Exploring syllable similarity across South African languages through self-supervised speech representation

Johannes Abraham Louw[0000−0002−8168−7857]

Natural Language Processing Research Group,
Next Generation Enterprises and Institutions,
CSIR,
Pretoria, South Africa
jalouw@csir.co.za
https://www.csir.co.za

Abstract. Syllables are fundamental units in speech production and carry prosodic information, but their acoustic and linguistic properties across different language families are not well understood. This study examines syllable discovery approaches across South African languages using algorithmic syllabification and S5-HuBERT, a self-supervised speech representation model that demonstrates emergent syllabic organization. We analyzed speech recordings from eleven languages representing five language families in South Africa using a systematic comparison of rule-based and data-driven syllable discovery methods. We evaluated both approaches using cross-linguistic consistency measures and acoustic quality assessments across speakers.
Our analysis reveals fundamental differences between the two approaches. Algorithmic syllables demonstrate strong language-family clustering with predominantly language-specific units, while S5-HuBERT units show superior cross-linguistic sharing and weaker family effects. Speaker independence analysis across four experimental phases demonstrates that data-driven methods achieve better acoustic consistency, with the fully data-driven approach reaching near-optimal speaker generalization. These results provide empirical guidance for implementing syllable-based semantic units in multilingual text-to-speech systems for resource-scarce languages.

Keywords: Self-supervised learning · Speech representation · Syllable segmentation · S5-HuBERT · Cross-linguistic analysis · Text-to-speech synthesis

1 Introduction

Developing effective text-to-speech systems for resource-scarce languages requires architectures that can leverage limited high-quality data efficiently. Cross-linguistic generalizability enables such systems to benefit from resource-rich languages and to handle the code-switching scenarios common in multilingual environments. This challenge is particularly pronounced in multilingual contexts

such as South Africa, where ten of the eleven official spoken languages[1], spanning diverse language families, are considered resource-scarce and have limited speech corpora. The term "resource-scarce" refers to languages that possess only a minimal set of essential text-processing tools, such as basic word segmentation, text normalization, and grapheme-to-phoneme (G2P) conversion, together with a small corpus of relatively high-quality audio recordings, typically amounting to only one to five hours.

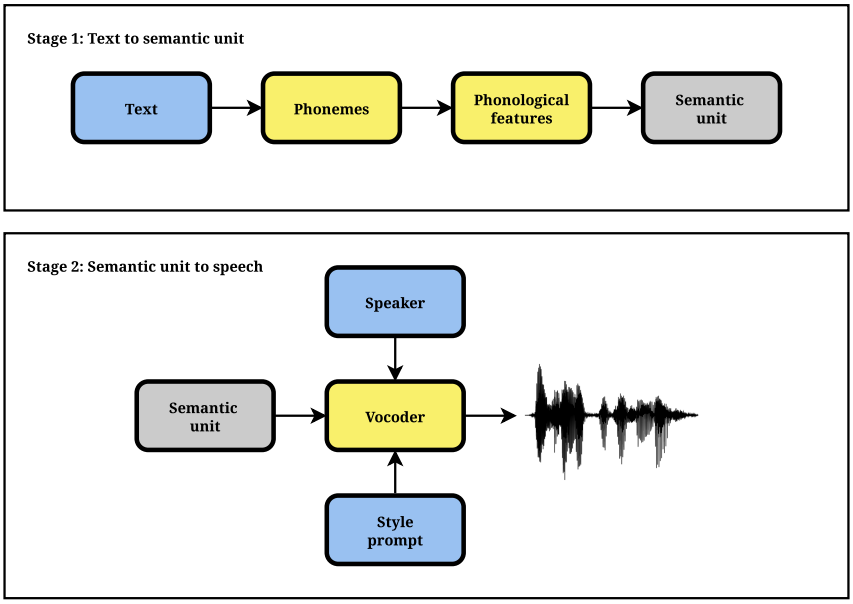

Fig. 1. The two stage, text-to-semantic and semantic-to-waveform, TTS model.

A promising solution involves decomposing speech synthesis into two distinct stages that separate linguistic content from acoustic realization, as depicted in Figure 1. In this architecture, the first stage transforms text through phonemes and phonological features into semantic units that capture linguistic content in a language-independent representation. The second stage then generates waveforms from these semantic units combined with style and speaker information. The choice of semantic unit in this architecture fundamentally affects system performance and cross-linguistic applicability.

While phonemes represent the traditional atomic units of phonological analysis, syllables function as the primary carriers of prosodic information in speech production and perception. Syllables bridge segmental phonology and suprasegmental prosody by providing coherent units that span multiple phonemes while maintaining prosodic structure. This prosodic role suggests that syllable-based

[1] South African Sign Language was recently added as the twelfth official language

semantic units may better capture the suprasegmental properties needed for natural speech synthesis across diverse languages.

