

Comparing Unsupervised and Supervised Semantic Speech Tokens: A Case Study of Child ASR

Mohan Shi, Natarajan Balaji Shankar, Kaiyuan Zhang, Zilai Wang, Abeer Alwan

Department of Electrical and Computer Engineering

University of California Los Angeles

Los Angeles, USA

{shimohan, balaji1312, kaiyuanzhang, zilaiwang2001}@ucla.edu, alwan@ee.ucla.edu

Abstract—Discrete speech tokens have gained attention for their storage efficiency and integration with Large Language Models (LLMs). They are commonly categorized into acoustic and semantic tokens, with the latter being more advantageous for Automatic Speech Recognition (ASR). Traditionally, unsupervised K-means clustering has been used to extract semantic speech tokens from Speech Foundation Models (SFMs). Recently, supervised methods, such as finite scalar quantization (FSQ) trained with ASR loss, have emerged for speech generation. Both approaches leverage pre-trained SFMs, benefiting low-resource tasks such as child ASR.

This paper systematically compares supervised and unsupervised semantic speech tokens for child ASR. Results show that supervised methods not only outperform unsupervised ones but even unexpectedly surpass continuous representations, and they perform well even in ultra-low bitrate settings. These findings highlight the advantages of supervised semantic tokens and offer insights for improving discrete speech tokenization.

Index Terms—Semantic Speech tokens, Speech Discretization, Children’s Speech Recognition, Finite Scalar Quantization

I. INTRODUCTION

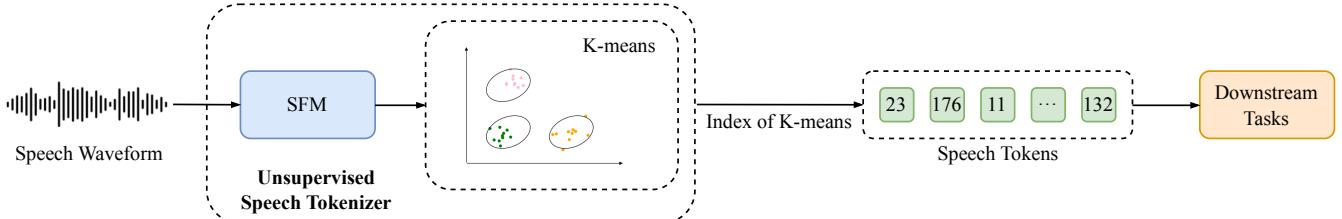
In recent years, the rapid advancement of deep learning [1] has driven remarkable progress in speech processing, particularly in Automatic Speech Recognition (ASR) [2]–[8]. Conventional ASR systems typically rely on continuous acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs) or Filterbanks (Fbanks), or leverage high-dimensional representations learned via self-supervised or data-driven methods [9]–[12]. More recently, the use of discrete speech tokens has attracted growing attention [13]–[18]. These methods encode speech into sequences of discrete tokens, serving as compact and symbolic representations for downstream tasks. Discrete tokens offer several key advantages: (1) they significantly reduce storage and computational costs due to their low-bitrate nature; and (2) they enable seamless integration with large language models [19]–[21], thereby facilitating unified speech-language modeling in multi-modal frameworks.

In the literature, discrete speech tokens are generally classified into two main categories: *acoustic* and *semantic* tokens. Acoustic tokens are typically learned through signal reconstruction objectives, such as those used in neural audio codecs [13], [14]. These tokens are effective in capturing low-level acoustic details but often lack higher-level linguistic ab-

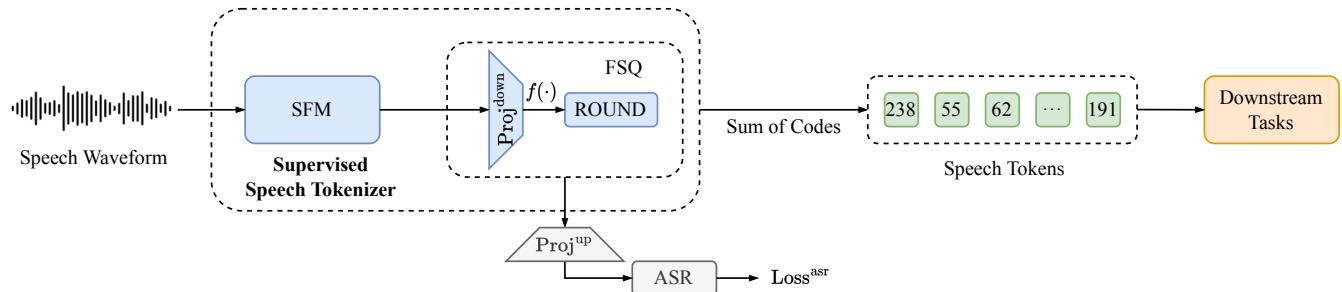
straction, thus conveying limited semantic content. In contrast, so-called ‘semantic’ tokens [15], [17], [18] are extracted from the representations of Speech Foundation Models (SFMs) [9]–[11]. As a result, semantic tokens inherently encode richer linguistic and contextual information. This makes them particularly well-suited for high-level speech tasks such as ASR, where capturing semantics is more crucial than preserving fine-grained acoustic fidelity.

