Dynamic Layer Selection for Efficient Tone Recognition in Self-Supervised Speech Models

Saint Germes Bienvenu Bengono Obiang, Norbert Tsopze, Paulin Melatagia Yonta
Department of Computer Science, University of Yaoundé I, Cameroon
Sorbonne Université - IRD - UMMISCO - F-93143, Bondy, France
{bengobiang, tsopze.norbert, paulinyonta}@gmail.com

Abstract

Low-resource tonal languages present significant challenges to speech processing technologies, due to limited training data and the critical role of pitch variation in expressing meaning. This paper applies established weighted layer combination methods to tone recognition in such languages, with a specific focus on Yoruba and Yemba. Building on our previous work with Wav2vec 2.0 representations and the weighted-sum methodology from Yang et al. (2024), we investigate layer specialisation in the SSA-HuBERT self-supervised speech model for tonal tasks. Our systematic analysis reveals significant performance differences between different layers, with middle layers generally outperforming both lower and upper layers for tonal recognition tasks. While typical approaches only use the output of the last layer, our experiments show that weighted layer combination outperforms the last layer by 20.4% and 15.8% relative improvement in tone error rate (TER) for Yoruba and Yemba, respectively. In addition to performance improvements, our approach provides dramatic computational efficiency gains, reducing the resources required by over 90% compared to evaluating each layer separately. Analysis of the learned layer weights reveals language-specific patterns, with Yoruba favouring middle layers and Yemba giving more weight to early layers. These results provide valuable insights into how tonal information is encoded in self-supervised speech models, and demonstrate a practical application of established layer combination methods in low-resource language contexts.

Keywords: speech processing, tone recognition, low-resource languages, self-supervised learning, dynamic layer selection

1. Introduction

Tonal languages convey lexical or grammatical meaning through pitch variations, where identical segmental sequences can differ semantically depending on tone. In Yoruba, for example, *Kí* "to salute", *Ki* "dense", and *Kì* "to press lightly" differ only by tone. Ignoring such contrasts severely limits the performance of automatic speech recognition systems (van Niekerk and Barnard, 2012; Odejobi, 2008).

Many sub-Saharan African languages are tonal and remain low-resource due to the scarcity of labelled data. Recent advances in self-supervised learning (SSL) for speech have provided effective representations learned from large unlabelled corpora (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022). These models, such as Wav2vec 2.0 and HuBERT, have achieved competitive performance across several downstream tasks, including speech recognition in low-resource settings.

Bengono OBIANG et al. (2024) showed that Wav2vec 2.0 representations outperform traditional spectral features for tone recognition in Yoruba, reducing the Tone Error Rate (TER) from 19.45% to 17.72%. This confirmed that SSL features trained on high-resource languages can transfer effectively to tonal tasks. However, that study relied solely on the final layer of the pretrained model, potentially overlooking complementary information from intermediate layers.

Prior analyses of SSL models have shown layer-specific specialisation: lower layers encode speaker and acoustic traits, middle layers capture phonetic and prosodic cues, and upper layers represent linguistic and semantic information (Chen et al., 2022). For tone recognition, tonal cues are likely concentrated in particular intermediate layers, yet this relationship remains largely unexplored, especially in low-resource tonal languages.

To address this gap, we apply weighted layer combination techniques (Yang et al., 2024) to investigate layer specialisation in the SSA-HuBERT model for tone recognition in Yoruba and Yemba. We systematically evaluate how tonal information is distributed across layers and how combining them affects performance and efficiency. The approach follows the weighted-sum methodology introduced in SUPERB, adapted here for tonal tasks.

The main contributions of this work are:

1. A systematic analysis of layer specialisation in SSA-HuBERT for tone recognition in Yoruba and Yemba, showing that middle layers outperform both lower and upper layers.

2. Application of weighted layer combination for tone modelling, achieving relative TER reductions of 20.4% for Yoruba and 15.8% for Yemba compared with the final layer.

3. Demonstration of over 90% reductions in training time and memory usage compared with

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10093–10101
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

training a separate model for each layer.

4. Evidence of language-specific weighting patterns: Yoruba emphasises middle layers, while Yemba distributes weights across early and middle layers.

