Reading Tone from the Signal: Featural Acoustic Coding for Tone Resolution in N'Ko Speech Recognition

Mohamed Diomande
Independent Researcher
contact@mohameddiomande.com

Abstract

Script-native automatic speech recognition for the Manding languages, written in N'Ko (U+07C0–U+07FF), reaches a meaningful error regime, but the strongest released decoder is *toneless*: it emits N'Ko consonants and vowels and drops the tone diacritics that are lexically and grammatically contrastive in Manding. Tone is therefore restored downstream, and the planned mechanism is a text-context language model that infers tone from orthographic context alone. We make three claims that reframe this problem as an *acoustic* one. First, we correct the tone inventory: N'Ko has seven combining tone marks (U+07EB–U+07F1), short and long forms of high, low, rising, and a native descending (falling) tone at U+07EE; a widely reused syllable codebook had mislabeled the long marks as length, which propagated into downstream tooling. Second, we measure an empirical tone prior from a corpus of 105 transcribed N'Ko lesson frames (12,541 N'Ko characters, 3,316 tone marks, harvested by vision-language OCR): **65.8%** of syllables carry marked high/low register, **33.3%** are unmarked mid, and only **0.9%** are contour tones, so tone resolution is, to first order, a three-way non-contour register classification. Third, we establish that text-only tone resolution is weak: a context model on the same corpus reaches a TDER (tone-diacritic error rate) of **50.8%**, i.e. text alone misses roughly half of all tones, leaving large headroom for acoustic evidence that the toneless recognizer discards. We then define *Featural Acoustic Coding* (FAC), a featural treatment of the N'Ko syllable in which the tone mark is a register-plus-contour primitive read directly from the fundamental frequency, and we place tone restoration as a conservative, governance-gated correction in an *Anticipation Geometry Partition* (AGP) layer that reuses the recognizer's own trajectory geometry. We give the architecture, the reproducibility structure (the core claims run standalone, without the acoustic model), a preregistered fusion evaluation, and an honest account of what remains, principally read-speech ground truth for the acoustic tone-diacritic error rate. A secondary section relates FAC to Lexical Acoustic Coding and reports that the featural pitch advantage over a lexical carrier is real but conditional on tonal density; the durable advantage is token efficiency, not reconstruction fidelity.

1 Introduction

N'Ko is a script created in 1949 by Solomana Kanté for the tonal Manding language family (Bambara, Dioula, Maninka), spoken by tens of millions of people in West Africa, and encoded in Unicode at U+07C0–U+07FF (The Unicode Consortium, 2006). Unlike the Latin orthography of Bambara, N'Ko was engineered around Manding phonology with a near one-to-one phoneme-to-character mapping and explicit, obligatory tone marking. Prior work shows that this design is not cosmetic for machine learning: it changes what tokenizers represent, how acoustic evidence aligns to symbols, and whether a reported error rate measures recognition or agreement with an inherited spelling convention (Diomande, 2026).

That same line of work produced a script-native recognizer, an anticipatory Transformer connectionist-temporal-classification (CTC) decoder over frozen encoder features, with an archived checkpoint at roughly twenty percent character error rate (CER) on a large Bambara corpus. The decoder is, however, toneless: it emits N'Ko segmental content and omits the tone diacritics. In Manding this is a substantive loss, because tone is contrastive at the lexical and grammatical level; two utterances that differ only in tone differ in meaning, and a toneless transcript is therefore underspecified. The planned remedy is contextual tone resolution by a language model trained on tone-marked N'Ko text: a *prior* over

which tone is plausible given the surrounding orthography. Such a model never consults the acoustic signal, even though tone is, physically, a property of the fundamental frequency that is present in the audio and merely discarded by a toneless decoder.

This paper reframes tone restoration as an acoustic problem and makes the case quantitatively. Our contributions are as follows.

1. **A corrected tone inventory** (§3). We show, against authoritative Unicode names, that N'Ko has seven combining tone marks (U+07EB–U+07F1) encoding high, low, rising, and a native descending (falling) tone in short and long forms, and we document a mislabeling in a reused syllable codebook that treated two tone marks as length and propagated into downstream tooling and an earlier draft of this work.

