Probing Discrete Speech Tokens of Spoken Language Models

Sven Naber, Julia Koch,
Pranav Singh, Alberto Saponaro, Ioanna Karagianni, Ngoc Thang Vu
Institute for Natural Language Processing (IMS) University of Stuttgart, Germany
{sven.naber, julia.koch, pranav.singh, alberto.saponaro, ioanna.karagianni, thang.vu}@ims.uni-stuttgart.de

Abstract

This paper presents a framework for systematic probing of discrete speech token representations in spoken language models (SLMs). We propose three complementary components: a distributional divergence analysis testing whether an attribute is reflected in token usage, token-based classifiers to quantify recoverability and an attribute-conditioned representation analysis revealing phonetic attribute realizations. As a demonstration we apply these probes to tokenizer outputs and model generations from CosyVoice2 and SparkTTS on LibriTTS-R and VCTK. We find that gender is encoded in their respective tokens but in different forms - the signal is more stable across stages and datasets in CosyVoice2, whereas SparkTTS shows weaker cross-stage consistency and stronger pause/prosody-related effects. Exploratory probes of valence, arousal, and dominance are weaker and less consistent. These results show that discrete speech tokens retain speaker-related information in different ways across architectures and that the proposed framework provides an interpretable basis for comparing token representations across spoken language modeling pipelines.

Keywords: Spoken language models, discrete speech tokens, model-agnostic probing methods

1. Introduction

Spoken Language Models (SLMs) have recently emerged as a framework for unifying speech understanding and generation by extending language-model style token prediction to spoken input and output (Arora et al., 2025). Many such systems operate on *discrete speech tokens*, obtained by quantizing continuous speech with neural codecs or self-supervised speech encoders (Défossez et al., 2024; Hsu et al., 2021). These tokens are not merely an implementation detail: they are the interface through which speech information enters the model and, in generation pipelines, a key intermediate representation through which that information is transformed before waveform synthesis.

Because discrete speech tokens act as an information pathway (Du et al., 2024b) and in some pipelines as an information bottleneck (Défossez et al., 2024), their structure constrains what an SLM can preserve, discard, or control. If speaker-related or phonetic information is absent from the token sequence, later components cannot reliably recover it. Conversely, if such information remains encoded in the tokens, it may influence generation even when the architecture is intended to factor it out into separate conditioning signals. Despite their central role, it is still not well understood which attributes are represented in discrete speech tokens, how those attributes are organized, and how token representations change between tokenizer outputs and tokens later generated by the SLM.

Prior work has mainly analyzed tokenization stages in isolation (Wells et al., 2022; Sadok et al., 2025) or evaluated token quality indirectly through downstream tasks such as phone classification or speaker verification (Chung et al., 2019). This leaves two important gaps. First, we lack direct analyses of token representations across multiple stages of a spoken language modeling pipeline. Second, we know little about how stable attribute-related token patterns are across datasets and across architectures with different conditioning mechanisms. These questions are particularly relevant, e.g., for controllable speech generation systems, where speaker-related information may be distributed across semantic tokens, reference signals, and separate global embeddings.

In this work, we propose a probing framework for discrete speech tokens and apply it to two token-based speech generation pipelines, CosyVoice2 (Du et al., 2024b) and SparkTTS (Wang et al., 2025). We study three token sets: `tokenizer` outputs extracted directly from audio, reference-conditioned (`zero-shot`) tokens generated from text plus reference speech, and `text-only` tokens generated from text alone. Our analysis asks three questions: **(1)** Which phonetic and speaker-related properties are encoded in these token sequences? **(2)** How do token representations change from tokenizer outputs to SLM-generated tokens? **(3)** How consistent are these patterns across datasets and models?

To answer these questions, we combine three complementary probes in our proposed probing framework. First, a distributional divergence analysis tests whether token usage differs across attribute groups. Second, token-based classifiers quantify how well attributes can be recovered from token identities. Third, an attribute-conditioned token–phoneme analysis examines how token–phone associations vary across attribute groups in

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2344–2354
11–16 May 2026.

a shared phonetic space. Together, these probes provide coarse-to-fine evidence about whether an attribute is reflected in token distributions, recoverable from sequences, and realized in phonetic structure.

