2022; Yarra & Ghosh, 2018; Yeung et al., 2021). The ASR system may have captured and utilized these more so than mean *F*0. Further investigation into the interaction between age, dialect density, acoustic measures, and specific phoneme characteristics (such as *F*1 transitions and harmonics) is necessary to fully understand how these factors influence ASR inaccuracies and to identify ways to improve transcription accuracy for child speakers of AAE.

Therefore, we recommend that developers of ASR systems account for varying dialect densities across different age groups. For instance, developers should consider anatomical and motor maturation when developing ASR systems for children 5 years and younger. Specifically, the system should be trained to handle higher dialect densities typical of this age group, especially when designing educational tools for young children. While dialectal variation continues to be present for older children, the dialect density may be lower when using educational resources due to environmental factors such as increased exposure to the mainstream dialect and increased ability to code-switch between dialects (Stockman, 1996). Hence, accessibility may be less of an issue for older children using ASR in homes and educational settings.

Implications and Directions for AI Development

The increasing prevalence of AI software in educational and clinical settings has highlighted significant challenges, particularly in the inaccuracy of speech-to-text systems for child speakers. Major ASR providers such as Google, Amazon, Apple, and IBM have shown reduced accuracies when processing the speech of adult speakers of AAE (Koenecke et al., 2020), which is posited to present a barrier to effective use in these critical settings for child AAE speakers. While expanding training data sets to include more children's speech and various dialects is a common approach to addressing this issue, it is a time-consuming process that requires extensive data collection. Therefore, it is imperative to explore alternative solutions that can bridge this gap more efficiently, such as the development of acoustic algorithms that are inclusive of nonmainstream American dialectal patterns (Johnson et al., 2023, 2024; Radha et al., 2022).

One promising avenue is to leverage the perceptual acoustic cues that enable humans to distinguish and understand different dialects and languages. Humans utilize a range of acoustic features, such as intonation, rhythm, and phoneme variation, to accurately perceive and interpret speech across diverse dialects and languages (Jackson, 2021; Pisoni & Remez, 2006). By incorporating these perceptual acoustic cues into ASR systems, developers could enhance the systems' ability to recognize and accurately transcribe children's speech without solely relying on large amounts of training data. This approach could involve the integration of advanced acoustic modeling techniques that mimic human auditory perception, thus improving the robustness of ASR systems in understanding the speech patterns of children.

In order for the system to gain the information required to predict and interpret dialectal differences, an immense amount of time and data is required to train ASR systems. There are methods that allow these systems to adapt preexisting models trained on adult speech to better recognize children's speech. One such method is transfer learning, which involves taking a model trained on a large data set (such as adult speech) and fine-tuning it with a smaller, more specific data set (such as children's speech; Tong et al., 2017). Domain adaptation, on the other hand, focuses on adjusting the model to perform well in a new domain (children's speech) based on knowledge gained from a related domain (adult speech; Fan et al., 2022). These methods could be helpful considering the availability of large data sets consisting of adult AAE speech that is available and used to train current models, which could then be fine-tuned to children's speech of the same dialect. Both techniques can significantly reduce the need for extensive new data collection while improving ASR accuracy for child speakers.

By combining these advanced techniques with an understanding of human perceptual cues, researchers and developers can create more effective and inclusive ASR systems that cater to the diverse speech characteristics of children. Collaboration with communities to collect these data ethically and inclusively will be essential. Furthermore, evaluating the effectiveness of these more representative models in both home and educational settings can help identify specific areas for improvement and ensure that ASR technology is equitable and accessible for all users, regardless of age and dialect.

Further consideration should be given to clinical challenges related to the use of ASR systems for AAE children. Considering the vast potential for ASR to be used as a tool to assess and treat speech sound disorders, accurate use of acoustic algorithms depends on the knowledge of how to provide the right settings to account for factors such as age, sex, quality of the recording, task, and measurements (Schiel & Zitzelsberger, 2018). Additionally, traditional ASR systems often rely on MFCCs, which does not tend to emphasize higher frequency ranges (Zhou et al., 2011). This can be problematic for accurately capturing fricative sounds and other phonetic nuances in children's speech that tend to be produced at higher frequency ranges than adults. As a result, ASR systems may struggle with correctly recognizing and processing the speech patterns of children, including those who speak AAE, leading to potential misdiagnoses or inadequate support for speech sound disorders.

