Interspeech 2026
27 September - 1 October 2026, Sydney, Australia

ZeroSyl: Simple Zero-Resource Syllable Tokenization for Spoken Language Modeling

Nicol Visser, Simon Malan, Danel Slabbert, Herman Kamper

Electrical and Electronic Engineering, Stellenbosch University, South Africa
{16986431, 24227013, 24051055, kamperh}@sun.ac.za

Abstract

Pure speech language models aim to learn language directly from raw audio without textual resources. A key challenge is that discrete tokens from self-supervised speech encoders result in excessively long sequences, motivating recent work on syllable-like units. However, methods like Sylber and SyllableLM rely on intricate multi-stage training pipelines. We propose ZeroSyl, a simple training-free method to extract syllable boundaries and embeddings directly from a frozen WavLM model. Using L2 norms of features in WavLM's intermediate layers, ZeroSyl achieves competitive syllable segmentation performance. The resulting segments are mean-pooled, discretized using K-means, and used to train a language model. ZeroSyl outperforms prior syllabic tokenizers across lexical, syntactic, and narrative benchmarks. Scaling experiments show that while finer-grained units are beneficial for lexical tasks, our discovered syllabic units exhibit better scaling behavior for syntactic modeling.
Index Terms: spoken language modeling, speech tokenization

1. Introduction

Advances in self-supervised learning (SSL) [1] have made it possible to train language models directly on audio data: instead of using tokenized text, a causal language model is trained on tokens derived from an SSL model [2]. Without relying on any textual resources, this framework would enable natural language processing for languages lacking the massive speech-text datasets required by traditional speech-driven technology. These pure speech language models showed early potential [3], but subsequent progress has slowed.

The decline in progress is most evident in syntactic modeling, where performance has plateaued since 2023. AudioLM [4] remains the state-of-the-art, yet it relies on a resource-intensive framework: a high-capacity language model trained on high-bitrate tokens from w2v-BERT XL [5]. Moreover, recent work indicates that pure speech modeling scales less favorably than text [6], suggesting that simply increasing model size and data volume will not solve the problem. This has resulted in a shift toward discovering more effective speech units in order to close the performance gap with textual models of comparable size [7].

One hypothesis is that the fine granularity of standard speech tokens creates a bottleneck [8]. Because discrete speech tokens are extracted at a much higher rate than text tokens, the resulting sequences are substantially longer, making it difficult to model long-range dependencies. To mitigate this, studies have attempted to utilize larger phone-, syllable-, or word-like units [8–11]. Syllable-based modeling has proven to be particularly successful [10, 11]. However, existing syllabic tokenizers rely on complex multi-stage pipelines where an SSL model must be fine-tuned with a specialized objective.

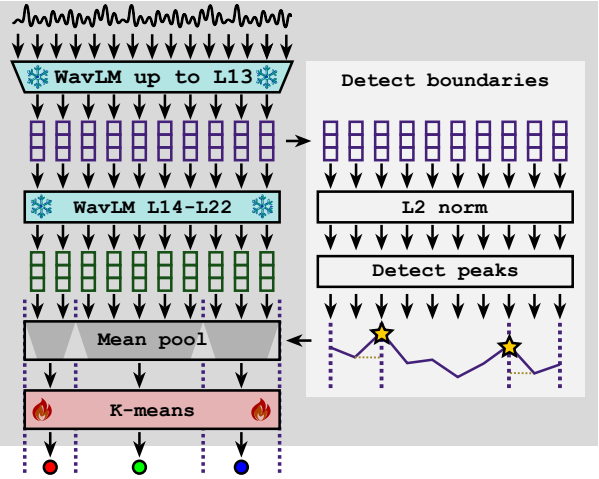


Figure 1: *ZeroSyl detects syllabic boundaries using prominence-based peak detection on features from layer 13 of a frozen WavLM. It then mean pools semantic features from layer 22 within the discovered boundaries and clusters these using spherical K-means. A language model is trained on these cluster IDs.*

