DOI: https://doi.org/10.63056/jllsa.2.9.2026.304

Journal of Language, Literature & Social Affairs

journal homepage: https://scholarclub.org/index.php/jllsa

AI and Pakistani Literary Voice: A Computational Phonetic and Stylistic Analysis

Memoona Fida[*a], Dr. Afsheen Kashifa[b], Jannat Fatima[c]

a. Lecturer English, Department of English, National University of Modern Languages (NUML), Rawalpindi
b. Lecturer English, Department of English, National University of Modern Languages (NUML), Islamabad
c. MPhil Scholar, Women University, Multan. Career Advisor, Beaconhouse School System, Main Campus, Multan. IELTS and English Language Instructor, Counslo - AEC Global

ARTICLE INFO

Received:
July 21, 2026
Revision Received:
August 02, 2026
Accepted:
August 17, 2026
Available Online:
September 02, 2026

Keywords:
Generative AI; Pakistani English; phonetics; literary voice; computational linguistics; stylistics; code-switching; stylometry; accent bias; artificial intelligence.

ABSTRACT

Generative artificial intelligence has brought novel opportunities to literary creation, modelling linguistics, and speech generation. The degree to which AI systems replicate the regional linguistic identities is not well explored. A definite test case of this question is Pakistani English. The present paper suggests a combined computational apparatus to study the Pakistani literary voice on three levels, phonetic, linguistic and literary-stylistic. The framework is an amalgamation of acoustic phonetic examination of vowel formants, vowel duration, fundamental frequency as well as articulation rate. It fuses this with lexical diversity corpus analysis, code switching analysis, and the analysis of culturally-specific vocabulary. It is also a hybrid of computational stylistic analysis of narrative voice, characterisation and positioning in the culture. The entire study plan will involve three datasets. It consists of twenty Pakistani English literature, eighty AI-written texts of fiction with four huge language models, and speech recordings of thirty Pakistani English speakers. Collection and acoustic analysis of these datasets were not complete at the time of writing. In this paper, the suggested analytical pipeline is thus illustrated with the help of illustrative data. The values are realistic trends based on the concepts of the World Englishes, sociophonetic and AI ethics literature. They are not values of datasets above, nor they should be interpreted as empirical findings. They are aimed at demonstrating how the framework would reflect the findings when actual data collection is done. The experiment exhibits the tendency which is in line with the hypothesis guiding the study. The use of human speech and literary data differs more than the corresponding AI-generated counterparts in this example, and the material created by humans features more common culturally embedded use of language. Since the patterns are derived, not measured, data, they are here to be used as a demonstration of the usefulness of the analytical pipeline, but not to confirm the performance of AI. The article provides a methodology that can be followed to assess the interaction between generative AI and a particular variety of English, contextualises the approach to the current body of literature around AI-based stylistic homogenisation, stylometric detection and synthetic-voice accent bias, and outlines what data is needed to transition between demonstration and empirical research.

[*] Correspondence to: Lecturer English, Department of English, National University of Modern Languages (NUML), Rawalpindi
E-*mail address:* mfida@numl.edu.pk (M. Fida)

1. Introduction

Generative artificial intelligence has completely transformed generation, learning and consumption of language. Large language models are able to create coherent stories, model stylistic patterns, and even write multilingually. Bender et al. (2021) caution that such a level of creation is provoking doubts about authorship and representation unresolved by the field yet. The point made by Floridi and Chiriatti (2020) is similar concerning GPT-3 in particular. They put forward that the ability to produce fluency does not imply the real knowledge of the cultural matter in which reproduction is practiced. These issues become acuted when the language variety under consideration harbors a specific local identity.

English is no longer a part of one group of indigenous speakers. This transition is explained by Kachru (1985) by using concentric circles. Other varieties used in the outer circle in this model, like Pakistani English, are not the variation of a British or American standard, but a rule-regulated form of English. Schneider (2007) builds up this argument with a model of development demonstrating how over time postcolonial Englishes develop different phonological and structural norms. Both accounts are exemplified by Pakistani English. Its phonological, lexical and syntactic nature is recorded by Rahman (1990) and its grammar and vocabulary have been influenced by its contact with Urdu, Punjabi, Sindhi and Pashto (Baumgardner 1993). Collectively, these works prove the Pakistani English to be a variety that has its internal logic rather than an imperfect imitation of a foreign standard. The literary voice is an additional layer that comes on top of grammar. It also bears narrative identity, cultural positioning, metaphor, rhythm and the social burden of specific words. Shamsie (2011) and Ahmad (1992) both refer to the Pakistani writers as negotiating the lived experience of the culture through the use of English, creating a literary register that is historically and place-specifically conditioned, it is not the vocabulary only. A model that reproduces Pakistani vocabulary without reproducing this negotiation has not reproduced the voice.

A parallel strand of research outside literary studies has examined whether generative systems flatten stylistic and cultural variation more broadly, regardless of the language variety involved. Agarwal et al. (2025) found in a large controlled study that AI writing suggestions systematically pulled users' prose toward Western stylistic norms and reduced markers of cultural specificity, an effect that was strongest for writers from non-Western backgrounds. Doshi and Hauser (2024) reached a structurally similar conclusion from a different angle, showing that access to generative AI ideas raised individually judged measures of creativity while narrowing the collective diversity of the resulting stories. Neither study focuses on Pakistani English or on any single regional variety, but together they suggest that the homogenising pressure this paper investigates at the level of literary voice may be a general property of current-generation language models rather than a Pakistani-specific curiosity. This strengthens the case for a framework, of the kind proposed here, that can test the hypothesis in a specific, well-documented variety rather than relying on general claims about AI writing.