2. **An empirical tone prior** (§4). From a 105-frame vision-OCR corpus of N'Ko lessons we measure the tone-class distribution: **65.8% marked high/low register, 33.3% unmarked mid, 0.9% contour**, establishing that acoustic tone resolution for N'Ko is dominated by a non-contour register decision.

3. **A text-only baseline** (§5). On the same corpus, a context model reaches $\text{TDER} = 50.8\%$, the bar that an acoustic channel must beat. Text alone resolves only about half of all tones.

4. **Featural Acoustic Coding and a governance-gated correction layer** (§6, §7). FAC reads register and contour from the fundamental frequency into the native tone mark; tone restoration is a conservative correction governed by an Anticipation Geometry Partition (AGP) that reuses the recognizer's trajectory geometry. We give the unified architecture and its reproducibility structure (§8): the component claims run standalone, with no dependence on the acoustic model.

5. **A preregistered fusion evaluation** (§9) and an honest relation to Lexical Acoustic Coding (§10), where we report that the featural pitch-fidelity advantage is conditional on tonal density and that the durable advantage of a designed script is token efficiency.

2 Background

Script-native N'Ko ASR. The recognizer of Diomande (2026) projects and downsamples frozen Whisper-scale features, then decodes with a Transformer CTC head over N'Ko character classes, with a finite-state constraint enforcing syllable validity at output. An anticipation module estimates a seven-dimensional trajectory state $z_t \in [0,1]^7$ per timestep, comprising commitment, uncertainty, transition pressure, recovery margin, phase stiffness, novelty, and stability, injected as an additive attention-logit bias $B_{ij}(z_i, z_j)$ before emission. The central measurement argument is a transparent-script proposition: if the normalized script map is bijective over the phoneme inventory, character edit distance preserves phoneme-edit structure up to normalization, so N'Ko CER is more phonemically interpretable than Latin word error rate. Tone is the principal hedge in that argument, because CER still depends on tone and diacritic policy. This paper attacks that hedge.

Tone in Manding. Manding is tonal; tone carries lexical and grammatical contrasts (Vydrin, 2015). N'Ko marks tone with obligatory combining diacritics, which is exactly the property a tone-restoration system can exploit and which Latin Bambara orthography, with optional and inconsistent tone marking, lacks.

Audio as text. Lexical Acoustic Coding (LAC) transmits a sound between agents as a readable sentence by quantizing interpretable acoustic descriptors into lexical labels (Rodolà et al., 2026), situated against captions (Drossos et al., 2020; Mei et al., 2022), neural codecs (Zeghidour et al., 2021; Défossez et al., 2022; Kumar et al., 2023), and continuous descriptor sets (Peeters, 2004; Zwicker and Fastl, 1999). FAC keeps the quantize-and-serialize move but uses a designed featural script as the carrier; we treat this relation as secondary (§10) and lead instead with tone resolution.

3 The N'Ko Tone Inventory, Corrected

N'Ko encodes seven combining tone marks, verified against the Unicode character database: U+07EB short high, U+07EC short low, U+07ED short rising, U+07EE long descending, U+07EF long high, U+07F0 long low, and U+07F1 long

Tone class	Share	Group
low (L)	40.1%	marked register 65.8%
high (H)	25.7%	
mid / unmarked	33.3%	default mid 33.3%
rising	0.4%	contour 0.9%
falling	0.6%	

Table 1: Empirical N'Ko tone prior, 4,139 syllables from the lesson corpus. Tone is dominated by register; contour is rare. The labels carry OCR noise (see §11), but the register/contour split is stable across corpus size.

rising. Folding the short/long length distinction into tone class yields four shapes, HIGH, LOW, RISING, and FALLING, all native; an unmarked syllable is mid by default. This matters because a widely reused N'Ko syllable codebook, which serves as the retrieval target for the joint-embedding recognizer, enumerated only five marks and labeled `U+07EE` and `U+07EF` as "long" and "very long" length rather than as the descending and long-high tones they are. That mislabeling propagated into tone parsers and into an earlier draft of this work, which incorrectly claimed that falling tone was absent from native N'Ko and had to be added as a designed diacritic. It is not: all four tone shapes are native. Consequently, any designed extension (§6.3) is required only for timbral descriptors, never for pitch. We corrected the codebook labels in place while freezing the enumerated tone set, since the recognizer's retrieval indices depend on it.