However, practical implementation of syllable-based semantic units faces significant challenges in syllabification methodology. Traditional approaches rely on phonological rules derived from linguistic theory, requiring extensive linguistic expertise and potentially failing to capture acoustic realities across diverse language families. Recent advances in self-supervised speech representation learning offer an alternative paradigm for discovering syllabic structure directly from acoustic data, with speaker-disentangled variants providing cleaner syllabic units for cross-linguistic applications.

The linguistic diversity of South Africa provides an ideal testing ground for comparing these approaches. Afrikaans exhibits relatively simple Germanic syllable structures, while Nguni languages present complex organization with tonal properties and click consonants. Sotho-Tswana languages display similar complexity with different tonal patterns, and Tshivenda and Xitsonga represent additional language families with distinct phonological characteristics. This study presents the first systematic comparison of phonological and data-driven syllable discovery methods for multilingual TTS applications. We evaluate both approaches across eleven South African languages using a comprehensive experimental framework that assesses cross-linguistic consistency, acoustic quality, and preservation of linguistically meaningful structure. Our contributions include establishing the relative performance of phonological versus data-driven approaches for cross-linguistic syllable discovery, identifying optimal combinations of unit discovery and feature extraction methods, and providing empirical guidance for implementing syllable-based semantic units in resource-scarce TTS systems.

2 Related Work

The development of syllable-based semantic units for multilingual TTS systems draws from traditional phonological approaches and recent data-driven alternatives. Understanding of both paradigms is needed for effective cross-linguistic syllabification methods.

2.1 Traditional Syllabification Approaches

Traditional syllabification approaches have relied on phonotactic rules and algorithmic methods derived from linguistic theory. The maximum onset principle, which assigns consonants to syllable onsets when possible, forms the basis of many automatic syllabification systems [10, 2]. Rule-based approaches typically incorporate language-specific constraints on syllable structure, including onset complexity limits, coda restrictions, and sonority sequencing principles [13]. These methods employ principles such as maximum onset assignment and language-specific phonotactic constraints to determine syllable boundaries from phoneme sequences.

Machine learning approaches have extended traditional methods by jointly modeling grapheme-to-phoneme conversion and syllabification through probabilistic frameworks [3]. These systems learn to predict both phoneme sequences and syllable boundaries simultaneously, capturing dependencies between orthographic patterns and syllabic structure. Joint probability models can incorporate contextual information and learn language-specific patterns from data rather than relying solely on theoretical rules.

Various machine learning architectures have been applied to syllabification tasks. Conditional Random Fields (CRFs) model syllable boundary prediction by learning from features extracted from character or phoneme sequences [17]. Hidden Markov Models (HMMs) provided early probabilistic sequence models for syllabification. With the rise of deep learning, architectures like Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) networks have been employed to capture long-range dependencies in syllabification tasks [16].

However, these approaches require substantial amounts of annotated training data, which is problematic for resource-scarce languages that lack large syllable-annotated corpora. The supervised nature of these methods necessitates manual annotation of syllable boundaries, a time-consuming process that requires linguistic expertise and is often inconsistent across annotators.

Speech rhythm guided approaches provide an alternative perspective by detecting syllable nuclei based on acoustic properties rather than orthographic patterns. These methods analyze energy patterns, fundamental frequency contours, and spectral characteristics to identify syllabic peaks corresponding to vowel centers [19]. These approaches require no training data as they rely purely on signal processing algorithms and acoustic heuristics. However, they are sensitive to signal quality and recording conditions, and depend on manually tuned heuristics that may not generalize well across different languages or speaking styles.

While rule-based methods work well for languages with established phonological descriptions, they require extensive linguistic expertise and may not adequately capture acoustic realities across diverse language families. Many languages lack established syllabification algorithms, and existing approaches often produce inconsistent results when applied to morphologically complex languages with intricate phonological processes. The rule-based nature of these systems means they cannot adapt to patterns that emerge from acoustic data but are not captured by theoretical phonological descriptions.

Forced alignment techniques provide a complementary method for obtaining time-aligned phoneme boundaries from speech and text pairs. HMMs have traditionally been used for this task, with systems like HTK and Kaldi providing robust implementations [18, 15]. More recent approaches incorporate neural acoustic models and attention mechanisms to improve alignment accuracy. The quality of forced alignment depends on the acoustic model and pronunciation dictionary used, with particular challenges arising for languages with limited training data or complex morphophonological processes. These phoneme-level

alignments can then be converted to syllable boundaries using phonological rules, creating a pipeline from raw audio to syllabic structure.

2.2 Self-Supervised Speech Representation Learning

Self-supervised speech representation learning has emerged as an effective paradigm for discovering linguistic structure from raw audio without explicit labels. These approaches learn representations through pretext tasks that require no manual annotation, enabling the use of large amounts of unlabeled speech data. Early contrastive methods like wav2vec 2.0 learn representations by distinguishing between true and false future frames [1]. These models capture phonetic distinctions and demonstrate improved performance on downstream tasks compared to traditional acoustic features.

