Linguistically Informed Evaluation of Multilingual ASR for African Languages

Fei-Yueh Chen[*]   **Lateef Adeleke**[*]   **C.M. Downey**
Department of Linguistics
University of Rochester
{fchen27, ladeleke, cdowney4}@ur.rochester.edu

Abstract

Word Error Rate (WER) mischaracterizes ASR models' performance for African languages by combining phonological, tone, and other linguistic errors into a single lexical error. By contrast, Feature Error Rate (FER) has recently attracted attention as a viable metric that reveals linguistically meaningful errors in models' performance. In this paper, we evaluate three speech encoders on two African languages by complementing WER with CER, and FER, and add a tone-aware extension (TER). We show that by computing errors on phonological features, FER and TER reveal linguistically-salient error patterns even when word-level accuracy remains low. Our results reveal that models perform better on segmental features, while tones (especially mid and downstep) remain the most challenging features. Results on Yoruba show a striking differential in metrics, with WER=0.788, CER=0.305, and FER=0.151. Similarly for Uneme (an endangered language absent from pretraining data) a model with near-total WER and 0.461 CER achieves the relatively low FER of 0.267. This indicates model error is often attributable to individual phonetic feature errors, which is obscured by all-or-nothing metrics like WER.

1 Introduction

Recent models for multilingual ASR, such as wav2vec2.0 (Baevski et al., 2020) or multilingual HuBERT variants (Boito et al., 2024), have made ASR for low-resource languages more tractable. Word Error Rate (WER) is still the standard for ASR model evaluation, but while useful, it has known limitations such as its treatment of all errors as the same, and its inability to provide linguistically meaningful information for error analysis. K et al. (2025) make the case that WER is fundamentally inadequate as a general evaluation metric for multilingual ASR, since its word-based evaluation

leads to high error scores that do not reflect actual transcription quality or human judgment. Similarly, recent Afro-centric surveys have shown that the issues with African ASR may not only be attributable to data scarcity, but also to the metrics for measuring models' performance. Imam et al. (2025a) confirm that Word Error Rate (WER) is the dominant evaluation metric in ASR for African languages despite its limitations in capturing tonal errors, diacritic variation, and morphological richness. Thus, evaluation via WER obscures insights into whether the model is failing primarily on a particular tone or segment contrast. Recent work advocates for metrics that consider smaller units like characters (CER) (K et al., 2025), phonemes (Mortensen et al., 2016; Imam et al., 2025b), and features (FER) (Zhu et al., 2021).

In this work, we evaluate multilingual ASR for two low-resource African languages by complementing WER with CER, as well as FER, which computes errors over phonological features of each segment. We also add tones to the feature vector to capture Tone Error Rate (TER), which is particularly significant for African languages, where tones are contrastive linguistic units. We conduct experiments on two typologically-related languages — Yoruba and Uneme — using mHuBERT-25-Hz from TWIST (Hassid et al., 2023) and mHuBERT-147 (Boito et al., 2024) before and after language finetuning. For Yoruba, we use the FLORES dataset (Conneau et al., 2022), and for Uneme, we use a new dataset developed from our fieldwork with the speech community in Edo state, Nigeria. This way, we are able to compare the model performance between a language that was seen by the pretrained model (Yoruba), and one that was not (Uneme), both before and after finetuning. For a baseline, we also evaluate our framework on English data from FLORES.

Our contributions include: (1) Providing the first ASR baseline and evaluation for Uneme, an en-

---

[*]These authors contributed equally to this work.

dangered Edoid language; (2) Extending Feature Error Rate (FER) evaluation to include Tone Error Rate (TER) by adding suprasegmental tone features to the phonological feature vector; (3) Demonstrating empirically that WER systematically mischaracterizes models' linguistic performance in tonal African languages by collapsing phonological learning into lexical failure; (4) Showing that phonological features are learned even when models struggle with lexical accuracy on unseen endangered languages; (5) Showing that careful, word-by-word speech degrades ASR performance compared to natural speech when models are trained on conversational data — evidence of domain-shift sensitivity. Overall, we demonstrate that active collaboration between field linguists, speech communities, and NLP practitioners enables richer evaluation frameworks that respect the linguistic typology of African languages.

2 Related Work

2.1 Objective Metrics for Speech Recognition

Several works have shown that representing phones with sub-segmental phonological vectors improves multilingual ASR and provides evaluation metrics that offer deeper linguistic insights into models' performance.

Mortensen et al. (2016) introduce PanPhon, a tool that maps IPA symbols to vectors of articulatory features. Because PanPhon treats segments as divisible into articulatory features, it allows for feature-aware modeling and evaluation across thousands of segments in a shared phonological space. PanPhon enables computation of feature-based edit distance, allowing diagnosis of phonological errors in models' performance. However, PanPhon only captures segmental features and does not consider suprasegmental features such as tone in its feature vectors.

Zhu et al. (2021) extends the use of phonological features through their JoinAP architecture, which combines phonological feature embeddings with neural acoustic models. They represent each segment as a vector of phonological features modeled after PanPhon, and then transform this to a phone embedding used in the acoustic output layer of the model. They show that a feature-based crosslingual ASR method enables sounds that are similar across languages to be linked by shared features rather than treated as unrelated, thereby improving recognition performance in multilingual ASR.

These works demonstrate that phonological feature representations are effective for evaluating multilingual ASR, and can provide linguistic insight during error analysis. We build on these ideas by extending evaluation metrics to include FER in addition to the WER and CER to African languages, and because tone evaluation is essential for most African languages, such as Uneme and Yoruba, where tones are distinctive and overtly marked on vowels, we added tones to the feature vectors.

2.2 Speech Encoders

Speech encoders use neural networks to transform raw waveforms into high-dimensional latent representations. The wav2vec 2.0 framework (Baevski et al., 2020) combines a CNN feature extractor with a Transformer encoder (Vaswani et al., 2023). This has become a dominant approach in recent years. While wav2vec 2.0 uses contrastive learning, HuBERT (Hsu et al., 2021) uses a masked prediction objective. The model is trained to predict discrete acoustic units, which are generated by clustering audio features (e.g., via K-means).