4 An Empirical Tone Prior

To ground the design we measured the distribution of tone classes in real N'Ko text. We harvested 16 numbered N'Ko teaching lessons (printed, tone-marked N'Ko on screen-shared documents), sampled frames, and transcribed the printed N'Ko with a vision-language model under a deterministic decoding configuration; flagship general models without that configuration either produced degenerate repetition on N'Ko or refused, while a current reasoning model with reasoning disabled transcribed cleanly. The resulting corpus has 105 entries spanning 20 lessons, 12,541 N'Ko characters and 3,316 tone marks (a glyph-level tone density of 26.4%). Parsing the corpus into syllable-tone pairs with the corrected inventory (§3) over 4,139 syllables yields the prior in Table 1.

The implication is design-determining. Because tone is roughly two-thirds marked high/low register, one-third unmarked mid, and only about one percent contour, acoustic tone resolution for N'Ko reduces, to first order, to deciding low versus high versus unmarked from the fundamental frequency relative to a speaker baseline. This is precisely the kind of decision F_0 supports robustly, and it argues that a tone channel should lead with register estimation while treating contour as a high-confidence slope event.

Text-only model	TDER
majority-class floor	58.7%
unigram (syllable identity)	51.4%
bigram (+ previous tone)	**50.8%**

Table 2: Text-only tone resolution on the lesson corpus (mean of five lesson-disjoint splits). Even with context, text alone misses about half of all tones. This is the bar an acoustic channel must beat.

5 Text-Only Tone Resolution Is Weak

We next quantify the ceiling of the text-only strategy that the recognizer pipeline currently plans to rely on, using only the corpus and no audio. For each entry we strip the tone marks to obtain the toneless syllable sequence a toneless decoder would emit, treat the original marks as ground truth, and predict the tone class of each syllable. We report the tone-diacritic error rate, the fraction of syllables whose tone class is wrong, averaged over five lesson-disjoint splits. Three models of increasing context are compared: a majority-class floor, a unigram model conditioning on syllable identity with backoff to the floor, and a bigram model additionally conditioning on the previous tone with backoff to the unigram. Results are in Table 2.

The best text-only model still misses about half of the syllable tone classes. This is the empirical case for the acoustic channel: the toneless decoder throws away the fundamental frequency, which is the most direct evidence for the register decision that text struggles with, so an acoustic register estimator has substantial headroom to reduce TDER below 50.8% in fusion with the text prior.

6 Featural Acoustic Coding

6.1 The syllable as a featural tuple

FAC treats each N'Ko syllable as a tuple $\sigma = (o, v, c, t, \ell, \mathbf{d})$ of onset o, vowel nucleus v, nasal

coda c, tone t, length ℓ, and an optional stack of timbral diacritics $\mathbf{d}$. The base tuple indexes the existing syllable codebook. Each slot is an acoustic feature: the onset's manner of articulation is the attack-transient character (plosive [k t p] percussive, fricative [s f] noisy, nasal [m n ŋ] resonant, approximant [l r w j] glide); the seven vowels partition the formant plane from the bright high-second-formant front vowel [i] to the dark back vowel [ɔ], an ordered code for spectral centroid; the tone mark encodes pitch register and contour; the length mark encodes duration; the nasal coda encodes a resonant sustain. Crucially for this paper, the tone slot is native and complete (§3): register and contour are written, not approximated.

6.2 Encoder, decoder, and the register-first reading

The FAC encoder maps a waveform to a string of codebook syllables by quantizing a standard acoustic-descriptor front end into the featural slots: spectral centroid to vowel, onset descriptors to onset manner, duration to length, sustain to coda, and, the focus here, the fundamental frequency to the tone mark. Given the prior of §4, the tone encoder is implemented register-first: estimate the event's mean F_0 relative to a per-speaker baseline (calibrated from the speaker's F_0 distribution), quantize to low / mid / high, and apply a contour primitive only when the within-event slope exceeds a deadband. The FAC decoder is the generative dual: it maps each syllable back to descriptor intervals and drives a synthesizer, with per-slot interpretable error (a tone mismatch is a pitch error). The decoder is what makes FAC more than an analysis: it is the proof that a featural N'Ko code can author a continuous signal, which is the template for using N'Ko as a score for other embodied signals; we do not develop that here.