These findings reveal how tonal information is encoded in self-supervised speech models and provide a practical means of exploiting it for tone recognition in low-resource settings. The proposed weighted-layer approach improves accuracy while substantially reducing computational cost, enabling efficient experimentation for researchers with limited resources.

2. Related Work

Tone recognition has been studied extensively, mainly for Mandarin Chinese. Early systems relied on explicit pitch features, such as F0 contours. Li et al. (2006) employed artificial neural networks with F0 features and achieved 90% accuracy on monosyllabic sequences. Subsequent work explored spectral representations to overcome F0 limitations. Ryant et al. (2014); Chen et al. (2016) showed that MFCC-based features achieved a lower Frame Error Rate (16.36%) than F0-based approaches (24.22%) on the Mandarin Broadcast News corpus. Integrating multiple cues, including duration and energy, further improved accuracy to around 78% (Chen et al., 2014). Lugosch and Tomar (2018) later introduced cepstrograms for continuous speech tone modelling, reporting a Tone Error Rate (TER) of 11.7% on AISHELL-1, surpassing traditional feature-based methods.

Most of these methods depend on accurate syllable audio alignment, which is difficult to obtain for low-resource tonal languages. Adams et al. (2017) proposed an alignment-free approach using Filter Banks and F0 features within an LSTM–CTC framework. Bengono OBIANG et al. (2024) extended this line of work to Yoruba, showing that Wav2vec 2.0 representations outperform handcrafted features, reducing the TER from 19.45% to 17.72%. However, only the final layer of the model was used, potentially overlooking information encoded in intermediate layers.

Recent progress in self-supervised learning (SSL) has transformed speech processing. These models learn universal acoustic representations from unlabelled data using objectives such as autoencoding (Vincent et al., 2010; Le et al., 2018), contrastive learning (Baevski et al., 2020; Schneider et al., 2019), and masked prediction (Hsu et al., 2021). HuBERT introduced a masked prediction strategy guided by offline clustering, while WavLM (Chen et al., 2022) added denoising objectives and achieved state-of-the-art results on the SUPERB benchmark.

Analyses of SSL models have revealed that their internal layers specialise in distinct aspects of speech (Chmiel et al., 2023). Lower layers capture speaker and acoustic information, middle layers encode phonetic and prosodic features, and upper layers represent high-level linguistic or semantic content (Chen et al., 2022). Traditional fine-tuning often relies on the final layer, but recent approaches employ weighted combinations of all layers to exploit complementary information. In the SUPERB benchmark (Hsu et al., 2021), this weighted-sum approach achieved consistent gains across tasks and models.

Yang et al. (2024) provided a large-scale analysis of 36 SSL models and 15 tasks, showing that trainable layer weighting consistently outperforms single-layer extraction. Their results confirmed that different tasks rely on distinct layers, although learned weights do not always correlate with intrinsic layer quality due to optimisation biases.

Building on these insights, we apply the weighted layer combination framework to tone recognition in low-resource African languages. While the mathematical formulation follows the SUPERB protocol, our study focuses on how tonal information is distributed across layers in SSL-HuBERT. This work constitutes the first systematic analysis of layer specialisation for tonal phenomena in low-resource contexts.

3. Proposed Method

We propose an efficient tone recognition framework for low-resource tonal languages based on dynamic layer weighting in self-supervised speech models.

3.1. Problem Formulation

Given an input speech sequence $x = (x_1, \ldots, x_T)$, tone recognition aims to predict the tone sequence $y = (y_1, \ldots, y_L)$, where T and L denote the lengths of the speech and tone sequences. Each tone is drawn from a finite inventory (e.g., high, mid, low for Yoruba).

Let $f_\theta(x)$ denote a pre-trained self-supervised model producing a hierarchy of hidden representations $H = \{h^0, h^1, \ldots, h^N\}$, with $h^i \in \mathbb{R}^{T' \times d}$. Conventional approaches typically use the final layer h^N for downstream tasks. We instead learn a function $g_\phi(H)$ that computes a weighted combination of all layers to obtain $h_{\text{combined}} \in \mathbb{R}^{T' \times d}$ (Yang et al., 2024), optimised for tone recognition. The training objective minimises the Tone Error Rate (TER):

$$\min_{\phi} \text{TER}(y, \hat{y}), \tag{1}$$

where $\hat{y}$ is the predicted tone sequence derived from h_{combined}.