Finally, the limitations of current technological hardware and software acquisition and preprocessing of signals can significantly impact the accurate collection of child speech data. Many ASR systems and their associated hardware are designed with adult speakers in mind, lacking the sensitivity and precision needed for younger users. Children's voices, with their higher *F*0 and formants (Lee et al., 1997) and greater variability (Hill et al., 2013), often require more sophisticated and sensitive recording equipment to ensure accurate data capture. Both software and hardware developers should follow guidelines set forth by expert speech and voice scientists to ensure high-quality acoustic recordings (e.g., microphone specifications, sampling rates appropriate for child speech, and data compression concerns on smart devices; Cavalcanti et al., 2023; Patel et al., 2018; Vogel & Morgan, 2009) to lay the foundation for ASR technology to work with the acquired acoustic signals. Therefore, advancing ASR technology for children, especially those speaking AAE, necessitates improvements in both algorithmic approaches and the hardware used for data collection to ensure more reliable and clinically useful outcomes.

Limitations

Due to this being a pilot investigation, our study had a relatively small number of participants that may have limited statistical power and the ability to detect significance. A larger sample size could provide more robust data, explore interaction effects, and potentially reveal significant patterns that were not apparent in this study. Although interaction terms were examined, they often encountered estimation issues, such as convergence failures or negative variance estimates, which is likely due to the sample size and a limitation of this study. Therefore, a larger group would allow for incorporation and interpretation of more representative statistical interactions.

AAE-speaking boys consistently exhibit higher dialectal density than girls (Washington & Craig, 1998). However, our sample does not have an equal representation of both genders. To assess potential skewness in the data, we conducted a correlational analysis between age and dialectal density for the entire sample of boys and girls (N = 11) and compared it to the analysis for the subset of girls only (N = 9). We found no significant differences between the two analyses and therefore included the entire sample in our study. Future research should aim to match participants by both age and gender to better control for these variables.

The researcher was bidialectal and used a range of both AAE and MAE phonological and morphosyntactic patterns. When providing nonimmediate repetition of stimuli during minimal pair phrases, the researchers range was not considered when eliciting the speech of the child AAE speakers. This could have impacted the consistency of the stimuli presentation, specifically related to AAE productions, potentially introducing bias or variability that may have influenced the participants' responses and the overall validity of the study's findings. However, we think this is unlikely given that children tend to use their own patterns given delayed repetition of speech patterns. Nevertheless, future research should aim to control these variables more rigorously. Increasing the sample size and controlling how stimuli is prompted or modeled by administers could help in isolating the effects of dialect use on ASR accuracy.

Collecting data within the natural home environment posed limitations when evaluating the effectiveness of AI tools in educational environments. However, data collection in the home also provided a broader representation of AAE beyond the educational context. Clinical and educational tools can also be utilized at home, highlighting the importance of gathering data in diverse settings. Still, collecting data in natural environments introduced challenges in acoustic analysis due to noise impacting the acoustic signals and subsequent formant estimations. Therefore, data collection in a sound booth would mitigate these acoustic interferences, allowing for a more accurate detection of formant data and reduce missing data from the acoustic analyses.

Dialectal density was calculated using the DELV screener, which provided scores related to the phonological, morphological, and syntactic variation of the speakers. This score may have been more reflective of the variation within the task by being derived from the percentage of phonological productions from the target minimal pairs used in this study. Therefore, a perceptual analysis of stimuli, to determine AAE variation, or calculation of dialectal density based on the stimuli may better represent dialectal variation for the specific task.

Additionally, our study only examined one AAE feature of voiced plosive variation; however, there are many additional aspects of AAE speech that may contribute to ASR inaccuracies. As mentioned, even adult speakers of AAE experience difficulties with using ASR to interpret their speech (Koenecke et al., 2020; Mengesha et al., 2021). Characteristics of interdental fricative variations (e.g., "they" pronounced as /deɪɣ/, consonant cluster variations; e.g., "street" pronounced as /skrit/) and plosives not produced in consonant clusters of plural words (e.g., "desks" pronounced as /dɛsʌz/) are additional patterns that could be investigated further (Lee, 2021). Within these, specific spectral acoustic parameters such as vowel inherent spectral change and fricative center of gravity may be useful to examine and characterize variation in these features (Holt & Ellis, 2018).

Lastly, enrolling more participants within a targeted age range that is reflective of speech developmental maturation could be beneficial for developing educational tools targeted for specific ages. Child consonant acquisition occurs around 5 years, and our results showed differences in ASR transcription accuracy between 5 and 7 years. This range includes kindergarten age where children are first in schools and acquiring literacy skills and undergoing critical phonological development. As children continue in schooling, maturation of code-switching is likely to occur, and therefore, we could investigate the interaction between all factors including age, dialect density, and code-switching abilities.

Conclusions

This study investigated the impact of acoustic measures (mean $F0$, $F1$, and vowel duration), age, and dialect density on Google speech-to-text transcription accuracy in voiced and voiceless plosive productions within child AAE speech. We found that our specific acoustic measures did not impact ASR accuracy. However, age and dialect

density influenced transcription accuracy for voiceless plosives, with age having a greater effect than dialect density.

