Beyond the Feature: Comparing Spectral, Cepstral, and Self-Supervised Representations for Sesotho Tone Classification, and What It Means for a Southern African Methodological Commons

Tshepo Mashishi
Institute for Artificial Intelligence Systems,
University of Johannesburg,
South Africa
mashists@gmail.com

Makhamisa Senekane
Institute for Artificial Intelligence Systems,
University of Johannesburg,
South Africa,
makhamisas@uj.ac.za

Mpho Primus
SARChI Chair in Big Data and the Social Sciences
University of Johannesburg
South Africa
mraborife@uj.ac.za

Abstract

Sesotho, a Southern Bantu language spoken by roughly 5.6 million people in South Africa and Lesotho, uses a binary High/Low pitch system to distinguish word meanings and grammatical forms (for example, the person of a verb's subject), a system that most speech and language technologies are not built to recognise. This paper reports a controlled comparison of four feature-extraction paradigms for automated Sesotho tone classification – Fast Fourier Transform (FFT) spectral features, a Librosa acoustic pipeline, Mel-Frequency Cepstral Coefficients (MFCC), and Wav2Vec2 self-supervised embeddings – each evaluated with five classifiers on 4,056 syllable segments from 15 speakers, of which 1,278 carry expert tone labels. Evaluation uses a speaker-disjoint split: 11 speakers for training and four held-out speakers (297 syllables) for testing. A 156-dimensional MFCC feature set paired with a neural network achieves the best held-out result (F1 = 0.9190, ROC-AUC = 0.9583), ahead of a 24-dimensional FFT feature set (F1 = 0.8718, ROC-AUC = 0.9358), Wav2Vec2 (F1 = 0.7966, ROC-AUC = 0.8938), and the Librosa baseline (F1 = 0.6916, ROC-AUC = 0.7740). The ranking holds when classifiers are chosen by cross-validation instead of on the test set. Feature dimensionality alone does not predict performance: the 24-dimensional FFT set outperforms the 776-dimensional Wav2Vec2 set and trails MFCC by 0.047 F1. The feature representation matters more than the classifier: the gap between the best and worst feature sets (0.227 F1) exceeds the spread across classifiers within any feature set ($0.041 - -0.072$). An earlier MFCC extraction lacked features for most syllables; the results reported here use a corrected extraction. Beyond these findings, we situate the work within the Digital Humanities Association of Southern Africa's call for a *methodological commons*, arguing that reproducible, low-resource benchmarking of this kind is a form of infrastructure-building for linguistic justice in the region.

1 Introduction

1.1 Background and Motivation

Sesotho distinguishes High (H) and Low (L) tones at the syllable level, so that identical segmental sequences can carry different meanings depending solely on pitch pattern (Mohasi et al., 2011; Wissing and Roux, 2017). This is not a peripheral phonetic detail: Sesotho uses pitch to convey lexical *and* grammatical meaning (Wissing and Roux, 2017). Grammatical meaning here refers to morphosyntactic information carried by tone rather than by segments: as in many Bantu languages, tone marks categories such as tense and polarity (Nurse, 2008), and in Sesotho the tone on a subject concord, for instance, separates second-person from third-person subjects in otherwise identical questions (*o batlang?* versus *ó batlang?*; Wissing and Roux, 2017). For automatic speech recognition, text-to-speech synthesis, and language-learning technologies, failing to model tone means failing to model meaning. Yet African tone languages remain conspicuously under-served by speech and language technology research relative to their number of speakers (Besacier et al., 2014; Davel and Barnard, 2008; Joshi et al., 2020), a disparity that recent large-scale audits of the NLP field have quantified directly (Joshi et al., 2020).

This asymmetry is also a Digital Humanities (DH) problem, not only an engineering one. Machine learning systems are not neutral instruments: the representations they learn, the corpora they are trained on, and the languages they are built for encode assumptions about which forms of speech and communication count as worth modelling (Birhane et al., 2022; Benjamin, 2019). Where a language is digitally under-resourced, the absence of tone-aware tooling is itself a form of exclusion that computational and archival practice can either entrench

This work is licensed under CC BY SA 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/4.0/deed.en
The copyright remains with the authors.

or begin to redress (Risam, 2019). This paper treats a technical comparison of feature-extraction methods as a contribution to that broader project: establishing, with reproducible rigour, which representations of Sesotho speech actually carry tone-discriminative information, so that future tools do not have to guess.

1.2 Research Aim and Questions

This study systematically compares feature-extraction methods and machine learning models for automated Sesotho binary tone classification, asking: *which feature-extraction methodology provides optimal performance for Sesotho tone prediction, and does that optimum depend on feature dimensionality or model sophistication?* Four paradigms are compared: (1) FFT, a traditional frequency-domain technique capturing harmonic structure; (2) a Librosa acoustic pipeline extracting pitch, energy, and spectral centroid statistics; (3) MFCC, perceptually motivated cepstral coefficients; and (4) Wav2Vec2, self-supervised deep speech embeddings (Baevski et al., 2020). Each representation is evaluated with the same five classifiers, chosen to span distinct modelling paradigms: Logistic Regression (linear), Random Forest (bagged ensemble), a Support Vector Machine with RBF kernel (kernel-based), XGBoost (gradient-boosted trees), and a Multi-Layer Perceptron (neural network). This design separates the effect of the feature representation from the effect of the classifier.

1.3 Contributions

We distinguish three claims. They are related, but they are not interchangeable, and they carry different levels of support.

Primary contribution (technical). A controlled, speaker-disjoint comparison of four representations and five classifiers for Sesotho H/L classification. Its main finding is that a 156-dimensional MFCC representation, after correcting its extraction, gives the best held-out performance at this data scale, with a 24-dimensional spectral representation a close second. The comparison is new for Sesotho: we know of no prior study that benchmarks classical spectral features against pre-trained embeddings for a Southern Bantu language (Section 2). The closest tone-recognition evidence comes from Yoruba, where Wav2Vec2 features improved tone recognition under a different task and metric (Bengono Obiang et al., 2024). Our result is therefore a new benchmark, not a replication.