3.2. SSA-HuBERT

Experiments are conducted using SSA-HuBERT (Caubrière and Gauthier, 2024), a variant of HuBERT (Hsu et al., 2021) pre-trained on nearly 60 000 hours of unlabelled speech from 21 African languages and dialects. The model follows the HuBERT-base configuration with 12 Transformer layers and 90 M parameters. Training involves two iterations of masked prediction using cluster-based pseudo-labels, enabling the network to learn both acoustic and phonetic regularities. Its Africa-centric pre-training provides strong coverage of prosodic and phonological patterns relevant to tonal languages, making it well suited to our task.

3.3. Model Architecture

The proposed architecture (Figure 1) comprises three components: the pre-trained SSA-HuBERT encoder, a dynamic layer weighting module, and a bidirectional RNN decoder trained with CTC loss.

3.3.1. Weighted Layer Combination

Following the SUPERB methodology (Yang et al., 2024), each layer representation h^i receives a learnable weight w_i, producing a weighted sum:

$$h_{\text{combined}} = \sum_{i=0}^{N} w_i \cdot h^i. \tag{2}$$

Weights are normalised using a softmax function:

$$w_i = \frac{\exp(\lambda_i / T)}{\sum_{j=0}^{N} \exp(\lambda_j / T)}, \tag{3}$$

where λ_i are trainable parameters and T is a temperature controlling sharpness. Lower T values emphasise a few dominant layers, while higher T encourages broader combinations. This mechanism enables the model to automatically identify layers most informative for tonal cues, providing both interpretability and efficiency.

3.3.2. Bidirectional RNN Decoder

The combined representation h_{combined} is passed through a bidirectional RNN with GRU units to capture contextual dependencies in both temporal directions. Forward and backward hidden states are concatenated and fed to a linear projection layer that outputs tone-class logits. This structure effectively models the suprasegmental nature of tone by integrating neighbouring contextual information.

3.3.3. Connectionist Temporal Classification

The decoder is trained with Connectionist Temporal Classification (CTC) (Graves et al., 2006), which aligns tone sequences without explicit segmentation. CTC naturally handles variable-length inputs, permits blank symbols for non-tonal frames, and eliminates the need for manual alignment resources. The training loss is defined as:

$$\mathcal{L}_{\text{CTC}} = -\log p(y|h_{\text{combined}}), \tag{4}$$

where $p(y|h_{\text{combined}})$ denotes the marginal likelihood of the target sequence. During inference, tone sequences are obtained using beam search decoding to identify the most probable alignment.

This architecture jointly optimises tone prediction accuracy and computational efficiency. Dynamic layer weighting exploits the most informative layers while avoiding the cost of training a separate model for each layer.

4. Experiments

This section details the datasets, training configurations, and results used to evaluate our dynamic layer weighting approach for tone recognition in low-resource tonal languages.

4.1. Datasets

Experiments were conducted on Yoruba and Yemba, two tonal African languages differing in resource availability. Yoruba provides around 4 h of labelled data, while Yemba offers only 47 min, allowing evaluation under progressively constrained conditions.

4.1.1. Yoruba Dataset

We used the Yoruba corpus from Gutkin et al. (2020), containing 4 h of speech recorded at 48 kHz by 36 male and female speakers. Originally designed for TTS and ASR, it was adapted for tone recognition through automatic syllabification (van Niekerk and Barnard, 2012) and tone extraction from each syllable. Yoruba has three tones: high (◌́), mid (unmarked), and low (◌̀). Manual checks confirmed tone accuracy. After filtering utterances longer than 6 s and resampling to 16 kHz, the dataset comprised 3 223 utterances (≈3.3 h).