Recent work has extended the HuBERT architecture to different domains. TWIST (Hassid et al., 2023) proposes a pipeline for audio-language models, including speech encoders, quantizers, language models, and vocoders. They release multiple HuBERT-based models trained with additional English data. mHuBERT-147 (Boito et al., 2024) scales the model to a multilingual setting, training on 147 languages to learn universal speech patterns.

3 Data

The data for this work are from two different sources. **The Yoruba dataset** is from the FLORES multilingual dataset (Conneau et al., 2022), which includes sentence-level audio and transcript pairs for more than 100 languages, including Yoruba. The Yoruba portion consists of more than 10 hours of audio-transcript pairs, with audio sampled at 16kHz. While Yoruba is not a high-resource language in an absolute sense, it has better description and public datasets than Uneme, which is an understudied, endangered language with no existing public ASR dataset.

The Uneme dataset is newly created and entirely from our 2025 fieldwork in the speech community in the northern part of Edo state, Nigeria. The recordings and transcripts were collected as part of a broader documentation project on Uneme

Language	Split	Utterances
Yorùbá	Train	2,339
	Validation	378
	Test	831
	Total	**3,548**
English	Train	2,602
	Validation	394
	Test	647
	Total	**3,643**

Table 1: Number of utterances in the English and Yorùbá subsets of the FLEURS speech corpus.

Language	Split	Utterances
Uneme	Train	4,729
	Validation	880
	Test-natural	1027
	Test-careful	1051
	Total	**7,687**

Table 2: Number of utterances in our newly created Uneme dataset.

language, culture, and traditions, archived at the Endangered Languages Archive (Adeleke, 2025). The dataset includes about 8 hours of recordings and captures diverse genres, including narratives, interviews, and elicitation. There are seven speakers (4 males and 3 females) in the portion used for this work.

Uneme Data Processing The recordings were segmented by pauses using Audacity (Audacity Team, 2025) and SayMore(SIL International, 2025), and manually transcribed by trained linguists who are familiar with the language using SayMore. Our transcription preserves the orthography of the language by explicitly marking tones on vowels and syllabic consonants.

In an attempt to test whether speech style can affect the model's performance, we collected a separate, careful speech dataset for Uneme. In this subset, a different speaker reproduces slow, clear, and deliberate renditions of selected utterances from the original recordings by reading them word for word. The careful speech data is used only for testing, and it is not included in model training or validation because it is a very small dataset. In this way, we gauge whether models' performance would improve with utterances produced word-for-word over the baseline of natural speech.

Thus, we have two separate test datasets: Test-natural to evaluate on the usual pace and style of speaking, and test-careful to test on careful speech. Overall, the Uneme dataset has the fewest hours of audio among the three languages. The representations in Tables 1 and 2 are at the utterance level and not by the number of hours. The Uneme data has more utterances only because we use Say-more and Audacity for automatic segmentation by pauses, unlike the dataset from FLORES, which is segmented by sentences.

4 Methodology

We conduct monolingual ASR training and evaluation on Uneme, Yoruba, and English with three types of encoders and two types of decoders.

4.1 Representation

A pretrained encoder, given a batch of audio waveforms $w \in \mathbb{R}^{batch \times samples}$ with a fixed sample rate such as 16 kHz, transforms the waveforms into a high-dimensional representation $v \in \mathbb{R}^{batch \times frames \times dim}$.

A trainable decoder, given the vocabulary size V of the training language, maps the representation v to the distribution of the vocabulary, where the output is $q \in \mathbb{R}^{batch \times frames \times V}$. CTC loss and decoding are subsequently computed for loss calculation and prediction (Graves et al., 2006), and the final output sequence consists of characters with a sequence length less than or equal to $frames$.

4.2 Model Architecture

Encoders We select mHuBERT-25-Hz (Hassid et al., 2023) as the baseline. It in turn is based on a HuBERT-base-95M (Hsu et al., 2021) checkpoint, but is trained with additional English data. In addition, since it has a larger hop size than the other checkpoints, we expect it to be less sensitive than the other two checkpoints. Comparable multilingual models such as wav2vec2.0 XLS-R (Baevski et al., 2020) and mHuBERT-147 (Boito et al., 2024) have a hop size of 320 (50 Hz) and were trained on more than a hundred languages.

Decoders During training, the encoder is frozen, and only the decoder is trainable. We design two types of decoders: a Linear Layer and a 2-Transformer Block model. The former contains only one fully connected layer, which is designed to better test the ability of the encoders, while the latter consists of two Transformer Blocks and is expected to yield better results.

For each Transformer block, we apply a K-V cache, attention with Rotary Embedding (Su et al., 2024), SWiGLU (Shazeer, 2020) as the activation function, and RMSNorm (Zhang and Sennrich, 2019) for pre-normalization. The design is inspired by GPT-OSS (OpenAI, 2025).

4.3 Phonological Features

To effectively capture the phonological differences between Uneme, Yoruba, and English, we use a sparse feature representation inspired by PanPhon (Mortensen et al., 2016). Before feature extraction, we perform Grapheme-to-Phoneme (G2P) conversion. For Uneme and Yoruba, the orthography corresponds closely to the International Phonetic Alphabet (IPA), and the segments that are otherwise represented with symbols other than IPA have one-to-one mapping with segments, enabling us to use direct rule-based mapping. For English, we utilize the g2p-en toolkit[1] to obtain phonetic transcriptions. After the G2P transformation, we map each phonetic segment to a 24-dimensional vector. The values in the vector are strictly from the set $\{-1, 0, +1\}$. A value of $+1$ indicates the presence of a feature, -1 indicates its absence, and 0 means the feature is not applicable or unspecified.

This sparsity design allows us to distinguish between a wrong prediction and an irrelevant feature. For example, English phonemes do not have lexical tones, so all tone-related dimensions for English are set to 0. This prevents the model from being penalized for undefined attributes when we compare languages with different phonological systems.