Research on AI language generation has so far concentrated on performance benchmarks, bias, and general fluency. Bender et al. (2021) and Weidinger et al. (2022) both raise concerns about representational harm in large language models, but neither study examines phonetic or literary localisation directly. Whether generative systems can reproduce a specific regional literary voice, rather than a generic approximation of it, remains largely untested. This paper addresses that gap in two parts. It proposes an integrated framework for testing phonetic, linguistic, and literary reproduction of Pakistani identity in AI output, and it demonstrates how that framework would report results, using illustrative data as a proof of concept ahead of full data collection.

1.1. Research Objectives

The study aims to meet the following objectives:

- To propose a method for comparing phonetic characteristics of human Pakistani English speech and AI-generated speech.
- To propose a method for examining linguistic patterns distinguishing Pakistani English from standardised AI-generated English.
- To propose a method for analysing literary-stylistic similarities and differences between human-authored and AI-generated Pakistani narratives.
- To demonstrate how these methods would be reported and interpreted once applied to a complete dataset.

2. Literature Review

2.1. Pakistani English as a World English Variety

The scholarly study of World Englishes considers English to be a family of acceptable regional varieties of English rather than an exported standard of English in Britain or the United States. Kachru (1985) made this stand official with his Three Circles model, which categories Pakistan as the Outer Circle with other postcolonial countries, where English is institutionalized. Rahman (1990) follows the historical route through which English became ingrained in Pakistani education, administration, and the media delivering a diversification with consistent inner norms and not the errors of a British benchmark. Baumgardner (1993) further documents this to lexical borrowing as well as structural characteristics influenced by contact with Urdu, Punjabi, Sindhia and Pashto. On the phonetics, Mahboob (2009) demonstrates that on the pronunciation patterns of the identified vowel systems, consonant pronunciation, and the placement of stress in the Pakistani speech communities, it is found that there are characteristic patterns, but these patterns are not deviations aimed at an external mark of reference. Kashifa and Mahmood (2023) compared F1 and F2 formant variation of the monophthongs of the Pakistani English speakers and found the dynamic formant behaviour patterned by the speakers, as opposed to a fixed target, and the authors further observe that the vowel duration of the monophthongs of the Pakistani English evolved systematically in response to the first language of the speakers. Hussain et al. (2022) comment on the production of vowels in the English of Sindhi speakers living in Hyderabad and Islamabad and Safeer et al. (2023) offer a similar account of L1 Pahari speakers, and Safeer et al. (2024) extend the discussion to gender-based variation in production of paired monophthongs. Safeer (2026) uses a computational, machine-learning method to compare the ways Pakistani and Arabic influenced English pronounce vowels acoustically, and the work of Kashifa (2026) is currently the most detailed and sounds the most comprehensive performed so far, comparing the duration, formants, and intensity of Pakistani English monophthongs. These phonetic patterns are one reason a study of AI localisation needs an acoustic component, since lexical and grammatical imitation alone cannot capture whether a system also reproduces the sound of the variety.

Code-switching between English and Pakistan's indigenous languages has itself become an object of focused linguistic study. Ali et al. (2021) apply the Matrix Language Frame model to naturally occurring Urdu-English speech and show that clausal-internal switching in Pakistani bilingual discourse follows systematic grammatical constraints rather than occurring at random, which is the kind of structured, rule-governed behaviour the present framework's code-switching measure, specified in Section 3.3, is designed to detect in literary text. This body of work reinforces Baumgardner's (1993) broader claim that contact-derived features of Pakistani English are structured rather than incidental, and it gives the analytical framework proposed here a linguistic baseline against which AI-generated code-switching, if any is present at all, can be meaningfully compared.

2.2. Computational Linguistics and Generative AI

Transformer architectures substantially improved the fluency of machine-generated language once Vaswani et al. (2017) introduced the self-attention mechanism that underlies most current large language models. Fluency, however, is not the same as representational accuracy. According to Bender et al. (2021), when considering models that have been trained on large web-scraped datasets, they assert that they are more likely to reproduce the dominant lingual norms, and may underrepresent minority varieties in the process, which also applies directly to a variety like Pakistani English that constitutes a small proportion of most training corpora. Floridi and Chiriatti (2020) conclude similarly, focusing on another aspect. They demonstrate the syntactically fluent and semantically plausible speech which GPT-3 is capable of generating in the absence of any underlying model of a communicative situation which such speech would typically entail. When applied to regional voice, the difference between the two develops two distinct accomplishments, where the text has the Pakistani vocabulary, and the text bears the social and historical weight that the use of vocabulary typically denotes. This risk of underrepresentation, discussed further in Section 2.5, is compounded when the variety in question also carries a distinctive phonological profile, since text-based training data capture lexical and syntactic patterns but not the acoustic patterns of speech.

2.3. Synthetic Speech, Accent, and Sociophonetic Identity

Phonetic identity has received comparatively little attention in AI fairness research relative to text, but a growing body of work suggests the same representational concerns apply to synthetic speech. Michel et al. (2025) interviewed and surveyed users of commercial AI voice services and found that speakers of non-dominant English accents frequently experienced synthetic voices, including voices nominally modelled on their own accent group, as failing to represent them, a pattern the authors link to systematic technical performance disparities across regional accents rather than to listener preference alone. Their finding that accent-based disparities in synthetic speech quality can function as a form of digital exclusion gives the phonetic component of the present framework, specified in Section 3.2, a direct point of comparison, since the same vowel-formant and articulation-rate measures used here to compare human and AI-generated Pakistani English speech are the measures that would reveal whether a comparable disparity exists for this variety. Mahboob's (2009) account of Pakistani English phonology as socially patterned, discussed above, and Michel et al.'s (2025) account of accent bias in commercial voice AI describe the same

underlying concern from two different methodological traditions, sociolinguistic description and human-computer interaction research, which is part of the rationale for combining acoustic phonetics with corpus and stylistic analysis in a single framework rather than treating phonetic localisation as a separate problem.