Secondary contribution (reproducibility). The protocol is specified in enough detail to re-run: the split, hyperparameter grids, seeds, and feature definitions are given, and code and run logs are available on request. This supports comparability.

Framing (Digital Humanities). We argue that benchmarks of this kind contribute to a methodological commons for Southern African languages. The experiments motivate this position but do not establish it.

1.4 Research Propositions

Three propositions, grounded in phonetic and machine-learning theory, guide the analysis:

P1 (Frequency-domain superiority). Frequency-domain methods (FFT, MFCC) will outperform the time-domain-oriented Librosa pipeline, because tonal distinctions are fundamentally frequency phenomena.

P2 (Deep-learning competitiveness). Wav2Vec2, via transfer learning from large-scale pre-training, will produce results competitive with traditional handcrafted features even on a modest labelled set.

P3 (Feature parsimony). Feature dimensionality will not correlate directly with classification performance; a concise representation capturing essential tonal cues may outperform larger feature spaces.

2 Related Work

2.1 Tone in Bantu Languages

Bantu languages exhibit tonal systems ranging from binary to complex multi-level inventories (Doke, 1992; Doke and Cole, 1961). Tone is not purely lexical: grammatical (morphological) tone marks tense, aspect, mood, and polarity in verb conjugation (Nurse, 2008). Sesotho's binary H/L system is comparatively simple within the Bantu family, yet computational modelling remains complicated by tonal spreading and context-dependent variation (Mohasi et al., 2011; Mixdorff et al., 2011; Mohasi, 2015). Prior linguistic work has modelled Sesotho prosody using the Fujisaki model and compared it with HMM-based modelling (Mohasi, 2015; Mohasi et al., 2014); related Bantu languages have been studied for the relationship between tone and pitch in Sepedi (Barnard and Zerbian, 2010) and the effect of tone marking on reading comprehension in Setswana (Pule and Theledi,

2023). The most directly relevant prior computational effort, a rule-based Sesotho tone-labelling algorithm (Raborife et al., 2015), targeted annotation rather than machine-learning classification. No published study, to our knowledge, has systematically compared multiple feature-extraction paradigms – spanning classical signal processing and self-supervised deep learning – for Sesotho tone classification under a controlled evaluation protocol, nor benchmarked classical spectral features against pre-trained embeddings for any Southern Bantu language. This gap motivates the present comparison.

2.2 Feature Extraction for Tone Recognition

Spectral analysis. The Fast Fourier Transform decomposes a time-domain signal $x(n)$ into frequency components via the Discrete Fourier Transform:

$$X(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-j2\pi kn/N} \qquad (1)$$

where $X(k)$ is the frequency-domain representation, N the number of samples, k the frequency-bin index, and j the imaginary unit. The magnitude spectrum $|X(k)|$ reveals the fundamental frequency (F0) and its harmonics, which correlate directly with perceived pitch and therefore tone category (Rabiner and Juang, 1993). Spectral centroid, rolloff, flatness, and harmonic-to-noise ratio derived from FFT analysis have been applied across speaker, emotion, and phoneme recognition tasks and, for tone classification specifically, harmonic energy ratios provide a compact spectral-shape representation (Davis and Mermelstein, 1980). Pitch-extraction algorithms range from autocorrelation and YIN through Probabilistic YIN (PYIN), which uses Hidden Markov Models to track pitch through voicing transitions, to deep-learning approaches such as CREPE (Rabiner and Juang, 1993).

Cepstral analysis. MFCCs remain a de facto standard in speech and speaker recognition owing to their compact, perceptually motivated encoding of the spectral envelope (Davis and Mermelstein, 1980). The mel-scale transformation from frequency f (Hz) to mel value m is

$$m = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \qquad (2)$$

Extraction proceeds through pre-emphasis, framing and windowing, FFT power-spectrum computation, mel-filterbank application, logarithmic compression, and a discrete cosine transform to decorrelate coefficients (Davis and Mermelstein, 1980). Delta and delta-delta coefficients add temporal dynamics relevant to pitch contours. MFCCs are among the feature extractors that have been compared for tone recognition in low-resourced African tone languages such as Yoruba (Bengono Obiang et al., 2024).

Self-supervised representations. Wav2Vec2 learns hierarchical speech representations from raw waveforms without linguistic annotation, using a convolutional feature encoder, a Transformer context network, and a contrastive quantisation objective during pre-training (Baevski et al., 2020). Because it requires only limited labelled data downstream, self-supervised pre-training is attractive for under-resourced languages (Baevski et al., 2020), and cross-lingual variants (XLSR, XLS-R) extend this to unseen languages (Conneau et al., 2021). For African tone languages specifically, Wav2Vec2-based representations have improved Yoruba tone recognition despite the base model's pre-training on predominantly non-tonal speech (Bengono Obiang et al., 2024).

2.3 Machine Learning for Tone Classification, and a Note on the Methodological Commons

Supervised classifiers – SVMs, ensemble trees, and neural networks – have been applied to tone recognition across languages, with hyperparameter optimisation via cross-validation particularly important where labelled data is scarce (Cortes and Vapnik, 1995; Breiman, 2001). What is comparatively rare, in African-language NLP as in DH more broadly, is *cross-method* benchmarking under a fixed protocol: most published tone-recognition work reports a single pipeline rather than a controlled comparison. This matters for the DH conception of a methodological commons: openly articulated, shared methods and evaluation frameworks that a community of researchers can build on, rather than isolated, non-comparable pipelines (Risam, 2019). A systematic, reproducible comparison across feature paradigms – of the kind reported here – is one concrete way computational and linguistic scholarship on Southern African languages can converge on shared infrastructure rather than duplicating incompatible one-off systems, in the spirit of calls for participatory, accountable AI research practice

(Birhane et al., 2022).