Our feature inventory is designed to support Edoid and Yoruboid phonology. We treat complex consonants as single segments rather than sequences or consonant clusters. This includes double-articulated labial-velars (e.g., *kp*, *gb*)specific fricatives in Uneme (*vb*, *gh*, *kh*), and affricates (*ch*). The complete list of feature definitions and segment mappings is provided in Appendix Tables 8, 9, 10, 11.

4.4 Tone Features

Since Uneme and Yoruba are tonal languages, where tones are consistently marked on the vowels or syllabic nasals, we model tone as a suprasegmental property that is anchored to vowels and syllabic nasals. We dedicate specific dimensions in the feature vector to represent tone levels (TONE_HIGH,

TONE_LOW, TONE_MID).

We apply language-specific rules to map orthography to these tone features:

- **Uneme:** Based on the orthography, we map acute accents (´) to HIGH and grave accents (`) to LOW. Importantly, vowels without any diacritic marks are mapped to DOWNSTEP.

- **Yoruba:** We distinguish between the three distinctive tone levels in the language: HIGH, LOW, and MID (typically unmarked as in the orthography of the language).

- **English:** Since English is a non-tonal language, all tone features are masked as 0.

By separating tone features from other articulatory features, we can independently evaluate the model's performance on intonation.

4.5 Phonetic Feature Error Rate

To evaluate the quality of the generated sequences, we use the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) to align the hypothesis sequence with the reference sequence. We modify the standard substitution cost to fit our sparse feature representation. We calculate an "NA-masked normalized distance" between a reference vector v_{ref} and a hypothesis vector v_{hyp}.

The distance is computed only over the dimensions where the feature is active (non-zero). The formula is:

$$Cost(v_{ref}, v_{hyp}) = \frac{\sum_i \mathbb{I}(v_{ref}^{(i)} \neq v_{hyp}^{(i)}) \cdot M^{(i)}}{\sum_i M^{(i)}}$$

(1)

where $M^{(i)}$ is a mask that equals 1 if either $v_{ref}^{(i)}$ or $v_{hyp}^{(i)}$ is non-zero, and 0 otherwise. This results in a cost between 0 and 1, representing the ratio of mismatched features among the relevant ones.

The final Feature Error Rate (FER) is the sum of the alignment costs divided by the total number of segments in the reference. In addition, we compute a Tone Error Rate (TER) by strictly considering the tone-related dimensions.

5 Results

We evaluate the performance of the models across three languages: English (high-resource baseline, lots of data in pre-training), Yoruba (low-resource, seen in pre-training), and Uneme (endangered, unseen in pre-training). We compare two decoder

---

[1] https://pypi.org/project/g2p-en/

architectures (Linear vs. Transformer) and different encoder checkpoints. The experimental setup is shown in Appendix A.

5.1 English (Baseline)

As a high-resource benchmark, English allows us to validate the feature extraction pipeline. Among all configurations, the mHuBERT-147 + Transformer setting achieved the best performance, yielding a WER of 0.511 and a CER of 0.167. This represents a substantial improvement (approximately 20% relative reduction in WER) over the standard mHuBERT baseline.

Evaluating the FER provides deeper insights into the specific phonetic challenges. We find that while consonant classes (stops, fricatives, nasals) showed moderate stability (FER $\approx$0.11–0.18), vowel-related features proved more error-prone. Specifically, back vowels (fer_V_BACK $\approx$ 0.27) and rhotic vowels (fer_V_RHOTIC $\approx$ 0.35) were sources of significant error. More specifically, diphthongs exhibited the highest error rate ($\approx$0.45), reflecting the difficulty of modeling dynamic spectral changes and multi-target vowels in English.

As expected, linear decoders fail to produce competitive results. The wav2vec + Linear configuration collapses into degenerate outputs (e.g., repetitive sequences), resulting in a WER of 1.0. Even with the stronger mHuBERT-147 representations, the linear decoder produces high-entropy, vowel-heavy sequences, confirming that a simple linear mapping is insufficient for decoding English orthography from acoustic embeddings.

5.2 Yoruba (Seen African Language)

Similar to English, the mHuBERT-147 + Transformer model outperforms other configurations, achieving the lowest WER (0.788) and TER (0.372). While segmental accuracy is relatively robust—with consonantal and basic vowel FERs ranging between 0.06 and 0.12—suprasegmental features remain a challenge.

The Tone Error Rate (TER) of 0.372 indicates that over one-third of tone-bearing units are misclassified. A breakdown of feature errors reveals that the mid tone (fer_TONE_M $\approx$ 0.34) was more challenging than High or Low tones. Furthermore, specific vowel qualities such as Advanced Tongue Root (ATR) and nasalization showed elevated error rates (0.14 and 0.11, respectively). Qualitative analysis suggests that while the model captures the segmental skeleton of Yoruba, it struggles to dis-

ambiguate lexical items that rely solely on tone and fine-grained vowel contrasts.

5.3 Uneme (our novel dataset)

We evaluate Uneme on two distinct test sets: natural speech (uneme-ASR) and careful speech (careful_speech). The results are presented in Table 4. On the natural speech test set, the mHuBERT-147 + Transformer model achieves a WER of 0.997. While the WER is near-total, the CER (0.461) and FER (0.267) indicate that the model is learning phonological structures despite failing to output correct lexical items. The Tone Error Rate is 0.402.

Surprisingly, performance on the careful_speech dataset was significantly worse across all metrics. The same model yielded a WER of 1.622, CER of 0.837, and FER of 0.515. We discuss the implications of this counter-intuitive result in Section 6.

Comparing mHuBERT-147 + Transformer results across both datasets, we observe a consistent trend in feature errors. The worst performing feature is consistently V_BACK (FER=0.280 for uneme-ASR, FER=0.373 for careful_speech), followed closely by other vowel features such as V_HEIGHT_MID and V_ROUND. This indicates that the model systematically struggles with disambiguating vowel backness contrasts (e.g., /u/ vs. /i/), regardless of the speech style. Similarly, for tone prediction, DOWNSTEP (the unmarked tone) is consistently the most challenging category or is nearly tied with High Tone errors. For instance, on uneme-ASR, DOWNSTEP FER is 0.565 while High Tone FER is 0.559. This suggests that the model has particular difficulty learning the absence of marked tone as a distinct prosodic feature.