The introduction of Bidirectional Encoder Representations from Transformers (BERT) in natural language processing demonstrated the effectiveness of masked prediction objectives for learning contextual representations [5]. Building on this foundation, Hidden-Unit BERT (HuBERT) adapted the masked prediction approach to speech representation learning [7]. The model learns by predicting discrete acoustic units for masked portions of the input, iteratively refining the target representations through clustering. This approach has shown strong performance across multiple speech processing tasks and provides representations that capture both phonetic and phonological information.

Extensions to HuBERT have explored various aspects of speech representation, including multilingual training, improved clustering methods, and incorporation of additional modalities. These developments have demonstrated that self-supervised models can learn hierarchical representations that capture multiple levels of linguistic organization, from phonetic features to higher-level prosodic patterns. The effectiveness of these approaches has motivated investigation into what linguistic knowledge emerges naturally from self-supervised objectives applied to speech data.

2.3 Data-Driven Syllable Discovery

Speaker-Disentangled HuBERT (SD-HuBERT) specifically targets sentence-level representation learning through self-distillation, where a teacher model provides targets for learning aggregated sentence representations [4]. This approach induces emergent syllabic organization in intermediate layer representations, with clear boundaries appearing between syllabic units without explicit syllable supervision during training. Analysis of these representations reveals temporal segmentation patterns that correspond to syllable-like units, suggesting that syllabic organization emerges naturally from speech data when appropriate learning objectives are applied.

However, analysis has revealed that sentence-level representations correlate strongly with speaker characteristics rather than purely linguistic content. The Self-Supervised Speaker-Separated Syllable HuBERT (S5-HuBERT) addresses this limitation by explicitly separating speaker identity from linguistic structure

during the fine-tuning process [11]. The resulting S5-HuBERT units demonstrate improved consistency across speakers while maintaining syllabic temporal organization. These units are extracted through clustering of speaker-disentangled representations, producing discrete symbols that can serve as intermediate representations for speech synthesis applications.

The speaker disentanglement process involves fine-tuning pre-trained HuBERT models with additional objectives that encourage speaker-invariant representations while preserving linguistic content. This approach produces dense feature vectors that capture phonological and prosodic information without speaker-specific characteristics. The clustering of these features into discrete units creates what we term `S5-HuBERT units`, following the methodology of Komatsu et al. [11]. These units represent a data-driven approach to syllable discovery that requires no linguistic expertise or language-specific rules, potentially providing universal acoustic patterns that transcend language-specific phonological descriptions.

2.4 Cross-Linguistic Evaluation Gap

While both traditional and data-driven approaches to syllable discovery have been developed and evaluated, systematic cross-linguistic comparisons remain limited. Most evaluations focus on single languages or small sets of closely related languages, making it difficult to assess the generalizability of different approaches across diverse linguistic contexts. Traditional syllabification methods have been primarily developed and tested on well-resourced languages like English, with limited validation on typologically diverse language families.

Similarly, data-driven methods have been predominantly evaluated on English and a few other major languages, with limited exploration of their performance across language families with different phonological characteristics. The emergence of syllabic organization in self-supervised models has been demonstrated primarily through visualization and downstream task performance, rather than through systematic comparison with established syllabification methods across multiple languages.

This evaluation gap is particularly pronounced for African languages, which present unique challenges including complex morphophonological processes, tonal systems, and click consonants. The lack of systematic cross-linguistic evaluation makes it difficult to determine whether data-driven methods can capture the linguistic diversity present in these language families or whether traditional rule-based approaches remain necessary for accurate syllabification. Understanding the relative performance of phonological versus data-driven syllabification approaches is needed for implementing effective two-stage TTS architectures in multilingual settings.

2.5 Multilingual TTS Development

Syllable-based approaches to speech synthesis have been motivated by the observation that syllables capture important prosodic and coarticulatory patterns [6].

Unit selection synthesis systems have successfully used syllables as basic units, though they require large databases to achieve good coverage [8]. Parametric approaches have incorporated syllable-level features for duration and prosody modeling, showing improvements over purely phoneme-based systems. Recent neural approaches have explored syllable-level sequence-to-sequence models and have demonstrated that syllabic representations can improve naturalness and prosodic accuracy in synthetic speech.

In metrical phonology frameworks, syllables serve as the foundation for hierarchical stress patterns, with prominence relationships operating primarily through syllabic infrastructure. When a word receives emphasis in connected speech, this prominence manifests through the lexically stressed syllable rather than being distributed across individual phonemes. This organization reflects the fact that prosodic structure influences segmental realization through effects such as syllable-final lengthening, stress-dependent vowel reduction, and temporal coordination of articulatory gestures within syllabic boundaries.

The development of multilingual TTS systems faces additional challenges in unit selection and cross-linguistic transfer. Traditional approaches require separate syllabification systems for each target language, increasing development complexity and requiring language-specific expertise. Data-driven approaches offer the potential for language-independent syllable discovery, but their effectiveness across diverse language families remains underexplored. The choice between phonological and data-driven syllabification methods affects not only synthesis quality but also the feasibility of developing systems for resource-scarce languages.