3 Methodology

3.1 Experimental Workflow

The pipeline has seven stages: (1) syllable segmentation from Praat TextGrid annotations (Boersma and Weenink, 2025); (2) audio loading at 16 kHz mono (McFee et al., 2015); (3) speaker-gender assignment from pitch; (4) feature extraction per method; (5) a speaker-disjoint train/test split of the labelled syllables (Section 3.2), with feature standardisation fitted only on training data; (6) hyperparameter tuning by speaker-grouped 5-fold cross-validation; and (7) refitting the selected configuration on the training speakers, evaluating it on the held-out speakers, and predicting tones for the unlabelled syllables. All five classifiers received identical feature matrices, splits, and cross-validation folds within each feature-extraction method, so that observed differences are attributable to the representation or the classifier rather than to sampling.

3.2 Dataset

Speakers and recordings. The corpus comprises recordings from 15 heritage Sesotho speakers, aged 18–28 and fluent in more than one language, across three South African regions (Free State, Soweto, and Vaal), totalling 540 audio files: 29–31 files for each of 13 speakers, 60 for speaker TL, and 90 for speaker NM (Appendix A, Table 3). Speaker gender was assigned automatically from mean pitch (male: 85–180 Hz; female: 165–255 Hz), with file-level majority voting. Twelve speakers are predominantly classified as female and three (OR, PL, TL) as male, and 83.2% of syllables are classified as female. These are pitch-based classifications rather than self-reports, and we use them descriptively only.

Sentence design. The sentence material was designed to cover three tonal phenomena (Raborife, 2022): high-tone spread within the clitic phrase, using polysyllabic verb stems in present and future tenses; lexical and grammatical minimal pairs, such as third- versus second-person subject markers and possessive versus noun; and grammatical tone insertion in imperative, negative, perfect, and subjunctive forms. Monosyllabic verb stems were excluded because their tonal behaviour is complex (Khoali, 1991). The 15-speaker corpus uses this sentence set. Speech is read rather than spontaneous: participants read sentences displayed on a screen, each with its English meaning.

Syllables and labels. Syllable boundaries were identified from Praat TextGrid annotations, yielding 5,391 boundaries. After excluding segments with missing or inaccessible audio, 4,056 syllables were retained. Of these, 1,278 carry expert tone labels (589 High, 689 Low) and 2,778 are unlabelled. Three labellers independently transcribed the sentences in orthography with tone labels and each submitted a brief account of their method. Disagreements were resolved by majority agreement. Agreement among the three labellers was moderate (free-marginal $\kappa = 0.41$). The unlabelled syllables were not verified. They receive predicted tones from the best classifier (Section 3.5) and are used only in the distribution analysis (Section 5.3).

Speaker-disjoint split. The labelled syllables were partitioned with `StratifiedGroupKFold` (five folds, shuffled, `random_state`=42), where each group is one recording unit (16 units in the labelled set, each belonging to a single speaker), and the first fold was taken as the test set. Its four speakers (KN, MM, NN, VS) contribute 297 syllables (137 High, 160 Low; 23% of the labelled set). The remaining 11 speakers provide 981 training syllables (452 High, 529 Low). No speaker appears in both partitions. A syllable-level random split would allow the same speaker's syllables in both partitions, letting speaker-specific acoustics inflate scores, and the speaker-disjoint split is designed to prevent this. Per-speaker counts are in Appendix A. The High-tone share ranges from 42% to 49% across speakers, so no speaker dominates the tone balance.

3.3 Feature Extraction Protocols

Table 5 in Appendix C summarises the four protocols. Each tests a different hypothesis about where H/L information lies: harmonic and F0 structure (FFT), pitch and energy statistics alone (Librosa), a perceptually motivated spectral envelope with temporal dynamics (MFCC), and a learned, phonetically unspecific representation (Wav2Vec2). Dimensionality is stated here rather than in the table. The FFT vector has 24 dimensions: 16 FFT-specific spectral features plus eight features shared by all protocols. The Librosa vector is those eight shared features. The MFCC vector has 156 dimensions,

and the Wav2Vec2 vector has 776 (768 embedding dimensions plus the eight shared features).

MFCC window correction. The MFCC delta and delta-delta statistics are computed with a nine-frame window by default, which requires a segment of at least 0.256 s at the 512-sample hop used here. Shorter syllables cannot use that window. An earlier extraction run failed on these syllables and left their 144 MFCC-derived statistics missing for 3,226 of 4,056 syllables (79.5%), with missing values replaced by zero. The reported results use a corrected extraction in which the window is the largest odd length that fits the segment (at most nine frames, at least three). Two syllables are shorter than three frames and have zero delta features; all other syllables have complete features. The earlier results are not reported.

3.4 Data Preprocessing

Audio was loaded at 16 kHz mono using Librosa (McFee et al., 2015). Feature standardisation (`StandardScaler`) is part of each classifier's pipeline. It is fitted only on the training folds and then applied to the held-out speakers, which avoids data leakage (Pedregosa et al., 2011). Missing MFCC values are zero-filled (Section 3.3).

3.5 Machine Learning Models

Five classifiers, each representing a distinct modelling paradigm, were evaluated per feature-extraction method: Logistic Regression (linear baseline) (Hosmer et al., 2013), Random Forest (bagged ensemble) (Breiman, 2001), Support Vector Machine with an RBF kernel (max-margin, kernelised) (Cortes and Vapnik, 1995; Schölkopf and Smola, 2002), XGBoost (gradient-boosted trees) (Chen and Guestrin, 2016), and a Multi-Layer Perceptron (feed-forward neural network) (Hornik et al., 1989). All classifiers used `scikit-learn` (Pedregosa et al., 2011), except XGBoost, which used the `xgboost` package (Chen and Guestrin, 2016). Seed 42 is passed as `random_state` to every stochastic component (Section 3.6).