We also note the extreme performance degradation of the wav2vec + Linear baseline. With a WER of 1.0 and CER > 0.9, this model essentially fails to learn any meaningful mapping, which is reflected in its near-random FER scores (e.g., V_BACK $>$ 0.99). This highlights that simply applying large multilingual encoders without adequate decoder depth or finetuning data is insufficient for low-resource, tonal languages like Uneme.

6 Discussion

6.1 The Validity of FER over WER

Our results highlight the inadequacy of Word Error Rate (WER) for low-resource African languages.

Encoder	Decoder	WER↓	CER↓	FER↓	TER↓	Worst F	Worst T
		English (en_us)					
mHuBERT-25-Hz	Transformer	0.642	0.216	0.212	–	Diphthong	–
mHuBERT-147	**Transformer**	**0.511**	**0.167**	**0.144**	–	Diphthong	–
mHuBERT-25-Hz	Linear	0.802	0.272	0.254	–	Diphthong	–
mHuBERT-147	Linear	1.000	0.765	0.740	–	Lateral	–
XLR-S	Linear	1.000	0.941	0.924	–	V_Rhotic	–
		Yoruba (yo_ng)					
mHuBERT-25-Hz	Transformer	0.827	0.335	0.170	0.385	V_ATR	Mid
mHuBERT-147	**Transformer**	**0.788**	**0.305**	**0.151**	**0.372**	V_ATR	Mid
mHuBERT-25-Hz	Linear	0.911	0.400	0.199	0.468	V_ATR	Mid
mHuBERT-147	Linear	1.000	0.840	0.630	0.614	Dorsal	Mid
XLR-S	Linear	0.999	0.934	0.901	0.949	V_Back	Mid

Table 3: ASR performance on FLEURS English and Yoruba test sets. TER is reported only for Yoruba. We also report tone and feature (except for tone) with the worst score (Worst F, Worst T).

Dataset	Decoder	WER ↓	CER ↓	FER ↓	TER ↓	Worst F	Worst T
		Encoder: mHuBERT-25-Hz					
careful_speech	Transformer	1.040	**0.668**	**0.417**	**0.490**	V_Back	Downstep
	Linear	1.006	0.699	0.446	0.596	V_Back	Downstep
uneme-ASR	Transformer	1.058	0.512	0.295	0.469	V_Back	Downstep
	Linear	1.003	0.589	0.366	0.596	V_Back	Downstep
		Encoder: mHuBERT-147					
careful_speech	Transformer	1.622	0.837	0.515	0.510	V_Back	Downstep
	Linear	1.006	0.791	0.591	0.772	V_Back	Tone_H
uneme-ASR	Transformer	**0.997**	**0.461**	**0.267**	**0.402**	V_Back	Downstep
	Linear	1.000	0.627	0.494	0.702	V_Back	Downstep
		Encoder: XLS-R					
careful_speech	Linear	1.000	0.918	0.858	0.997	V_Back	Downstep
uneme-ASR	Linear	1.000	0.906	0.873	0.993	V_Back	Downstep

Table 4: ASR performance on Uneme datasets (`careful_speech` and `uneme-ASR`) across different encoders. We report tone and feature (excluding tone) with the worst score (Worst F, Worst T). Bold indicates the best result per dataset.

For Uneme, the WER hovers around 1.0 (100%) for most models, implying a complete failure of the system. However, the Feature Error Rate (FER) paints a different picture. The best Uneme model achieves an FER of 0.267, meaning that nearly 74% of the phonological features were correctly predicted. This disparity indicates that the model is successfully learning the acoustic-to-phonetic mapping but misses some of the linguistic features — e.g.tone and vowel heights — required to resolve the sounds into valid words. For this reason, FER provides a more nuanced (and less pessimistic) metric of progress, especially with endangered languages where WER is often 100%. FER provides insights that would be helpful to im-

prove data collection and language features that need more attention.

6.2 Rethinking CER

Comparing CER and WER scores, our results support the same conclusion as K et al. (2024); Mortensen et al. (2016): that WER doesn't adequately evaluate a model's performance, and may be better replaced by CER for multilingual ASR. Our results show that WER scores perform badly under the low-resource training, while CER and FER demonstrate more stability.

However, relying solely on CER presents significant limitations in evaluation granularity. While CER serves as a proxy for phonetic accuracy to

Dataset	Decoder	TER ↓	FER (H) ↓	FER (L) ↓	FER (DS) ↓
Encoder: mHuBERT-25-Hz					
careful_speech	Transformer	**0.490**	**0.603**	**0.574**	**0.693**
	Linear	0.596	0.664	0.674	0.736
uneme-ASR	Transformer	0.469	0.601	0.533	0.631
	Linear	0.596	0.693	0.636	0.718
Encoder: mHuBERT-147					
careful_speech	Transformer	0.510	0.584	0.621	0.745
	Linear	0.772	0.824	0.793	0.824
uneme-ASR	Transformer	**0.402**	**0.559**	**0.451**	**0.565**
	Linear	0.702	0.696	0.637	0.848
Encoder: XLS-R					
careful_speech	Linear	0.997	0.993	0.996	1.000
uneme-ASR	Linear	0.993	0.996	0.982	0.999

Table 5: Tone Error Rate (TER) and Tone Feature Error Rates (FER) for Uneme datasets using different encoders and decoders. FER columns break down errors by tone category: High (H), Low (L), and Downstep (DS).