4.1.2. Yemba Dataset

The YembaTones corpus (Kenfack Jeuguim et al., 2024) is a syllable-tone annotated dataset designed for tone analysis. It includes 3 420 recordings (47 min) from 11 native speakers (4 men, 7 women) covering 344 words arranged in 149 tonal

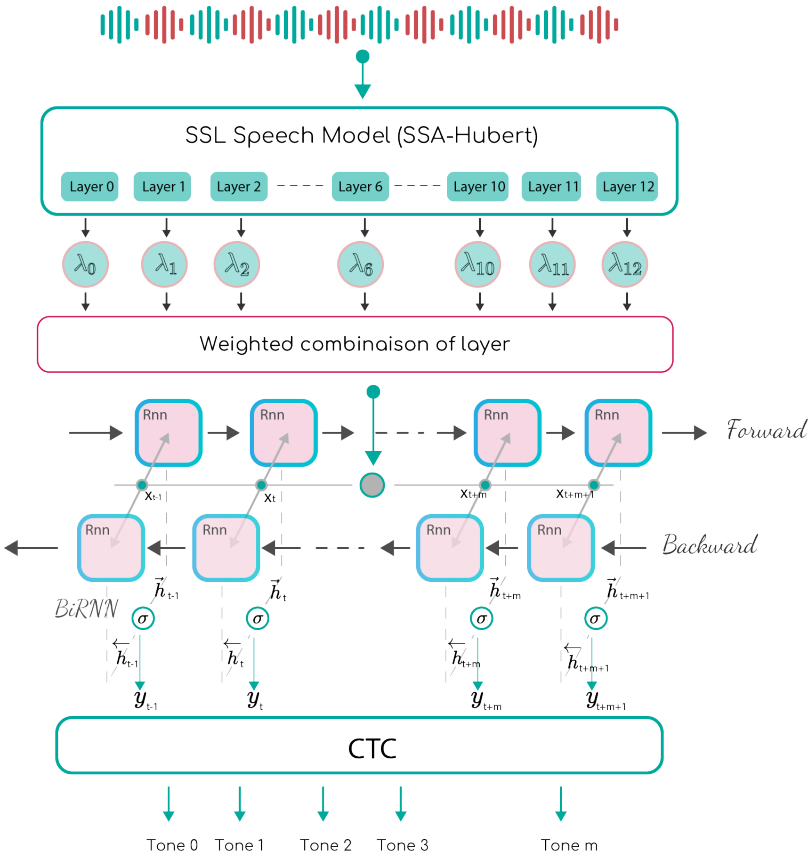


Figure 1: Proposed tone recognition model. Layer representations from SSA-HuBERT are dynamically weighted, combined, and processed by a bidirectional RNN with CTC decoding.

minimal sets. Recordings were made at 44.1 kHz and manually annotated with Praat TextGrid files, verified by linguists. After resampling to 16 kHz and cleaning, we retained 2 012 utterances. Preprocessing used the open-source YembaTones Toolkit.[1] This dataset represents an extremely low-resource setting, serving as a benchmark for evaluating method robustness under minimal supervision.

4.2. Evaluation Metric

Model performance was measured using the Tone Error Rate (TER), computed as the normalised Levenshtein distance between predicted and reference tone sequences:

$$\text{TER} = \frac{I + D + S}{N} \times 100\%, \tag{5}$$

where I, D, and S represent insertions, deletions, and substitutions, and N denotes the reference length.

[1] https://github.com/germes96/YembaTones-Toolkit

4.3. Implementation Details

All models were implemented with SpeechBrain (Ravanelli et al., 2021). Representations were extracted from the 13 SSA-HuBERT layers (including the input embedding). Unless otherwise stated, temperature was set to $T = 1.0$. The decoder used bidirectional GRU units. Training used the Adam optimiser with an adaptive scheduler (initial learning rate 1.0, decayed by 0.8 when no improvement was observed). Models were trained for 50 epochs with a batch size of 2 on a single NVIDIA V100 GPU.

4.4. Training Procedure

In all experiments, the SSA-HuBERT encoder was fine-tuned end-to-end jointly with the dynamic layer weights and BiRNN decoder; no layers were frozen. Audio was normalised and resampled to 16 kHz. To improve robustness, we applied data augmentation: (i) additive noise from MUSAN (SNR 5–15 dB) (Ko et al., 2017), (ii) time masking and warping (Park et al., 2019), and (iii) pitch perturbation (±5%) to preserve tonal patterns. Validation TER was tracked at each epoch, and early

stopping (patience = 1) retained the best checkpoint. We also monitored layer-weight evolution to analyse how layer importance developed during training for each language.

4.5. Single-Layer Performance

To identify which layers encode tonal information most effectively, we first trained a separate model for each SSA-HuBERT layer.