In this paper, we propose an unsupervised approach to syllable tokenization that does not require the complex methods of previous work. Figure 1 shows our simple pipeline. We find that high-quality syllabic boundaries can be extracted by simply applying peak detection to the L2 norms of frozen WavLM [12] features. We then mean-pool and cluster features within these boundaries to generate a discrete vocabulary. A symbolic language model [13] is trained on the resulting sequences. This approach, which we call ZeroSyl, surpasses the syntactic (sBLIMP) performance of comparable syllabic models and achieves superior performance on the Topic StoryCloze (tSC) narrative benchmark. It also outperforms syllabic systems on the lexical sWUGGY benchmark. But in scaling experiments, it does not surpass the performance of a system utilizing fine-grained units [7]. Nevertheless, our results demonstrate that high-quality syllable discovery is possible without complex multi-stage training pipelines, paving a simple and effective path for future research in spoken language modeling. Models and code available at: https://github.com/nicolvisser/ZeroSyl/

2. Related work

This work focuses on *pure* speech language models, where text is not used or incorporated at all [14]. Standard benchmarks have been developed that assess how well these models capture lexical, syntactic, and semantic properties [15, 16].

10.21437/Interspeech.2026-315

To date, AudioLM [4] achieves state-of-the-art performance in syntactic modeling. However, since it uses a 25 Hz-tokenizer that produces long sequences, a large language model (300M parameters) is subsequently required to capture long-range dependencies. SpidR [7] is a recent SSL-based approach that is more lightweight. It uses online clustering and a novel training objective to improve the SSL representations and resulting speech units. While SpidR achieves the best lexical modeling scores, its units remain close to the frame level. As with AudioLM, the sequence lengths are therefore very long, and SpidR has not surpassed AudioLM's syntactic scores.

Instead of directly discretizing SSL features, two noteworthy approaches have turned to syllable units to deal with the long sequence lengths. Sylber [10] utilizes a sentence-level distillation bjective [17] to finetune HuBERT [18]. After distillation, syllable boundaries are accessible using cosine similarity between the framewise features. SyllableLM [11] identifies boundaries by analyzing the HuBERT loss objective. The idea is that fully masking a syllable results in a distinct disruption of the loss compared to partial masking. Syllable boundaries are predicted by monitoring the loss as a masked interval sweeps across the utterance. To reduce computational load at inference time, both approaches distill the boundary information: they train an encoder to predict piecewise-constant targets, derived by mean-pooling the teacher model's embeddings over the discovered segments. Both approaches therefore require multiple training stages with specially designed architectures and objectives.

The syllabic methodology shows clear promise for spoken language modeling, but we want to see if a simpler approach is possible. For this, we build on a simple approach for unsupervised *word* discovery [19]: prominent peaks in a dissimilarity curve (typically cosine distance) between adjacent features [20] from a frozen SSL model are taken to indicate word boundaries. So far, this approach has been optimized for lexical units [19, 21] and underperforms when used directly to find syllabic units [22].

3. Models and methodology

Our work builds on prominence-based segmentation [19]. However, where previous work used the cosine similarity between adjacent frames to find word boundaries, we find that the L2 norm of the hidden representations themselves provides a signal corresponding to syllabic structure. In contrast to Sylber [10] and SyllableLM [11], our method is much simpler and does not require further training of the SSL encoder.

3.1. Boundary detection

For a given speech utterance, we extract framewise embeddings $H_{1:T}^{[13]} = [h_1^{[13]}, h_2^{[13]}, \ldots, h_T^{[13]}]$ from layer 13 of WavLM Large [12]. We compute the L2 norm for each frame, $n_t = \|h_t^{[13]}\|_2$, and apply a 3-point moving average filter to reduce high-frequency noise. We then apply prominence-based peak detection on the smoothed signal: a peak is identified as a boundary if it has a prominence of at least $\delta = 0.45\sigma$, where σ is the standard deviation of the signal n_t.

Example predictions are shown in Figure 2. R-value and token-F1 scores on development data (described later) are best using layer 13, a filter length of 3, and a prominence of 0.45σ.