Language	Type	Transcription
Yorùbá	Reference	roland mendoza yin ìbọn rẹ m16re mọ́ àwọn arìnrìnàjò
	Hypothesis	ro land mẹdósáyí ìbn rẹ̀ ẹnsistire mọ́ àwọn arìrìn àjò
	Metrics	WER=88.89% CER=45.45% FER=24.14% TER=26.32%
Uneme	Reference	è kwágù mariki ọ́mọ́ kirì ọ̀ọ́furinì ọremọ̀nì
	Hypothesis	ekwá gù marekí ọ́mọ́ kerè ọ̀ọ́fúri nọremọ̀rì
	Metrics	WER=85.71% CER=32.50% FER=6.22% TER=21.05%

Table 6: **Example ASR Outputs with Error Type Analysis.** Errors are color-coded (segmentation not highlighted): tone error only , featural error , deletion/insertion , mixed errors , correct . The model captures most segmental features correctly while struggling primarily with tones and vowel quality.

some extent, it treats all character substitutions equally, regardless of phonetic distance. A substitution error involving a single feature mismatch (e.g., voicing) is penalized identically to a substitution involving multiple feature mismatches. This lack of nuance makes fine-grained evaluation infeasible and obscures the model's partial success in learning phonological structures.

Furthermore, we observe discrepancies where models achieve high FER scores despite poor CER. This phenomenon occurs when a model consistently mis-predicts a specific, high-frequency feature while correctly predicting the majority of other features. In such cases, CER penalizes the entire character for the single feature failure, whereas FER accurately reflects that the bulk of the phonological information was preserved.

Crucially, when comparing CER with TER, it becomes evident that CER cannot isolate the impact of tonal errors within the overall error rate. Our experiments on tonal languages, such as Un-

eme and Yoruba, reveal that the TER is consistently higher than the error rates for consonantal and vocalic features. For instance, in the uneme-ASR test set using the mHuBERT-147 encoder, the model had a TER of 0.402, while the consonantal and vocalic FERs were significantly lower (e.g., syllabic and consonantal FER at 0.145). This confirms that a substantial portion of character-level errors do not stem from fundamental phonetic misspellings, but rather from mis-predictions of intonational contours. These are vital linguistic insights that FER and TER can provide, but which remain invisible when relying on CER alone.

6.3 Tone and Suprasegmental Challenges

Tones remain a significant challenge. In Yoruba, the Tone Error Rate was 0.372, which is higher than the overall FER (0.151), with the mid tone contributing the highest error rate.

Similarly, in Uneme, the Tone Error Rate (TER) was 0.402, with the downstep tone contributing the highest error rate. The TER is significantly higher

than the overall FER (0.267). These facts, taken together, confirm that standard spectral features in architectures like HuBERT may not sufficiently capture pitch significance without explicit pitch-aware pre-training or augmentation.

6.4 The "Careful Speech" Paradox

A striking observation is the performance degradation on the `careful_speech` dataset. Intuitively, slow and articulated speech should be easier to recognize. However, our best model degraded from an FER of 0.267 (natural) to 0.515 (careful).

We hypothesize two reasons for this. First, **domain mismatch**: The model was fine-tuned on natural, conversational data, and careful speech is not in our training set. The prosody of "careful speech" characterized by exaggerated pauses, distinct syllabification, and altered pitch contours—likely constitutes an out-of-distribution shift for the model. Second, **overfitting to speaking style**: The model may have overfit to the specific speaking rate and co-articulation patterns of the training speakers. This finding suggests that for low-resource languages, "clean" and word-by-word read speech is not always the best test set if the training data is spontaneous; the model expects the messiness of natural speech.

6.5 Transformer vs. Linear Decoders

Across all languages, the Transformer decoder significantly outperformed the Linear decoder. In English, the Linear decoder on mHuBERT-147 yielded a WER of 1.074, while the Transformer reduced this to 0.511. This confirms that even with powerful pre-trained encoders, a simple linear projection is insufficient for mapping acoustic representations to discrete tokens, especially when the target script (orthography) requires complex tonal integration (as in Yoruba/Uneme).

7 Limitations

A primary limitation of our evaluation framework is its heavy reliance on the quality of Grapheme-to-Phoneme (G2P) conversion. Our methodology relies on a transparent mapping between the target language's orthography and its phonetic realization. This assumption holds for Uneme and Yoruba, where the writing systems align closely with the International Phonetic Alphabet (IPA) and symbols are faithful to sounds. However, applying this framework to languages with deep orthographies or opaque sound-to-symbol correspondence would require highly accurate, language-specific G2P models, which may not be available for many low-resource languages.

Furthermore, our experimental design focuses on linguistic feature accuracy but does not account for extralinguistic variability. We did not explicitly model speaker-specific acoustic characteristics, such as age, gender, accents, and idiolects, which can significantly influence model performance. Similarly, since our recordings come from long hours of recording, it is hard to segment by meaningful sentences. Instead, we use Saymore and Audacity to segment by pauses. These factors introduce acoustic and lexical variability that may affect feature prediction accuracy, but fall outside the scope of our current phonological evaluation.

8 Conclusion

This work presents a linguistically-informed evaluation of multilingual ASR for African languages, focusing on Yoruba and Uneme. By utilizing Feature Error Rate (FER) and Tone Error Rate (TER), we demonstrate that standard metrics like WER obscure meaningful progress in low-resource settings. Our experiments with mHuBERT-147 show that while the model captures segmental features relatively well (FER $\approx$ 0.15–0.27), it struggles with suprasegmental features like tone and shifts in speaking styles.

For the Uneme language, we provide the first baseline results, establishing that while word-level recognition remains a challenge, phonological reconstruction is feasible. Future work must focus on integrating pitch-aware encoders and expanding text data to bridge the gap between phonetic accuracy and lexical correctness. We conclude that active collaboration between linguists and NLP practitioners is essential to creating dataset and evaluation frameworks that respect the linguistic typology of African languages.

References

Lateef Adeleke. 2025. Uneme iron technology. Endangered Languages Archive. Accessed: 2025-10-05.

Audacity Team. 2025. Audacity: Free, open source, cross-platform audio software. https://www.audacityteam.org/. Version 3.7.4.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. 2024. mhubert-147: A compact multilingual hubert model. *arXiv preprint arXiv:2406.06371*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *Preprint*, arXiv:2205.12446.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, Roy Schwartz, and Yossi Adi. 2023. Textually pretrained speech language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *Preprint*, arXiv:2106.07447.