Two-stage architectures that separate linguistic content from acoustic realization have shown promise for multilingual applications, but the optimal choice of semantic units remains an open question. The first stage must produce representations that capture linguistic content consistently across languages while remaining independent of speaker characteristics. The second stage then maps these representations to acoustic realizations, potentially enabling cross-linguistic transfer of acoustic modeling components.

2.6 South African Language Processing Challenges

Speech processing for African languages faces unique challenges due to the diversity of phonological systems and limited availability of linguistic resources. Indigenous African languages exhibit complex morphophonological processes, tonal systems, and syllable structures that differ significantly from well-studied Indo-European languages [14]. The linguistic diversity of South Africa spans multiple language families: Afrikaans exhibits relatively simple Germanic syllable structures, while Nguni languages (isiZulu, isiXhosa, isiNdebele, siSwati) present complex organization with tonal properties, click consonants, and intricate morphophonological processes.

Click consonants in Nguni languages present particular challenges for both acoustic modeling and syllabification, as these segments involve complex articulatory coordination that may not align with traditional syllable boundary

principles. Sotho-Tswana languages (Sepedi, Sesotho, Setswana) display similar complexity with different tonal patterns, and Tshivenda and Xitsonga represent additional language families with distinct phonological characteristics. This diversity challenges both traditional syllabification algorithms and data-driven approaches, providing robust evaluation conditions for cross-linguistic syllable discovery.

Automatic speech recognition and synthesis systems for these languages often struggle with tonal processing, morphological complexity, and limited training data. The agglutinative nature of many indigenous African languages results in complex word structures that challenge traditional text processing pipelines. Tonal patterns interact with syllabic organization in ways that may not be captured by phoneme-level representations, suggesting potential advantages for syllable-based approaches.

Recent work has explored cross-lingual transfer methods and self-supervised approaches as ways to address resource constraints [12], though syllable-level analysis specifically has received limited attention in this context. The linguistic diversity present in South Africa provides both challenges and opportunities for evaluating syllabification methods, as it includes representatives from multiple language families with distinct phonological characteristics.

3 Methodology

We compare algorithmic and data-driven syllable discovery methods across eleven South African languages using a 2×2 experimental design. The comparison evaluates both unit discovery approaches and feature extraction methods to assess cross-linguistic syllabification performance.

3.1 Dataset

The dataset contains speech recordings from eleven South African languages spanning five language families. Afrikaans and English represent the Germanic family. The Nguni family includes IsiZulu, IsiXhosa, IsiNdebele, and SiSwati, which feature complex tonal systems and click consonants. The Sotho-Tswana family encompasses Sepedi, Sesotho, and Setswana, characterized by agglutinative morphology and distinct tonal patterns. Tshivenda and Xitsonga represent the Venda and Tsonga families respectively. The indigenous African languages exhibit complex morphophonological processes and syllable structures that differ from Indo-European languages.

The speech data consists of internal datasets recorded previously. Sixteen speakers across the eleven languages provide recordings totaling approximately 70 hours[2]. Table 1 shows speaker demographics and data quantities for each language. Multiple speakers per language enable analysis of within-language variation where data permits. The dataset includes both male and female speakers for examining gender effects on syllabification consistency.

[2] The isiZulu speaker "Johnny" also provided English recordings.

Table 1. Speaker summary sorted by language and speaker

Language	Speaker	Number of WAVs	Total Duration (HH:MM:SS)
Afrikaans	Antowan	3765	04:45:04
	Nadia	4757	06:53:04
English	Johnny	477	00:23:16
	Mel	4083	05:20:00
	Nathan	3274	04:21:17
IsiNdebele	Cedric	1452	03:58:07
IsiXhosa	Dumza	1380	02:56:57
	Kholeka	1704	05:44:26
IsiZulu	Bongiwe	1624	05:02:13
	Johnny	1322	02:29:08
Sesotho	Palesa	1314	02:29:15
Setswana	Didi	2276	04:30:25
Sepedi	Marcia	2257	04:44:53
	Roelf	1778	03:48:38
SiSwati	Sakhile	1492	05:06:43
Tshivenda	Eugene	2409	04:16:23
Xitsonga	Hlamalani	916	01:47:59

All recordings were captured at 16 kHz sampling rate in controlled acoustic environments. Text transcriptions accompany all recordings, enabling extraction of algorithmic syllable boundaries through forced alignment and rule-based G2P methods. The limited data per language reflects real-world constraints for developing TTS systems in resource-scarce contexts.

3.2 Experimental Design

The experimental framework uses a 2×2 design comparing unit discovery methods and feature extraction approaches. The two dimensions are: (1) Unit Discovery Method: algorithmic syllables versus S5-HuBERT units, and (2) Feature Extraction: MFCC features versus S5-HuBERT dense features. This yields four experimental phases for evaluating different syllabification approaches.