These findings highlight the importance of considering developmental and dialectal variations in ASR performance. Future research should explore the relationship between age and dialect density in greater detail and examine these factors within a larger sample. Additionally, continued evaluation of AI and ASR technologies tailored for diverse American dialects, including AAE, is necessary to improve digital inclusivity. This includes developing AI-driven language assessment and intervention tools that account for dialectal phonological and syntactic features. Enhancing transcription accuracy for AAE speakers could lead to more effective educational and clinical applications, benefiting speech recognition services across various domains.

Data Availability Statement

Data will be made available upon reasonable request to the authors.

Acknowledgments

This work was supported by the National Institutes of Health through the T32 Grant DC013017 and by the Ohio Speech-Language-Hearing Association under the 2023 Research Grant Seed Money Award. Portions of this work were presented at the Voice, Speech and Technology Symposium, December 2023; the National Black Association for Speech-Language and Hearing Conference, March 2024; and the American Speech-Language-Hearing Association Conference, December 2024. The authors would like to thank all the children and families who participated in the study. Without their participation, we would not be able to improve the representation of AAE speakers. Last, thank you to Jenny Vojtech for assisting in normalization of all audio files and Andres Llico and Elizabeth Heller Murray for assistance in semiautomatic acoustic analysis.

References

Abramson, A. S., & Whalen, D. H. (2017). Voice onset time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. *Journal of Phonetics, 63,* 75–86. https://doi.org/10.1016/j.wocn.2017.05.002

Albudoor, N., & Peña, E. (2022). Identifying language disorder in bilingual children using automatic speech recognition. *Journal of Speech, Language, and Hearing Research, 65*(7), 2648–2661. https://doi.org/10.1044/2022_JSLHR-21-00667

Attia, A. A., Liu, J., Ai, W., Demszky, D., & Espy-Wilson, C. (2024). *Kid-whisper: Towards bridging the performance gap in automatic speech recognition for children vs. adults*. arXiv. https://doi.org/10.48550/arXiv.2309.07927

Bailey, Z., Krieger, N., Agenor, M., Graves, J., Linos, N., & Bassett, M. (2017). Structural racism and health inequities in the USA: Evidence and interventions. *Lancet, 389*(10077), 1453–1463. https://doi.org/10.1016/S0140-6736(17)30569-X

Benki, J. (2001). Place of articulation and first formant transition pattern both affect perception of voicing in English. *Journal of Phonetics, 29*(1), 1–22. https://doi.org/10.1006/jpho.2000.0128

Bhardwaj, V., Othman, M., Kukreja, V., Belkhier, Y., Bajaj, M., Goud, B., Rehman, A., Shafiq, M., & Hamam, H. (2022). Automatic speech recognition (ASR) systems for children: A systematic literature review. *Applied Sciences, 12*(9), Article 4419. https://doi.org/10.3390/app12094419

Boersma, P., & Weenink, D. (2018). Praat (Version 6.13.17) [Computer software]. https://www.praat.org

Cavalcanti, J. C., Englert, M., Oliveira, M., & Constantini, A. C. (2023). Microphone and audio compression effects on acoustic voice analysis: A pilot study. *Journal of Voice, 37*(2), 162–172. https://doi.org/10.1016/j.jvoice.2020.12.005

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4960–4964). https://doi.org/10.1109/ICASSP.2016.7472621

Charity, A. (2008). African American English: An overview. *Perspectives on Communication Disorders and Sciences in Culturally and Linguistically Diverse (CLD) Populations, 15*(2), 33–42. https://doi.org/10.1044/cds15.2.33

Charity Hudley, A. H., Mallinson, C., Sudler, K., & Fama, M. (2018). The sociolinguistically trained speech-language pathologist: Using knowledge of African American English to aid and empower African American clientele. *Perspectives of the ASHA Special Interest Groups, 3*(1), 118–131. https://doi.org/10.1044/persp3.SIG1.118

Chiu, C.-C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, E., Jaitly, N., Li, B., Chorowski, J., & Bacchiani, M. (2018). State-of-the-art speech recognition with sequence-to-sequence models. *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4774–4778). https://doi.org/10.1109/ICASSP.2018.8462105

Chung, H. (2020). Acquisition and acoustic patterns of Southern American English /l/ in young children. *Journal of Speech, Language, and Hearing Research, 63*(8), 2609–2624. https://doi.org/10.1044/2020_JSLHR-19-00040

Cohen, W., & Anderson, C. (2011). Identification of phonological processes in preschool children's single-word productions. *International Journal of Language & Communication Disorders, 46*(4), 481–488. https://doi.org/10.1111/j.1460-6984.2011.00011.x

Craig, H. K., Thompson, C. A., Washington, J. A., & Potter, S. L. (2003). Phonological features of child African American English. *Journal of Speech, Language, and Hearing Research, 46*(3), 623–635. https://doi.org/10.1044/1092-4388(2003/049)