Traditionally, unsupervised K-means clustering [22] has been the predominant approach for extracting semantic tokens from the intermediate representations of SFMs [15], [17], [18]. This method discretizes the continuous speech representations by assigning each frame-level embedding to one of the cluster centroids, which are learned from the training set. While simple and effective, it does not incorporate task-specific objectives and thus may overlook semantically important distinctions. More recently, attention has been directed toward leveraging supervised learning objectives to train quantizers based on SFM representations [20], [21]. Among these methods, the Finite Scalar Quantization (FSQ) technique [23], which was proposed as a core component of the supervised semantic speech (S^3) tokenizer [21], demonstrates higher codebook utilization than traditional Vector Quantization (VQ) [20]. The supervised FSQ tokenizer, optimized with an ASR loss during training, has demonstrated strong empirical performance in text-to-speech (TTS) [21] as the learned tokens retain high-level semantic information beneficial for next-token prediction.

Both unsupervised and supervised discrete speech tokens leverage the pre-trained knowledge embedded in SFMs, making them particularly suitable for low-resource ASR scenarios, such as children’s speech recognition [24], [25]. Prior work [26] has benchmarked children’s ASR using SFMs, while other studies have explored unsupervised K-means-based semantic tokens for this domain [27], [28]. However, these unsupervised approaches often result in degraded ASR performance compared to continuous features. Despite growing interest, there has been no systematic evaluation comparing discrete semantic tokens with continuous features for children’s ASR. In particular, the relative strengths of unsupervised versus supervised semantic tokens for ASR remain underexplored. A comprehensive investigation is thus needed to better understand the trade-offs between token types and their potential for improving ASR.



(a) Extraction pipeline of unsupervised semantic speech tokens using K-means clustering.



(b) Extraction pipeline of supervised semantic speech tokens using Finite Scalar Quantization (FSQ).

Fig. 1: Schematic illustration of the extraction pipeline for both unsupervised and supervised semantic speech tokens. $f(\cdot)$ denotes the bounding function.

To address these gaps, this study systematically compares semantic speech tokens derived from unsupervised K-means and supervised FSQ with continuous representations from SFMs, in the context of child ASR as a representative low-resource setting. Experimental results and in-depth analysis uncover unexpected performance and yield meaningful insights into speech tokenization.

In particular, the main contributions of this paper are:

- We conduct a comprehensive comparison between unsupervised and supervised semantic speech tokens on the ASR task across multiple child speech datasets. Our analysis covers not only ASR performance but also bitrate efficiency, codebook distribution, and cross-domain generalization, providing new insights into the characteristics of these tokenization methods.
- Our results demonstrate that supervised FSQ tokens not only outperform unsupervised K-means tokens but also, unexpectedly, **surpass continuous high-dimensional representations extracted from in-domain SFMs**, revealing a previously underexplored advantage of supervised discrete tokenization.
- Through in-depth analysis of codebook usage and extended experiments, we demonstrate that **discrete semantic speech tokens retain competitive ASR performance even at ultra-low bitrates**, underscoring their potential for efficient and effective speech representation. We further evaluate explore their generalizability across different speaker styles and age groups, revealing key limitations and future research directions.

II. SEMANTIC SPEECH TOKENS

Semantic speech tokens, commonly referred to as such in the literature despite capturing a mix of phonetic and linguistic cues [29], are derived from speech foundation models and are primarily used in semantics-oriented tasks such as ASR. Their extraction methods broadly fall into two categories: unsupervised and supervised approaches, as illustrated in Figure 1. For clarity and consistency, we adopt the term *semantic tokens* throughout this paper to distinguish them from low-level acoustic tokens.