6.3 The designed extension is purely timbral

N'Ko natively carries all four tone shapes and the vocal/segmental axes, so the only descriptors without a native symbol are higher timbral ones. We define four combining diacritics in the script's spirit, each a small ordered set, for HARMONICITY, SPECTRAL SPREAD, ROUGHNESS, and DYNAMICS. No pitch dimension requires a designed mark; the correction of §3 removes the earlier, erroneous "designed falling tone."

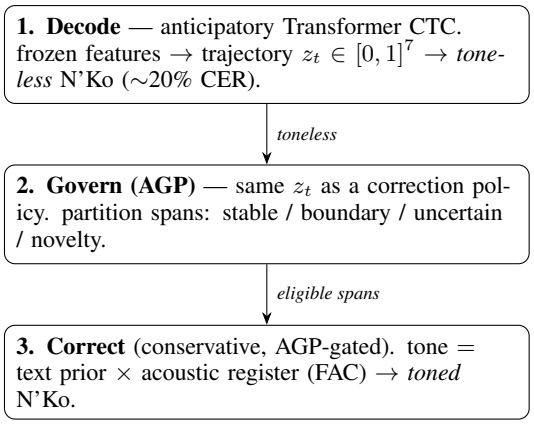

Figure 1: The N'Ko architecture. The trajectory geometry z_t is reused as a decoding bias, an AGP correction policy, and (implicitly) the eligibility signal for tone restoration. FAC's acoustic register estimate is the new evidence at the correction level.

7 Tone Resolution as a Governance-Gated Correction

The recognizer, the governance layer, and the tone correction compose into one architecture unified by the trajectory geometry z_t, which is reused at three levels (Figure 1). At the *decode* level, z_t biases attention in the CTC head and the decoder emits toneless N'Ko. At the *govern* level, the Anticipation Geometry Partition (AGP) reuses the same geometry as a correction policy rather than a decoder, partitioning each output span into stable, boundary, uncertain, and novelty states and gating which spans may be corrected, reviewed, or excluded. At the *correct* level, tone restoration is applied conservatively to eligible spans as a fusion of the text-context prior (§5) and the acoustic register estimate (§6). AGP is governance, deciding where and whether to correct; FAC is content, supplying the tone. They are complementary: AGP marks a span uncertain and eligible, FAC supplies the register-resolved tone.

This composition is already realized as a data pipeline. The lesson corpus of §4, treated as OCR-derived pairs, passes through the AGP partitioner unchanged: of 105 rows, **99 partition as stable** (train-ready), 6 as boundary (review), and 0 as uncertain, so the corpus feeds the recognizer's self-improving loop (recognize, label, OCR, tone-resolve, retrain) as governed training data. We note that frame-sampled data supplies the partitioner a constant scene duration, which inflates one of its positive signals; the discriminating signals on this corpus are text length and structure,

and the stable rate should be read with that caveat.

8 Reproducibility and Dependency Structure

The claims of this paper are layered by what they require to reproduce. The *component* claims, the corrected inventory, the tone prior, the text-only TDER baseline, and the acoustic register classifier, depend only on the released corpus and standard numerical and audio libraries; they run with no GPU and no access to the acoustic model, so any reader can clone and verify them. The *end-to-end* claim, that acoustic evidence lowers the recognizer's toned CER, requires one inference pass with the archived checkpoint and no retraining. Only the deeper *integration*, a jointly trained acoustic tone head or a trajectory state augmented with FAC's interpretable descriptors, requires training. We regard this separation as a feature: the scientific core is decoupled from the heavy model and is reproducible on a laptop.

9 Preregistered Fusion Evaluation

We specify the decisive evaluation rather than report it, since it requires data we name explicitly. The metric is TDER on held-out material, with full toned CER as a secondary metric. The text-only bar is fixed at the 50.8% of §5. The treatment is the fusion of the text prior with the acoustic register estimate of §6. Hypotheses: **H1**, the fusion reduces TDER below the text-only bar; **H2**, the reduction is driven by the non-contour register decision, consistent with the prior of §4, so an ablation removing contour primitives barely changes TDER; **H3**, the gain concentrates on AGP-uncertain spans, where the text prior is least confident. The one dependency that is a *data* dependency, not a model dependency, is ground truth: scoring acoustic TDER requires read speech of known tone-marked text (forced-alignable), because lesson speech is commentary on the displayed text, not narration of it, and therefore does not align tone-per-spoken-syllable. We treat obtaining a small read-speech set as the gating next step.