Craig, H. K., Thompson, C. A., Washington, J. A., & Potter, S. L. (2004). Performance of elementary-grade African American students on the Gray Oral Reading Tests. *Language, Speech, and Hearing Services in Schools, 35*(2), 141–154. https://doi.org/10.1044/0161-1461(2004/015)

Craig, H. K., Zhang, L., Hensel, S. L., & Quinn, E. J. (2009). African American English–speaking students: An examination of the relationship between dialect shifting and reading outcomes. *Journal of Speech, Language, and Hearing Research*, *52*(4), 839–855. https://doi.org/10.1044/1092-4388(2009/08-0056)

Crowe, K., & McLeod, S. (2020). Children's English consonant acquisition in the United States: A review. *American Journal of Speech-Language Pathology, 29*(4), 2155–2169. https://doi.org/10.1044/2020_AJSLP-19-00168

Deng, L., & Li, X. (2013). Machine learning paradigms for speech recognition: An overview. *IEEE Transactions on Audio, Speech, and Language Processing, 21*(5), 1060–1089. https://doi.org/10.1109/TASL.2013.2244083

Derdemezis, E., Vorperian, H. K., Kent, R. D., Fourakis, M., Reinicke, E. L., & Bolt, D. M. (2016). Optimizing vowel formant measurements in four acoustic analysis systems for diverse speaker groups. *American Journal of Speech-Language Pathology, 25*(3), 335–354. https://doi.org/10.1044/2015_AJSLP-15-0020

Dixit, R. P. (1989). Glottal gestures in Hindi plosives. *Journal of Phonetics, 17*(3), 213–237. https://doi.org/10.1016/S0095-4470(19)30431-0

Dodd, B., Hua, Z., Crosbie, S., Holm, A., & Ozanne, A. (2006). Diagnostic Evaluation of Articulation and Phonology. *Pearson*. https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Speech-&-Language/Diagnostic-Evaluation-of-Articulation-and-Phonology/p/100000295.html

Du, Y. (2023). '*Alexa, please start our speech session*.' ASHA LeaderLive. https://leader.pubs.asha.org/do/10.1044/leader.FTR1a.28012023.voice-assistants-slp.48/full/

Evans, S., Neave, N., Wakelin, D., & Hamilton, C. (2008). The relationship between testosterone and vocal frequencies in human males. *Physiology & Behavior, 93*(4–5), 783–788. https://doi.org/10.1016/j.physbeh.2007.11.033

Fan, R., Zhu, Y., Wang, J., & Alwan, A. (2022). Towards better domain adaptation for self-supervised models: A case study of child ASR. *IEEE Journal of Selected Topics in Signal Processing, 16*(6), 1242–1252. https://doi.org/10.1109/JSTSP.2022.3200910

Fischer, V., Kunzmann, S., Janke, E.-W., & Tyrrell, J. (2006). *Method and apparatus for phonetic context adaptation for improved speech recognition (U.S. Patent No. US 6,999,925 B2)*. U.S. Patent and Trademark Office.

García, O., & Lin, A. M. Y. (2016). Translanguaging in bilingual education. In O. Garcia, A. Lin, & S. May (Eds.), *Bilingual and multilingual education* (pp. 1–14). Springer. https://doi.org/10.1007/978-3-319-02324-3_9-1

Gelfer, M. P., & Denor, S. L. (2014). Speaking fundamental frequency and individual variability in Caucasian and African American school-age children. *American Journal of Speech-Language Pathology, 23*(3), 395–406. https://doi.org/10.1044/2014_AJSLP-13-0016

Gerosa, M., Giuliani, D., & Brugnara, F. (2007). Acoustic variability and automatic recognition of children's speech. *Speech Communication, 49*(10–11), 847–860. https://doi.org/10.1016/j.specom.2007.01.002

Gilbert, J. H. V., & Purves, B. A. (1977). Temporal constraints on consonant clusters in child speech production. *Journal of Child Language, 4*(3), 417–432. https://doi.org/10.1017/S030500090000177X

Giles, H., & Hewstone, M. (1982). Cognitive structures, speech, and social situations: Two integrative models. *Language Sciences, 4*(2), 187–219. https://doi.org/10.1016/S0388-0001(82)80005-3

Graves, A., Fernandez, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd International Conference on Machine Learning* (pp. 369–376). https://doi.org/10.1145/1143844.1143891

Green, L. J. (2010). *Language and the African American child*. Cambridge University Press. https://doi.org/10.1017/CBO9780511975561

Grigos, M. I., Saxman, J. H., & Gordon, A. M. (2005). Speech motor development during acquisition of the voicing contrast. *Journal of Speech, Language, and Hearing Research, 48*(4), 739–752. https://doi.org/10.1044/1092-4388(2005/051)