4.5.1. Yoruba

Figure 2 shows the TER per layer. Middle layers (6–9) achieved significantly better results than both early and late layers. Layer 8 obtained the lowest TER (12.84%), whereas the final layer reached 15.66%, corresponding to a 21.8% relative difference. This U-shaped trend indicates that tonal cues are best captured in intermediate representations, while upper layers prioritise higher-level linguistic information.

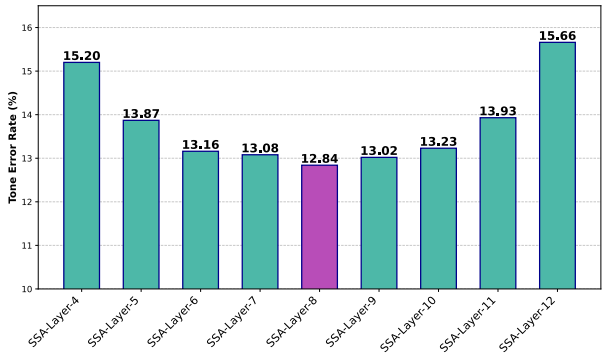


Figure 2: TER performance across SSA-HuBERT layers for Yoruba. Middle layers (6–9) outperform early and late layers.

4.5.2. Yemba

As shown in Figure 3, the Yemba dataset displays a similar but more irregular pattern. Layer 7 achieved the best TER (25.37%), while the last layer reached 30.04%, an 18.5% relative difference. This greater variability likely reflects the smaller dataset, but the trend of superior middle-layer performance remains consistent.

4.6. Dynamic Layer Weighting Results

Table 1 compares dynamic weighting against the baseline and best single layers. For Yoruba, dynamic selection achieved 12.46% TER, outperforming the baseline (15.66%) by 20.4% relative and the best single layer (12.84%) by 3.0%. For Yemba, TER decreased from 30.04% to 25.29%,

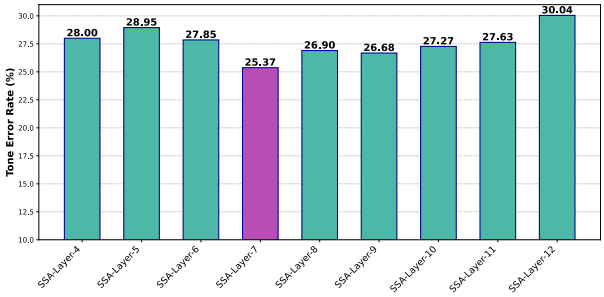


Figure 3: TER across SSA-HuBERT layers for Yemba. Middle layers outperform the last layer despite small data size.

yielding 15.8% relative improvement. These results confirm that dynamically combining layers consistently outperforms static selection.

Table 1: TER (%) for selected single-layer configurations and dynamic layer weighting. Full per-layer results are shown in Figures 2 and 3.

Method	Yoruba	Yemba
Baseline (L12)*	15.66	30.04
Worst non-baseline (L4 / L5)	15.20	28.95
Median layer (L10 / L6)	13.23	27.85
Best single layer (L8 / L7)	12.84	25.37
Dynamic Layer Weighting	**12.46**	**25.29**

*Standard end-to-end fine-tuning using the last transformer layer output (Layer 12), as is conventional in HuBERT-based systems.

4.7. Layer Weight Distribution

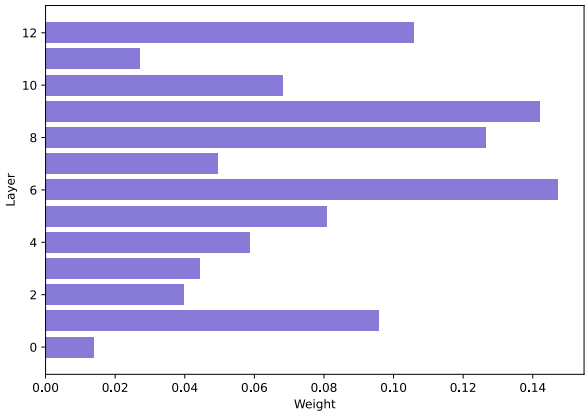
Figure 4 illustrates the learned layer weights. Yoruba shows a strong focus on middle layers (6–9), with peaks at layers 6 (0.1474) and 8 (0.1265). Layer 12 also contributes moderately (0.1058). Yemba displays a more balanced distribution, assigning weight to both early (0, 5) and middle (6, 7) layers. These distinct patterns reflect adaptation to language-specific signal structures and dataset size.