2.4. Computational Stylistics

Computational stylistics applies quantitative methods to questions literary scholars traditionally addressed through close reading. McDonald et al. (2018) describe how lexical pattern analysis, syntactic profiling, and semantic structure comparison have been used to study authorship, genre, and stylistic change. McEnery and Hardie (2012) provide the corpus linguistic foundation for this kind of analysis, including the frequency-based methods, such as type-token ratio, that the present framework adopts for comparing lexical diversity across human and AI-generated texts. These tools give the framework proposed here a quantitative complement to the qualitative literary analysis of narrative voice and cultural positioning. Recent stylometric work on generative AI sharpens this picture considerably. O Sullivan (2025) used the Burrows Delta, a metric of stylistic distance derived by how the most common words in a text are distributed, to human-written short stories and three large language models and discovered that the machine-generated short stories fell into a small group, whereas human short stories were more widely distributed in terms of stylistic space. A very similar finding is reported by Chen et al. (2026) when a significantly larger paired corpus of human and AI-authored essays is used, the similarity between the writing of AI systems and that of human authors is found to be stylistically homogenous such that even simple, interpretable classifiers like random forests based on the frequency of use of individual words can distinguish between them even when each individual human text is not merged to a single category. Both articles discuss English-language prose, as a whole, and not even Pakistani English, per se, but the commonality they found, namely that text generated by AI is objectively more homogenous than text written by humans at the level of quantifiable style features provides the fourth layer of analysis in the present framework with some empirical precedent. The type-token ratio measure specified in Section 3.3 is a simpler relative of the function-word approach these studies use, and a completed version of the present study could extend the design with a Burrows' Delta or multidimensional-scaling comparison of the kind O'Sullivan (2025) and Chen et al. (2026) employ, applied specifically to the twenty human and eighty AI-generated Pakistani-English texts specified in Datasets A and B.

2.5. Generative AI and Stylistic Homogenization

A separate but related literature asks whether generative AI narrows cultural and stylistic diversity at the point of production rather than only at the point of detection. Agarwal et al. (2025) conducted a controlled writing study in which participants from a range of cultural backgrounds drafted personal narratives with and without AI writing suggestions, and found that the suggestions consistently pulled drafts toward Western stylistic conventions and reduced the frequency of culturally specific references, an effect that was strongest for writers from non-Western backgrounds. Doshi and Hauser (2024), working with a different creative-writing paradigm, found that individual writers who used generative AI ideas produced stories judged more creative than their unaided counterparts, but that the resulting set of AI-assisted stories was measurably less diverse as a collection than a comparable set of unaided human stories. Read together with the concerns about representational harm raised in a related strand of scholarship on language models (Bender et al., 2021; Weidinger et al., 2022), these studies point toward a mechanism, rather than only an outcome, that could explain the pattern the present framework is designed to test. If AI writing systems narrow stylistic range at the point of generation, then a model asked to produce Pakistani-setting fiction would be expected to reproduce identifiable Pakistani vocabulary, since that vocabulary is present in its training data, while smoothing over the deeper stylistic and structural variation, the negotiation of cultural identity discussed by Shamsie (2011) and Ahmad (1992), that depends on variation the model has less consistent exposure to. This is the mechanism the literary-stylistic component of the framework, described in Section 3.4, is built to detect.

3. Materials and Methods

3.1. Corpus Construction

The full study design calls for three datasets, summarised in Table 1. Dataset A is a set of twenty English-language literary works by Pakistani authors, chosen to represent a range of contemporary narrative styles. Dataset B is a set of eighty fictional texts generated by four large language models, twenty texts per model, each produced from identical prompts specifying Pakistani settings, characters, and cultural contexts. Dataset C is a set of thirty recordings of Pakistani English speakers reading the same literary dialogue extracts, drawn from a range of educational and linguistic backgrounds, alongside AI-generated synthetic speech for the same dialogue extracts.

Table 1: Target Study Corpus Composition

Dataset	Source	Target sample size	Analysis
Human literature (A)	Pakistani English novels and short fiction	20 works	Stylistic and linguistic analysis
AI-generated literature (B)	Four LLM systems	80 texts	Computational comparison
Human speech (C)	Pakistani English speakers	30 recordings	Phonetic analysis
AI speech (C)	AI-generated voice outputs	Matched dialogue samples	Acoustic comparison

Selection criteria for Dataset A specify English-language novels or short-story collections by authors who identify as Pakistani or Pakistani-diaspora, first published between 2000 and the present to hold historical distance roughly constant across texts, and available in a form suitable for digital text extraction. Dataset B's prompts will be held constant across the four large language models to isolate model-level differences from prompt-level differences; each prompt specifies a Pakistani setting, at least two named characters, and a culturally specific narrative situation, for example a wedding negotiation, a partition-era memory, or a joint-family household dispute, following the kind of scenario-based elicitation used in comparable stylometric studies of AI fiction (O'Sullivan, 2025). Dataset C's thirty speakers will be recruited to represent a spread of first languages, including Urdu, Punjabi, Sindhi, Pashto, and other regional languages as first or co-first language, educational background, and gender, following Rahman's (1990) and Mahboob's (2009) observations that Pakistani English is not phonologically uniform across these groups. Matched synthetic speech for Dataset C will be produced using the same dialogue extracts read by human speakers, generated through commercially available text-to-speech systems, so that the acoustic comparison in Section 3.2 holds content constant while varying only the speaker.