Phase 1 combines algorithmic syllables with MFCC features, representing traditional syllable-based speech processing. Algorithmic syllables are extracted using rule-based algorithms applied to forced alignment outputs. MFCC features use standard parametrization with 12 coefficients. Similarity computation employs Dynamic Time Warping (DTW) with Mel Cepstral Distortion (MCD) as the distance metric.

Phase 2 pairs algorithmic syllables with S5-HuBERT dense features, examining whether modern feature representations improve traditional syllabification. The same algorithmic syllable boundaries are used, but acoustic similarity is computed using cosine similarity between S5-HuBERT dense feature vectors.

Phase 3 combines S5-HuBERT units with MFCC features, evaluating data-driven syllable discovery using traditional acoustic features. S5-HuBERT units are extracted through clustering of speaker-disentangled HuBERT representations. Acoustic similarity between units uses DTW and MCD on MFCC features.

Phase 4 pairs S5-HuBERT units with S5-HuBERT dense features, representing the fully data-driven approach. Both syllable discovery and feature extraction rely on self-supervised representations, with similarity computed using cosine distance between dense feature vectors.

3.3 Syllable Unit Extraction

Algorithmic syllables are extracted through forced alignment of speech and text using language-specific acoustic models. Phoneme-level alignments are converted to syllable boundaries using the maximum onset principle and language-specific phonotactic constraints. Established algorithms are adapted for each language family, accounting for differences in consonant cluster restrictions and vowel systems.

S5-HuBERT units are extracted using the pre-trained speaker-disentangled HuBERT model from the original authors. The model directly outputs discrete unit IDs and corresponding dense feature representations for input speech segments. The unit IDs provide syllabic segmentation while the dense features capture linguistic content without speaker-specific characteristics.

Unit extraction quality was assessed through diagnostic procedures. Algorithmic syllable boundaries were manually inspected for a subset of utterances to verify consistency with linguistic principles. S5-HuBERT unit extraction was evaluated by analyzing unit duration distributions across languages and speakers to identify potential segmentation errors.

3.4 Feature Extraction and Similarity Computation

MFCC features use identical parametrization across all languages for direct comparison. Feature vectors include 12 mel-frequency cepstral coefficients. Frame-level features are computed every 10ms with 25ms windows for detailed temporal resolution.

S5-HuBERT dense features are extracted directly from the pre-trained speaker-disentangled HuBERT model. The model outputs dense feature representations that capture phonetic content and prosodic characteristics without speaker-specific information.

Similarity computation methods are tailored to feature types. DTW alignment with MCD distance provides temporally flexible comparison for MFCC features, accounting for natural variation in syllable duration and speaking rate. Cosine similarity between dense feature vectors enables direct comparison without requiring temporal alignment. Similarity thresholds are established through cross-validation to optimize performance across languages.

3.5 Analysis Framework

Cross-linguistic consistency is quantified using the *Jaccard similarity coefficient* [9], a widely adopted measure of set similarity. It is defined as the ratio of the size of the intersection to the union of two sets. Formally, for two sets A and B:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

This coefficient ranges from 0, indicating no overlap, to 1, indicating identical sets. In this study, higher Jaccard similarity between unit inventories across languages and language families reflects greater cross-linguistic generalizability of discovered units.

Language family analysis examines within-family versus across-family unit sharing patterns to assess whether syllabification methods capture genealogical relationships. Family effect ratios quantify the relative strength of within-family clustering compared to across-family similarity.

Speaker independence analysis compares similarity distributions within and across speakers to evaluate acoustic consistency. Within-speaker similarities indicate consistency of units for individual voices, while across-speaker similarities reflect generalizability across different speakers. Speaker independence ratios near 1.0 indicate optimal speaker-invariant performance.

Statistical significance testing employs appropriate non-parametric methods given the characteristics of the similarity data. Unit inventory analysis tracks the distribution of universal, language-specific, and family-specific units to characterize cross-linguistic sharing patterns.

4 Results

Cross-linguistic analysis across eleven South African languages reveals fundamental differences between algorithmic and S5-HuBERT syllable discovery approaches. Data-driven methods achieve substantially better generalization while maintaining acoustic consistency across speakers.

4.1 Universal versus Language-Specific Units

The two approaches differ markedly in their discovery of universal syllabic patterns. Algorithmic methods found 42,093 unique syllables across 70 hours of speech, with significant language specificity. Only 10 syllables (0.02%) appeared across all eleven languages, while 34,944 syllables (83.02%) were language-specific. This distribution reflects the theoretical basis of rule-based syllabification in language-particular phonological principles.

SS5-HuBERT methods discovered a total of 14,165 units, of which 1,640 (11.58%) were universal units. Language-specific units accounted for 1,483 (10.47%) of the inventory. The reduced total unit count, coupled with a higher degree of universal sharing, suggests that data-driven methods effectively identify acoustic patterns that cross individual language boundaries.