Hamilton, M.-B., Mont, E. V., & Cameron, M. (2018). Deletion, omission, reduction: Redefining the language we use to talk about African American English. *Perspectives of the ASHA Special Interest Groups, 3*(1), 107–117. https://doi.org/10.1044/persp3.SIG1.107

Heller Murray, E. (2024). Conducting high-quality and reliable acoustic analysis: A tutorial focused on training research assistants. *The Journal of the Acoustical Society of America, 155*(4), 2603–2611. https://doi.org/10.1121/10.0025536

Hendricks, A. E., & Adlof, S. M. (2017). Language assessment with children who speak nonmainstream dialects: Examining the effects of scoring modifications in norm-referenced assessment. *Language, Speech, and Hearing Services in Schools, 48*(3), 168–182. https://doi.org/10.1044/2017_LSHSS-16-0060

Hermansky, H., & Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing, 2*(4), 578–589. https://doi.org/10.1109/89.326616

Higgs, J. A. W., & Hodson, B. W. (1978). Phonological perception of word-final obstruent cognates. *Journal of Phonetics, 6*(1), 25–35. https://doi.org/10.1016/S0095-4470(19)31083-6

Hill, C. A., Ojha, S., Maturo, S., Maurer, R., Bunting, G., & Hartnick, C. J. (2013). Consistency of voice frequency and perturbation measures in children. *Otolaryngology—Head and Neck Surgery, 148*(4), 637–641. https://doi.org/10.1177/0194599813477829

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America, 97*(5), 3099–3111. https://doi.org/10.1121/1.411872

Holt, Y. (2018). Mechanisms of vowel variation in African American English. *Journal of Speech, Language, and Hearing Research, 61*(2), 197–209. https://doi.org/10.1044/2017_JSLHR-S-16-0375

Holt, Y., & Ellis, C. (2018). African American women's speech: Vowel inherent spectral change. *Acoustical Science and Technology, 39*(2), 160–162. https://doi.org/10.1250/ast.39.160

Hunt, M. (2006). *Phonetically based speech recognition system and method (U.S. Patent No. 7,146,319 B2)*. U.S. Patent and Trademark Office.

Ivy, L. J., & Masterson, J. J. (2011). A comparison of oral and written English styles in African American students at different stages of writing development. *Language, Speech, and Hearing Services in Schools, 42*(1), 31–40. https://doi.org/10.1044/0161-1461(2010/09-0069)

Jackson, L. L. (2021). Assessment of cross-language speech perception in adults: Strongly considering phonetic and acoustic contributions. *Journal of Linguistics and Languages, 3*(2), 1–10.

Johnson, A., Shankar, N. B., Ostendorf, M., & Alwan, A. (2024). An exploratory study on dialect density estimation for children and adult's African American English. *The Journal of the Acoustical Society of America, 155*(4), 2836–2848. https://doi.org/10.1121/10.0025771

Johnson, A., Shetty, V. M., Ostendorf, M., & Alwan, A. (2023). Leveraging multiple sources in automatic African American English dialect detection for adults and children. *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1–5). https://doi.org/10.1109/ICASSP49357.2023.10096614

Keating, C. (2023, October 13). *Data drop: Gen Z leading voice assistant growth*. EMARKETER. https://www.emarketer.com/content/data-drop-gen-z-leading-voice-assistant-growth

Kehoe, M., & Philippart de Foy, M. (2023). The development of alveolar and alveopalatal fricatives in French-speaking monolingual and bilingual children. *Journal of Speech, Language, and Hearing Research, 66*(2), 475–502. https://doi.org/10.1044/2022_JSLHR-22-00387

Kendall, T., & Farrington, C. (2018). *The corpus of regional African American language*. (Version 2018.10.06).

Kendall, T., & Farrington, C. (2024). *The corpus of regional African American language* (Version 2023.06). https://oraal.github.io/coraal

Kent, R. D. (1976). Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies. *Journal of Speech and Hearing Research, 19*(3), 421–447. https://doi.org/10.1044/jshr.1903.421

Kent, R. D., Pagan-Neves, L., Hustad, C., & Wertzner, H. F. (2010). Children's speech sound disorders: An acoustic perspective. In R. Paul & P. Flipsen, Jr., *Speech sound disorders in children* (pp. 93–114). Plural.