Logistic Regression models the log-odds of the High-tone class linearly, with L2 regularisation:

$$P(y{=}1 \mid x) = \frac{1}{1+e^{-z}}, \quad z = \beta_0 + \sum_i \beta_i x_i \tag{3}$$

The SVM's RBF kernel is

$$K(x, x') = \exp\!\left(-\gamma \|x - x'\|^2\right) \tag{4}$$

and XGBoost's sequential boosting update is

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{5}$$

where f_t is chosen at iteration t to minimise $L(y, \hat{y}^{(t-1)}) + \Omega(f_t)$, with Ω a complexity-penalising regularisation term (Chen and Guestrin, 2016).

The MLP was trained with the Adam optimiser for up to 500 iterations (`max_iter=500`) without early stopping; overfitting given the modest training size (981 training syllables) was controlled through the L2 penalty α and the architecture grid in Appendix B. SVM probability estimates were calibrated via Platt scaling to support ROC-AUC computation.

3.6 Hyperparameter Optimisation and Reproducibility

Hyperparameters were tuned by exhaustive `GridSearchCV` on the 981 training syllables. The inner cross-validation used `StratifiedGroupKFold` (five folds, shuffled, `random_state`=42), which keeps each recording unit, and therefore each speaker, within a single fold, and weighted F1 was the scoring function. Each method's grid contains 242 configurations, giving 1,210 cross-validation fits per method and 4,840 across the four methods (excluding final refits). The grids and the selected configurations are listed in Appendix B. The selected configuration was refitted on all training syllables and evaluated once on the held-out speakers.

Seed 42 is passed as `random_state` to the splitter and to every stochastic estimator (Random Forest; MLP initialisation and mini-batch shuffling; SVM probability estimation; XGBoost). For the scikit-learn components, an integer `random_state` seeds a NumPy `RandomState`, i.e. the Mersenne Twister (MT19937) generator; XGBoost uses its own internal generator seeded with the same value. NumPy's global generator, Python's `random` module, and PyTorch were not seeded. Wav2Vec2 embeddings were extracted with the pre-trained model in evaluation mode (`model.eval()`, dropout disabled) under `torch.no_grad()`, so extraction involves no stochastic layers. The FFT, Librosa (PYIN), and MFCC pipelines are deterministic. Results should reproduce under the same data and library versions. Python 3.13.18 was used. Library versions: scikit-learn 1.6.1, XGBoost 3.4.1,

Librosa 0.11.0, and PyTorch 2.11.0 (CPU build). The notebooks install their dependencies without pinning, so we release a `requirements.txt` recording the environment (see Data and Code Availability).

3.7 Evaluation Metrics

Accuracy, weighted precision, weighted recall, weighted F1-score, and ROC-AUC (computed on the H-class probability) are reported on the held-out speakers. Each method's "best" classifier is the one with the highest held-out F1, which is the criterion used throughout. Because that criterion selects on the test set, we also report the classifier selected by cross-validation; it yields the same method ranking (Section 5.2). With $n = 297$, the approximate 95% Wald interval for accuracy is ± 3.1 percentage points for MFCC, ± 3.8 for FFT and ± 5.2 for Librosa. These intervals treat syllables as independent and understate uncertainty when only four speakers are tested.

4 Results

4.1 Cross-Method Performance Comparison

Table 1 reports the best classifier per method on the held-out speakers, with the cross-validation F1 of that same classifier. MFCC with a neural network is strongest on every held-out metric, using 156 features against 24 (FFT), 776 (Wav2Vec2), and 8 (Librosa). The ROC-AUC ordering matches the F1 ordering (MFCC > FFT > Wav2Vec2 > Librosa), so the ranking does not depend on the metric. Because the neural network was chosen on the test set, the cross-validation-selected classifier is reported in Section 5.2 as well. We also ran exact paired McNemar tests on the 297 held-out syllables. MFCC differs significantly from FFT ($p = 0.034$; 26 vs. 12 discordant syllables), Wav2Vec2 ($p < 0.001$; 49 vs. 14), and Librosa ($p < 0.001$; 72 vs. 8). FFT differs significantly from Wav2Vec2 ($p = 0.017$; 46 vs. 25) and from Librosa ($p < 0.001$; 63 vs. 13). These tests treat syllables as independent, which is optimistic when only four speakers are tested. The MFCC–FFT difference is the smallest of these and rests on a single held-out split.

4.2 Model Consistency and Classifier Effects

Appendix D (Table 6) reports F1 and ROC-AUC for all 20 method–classifier pairs. Neural networks lead for FFT and MFCC, and RBF-kernel SVMs for Wav2Vec2 and Librosa. For Wav2Vec2 the linear model (LR, F1 0.7833) outperforms the neural network (0.7732), suggesting that the pretrained embeddings are fairly linearly separable. Within each feature set, changing the classifier moves F1 by only 0.041–0.072. By contrast, the best-classifier F1 differs by 0.227 between MFCC and Librosa, and the weakest FFT classifier (RF, 0.8112) still exceeds the best Wav2Vec2 classifier (0.7966). The feature representation therefore explains more of the variation in performance than the classifier does.