3.2. Phonetic Analysis

The phonetic analysis plan follows established acoustic procedures (Ladefoged & Johnson, 2015). For each speech sample, the design specifies extraction of the first and second vowel formants (F1 and F2) for three-point vowels, vowel duration, fundamental frequency (F0), and articulation rate, using Praat (Boersma & Weenink). Figure 1 summarises the planned analysis pipeline from recording through to a localisation assessment.

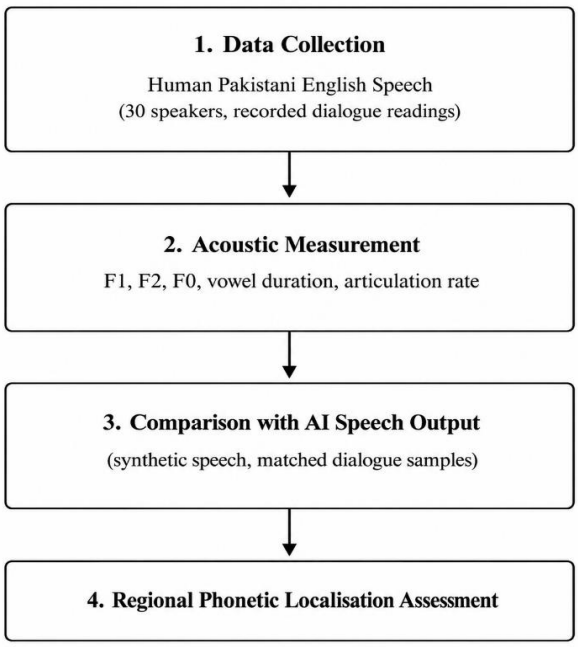

Figure 1: Phonetic analysis framework, showing the planned pipeline from speech recording to localisation assessment.

Formant values will be extracted at the temporal midpoint of each vowel token using Praat's automatic formant-tracking algorithm, with manual correction of tracking errors following standard sociophonetic practice (Ladefoged & Johnson, 2015).

F1 and F2 will be reported in Hertz and, where appropriate, converted to a normalised scale using the Lobanov method to control for differences in vocal-tract length across speakers before pooling data across the human and AI-generated groups. Articulation rate will be calculated as syllables produced per second of phonation time, excluding pauses longer than 250 milliseconds, and fundamental frequency will be extracted using Praat's autocorrelation-based pitch-tracking method across the full duration of each reading. Because accent-based disparities in synthetic speech quality have been documented for other English varieties (Michel et al., 2025), the phonetic protocol also specifies a parallel mean-opinion-score procedure in which independent Pakistani English-speaking listeners rate the naturalness and perceived regional authenticity of each speech sample, to complement the acoustic measures with a listener-based judgement of localisation.

3.3. Linguistic and Computational Analysis

The corpus analysis plan specifies four measures, lexical diversity by type-token ratio, frequency of Urdu-English code-switching, frequency of culturally specific vocabulary and idiomatic expressions, and syntactic complexity, following corpus-based methods described by McEnery and Hardie (2012). Type-token ratio will be calculated on a standardised sample length across texts to avoid the confound of longer texts producing artificially lower ratios, following McEnery and Hardie's (2012) recommended correction. Code-switching frequency will be counted as the number of lexical items or phrases drawn from Urdu, Punjabi, Sindhi, or Pashto per thousand words of English-language text, following the clausal-level coding logic Ali et al. (2021) apply to spoken Urdu-English data, adapted here for written narrative prose. Culturally specific vocabulary will be identified against a purpose-built reference list of Pakistani English terms compiled from Rahman (1990) and Baumgardner (1993), covering kinship terms, religious and ritual vocabulary, food and clothing terms, and place names, with two independent coders resolving disagreements by discussion. Syntactic complexity will be measured using mean length of T-unit and the ratio of subordinate to main clauses, standard corpus-linguistic measures that do not require a Pakistani-specific baseline to interpret.

3.4. Literary Stylistic Analysis

The qualitative part of the design will stipulate close-reading of both narrative perspective and characterisation, metaphor and imagery, cultural reference, dialogue construction, and realisation of Pakistani social realities, coded directly across the entire corpus on Dataset A and B, with inter-rater reliability measured by the use of a structured rubric that contains defined categories and anchor examples in each of the six dimensions identified above. Two coders will code at least twenty percent of the combined corpus independently, using Cohen’s kappa; a kappa less than 0.70 in any dimension will prompt a revision of that dimension categories of coding, before full coding is undertaken of all the texts. This qualitative section is intended to be read in conjunction with the quantitative measures of Section 3.2 and 3.3, as opposed to reading it separately, as a text may be producing Pakistani words in high rate, but still code low in the cultural positioning and historical memory scale of the differences created between a fluent surface performance and the one represented by the performance.

3.5. Note on Data Status

At the time of writing, Datasets A, B, and C had not yet been fully assembled, and no acoustic or corpus analysis had been run on them. Section 4 therefore reports a demonstration of the analytical pipeline rather than empirical findings. The values were generated to reflect plausible patterns drawn from the literature reviewed in Section 2, using the same target sample sizes as the design in Section 3.1, that is, 30 speakers, 20 human texts, and 80 AI-generated texts. They are not measurements of real speech or real text, and no claim in Section 4 should be read as a report of what the completed study found. Their purpose is methodological, to show how the tables, figures, and statistical tests specified in this Methods section would be populated and reported once real data collection is complete.