4.8. Computational Efficiency

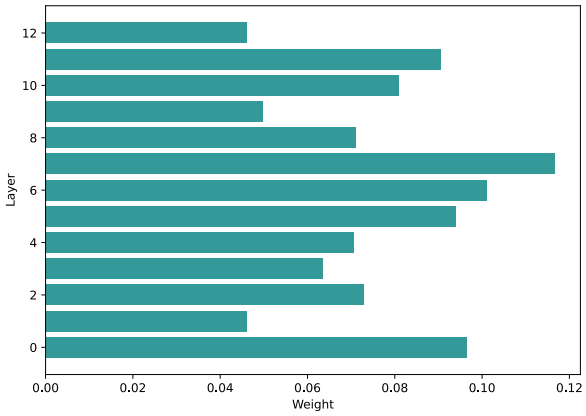
A key advantage of the proposed approach is computational efficiency. We compared dynamic weighting with the static single-layer baseline, where a separate model is trained for each of the 13 layers.

4.8.1. Training Time

Table 2 presents the training time comparison. For Yoruba, training 13 independent single-layer mod-



(a) Yoruba



(b) Yemba

Figure 4: Normalized layer weight distributions for Yoruba and Yemba.

Table 2: Training time (hours) comparison between static and dynamic approaches.

Approach	Yoruba	Yemba
Static (single-layer ×13)	37.05	28.60
Dynamic weighting	3.33	2.22
Time Savings	91.0%	92.2%

els took 37.05 h in total, whereas the dynamic weighting model completed training in 3.33 h, reducing time by 91.0%. For Yemba, training time dropped from 28.6 h to 2.22 h, representing a 92.2% reduction. These results confirm that dynamic layer weighting substantially improves training efficiency without compromising accuracy.

4.8.2. Memory and Storage

Both approaches used about 60% of GPU memory. However, the static setup required 13 separate models (≈6.1 GB total), while the dynamic model required only one (468 MB), yielding a 92.3% reduction in storage. These savings enable large-scale experimentation even on modest hardware, supporting low-resource research.

4.9. Error Analysis

Figure 5 compares error metrics between static and dynamic models. For Yoruba, dynamic weighting reduces deletion errors by 20.3% and substitution errors by 22.5%, improving overall TER by 20.4% relative. For Yemba, deletion errors decrease by 54.5%, and sentence error rate drops by 20.5%. In both languages, tone-specific accuracy improved, particularly for mid tones (Yoruba: 88.80%→91.92%; Yemba: 69.40%→77.60%). These results confirm that dynamic layer weighting enhances tonal transition modelling and reduces common recognition errors.