4.2 Language Family Structure

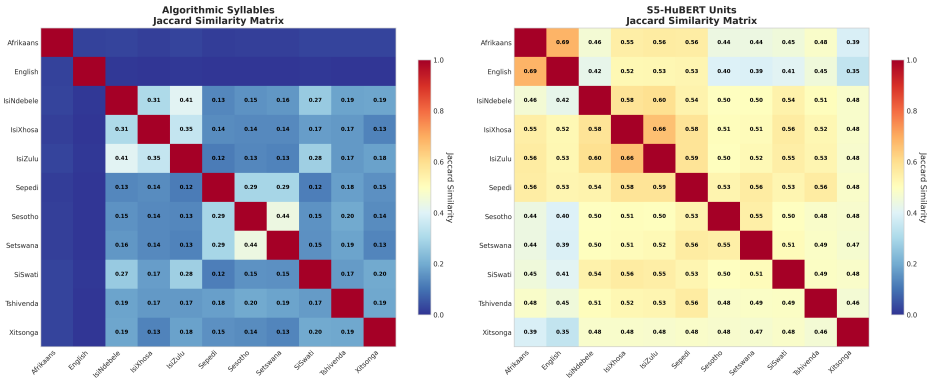

Fig. 2. Cross-linguistic Jaccard similarity matrices comparing unit sharing patterns.

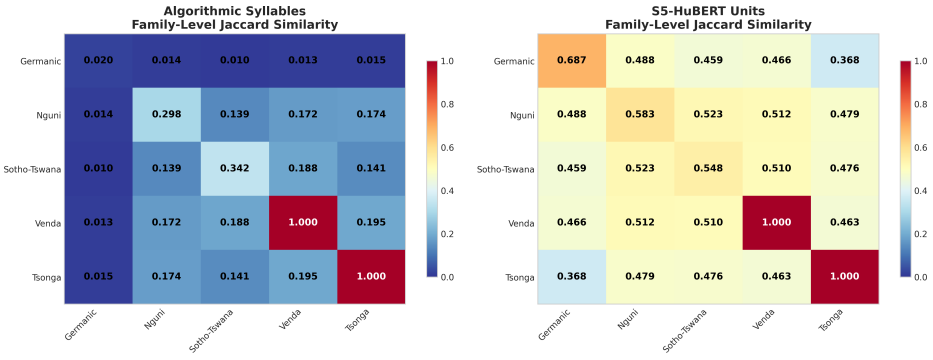

Fig. 3. Family-level Jaccard similarity matrices aggregating within-family and across-family unit sharing patterns.

Both approaches capture language family relationships but with different sensitivities. Algorithmic syllables exhibit strong family clustering, with within-family similarity (0.2835) substantially exceeding across-family similarity (0.0991). The resulting family effect ratio of 2.86 indicates that syllables are nearly three times more similar within language families than across them ($p < 0.001$).

S5-HuBERT units show weaker family effects despite higher overall similarities. Within-family similarity (0.5832) exceeds across-family similarity (0.4887), producing a family effect ratio of 1.19 ($p < 0.001$). This compressed ratio indicates that data-driven units maintain more consistent properties across diverse language families.

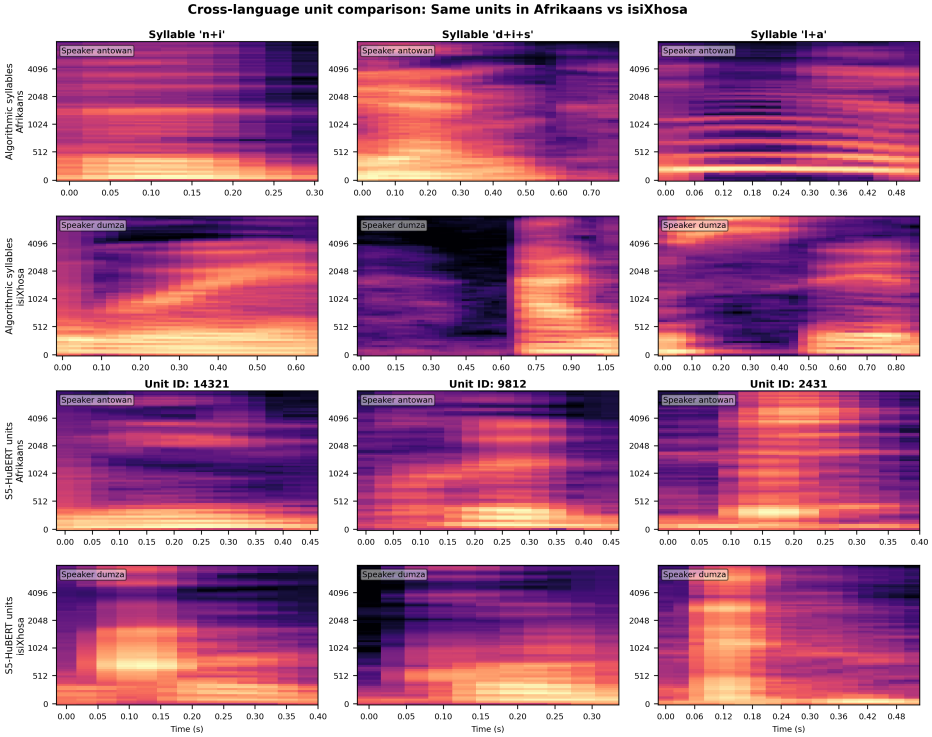

Fig. 4. Cross-language spectrogram comparison between frequent algorithmic syllables and S5-HuBERT units in Afrikaans and isiXhosa.