Sukairaj Hafiz Imam, Tadesse Destaw Belay, Kedir Yassin Husse, Ibrahim Said Ahmad, Idris Abdulmumin, Hadiza Ali Umar, Muhammad Yahuza Bello, Joyce Nakatumba-Nabende, Seid Muhie Yimam, and Shamsuddeen Hassan Muhammad. 2025a. Automatic speech recognition (asr) for african low-resource languages: A systematic literature review. *Preprint*, arXiv:2510.01145.

Sukairaj Hafiz Imam, Babangida Sani, Dawit Ketema Gete, Bedru Yimam Ahmed, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Seid Muhie Yimam, Muhammad Yahuza Bello, and Shamsuddeen Hassan Muhammad. 2025b. Automatic speech recognition for african low-resource languages: Challenges and future directions. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 89–94.

Thennal D K, Jesin James, Deepa P Gopinath, and Muhammed Ashraf K. 2024. Advocating character error rate for multilingual asr evaluation. *Preprint*, arXiv:2410.07400.

Thennal D K, Jesin James, Deepa Padmini Gopinath, and Muhammed Ashraf K. 2025. Advocating character error rate for multilingual ASR evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4926–4935, Albuquerque, New Mexico. Association for Computational Linguistics.

David R Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. Panphon: A resource for mapping ipa segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484.

Saul B Needleman and Christian D Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b Model Card. *Preprint*, arXiv:2508.10925.

Noam Shazeer. 2020. Glu variants improve transformer. *Preprint*, arXiv:2002.05202.

SIL International. 2025. Saymore: A tool for managing and analyzing language documentation data. https://software.sil.org/saymore/. Version 3.7.2.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Biao Zhang and Rico Sennrich. 2019. *Root mean square layer normalization*. Curran Associates Inc., Red Hook, NY, USA.

Chengrui Zhu, Keyu An, Huahuan Zheng, and Zhijian Ou. 2021. Multilingual and crosslingual speech recognition using phonological-vector based phone embeddings. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1034–1041. IEEE.

A   Experimental Setup

We provide the detailed hyperparameters and configurations used in our experiments in Table 7.

B   Feature Inventory

Table 8 presents the comprehensive inventory of the 24 articulatory phonetic features used for FER evaluation. These features are categorized into major classes, laryngeal settings, manner and place of articulation, vowel properties, and suprasegmental features.

Each feature can take a ternary value from $\{-1, 0, +1\}$, representing the absence, undefined status, or presence of a specific articulatory property. Note that certain features (e.g., TONE_M, DOWNSTEP, V_ATR) are language-specific and are active only for Yoruba or Uneme, while remaining zero-padded for English.

Table 7: **Hyperparameters and Model Configurations.** All encoders were frozen during training.

Category	Parameter	Value
Optimization	Optimizer	AdamW
	Learning rate	5×10^{-5}
	Weight decay	10^{-4}
	Gradient clipping	3.0
	Precision	16-bit mixed
Scheduling	Max training steps	20,000
	Warmup steps	1,000
	Batch size (per GPU)	16
	Gradient accumulation	4
	Effective batch size	64
Architecture	Transformer Decoder Layers	2
	Attention Heads	1
	Decoder Dimension	1,024
	Dropout	0.2
Encoder Dimensions	mHuBERT-25-Hz	768
	mHuBERT-147	1,024
	wav2vec2-XLSR-53	1,024
SpecAugment	Mask time probability	0.05
	Mask time length	10
	Mask feature probability	0.01
	Mask feature length	64

C Feature Maps

C.1 Uneme

Table 9 provides feature maps for Uneme. The segments are based on the orthography of the language, but most symbols are IPA-like, and those that are not still represent unique sounds.

C.2 Yoruba

Table 10 provides feature maps for Yoruba. The segments are based on the orthography of the language, but most symbols are IPA-like, and those that are not still represent unique sounds.

C.3 English

Table 11 provides feature maps for English. The segments are base on ARPABET format.

Table 8: **Complete Inventory of Articulatory Features.** The model utilizes a total of 24 features. The *Scope* column indicates whether a feature is universally applied or specific to certain languages in our dataset (English, Yoruba, Uneme).

Category	Feature Set (Names)	Scope / Notes
Major Class	Syllabic, Consonantal Sonorant, Approximant	Universal Universal
Laryngeal	Voice, Aspirated	Universal
Place of Articulation	Labial, Coronal, Dorsal, Labial_Velar	Universal (Lb_Vel for /kp/, /gb/)
Manner of Articulation	Stop, Nasal, Fricative, Lateral	Universal
Vowel Features	*Height:* High, Mid, Low *Quality:* Back, Round *Secondary:* ATR, Nasalized	Universal Universal Primarily Yoruba & Uneme
Suprasegmental	Tone_H, Tone_L Tone_M, Downstep	Yoruba & Uneme Tone_M (Yoruba), Downstep (Uneme)