3.2. Tokenization

Once the boundaries $[b_1, b_2, ..., b_N]$ are established, we extract semantic features from layer 22 of WavLM Large and apply

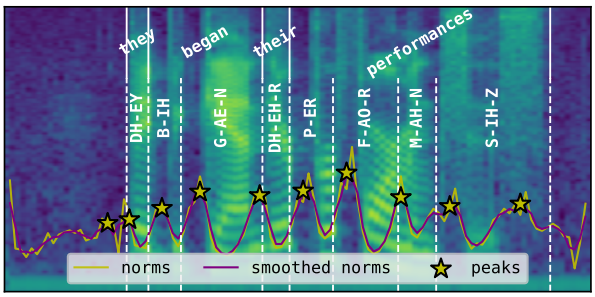


Figure 2: *Peak detection on the smoothed L2 norms of framewise features gives our predicted syllable boundaries. The dashed lines indicate the ground truth syllables.*

mean-pooling within each segment:

$$e_j = \text{MeanPool}(H_{b_j:b_{j+1}}^{[22]}) \quad (1)$$

We select this layer because higher WavLM layers are known to capture richer semantic information than those used for boundary detection [23]. Empirically, we find that layer 22 yields the highest mutual information with syllable labels after clustering.

The resulting embeddings e_j are clustered using spherical K-means with K-means++ initialization. The K-means model is trained until convergence on 100 hours of LibriSpeech [24] using the `faiss` library [25]. We set the vocabulary size to $K = 10\,000$ following Sylber [10], which reports this as the optimal size for syntactic modeling.

After clustering, multiple centroids correspond to silences. We find that language modeling performance improves when these entries are collapsed to a single vocabulary item (see Table 2 below). We do this in an unsupervised manner by performing hierarchical clustering on the centroids. From informal inspection, we know that the two main branches in agglomerative hierarchical clustering correspond to silences and non-silences, respectively. We pick the smaller branch, which represents silences, and map these items to one vocabulary item. This reduces the vocabulary size from 10 000 to 9 116.

3.3. Language modeling

We train a causal language model on the discovered syllabic units. Following prior work [7, 11], we use the OPT-125M architecture [13] with a batch size of 81 920 tokens, and a context length of 2 048. The learning rate increases linearly from 0 to $2 \cdot 10^{-4}$ for the first 8% of steps followed by a cosine annealing schedule. For our comparative experiments in Section 4, we train the ZeroSyl language model on a single GPU on 6k hours of Libri-Light [26] for 25k steps. For our final scaled experiments in Section 5, we train on the full 60k hours for 200k steps.

3.4. Implementation of previous systems

For our comparisons against Sylber [10] and SyllableLM [11], we strictly follow the code implementations released by the authors. Since Sylber did not release a clustering model, we train a spherical K-means model with $K = 10\,000$ on their embeddings. As per their specifications [10], we include an explicit silence token when more than seven frames correspond to silence. For SyllableLM, we utilize the official extraction pipeline, including agglomerative clustering and deduplication. SyllableLM has multiple variants. To ensure the strongest possible baseline, we consider all configurations (5.0, 6.25, and 8.33 Hz) and report the best result. We train OPT-125M [13] language models on top

Table 1: *Boundary detection performance on the combined test sets of LibriSpeech. A boundary tolerance of 50 ms is allowed. Pr: precision, Re: recall, R: R-value, OS: over segmentation.*

	Boundary					Token		
	Pr	Re	F1	OS	R	Pr	Re	F1
SyllableLM 5.0 Hz	**71**	79	**75**	10	**77**	**54**	58	**56**
SyllableLM 6.25 Hz	62	87	72	41	59	44	**59**	50
SyllableLM 8.33 Hz	50	**91**	65	81	27	31	52	39
Sylber	65	74	69	13	71	48	54	51
PromSeg	62	65	64	**4**	69	40	41	41
ZeroSyl	69	75	72	10	75	52	56	54

of the resulting units, following the same setup as for ZeroSyl above, except for SyllableLM, which is trained on two GPUs for 50% more steps to account for the higher token rate. For scaled comparison to SpidR [7], we use results directly from the paper.

4. Experiments: Syllabic system comparison

We perform a range of experiments to benchmark ZeroSyl's units. In this section, we focus on systems that operate at the syllable level; comparisons against other systems are reserved for Section 5. We therefore compare ZeroSyl against Sylber [10] and SyllableLM [11] when trained on 6k hours of Libri-Light. This is a reasonable training set size enabling reliable comparisons without extensive training time. Before describing language modeling results, we first perform intrinsic evaluations of the predicted syllable boundaries and discovered syllabic clusters.