Figure 2 illustrates these cross-linguistic sharing patterns. The algorithmic syllables matrix (left) shows stark family-based clustering, with Germanic languages (Afrikaans, English) exhibiting minimal unit sharing with Bantu language families, appearing as dark regions indicating near-zero Jaccard similarities. This reflects the fundamental phonological differences between Germanic and indigenous African languages syllable structures. In contrast, S5-HuBERT units (right) demonstrate more uniform similarities across all language pairs, including substantial cross-family sharing. The family-level aggregation in Figure 3 further emphasizes the difference in family effect magnitudes between the two approaches. Figure 4 gives a visual comparison of the spectrograms of the same algorithmic syllables in Afrikaans and isiXhosa in the top two rows and the same S5-HuBERT units in Afrikaans and isiXhosa in the bottom two rows.

4.3 Speaker Generalization Performance

This pattern demonstrates that speaker-disentangled representations improve acoustic consistency regardless of unit type. The best performance emerges when both unit discovery and feature extraction employ data-driven methods, suggesting effective separation of linguistic content from speaker characteristics.

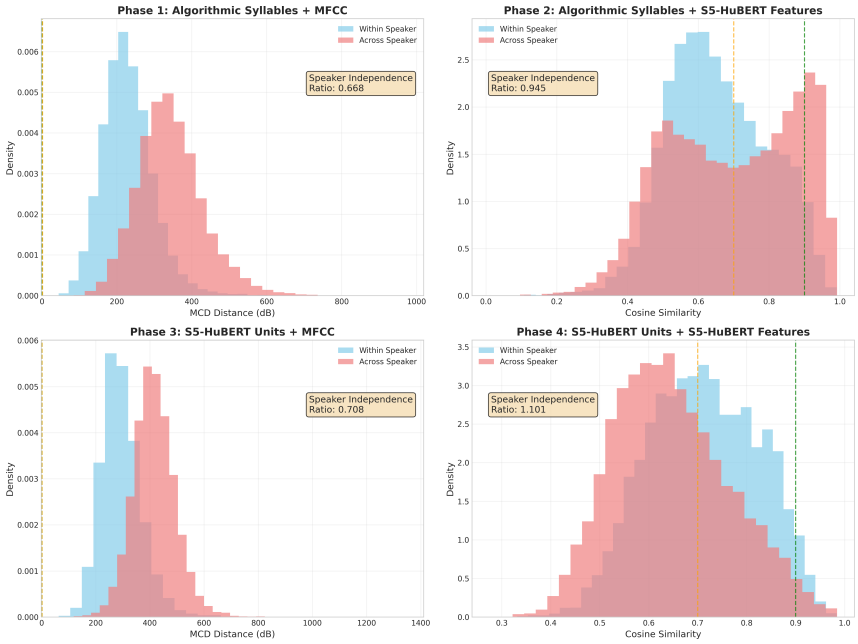

Fig. 5. Acoustic quality distributions across experimental phases showing within-speaker versus across-speaker similarity patterns.

Figure 5 illustrates these acoustic quality patterns across all four experimental phases, with Phase 4 achieving near-optimal speaker independence.

Acoustic consistency across speakers varies systematically across the four experimental phases. Speaker independence ratios measure how similarly units behave across different speakers, with values near 1.0 indicating optimal generalization.

Traditional algorithmic syllables with MFCC features show poor speaker generalization (ratio: 0.668). Replacing MFCC features with S5-HuBERT dense representations improves performance substantially (ratio: 0.945). S5-HuBERT units paired with MFCC features achieve moderate speaker independence (ratio: 0.708). The fully data-driven combination of S5-HuBERT units and dense features achieves optimal performance (ratio: 1.101), with across-speaker similarities slightly exceeding within-speaker similarities.

This pattern demonstrates that speaker-disentangled representations improve acoustic consistency regardless of unit type. The best performance emerges when both unit discovery and feature extraction employ data-driven methods, suggesting effective separation of linguistic content from speaker characteristics. Figure 5 illustrates these acoustic quality patterns across all four experimental phases, with Phase 4 achieving near-optimal speaker independence.

5 Discussion

Algorithmic syllabification yields units that align closely with theoretical phonological boundaries, producing predominantly language-specific inventories that reflect language-particular phonological principles. In contrast, S5-HuBERT units exhibit a distinct pattern, capturing a broader set of cross-linguistic units that are not constrained by conventional syllable segmentation. Remarkably, these results are achieved despite S5-HuBERT having been trained solely on English data, suggesting that self-supervised models can access universal acoustic–phonetic patterns shared across human languages, albeit with reduced linguistic interpretability compared to rule-based methods.