Kewley-Port, D., & Preston, M. S. (1974). Early apical stop production: A voice onset time analysis. *Journal of Phonetics, 2*(3), 195–210. https://doi.org/10.1016/S0095-4470(19)31270-7

Kim, E. (2006). Reasons and motivations for code-mixing and code-switching. *TESOL, 4*(1), 43–61.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences, 117*(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Krishna, G., Krishnan, R., & Mittal, V. (2020). A system for automatic regional accent classification. *Proceedings of the 2020 IEEE 17th India Council International Conference (INDICON)* (pp. 1–5). https://doi.org/10.1109/INDICON49873.2020.9342577

Kumar, H. K., Kadiri, S. R., Alku, P., & Kurimo, M. (2020). Study of formant modification for children ASR. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7429–7433). https://doi.org/10.1109/ICASSP40776.2020.9053334

Lee, S. (2021). *African American English Articulation Tool and its impact on differentiating speech sound delays/disorders* [Doctoral dissertation, Howard University]. https://www.proquest.com/openview/3559202b57f0864916bb236e35bfc33e/1?pq-origsite=gscholar&cbl=18750&diss=y

Lee, S., Lucker, J., Picho-Kiroga, K., Wright-Harp, W., & Martinez, S. (2022). A compilation of speech sound features of African American English. *Journal of the National Black Association for Speech Language and Hearing, 17*(1), 6–29.

Lee, S., Potamianos, A., & Narayanan, S. S. (1997). Analysis of children's speech: Duration, pitch and formants. *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech 1997)* (pp. 473–476). https://doi.org/10.21437/Eurospeech.1997-161

Levy, H., & Hanulíková, A. (2019). Variation in children's vowel production: Effects of language exposure and lexical frequency. *Laboratory Phonology, 10*(1), Article 9. https://doi.org/10.5334/labphon.131

Lin, H., Deng, L., Yu, D., Gong, Y.-F., Acero, A., & Lee, C.-H. (2009). A study on multilingual acoustic modeling for large vocabulary ASR. *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4333–4336). https://doi.org/10.1109/ICASSP.2009.4960588

Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech, 29*(1), 3–11. https://doi.org/10.1177/002383098602900102

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*(3), 384–422. https://doi.org/10.1080/00437956.1964.11659830

Llamas, C., & Watt, D. (2010). *Language and identities*. Edinburgh University Press. https://doi.org/10.1515/9780748635788

Lo, T.-H., Chao, F.-A., Weng, S.-Y., & Chen, B. (2020). *The NTNU system at the Interspeech 2020 non-native children's speech ASR challenge*. arXiv. https://doi.org/10.48550/arXiv.2005.08433

Löfqvist, A. (1980). Interarticulator programming in stop production. *Journal of Phonetics, 8*(4), 475–490. https://doi.org/10.1016/S0095-4470(19)31502-5

Lulich, S. M., & Charles, S. D. (2020). Development of formant frequency distributions in American English–speaking elementary school-aged children: A longitudinal study. *Proceedings of Meetings on Acoustics, 42,* Article 060016. https://doi.org/10.1121/2.0001459

Marencin, N. C., Edwards, A. A., & Terry, N. P. (2024). African American preschoolers' performance on norm-referenced language assessments: Examining the effect of dialect density and the use of scoring modifications. *Language, Speech, and Hearing Services in Schools, 55*(3), 918–937. https://doi.org/10.1044/2024_LSHSS-23-00134

Matassoni, M., Gretter, R., Falavigna, D., & Giuliani, D. (2018). *Non-native children speech recognition through transfer learning (arXiv:1809.09658)*. arXiv. https://doi.org/10.48550/arXiv.1809.09658

Mengesha, Z., Heldreth, C., Lahav, M., Sublewski, J., & Tuennerman, E. (2021). "I don't think these devices are very culturally sensitive."—Impact of automated speech recognition errors on African Americans. *Frontiers in Artificial Intelligence, 4,* Article 725911. https://doi.org/10.3389/frai.2021.725911

Modolo, D. J., Berretin-Felix, G., Genaro, K. F., & Brasolotto, A. G. (2010). Oral and vocal fold diadochokinesis in children. *Folia Phoniatrica et Logopaedica, 63*(1), 1–8. https://doi.org/10.1159/000319728

Moland, C. W., & Oetting, J. B. (2021). Comparison of the Diagnostic Evaluation of Language Variation–Screening Test Risk Subtest to two other screeners for low-income prekindergartners who speak African American English and live in the urban south. *American Journal of Speech-Language Pathology, 30*(6), 2528–2541. https://doi.org/10.1044/2021_AJSLP-20-00270

Munhall, K. G., Löfqvist, A., & Kelso, J. A. (1994). Lip–larynx coordination in speech: Effects of mechanical perturbations to the lower lip. *The Journal of the Acoustical Society of America, 95*(6), 3605–3616. https://doi.org/10.1121/1.409929

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE Access, 7,* 19143–19165. https://doi.org/10.1109/ACCESS.2019.2896880

National Center for Technology Innovation. (2010). *Speech recognition for learning*. Reading Rockets. https://www.readingrockets.org/article/speech-recognition-learning