Method	Model	Acc.	F1	ROC-AUC	CV F1
MFCC	NN	.9192	**.9190**	.9583	.8749
FFT	NN	.8721	**.8718**	.9358	.8549
Wav2Vec2	SVM	.8013	**.7966**	.8938	.7681
Librosa	SVM	.7037	**.6916**	.7740	.7272

Table 1: Best classifier per method, selected by held-out F1 (bold), on the 297-syllable test set. CV F1 is the weighted speaker-grouped 5-fold cross-validation F1 of the same classifier. NN = neural network (MLP); XGB = XGBoost; SVM = RBF-kernel support vector machine.

Robustness to classifier selection. If each method's classifier is instead chosen by cross-validation, XGBoost is selected for all four feature sets (CV F1 0.8564, 0.9027, 0.8071, and 0.7554 for FFT, MFCC, Wav2Vec2, and Librosa). Its held-out F1 is 0.8511 (FFT), 0.8890 (MFCC), 0.7865 (Wav2Vec2), and 0.6863 (Librosa), with ROC-AUC 0.9219, 0.9582, 0.8901, and 0.7776. The method ranking is the same under both rules (MFCC, FFT, Wav2Vec2, Librosa), so the MFCC lead does not depend on selecting the neural network on the test set.

4.3 Tone Distribution Analysis

Table 2 reports the High-tone share of the expert-labelled syllables and of the 2,778 unlabelled syllables, where the tone is the best classifier's prediction. On the unlabelled syllables, the four classifiers predict 32–36% High, against 46% High among the expert-labelled syllables. Because the unlabelled syllables have no reference, these figures cannot show which method is closer to the truth, and the comparison is not a measure of accuracy. A difference in the composition of labelled and unlabelled syllables would produce the same pattern, and our data do not separate that explanation from a systematic under-prediction of High tones. Connected-speech context effects (Mixdorff

Source	All (%)	Unlabelled (%)
Expert labels (n = 1,278)	46.09	–
FFT (NN)	36.42	31.97
MFCC (NN)	38.66	35.24
Wav2Vec2 (SVM)	39.35	36.25
Librosa (SVM)	38.78	35.42

Table 2: High-tone share (%) of all 4,056 syllables (expert labels where available, otherwise the best classifier's prediction) and of the 2,778 unlabelled syllables, which are predictions only.

et al., 2011) could also contribute, but we do not test them.

4.4 Computational Considerations

Qualitatively, Librosa is the fastest and lightest method, and FFT is likewise fast with minimal overhead. MFCC incurs moderate cost from its multiple feature types, and Wav2Vec2 is by far the most expensive, owing to forward passes through a transformer with roughly 95M parameters (the BASE configuration; Baevski et al., 2020) on a GPU. Weighing performance against cost, FFT occupies a Pareto-optimal position: low computational cost, comparable to Librosa, combined with the best observed performance. Runtime was not profiled systematically.

5 Discussion

5.1 Interpretation of Results

MFCC's lead (F1 0.9190, ROC-AUC 0.9583) is consistent with tone information lying partly in the spectral envelope and its temporal dynamics, which the delta features capture, as well as in F0 and harmonic structure (Rabiner and Juang, 1993). The FFT set, with 24 features, comes within 0.047 F1 of it with far fewer dimensions, which suggests that explicit F0 and harmonic-energy measures capture most of the H/L contrast. Because the MFCC vector also includes the pitch statistics, the lead cannot be attributed to cepstral information alone; an ablation that removes those statistics would be needed to show that, and we did not run one.

Wav2Vec2's result is competitive with the handcrafted sets but not superior: it beats only Librosa and is informative in its own right. Four factors plausibly explain it: (1) *task specificity*, since Wav2Vec2 was optimised for phoneme discrimination rather than suprasegmental tone (Baevski et al., 2020); (2) *language mismatch*, since pretraining on predominantly non-tonal speech may not induce tone-relevant representations in the layers used here; (3) the *dimensionality curse*, since 768 embedding dimensions against 981 training syllables invites overfitting, consistent with the linear model outperforming the neural network on this feature set; and (4) *binary task simplicity*, since the H/L system may be simple enough that explicit F0 extraction suffices. The Librosa baseline's weaker performance shows that pitch statistics alone, without harmonic or spectral-envelope information, lose tone-discriminative information.

5.2 Validation of Research Propositions

P1 is supported: FFT and MFCC exceed the Librosa pipeline by 18.0 and 22.7 F1 points, and both differences are significant under McNemar's test (Section 5.1). **P2** is partially supported: Wav2Vec2 (F1 0.7966) is ahead of Librosa, but it trails FFT by 7.5 points (significant) and MFCC by 12.2 points (significant, $p < 0.001$). Transfer learning does not substitute for task-specific feature engineering at this data scale. **P3** is partially supported: the 24-dimensional FFT set outperforms the 776-dimensional embedding and comes within 0.047 F1 of the 156-dimensional MFCC set, but the larger MFCC set is ahead. Feature parsimony helps relative to the largest representation, not relative to the best one.

5.3 Implications for Tone-Recognition Systems and the Methodological Commons

For resource-constrained deployments, which are likely to be the majority of practical Sesotho language-technology contexts, FFT-based features with a neural-network classifier give the strongest accuracy per feature and require no external pretrained model, while MFCC features with a neural network give the highest accuracy at moderate extraction cost. Wav2Vec2 required a GPU for feature extraction and was not more accurate here. We did not test low-data regimes, so the present evidence does not support recommending Wav2Vec2 for settings with very little labelled data. MFCC's behaviour under varied recording conditions was not tested,. The Librosa pipeline remains usable for low-latency applications at lower accuracy.

More broadly, this comparison is offered as a candidate building block for the methodological commons. It is a documented, seed-fixed, speaker-disjoint pipeline that other researchers working on Sesotho, or on related Southern Bantu tone systems

such as Sepedi, Setswana, Zulu, or Xhosa, can reuse, extend, or challenge with new data, rather than starting each effort from an unpublished, non-comparable pipeline. Any sharing of the corpus, code, and hyperparameter logs is subject to the speaker-consent and access-control requirements that any Sesotho speech corpus must observe.