Both approaches capture language family relationships, but with differing sensitivities. Algorithmic syllables produce strong family-level clustering, while S5-HuBERT units show attenuated family effects, emphasizing acoustic–phonetic similarity over genealogical affiliation.

Analysis of speaker independence reveals systematic improvement across experimental phases. The fully data-driven approach achieves near-optimal performance, demonstrating that speaker-disentangled representations enhance acoustic consistency in both unit discovery and feature extraction, yielding normalized representations that improve cross-speaker robustness.

The superior cross-linguistic consistency and speaker independence of S5-HuBERT units highlight their potential as universal semantic units for multilingual TTS architectures. However, this comes with a trade-off, while S5-HuBERT units excel in acoustic generality, algorithmic syllables maintain transparent links to phonological theory, enabling more direct integration with existing linguistic resources.

These findings are subject to several limitations. The dataset covers a limited portion of global linguistic diversity, focusing on South African languages that may not generalize to language families with different phonological characteristics. The use of forced alignment for algorithmic syllable boundaries introduces potential systematic errors, and the speaker-disentangled S5-HuBERT model represents only one possible approach to self-supervised syllable discovery. Future work should examine downstream TTS performance, extend evaluation to a broader set of languages, and explore hybrid models that combine phonological structure with self-supervised acoustic universals.

References

1. Baevski, A., Zhou, H., Mohamed, A., Auli, M.: wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In: Advances in Neural Information Processing Systems (NeurIPS 2020). vol. 33, pp. 12449–12460 (2020)
2. Bartlett, S., Kondrak, G., Cherry, C.: Automatic syllabification with structured SVMs for letter-to-phoneme conversion. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics pp. 568–576 (2008)
3. Bisani, M., Ney, H.: Joint-sequence models for grapheme-to-phoneme conversion. Speech communication **50**(5), 434–451 (2008)
4. Cho, C.J., Mohamed, A., Li, S.W., Black, A.W., Anumanchipalli, G.K.: Sd-hubert: Sentence-level self-distillation induces syllabic organization in hubert. In: 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 12076–12080. IEEE, Seoul, Korea (Apr 2024)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). vol. 1, pp. 4171–4186. Association for Computational Linguistics, Minnesota, USA (Jun 2019), https://aclanthology.org/N19-1000/
6. Donovan, R.E.: Trainable speech synthesis. Ph.D. thesis, University of Cambridge (1996)
7. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3451–3460 (2021)
8. Hunt, A.J., Black, A.W.: Unit selection in a concatenative speech synthesis system using a large speech database. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). vol. 1, pp. 373–376. IEEE, Atlanta, Georgia, USA (May 1996)
9. Jaccard, P.: Étude comparative de la distribution florale dans une portion des alpes et des jura. Bull Soc Vaudoise Sci Nat **37**, 547–579 (1901)
10. Kahn, D.: Syllable-based generalizations in English phonology. Ph.D. thesis, Massachusetts Institute of Technology (1976)
11. Komatsu, R., Shinozaki, T.: Self-supervised syllable discovery based on speaker-disentangled hubert. In: 2024 IEEE Spoken Language Technology Workshop (SLT). pp. 1131–1136. IEEE, Macao, China (Dec 2024)
12. Louw, J.A.: Cross-lingual transfer using phonological features for resource-scarce text-to-speech. In: Proceedings of the 12$^{\text{th}}$ Speech Synthesis Workshop (SSW). Grenoble, France (9 2023)
13. Müller, K.: Automatic Detection of Syllable Boundaries Combining the Advantages of Treebank and Bracketed Corpora Training. In: Proceedings of the 39$^{\text{th}}$ Annual Meeting of the Association for Computational Linguistics. pp. 410–417. Association for Computational Linguistics, Toulouse, France (Jul 2001), https://aclanthology.org/P01-1053/
14. Nurse, D., Philippson, G.: The Bantu languages. Routledge, London (2008)
15. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al.: The Kaldi speech recognition toolkit. In: IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU 2011). vol. 1, pp. 5–1. IEEE, Hawaii,USA (Dec 2011)

16. Rao, K.S.: Modeling supra-segmental features of syllables using neural networks. In: Speech, audio, image and biomedical signal processing using neural networks, pp. 71–95. Springer (2008)
17. Rogova, K., Demuynck, K., Van Compernolle, D.: Automatic syllabification using segmental conditional random fields. Computational Linguistics in the Netherlands Journal **3**, 34–48 (2013)
18. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK book. Manual 3.4, Cambridge University Engineering Department, Cambridge, USA (Mar 2002)
19. Zhang, Y., Glass, J.R.: Speech rhythm guided syllable nuclei detection. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 3797–3800. IEEE, Taipei, Taiwan (may 2009)