4.1. Boundary and token detection

Setup: We start by evaluating how closely the predicted boundaries match true syllables. Ground truth syllable boundaries are obtained using phone-level forced alignments [27] with a rule-based method [28] that converts phone sequences to syllables. In addition to precision, recall, and F1, boundaries are also evaluated using R-value and token scores. R-value measures how close predicted syllable boundaries are to an ideal operating point with 100% recall and 0% over-segmentation [29]; over-segmentation is the ratio of the number of predicted boundaries relative to the number of ground-truth boundaries [30]. Token boundary scores (precision, recall, and F1) require both the start and end boundaries to be correct. For all metrics, we allow a boundary tolerance of 50 ms on either side of a true boundary [10, 11]. Boundaries that border silences are not evaluated. As in [10], we allow for a constant time shift in the predicted boundaries to account for potential latency in the representations; these shifts are tuned for each system on development data.

Results: Table 1 presents the boundary prediction results on the combined test sets of LibriSpeech. Overall, the 5.0 Hz version of SyllableLM performs best on several of the metrics. However, ZeroSyl performs competitively despite not requiring any model training, achieving very similar token F1 scores (54% vs. 56%). With an R-value of 75% and a token F1-score of 54%, ZeroSyl also outperforms Sylber (71% and 51%, respectively). ZeroSyl maintains a low over-segmentation rate of 10%, substantially lower than the higher-frequency SyllableLM variants. This indicates that our method avoids fragmenting syllables into smaller acoustic events. We also include a prominence-based baseline (PromSeg) that uses the same layer, window size, and prominence as ZeroSyl, but with peak detection on the cosine distance between adjacent frames [19]. ZeroSyl outperforms this baseline on all metrics except over-segmentation. This shows that the L2 norm provides a better signal for syllable boundaries.

4.2. Syllable discovery

Setup: To evaluate the quality of the discovered syllabic units, we assess how well the clusters map to true syllable labels. As in [10], we map predicted segments to ground-truth syllables with the most overlap. High-quality clusters should exhibit high per-cluster purity, where each cluster ID maps mostly to one syllable label, as well as high inverse purity, where each syllable is found mostly in one cluster. Syllable-normalized mutual information (SNMI) measures the percentage of uncertainty about the syllable label eliminated after observing the cluster ID [18]. While smaller vocabularies are often used to evaluate syllabic clusters [10, 11, 17], we evaluate the systems using the actual vocabulary sizes employed in the downstream language model (8k–10k). We compare system efficiency in terms of bitrate (using the entropic formulation [2]) and token frequency. Evaluations are done on the LibriSpeech test set.

Results: Table 2 shows syllable discovery results in its middle columns. By default, ZeroSyl collapses silences (Section 3.2), but here we also report scores without collapsing. ZeroSyl outperforms all other systems on all purity and mutual-information metrics, while maintaining the lowest bitrate (52 bps). Specifically, ZeroSyl achieves an SNMI of 88.9%, higher than Sylber (83.5%) and SyllableLM (82.6%). The rich semantic information in the later layers of WavLM Large likely boosts the quality of the representations compared to the base versions of data2vec [31] and SD-HuBERT [17] that are used as backbones in SyllableLM and Sylber, respectively. Comparing the uncollapsed and collapsed versions of ZeroSyl shows that merging redundant silence tokens drastically improves inverse purity (from 20.0% to 32.0%) while maintaining high per-cluster purity and SNMI.

4.3. Spoken language modeling

Our main question is whether syllabic units from the simple ZeroSyl approach (Sections 3.1 and 3.2) are useful for spoken language modeling. We train language models on the syllabic units of ZeroSyl (Section 3.3) and the other systems (Section 3.4), and probe the models' lexical, syntactic, and narrative capabili-

Table 2: *Syllable discovery and language modeling performance. Syllable discovery metrics are reported on the combined LibriSpeech test set. Language models are trained on 6k hours of Libri-Light and evaluated on the sLM21 development set.*

				Syllable discovery (%)			Language modeling (%)			
	Vocab. size	Bitrate (bps)	Freq. (Hz)	Purity	Inverse purity	SNMI	sWUGGY (all)	sWUGGY (IV)	sBLIMP	tSC
SyllableLM 6.25Hz	8k	73	5.87	67.5	23.2	82.6	66.1	74.2	56.4	67.6
Sylber	10k	53	4.46	73.5	28.4	83.5	66.0	74.7	59.1	65.8
ZeroSyl (uncollapsed sil.)	10k	58	4.55	**80.5**	20.0	**89.5**	67.0	76.3	58.6	67.5
ZeroSyl	9k	**52**	**4.35**	79.9	**32.0**	88.9	**68.0**	**78.6**	**60.5**	**68.1**

ties. All the benchmarks below rely on likelihood estimates from the language model, for which we use the mean log-likelihood per token to account for varying sequence lengths.

