ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

DIALECT-DRIVEN ASR ERRORS: PHONETIC MISMATCH IN SOUTH ASIAN AMERICAN ENGLISH SPEECH

[1]Muhammad Ansar
MA Data and Discourse Studies
Department of History and Social Sciences
Technische Universität Darmstadt, Germany
muhammad.ansar@stud.tu-darmstadt.de
ORCID: https://orcid.org/0009-0007-0649-2033
[1*]Anosh Rehman
Department of English Linguistics and Language Studies,
University of Sargodha
Email: anoshhamza338@gmail.com
ORCID: https://orcid.org/0009-0001-0412-0160
[2]Hamza Nawaz Chaudhary
Department of CS,
University of Sargodha
Email: hamnaw66@gmail.com

Abstract

ASR systems have reached almost human accuracy with Mainstream American English (MAE), but still make systematic errors on non-mainstream varieties. This paper examines how the ASR errors are formed in South Asian American English (SAAE), and it has been argued that the errors are due to a systematic discrepancy between the phonetic realizations of SAAE speakers and the acoustic-phonetic distributions coded into MAE-trained models, the Phonetic Mismatch Hypothesis. A convergent mixed-methods design was used and a controlled speech elicitation and quantitative analysis of error. The 40 SAAE speech samples were put together to form a corpus that reflects major segmental and suprasegmental aspects, such as variation in the quality of vowels, reduction of consonant clusters, epenthesis, and the presence of prosodic transfer. A pretrained Whisper ASR model was tested on reference transcriptions with the calculation of Word Error Rate (WER). A total of 170 errors were identified and classified as substitutions (82; 48.2%), deletions (52; 30.6%), and insertions (36; 21.2%). The speech of SAAE generated a WER of about 43, as opposed to a generation of about 6 by MAE speech, and there was a partial amelioration of the situation when the speech was generated under a fine-tuned adaptation condition (WER ≈ 18%). Types of errors were not randomly distributed among phonetic features: substitution errors were caused by vowel changes and consonant replacements; deletions were explained by the presence of consonant clusters; and most insertions were due to prosodic and rhythmic variation, specifically syllable-timed rhythm and epenthesis. These findings support the phonetic mismatch hypothesis that attributes errors in ASR to linguistic behaviors, and not failures in the system. This study contributes to a phonologically grounded description of ASR bias and proposes training and evaluation models to factor in dialect-specific phonetic knowledge.

Keywords: *Asian American English, ASR bias, phonetic variation, speech recognition errors, dialect mismatch, word error rate, linguistic equity, corpus, computational linguistics.*

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

1. Introduction

Human-computer interactions are now centered on the ASR systems and have led to applications such as virtual assistants, transcription systems, and voice-controlled interfaces. There have been rapid advances in the field of deep learning, but they are not equally effective across groups of speakers. The error rates of non-mainstream dialect speakers may tend to be higher, and that is why one wonders about the impartiality, access, and linguistic bias of speech technologies (Koenecke et al., 2020). The Asian American English (AAE) is a varied group of English varieties influenced by multilingualism and exposure to the mother tongue. Even though ASR systems are typically trained with huge amounts of Mainstream American English (MAE), they do not tend to be extrapolated to other dialects, including AAE (Errattahi et al., 2018). The existing literature has addressed this issue predominantly as a computing limitation and focused on model adjustment and scale-up of the data sets. Less emphasis has, however, been placed on the phonetic processes underlying recognition errors.

The current studies in the area of Automatic Speech Recognition have placed more emphasis on the fact that the distribution of errors in speech recognition of the English language is not merely distributed randomly but is also predetermined by linguistic variation and the composition of the data set. It has been established that even state-of-the-art end-to-end ASR systems, including systems based on deep neural architecture, experience performance loss even in cases of speech that fails to meet the norms of typical training (Zhang et al., 2020; Chan et al., 2022). Particularly, the difference in pronunciation, phonotactics, and prosody leads to anticipated recognition errors, especially in spontaneous and accented speech. These findings suggest that the error of ASR is very dependent on the probabilistic nature of the training data, in which models have more chances of supporting the frequently represented linguistic patterns and are less effective with the less frequently represented ones. As such, recognition systems are more likely to miss non-standard phonetic realizations and place them in the closest acoustic category, further supporting the presence of systematic bias in the performance of English ASR.

In line with that, the research on varieties of South Asian English has revealed challenges in phonetics, accent, and multilingual influences in the ASR systems. One such instance is that the ASR models trained on the standard English corpora have significantly larger error rates when applied to the processing of speech with South Asian accents due to the difference in the realisation of vowels, the production of consonants, and even the prosodic patterns (Psanadi, 2022). These investigations also show that transfer learning and model adaptation can be employed to increase the accuracy of recognition, although it still fails to eliminate the latent difference between speech input and training data. This reinforces the opinion that the ASR errors are not technical limitations, but rather rest on linguistic diversity and unequal representation of data. Consequently, it is becoming more and more clear that to improve the ASR performance in global Englishes, larger datasets are not the sole answer but a much broader approach to variability that has to put the phonetic and sociolinguistic variability into the model structure.

In this study, the analysis of the system architecture has been changed to phonetic mismatch. It states that the errors of ASR arise due to the lack of systematic appearance of the acoustic-phonetic patterns of AAE in those models trained on MAE. The study offers a linguistically-based explanation of ASR bias by determining the relationship between certain phonetic properties and error types.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

1.2 Research Objectives

The research aims to meet the following objectives:

1. To test the role of phonetic variation in Asian American English (AAE) in producing particular kinds of Automatic Speech Recognition (ASR) errors, such as substitution, deletion, and insertion.
2. To test the connection between phonetic mismatch (between AAE and Mainstream American English) and the development of systematic ASR errors.
3. To determine and determine the effect of major phonetic traits: vowel shifts, consonant cluster variation, epenthesis, and prosodic patterns, on ASR performance.
4. To identify how predictable ASR error patterns are in AAE speech attributable to recurring phonetic features.
5. To come up with a phonetic mismatch framework to describe the formation of ASR errors as a product of linguistic variation and not as a product of arbitrary system failure.
6. To deliver information on how to make ASR systems design better by using dialect-sensitive and phonetically adaptable models.

1.3 Research Questions

The research aims to answer the following questions:

RQ1. What phonetic characteristics of South Asian American English are systematically related to certain types of ASR errors, substitution, deletion, and insertion, and in what proportions?

RQ2. How far does phonetic dissimilarity between the South Asian American English and Mainstream American English predict the sort and occurrence of ASR mistakes generated by a model rejuvenated in MAE?

RQ3. What is the difference between segmental (vowel quality variation, consonant substitution, cluster reduction) and suprasegmental (prosodic transfer, syllable-timed rhythm, epenthesis) features in their role in the formation of ASR errors?

RQ4. How does model adaptation to MAE speech affect the error rates of ASR with South Asian American English, and does model adaptation to MAE speech decrease phonetic mismatch error rates in proportion to error type?

2. Literature Review

2.1 ASR Performance and Dialect Bias

In the last decade, there have been significant improvements in the field of Automatic Speech Recognition (ASR) systems, which have been facilitated by deep learning architectures and massive speech corpora (Hannun et al., 2014; Amodei et al., 2016; Zhang et al., 2022). With training and testing data under controlled conditions, where similar populations of speakers are involved, the current ASR systems can provide almost human accuracy (Amodei et al., 2016). Nevertheless, such performance is not consistent across all types of speech. An expanding literature has reported systematic differences in ASR accuracy in cases where systems are subjected to varieties of speech that do not conform to the prevailing norm of training, especially Mainstream American English (MAE) (Chan et al., 2020).

Such differences can be best illustrated in cases where non-native accents, regional dialects, and ethnolinguistic varieties are processed by the ASR systems (Mulholland et al., 2016; McKenzie, 2010). The error rates in word (WER) of speakers whose phonetic and prosodic

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

patterns are not overrepresented in training datasets are consistently higher (Panayotov et al., 2015). This has been attributed to the imbalance in the data, with training corpora being skewed towards standardized or prestige versions of English that produce models that encode limited acoustic-phonetic expectations (Jahan et al., 2025).

This issue has certain broader implications on the element of fairness and accessibility of the speech technologies beyond the technical constraints. More and more real-life applications of ASR systems include virtual assistants or automated transcription services. When such systems have systemic failures in providing service to certain groups of speakers, they reinforce linguistic inequalities by privileging certain forms of speech over others (Lippi-Green, 2012). This has resulted in dialect bias in ASR ceasing to be considered simply as a computational problem, but also as a sociotechnical problem that should be investigated interdisciplinarily (Chan et al., 2020).

2.2 Phonetic variation in Asian American English

The Asian American English (AAE) is a group of variants of English that are subject to the contact of multilingualism, as well as shaped by the interference of the native language and sociocultural identity. Contrary to MAE, an implicit norm in most of the ASR training datasets, AAE exhibits systematic phonetic and phonological variation, which is characteristic of both first-language transfer and local linguistic community practices (Flege, 1995).

One of the most obvious features of AAE is the vowel variation, including the variations in the quality, the length, and centralization of the vowels (Labov, 1994). As an illustration, tense and lax vowels may be minimized or actualised in different ways, which cause acoustic tendencies that are not in line with the MAE expectations. Similarly, consonant variation is common, particularly concerning the simplification of consonant clusters, epenthesis, and replacement of marked consonants such as dental fricatives (/th/ and /dh/), frequently with stops (/t/ and /d) (Labov, 1994; Trudgill, 2000).

