Switch-Point Prosody: A Research Proposal for Code-Switched Urdu-English Text-to-Speech in Pakistani Advertising

Why AI voices break where Pakistani ads switch languages — and how to fix it

Muhammad Usman — BS Artificial Intelligence, Superior University, Lahore

October 2026 — Research Proposal (v1.0)

Abstract

Code-switching between Urdu and English is the default spoken register of Pakistani advertising: "Yeh product *best* hai, *order now!*" Modern text-to-speech (TTS) systems handle monolingual input well, but at language switch points they produce audible prosodic breaks — misplaced stress, pitch resets, and mis-phonemized words — that undermine the persuasive delivery an advertisement requires. Existing code-mixed TTS research targets Hindi-English and Mandarin-English and optimizes for intelligibility, not persuasive prosody; published acoustic measurements confirm that switch points measurably distort pitch contours, yet no published work addresses Urdu-English switch-point prosody for advertising speech. This paper proposes a three-phase research program: (1) a corpus of Pakistani code-switched advertising speech, (2) acoustic measurement of switch-point prosody anomalies against human delivery, and (3) prosody-aware synthesis with perceptual evaluation. We present supporting evidence from the literature, a live synthesis demonstration of the failure mode, a worked failure analysis of representative ad scripts, and a fully reproducible experimental protocol. The study is proposed, not yet executed; all claims about expected outcomes are hypotheses to be tested.

1. Introduction

Walk through any commercial break on Pakistani television, or scroll through sponsored video ads on Instagram and TikTok, and a linguistic pattern repeats: the pitch is neither pure Urdu nor pure English. It is a deliberate mix — Urdu as the matrix language carrying the message, English insertions carrying prestige, urgency, and the call to action:

> *"Yeh phone **best** hai — **limited offer**, abhi **order** karo!"*
> *"Naya **smart TV** sirf Rs. 85,000 me — **free delivery** ke saath!"*

This is not sloppy speech. In Myers-Scotton's Matrix Language Frame model, bilingual speakers systematically embed one language's content morphemes inside another language's grammatical frame, and advertisers exploit this: English words signal modernity and trigger the purchase impulse, while Urdu carries trust and cultural proximity. The persuasive force of the ad lives *exactly* at the switch points — the emphatic stress on "*best*", the rising pitch into "*order now!*"

AI-generated voiceovers are now cheap enough for Pakistan's small businesses to use in video ads. But every practitioner who has tried knows the failure: feed a code-switched script to a TTS engine and the voice stumbles precisely where the ad needs it to soar. The Urdu words come out with English phonemes, the stress lands on the wrong syllable, and a flat pitch contour replaces the salesman's rise.

The result sounds foreign to the very audience it must persuade.

This paper argues that **switch-point prosody** — the pitch, stress, and timing behavior of speech at language-switch boundaries — is an unaddressed research problem for Urdu-English, and that solving it is a precondition for believable AI voiceovers in Pakistani advertising.

2. Background and Related Work

2.1 Code-mixed speech synthesis

That TTS systems struggle with mixed-language input is well documented. Sitaram and Black (CMU) built an early framework for synthesizing code-mixed Hindi-English text, identifying language identification, spelling normalization, and pronunciation modeling as the core challenges, and found significant listener preference for systems that correctly identify and pronounce words in each language. Joshi and Garera (2023) took this to production quality for Hindi-English e-commerce voice bots, using transliteration to a common script plus monolingual data combination with Tacotron2 and WaveGlow, reporting a positive comparative MOS against a commercial baseline. For Mandarin-English, researchers combined monolingual recordings from two speakers with cross-lingual voice conversion to build bilingual and code-switched voices with Tacotron2, Transformer, and FastSpeech architectures (Interspeech 2020). An ICASSP 2020 system mapped text to bilingual phonetic posteriorgrams while explicitly predicting log-F0, noting that phonetic representations alone lack the prosody information code-switched synthesis needs.

The pattern across this literature: the language pairs are Hindi-English or Mandarin-English, and the optimization target is *intelligibility* — can the listener understand the words? — not *persuasiveness*: does the delivery sell?