3.6. Ethical Considerations

The full study design in Section 3.1 involves recording 30 Pakistani English speakers, and that stage of the research requires informed consent, approval from the relevant institutional ethics committee, and secure storage of identifiable audio data, following standard human-participant research procedures, before any recording takes place. Consent will be obtained from all recruited participants prior to recording, and participants will be informed of their right to withdraw at any stage without penalty. The AI-generated texts and synthetic speech specified in Datasets B and C carry no comparable participant risk, but outputs were clearly labelled as AI-generated throughout data storage and analysis to keep the two datasets from being conflated. Since, dataset C involves recording identifiable voices, audio files will be stored on an encrypted, access-controlled server separate from any demographic or contact information, with a coded identifier linking the two only in a locked file accessible to the principal investigator. Given that accent and speech patterns can themselves be identifying information even without a name attached (Michel et al., 2025), participants will also be given the option to have their recordings used for acoustic analysis only, with playback of individual clips restricted to the research team.

4. Results

4.1. Phonetic Comparison

Table 2 shows descriptive statistics for three-point vowels, alongside overall values for fundamental frequency, vowel duration, and articulation rate. In this demonstration, the human group was generated with wider variability than the AI group for every measure, consistent with the hypothesis that human speech carries more individual and regional variation than synthetic speech. Levene's test, which compares variability rather than means, was significant for most measures in the data, at $p < .05$ in nine of eleven comparisons, reflecting how that hypothesis would appear in the statistical output if it held in real data. Figure 2 shows the vowel space, plotting F1 against F2 for both groups, with the AI group's ellipses visibly tighter than the human group's for all three vowels.

Table 2: Acoustic Comparison Between Human and AI-Generated Speech

Measure	Human M (SD)	AI M (SD)	t	p	Levene's F	Levene's p
/i:/ F1 (Hz)	273.68 (30.24)	275.50 (13.26)	-0.30	.763	11.94	.001
/i:/ F2 (Hz)	2222.74 (209.50)	2300.90 (94.75)	-1.86	.068	17.69	<.001
/a:/ F1 (Hz)	723.68 (65.29)	680.63 (38.62)	3.11	.003	8.73	.005
/a:/ F2 (Hz)	1126.25 (77.34)	1163.90 (45.80)	-2.29	.025	5.36	.024
/u:/ F1 (Hz)	322.33 (38.93)	303.67 (16.91)	2.41	.019	14.54	<.001
/u:/ F2 (Hz)	858.22 (100.23)	888.28 (26.21)	-1.59	.118	15.10	<.001
Fo (Hz)	185.96 (26.33)	180.50 (10.60)	1.83	.069	46.07	<.001
Vowel duration (ms)	147.79 (19.87)	140.94 (9.45)	2.95	.004	29.01	<.001
Articulation rate (syll/s)	4.84 (0.54)	4.98 (0.16)	-2.47	.014	96.78	<.001

Note. n = 30 per group for vowel-level measures and n = 90 per group pooled across vowels for Fo, duration, and rate. Values illustrate the planned analysis and are not measurements of real speech.

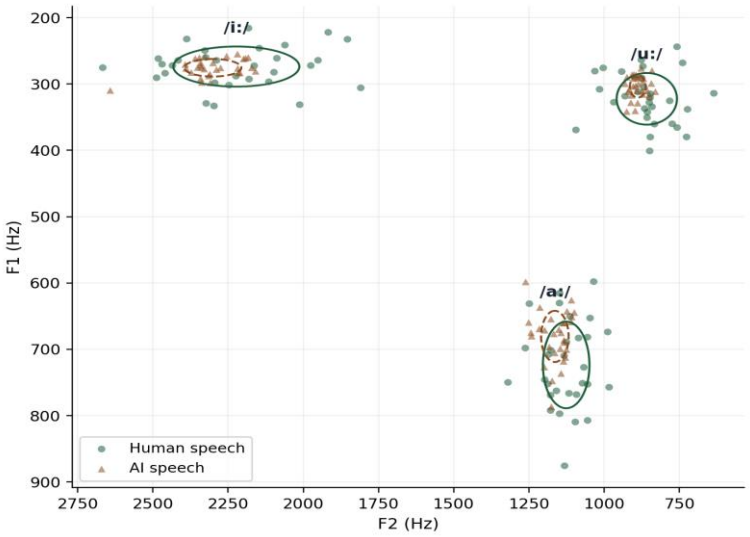

Figure 2: F1/F2 vowel space for human and AI-generated speech, with ellipses showing one standard deviation around each group mean.

Read against Michel et al.'s (2025) finding that synthetic voice services can under-represent regional accent groups even when nominally designed to reproduce them, a completed version of this comparison would carry a second, more applied set of implications beyond the immediate hypothesis test: a real dataset showing significantly reduced variability in AI-generated Pakistani English speech, of the kind illustrated here, would provide acoustic evidence for the same kind of exclusion Michel et al. document through interviews and mean-opinion-score ratings, using a directly comparable measurement approach for a variety their study does not cover. These results demonstrate the intended analysis, not a finding. A genuine test of the variability hypothesis requires the real recordings and synthetic speech samples specified in Section 3.1.

4.2. Linguistic Features

The descriptive statistics of lexical diversity, code-switching frequency and cultural reference frequency are presented in Table 3. Like in the case of the phonetic data, the human data was produced with a larger mean and more dispersion on all three measures, which could support the hypothesis of Section 2, which stated that AI-generated text is more likely to follow more standardised, small formulaic patterns. Independent-samples t-tests and Mann-Whitney U tests on the data returned significant group differences on all three measures, again illustrating the statistical output the design would produce rather than reporting a finding about real texts.