A. Unsupervised Semantic Tokens Based on K-means

K-means clustering [22] is one of the most widely used unsupervised methods for extracting semantic speech tokens. The overall extraction pipeline is illustrated in Figure 1a. Given a speech training dataset X , we first extract continuous representations using a pre-trained speech foundation model (SFM):

$$H = \text{SFM}(X) \quad (1)$$

where $H = \{h_n \in \mathbb{R}^d \mid n = 1, \dots, N\}$ represents the frame-wise continuous representations, with d denoting the feature dimension and N the total number of frames.

Next, we train a K-means model on H :

$$C = \text{K-means}(H) \quad (2)$$

where $C = \{c_k \in \mathbb{R}^d \mid k = 1, \dots, K\}$ represents the K cluster centroids computed based on the Euclidean distance.

For each frame, we assign the closest cluster index to obtain the discrete token set $Z^{\text{km}} = \{z_n^{\text{km}} \in \mathbb{Z} \mid n = 1, \dots, N\}$:

$$z_n^{\text{km}} = \underset{k \in \{1, \dots, K\}}{\operatorname{argmin}} \|x_n - c_k\|^2 \quad (3)$$

The clustering model is trained on the training set and applied consistently across all subsets. The resulting tokens serve as discrete inputs for downstream tasks.

B. Supervised Semantic Tokens Based on FSQ

Unlike unsupervised tokenizers that rely on K-means clustering to obtain discrete tokens, the Supervised Semantic Speech (S^3) tokenizer [20], [21] employs the ASR loss during training to guide the speech discretization process, as illustrated in Figure 1b.

Specifically, given a speech training dataset X , the SFM first generates intermediate representations H , as shown in Equation (1). These representations are then passed through the finite scalar quantization (FSQ) [23] module for discretization.

In the FSQ module, the intermediate representations H are first projected into a low-rank space of dimension d^{low} . Each dimension is then quantized into a fixed number of discrete values using a bounding function $f(\cdot)$ (e.g., \tanh) followed by a rounding operation.

$$H^{\text{down}} = \text{Proj}^{\text{down}}(H) \quad (4)$$

$$H^{\text{code}} = \text{ROUND}(f(H^{\text{down}})) \quad (5)$$

The resulting quantized representation $H^{\text{code}} = \{h_n^{\text{code}} \in \mathbb{Z}^{d^{\text{low}}} \mid n = 1, \dots, N\}$ forms an implicit codebook. Given the hyperparameters of FSQ, denoted as $\mathcal{L} = [L_0, L_1, \dots, L_{d^{\text{low}}-1}]$, where d^{low} represents the number of low-rank channels, and L_i corresponds to the number of quantization levels per channel. Each quantized code $h_{n,i}^{\text{code}}$ is an integer within the range $[0, L_i]$.

The final speech token z_n^{fsq} is computed as a weighted sum of the quantized low-rank codes h_n^{code} :

$$z_n^{\text{fsq}} = h_{n,0}^{\text{code}} + \sum_{i=1}^{d^{\text{low}}-1} (h_{n,i}^{\text{code}} \prod_{j=0}^{i-1} L_j) \quad (6)$$

The size of the implicit codebook is determined by the product of all elements in \mathcal{L} . During tokenizer training, the quantized representation H^{code} is projected back to its original dimensional space using an up-projection function:

$$\hat{H} = \text{Proj}^{\text{up}}(H^{\text{code}}) \quad (7)$$

The reconstructed representation, \hat{H} is then fed into an ASR module to generate the ASR hypothesis, which is then compared with the ground truth transcriptions to compute the ASR loss for optimizing the supervised speech tokenizer:

$$\hat{Y} = \text{ASR}(\hat{H}) \quad (8)$$

$$\text{Loss}^{\text{fsq}} = \text{Loss}^{\text{asr}}(\hat{Y}, Y) \quad (9)$$

where \hat{Y} and Y denote the predicted and ground truth transcriptions, respectively. After training the entire system with

the ASR loss, the discrete speech tokens z_n^{fsq} are extracted using Equation (6) and can then be utilized for downstream tasks.