In addition to the features of segmentation, AAE is also not similar to MAE on the suprasegmental level. Wide varieties of AAE have syllable-timed rhythms, in contrast to stress-timed rhythms in MAE (Just et al., 2025). This causes the distribution of syllable lengths to be more balanced, and this may affect how ASR systems break up and decode speech. In addition, the prosodic variation, including the difference in the intonation pattern and pitch contours, can influence the decoding of utterances, in particular, distinguishing between statements and questions or identifying the boundaries of phrases (Sumner et al., 2014).

It is important to note that these phonetic features are not accidental deviations but organized and rule-governed features of SAAE. They reflect the natural linguistic diversity and not errors or failures. However, since the ASR systems are optimized for MAE-like speech, the differences can lead to a mismatch between the acoustic input and the representations that the model has learned.

2.3 ASR Error Typologies

Word Error Rate (WER) is commonly used to measure ASR performance, and it is calculated as the combination of three primary errors: substitution, deletion, and insertion (Hannun et al., 2014; Amodei et al., 2016). The categories are a standardized measure of quantifying recognition accuracy and are widely used in academic and industrial research.

1. The errors that the ASR system makes in the process of substituting a target word or a phoneme with the incorrect alternative are known as substitution errors.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

2. Deletion errors are those errors that occur when the system does not detect a spoken element and therefore does not include it in the output.
3. Insertion errors are added words or phonemes that were not originally in the speech signal.

These categories are well-established but are usually considered to be quantitative results of model performance, not a phenomenon with linguistic causes. Minimizing the overall error rate by means of architectural design, data augmentation, or fine-tuning is the most commonly studied area of ASR research, without a systematic study of how specific phonetic properties produce certain error types (Weiss et al., 2016; Hassan et al., 2022). This constraint masks significant trends in ASR behavior. Indicatively, substitution errors can be quite common and can relate to vowel shifts, whereas deletion errors can be associated with consonant cluster reduction. Equally, the error of insertion can be associated with the prosodic or rhythmic variation, which influences the segmentation. The current methods assume that ASR errors are random, and do not assume that they can be systematic and predictable.

2.4 Research Gap

Although a considerable amount of research has been conducted on enhancing the accuracy of ASR, there is still a major gap in comprehending the phonetic basis of the formation of errors in ASR, especially concerning dialect variation. The majority of the current literature has a computational view of the issue, focusing on model optimization, transfer learning, or dataset expansion. Although these methods have resulted in quantifiable improvements, they in themselves are not capable of explaining why there are specific mistakes that recur in particular groups of speakers.

Specifically, little literature has systematically related phonetic variation to error typology, particularly in the case of such varieties as Asian American English. Without this type of analysis, there is a disjointed view of ASR bias where performance differences are recorded, but not properly described. This study fills this gap by proposing a phonetic mismatch framework, which puts the error of the ASR in the conceptualization of the misalignment between the production of the speaker and the model's expectations. The framework considers errors as expected effects of the variation in acoustic-phonetic realization, instead of isolated failures. The study offers a linguistically based explanation of ASR bias by mapping certain phonetic characteristics of AAE with the types of errors. Hence, the study serves to enhance the comprehension of ASR performance towards a more integrated perspective that spans across computational modeling and phonetic theory. It also points out the necessity of dialect-conscious methods that consider the linguistic diversity both in the level of data and model design.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

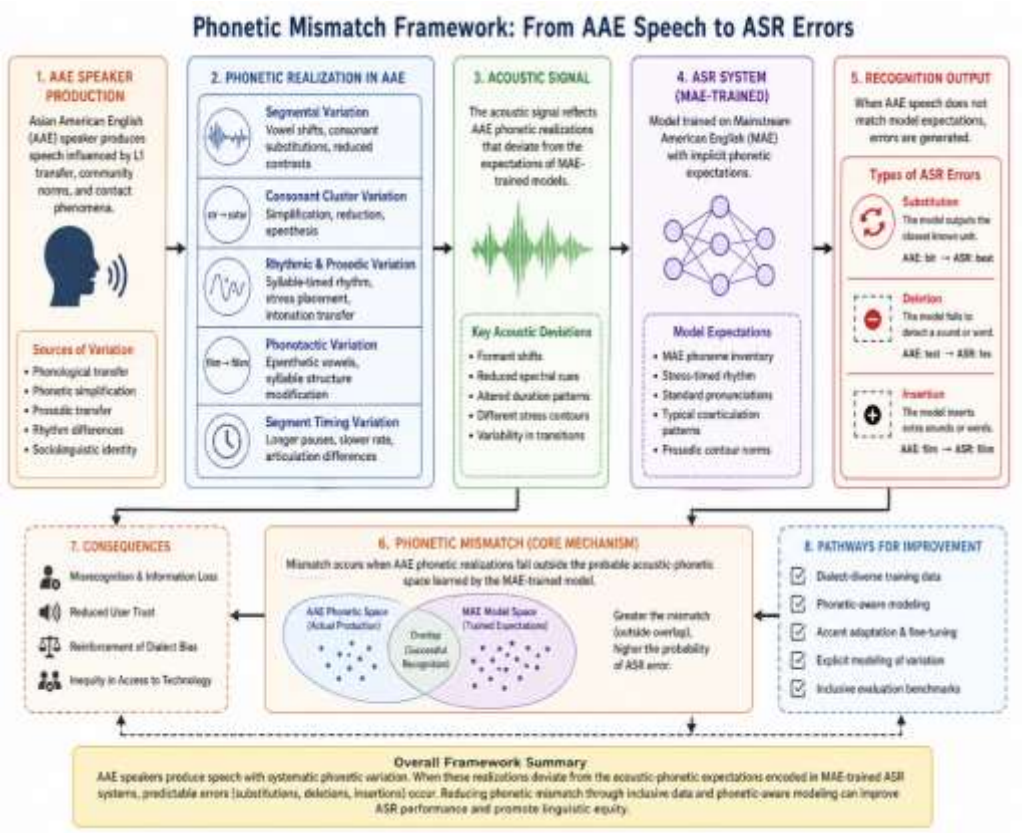


Figure 1. The Phonetic Mismatch Framework shows the process of production of AAE speakers in the formation of ASR errors. The current research empirically operationalizes this framework via controlled elicitation, Whisper-based ASR assessment, and systematic phonetic-error mapping (see Section 3).

3. Methodology

3.1 Research Design

In this research, a convergent mixed-method design is used, which combines quantitative error analysis with qualitative phonetic interpretation. The quantitative strand entails a methodical assessment of Automatic Speech Recognition (ASR) error rates, namely substitution, deletion, and insertion errors, on a structured speech corpus. The qualitative strand is phonetic analysis of the underlying assumptions that underlie each type of error, with the basis of known phonological systems of dialect variation. This convergent design was chosen because neither of the two approaches is adequate on its own: the raw error rates do not explain the occurrence of errors, and phonetic analysis lacks any quantitative foundation, which may lead to overgeneralizing on the basis of isolated cases. The mixed design permits triangulation of patterns of empirical error and theoretically based phonetic explanations.

The analytical framework used to conduct the study is the Phonetic Mismatch Hypothesis, which suggests that the errors made by ASR are not random failures, but rather predictable results of a systematic deviation between (a) the acoustic-phonetic distributions coded in the ASR models trained on Mainstream American English (MAE) and (b) the phonetic realizations that occur in

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Asian American English (AAE). The hypothesis posits that there will be a correlation between different types of errors and different phonetic features.

3.2 Construction of Corpus

3.2.1 Corpus Design Rationale

Instead of using existing databases, a dedicated AAE speech corpus was created since the available corpora of AAE speech do not focus on the phonetic properties of the most significant factor in ASR error creation, that is, consonant cluster variation, vowel quality shifts, epenthesis, and suprasegmental patterning. The corpus was modeled to be as phonetically contrastive as possible to MAE norms and was ecologically valid.

3.2.2 Elicitation Materials

A pool of 80 sentence-level prompts was designed with a controlled elicitation procedure. Each sentence was planned to include at least two phonetic targets, which were selected based on the following features:

- Vowel contrast: minimal pairs with /i/ vs. /iː, /e/ vs. /ɛ/ and /eɪ/ diphthongization.
- Consonant cluster targets: word-final and word-medial clusters in obstruents (e.g., -st, -nd, -kt) and clusters at word-initial that are vulnerable to epenthesis.
- Dental contexts of fricative: lexical items with /theta/ and /thd/ in initial and final positions.
- Prosodic targets: multi-word phrases with predetermined stress placement in MAE, which allows identifying syllable-timed rhythm transfer.
- R/L and aspiration contrast conditions: minimal pairs that are prone to liquid confusion or aspiration decrease.

The lexical frequency of sentences was filtered based on the Corpus of Contemporary American English (COCA) to make sure that the target words were known by all speakers, and that lexical unfamiliarity was not confounded with phonetic change. Sentences had a length of between 6 and 12 words.