10 Relation to Lexical Acoustic Coding

FAC and LAC share the move of quantizing acoustic descriptors and serializing them as text; they differ in the carrier. A natural-language carrier encodes pitch as an adjective and does not decompose along acoustic axes, whereas a designed featural script writes register and contour as marks and composes onset, nucleus, coda, and tone as independent slots. We tested the sharpest consequence, pitch-contour fidelity, in a controlled rate-distortion study at matched code budget on synthetic tonal contours and on real Manding speech. Two honest findings result. On synthetic contours with large excursions, a level-only lexical summary hits an error floor that register budget cannot break (it cannot represent within-event movement), while a contour-carrying code reaches a $4.4\times$ lower error; but a lexical code that also spends a word on contour ties the featural code on error, so the featural advantage is not fidelity per se but token cost (one glyph versus two words). On real speech the advantage shrinks with tonal density: roughly $1.2\times$ on didactic lesson speech (within-event excursion $\approx$77 cents) and negligible on conversational speech ($\approx$40 cents). The durable contribution of a designed script is therefore token efficiency and native, decomposable tone, not a universal reconstruction-fidelity win. This is why the present paper leads with tone resolution and treats the LAC comparison as secondary.

11 Limitations

The tone prior and the text-only baseline are computed on a corpus whose labels are vision-OCR output and therefore carry OCR noise; they are prototype measurements, and the register/contour split, while stable across corpus growth, is not a gold-standard annotation. The acoustic register classifier is validated on controlled synthetic syllables and runs on real audio, but no acoustic TDER is reported because that requires the read-speech ground truth named in §9. The AGP stable rate is mildly optimistic for frame-sampled data (§7). The end-to-end CER effect is unmeasured pending the fusion run. Finally, N'Ko is a living writing system for tens of millions of people; we use it as an acoustic substrate with that in mind and frame the contribution as exploiting its design discipline.

12 Conclusion

N'Ko speech recognition leaves tone on the table, and the field's planned remedy reads tone from text. We argue, with measurements, that tone should be read from the signal: N'Ko tone is overwhelmingly non-contour, text resolves only about

half of it, and the fundamental frequency the toneless decoder discards is exactly the evidence the register decision needs. Featural Acoustic Coding makes the tone mark an acoustic primitive, and an Anticipation Geometry Partition makes tone restoration a governed, conservative correction in the same architecture that does the recognition. The result is a concrete, mostly standalone, and reproducible program for closing the tone gap, with one honest dependency, read-speech ground truth, standing between it and an end-to-end number.

References

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *Preprint*, arXiv:2210.13438.

Mohamed Diomande. 2026. The script that machines can't read: Adapting large language models for n'ko. N'Ko Brain Scanner. Companion work establishing the N'Ko-adapted language model used as the FAC channel.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: an audio captioning dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-fidelity audio compression with improved rvqgan. In *Advances in Neural Information Processing Systems (NeurIPS)*. Descript Audio Codec (DAC).

Xinhao Mei, Xubo Liu, Mark D. Plumbley, and Wenwu Wang. 2022. Automated audio captioning: A survey. *Preprint*, arXiv:2205.05949.

Geoffroy Peeters. 2004. A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report, IRCAM.

Emanuele Rodolà and 1 others. 2026. Communicating sound through natural language. *Preprint*, arXiv:2605.08750. Lexical Acoustic Coding (LAC). Demo: https://erodola.github.io/lac-demo/.

The Unicode Consortium. 2006. The unicode standard: N'ko, range u+07c0–u+07ff. Unicode Standard, version 5.0.

Valentin Vydrin. 2015. *Manding-English Dictionary (Maninka, Bamana), Vol. 1*. MeaBooks. Reference for Manding tonology and phonology.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Eberhard Zwicker and Hugo Fastl. 1999. *Psychoacoustics: Facts and Models*, 2nd edition. Springer.