5.4 Comparison with Related Work

Direct comparison is limited by the scarcity of computational tone-recognition research specific to Sesotho, but the results are broadly consistent with related work. Work on Yoruba, a three-tone language, has shown that Wav2Vec2-based representations can improve tone recognition in a low-resourced language, reporting a tone error rate (TER) of 17.72% with Wav2Vec2 against 19.54% for the best combination of conventional features (Bengono Obiang et al., 2024). That study uses a different language, tone inventory, task formulation (continuous speech rather than isolated syllables), and metric (TER rather than F1), so its figures cannot be compared directly with ours. The shared finding is that pre-trained self-supervised representations are competitive for tone recognition in a low-resourced language. Our results add that, at our data scale and for a binary tone system, explicit spectral features can match or exceed them. Work on automated processing of other under-resourced African languages, such as Igbo (Onyenwe, 2017), and recent multilingual generation systems spanning hundreds of African languages (Adebara et al., 2024) point to growing momentum in African-language technology, even as tone recognition remains a comparatively specialised sub-problem within it.

6 Conclusion and Future Work

MFCC features with a neural-network classifier achieve the best performance for Sesotho binary tone classification under a speaker-disjoint evaluation (F1 0.9190, ROC-AUC 0.9583), ahead of a 24-dimensional FFT representation (F1 0.8718), Wav2Vec2 (F1 0.7966), and a Librosa pitch-and-energy pipeline (F1 0.6916). The ranking holds when classifiers are chosen by cross-validation. The central empirical finding is that feature dimensionality alone does not predict performance: a compact, phonetically motivated spectral set comes within 0.047 F1 of the best representation and outperforms a 776-dimensional self-supervised embedding. We recommend MFCC with a neural network where accuracy matters most, and FFT features where the feature set must stay small. Testing whether the MFCC lead survives removing its pitch statistics, and evaluating on additional speakers, are the immediate next steps.

Future work should explore hybrid feature combinations (for example FFT harmonics plus Wav2Vec2 embeddings), contextualised models that incorporate syllable- and word-level context, end-to-end fine-tuning of self-supervised models on Sesotho tone data, and temporal architectures that model pitch-contour dynamics beyond static features. It should also expand the labelled corpus with more speakers and use leave-speaker-out evaluation, test robustness under noise, and extend the methodology to Zulu, Xhosa, Sepedi, and Setswana. Within a Southern African methodological commons, we see the more durable contribution of this study as the reproducible, speaker-disjoint evaluation protocol itself: a shared point of comparison against which future Sesotho and Southern Bantu tone-recognition research can be measured.

Data and Code Availability

A methodological commons depends on shared artefacts, so the following are available from the authors on request: (i) the feature-extraction scripts for all four pipelines and the training and evaluation code for all five classifiers, including the speaker-disjoint split; (ii) a `requirements.txt` recording the library versions used; (iii) the hyperparameter grids, best-parameter records, and cross-validation logs for all 20 method–classifier combinations; and (iv) the derived feature matrices and tone labels, which contain speaker and file identifiers and are available on request, subject to the conditions in the Ethical Considerations section.

Limitations

The held-out set is small: four speakers and 297 syllables. The intervals and McNemar tests treat syllables as independent and so understate speaker-level uncertainty, which makes the FFT–MFCC difference (McNemar $p = 0.034$) suggestive rather than established. The "best" classifier is chosen on held-out F1, which is optimistic, although the cross-validation-selected XGBoost gives the same ranking.

The labelled set is modest (1,278 syllables) and uneven across speakers: NM and TL contribute 152

and 146 labelled syllables, and the others 71–76. Regional coverage is not reported per speaker. We address only binary H/L classification of isolated syllables, so tone sandhi, downstep, and coarticulation in connected speech were not evaluated. The 2,778 unlabelled syllables were not validated, and gender labels are pitch-based rather than self-reported. The notebooks install unpinned dependencies, so exact reproduction also depends on the released requirements file.

Ethical Considerations

The study reports speech data from 15 speakers. The study comprises 15 heritage Sesotho speakers aged 18–28 who are fluent in more than one language. Participants read sentences displayed on a screen, after reading five sample sentences to set microphone placement. Before recording, they were told the study's purpose, assured of confidentiality through pseudonyms, took part voluntarily, and signed a formal consent form and completed a language profile questionnaire. No formal ethics clearance reference number was issued for this study. Redistribution of recordings and derived data is limited to what the consent terms permit, and the feature matrices and tone labels are released on the same basis.

References

I. Adebara, A. Elmadany, and M. Abdul-Mageed. 2024. Cheetah: Natural language generation for 517 African languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 12798–12823.

A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 12449–12460.

Etienne Barnard and Sabine Zerbian. 2010. From tone to pitch in Sepedi. In *Proceedings of the 2nd International Workshop on Spoken Languages Technologies for Under-Resourced Languages (SLTU'10)*, pages 29–34, Penang, Malaysia.

S. G. B. Bengono Obiang, N. Tsopzé, P. M. Yonta, J.-F. Bonastre, and T. Jiménez. 2024. Improving tone recognition performance using Wav2vec 2.0-based learned representation in Yoruba, a low-resourced language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(12):172:1–172:11.

Ruha Benjamin. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*. Polity Press, Cambridge, UK.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 173–184.

P. Boersma and D. Weenink. 2025. Praat: Doing phonetics by computer. `http://www.praat.org/`. Accessed: 16 October 2025.

L. Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

T. Chen and C. Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 785–794.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Proceedings of INTERSPEECH*, pages 2426–2430.

C. Cortes and V. Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Marelie Davel and Etienne Barnard. 2008. Pronunciation prediction with Default&Refine. *Computer Speech & Language*, 22(4):374–393.