III. EXPERIMENTAL SETUPS

A. Dataset

To ensure a fair and consistent comparison with previous benchmark studies, we conduct experiments on two child speech corpora as used in [26]: the My Science Tutor (MyST) corpus of **spontaneous** speech [24], and the CSLU OGI Kids (OGI) corpus of **scripted**, read speech [25]. We follow the preprocessing procedure described in [26].

The MyST corpus comprises 240 hours of speech transcribed from children in grades 3–5 (aged 8–10 years) recorded during virtual tutoring sessions. Following [30], we perform quality filtering using Whisper-largeV2, and discard utterances that yield a WER greater than 50% or contain fewer than three words. We also exclude utterances longer than 30 seconds. After filtering, we obtain 133 hours of training data, with development and test sets comprising 21 and 25 hours, respectively.

The OGI Kids corpus comprises 50 hours of speech collected from 1,100 children aged 4–15, reading isolated words, sentences, or sequences of digits. The data is divided into training (70%), development (15%), and test (15%) sets, with no speaker overlap across splits [26].

B. Configurations

For the speech foundation model, we adopt WavLM Large [11], a leading self-supervised learning (SSL) model, as the backbone for both the unsupervised K-means tokenizer and the supervised FSQ tokenizer. K-means clustering is performed with the default setting of 2000 clusters, following [18]. For FSQ, we configure the codebook granularity as $\mathcal{L} = [5, 5, 5, 4, 4]$, yielding a total codebook size of 2000, thereby aligning with the number of K-means clusters for fair comparison.

Both K-means and FSQ tokenizers are trained on the full training set of each dataset. For K-means, we extract speech representations from the 21st layer of a pre-trained WavLM model and the final (24th) layer of a WavLM model fine-tuned on the respective training set^{1,2}. For FSQ training, we initialize the model with a fine-tuned WavLM and optimize it using CTC loss [2], during which the transformer layers in WavLM are jointly trained.

We use ASR as the downstream evaluation task throughout this work. For consistency, we adopt the default 12-layer E-Branchoformer architecture [7] from ESPnet [31], and perform CTC-only decoding using greedy search. When using discrete speech tokens as input, their embeddings are randomly initialized. For text targets, we employ byte pair encoding (BPE)

¹<https://huggingface.co/FSQChildASR/wavlm-large-myst>

²<https://huggingface.co/FSQChildASR/wavlm-large-ogi>

TABLE I: WER (%) comparison of various methods on MyST and OGI. The first three rows show results from previous works, where fine-tuned SSL SFMs were directly used for inference as reference baselines. The remaining rows present results obtained by feeding different input features (continuous or discrete) into the downstream ASR model. Bitrate (bits/sec) serves as a reference for input data efficiency. The downstream ASR adopts an E-Branchformer with a CTC-only architecture and greedy decoding. The bold font indicates the best performance.

Approaches	ASR Model	Feature	WER ↓				Feature Bitrate ↓
			Myst dev	Myst test	OGI dev	OGI test	
Fine-tuned SSL SFMs [26]	Wav2vec 2.0-CTC	-	10.6	11.1	2.1	2.5	-
	HuBERT-CTC	-	10.5	11.3	2.2	2.5	-
	WavLM-CTC	-	9.6	10.4	1.7	1.8	-
Feature (Continuous/Discrete) + Downstream ASR	E-Branchformer-CTC	<i>Continuous Feature & Representation</i>					
		Fbank	16.4	16.5	2.2	2.9	256000
		WavLM	10.9	11.6	3.3	3.9	1638400
		Fine-tuned WavLM	9.6	10.1	1.6	1.7	1638400
		<i>Unsupervised Semantic Tokens</i>					
		WavLM K-means	11.4	11.9	5.0	6.4	548.3
		Fine-tuned WavLM K-means	10.1	10.7	2.8	3.1	548.3
		<i>Supervised Semantic Tokens</i>					
		Fine-tuned WavLM FSQ	9.3*	10.0*	1.5	1.5*	548.3

*: Statistical significance is confirmed with $p < 0.05$

with a vocabulary size of 5000 on MyST, and a smaller size of 200 for OGI to reflect its limited lexical diversity.

C. Metrics

In addition to Word Error Rate (WER) as the primary evaluation metric, we also report bitrate (bits per second) to provide an auxiliary measure of the encoding efficiency of different continuous and discrete speech representations. The bitrate is computed following the methodology outlined in [18], allowing a consistent and fair comparison across various encoding schemes.