Table 9: Uneme Phonetic Feature Matrix (24-vector) using Abbreviations

Segment	SYL	CNS	SON	APR	STP	NAS	FRI	LAT	LAB	COR	DOR	LBV	VOI	ASP	VHH	VHM	VHL	VBK	VRD	ATR	VNS	TH	TL	DWS
a	+	−	+	o	o	o	o	o	o	o	o	o	+	o	−	−	+	o	−	−	−	o	o	o
b	−	+	−	−	+	−	−	−	+	−	−	−	+	o	o	o	o	o	o	o	o	o	o	o
ch	−	+	−	−	+	−	+	−	−	+	−	−	−	o	o	o	o	o	o	o	o	o	o	o
d	−	+	−	−	+	−	−	−	−	+	−	−	+	o	o	o	o	o	o	o	o	o	o	o
e	+	−	+	o	o	o	o	o	o	o	o	o	+	o	−	+	−	−	−	+	−	o	o	o
f	−	+	−	−	−	−	+	−	+	−	−	−	−	o	o	o	o	o	o	o	o	o	o	o
g	−	+	−	−	+	−	−	−	−	−	+	−	+	o	o	o	o	o	o	o	o	o	o	o
gb	−	+	−	−	+	−	−	−	+	−	+	+	+	o	o	o	o	o	o	o	o	o	o	o
gh	−	+	−	−	−	−	+	−	−	−	+	−	+	o	o	o	o	o	o	o	o	o	o	o
h	−	+	−	−	−	−	+	−	−	−	−	−	−	o	o	o	o	o	o	o	o	o	o	o
i	+	−	+	o	o	o	o	o	o	o	o	o	+	o	+	−	−	−	−	+	−	o	o	o
k	−	+	−	−	+	−	−	−	−	−	+	−	−	o	o	o	o	o	o	o	o	o	o	o
kh	−	+	−	−	+	−	+	−	−	−	+	−	−	o	o	o	o	o	o	o	o	o	o	o
kp	−	+	−	−	+	−	−	−	+	−	+	+	−	o	o	o	o	o	o	o	o	o	o	o
l	−	+	+	+	−	−	−	+	−	+	−	−	+	o	o	o	o	o	o	o	o	o	o	o
m	−	+	+	−	−	+	−	−	+	−	−	−	+	o	o	o	o	o	o	o	o	o	o	o
n	−	+	+	−	−	+	−	−	−	+	−	−	+	o	o	o	o	o	o	o	o	o	o	o
o	+	−	+	o	o	o	o	o	o	o	o	o	+	o	−	+	−	+	+	+	−	o	o	o
p	−	+	−	−	+	−	−	−	+	−	−	−	−	o	o	o	o	o	o	o	o	o	o	o
r	−	+	+	+	−	−	−	−	−	+	−	−	+	o	o	o	o	o	o	o	o	o	o	o
rh	−	+	+	+	−	−	+	−	−	+	−	−	−	o	o	o	o	o	o	o	o	o	o	o
s	−	+	−	−	−	−	+	−	−	+	−	−	−	o	o	o	o	o	o	o	o	o	o	o
sh	−	+	−	−	−	−	+	−	−	+	−	−	−	o	o	o	o	o	o	o	o	o	o	o
t	−	+	−	−	+	−	−	−	−	+	−	−	−	o	o	o	o	o	o	o	o	o	o	o
u	+	−	+	o	o	o	o	o	o	o	o	o	+	o	+	−	−	+	+	+	−	o	o	o
v	−	+	−	−	−	−	+	−	+	−	−	−	+	o	o	o	o	o	o	o	o	o	o	o
vb	−	+	−	−	+	−	−	−	+	−	−	−	+	o	o	o	o	o	o	o	o	o	o	o
w	−	+	+	+	−	−	−	−	+	−	+	+	+	o	o	o	o	o	o	o	o	o	o	o
y	−	+	+	+	−	−	−	−	−	−	+	−	+	o	o	o	o	o	o	o	o	o	o	o
z	−	+	−	−	−	−	+	−	−	+	−	−	+	o	o	o	o	o	o	o	o	o	o	o
e.	+	−	+	o	o	o	o	o	o	o	o	o	+	o	−	+	−	−	−	−	−	o	o	o
o.	+	−	+	o	o	o	o	o	o	o	o	o	+	o	−	+	−	+	+	−	−	o	o	o

160

Table 10: Yoruba Phonetic Feature Matrix (24-vector) using Abbreviations

Segment	SYL	CNS	SON	APR	STP	NAS	FRI	LAT	LAB	COR	DOR	LBV	VOI	ASP	VHH	VHM	VHL	VBK	VRD	ATR	VNS	TH	TL	TM
a	+	−	+	o	o	o	o	o	o	o	o	o	+	o	−	−	+	o	−	−	−	o	o	o
b	−	+	−	−	+	−	−	−	+	−	−	−	+	−	o	o	o	o	o	o	o	o	o	o
d	−	+	−	−	+	−	−	−	−	+	−	−	+	−	o	o	o	o	o	o	o	o	o	o
e	+	−	+	o	o	o	o	o	o	o	o	o	+	o	−	+	−	−	−	+	−	o	o	o
f	−	+	−	−	−	−	+	−	+	−	−	−	−	−	o	o	o	o	o	o	o	o	o	o
g	−	+	−	−	+	−	−	−	−	−	+	−	+	−	o	o	o	o	o	o	o	o	o	o
gb	−	+	−	−	+	−	−	−	+	−	+	+	+	−	o	o	o	o	o	o	o	o	o	o
h	−	+	−	−	−	−	+	−	−	−	−	−	−	o	o	o	o	o	o	o	o	o	o	o
i	+	−	+	o	o	o	o	o	o	o	o	o	+	−	+	−	−	−	−	+	−	o	o	o
k	−	+	−	−	+	−	−	−	−	−	+	−	−	−	o	o	o	o	o	o	o	o	o	o
kp	−	+	−	−	+	−	−	−	+	−	+	+	−	−	o	o	o	o	o	o	o	o	o	o
l	−	+	+	+	−	−	−	+	−	+	−	−	+	−	o	o	o	o	o	o	o	o	o	o
m	−	+	+	−	−	+	−	−	+	−	−	−	+	−	o	o	o	o	o	o	o	o	o	o
n	−	+	+	−	−	+	−	−	−	+	−	−	+	−	o	o	o	o	o	o	o	o	o	o
o	+	−	+	o	o	o	o	o	o	o	o	o	+	o	−	−	−	+	+	+	−	o	o	o
p	−	+	−	−	+	−	−	−	+	−	−	−	−	−	o	o	o	o	o	o	o	o	o	o
r	−	+	+	+	−	−	−	−	−	+	−	−	+	−	o	o	o	o	o	o	o	o	o	o
s	−	+	−	−	−	−	+	−	−	+	−	−	−	−	o	o	o	o	o	o	o	o	o	o
t	−	+	−	−	+	−	−	−	−	+	−	−	−	−	o	o	o	o	o	o	o	o	o	o
u	+	−	+	o	o	o	o	o	o	o	o	o	+	o	+	−	−	+	+	+	−	o	o	o
w	−	−	+	+	−	−	−	−	+	−	+	+	+	−	o	o	o	o	o	o	o	o	o	o
y	−	−	+	+	−	−	−	−	−	−	+	−	+	−	o	o	o	o	o	o	o	o	o	o
ṣ	−	+	−	−	−	−	+	−	−	+	−	−	−	−	o	o	o	o	o	o	o	o	o	o
ẹ	+	−	+	o	o	o	o	o	o	o	o	o	+	o	−	+	−	−	−	−	−	o	o	o
ọ	+	−	+	o	o	o	o	o	o	o	o	o	+	o	−	+	−	+	+	−	−	o	o	o