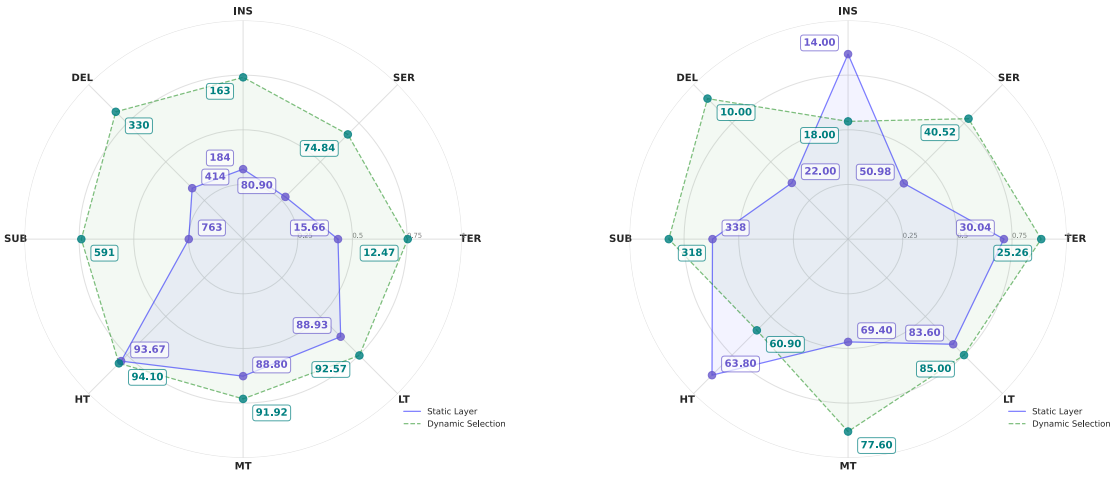
5. Discussion

5.1. Low-Resource Validation and Methodological Strengths

Our experimental design intentionally targets the most challenging conditions for tone recognition in African languages. The choice of two low-resource languages (Yoruba: 4 hours; Yemba: 47 minutes) provides a stringent and realistic validation scenario for typical African tonal language contexts. Rather than relying on well-resourced languages, the progressive reduction in available data serves as a stress test that highlights the robustness of our method under increasingly scarce data conditions.

The strong performance of the weighted layer combination approach using only 47 minutes of Yemba speech confirms its suitability for the majority of African languages, which generally lack extensive corpora. This validation is critical since it aligns with the real-world situation of tone language research on the continent, where annotated data are minimal. Including well-resourced languages would have reduced the ecological validity of our evaluation, diverting it from the actual low-resource challenge we aim to address.

Furthermore, the contrast between the Yemba dataset of isolated words and the Yoruba dataset of continuous speech strengthens the robustness of our validation. The ability of our approach to perform well across these varied conditions demonstrates that the weighted layer combination technique generalises effectively to both word-level and continuous speech tone recognition, which is essential for broad deployment across African languages.



(a) Yoruba (b) Yemba

Figure 5: Error metric analysis: Dynamic vs. static layer selection in Yoruba and Yemba tone recognition. Metrics: **TER** (Tone Error Rate), **SER** (Sentence Error Rate), **INS** (Insertions), **DEL** (Deletions), **SUB** (Substitutions), and per-tone accuracies (**HT**, **MT**, **LT**).

5.2. Computational Accessibility for Resource-Constrained Research

The significant computational efficiency gains, with over 90% reduction in training time and 92.3% reduction in storage requirements, make our method particularly accessible for researchers operating with limited resources. Such efficiency enables wider participation in speech technology development for underrepresented languages and lowers entry barriers for institutions with modest computational capacity. This accessibility is a crucial step towards equitable research participation and sustainable development of speech technologies in Africa.

5.3. Limitations and Future Directions

While our study on Yoruba and Yemba provides compelling evidence for the effectiveness of weighted layer combination in low-resource tonal contexts, it remains limited in linguistic coverage. Future work should aim to extend this analysis to a broader range of African languages as community datasets become available.

Additionally, our use of a fixed temperature parameter ($T = 1.0$) simplifies the model and could be optimised for further gains. Investigating a frozen-encoder setting, where only the layer weights and BiRNN decoder are learned while SSA-HuBERT parameters remain fixed, would help disentangle the contribution of layer weighting from encoder adaptation. Alternative strategies such as uniform averaging, layer-wise attention, top-K sparse selection, and concatenation with dimensionality reduction warrant investigation. We favoured learned scalar weights for their minimal parameter overhead and interpretability in low-resource settings. Comparing these approaches is an important direction for future work.

6. Conclusion

This study examined layer specialisation in self-supervised speech models for tone recognition in low-resource tonal languages. Using the weighted layer combination approach of Yang et al. (2024), we conducted the first systematic analysis of how different layers in SSA-HuBERT encode tonal information for African languages, focusing on Yoruba and Yemba.

Our findings show that middle layers of SSA-HuBERT capture tone information more effectively than either lower or higher layers. The weighted combination approach achieved **20.4%** and **15.8%** relative Tone Error Rate (TER) improvements over last-layer baselines for Yoruba and Yemba, respectively, while reducing computational requirements by over **90%** compared to training a separate model for each layer.

Analysis of learned weights revealed language-specific trends: Yoruba favoured middle layers, while Yemba exhibited a more even distribution between early and middle layers. These results underscore both the adaptability of weighted layer combination methods and the language-dependent nature of tone representation in self-supervised models.