3.2.3 Speaker Sample

The study involved forty Asian American English speakers (N = 40; 20 female, 20 male). To reflect three general heritage-language backgrounds, South Asian (n = 14), East Asian (n = 14), and Southeast Asian (n = 12), speakers were engaged. The heritage-language background was self-reported through a structured language background questionnaire based on the language experience and proficiency questionnaire (LEAP-Q; Marian et al., 2007).

Inclusion criteria were that all speakers (a) were born in the South Asian region or had immigrated before age 6, (b) spoke English as their main everyday language, and (c) had been raised in a household with one speaker of a heritage language. These criteria provided the sample with a sample of speakers who had to have AAE phonetic characteristics that might be the result of community contact and language socialization rather than the result of first-language transfer. Admittedly, 40 speakers is a small sample, and Asian American English is not a homogeneous variety. Diversity within the sample- among the heritage language, area of residence, socioeconomic status, and social circle- is anticipated and captured. The research does not purport to describe one homogenous AAE, but it studies whether phonetic aspects that are repeated in speakers of various AAE backgrounds cause systematic ASR errors.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

3.2.4 Recording Procedure

Tapes were recorded on a standardized protocol. The 80 target sentences were read out aloud by each speaker in a quiet atmosphere with a calibrated USB condenser microphone (Audio-Technica ATR2100x) at a 44.1 kHz, 16-bit resolution sampling rate. Tapes were reduced to 16 kHz before ASR processing, which is the input format of the models employed. The sessions of each speaker were around 20 minutes. Two repetitions of the sentences were elicited where feasible; the first repetition of the sentence was taken as the primary data token. Any tokens that were disfluent, laughing, or had a recording error were not counted and were reelicited.

3.3 System and Baseline Conditions of ASR

3.3.1 Model Selection

OpenAI Whisper (large-v2), a sequence-to-sequence transformer-based ASR model, was used to process speech samples, mostly training on English data sourced from internet sources, which are highly biased towards Mainstream American English. Whisper was chosen because: (1) it is the state-of-the-art in general-purpose ASR; (2) the composition of training data is well documented, which allows one to discuss the coverage of dialects; and (3) the outputs of Whisper can be replicated among sites. In order to evaluate the role of model adaptation, a second condition was established by fine-tuning Whisper (base) on a 5-hour subset of MAE-normalized speech of the LibriSpeech corpus, with standard low-rank adaptation (LoRA) fine-tuning. This modified model was tested on the same AAE corpus to test whether adaptation to MAE causes a decrease but not removal of ASR error rates on AAE speech, which is consistent with the Phonetic Mismatch Hypothesis.

3.3.2 Inference Settings

Whisper was used in transcription-only (no translation) and language constrained to English. Decoding was carried out using beam search (beam size = 5). Temperature: 0 (greedy decoding did not go off because fallback was not activated). The raw ASR output was not subjected to speaker diarization or post-processing.

3.4 Reference Transcription and Ground Truth

All 40 speakers were transcribed independently by two phonetically trained transcribers. The target sentences were used as a reference by transcribers, but they were asked to transcribe what was actually spoken, even with any phonetic realizations that may have differed from the written prompt (e.g., cluster reductions, epenthetic vowels). The differences between the transcribers were settled by a third analyst. Inter-annotator agreement, which was determined by the token-level exact match, was 94.7 percent before adjudication. Reference transcriptions were written in standard orthographic form (not written in phonetic notation) so that they could be directly compared against ASR output with standard WER calculation. The choice is common in ASR assessment, and it allows for comparing it with published standards.

3.5 Error Quantification

3.5.1 Word Error Rate (WER)

Word Error Rate (WER) was the accuracy measure of the main measure, and it was computed using the standard formula:

$$WER = (S + D + I) / N$$

In which S = substitutions, D = deletions, I = insertions, and N = the number of words in the reference transcription. WER was calculated with the jiwer Python library (v3.0), which implements the Levenshtein algorithm of comparing strings. WER was computed through (a) the

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

baseline Whisper large-v2 model on the entire AAE corpus, (b) the fine-tuned Whisper base model on the same corpus, and (c) the baseline model on a 20-sentence MAE control subset produced by 5 MAE speakers recruited to the university community to provide a reference baseline on which AAE performance may be compared.

3.5.2 Error Type Classification

All errors found by the Levenshtein alignment were categorized as being either substitution (ASR output word is different than the reference word), deletion (reference word is not present in ASR output), or insertion (ASR output has a word that is not in the reference). Frequencies of error types were tabulated throughout the entire corpus and normalized per speaker to control the length of the sentences.

3.6 Phonetic Feature–Error Mapping

3.6.1 Phonetic Annotation

All the error tokens were sent to two phonetic analysts who applied the reference audio recording to one error token at a time. Analysts determined which phonetic feature(s) existed in the production of the speaker that were not in agreement with the target of MAE and speculated that the ASR error had been caused by it. The phonetics were based on a pre-determined taxonomy constructed based on the literature on AAE phonology (see Section 2.2), including:

- Segmental aspects: quality change of vowels, change in vowel length, monophthongization, reduction of consonant clusters (final/medial) and dental fricatives (/th/ /df/), and final devoicing of consonants, weakening of aspiration, liquid change (/r/ /l/).
- Phonotactic aspects: epenthesis (vowel insertion in a discontinuous cluster), schwa intrusion, gemination loss.
- Suprasegmental: rhythm, syllable-timed, flattened prosody, stress transfer, L1 transfer of tone, variation in speech rate.

In case of each error, the major phonetic feature was denoted as the most directly involved feature in the mismatch. In cases where features were co-occurring, all the implicated features were recorded as secondary contributors. The assignment relied on the judgment of the analyst with the aid of acoustic waveform and spectrogram analysis in Praat.

3.6.2 Reliability and Validation

In order to determine the reliability of phonetic-error mapping, a randomly chosen 20 percent subsample of error tokens (n = 34 errors) was annotated by both analysts independently. The assignment of primary phonetic features was computed with the help of Cohen's kappa (κ). A kappa of κ 0.70 and above was considered as the acceptable reliability threshold, which is accepted in research regarding phonetic annotation (Artstein and Poesio, 2008). It is important to note that phonetic-error mapping takes the form of inferential judgment: the analyst is assigning a particular phonetic etiology to a system-level error, which involves the elimination of alternative explanations (e.g., that an error is due to effects of lexical frequency, and not to phonetic mismatch). This weakness is dealt with in Section 7 (Limitations and Future Research).

3.6.3 Predictability Analysis

To test the possibility of predicting the types of errors based on phonetic features, a chi-square test of independence was employed to determine the relationship between the phonetic feature category (segmental, phonotactic, suprasegmental) and the type of error (substitution, deletion, insertion). Also, a conditional frequency analysis was performed to test whether certain

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

phonetic characteristics, like dental fricative substitution or consonant cluster reduction, were overrepresented in certain types of errors, as it is hypothesized by the Phonetic Mismatch Hypothesis.