161

Table 11: English (ARPABET) Phonetic Feature Matrix (24-vector) using Abbreviations

Segment	SYL	CNS	SON	APR	STP	NAS	FRI	LAT	LAB	COR	DOR	VOI	ASP	VHH	VHM	VHL	VBK	VRD	VTN	RHO	DIP	N/A	N/A	N/A
AA	+	−	+	0	0	0	0	0	0	0	0	+	0	−	−	+	+	−	+	0	0	0	0	0
AE	+	−	+	0	0	0	0	0	0	0	0	+	0	−	+	+	−	−	−	0	0	0	0	0
AH	+	−	+	0	0	0	0	0	0	0	0	+	0	−	+	−	−	−	−	0	0	0	0	0
AH0	+	−	+	0	0	0	0	0	0	0	0	+	0	−	+	−	+	−	−	0	0	0	0	0
AO	+	−	+	0	0	0	0	0	0	0	0	+	0	−	−	+	+	+	+	0	0	0	0	0
AW	+	−	+	0	0	0	0	0	0	0	0	+	0	−	−	+	+	+	+	0	+	0	0	0
AY	+	−	+	0	0	0	0	0	0	0	0	+	0	−	−	+	−	−	+	0	+	0	0	0
B	−	+	−	0	+	0	0	0	+	0	0	+	0	0	0	0	0	0	0	0	0	0	0	0
CH	−	+	−	0	+	0	+	0	0	+	0	−	0	0	0	0	0	0	0	0	0	0	0	0
D	−	+	−	0	+	0	0	0	0	+	0	+	0	0	0	0	0	0	0	0	0	0	0	0
DH	−	+	−	0	0	0	+	0	0	+	0	+	0	0	0	0	0	0	0	0	0	0	0	0
EH	+	−	+	0	0	0	0	0	0	0	0	+	0	−	+	−	−	−	−	0	0	0	0	0
ER	+	−	+	0	0	0	0	0	0	0	0	+	0	−	+	−	−	−	0	+	0	0	0	0
EY	+	−	+	0	0	0	0	0	0	0	0	+	0	−	+	−	−	−	+	0	+	0	0	0
F	−	+	−	0	0	0	+	0	+	0	0	−	0	0	0	0	0	0	0	0	0	0	0	0
G	−	+	−	0	+	0	0	0	0	0	+	+	0	0	0	0	0	0	0	0	0	0	0	0
HH	−	−	+	0	0	0	+	0	0	0	0	−	0	0	0	0	0	0	0	0	0	0	0	0
IH	+	−	+	0	0	0	0	0	0	0	0	+	0	+	−	−	−	−	−	0	0	0	0	0
IY	+	−	+	0	0	0	0	0	0	0	0	+	0	+	−	−	−	−	+	0	0	0	0	0
JH	−	+	−	0	+	0	+	0	0	+	0	+	0	0	0	0	0	0	0	0	0	0	0	0
K	−	+	−	0	+	0	0	0	0	0	+	−	0	0	0	0	0	0	0	0	0	0	0	0
L	−	+	+	+	0	0	0	+	0	+	0	+	0	0	0	0	0	0	0	0	0	0	0	0
M	−	+	+	0	0	+	0	0	+	0	0	+	0	0	0	0	0	0	0	0	0	0	0	0
N	−	+	+	0	0	+	0	0	0	+	0	+	0	0	0	0	0	0	0	0	0	0	0	0
NG	−	+	+	0	0	+	0	0	0	0	+	+	0	0	0	0	0	0	0	0	0	0	0	0
OW	+	−	+	0	0	0	0	0	0	0	0	+	0	−	+	−	+	+	+	0	+	0	0	0
OY	+	−	+	0	0	0	0	0	0	0	0	+	0	−	+	−	+	+	+	0	+	0	0	0
P	−	+	−	0	+	0	0	0	+	0	0	−	0	0	0	0	0	0	0	0	0	0	0	0
R	−	+	+	+	0	0	0	0	0	+	0	+	0	0	0	0	0	0	0	0	0	0	0	0
S	−	+	−	0	0	0	+	0	0	+	0	−	0	0	0	0	0	0	0	0	0	0	0	0
SH	−	+	−	0	0	0	+	0	0	+	0	−	0	0	0	0	0	0	0	0	0	0	0	0
T	−	+	−	0	+	0	0	0	0	+	0	−	0	0	0	0	0	0	0	0	0	0	0	0
TH	−	+	−	0	0	0	+	0	0	+	0	−	0	0	0	0	0	0	0	0	0	0	0	0
UH	+	−	+	0	0	0	0	0	0	0	0	+	0	+	−	−	+	+	−	0	0	0	0	0
UW	+	−	+	0	0	0	0	0	0	0	0	+	0	+	−	−	+	+	+	0	0	0	0	0
V	−	+	−	0	0	0	+	0	+	0	0	+	0	0	0	0	0	0	0	0	0	0	0	0
W	−	+	+	+	0	0	0	0	+	0	+	+	0	0	0	0	0	0	0	0	0	0	0	0
Y	−	+	+	+	0	0	0	0	0	0	+	+	0	0	0	0	0	0	0	0	0	0	0	0
Z	−	+	−	0	0	0	+	0	0	+	0	+	0	0	0	0	0	0	0	0	0	0	0	0
ZH	−	+	−	0	0	0	+	0	0	+	0	+	0	0	0	0	0	0	0	0	0	0	0	0