Our experiments on LibriTTS-R and VCTK show a clear pattern for gender. In both models, gender is strongly recoverable from `tokenizer` and `zero-shot` tokens, while `text-only` generations are at or near chance. However, the organization of this signal differs across architectures. CosyVoice2 exhibits a more stage-stable token-level gender code, with recurring markers across datasets and strong similarity between `tokenizer` and `zero-shot` tokens. SparkTTS also preserves gender information, but in a weaker and more probe-sensitive form that is more strongly tied to pause and prosodic structure. Exploratory analyses of valence, arousal, and dominance are substantially weaker and less consistent.

In sum, our contributions are as follows: 1) we introduce a probing framework for comparing discrete speech tokens across multiple stages of token-based spoken language modeling pipelines; 2) we show that analyzing tokenizer outputs and SLM-generated tokens side by side reveals which speaker-related information survives conditioning and generation; 3) we provide evidence that different architectures organize similar attributes differently in token space, with implications for interpretability, controllability, and representation design in spoken language models.

2. Related Work

An attempt at analyzing information content of speech audio representations has been made by Wells et al. (2022) who showed that speech tokens derived from HuBERT (Hsu et al., 2021) can be assigned to articulatory properties of phonemes. However, their analyses do not include insights into prosodic or paralinguistic attributes. Sadok et al. (2025) analyze how content, speaker identity, and pitch are encoded in speech tokens. They investigated how neural audio codecs encode these attributes, revealing that such features are often entangled and difficult to interpret. Although they mapped acoustic tokens extracted from codecs to semantic (also called phonetic) tokens to help with interpretability, their conclusions mainly referenced acoustic tokens and may not generalize to semantic tokens as differentiated in Zhang et al. (2024).

Chung et al. (2019) present a model for speech representation learning suitable for a wide range of downstream tasks. They evaluate the information conserved in their tokens on the tasks of phone classification and speaker verification and further analyze how speaker information is encoded across different layers of their model. On a similar note, Onda et al. (2026) propose a speech tokenizer that combines the advantages of acoustic and phonetic tokens, aiming to preserve both phonetic and prosodic information while discarding speaker identity. They evaluate performance of their Phonological Tokenizer on various downstream tasks, including some involving an SLM for speech continuation. Chang et al. (2025) point out that SLM-optimized tokens need to fulfill mainly three criteria: preservation of phonetic and semantic information, retention of acoustic details, and robustness to perturbations. Similar to Chung et al. (2019) and Onda et al. (2026), they assess the information retained in the tokens via downstream tasks instead of looking at the tokens directly.

For our attribute-conditioned representation analysis we take inspiration from Lin et al. (2024b) who propose a similar analysis to identify property neurons in self-supervised speech transformers, i.e. neurons selectively responsive to specific attributes. While they use this method to examine the inner workings of transformer models, we instead apply our statistical analysis directly on the raw speech tokens to gain insights on how phoneme and speaker attributes are structured in the tokens.

3. Spoken Language Models

3.1. CosyVoice2

In this work, we use CosyVoice2 (Du et al., 2024b) as an example SLM. Its architecture consists of the following components.

1) Text tokenizer: The raw text for each utterance is first tokenized using Byte-Pair Encoding.

2) Semantic speech tokenizer: The audio file is converted into semantic speech tokens generated at a rate of 25 tokens per second.

3) Text-speech language model: The pre-trained textual Large Language Model Qwen2.5-0.5B takes the text tokens and semantic speech tokens as input and generates semantic speech tokens which unify the text and audio information as the output.

4) Flow matching: The causal flow-matching model takes Qwen's output, now "text-aware" semantic speech tokens, and the semantic speech tokens from the reference speech as its initial input. The speaker embedding from the reference speech is then used to generate a Mel spectrogram.

5) Vocoder: Finally, a Vocoder is used to synthesize the speech signal from the Mel spectrogram.

In the `zero-shot` mode the LM is conditioned on the prompt text and prompt speech tokens extracted from a reference utterance together with the target text, and autoregressively predicts the target semantic speech tokens (Du et al., 2024b).