S. Davis and P. Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

C. M. Doke and D. T. Cole. 1961. *Contributions to the History of Bantu Linguistics*. Witwatersrand University Press, Johannesburg.

M. Doke. 1992. *Textbook of Zulu Grammar*, 6th edition. Maskew Miller Longman, Cape Town.

K. Hornik, M. Stinchcombe, and H. White. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366.

D. W. Hosmer, Jr., S. Lemeshow, and R. X. Sturdivant. 2013. *Applied Logistic Regression*, 3rd edition. Wiley, Hoboken, NJ, USA.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online.

B. Khoali. 1991. *A Sesotho Tonal Grammar*. Phd thesis, University of Illinois.

B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto. 2015. librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference (SciPy)*, pages 18–24.

Hansjörg Mixdorff, Lehlohonolo Mohasi, M. Machobane, and Thomas Niesler. 2011. A study on the perception of tone and intonation in Sesotho. In *Proceedings of Interspeech 2011*, pages 3181–3184.

Lehlohonolo Mohasi. 2015. *Prosody Modelling for a Sesotho Text-to-Speech System Using the Fujisaki Model*. Ph.D. thesis, Stellenbosch University. PhD thesis, Stellenbosch University. Handle: 10019.1/97050.

Lehlohonolo Mohasi, Hansjörg Mixdorff, M. Machobane, and Thomas Niesler. 2011. An acoustic analysis of tone in Sesotho. In *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII)*, pages 1402–1405.

Lehlohonolo Mohasi, Milan Sečujski, Robert Mak, and Thomas Niesler. 2014. A comparison of two prosody modelling approaches for Sesotho and Serbian. In *Speech and Computer: 16th International Conference, SPECOM 2014, Novi Sad, Serbia, October 5–9, 2014, Proceedings*, volume 8773 of *Lecture Notes in Computer Science*, pages 34–41, Cham. Springer.

D. Nurse. 2008. *Tense and Aspect in Bantu*. Oxford University Press, Oxford, UK.

I. E. Onyenwe. 2017. *Developing Methods and Resources for Automated Processing of Igbo*. Ph.D. thesis, University of Sheffield.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Violet M. S. Pule and Kgomotso Theledi. 2023. The impact of the presence of prosodic features (tone markings) on comprehending Setswana words in reading. *African Journal of Inter/Multidisciplinary Studies*, 5(1):1–12.

L. Rabiner and B.-H. Juang. 1993. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ.

Mpho Raborife. 2022. Motivation for corpus selection. Unpublished manuscript, Department of Applied Information Systems, University of Johannesburg.

Mpho Raborife, Sigrid Ewert, and Sabine Zerbian. 2015. Improving a tone labelling algorithm for Sesotho. *Language Resources and Evaluation*, 49(1):19–50.

Roopika Risam. 2019. *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy*. Northwestern University Press, Evanston, IL, USA.

B. Schölkopf and A. J. Smola. 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA.

D. Wissing and J. C. Roux. 2017. The status of tone in Sesotho: A production and perception study. *Nordic Journal of African Studies*, 26(4):384–402.

A Speaker-Level Counts

Table 3 gives the per-speaker composition of the corpus and the partition to which each speaker was assigned. "Predominant gender" is the majority pitch-based label across that speaker's syllables (F = female, M = male).

B Hyperparameter Grids and Selected Configurations

Grids were searched exhaustively with speaker-grouped 5-fold cross-validation on the training partition. Logistic regression: $C \in \{0.01, 0.1, 1.0, 10.0\}$, solver $\in$ {lbfgs, liblinear} (8 configurations). Random forest: $n_{\text{estimators}} \in \{50, 100, 200\}$, max_depth $\in$ {None, 10, 20, 30}, min_samples_split $\in \{2, 5, 10\}$, min_samples_leaf $\in \{1, 2, 4\}$ (108). SVM: $C \in \{0.1, 1.0, 10.0\}$, $\gamma \in$ {scale, auto, 0.001, 0.01}, kernel $\in$ {rbf, linear} (24). XGBoost: $n_{\text{estimators}} \in \{50, 100, 200\}$, max_depth $\in \{3, 5, 7\}$, learning rate $\in \{0.01, 0.1, 0.3\}$, subsample $\in \{0.8, 1.0\}$ (54). MLP: hidden_layer_sizes $\in \{(50), (100), (100, 50), (100, 100)\}$, activation $\in$ {relu, tanh}, $\alpha \in \{0.0001, 0.001, 0.01\}$, learning rate $\in$ {constant, adaptive} (48). Table 4 lists the configuration selected for each method and classifier.

C Feature Composition and Algorithm Summaries

FFT-based approach. A Hamming window is applied per segment before computing the FFT magnitude spectrum, from which spectral centroid, spread, rolloff, flatness, peak frequency, F0 within the 80–400 Hz speech range, harmonic energy ratios for the first five harmonics, four-band energy distribution, and spectral flux are extracted. *Strengths:* direct harmonic capture, fast computation, no dependency on pre-trained models, phonetically interpretable features. *Limitations:* sensitivity to window/parameter choice and to recording noise.

Speaker	Partition	Files	Syllables	High	Low	Unlabelled	Predominant gender
KM	Train	30	225	35	41	149	F
KN	Test	30	225	32	39	154	F
LL	Train	31	235	35	40	160	F
LM	Train	29	214	36	40	138	F
LP	Train	30	226	35	41	150	F
MM	Test	30	225	34	40	151	F
NM	Train	90	675	72	80	523	F
NN	Test	30	226	37	39	150	F
OR	Train	30	226	36	40	150	M
PL	Train	30	226	34	42	150	M
RM	Train	30	224	32	44	148	F
SM	Train	30	226	34	42	150	F
TL	Train	60	452	68	78	306	M
TQ	Train	30	225	35	41	149	F
VS	Test	30	226	34	42	150	F
Total		540	4,056	589	689	2,778	

Table 3: Per-speaker composition. Test partition: 137 High and 160 Low labelled syllables (297). Training partition: 452 High and 529 Low labelled syllables (981).