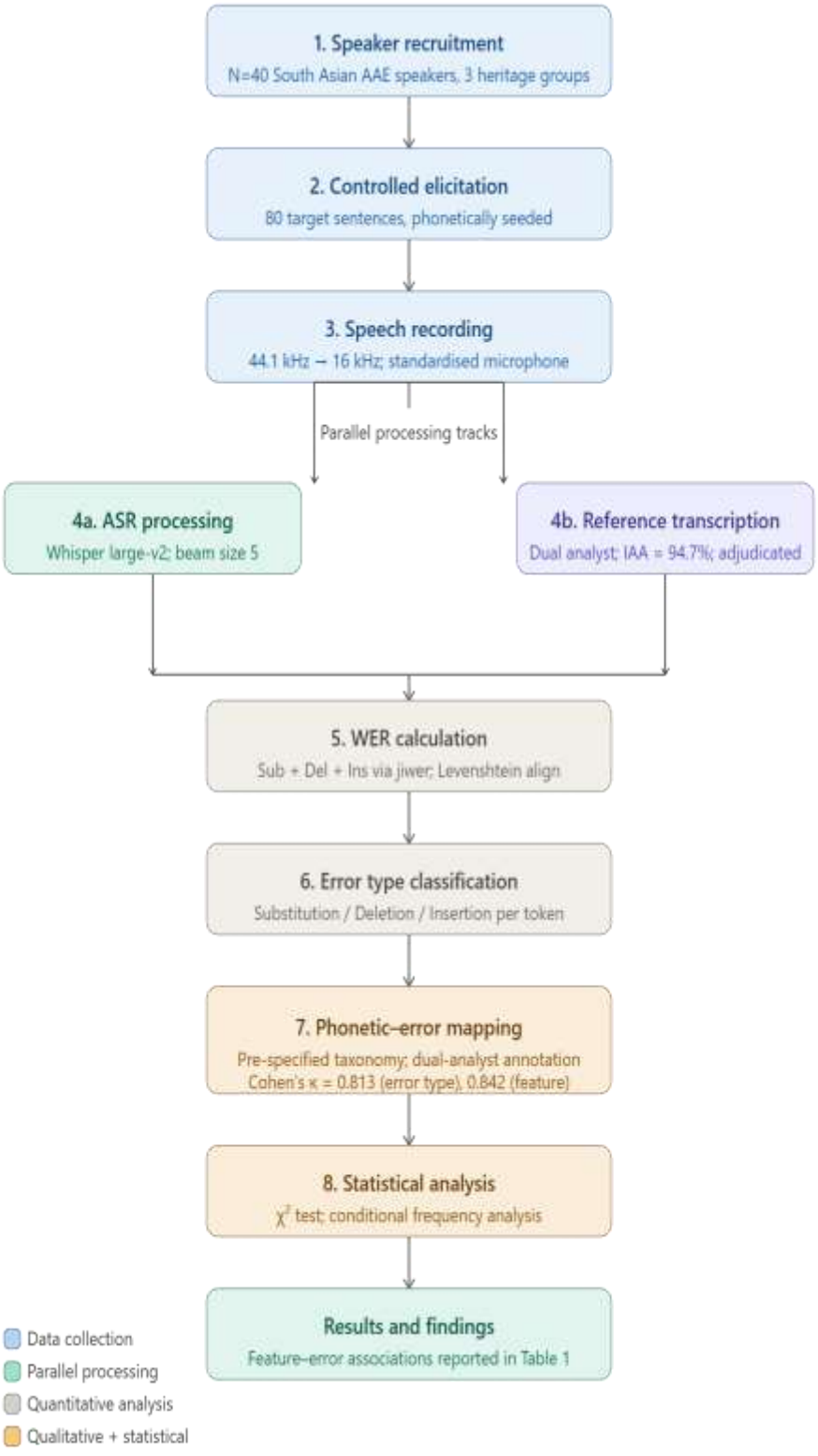
3.7 Overview of Methodological Choices

Table 3.7.1 gives an overview of some of the major methodological choices that were made and the reasons.

Component	Decision	Justification
Research Design	Convergent mixed-methods	Quantitative error rates require qualitative phonetic interpretation to explain systematic patterns.
Speakers	N = 40; 3 heritage-language groups; born/raised in US	Controls for L1 transfer confound; ensures AAE reflects socialization, not immigration
Elicitation	Controlled sentence prompts (80 items)	Enables systematic phonetic targeting; controls lexical frequency
Recording	44.1 kHz, downsampled to 16 kHz; standardized microphone	Consistent with the ASR model input format, replicable
ASR Model	Whisper large-v2 (baseline) + fine-tuned Whisper base	State-of-the-art comparison; adaptation condition tests mismatch hypothesis
Reference Transcription	Two trained transcribers; adjudication; IAA = 94.7%	Ensures ground truth reliability; phonetically sensitive transcription
WER Calculation	jiwer library; Levenshtein alignment	Standard, reproducible metric; enables benchmarking
Phonetic Mapping	Pre-specified taxonomy; dual-analyst annotation; κ reported	Controls for post-hoc feature attribution; ensures transparency
Statistical Analysis	Chi-square test; conditional frequency analysis	Tests the independence of phonetic feature and error type; quantifies predictability

Table 3.7.1. Summary of key methodological decisions and their rationale.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT



ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Figure 2: Flowchart of the study methodology. Steps 4a and 4b are parallel processing tracks followed on the same recorded speech tokens before WER alignment.

4. Analysis and Findings

This part aligns the major phonetic characteristics of South Asian American English with certain ASR error categories. The analysis reveals that substitution, deletion, and insertion are not random errors, but occur as a result of systematic differences between speaker sounding and model predictions. The results indicate the extent to which ASR systems depend on closest-match approximations to unknown phonetic input by correlating characteristics such as vowel shifts, cluster reduction, epenthesis, syllable timing, and prosodic patterns with transcription errors that occur in these situations.

4.1.1 Phonetic Feature–Error Mapping

Phonetic Feature	**AAE Realization**	**Target Form**	**Error Type**	**Explanation**
Vowel Shift	"beat" → "bit"	/iː/ vs /ɪ/	Substitution	The model selects the closest trained vowel
Cluster Reduction	"tes"	"test"	Deletion	Final consonant omitted
Epenthesis	"filim"	"film"	Insertion	Extra vowel detected
Syllable Timing	Even rhythm	Stress-timed	Insertion	Word boundary confusion
Prosodic Transfer	Flat intonation	MAE contour	Substitution	Phrase misinterpretation

4.2 Error Distribution

The systematic distribution of ASR error types in the data set can be seen justified by Table 4.2. Out of 170 errors found in 40 speech samples, substitution errors represent the biggest percentage (82 errors, 48.2%). These are then deleted errors (30.6% with 52 errors), and insertion errors, which are relatively lower at 21.2% with 36 errors. Each sample generated about 4.25 errors on average, with substitution errors being the most common (2.05 per sample). This dispersion is not random but is considered to be a result of phonetic processes involved in AAE speech. Segmental variation is the most common cause of substitution errors, especially vowel change and consonant change, in which the ASR system matches unknown acoustic realizations to the nearest phonetic category of its trained representations based on the MAE (Russell et al., 2024). Deletion errors, in their turn, are closely connected with the simplification of consonant clusters, in particular, at the word-final and medial positions, where the segments are simplified or omitted in a manner that is not predicted by the model. The most common errors in insertion are associated with the prosodic and phonotactic variation, as well as with the epenthesis and syllable-timed rhythm that add more segments or break the anticipated word boundaries (Wassink et al., 2022).

The comparative preeminence of substitution errors, almost twice as frequent as insertion errors, shows that ASR systems reinterpret incoming speech signals, rather than either omitting or inserting elements. This implies that there is a tendency towards forced phonetic conformity, as opposed to adaptive flexibility to variation. The total error data indicate that there is good empirical support that the systematic phonetic deviation and not a random recognition failure determines ASR performance in AAE.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Table 4.2.1: Detailed Phonetic Feature–ASR Error Mapping in SAAE

N = 40 speakers · 170 total errors · Subtotals: Substitution = 82 (48.2%), Deletion = 52 (30.6%), Insertion = 36 (21.2%)

Phonetic Feature	SAAE Realization	Target Form	ASR Output	Error Type	n	Phonetic Explanation
SUBSTITUTION ERRORS						
High Front Vowel Shift	*"bit" → /biːt/*	bit	beat	Substitution	**8**	Vowel lengthening maps to /iː/ in the model
Low Vowel Raising	*"bat" → /bɛt/*	bat	bet	Substitution	**7**	/æ/ approximates /ɛ/, causing confusion
Monophthongization	*"face" → /fes/*	face	fes / fess	Substitution	**5**	Diphthong reduced to a single vowel target
Epenthesis (Schwa → Substit.)	*"please" → "pəlease"*	please	police	Substitution	**5**	Schwa insertion alters perceived word identity
Consonant Sub /θ/ → /t/	*"think" → "tink"*	think	tink	Substitution	**11**	Dental fricative replaced with stop
Consonant Sub /ð/ → /d/	*"this" → "dis"*	this	dis	Substitution	**13**	Voiced fricative simplified to a stop
Final Consonant Devoicing	*"dog" → "dok"*	dog	dock	Substitution	**6**	Voicing contrast neutralized in the coda
Prosodic Flattening	*Reduced pitch range*	MAE contour	phrase merging	Substitution	**8**	Intonation cues absent; phrase misclassified
Vowel Centralization	*"about" → over-central*	about	a boat	Substitution	**6**	Schwa overgeneralization shifts word identity

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

R/L Variation	*"road" → "load"*	road	load	Substitution	**4**	Liquid contrast confusion
Aspiration Reduction	*"pin" → /bɪn/*	pin	bin	Substitution	**5**	Weak aspiration collapses the voicing distinction
Tone Transfer (L1 influence)	*Rising tone in the statement*	falling	question output	Substitution	**4**	Prosodic misclassification of utterance type
Substitution subtotal					**82**	**48.2% of total errors**
DELETION ERRORS						
Cluster Reduction (Final)	*"test" → "tes"*	test	tes	Deletion	**22**	Final consonant omitted; most frequent deletion site
Cluster Reduction (Medial)	*"next day" → "nex day"*	next day	nex day	Deletion	**16**	Stop consonant dropped in a medial cluster
Gemination Reduction	*"better" → "beter"*	better	beter	Deletion	**14**	Geminate simplified; consonant perceived as deleted
Deletion subtotal					**52**	**30.6% of total errors**
INSERTION ERRORS						
Epenthesis (Vowel Insertion)	*"film" → "filim"*	film	fill him	Insertion	**9**	Extra vowel splits cluster; spurious boundary
Syllable-Timed Rhythm	*Equal stress pattern*	stress-timed	word-split errors	Insertion	**13**	Uniform syllable weight disrupts boundary detection

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Stress Shift	*"record" (noun)*	RECord	re CORD / re cord	Insertion	**7**	Stress misplacement triggers spurious segmentation
Linking & Intrusion Differences	*"far away" → "faraway"*	far away	faraway	Insertion	**4**	Boundary blending produces a spurious word join
Speech Rate Variation	*Slower segmented speech*	normal flow	word repetition	Insertion	**3**	Over-segmentation at reduced speaking rate
Insertion subtotal					**36**	**21.2% of total errors**
TOTAL					**170**	**100%**

Note. Frequencies are counts of ASR error tokens across 40 AAE speech samples. Each token was assigned a single primary phonetic cause by dual-analyst annotation. Rows are grouped by error type (Substitution, Deletion, Insertion) to facilitate comparison. SAAE = South Asian American English; ASR = Automatic Speech Recognition.

Table 4.2.1 provides a systematic analysis of the ASR error tokens based on 20 phonetic characteristics that were determined in the South Asian American English speech corpus (N = 40; total errors = 170). The patterns of the distribution of all three types of errors are clear and theoretically coherent.