Most notably, the Yemba results, achieved with just **47 minutes** of training data, demonstrate that our approach is viable for the highly resource-

limited environments characteristic of African language research. The method's dramatic computational efficiency makes it accessible to research groups with limited infrastructure, paving the way for scalable tone recognition across hundreds of underrepresented African tonal languages.

Overall, this work establishes a practical foundation for explainable and efficient tone recognition in low-resource contexts. Its adaptability across speech types (isolated versus continuous) and tonal systems positions it as a promising solution for future African language technology initiatives.

Future research should pursue the community-driven extension of datasets, temperature optimisation, and exploration of advanced layer combination strategies to further enhance performance in resource-constrained environments.

7. Bibliographical References

Oliver Adams, Trevor Cohn, Graham Neubig, and Alexis Michaud. 2017. Phonemic transcription of low-resource tonal languages. In *Proceedings of the Australasian Language Technology Association Workshop 2017*, pages 53–60, Brisbane, Australia.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

S.G.B. Bengono OBIANG, Paulin Melatagia Yonta, Norbert Tsopze, Jean-Francois Bonastre, and Tania Jimenez. 2024. Improving tone recognition performance using wav2vec 2.0-based learned representation in yoruba, a low-resourced language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 1(1).

Antoine Caubrière and Elodie Gauthier. 2024. Africa-centric self-supervised pre-training for multilingual speech representation in a sub-saharan context.

Charles Chen, Razvan C. Bunescu, Li Xu, and Chang Liu. 2016. Tone classification in mandarin chinese using convolutional neural networks. In *Interspeech*.

Mingming Chen, Zhanlei Yang, and Wen-Ju Liu. 2014. Deep neural networks for mandarin tone recognition. pages 1154–1158.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:3451–3460.

Wojciech Chmiel, Joanna Kwiecień, and Kacper Motyka. 2023. Saliency map and deep learning in binary classification of brain tumours. *Sensors*, 23(9).

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.

Alexander Gutkin, Işın Demirşahin, Oddur Kjartansson, Clara Rivera, and Kọ́lá Túbọ̀sún. 2020. Developing an Open-Source Corpus of Yoruba Speech. In *Proceedings of Interspeech 2020*, pages 404–408, Shanghai, China. International Speech and Communication Association (ISCA).

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdel rahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Marc Sturm Kenfack Jeuguim, Paulin Melatagia Yonta, and Etienne Sandembouo. 2024. Yembatones: A syllable-tone annotated dataset for speech recognition and prosodic analysis of the yemba language. *Data in Brief*, 52:109860.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur. 2017. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5220–5224.

Tuan Anh Le, Maximilian Igl, Tom Rainforth, Tom Jin, and Frank Wood. 2018. Auto-encoding sequential monte carlo. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Xu Li, Zhang Wenle, Zhou Ning, Lee Chaoyang, Li Yongxin, Chen Xiuwu, and Zhao Xiaoyan.

2006. Mandarin chinese tone recognition with an artificial neural network. *Journal of Otology*, 1(1):30–34.

Loren Lugosch and Vikrant Singh Tomar. 2018. Tone recognition using lifters and ctc. In *Interspeech*.

Odetunji Ajadi Odejobi. 2008. Recognition of tones in yorùbá speech: Experiments with artificial neural networks. In *Speech, Audio, Image and Biomedical Signal Processing using Neural Networks*.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*. ISCA.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Neville Ryant, Jiahong Yuan, and Mark Liberman. 2014. Mandarin tone classification without pitch tracking. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4868–4872.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *Interspeech 2019*, pages 3465–3469.

Daniel R. van Niekerk and Etienne Barnard. 2012. Tone realisation in a yorùbá speech recognition corpus. In *Proc. 3rd Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2012)*, pages 54–59.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408.

Shu-wen Yang, Heng-Jui Chang, Zili Huang, Andy T. Liu, Cheng-I Lai, Haibin Wu, Jiatong Shi, Xuankai Chang, Hsiang-Sheng Tsai, Wen-Chin Huang, Tzu-hsun Feng, Po-Han Chi, Yist Y. Lin, Yung-Sung Chuang, Tzu-Hsien Huang, Wei-Cheng Tseng, Kushal Lakhotia, Abdelrahman Mohamed, Shang-Wen Li, Shinji Watanabe, and Hung-yi Lee. 2024. A large-scale evaluation of speech foundation models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2884–2899.