Lexical task (sWUGGY): This task tests whether a model can distinguish real lexical items from phonologically plausible fakes. The model processes pairs containing one actual word, such as "manufacturer," and one matched non-word, like "manifelturer," and must assign a higher likelihood to the real word. We report metrics for *all* pairs in the SLM21 development set [15], and for *in-vocabulary (IV)* pairs, where the real word is guaranteed to appear in the LibriSpeech vocabulary.

Results: Table 2 shows that ZeroSyl outperforms all the other systems on the sWUGGY benchmark: it achieves the highest accuracy on both the general (68.0%) and in-vocabulary (78.6%) subsets. The silence collapsing step in ZeroSyl also yields a clear improvement over the uncollapsed version.

Syntactic task (sBLIMP): This benchmark evaluates a model's grasp of sentence structure [15]. Each test item consists of two spoken sentences: one is grammatically correct, while the other contains a single syntax error. For example, the model compares "What *dentist* is Angela working with?" against "What is Angela working with *dentist*?" and must identify the valid sentence by assigning it a higher likelihood.

Results: Table 2 shows that ZeroSyl achieves an sBLIMP score of 60.5%, higher than both SyllableLM (56.4%) and Sylber (59.1%).[1] Again, the collapsed version of ZeroSyl improves upon the one where silences are not collapsed.

Narrative task: The Topic StoryCloze (tSC) benchmark evaluates long-range language modeling performance by measuring narrative coherence [16]. The model receives a four-sentence spoken prompt and must identify the original fifth sentence from a negative sample randomly selected from a different story. This task tests whether the model can pick a logically consistent continuation in a spoken context.

Results: As shown in Table 2, ZeroSyl achieves 68.0% accuracy on the tSC task, outperforming all other systems.

Overall, ZeroSyl outperforms other syllabic units systems across the lexical, syntactic, and narrative benchmarks. Given that ZeroSyl also has the lowest bitrate, these results show that the discovered units provide a more efficient representation for spoken language modeling than previous work.

5. Experiments: Scaling behavior

Finally, we investigate the performance of ZeroSyl as training data scales from 600 to 6k to 60k hours of Libri-Light. Here we compare mainly to SpidR [7], because it represents a well-performing non-syllabic approach which is still simple (like ZeroSyl): SpidR uses K-means to discretize features from a proposed SSL model. However, unlike ZeroSyl, this system produces tokens close to the frame-level,[2] resulting in a much higher bitrate than ZeroSyl. We also include the text topline from [7] that uses byte-pair encoding (BPE) on characters. ZeroSyl, SpidR, and the topline all use the same language model [13].

Results: sWUGGY, sBLIMP, and tSC results are given in Figure 3. In the sWUGGY task (Figure 3 top-left), ZeroSyl trails the fine-grained SpidR. This reflects a trade-off: while SpidR units preserve the frame-level phonetic detail required in the lexical task, ZeroSyl compresses this information into syllabic units, sacrificing acoustic granularity. However, this trade-off proves advantageous at the syntactic level: in the sBLIMP evaluation (Figure 3 top-right), the performance of SpidR saturates as data scales, whereas ZeroSyl maintains a steep upward trajectory, outperforming SpidR at 6k and 60k hours. On the tSC narrative task (Figure 3 bottom-left), ZeroSyl effectively matches SpidR at the 60k-hour mark, with a steeper upward trend.

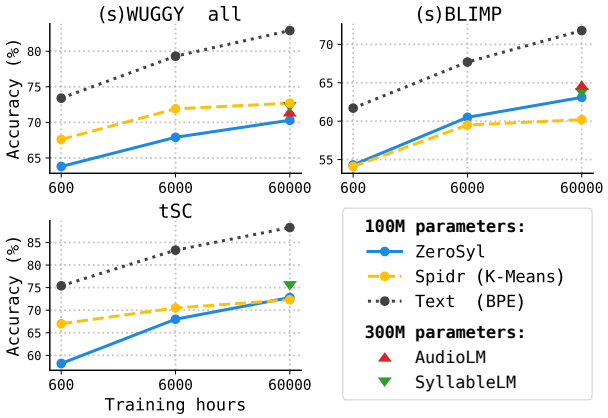


Figure 3: *Scaling behavior of ZeroSyl compared to speech (SpidR with K-means) and text (BPE) systems on the Libri-Light corpus.*