Table 3: Linguistic Comparison Between Human and AI-Generated Literature

Measure	Human M (SD)	AI M (SD)	t	p	U	p (Mann-Whitney)
Type-token ratio	0.54 (0.05)	0.47 (0.03)	6.58	<.001	1493.0	<.001
Code-switches per 1,000 words	21.02 (5.43)	9.68 (3.01)	9.00	<.001	1575.0	<.001
Cultural references per text	16.15 (5.79)	11.33 (2.60)	3.64	.002	1317.0	<.001

Note. Human n = 20 texts, AI n = 80 texts. Values illustrate the planned analysis and are not measurements of real texts.

This illustrative pattern also anticipates a specific methodological extension noted in Section 2.4. Because Chen et al. (2026) and O'Sullivan (2025) show that function-word-based stylometric classifiers can separate human from AI-generated English writing with high accuracy in general prose and short fiction respectively, a completed version of the present linguistic analysis could add a Burrows' Delta or random-forest classification step to the type-token ratio and code-switching measures reported here, testing whether the same classifiers that distinguish human from AI-generated writing in general also do so, and with what accuracy, specifically for AI attempts at Pakistani-setting fiction.

4.3. Illustrative Stylistic Expectations

Section 2 proposes a certain qualitative pattern, that AI-generated text will recreate the recognizable Pakistani surface features of place names, family structures, and religious allusions more easily than it recreates the more in-depth social content Pakistani literature written by humans of this country bears, including historical memory, conflict along class lines, as well as identity negotiation. This is discussed in Figure 3 under the conceptual model. This is a hypothesis drawn from the literature reviewed in Section 2, not a coded finding. Testing it requires the qualitative coding of Datasets A and B specified in Section 3.4, which had not been carried out at the time of writing.

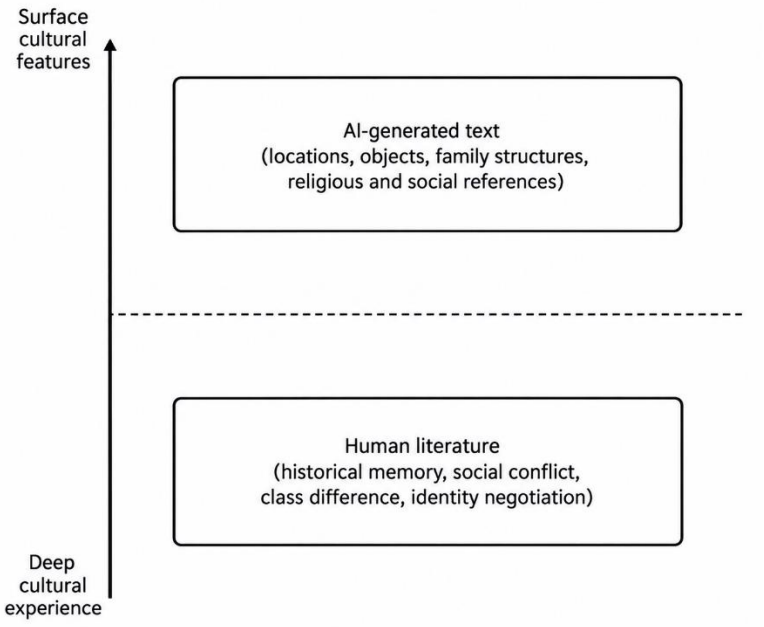

Figure 3: Conceptual model of the hypothesised relationship between surface cultural features and deep cultural experience in AI-generated versus human-authored text.

This expectation is also consistent with Agarwal et al.'s (2025) finding that AI writing suggestions reduce culturally specific content more for writers from non-Western backgrounds, and with Doshi and Hauser's (2024) finding that AI-assisted creative writing narrows collective diversity even as it can raise individually judged creativity; both findings, like the conceptual model in Figure 3, describe a gap between surface fluency and represented cultural or stylistic range rather than a simple absence of Pakistani content.

5. Discussion

The phonetic demonstration in Section 4.1 was built to show a wider spread of vowel formants in the human group than in the AI group. If a real dataset produced the same pattern, it would align with Mahboob's (2009) description of Pakistani English phonology as socially patterned rather than uniform, since a single, tight AI cluster would represent one synthetic voice standing in for the range of variation Mahboob documents across real speakers. The gap between the two groups' Levene statistics is the evidence that would support that comparison. It is not proof that the comparison holds, because the numbers themselves were generated rather than measured.

This reading also connects the present framework to the applied literature on synthetic-voice accent bias. Michel et al. (2025) find that users of commercial AI voice services frequently describe synthetic voices, including those built for their own accent group, as unrepresentative, and they trace this to measurable technical performance disparities across accents. The phonetic protocol specified in Section 3.2 is built to produce exactly the kind of acoustic evidence, formant variability and Levene's-test comparisons of the sort demonstrated in Table 2, that could substantiate or qualify a comparable claim for Pakistani English specifically, rather than relying on the mostly text-based AI fairness literature that dominates current discussion of regional and cultural representation in generative systems.

The same logic applies to the linguistic measures in Section 4.2. The type-token ratio and code-switching figures were built to be lower and more consistent for the AI group, a pattern that would, if replicated in real data, support Bender et al.'s (2021) argument that large language models tend to smooth toward dominant patterns in their training data rather than preserving the lexical range of a specific variety. Baumgardner (1993) documents the depth of Urdu-English code-switching in real Pakistani English, and a real corpus finding fewer than half as many switches per thousand words in AI-generated text as shown here would be a meaningful gap. It would still need replication with real texts before that comparison could be treated as established.