IV. EXPERIMENTAL RESULTS

A. Overall Comparison on Child ASR

Table I compares various approaches on the MyST and OGI datasets. Baseline results from prior work, using fine-tuned self-supervised learning (SSL) speech foundation models (SFMs) for direct inference, are reported in the first three rows. The remaining rows evaluate different input representations, both continuous and discrete, within the downstream E-Branchformer-CTC ASR model.

Among these approaches, Fbank does not show promising results due to the lack of pre-training benefits. In contrast, using WavLM representations, especially those fine-tuned on in-domain data, yields substantial performance improvements, achieving WERs of 9.6% / 10.1% on MyST-dev/test and 1.6% / 1.7% on OGI-dev/test. However, the high dimensionality of WavLM Large (1024) combined with its 32-bit floating-point representation leads to an exceptionally high bitrate of 1,638,400 bits per second.

Discrete semantic tokens offer a compelling alternative by combining the benefits of SFM pre-training with efficient

compression. Notably, supervised FSQ tokens achieve the best overall performance, reaching WERs of 9.3% / 10.0% on MyST and 1.5% / 1.5% on OGI, outperforming both Fbank and unsupervised K-means tokens. Remarkably, FSQ tokens **outperform in-domain fine-tuned WavLM representations** as well as the direct inference results from fine-tuned SSL SFMs, despite operating at a much lower bitrate and using the same WavLM backbone.

This finding is counterintuitive, as continuous upstream features are typically expected to outperform their discrete counterparts in ASR due to information loss during discretization [15]–[18]. We hypothesize that this advantage arises because the WavLM backbone used in supervised FSQ tokenizer was explicitly adapted during supervised training to facilitate discretization, thereby enhancing its effectiveness for downstream ASR tasks.

B. Comparison of De-duplication and Sub-word Modeling on Discrete Speech Tokens

Table II presents the ASR performance and bitrate when applying de-duplication (DD) and Byte-Pair Encoding (BPE) sub-word (SW) modeling [15] as post-processing strategies on speech token sequences. Both techniques aim to reduce redundancy: DD eliminates consecutive repeated tokens, while SW compresses frequent token patterns into subword units.

The results show that both DD and SW modeling effectively reduce the bitrate by shortening the overall token sequence length. However, this compression-induced reduction leads to a slight degradation in ASR performance, highlighting a trade-off between compactness and recognition accuracy. Notably, applying DD to supervised FSQ tokens incurs only a modest accuracy loss while achieving a meaningful reduction in bitrate.

TABLE II: Comparison of WER (%) and Bitrate (bits/sec) using de-duplication (DD) and BPE sub-word (SW) modeling on speech tokens. Bitrate with DD and SW modeling is based on MyST, with an SW vocabulary of 6000 for a codebook size of 2000.

Token Type	DD	SW	WER ↓				Bitrate ↓
			MyST		OGI		
	dev	test	dev	test			
K-means	✗	✗	10.1	10.7	2.8	3.1	548.3
	✓	✗	10.3	10.8	2.9	3.0	388.0
	✗	✓	10.4	10.8	3.0	3.1	351.2
	✓	✓	10.4	10.9	6.2	6.1	267.6
FSQ	✗	✗	9.3*	10.0*	1.5	1.5*	548.3
	✓	✗	9.4	10.2	1.5	1.7	329.1
	✗	✓	9.4	10.1	4.6	4.5	304.3
	✓	✓	9.5	10.2	5.6	5.1	239.2

*: Statistical significance is confirmed with $p < 0.05$

Across all post-processing configurations, FSQ tokens consistently achieve a lower bitrate than K-means tokens, indicating greater efficiency in discarding redundant information during compression. These findings suggest that supervised FSQ tokens provide superior encoding efficiency by preserving essential semantic content while minimizing redundancy.

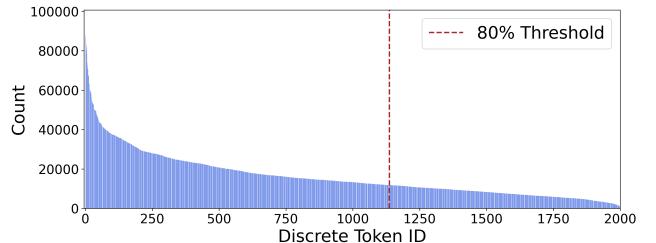
C. Frequency Distribution Analysis of Speech Tokens

Figure 2 illustrates the frequency distribution of both types of discrete tokens on the MyST dataset. Interestingly, we observe that K-means tokens are relatively uniformly distributed across the codebook, whereas FSQ tokens exhibit a sharp and highly skewed distribution, with a large proportion of occurrences concentrated in a small subset of tokens.