2.2 Measured prosody anomalies at switch points

Crucially, the prosodic disruption at switch points is not anecdotal — it has been measured. A study of Chinese-English bilingual speech compared the fundamental-frequency (F0) distribution of English words spoken in a bilingual corpus against the same words in a monolingual English corpus. English words inside code-switched utterances showed a substantially *wider* pitch range: the average standard deviation of F0 was 51.9 Hz in the bilingual context versus 29.8 Hz in the monolingual context, with the maximum pitch averaging 317.5 Hz against 271.5 Hz. In other words, speakers measurably alter their prosody around switch points — and a TTS system trained on monolingual prosody will therefore misrender exactly those regions.

2.3 The advertising prosody gap

Advertising speech is a distinct prosodic register: wider pitch excursions, emphatic stress on product attributes, and boundary tones engineered for urgency. No published work combines this register with code-switched synthesis, and none at all addresses Urdu-English — despite Urdu-English mixing being the everyday reality of hundreds of millions of speakers and the default voice of Pakistani commerce.

3. Supporting Evidence

We support the proposal with four kinds of evidence: published measurements, a live synthesis demonstration, a worked failure analysis, and a reproducible experimental protocol.

Proof A — Published acoustic measurements.

The Chinese-English F0 study (Section 2.2) provides the quantitative anchor: a near-doubling of pitch variability (51.9 Hz vs 29.8 Hz standard deviation) for words in code-switched versus monolingual contexts. If switch points distort prosody this severely in a tonal/non-tonal language pair, there is no reason to expect Urdu-English — a stress-timed/non-native-stress pairing with Perso-Arabic versus Latin phoneme inventories — to behave better. The burden of proof is on any claim that current TTS handles these boundaries gracefully.

Proof B — Live synthesis demonstration.

To verify the failure mode is real and reproducible with current tooling, we synthesized the representative ad line "Yeh product best hai, order now" through a production English TTS voice (Meta Shortwave API, October 2026). The Urdu function words receive English phonemization ("Yeh" → /j■/ as in "yeah"; "hai" → /he■/ as in "hay"), stress defaults to English patterns, and the persuasive pitch rise into the call to action is absent. The supplementary audio artifact (*tts_proof_codeswitch_en_voice.mp3*) lets any reader verify this in seconds. A monolingual-voice pipeline cannot help but fail here: the failure is architectural, not a tuning issue.

Proof C — Worked failure analysis.

Table 1 analyzes six ad-style utterances typical of Pakistani video advertising. For each, we mark the switch points (■) and predict the failure mode of a monolingual TTS pipeline from the mechanisms documented in Sections 2.1–2.2.

#	Ad script (switch points marked ■)	Predicted TTS failure
1	Yeh phone ■ *best* ■ hai	Stress on wrong syllable of "best"; flat pitch kills the emphatic peak
2	Sirf Rs. 5,000 me ■ *free delivery* ■ ke saath	Pitch reset at boundary; "free delivery" rushed, urgency lost
3	■ *Limited offer* ■ — abhi ■ *order* ■ karo	Triple switch: each boundary a prosodic cliff; call-to-action sounds hesitant
4	Naya ■ *smart TV* ■ sirf aap ke liye	"TV" phonemized per English; vowel mismatch audible to Urdu listeners
5	Dosto, yeh ■ *deal* ■ miss mat karna	Code-switch mid-phrase breaks intonation contour; warning tone flattened
6	■ *Cash on delivery* ■ available hai	Leading English phrase sets wrong prosodic frame for Urdu tail

Table 1: Representative code-switched ad utterances with predicted monolingual-TTS failure modes. Each row is a testable hypothesis for Phase 2.

Proof D — Reproducible experimental protocol.

To move from demonstration to measurement, we specify the experiment any lab — including an undergraduate lab — can run:

Phase 1 — Corpus. Collect 50 code-switched Pakistani ad scripts (TV, Instagram, TikTok). Record 5 native bilingual speakers reading each script naturally. Annotate switch points and the matrix language per utterance (Myers-Scotton framework).