Librosa pipeline. PYIN-derived pitch statistics, RMS energy, zero-crossing rate, spectral centroid, and syllable duration are combined into an 8-dimensional feature vector. *Strengths:* parsimonious, low computational cost. *Limitations:* no explicit spectral-envelope or harmonic modelling, limiting sensitivity to subtle tonal cues.

MFCC-based approach. Thirteen MFCCs (with mean/SD/min/max statistics), delta and delta-delta coefficients, spectral contrast, chroma, Tonnetz, and complementary pitch/ZCR/bandwidth/rolloff features are concatenated into a 156-dimensional vector. *Strengths:* perceptually grounded, captures temporal dynamics via delta features. *Limitations:* high dimensionality relative to the labelled sample size; possible loss of fine frequency resolution from mel binning.

Wav2Vec2-based approach. Audio is resampled to 16 kHz and passed through the pre-trained `facebook/wav2vec2-base` model (Baevski et al., 2020); the final hidden state (768-d) is mean-pooled across time and concatenated with pitch, RMS, ZCR, spectral centroid, and duration statistics for a 776-dimensional vector. *Strengths:* transfer learning from large-scale pre-training, no manual feature engineering, robustness to speaker variability. *Limitations:* high computational cost, elevated overfitting risk given 1,278 labelled samples against 768 embedding dimensions, and reduced interpretability.

D All Method–Classifier Results

F1 and ROC-AUC for every method–classifier pair on the 297-syllable test set.

Method	Classifier	Best CV F1	Selected hyperparameters
FFT	LR	.8352	C=1.0, solver=lbfgs
FFT	RF	.8482	n_{est}=200, max_depth=10, min_split=2, min_leaf=1
FFT	SVM	.8464	C=10, γ=scale, kernel=rbf
FFT	XGB	.8564	n_{est}=100, max_depth=7, lr=0.1, subsample=0.8
FFT	NN	.8549	hidden=(50), activation=tanh, α=0.01, lr=constant
MFCC	LR	.8764	C=0.1, solver=liblinear
MFCC	RF	.8824	n_{est}=50, max_depth=10, min_split=10, min_leaf=4
MFCC	SVM	.8828	C=1, γ=scale, kernel=rbf
MFCC	XGB	.9027	n_{est}=200, max_depth=3, lr=0.1, subsample=0.8
MFCC	NN	.8749	hidden=(50), activation=relu, α=0.0001, lr=constant
Wav2Vec2	LR	.7915	C=0.01, solver=lbfgs
Wav2Vec2	RF	.7464	n_{est}=200, max_depth=None, min_split=2, min_leaf=2
Wav2Vec2	SVM	.7681	C=10, γ=scale, kernel=rbf
Wav2Vec2	XGB	.8071	n_{est}=200, max_depth=5, lr=0.3, subsample=0.8
Wav2Vec2	NN	.7777	hidden=(100,100), activation=tanh, α=0.0001, lr=constant
Librosa	LR	.5938	C=10, solver=lbfgs
Librosa	RF	.7508	n_{est}=100, max_depth=10, min_split=2, min_leaf=1
Librosa	SVM	.7272	C=10, γ=scale, kernel=rbf
Librosa	XGB	.7554	n_{est}=100, max_depth=3, lr=0.3, subsample=1.0
Librosa	NN	.7413	hidden=(100,100), activation=tanh, α=0.01, lr=constant

Table 4: Selected hyperparameters for each method–classifier pair, with the corresponding speaker-grouped cross-validation F1. LR = logistic regression; RF = random forest; SVM = RBF or linear-kernel support vector machine; XGB = XGBoost; NN = MLP.

Method	Type	Key features
FFT	Spectral / harmonic	Spectral centroid, spread, rolloff, flatness; peak frequency; F0 (80–400 Hz); harmonic energy ratios (5 harmonics); energy in four frequency bands; spectral flux; plus the eight shared features
Librosa	Acoustic / temporal	The eight shared features only
MFCC	Cepstral / perceptual	13 MFCCs (mean, SD, min, max); delta and delta-delta (mean, SD); spectral contrast (7 bands); chroma (12 classes); Tonnetz (6 centroids); pitch, ZCR, spectral centroid, bandwidth, rolloff, RMS, and duration
Wav2Vec2	Deep embedding	768 mean-pooled `facebook/wav2vec2-base` hidden-state embeddings; plus the eight shared features

Table 5: Feature extraction methods and their composition. The eight shared features are pitch mean, SD, minimum and maximum (PYIN), RMS energy, zero-crossing rate (ZCR), spectral centroid, and syllable duration.

	FFT		MFCC		Wav2Vec2		Librosa	
Classifier	F1	ROC	F1	ROC	F1	ROC	F1	ROC
Neural net (NN)	.8718	.9358	.9190	.9583	.7732	.8846	.6817	.7595
XGBoost (XGB)	.8511	.9219	.8890	.9582	.7865	.8901	.6863	.7776
Random forest (RF)	.8112	.9147	.8782	.9459	.7377	.8622	.6679	.7750
Logistic reg. (LR)	.8191	.9157	.9089	.9486	.7833	.8813	.6196	.7106
SVM (RBF)	.8398	.9224	.9120	.9530	.7966	.8938	.6916	.7740
F1 range across classifiers	.0606		.0408		.0589		.0720	

Table 6: Held-out F1 and ROC-AUC for all 20 method–classifier combinations (297-syllable test set). The best-classifier entry for each method is the one in Table 1.