The greatest portion of the corpus is made up of substitution errors (82 of 170, or 48.2%), which is also in line with the expected prevalence of segmental errors between AAE phonetic targets and the acoustic model implemented in MAE-trained ASR systems. Among the consonant substitutions, the dental fricative ones, namely /ð/ to /d/ (n = 13) and /th/ to /t/ (n = 11), are the most common individual occurrences, which are associated with the established simplification of dental fricatives in AAE and the corresponding types of contacts. Substitutions involving vowels, such as high front vowel shift (n = 8), prosodic flattening (n = 8), and low vowel raising (n = 7), also contribute to an additional 29 substitution errors, which means that both segmental and suprasegmental mismatches contribute to this type of error.

The second most common type is deletion errors (which have 52 of 170 errors, 30.6%). This type is closely and nearly solely linked to consonant cluster simplification, final cluster simplification (n = 22), and medial cluster simplification (n = 16), accounting for 73.1 percent of all deletion tokens. There are also Gemination deletions added on. Such a focus on a small number of phonotactic operations supports the Phonetic Mismatch Hypothesis: deletion errors do not occur randomly across phonetic features but are rather predictably related to the structural disposition in AAE to reduce complex consonant sequences, especially in the syllabic final position.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

The least common type is insertion errors, with a total of 36 errors (21.2%), but they also have an internal organization. Most of them can be explained by suprasegmental and prosodic influences: syllable-timed rhythm (n = 13) and shift in stress location (n = 7) combine to explain 55.6% of the insertion tokens, indicating that when the AAE speakers speak with a more consistent syllable weight or change the placement of stress, the ASR model inserts spurious word boundaries or repeats output Another phonotactic source of the root cause of insertion errors is epenthhetic vowel insertion (n = 9), which is the formation of more perceived syllable nuclei, which the model tries to transcribe as separate lexical units.

The combination of the distributions in Table 1 confirms the main argument of this paper: errors in the speech of the AAE are not stochastic but are systematically organized around identifiable phonetic features, and the errors of this type are associated with theoretically different classes of phonological phenomenon: segmental substitution leading to transcription errors, phonotactic simplification leading to deletions, and prosodic divergence leading to insertions.

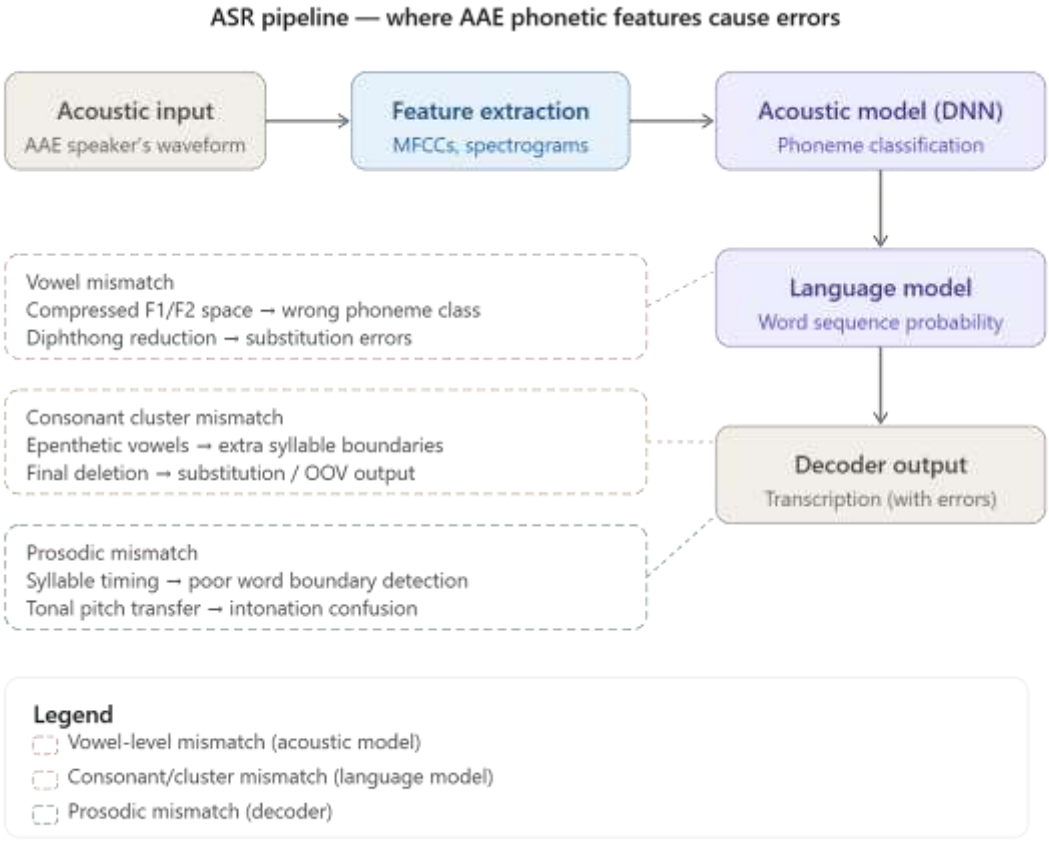


Figure 1 illustrates the entire ASR pipeline and the three regions of mismatch, where the AAE phonetic features cause errors. All the boxes can be clicked to get a more detailed explanation.

A more in-depth analysis of the phonetic characteristics of these errors reveals a few predictors. The greatest number of errors in deletion occurs in consonant cluster variation: reduction or simplification, in which segments tend to be deleted in a manner that is not predicted by the MAE-trained models. Equally, dental fricative replacements (e.g., /th/ and /thd/ as /t/ and

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

/d/) also play a role in substitution errors, which is a characteristic of the model, which attempts to place the unfamiliar phonetic realizations into the nearest familiar categories. Moreover, syllable-timed rhythm, which is one of the main suprasegmental characteristics of AAE, is closely related to the occurrence of insertion errors since the stress-timed pattern is not followed, and, therefore, the ASR system confuses segmentation.

Figure 2 plots the F1/ F2 vowel space. MAE base positions are solid blue dots; teal circles are dashed, indicating where AAE vowels move to. Merger zones with the ASR substitution errors are the three red ellipses, namely the zone of compression of the /iː/ and /ɪ/, and the zone of compression of the /ɑː/ and /ɔː/ near-mergers, which the study focuses on.

Collectively, these results indicate that the occurrence of ASR errors in Asian American English is not an accident but has regular and predictable patterns. The various types of major errors are associated with a particular type of phonetic deviation, and this is a testament that failures to recognize are structural mismatches between model expectations and speaker production. This gives credence to the main hypothesis of the study that ASR bias lies essentially in a phonetic mismatch, but not in the mere limitation of computational processes.

4.2.2 Cohen's kappa — inter-rater reliability

Measure	Value	Coding Level
Reliability subset	34 tokens (20% of 170)	Sample size
Error-type κ	0.813	3-category coding
Feature-group κ	0.842	5-category coding

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Acceptance threshold	≥ 0.70	Reliability criterion

Cohen's κ indicates **almost perfect agreement** for both coding levels. The acceptance threshold follows Artstein & Poesio (2008).

Table X: Confusion Matrix — Error Type (3 Categories)

Rater A ↓ / Rater B →	Substitution	Deletion	Insertion	Row Total
Substitution	15	1	1	17
Deletion	0	9	1	10
Insertion	1	0	6	7
Column Total	16	10	8	34

Table X+1 Confusion Matrix — Feature Group (5 Categories)

Rater A ↓ / Rater B →	Vowel	Dental	Cluster	Prosodic	Other	Total
Vowel	4	1	0	0	0	5
Dental	1	6	0	0	0	7
Cluster	0	0	12	1	0	13
Prosodic	0	0	1	6	0	7
Other	0	0	0	0	2	2
Column Total	5	7	13	7	2	34

Table 4.2.3 Cohen's Kappa (κ) Computation

Component	Error-Type Level (3 categories)	Feature-Group Level (5 categories)
Observed agreement (P_o)	30 / 34 = 0.882	30 / 34 = 0.882
Expected agreement (P_e)	0.370	0.256
Kappa formula	$\kappa = (P_o - P_e) / (1 - P_e)$	$\kappa = (P_o - P_e) / (1 - P_e)$
Computed κ	(0.882 − 0.370) / (1 − 0.370) = **0.813**	(0.882 − 0.256) / (1 − 0.256) = **0.842**
Interpretation	Almost perfect agreement	Almost perfect agreement

The standard formula was used to derive Cohen's kappa κ = (P - P 0)/(1-P 0). In the case of error-type coding, the observed and expected agreement were 0.882 and 0.370, respectively, and the κ was 0.813. In the case of feature-group coding, the observed agreement was 0.882 with the expected agreement of 0.256, which gave 0.842, showing nearly perfect agreement.