Ohio Department of Health. (2022). *Hearing screening guidelines and requirements.* https://odh.ohio.gov/know-our-programs/childrens-hearing-vision-program/requirements/2022-guidelines

O'Shaughnessy, D. (2024). Trends and developments in automatic speech recognition research. *Computer Speech and Language, 83,* Article 101538. https://doi.org/10.1016/j.csl.2023.101538

Patel, R. R., Awan, S. N., Barkmeier-Kraemer, J., Courey, M., Deliyski, D., Eadie, T., Paul, D., Švec, J. G., & Hillman, R. (2018). Recommended protocols for instrumental assessment of voice: American Speech-Language-Hearing Association Expert Panel to develop a protocol for instrumental assessment of vocal function. *American Journal of Speech-Language Pathology, 27*(3), 887–905. https://doi.org/10.1044/2018_AJSLP-17-0009

Pisoni, D. B., & Remez, R. E. (2006). *The handbook of speech perception*. Blackwell.

Potamianos, A., & Narayanan, S. (2003). Robust recognition of children's speech. *IEEE Transactions on Speech and Audio Processing, 11*(6), 603–616. https://doi.org/10.1109/TSA.2003.818026

Potamianos, A., Narayanan, S., & Lee, S. (1997). Automatic speech recognition for children. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech 1997)* (pp. 2371–2374). https://doi.org/10.21437/Eurospeech.1997-623

Puranik, C., Branum-Martin, L., & Washington, J. A. (2020). The relation between dialect density and the codevelopment of writing and reading in African American children. *Child Development, 91*(4), e866–e882. https://doi.org/10.1111/cdev.13318

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision*. arXiv. https://doi.org/10.48550/arXiv.2212.04356

Radha, K., Bansal, M., & Pachori, R. B. (2024). Speech and speaker recognition using raw waveform modeling for adult and children's speech: A comprehensive review. *Engineering Applications of Artificial Intelligence, 131,* Article 107661. https://doi.org/10.1016/j.engappai.2023.107661

Radha, K., Bansal, M., & Shabber, S. M. (2022). Accent classification of native and non-native children using harmonic pitch. *Proceedings of the 2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP)* (pp. 1–6). https://doi.org/10.1109/AISP53593.2022.9760588

Salau, A., Olowoyo, T., & Akinola, S. (2020). Accent classification of the three major Nigerian Indigenous languages using 1D CNN LSTM network model. In S. Jain, M. Sood, & S. Paul (Eds.), *Advances in computational intelligence techniques* (pp. 1–16). Springer. https://doi.org/10.1007/978-981-15-2620-6_1

Say It Labs. (2023). *Stutter Stars*. https://www.sayitlabs.com/stutter-stars

Schiel, F., & Zitzelsberger, T. (2018). Evaluation of automatic formant trackers. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association. https://aclanthology.org/L18-1449

Seymour, H., Roeper, T., & Villiers, J. (2003). *Diagnostic Evaluation of Language Variation (DELV)–Norm-referenced*. Harcourt.

Snow, D. (1997). Children's acquisition of speech timing in English: A comparative study of voice onset time and final syllable vowel lengthening. *Journal of Child Language, 24*(1), 35–56. https://doi.org/10.1017/S0305000996003029

Stevens, K. N., & Klatt, D. H. (1974). Role of formant transitions in the voiced–voiceless distinction for stops. *The Journal of the Acoustical Society of America, 55*(3), 653–659. https://doi.org/10.1121/1.1914578

Stockman, I. (1996). Phonological development and disorders in African American children. In A. G. Kamhi, K. E. Pollock, & J. L. Harris (Eds.), *Communication development and disorders in African American children: Research, assessment, and intervention* (pp. 117–153). Brookes.

Stockman, I. (2010). A review of developmental and applied language research on African American children: From a deficit to difference perspective on dialect differences. *Language, Speech, and Hearing Services in Schools, 41*(1), 23–38. https://doi.org/10.1044/0161-1461(2009/08-0086)

Story, B. H., & Bunton, K. (2016). Formant measurement in children's speech based on spectral filtering. *Speech Communication, 76,* 93–111. https://doi.org/10.1016/j.specom.2015.11.001

Sunil, Y., Prasanna, S. R. M., & Sinha, R. (2016). Children's speech recognition under mismatched condition: A review. *IETE Journal of Education, 57*(2), 96–108. https://doi.org/10.1080/09747338.2016.1201014

Terry, N. P., Connor, C. M., Petscher, Y., & Conlin, C. R. (2012). Dialect variation and reading: Is change in nonmainstream American English use related to reading achievement in first and second grades? *Journal of Speech, Language, and Hearing Research, 55*(1), 55–69. https://doi.org/10.1044/1092-4388(2011/09-0257)