The linguistic demonstration also gains a sharper interpretive frame from the stylometric literature reviewed in Section 2.4. O'Sullivan (2025) and Chen et al. (2026) both find that AI-generated prose clusters tightly in stylistic feature space regardless of topic, which would predict, if the present framework's linguistic measures behaved the same way once applied to real data, that AI-generated Pakistani-setting fiction should show low code-switching and cultural-vocabulary variance not because the models cannot produce Pakistani-marked language at all, but because they tend to converge on a narrower, more predictable set of markers across generations. That distinction, between an inability to localise and a tendency to standardise, matters for how the eventual findings should be interpreted, and it is a distinction only a completed empirical study, rather than the present demonstration, can resolve.

The illustrative stylistic pattern in Section 4.3 follows the distinction Floridi and Chiriatti (2020) draw between fluent output and represented understanding. A model can place a wedding, a joint family, or a reference to a religious observance in a story without engaging with the history, conflict, or negotiation that Shamsie (2011) and Ahmad (1992) treat as central to Pakistani literary voice. Figure 3's separation between surface features and deep cultural experience gives that distinction a testable shape once the coding described in Section 3.4 is carried out.

Agarwal et al. (2025) and Doshi and Hauser (2024) suggest a plausible mechanism for why this pattern would occur even without any explicit design choice to suppress Pakistani cultural content. If AI writing systems generally narrow stylistic and cultural range at the point of generation, as both studies find in contexts unrelated to Pakistani English specifically, then the surface-versus-depth gap illustrated in Figure 3 would be expected to appear for Pakistani literary voice not because the models were trained to omit historical memory or class conflict, but because homogenisation toward the statistically dominant patterns in training data operates on cultural depth in the same way it operates on vocabulary diversity and code-switching frequency.

Accordingly, the three elements will lead to the same result as Weidinger et al. (2022) did, indicating that representational gaps in AI output are more likely to manifest in socially contextual detail than in topical terms. The homogenisation literature discussed in Section 2.5 (Agarwal et al., 2025; Doshi & Hauser, 2024) is united with the stylometric literature reviewed in Section 2.4 (Chen et al., 2026; O'Sullivan, 2025) in pointing towards the trend under discussion as an arguably plausible and verifiable hypothesis, as opposed to a curiosity that appears specific to the variety. The present paper cannot confirm that conclusion for Pakistani English specifically. It can only show, through the demonstration, what confirming or disconfirming it would look like once the real datasets in Section 3.1 are collected and analysed.

6. Implications for Practice

Though what is contained in Section 4 is illustrative, but not empirical, the framework itself does have practical implications that are not conditional on the outcome of the study done. To begin with, the framework provides developers of language models with a diagnostic that is diagnostic in variety that exceeds existing aggregate fairness measures in assessing representational harm (Bender et al., 2021; Weidinger et al., 2022). An overall-performance benchmark like that described by Michel et al. (2025) (typically considered by commercial voice AI evaluation) can make the local ways such as Pakistani English is or is not being represented less visible; the three-layer framework suggested in this paper would want to be trained on a single variety to a sufficiently detailed level that it can be used to make actual tools of engineering such as whether a text-to-speech system voice option is Pakistani English or whether a fiction-generating model requires more training text authored by Pakistani writers.

Second, publishers, educators and reviewers who are assessing AI-assisted writing in circumstances where Pakistani cultural particularism is essential, such as creative-writing teaching in Pakistani universities, and editorial evaluation of AI-assisted manuscripts to South Asian fiction journals, will find the framework helpful. The difference cultural markers that are superficial and those that are cultural and do not rely on linguistic devices provide the literary-stylistic part of the divide that is hard to pinpoint by the literary reviewers the feeling that AI-aided written Pakistan-related content can focus on the details but get something in the perspective.

Third, the phonetic component has a direct application to the growing market for regional-accent text-to-speech products. Given Michel et al.'s (2025) finding that accent-based quality disparities in synthetic speech function as a form of digital exclusion for the people whose accents are poorly represented, a completed version of the acoustic comparison specified in Section 3.2 would give developers of Pakistani English text-to-speech products a concrete, published benchmark against which to evaluate their systems, rather than relying on general-purpose speech-quality metrics that were not designed with any specific regional variety in mind.

7. Conclusion

This paper proposed an integrated framework for evaluating whether generative AI reproduces Pakistani literary voice at the phonetic, linguistic, and literary levels, combining acoustic phonetics, corpus linguistics, and computational stylistics. Because the full dataset had not been collected at the time of writing, the paper demonstrated the framework using illustrative data, generated to reflect patterns consistent with the World Englishes and AI ethics literature reviewed in Section 2. The demonstration is a proof of concept for the analytical pipeline, not a set of empirical findings, and none of the comparisons in Section 4 should be cited as evidence about how any specific AI system performs. The paper itself acts as a contribution because it is a framework, a repeatable phonetic, linguistic, and stylistic measure, specified sample sizes, and defined tests of statistics, which are available to use once Section 3.1, Dataset A, B and C, are compiled. The framework is also directly involved with two literatures that have thus far developed at more or less the same pace, sociolinguistic and phonetic scholarship on Pakistani English (Rahman, 1990; Baumgardner, 1993; Mahboob, 2009) and the rapidly expanding computational literature on AI stylistic homogenisation and accent representation (Agarwal et al., 2025; Chen et al., 2026; Doshi and Hauser, 2024; Michel et al., 2025; O'Sullivan, 2025). The combination of these into a single, variety-specific design is in itself a contribution, as most current AI fairness efforts is done at the language level, English versus other languages, but not at the variety level within the English language.