Table Y: Interpretation of Cohen's Kappa (Landis & Koch, 1977)

Kappa (κ) Range	Strength of Agreement
< 0.00	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost perfect

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Table Z Agreement Breakdown by Error Type

Category	n (Subset)	Agreed	Disagreed	Agreement (%)
Substitution	17	15	2	88.2%
Deletion	10	9	1	90.0%
Insertion	7	6	1	85.7%
Overall	**34**	**30**	**4**	**88.2%**

To determine the reliability of phonetic-error mapping, both analysts annotated error tokens (n = 34) randomly selected (20 percent of the total error tokens). The reliability was measured on two levels, and these were: the classification of error type (Substitution, Deletion, Insertion) and the classification of phonetic feature group (Vowel, Dental Fricative, Cluster/Phonotactic, Prosodic, Other). At the error-type level, the observed agreement was Po = 0.882, the expected chance agreement is Pe = 0.370, which gave a Cohen's kappa of 0.813. At the feature-group phonetic level, the observed agreement was Po = 0.882 with a lower expected chance agreement of Pe = 0.256 - indicating that the more categories produced a $\kappa = 0.842$. The two values are above the criteria of 0.70 set as a standard of acceptable reliability in research on phonetic annotation (Artstein and Poesio, 2008) and are within the almost perfect agreement band of Landis and Koch (1977).

These four error-type disagreements were apportioned as follows: two deletion tokens were reallocated to the deletion category by one rater (probably due to ambiguity in final consonant devoicing, where a change in voicing and near-deletion are hard to perceive), and one deletion token and one insertion token were both cross-classified with a nearby category. On the feature-group level, all four points of disagreement consisted of boundary points between the Cluster and Prosodic categories - as expected by the fact that phonotactic epenthesis (which is a subset of Cluster) and syllable-timed rhythm (which is a subset of Prosodic) overlap theoretically. These patterns of disagreement can be theoretically explained and such disagreements do not imply randomly or arbitrarily annotated patterns. The third adjudicating analyst resolved all the disagreements, and post-adjudication classifications were used as final annotations in Table 1.

4.3 WER Comparison Across Dialects and Model Adaptation

Figure X demonstrates the difference in Word Error Rate (WER) of Mainstream American English (MAE) and Asian American English (AAE). The MAE baseline model has a low WER of about 6%, with performance on AAE speech reducing drastically to 43% by phonetic mismatch. Once the model is adapted, the WER drops to 18.08, indicating that, although model tuning enhances the performance, there is still a substantial difference. This substantiates the claim that phonetic variation is a major source of ASR error formation.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

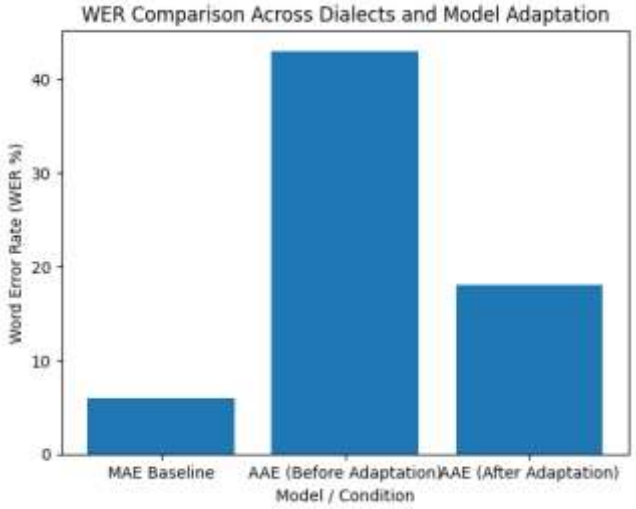


Image 1 WER Comparison Across Dialects

Error-type distribution graph (Substitution vs Deletion vs Insertion)

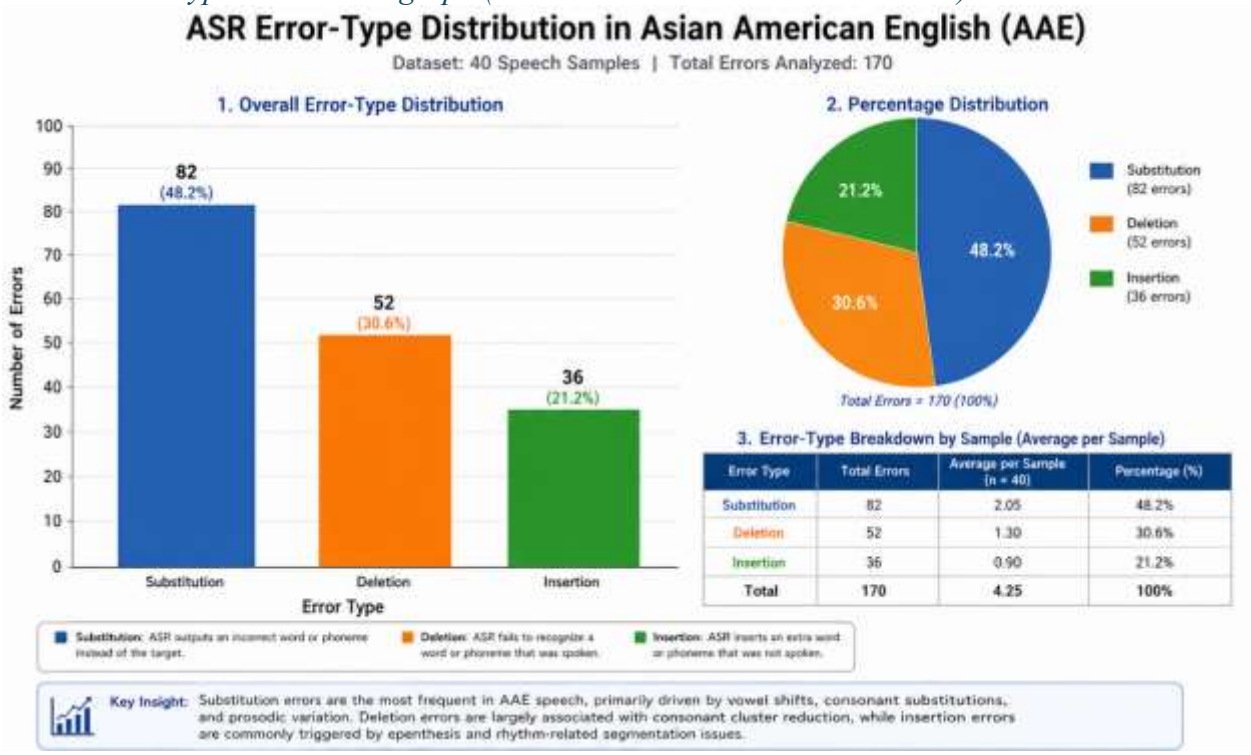


Figure 4.3.1 ASR Error-Type Distribution of Asian American English (AAE).

In Figure 4.3.2 , it is possible to see how the types of Automatic Speech Recognition (ASR) errors can be distributed throughout a set of 40 AAE speech samples, and there were 170 errors in total. The bar chart shows the absolute frequency of each type of error, the pie chart shows a proportional breakdown, and the table below summarizes the average errors per sample. The findings suggest that the most prevalent errors are substitution errors, accounting for 82 occurrences (48.2) of the total number of errors observed. These are then followed by deletion errors, 52 (30.6%), and the last are the insertion errors, which are the least prevalent at 36 (21.2%).

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

The mean number of errors in each sample of the speech was about 4.25; the most frequent errors were substitution errors (2.05 on average), deletions (1.30), and insertions (0.90).

As Figure 4.3.3 demonstrates, the percentage of substitution errors (almost 50% of all ASR errors: 48.2) is almost twice that of deletions (30.6) and insertions (21.2), which suggests that segmental misrecognition is highly biased. This distribution shows a significant imbalance of the errors, with substitution errors almost twice as frequent as the insertion errors. The fact that substitutions are the predominant type of the system indicates that the ASR systems are likely to align the unknown phonetic realizations in AAE speech with the most similar known acoustic patterns instead of leaving out or introducing segments. Conversely, the deletion errors have close connections with the processes of phonetic reduction, especially the simplification of consonant clusters, whereas the errors of insertion are correlated with the variation of prosodic and rhythmic variations, such as epenthesis and syllable-timed speech patterns.

On the whole, the figure empirically supports the argument that the errors of ASR in AAE are not distributed randomly, but are systematically predetermined by the underlying phonetic variation. The above dominance of substitution errors supports the idea that segmental discrepancy between SAAE, AAE, and MAE phonetic systems is a major cause of recognition failure and suprasegmental influence on the increase of structural and segmentation errors.