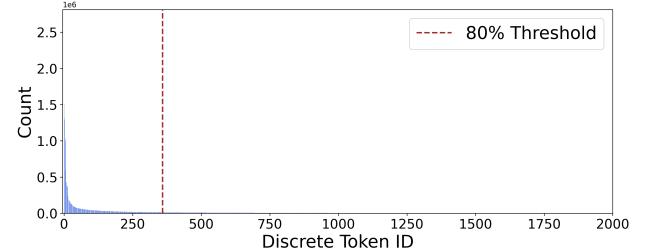
Remarkably, this sharp distribution correlates with better ASR performance, challenging the commonly held assumption that a more uniform token usage is beneficial [21], [32]. We hypothesize that this phenomenon may be attributed to the nature of speech phonemes, where only a limited set of sound categories including vowels, consonants and semi-vowels need to be effectively represented for successful transcription. As a result, a non-uniform token distribution that prioritizes frequently occurring phonetic units might actually enhance ASR performance. This observation motivates our further investigation into whether reducing the codebook size of discrete tokens can maintain or even improve ASR performance, as discussed in Section IV-D.

D. Exploring Ultra-Low Bitrate Speech Tokens

Building on the highly skewed FSQ token distribution observed in Section IV-C, we explore the impact of reducing the codebook size to a significantly lower level. Specifically, we set $\mathcal{L} = [8, 6, 5]$, resulting in a codebook of size 240 (approximately 2^8), which is also the default configuration in [23]. For fair comparison, we configure K-means clustering to produce 240 centers as well. The results are presented in



(a) Distribution of unsupervised K-means tokens on MyST



(b) Distribution of supervised FSQ tokens on MyST

Fig. 2: Frequency distribution of discrete tokens with a codebook size of 2000, based on the MyST corpus. The y-axis indicates the frequency of tokens, and the x-axis denotes token IDs sorted in descending order of frequency. The red line marks the 80% cumulative frequency threshold.

Table III, with the first row showing continuous fine-tuned WavLM representation as a reference baseline.

Surprisingly, both tokenizers with a codebook of only 240 entries achieve ASR performance comparable to that of the original 2000-entry codebook across both datasets. For K-means tokens, whose usage is more uniformly distributed, reducing the codebook size to 240 leads to a slight drop in ASR accuracy on the MyST test set, whereas it yields an improvement on the OGI dataset. FSQ tokens also achieve comparable or better performance on OGI. We attribute this to the characteristics of the OGI corpus, which involves younger child speakers with more constrained vocabularies, a smaller dataset size, and simpler scripted utterances. These factors inherently require fewer distinct token types, making a reduced codebook size more effective by eliminating redundant tokens.

These results reinforce our earlier observation (Section IV-C) that only a compact subset of semantic speech units may be sufficient for effective ASR. Moreover, applying additional techniques such as token de-duplication (DD) and BPE-based sub-word (SW) modeling enables further bitrate reduction with minimal impact on transcription quality.

E. Cross-Domain Evaluation of the Speech Tokenizers

We further assess the robustness and generalizability of speech tokenizers across different age groups and styles using different domains of tokenizers. Specifically, we define the domain of a speech tokenizer based on its training data (e.g., “Pretrained SSL” means that tokens are extracted from a pretrained SSL model). We train on the combined MyST and

TABLE III: Comparison of WER (%) and bitrate (bits/sec) for codebook sizes of 2000 and 240 on the MyST and OGI datasets. Bitrate with de-duplication (DD) and BPE sub-word (SW) modeling is reported on MyST, with SW vocabularies of 6000 (for 2000) and 600 (for 240), respectively.