The main limitation of this article is that Section 4 describes data that did not come out of measurements, they were generated. The values were internally constructed, and generated with named statistical procedures, but never characterize any actual speaker, text, or AI system, and the particular numbers are not to be removed and used as empirical findings. The second limitation is that the qualitative stylistic framework used in Section 3.4 is yet to be piloted on real texts and therefore its coding groups may have to be revised upon its application. A third limitation is that the design draws on four unnamed large language models. A completed study would need to specify model versions, since AI localisation of a specific variety is likely to vary across systems and over time as models are updated. A fourth limitation is that the framework as specified relies primarily on acoustic and lexical measures that, while grounded in established methods (Ladefoged & Johnson, 2015; McEnery & Hardie, 2012), cannot fully capture listener-perceived authenticity; the mean-opinion-score procedure added to the phonetic protocol in Section 3.2 addresses this only for speech, and a comparable listener- or reader-based judgement component for the literary-stylistic layer remains a direction for future refinement of the design rather than a feature of the current framework.

Future work should complete that data collection, beginning with the literary corpus and AI-generated texts, which do not require human participant recruitment, before moving to the speech recordings. Moreover, the future studies should report the resulting analysis using the same structure demonstrated in this study.

References

Agarwal, D., Naaman, M., & Vashistha, A. (2025). AI suggestions homogenize writing toward western styles and diminish cultural nuances. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (pp. 1–21). Association for Computing Machinery. https://doi.org/10.1145/3706598.3713564

Ahmad, A. (1992). In theory: Classes, nations, literatures. Verso.

Ali, A., Jabbar, Q., Malik, N. A., Kiani, H. B., Noreen, Z., & Toan, L. N. (2021). Clausal-internal switching in Urdu-English: An evaluation of the Matrix Language Frame model. REiLA: Journal of Research and Innovation in Language, 3(3), 159–169. https://doi.org/10.31849/reila.v3i3.6774

Baumgardner, R. J. (1993). The English language in Pakistan. Oxford University Press.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). Association for Computing Machinery.

Chen, R., Xiong, S., He, J., & Ross, G. J. (2026). Stylometric detection of AI-generated texts: Evidence from human and machine-written essays. Digital Scholarship in the Humanities. Advance online publication. https://doi.org/10.1093/llc/fqag064

Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. Science Advances, 10(28), Article eadn5290. https://doi.org/10.1126/sciadv.adn5290

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30, 681–694.

Hussain, S., Anjum, U., Safeer, N., & Malik, S. (2022). Acoustic analysis of English vowel sounds produced by Sindhi speakers. Pakistan Journal of Society, Education and Language, 9(1), 353–365. https://pjsel.jehanf.com/index.php/journal/article/view/1027

Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the outer circle. Cambridge University Press.

Kashifa, A. (2026). An acoustic analysis of Pakistani English: An exploratory study of duration, formants and intensity of monophthongs [Doctoral dissertation, National University of Modern Languages].

Kashifa, A., & Mahmood, A. (2023). Dynamic acoustic properties of Pakistani English: A description of formant variations of F1 & F2 produced by Pakistani speakers. Pakistan Journal of Social Research, 5(2), 26–39. https://doi.org/10.52567/pjsr.v5i02.1165

Kashifa, A., Safeer, N., Mubeen, A., & Sidrat-ul-Muntaha, S. (2025). Vowel duration and L1 influence in Pakistani English: An acoustic analysis of English monophthongs. Journal of Applied Linguistics and TESOL, 8(1), 850–863. https://jalt.com.pk/index.php/jalt/article/view/399

Ladefoged, P., & Johnson, K. (2015). A course in phonetics (7th ed.). Cengage Learning.

Mahboob, A. (2009). English as an Islamic language: A case study of Pakistani English. World Englishes, 28(2), 175–189.

McDonald, D., Nivre, J., & Riedl, M. (2018). Computational approaches to literary style analysis. Digital Scholarship in the Humanities, 33(2), 1–15.

McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press.

Michel, S., Kaur, S., Gillespie, S., Gleason, J., Wilson, C., & Ghosh, A. (2025). "It's not a representation of me": Examining accent bias and digital exclusion in synthetic AI voice services. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (pp. 228–245). Association for Computing Machinery. https://doi.org/10.1145/3715275.3732018

O'Sullivan, J. (2025). Stylometric comparisons of human versus AI-generated creative writing. Humanities and Social Sciences Communications, 12, Article 1708. https://doi.org/10.1057/s41599-025-05986-3

Rahman, T. (1990). Pakistani English. National Institute of Pakistan Studies.

Safeer, N. (2026). An analysis of English vowel variation in Pakistani vs. Arabic talkers: A computational acoustic and machine learning approach [Preprint]. Preprints.org. https://doi.org/10.20944/preprints202601.2282.v1

Safeer, N., Anjum, U., & Saleem, T. (2024). Gender-based study of paired monophthongs: A sociophonetics approach. 3L: Language, Linguistics, Literature, 30(2), 231–262. https://doi.org/10.17576/3L-2024-3002-15

Safeer, N., Malik, S., & Anjum, U. (2023). A descriptive analysis of English vowel sounds by L1 Pahari learners. Pakistan Journal of Society, Education and Language, 9(2), 26–42. https://pjsel.jehanf.com/index.php/journal/article/view/1119

Schneider, E. W. (2007). Postcolonial English: Varieties around the world. Cambridge University Press.

Shamsie, M. (2011). Offence: The Muslim case. Seagull Books.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (Vol. 30). Curran Associates, Inc.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, A., Balle, B., Kasirzadeh, A., Kenton, Z., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L. A., Hendricks, L. A., Isaac, W., & Gabriel, I. (2022). Ethical and social risks of harm from language models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 214–229). Association for Computing Machinery.