Phase 2 — Acoustic measurement. Extract F0 contours with pYAAPT and energy with librosa. For each switch point, compare a 300 ms window around the boundary against matched monolingual windows from the same speaker: test for differences in pitch range, pitch slope discontinuity, and pre-boundary lengthening (paired t-tests, $\alpha = 0.05$). Synthesize the same scripts with 3 commercial TTS engines and repeat the measurements on synthetic speech.

Phase 3 — Perceptual evaluation. ABX listening tests ($n \geq 40$ native listeners): human vs synthetic delivery of the same script, rated on naturalness (MOS 1–5) and purchase intent ("How likely would you be to order?" 1–5). Report CMOS with 95% confidence intervals.

Metric	What it measures	Success criterion
F0 discontinuity at switch	Pitch jump across the boundary (Hz)	Synthetic within 1 SD of human distribution
Pre-boundary lengthening	Syllable duration before switch (ms)	No significant human–synthetic gap ($p > 0.05$)
MOS naturalness	Listener rating 1–5	Synthetic MOS ≥ 4.0
Purchase-intent rating	Persuasive effect 1–5	Gap to human delivery < 0.5 points

Table 2: Evaluation metrics for the proposed study.

4. Expected Contributions

- The first acoustic characterization of switch-point prosody in **Urdu-English** speech, a pair with zero published TTS prosody data.
- A public corpus of annotated Pakistani code-switched advertising speech (scripts + recordings + switch-point labels).
- Quantified human–synthetic prosody gaps at switch points for current commercial TTS engines.
- A prosody-aware synthesis recipe (switch-conditioned pitch and duration modeling) evaluated on both naturalness and *purchase intent* — the metric advertisers actually care about.

5. Limitations and Honest Framing

This is a **research proposal**: the three-phase study has not yet been executed, and the expected outcomes in Table 2 are hypotheses, not results. The demonstration in Proof B uses a single English voice and one sentence — it is illustrative, not conclusive. Findings from Hindi-English and Mandarin-English systems may partially transfer to Urdu-English, which would narrow but not eliminate the novelty: the advertising-persuasion angle and the Urdu-English pair remain unaddressed. All literature claims above are verifiable through the references below.

References

[1] S. Sitaram and A. W. Black, "Speech synthesis of code-mixed text," Proc. LREC, 2016. Early framework for Hindi-English code-mixed TTS; language identification and pronunciation modeling as core challenges. https://preview.aclanthology.org/override-month/L16-1546.pdf

[2] R. Joshi and N. Garera, "Code-mixed text to speech synthesis under low-resource constraints," arXiv:2312.01103, 2023. Production Hindi-English TTS via transliteration plus monolingual data combination; positive CMOS vs. commercial baseline. https://arxiv.org/abs/2312.01103

[3] "Towards natural bilingual and code-switched speech synthesis based on mix of monolingual recordings and cross-lingual voice conversion," Proc. Interspeech, 2020. Mandarin-English bilingual/code-switched TTS with Tacotron2, Transformer, FastSpeech. https://arxiv.org/abs/2010.08136v1

[4] "Code-switched speech synthesis using bilingual phonetic posteriorgram with only monolingual corpora," Proc. ICASSP, 2020. Bilingual PPG mapping with explicit log-F0 prediction for code-switched prosody. http://dihana.cps.unizar.es/proceedings/ICASSP/2020/pdfs/0007614.pdf

[5] F0 distribution study of English words in bilingual vs. monolingual corpora (Chinese-English). Reports AST D(F0) of 51.9 Hz (bilingual) vs 29.8 Hz (monolingual); μFmax 317.5 Hz vs 271.5 Hz. https://nlpr.ia.ac.cn/2008papers/gjhy/gh162.pdf

[6] C. Myers-Scotton, *Duelling Languages: Grammatical Structure in Codeswitching*, Oxford University Press, 1993. Matrix Language Frame model.

Supplementary artifact: tts_proof_codeswitch_en_voice.mp3 — English-voice synthesis of "Yeh product best hai, order now" demonstrating the switch-point failure (Proof B).