Codebook Size	WER ↓				Bitrate ↓
	MyST ¹		OGI ²		
	dev	test	dev	test	
<i>Continuous Fine-tuned WavLM Representation</i>					
-	9.6	10.1	1.6	1.7	1638400
<i>Unsupervised K-means Tokens</i>					
2000	10.1	10.7	2.8	3.1	548.3
240	10.3	10.9	1.9	2.2	395.3
<i>Supervised FSQ Tokens</i>					
2000	9.3	10.0*	1.5	1.5	548.3
240	9.4	10.2	1.4	1.5	395.3
+ DD	9.5	10.2	1.6	1.8	224.2
+ SW	9.6	10.3	2.0	2.2	186.9

*: Statistical significance is confirmed with $p < 0.05$

OGI training sets using each domain of speech tokenizers, and evaluate on test sets across different styles (the scripted OGI test set and the spontaneous MyST test set) and age groups (4–7, 8–10, and 11–15 years old). The WER results are summarized in Table IV.

As expected, the best performance on both test set styles is achieved using an in-domain tokenizer, while noticeable degradation is observed when using out-of-domain tokenizers. Notably, using a pretrained SSL model to extract speech tokens maintains a relatively good balance across the two test set styles, as it is not fine-tuned for any specific domain. Using tokenizers fine-tuned on spontaneous child speech can improve performance across both styles while still maintaining a good balance. In contrast, tokenizers fine-tuned on scripted child speech improve performance on the OGI scripted test set but significantly degrade performance on the spontaneous test set. This is because these tokenizers are overfitted to scripted child speech data, which features a more limited and less diverse vocabulary and generalizes poorly to other data.

Furthermore, while the supervised FSQ tokenizer yields overall better ASR results in the in-domain setting, its performance suffers more significantly in cross-domain scenarios compared to the unsupervised K-means tokenizer. This suggests that the supervised FSQ tokenizer may be more closely aligned with the specific characteristics of the training data, which in turn reduces its generalizability to other domains. In contrast, the K-means tokenizer is trained solely based on the distance relationships between frame-level representations, without incorporating any task-specific objectives.

Finally, for different age groups, we observe that younger children’s speech (4–7 years) is more challenging than that of older children. Due to the imperfect development of the vocal tract, their pronunciation differs significantly from older age groups. These findings highlight the importance of developing

TABLE IV: Evaluation of speech tokenizers trained on data from different domains, evaluated across styles (Scripted, Spontaneous) and age groups (4–7, 8–10, 11–15 years) in terms of WER (%).

	WER ↓			MyST (Spon)
	Age 4-7	OGI (Scripted)	Age 11-15	
<i>Training Domain of Unsupervised K-means Tokens</i>				
Pretrained SSL	11.0	2.5	1.8	11.7
MyST (Spon)	10.1	1.9	1.3	10.7
OGI (Scripted)	7.3	1.0	0.9	78.2
<i>Training Domain of Supervised FSQ Tokens</i>				
MyST (Spon)	12.0	2.4	1.4	10.0*
OGI (Scripted)	6.3*	0.8*	0.6*	88.0

*: Statistical significance is confirmed with $p < 0.05$

discrete speech tokenizers that are not only effective within a domain but also robust and adaptable across diverse speech styles and age groups. Future research may explore domain-agnostic and age-invariant training strategies and representations to enhance transferability in practical applications.

V. CONCLUSION

This paper presents a comprehensive comparative analysis between K-means-based unsupervised and FSQ-based supervised semantic speech tokens for low-resource child ASR, evaluating their ASR performance, bitrate efficiency, codebook distribution, and cross-domain generalization. Experiments on the MyST and OGI datasets demonstrate that FSQ-based supervised speech tokens not only surpass K-means-based unsupervised tokens but also unexpectedly outperform continuous representations from SFMs. Experimental results and analysis of codebook distributions further reveal that discrete speech tokens remain effective on child ASR even in ultra-low bitrate settings. Cross-domain evaluations, however, highlight a key limitation: both tokenizers show limited generalizability across speaking styles and age groups. In future work, we plan to investigate alternative speech foundation models and datasets to obtain more effective semantic tokens and to develop unified, robust, and efficient speech tokenizers aimed at improving both performance and generalization.