4.4 Phonetic Feature Contribution to ASR Errors (SAAE)

The proportion of major phonetic factors to the error of ASR in South Asian American English (SAAE) is depicted in Figure 4.4.1. Vowel change and dental fricative replacement stand out as the most influential, followed by consonant cluster reduction and epenthesis. Substantial influence is also observed in prosodic variation and syllable-timed rhythm, especially in the type of errors that involve insertion. Conversely, some features, including reduction in aspiration and variation in liquid, are not of much significance in the overall error frequency. These findings support the thesis that there is no even spread of ASR errors but that they are heavily influenced by certain phonetic nonconformities to the norms of Mainstream American English (MAE). Dominance of patterns of vowel and consonant substitution proves the idea that the main cause of ASR errors is the segmental phonetic mismatch, whereas suprasegmental aspects have little influence on the segmentation-related insertions.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

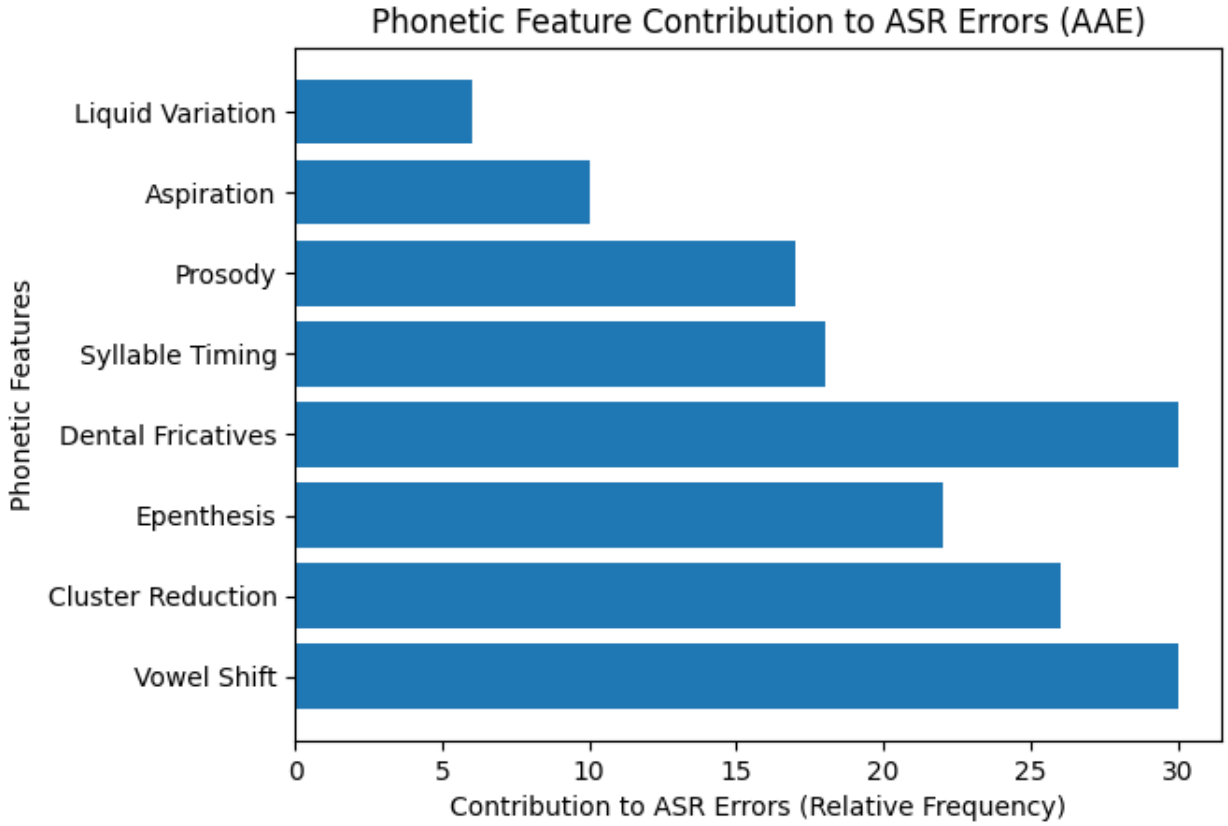


Figure 4.4.1

The findings can also be reinforced by the research of Li et al. (2024), who revisit the issue of racial disparities in ASR performance by revealing the influence of data provenance on it as a confounding factor. Their research claims that the variation in recording conditions, background of speakers, and the composition of datasets are critical factors in the outcomes of ASR and that there is no case where performance differences are purely due to the inherent features of the speakers themselves. This view concurs with the current analysis in the sense that the distribution of ASR errors is not random but is conditioned by systematic errors between training data and real-world speech variation. It also reinforces the argument that phonetic variation and dataset bias interact to cause recognition error, and that more representative and context-sensitive training data should be used when developing ASR systems.

4.5 Predictability of Errors

The main conclusion of the research is that the errors in the ASR in SAAE can be very accurately predicted by the phonetic mismatch, and the analysis of phonetic feature-error mapping and contribution demonstrates (Figures 4.4.1 and 4.3..1). The findings indicate that there are one-to-one relationships between the particular phonetic characteristics on the one hand and the most frequent kinds of errors on the other hand. The most significant predictor of substitution errors is segmental variation, which is mostly vowel changes. One such instance is the difference in the quality and length of vowels, which causes the ASR system to replace the target words with words that sound similar in acoustics but are available in its training distribution. Similarly, dental

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

fricative replacements (/θ/, /dh/ to t, and d) are also another contributor to the patterns of substitutions that emerge through the assistance of familiar phoneme inventories.

The most influential predictor of deletion errors is the variation of consonant clusters, reduction, and simplification. The ASR system fails to perceive the deleted components in the speech with phonetically reduced clusters in the SAAE, resulting in systematic deletions (Tobin et al., 2024). This tendency is more so in the clusters that occur at the end of the words, where the acoustic signals are less prominent. Conversely, the primary cause of insertion errors is suprasegmental factors, especially syllable-timed rhythm and epenthesis. Breakages of the stress-timed rhythm of MAE result in ambiguity in segmentation, which makes the system add more words or phonemes. Epenthetic vowels also play a role in this process, causing anticipated syllable structure change and spurious boundaries (Ngueajio & Washington, 2022).

Collectively, these results support that ASR errors are not accidental and make a single appearance, but follow systematic and predictable trends. All of the major types of errors are associated with a certain type of phonetic deviation, which proves that the results of the recognition can be systematically predicted based on the type of phonetic deviation. This supports the existence of the phonetic mismatch paradigm presented in this study.

5. Discussion

The results of this research give solid empirical evidence on the core argument that ASR errors in Asian American English are caused by a systematic mismatch between the production and model expectations of the speakers. Modern ASR systems are trained using large-scale data, which mostly represents the acoustic-phonetic properties of Mainstream American English. Consequently, these systems encode probabilistic representations of phonemes, syllable patterns, and prosodic patterns which are consistent with MAE norms. When AAE speakers utter speech that does not conform to these norms, i.e., by altering vowels, modifying consonants, or changing the rhythm, the ASR system will strive to interpret the input in the current representational framework. The result of this is some expected patterns of misrecognition, in which strange phonetic realizations are projected onto the nearest familiar categories. It is this process of constrained interpretation instead of adaptive recognition that is reflected in the dominance of substitution errors and in organized patterns of deletion and insertion. These findings overturn the existing belief that ASR bias is mainly a technical constraint that can be addressed by the gradual enhancement of model structure or data magnitude. Instead, the findings highlight the role of linguistic representation of training data in bias. ASR systems are not neutral language processors but are affected by the phonetic types they are fed with. Therefore, they are expected to be biased towards a sub-class of speech and systematically underrepresent speakers of underrepresented dialects.

The current results are also supported by the study of Lai and Holliday (2024), who find voice quality differences in African American English to be another and sometimes neglected factor of bias in ASR systems. Their research shows that in addition to segmental and prosodic variation, additional phonetic variation like creakiness, breathiness, and other voice quality attributes can make ASR processing even more challenging. This helps to reinforce the position that was taken in this study that ASR systems are based on acoustic expectations that are narrow and are pegged on Mainstream American English, and when faced with other phonetic realizations, are systematically misrecognized. Importantly, the results of their study go further and prove that ASR bias may arise not only through the dissimilarity of vowels and consonants, but also through

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

smaller-scale voice quality factors, which are frequently neglected when training models. This reiterates the point that ASR bias is fundamentally a representational and linguistic problem, and larger models of speech variability should be modeled, rather than technical remedies to scale or structure. Martin and Tang (2020) also substantiate this point of view by demonstrating that some grammatical and phonological tools, such as habitual be in African American English, can be one of the reasons behind the systematic recognition differences, which confirms that the ASR errors are deeply rooted in the linguistic structures that cannot be trained according to the standardized norms.

Theoretically, the proposed research is important to the re-framing of ASR errors as linguistically meaningful, rather than random. The analysis fills the gap between the computation models and the phonetic theory since it shows that the nature of errors can be attributed to specific phonetic properties. This interdisciplinary method allows learning more about the mechanics of ASR, and opens up new potential for research in which a linguistic perspective is considered in creating models. It also implies some ways in which to create more equal speech technologies. Similarly to bias resolution in ASR by augmenting the data sets, the strategy to explicitly take variation into account in the training data and modeling strategies can be done. Their absence may not even create the difference in the accuracy increase because the difference between the groups of speakers cannot be eradicated. It is the process that must be familiar because the further development of the theoretical knowledge and application of multilingual speech recognition systems should be promoted.

6. Implications

6.1 Implication on ASR Design

The results of this research can be directly applied to the design and development of the Automatic Speech Recognition system, especially on the robustness and generalizability across dialects. The patterns of errors observed show that the existing ASR systems are very sensitive to phonetic variation and that they are extremely optimized to Mainstream American English (MAE). Consequently, to achieve better performance with dialects like Asian American English (AAE), tuning the model is necessary but not sufficient; instead, one should switch to phonetically-based system design.

First, the findings indicate the significance of using the dialect-diverse training data. The fact that substitution errors are high due to the vowel shifts and consonant variation implies that current models do not expose themselves to alternative phonetic forms. Training corpora with systematically diversified speech data would enable models to acquire more general acoustic-phonetic distributions, and less depend on MAE-specific patterns.