Throughout this paper, we used the same 100M-parameter language model architecture [13] to allow for meaningful comparisons, and did not consider the considerably larger sizes also occasionally used [4, 11]. But we see in Figure 3 that the performance gains of the 300M-parameter AudioLM [4] and SyllableLM [11] systems are marginal considering the significant increase in computational cost. Interestingly, even with these bigger models, a single system does not perform best on all three tasks.[3] Returning to our main question of whether a simple syllable-based approach is possible, it is clear that even compared to these much larger models, ZeroSyl provides competitive performance without intricate training procedures.

6. Conclusion

In this work, we introduced ZeroSyl, a training-free framework for discovering syllabic units directly from unlabelled speech. By leveraging feature norms from a frozen SSL model (WavLM Large), ZeroSyl eliminates the complex multi-stage methods required by prior state-of-the-art syllabic tokenizers.

ZeroSyl outperforms existing syllabic tokenizers across lexical, syntactic, and narrative benchmarks when trained on 6k hours of speech. Scaling experiments reveal that while finer-grained units retain an advantage for lexical tasks, ZeroSyl's syllabic units are more effective for modeling long-range dependencies, showing a steep upward trajectory in syntactic performance as data increases. Ultimately, ZeroSyl demonstrates that high-quality spoken language modeling does not necessarily require increasingly complex tokenizers, but rather a deeper understanding of the signals already present in SSL features.

Future work should investigate why and how the L2 norm signal of SSL representations provides implicit positional encodings of syllables. While ZeroSyl's low rate encoding benefits syntactic and narrative modeling, the resulting syllabic units might be detrimental for encoding lexical items, especially if they are rare or unseen. This should be investigated.

[1] We report all scores using mean per-token log-likelihoods, since this gave better sWUGGY and tSC scores for all systems, and since this is what is reported in [4, 7]. However, when using unnormalized log-likelihoods, sBLIMP scores improve to 62.4%, 60.8%, 63.0%, and 63.8% for SyllableLM, Sylber, Zerosyl-uncollapsed, and ZeroSyl, respectively.

[2] With $K = 256$, the token rate is 26 Hz and the bitrate is 205 bps.

[3] AudioLM did not evaluate the tSC benchmark or release models.

7. Acknowledgements

Nicol Visser and Simon Malan are funded through the Google PhD Fellowship Program.

8. Generative AI use disclosure

Generative AI tools were utilized exclusively for linguistic refinement, copy-editing, and polishing of the author-drafted text to improve clarity. No scientific content, data analysis, or significant portions of the manuscript were generated with these tools; the authors retain full responsibility for the accuracy, originality, and intellectual integrity of the work.

9. References

[1] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath, and S. Watanabe, "Self-Supervised Speech Representation Learning: A Review," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1179–1210, 2022.

[2] E. Dunbar, N. Hamilakis, and E. Dupoux, "Self-Supervised Language Learning From Raw Audio: Lessons From the Zero Resource Speech Challenge," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1211–1226, 2022.

[3] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed, and E. Dupoux, "On Generative Spoken Language Modeling from Raw Audio," *Trans. Assoc. Comput. Linguist.*, vol. 9, pp. 1336–1354, 2021.

[4] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, and N. Zeghidour, "AudioLM: A Language Modeling Approach to Audio Generation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2523–2533, 2023.

[5] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "w2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training," in *Proc. ASRU*, 2021, pp. 244–250.

[6] S. Cuervo and R. Marxer, "Scaling Properties of Speech Language Models," in *Proc. EMNLP*, 2024, pp. 351–361.

[7] M. Poli, M. Luthra, Y. Benchekroun, Y. Higuchi, M. Gleize, J. Shen, R. Algayres, Y.-A. Chung, M. Assran, J. Pino, and E. Dupoux, "SpidR: Learning Fast and Stable Linguistic Units for Spoken Language Models Without Supervision," *Trans. Mach. Learn. Res.*, 2025.