REFERENCES

- [1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.
- [2] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.
- [3] A. Graves, “Sequence transduction with recurrent neural networks,” *International Conference on Machine Learning (ICML) 2012 Workshop on Representation Learning*, 2012.
- [4] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” *Advances in neural information processing systems*, vol. 28, 2015.
- [5] L. Dong, S. Xu, and B. Xu, “Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5884–5888.
- [6] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, “Conformer: Convolution-augmented transformer for speech recognition,” in *Interspeech 2020*, 2020, pp. 5036–5040.
- [7] K. Kim, F. Wu, Y. Peng, J. Pan, P. Sridhar, K. J. Han, and S. Watanabe, “E-branchformer: Branchformer with enhanced merging for speech recognition,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 84–91.
- [8] J. Li *et al.*, “Recent advances in end-to-end automatic speech recognition,” *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.
- [9] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [10] W.-N. 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 transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.
- [11] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [12] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, “Superb: Speech processing universal performance benchmark,” in *Interspeech 2021*, 2021, pp. 1194–1198.
- [13] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, “Soundstream: An end-to-end neural audio codec,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.
- [14] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, “High fidelity neural audio compression,” *Transactions on Machine Learning Research*, 2023.
- [15] X. Chang, B. Yan, Y. Fujita, T. Maekaku, and S. Watanabe, “Exploration of efficient end-to-end asr using discretized input from self-supervised learning,” in *Interspeech 2023*, 2023, pp. 1399–1403.
- [16] X. Chang, B. Yan, K. Choi, J.-W. Jung, Y. Lu, S. Maiti, R. Sharma, J. Shi, J. Tian, S. Watanabe *et al.*, “Exploring speech recognition, translation, and understanding with discrete speech units: A comparative study,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 481–11 485.
- [17] Y. Yang, F. Shen, C. Du, Z. Ma, K. Yu, D. Povey, and X. Chen, “Towards universal speech discrete tokens: A case study for asr and tts,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 401–10 405.
- [18] X. Chang, J. Shi, J. Tian, Y. Wu, Y. Tang, Y. Wu, S. Watanabe, Y. Adi, X. Chen, and Q. Jin, “The interspeech 2024 challenge on speech processing using discrete units,” in *Interspeech 2024*, 2024, pp. 2559–2563.
- [19] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, “Neural codec language models are zero-shot text to speech synthesizers,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 705–718, 2025.
- [20] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma *et al.*, “Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens,” *arXiv preprint arXiv:2407.05407*, 2024.
- [21] Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang *et al.*, “Cosyvoice 2: Scalable streaming speech synthesis with large language models,” *arXiv preprint arXiv:2412.10117*, 2024.
- [22] S. Lloyd, “Least squares quantization in pcm,” *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.
- [23] F. Mentzer, D. Minnen, E. Agustsson, and M. Tschannen, “Finite scalar quantization: VQ-VAE made simple,” in *The Twelfth International Conference on Learning Representations*, 2024.
- [24] W. Ward, R. Cole, D. Bolanos, C. Buchenroth-Martin, E. Svirsky, S. V. Vuuren, T. Weston, J. Zheng, and L. Becker, “My science tutor: A conversational multimedia virtual tutor for elementary school science,” *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 7, no. 4, pp. 1–29, 2011.
- [25] K. Shobaki, J.-P. Hosom, and R. Cole, “The ogi kids speech corpus and recognizers,” in *6th International Conference on Spoken Language Processing (ICSLP 2000)*, 2000, pp. vol. 4, 258–261.
- [26] R. Fan, N. B. Shankar, and A. Alwan, “Benchmarking children’s asr with supervised and self-supervised speech foundation models,” in *Interspeech 2024*, 2024, pp. 5173–5177.
- [27] V. N. Sukhadia and S. A. Chowdhury, “Children’s speech recognition through discrete token enhancement,” in *Interspeech 2024*, 2024, pp. 5143–5147.
- [28] S. Dutta, D. Irvin, and J. H. Hansen, “Exploring discrete speech units for privacy-preserving and efficient speech recognition for school-aged and preschool children,” *International Journal of Human-Computer Studies*, p. 103460, 2025.
- [29] K. Choi, A. Pasad, T. Nakamura, S. Fukayama, K. Livescu, and S. Watanabe, “Self-supervised speech representations are more phonetic than semantic,” in *Interspeech 2024*, 2024, pp. 4578–4582.
- [30] A. A. Attia, J. Liu, W. Ai, D. Demszky, and C. Espy-Wilson, “Kid-whisper: Towards bridging the performance gap in automatic speech recognition for children vs. adults,” in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, 2024, pp. 74–80.
- [31] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, “Esnet: End-to-end speech processing toolkit,” in *Interspeech 2018*, 2018, pp. 2207–2211.
- [32] D. Wang, M. Cui, D. Yang, X. Chen, and H. Meng, “A comparative study of discrete speech tokens for semantic-related tasks with large language models,” *arXiv preprint arXiv:2411.08742*, 2024.