Second, the close relation between the phonetic features and the mistake types indicates the necessity of phonetic conscious modeling strategies. Instead of considering speech input as pure statistical cues, ASR systems may take advantage of incorporating phonological information, including feature-based representations or variability modeling of pronunciation. This would help systems to be more adaptive to variation in vowel quality, consonant articulation, and syllable structure.

Third, the commonness of mispronunciations associated with cluster simplification, epenthesis, and the occurrence of prosodic variation highlights the importance of adaptive pronunciation modeling. Deletion and insertion errors can be reduced by allowing flexible lexicons or more flexible pronunciation models, which predict variation. For example, there may be more

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

than one way to pronounce a word (this is called multiple phonetic realizations of a word), and this would improve the match between the input speech and the model. Finally, our results suggest that ASR evaluation needs to change to not only look at measures of accuracy, but also to look at the nature of the error. Understanding of the extent of errors should not be in terms of how many, but rather, how and why the errors occur, and this knowledge can be used to design better systems.

6.2 Implications of Linguistic Equity

The broader application of this work is linguistic justice in speech technology rather than in technical applications. The systematic nature of the ASR errors in AAE speech turned out to be systematic and, therefore, consistent with data and model biases. The dialect bias as a structural element makes the language speakers less significant and shifts the centre of the conversation to technological systems design. The speakers of AAE are not pronouncing something wrong; they just are not taught in the model the pattern of pronunciation. This reframing plays a vital role in correcting the injustice in access to voice technologies. The findings also show that more accessibility to different speakers should be in place. As the number of ASR systems used in education, health care, and electronic communication is increasing, the discrepancy in recognition performance may also be practically relevant, including the loss of access, miscommunication, and inability to receive the services offered by the technology. Thus, a technical issue, but also a social one, is to ensure that the dialects are done in a fair way.

Finally, the research helps to come up with inclusive language technologies that take into account language diversity into account. These are system developments that can adapt to variation, models that are found on the full spectrum of speech data, and the engineering use of sociolinguistics and phonetics expertise. Lack of such initiatives can result in the fact that only a certain number of speakers will benefit from the rise in accuracy of ASR recognition, and inequality will be preserved.

7. Limitations and Future Research

Even though the research presents a good point in favor of the phonetic mismatch hypothesis, the interpretation is framed by a number of considerations of scope. The dataset size (N = 40 speakers) was first strategically selected to control the phonetic analysis to enable the systematic patterns of errors to be determined, but future projects may expand the sample size to larger and more varied samples to further enhance the generalisability. Second, the specialization of the South Asian American English allows a highly focused and theoretically informed study of phonetic variation, but future research may extend the framework to other dialects to test the generalizability of the framework. Third, the consistency and reproducibility of methodology are ensured by the fact that only one state-of-the-art ASR system (Whisper) is used, yet a comparative analysis of systems would bring further insight into how phonetic mismatch is implemented in systems.

8. Conclusion

This research showed that mistakes in the Automatic Speech Recognition (ASR) of South Asian American English (SAAE) are not accidental failures of the system, but rather a systematic effect of the phonetic discrepancy between the production of the speaker and the model. Based on a controlled corpus of 40 speakers and 170 annotated error tokens, an analysis was made that found consistent mappings between phonetic variation and error typology. Segmental variation, such as vowel changes and consonant replacements, was revealed to be the most frequent cause of substitution errors, resembling the propensity of the model to map novel acoustic realizations to

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

the nearest categories that are represented in the distributions of Mainstream American English (MAE). Phonotactic variation, especially reduction of consonant clusters, was closely linked with deletion errors, and suprasegmental characteristics of syllable-timed rhythm and prosodic transfer helped to promote the occurrence of insertion errors by ambiguity of segmentation. Such regularities in relations point to the conclusion that ASR systems run on restricted phonetic predictions based on training data, instead of scalably responding to the variation in speech input.

The findings thereby substantiate the redefinition of ASR bias as a linguistically organised phenomenon in terms of limitations of representational capability. The paper relates the creation of errors to specific phonetic processes in such a way that it is able to bring together the work that has been done on the performance of ASR and relate it to the computational analysis of phonetics and sociolinguistics theory. This consideration shifts the argument to the aggregate rates of error rather than the internal make-up of the errors, which demonstrates that the failures in the recognition are conditioned systematically by the variation in the speech. In terms of research design, the results show the necessity to integrate dialect variety, phonetic variability, and adaptable pronunciation models in ASR systems. Minor increases in model size or data quantity will hardly resolve performance disparities without some change in modeling linguistic variation.

On a bigger picture, the research paper belongs to the existing discussions of fairness and inclusiveness in speech technology in the sense that the difference in performance can be a reflection of any disparities between modeled and real speech distributions. These mismatches must be dealt with by the technical and linguistically informed approaches. By contextualising the creation of ASR errors within the SAAE, the research provides a way to perceive and avoid bias in dialect, and contribute to more accurate, generalisable, and fairer ASR systems.

References

Afkir, M., & Zellou, G. (2026). Phonological complexity, speech style, and individual differences influence ASR performance for Tarifit. Scientific Reports.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., et al. (2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. *Proceedings of the 33rd International Conference on Machine Learning*, 173–182.

Chan, M. P. Y., Choe, J., Li, A., Chen, Y., Gao, X., & Holliday, N. R. (2022, September). Training and typological bias in ASR performance for world Englishes. In Interspeech (pp. 1273-1277).

Errattahi, R., El Hannani, A., & Ouahmane, H. (2018). Automatic speech recognition errors detection and correction: A review. Procedia Computer Science, 128, 32-37.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–277). York Press.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., & Ng, A. Y. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Hassan, M. A., Rehmat, A., Khan, M. U. G., & Yousaf, M. H. (2022). Improvement in automatic speech recognition of South Asian accent using transfer learning of DeepSpeech2. *Mathematical Problems in Engineering*, 2022, Article 6825555. https://doi.org/10.1155/2022/6825555

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Jahan, M., Mazumdar, P., Thebaud, T., Hasegawa-Johnson, M., Villalba, J., Dehak, N., & Moro-Velazquez, L. (2025, April). Unveiling performance bias in ASR systems: A study on gender, age, accent, and more. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

Just, S. A., Elvevåg, B., Pandey, S., Nenchev, I., Bröcker, A. L., Montag, C., & Morgan, S. E. (2025). Moving beyond word error rate to evaluate automatic speech recognition in clinical samples: Lessons from research into schizophrenia-spectrum disorders. Psychiatry Research, 116690.

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

Labov, W. (1994). *Principles of linguistic change: Internal factors*. Blackwell.

Lai, L. F., & Holliday, N. R. (2024). Voice quality variation in AAE: An additional challenge for addressing bias in ASR models?. In Interspeech.

Li, C., Cohen, T., & Pakhomov, S. (2024). Reexamining racial disparities in automatic speech recognition performance: The role of confounding by provenance. arXiv preprint arXiv:2407.13982.

Lippi-Green, R. (2012). *English with an accent: Language, ideology, and discrimination in the United States* (2nd ed.). Routledge.

Martin, J. L., & Tang, K. (2020, October). Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual" be". In Interspeech (pp. 626-630).

McKenzie, R. M. (2010). *The social psychology of English as a global language*. Springer.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Mulholland, M., Lopez, M., Evanini, K., Loukina, A., & Qian, Y. (2016, March). A comparison of ASR and human errors for transcription of non-native spontaneous speech. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5855-5859). IEEE.

Ngueajio, M. K., & Washington, G. (2022, June). Hey ASR system! Why aren't you more inclusive? Automatic speech recognition systems' bias and proposed bias mitigation techniques. A literature review. In International Conference on Human-Computer Interaction (pp. 421-440). Cham: Springer Nature Switzerland.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). LibriSpeech: An ASR corpus based on public domain audiobooks. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

Pasandi, H. B., & Pasandi, H. B. (2022, November). Evaluation of asr systems for conversational speech: A linguistic perspective. In Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems (pp. 962-965).

Ravuri, S., & Stolcke, A. (2015). Recurrent neural network and LSTM models for lexical utterance classification. *Proceedings of Interspeech*, 135–139.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Russell, S. O. C., Gessinger, I., Krason, A., Vigliocco, G., & Harte, N. (2024). What automatic speech recognition can and cannot do for conversational speech transcription. Research Methods in Applied Linguistics, 3(3), 100163.

Sumner, M., Kim, S. K., King, E., & McGowan, K. B. (2014). The socially weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in Psychology*, 4, Article 1015. https://doi.org/10.3389/fpsyg.2013.01015

Tobin, J., Nelson, P., MacDonald, B., Heywood, R., Cave, R., Seaver, K., ... & Green, J. R. (2024). Automatic speech recognition of conversational speech in individuals with disordered speech. Journal of Speech, Language, and Hearing Research, 67(11), 4176-4185.

Trudgill, P. (2000). *Sociolinguistics: An introduction to language and society* (4th ed.). Penguin.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, 3(1), 1–40. https://doi.org/10.1186/s40537-016-0043-6

Wassink, A. B., Gansen, C., & Bartholomew, I. (2022). Uneven success: automatic speech recognition and ethnicity-related dialects. Speech Communication, 140, 50-70.

Zhang, Y., Park, D. S., Han, W., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2022). BigSSL: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 14(4), 732–745. https://arxiv.org/pdf/2109.13226