[8] R. Algayres, Y. Adi, T.-A. Nguyen, J. Copet, G. Synnaeve, B. Sagot, and E. Dupoux, "Generative Spoken Language Model Based on Continuous Word-Sized Audio Tokens," in *Proc. EMNLP*, 2023, pp. 3008–3028.

[9] N. Visser and H. Kamper, "Spoken Language Modeling with Duration-Penalized Self-Supervised Units," in *Proc. Interspeech*, 2025, pp. 1968–1972.

[10] C. J. Cho, N. Lee, A. Gupta, D. Agarwal, E. Chen, A. W. Black, and G. K. Anumanchipalli, "Sylber: Syllabic Embedding Representation of Speech from Raw Audio," in *Proc. ICLR*, 2025.

[11] A. Baade, P. Peng, and D. Harwath, "SyllableLM: Learning Coarse Semantic Units for Speech Language Models," in *Proc. ICLR*, 2025.

[12] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.

[13] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer, "OPT: Open Pre-trained Transformer Language Models," 2022. [Online]. Available: https://arxiv.org/abs/2205.01068

[14] S. Arora, K.-W. Chang, C.-M. Chien, Y. Peng, H. Wu, Y. Adi, E. Dupoux, H.-y. Lee, K. Livescu, and S. Watanabe, "On The Landscape of Spoken Language Models: A Comprehensive Survey," *Trans. Mach. Learn. Res.*, 2025.

[15] E. Dunbar, M. Bernard, N. Hamilakis, T.-A. Nguyen, M. de Seyssel, P. Rozé, M. Rivière, E. Kharitonov, and E. Dupoux, "The Zero Resource Speech Challenge 2021: Spoken language modelling," in *Proc. Interspeech*, 2021, pp. 1574–1578.

[16] M. Hassid, T. Remez, T.-A. Nguyen, I. Gat, A. Conneau, F. Kreuk, J. Copet, A. Defossez, G. Synnaeve, E. Dupoux, R. Schwartz, and Y. Adi, "Textually Pretrained Speech Language Models," in *Proc. NeurIPS*, 2023, pp. 63 483–63 501.

[17] C. J. Cho, A. Mohamed, S.-W. Li, A. W. Black, and G. K. Anumanchipalli, "SD-HuBERT: Sentence-Level Self-Distillation Induces Syllabic Organization in HuBERT," in *Proc. ICASSP*, 2024, pp. 12 076–12 080.

[18] W. Hsu, B. Bolte, Y. H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[19] A. Pasad, C.-M. Chien, S. Settle, and K. Livescu, "What Do Self-Supervised Speech Models Know About Words?" *Trans. Assoc. Comput. Linguist.*, vol. 12, pp. 372–391, 2024.

[20] L. ten Bosch and B. Cranen, "A Computational Model for Unsupervised Word Discovery," in *Proc. Interspeech*, 2007, pp. 1481–1484.

[21] S. Malan, B. van Niekerk, and H. Kamper, "Unsupervised Word Discovery: Boundary Detection with Clustering vs. Dynamic Programming," in *Proc. ICASSP*, 2025.

[22] ——, "Should Top-Down Clustering Affect Boundaries in Unsupervised Word Discovery?" 2025. [Online]. Available: https://arxiv.org/abs/2507.19204

[23] A. Pasad, B. Shi, and K. Livescu, "Comparative layer-wise analysis of self-supervised speech models," in *Proc. ICASSP*, 2023, pp. 1–5.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[25] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The Faiss library," 2024. [Online]. Available: https://arxiv.org/abs/2401.08281

[26] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "Libri-Light: A Benchmark for ASR with Limited or No Supervision," in *Proc. ICASSP*, 2020, pp. 7669–7673.

[27] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech*, 2017, pp. 498–502.

[28] K. Gorman, "Python module for syllabifying English ARPABET transcriptions," 2013, GitHub repository. [Online]. Available: https://github.com/kylebgorman/syllabify

[29] O. J. Räsänen, U. K. Laine, and T. Altosaar, "An improved speech segmentation quality measure: the R-value," in *Proc. Interspeech*, 2009, pp. 1851–1854.

[30] B. Petek, O. Andersen, and P. Dalsgaard, "On the robust automatic segmentation of spontaneous speech," in *Proc. ICSLP*, 1996, pp. 913–916.

[31] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language," 2022. [Online]. Available: https://arxiv.org/abs/2202.03555