Thomas, E. (2007). Phonological and phonetic characteristics of African American Vernacular English. *Language and Linguistics Compass, 1*(5), 450–475. https://doi.org/10.1111/j.1749-818X.2007.00029.x

Thompson, C. A., Craig, H. K., & Washington, J. A. (2004). Variable production of African American English across oracy and literacy contexts. *Language, Speech, and Hearing Services in Schools, 35*(3), 269–282. https://doi.org/10.1044/0161-1461(2004/025)

Tong, R., Wang, L., & Ma, B. (2017). Transfer learning for children's speech recognition. *Proceedings of the International Conference on Asian Language Processing (IALP)* (pp. 36–39). https://doi.org/10.1109/IALP.2017.8300540

Trudgill, P. (2008). Colonial dialect contact in the history of European languages: On the irrelevance of identity to new-dialect formation. *Language in Society, 37*(02), 241–254. https://doi.org/10.1017/S0047404508080287

Vallabha, G., & Tuller, B. (2002). Systematic errors in the formant analysis of steady-state vowels. *Speech Communication, 38*(1–2), 141–160. https://doi.org/10.1016/S0167-6393(01)00049-8

Vogel, A. P., & Morgan, A. T. (2009). Factors affecting the quality of sound recording for speech and voice analysis. *International Journal of Speech-Language Pathology, 11*(6), 431–437. https://doi.org/10.3109/17549500902822189

Vorperian, H. K., & Kent, R. D. (2007). Vowel acoustic space development in children: A synthesis of acoustic and anatomic data. *Journal of Speech, Language, and Hearing Research, 50*(6), 1510–1545. https://doi.org/10.1044/1092-4388(2007/104)

Waserblat, M., & Eilam, B. (2014). *Automatic speech analysis (U.S. Patent No. 8,725,518 B2)*. U.S. Patent and Trademark Office.

Washington, J. A., Branum-Martin, L., Sun, C., & Lee-James, R. (2018). The impact of dialect density on the growth of language and reading in African American children. *Language, Speech, and Hearing Services in Schools, 49*(2), 232–247. https://doi.org/10.1044/2018_LSHSS-17-0063

Washington, J. A., & Craig, H. K. (1998). Socioeconomic status and gender influences on children's dialectal variations. *Journal of Speech, Language, and Hearing Research, 41*(3), 618–626. https://doi.org/10.1044/jslhr.4103.618

Wilpon, J. G., & Jacobsen, C. N. (1996). A study of speech recognition for children and the elderly. *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, 1,* 349–352. https://doi.org/10.1109/ICASSP.1996.541104

Wyatt, T. A. (1995). Language development in African American English child speech. *Linguistics and Education, 7*(1), 7–22. https://doi.org/10.1016/0898-5898(95)90017-9

Yarra, C., & Ghosh, P. K. (2018). Automatic intonation classification using temporal patterns in utterance-level pitch contour and perceptually motivated pitch transformation. *The Journal of the Acoustical Society of America, 144*(5), EL471–EL476. https://doi.org/10.1121/1.5080466

Yeung, G., & Alwan, A. (2018). On the difficulties of automatic speech recognition for kindergarten-aged children. In *Proceedings of Interspeech 2018* (pp. 1661–1665). https://doi.org/10.21437/Interspeech.2018-2297

Yeung, G., Fan, R., & Alwan, A. (2021). *Fundamental frequency feature normalization and data augmentation for child speech recognition*. arXiv. https://doi.org/10.48550/arXiv.2102.09106

Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., Chen, N., Li, B., Axelrod, V., Wang, G., Meng, Z., Hu, K., Rosenberg, A., Prabhavalkar, R., Park, D. S., Haghani, P., Riesa, J., Perng, G., Soltau, H., . . . Wu, Y. (2023). *Google USM: Scaling automatic speech recognition beyond 100 languages*. arXiv. https://doi.org/10.48550/arXiv.2303.01037

Zhao, R., Choi, A. S. G., Koenecke, A., & Rameau, A. (2025). Quantification of automatic speech recognition system performance on d/deaf and hard of hearing speech. *The Laryngoscope, 135*(1), 191–197. https://doi.org/10.1002/lary.31713

Zhou, X., Garcia-Romero, D., Duraiswami, R., Espy-Wilson, C., & Shamma, S. (2011). Linear versus mel frequency cepstral coefficients for speaker recognition. *IEEE Workshop on Automatic Speech Recognition & Understanding, 2011,* 559–564. https://doi.org/10.1109/ASRU.2011.6163888

Zlatin, M., & Koenigsknecht, R. (1976). Development of the voicing contrast: A comparison of voice onset time in stop perception and production. *Journal of Speech and Hearing Research, 19*(1), 93–111. https://doi.org/10.1044/jshr.1901.93