3.2. SparkTTS

We additionally use SparkTTS (Wang et al., 2025) as a second example SLM. Its architecture consists of the following components.
1) Text tokenizer: The raw text for each utterance is first tokenized using Byte-Pair Encoding.
2) BiCodec speech tokenizer: The audio file is converted into two token streams: semantic speech tokens generated at a rate of 50 tokens per second and a fixed-length sequence of global tokens capturing speaker and other global speech characteristics.
3) Text-speech language model: The pre-trained textual Large Language Model Qwen2.5-0.5B takes the text tokens and conditioning speech tokens as input and generates semantic speech tokens.
4) BiCodec decoder: Finally, the decoder reconstructs the speech signal directly from the semantic and global tokens.

In the `zero-shot` mode the LM is in contrast to CosyVoice2 conditioned on the target text together with global tokens extracted from a reference utterance and then autoregressively predicts the target semantic speech tokens.

4. Probing Methods

4.1. Preprocessing

For each utterance, we align text and audio to obtain an aligned phoneme sequence $p_i = (p_{i1}, \ldots, p_{in})$ from which co-occurrence embeddings are computed (cf. Sections 4.1.1, 4.4). Associated attribute labels a_i are extracted from audio only (cf. Section 5.2). A discrete token sequence $t_i = (t_{i1}, \ldots, t_{in})$ is obtained either by extracting it from the respective pipeline stages or audio via the observed model's tokenizer (cf. Section 5.1).

4.1.1. MFA alignment

Phoneme alignment is performed with Montreal Forced Alignment (MFA) (McAuliffe et al., 2017) which gives phone intervals that are then discretized to the token time grid by center-time sampling, yielding equal-length token and phoneme sequences $(t_{i1}, p_{i1}), \ldots, (t_{in}, p_{in})$. This alignment enables the construction of phoneme-based embeddings for each token, used in the attribute-conditioned representation analysis.

4.2. Distributional divergence analysis

For each token set T and attribute group a, we estimate empirical token distributions $P_T(t \mid a)$ from normalized token counts. We then compute the Jensen–Shannon divergence (JSD) between high (H) and low (L) subsets to quantify distributional differences. Per-token frequency differences,

$$\Delta_T(t) = P_T(t \mid \mathrm{H}) - P_T(t \mid \mathrm{L}), \tag{1}$$

highlight tokens associated with either group. Cross-token-set JSDs provide a measure of consistency across token representations. We include a randomized baseline (20 label shuffles) to estimate expected divergence under the null hypothesis and report JSD mean. Only tokens with corpus frequency of at least 50 are included in the divergence vocabulary.

4.3. Token-based classifiers

The second component quantifies attribute recoverability from token identities. Utterances are represented as bags of discrete tokens, and a logistic regression classifier is trained to predict the high/low attribute bin. Three feature modes are used: raw token counts (*BOW*), L_1-normalized token shares (*SHARE*), and binary presence indicators (*SET*). For all classifier runs, token filtering is applied on the training split only: we retain only tokens that occur in both classes and whose training frequency is at least 0.002% of all training token occurrences. This reduces overfitting to rare tokens while keeping filtering comparable across datasets of different sizes.

Classifiers are evaluated on held-out data from 80/20 speaker-disjoint train/test splits, and token coefficients indicate the direction and strength of association with the attribute. Because speaker-disjoint splits are not perfectly class-balanced, we report balanced accuracy (mean recall across classes) as the primary metric.

4.4. Attribute-conditioned representation analysis

The final component examines attribute variation in the phonetic domain. Each token t is aligned with phones p within its utterance. For each attribute bin $b \in \{\mathrm{H}, \mathrm{L}\}$, we estimate the conditional phone distribution

$$P(p \mid t, b) = \frac{c(t, p, b)}{\sum_{p'} c(t, p', b)},$$

where $c(t, p, b)$ denotes the co-occurrence count of token t with phone p in bin b. Tokens are retained only if their total paired count is at least 400 and the class balance ratio satisfies $\min(c_H, c_L)/\max(c_H, c_L) \geq 0.2$. Phonetic divergence between bins is measured by the cosine similarity between the corresponding phone distributions:

$$\cos(t) = \mathrm{cosine}\big(P(\cdot \mid t, \mathrm{H}),\, P(\cdot \mid t, \mathrm{L})\big).$$

Low $\cos(t)$ values indicate strong attribute-dependent phonetic differences. In addition, for each token we compute per-phone differences

$$\Delta_t(p) = P(p \mid t, \mathrm{H}) - P(p \mid t, \mathrm{L}),$$

and define a signed bias score

$$\lambda(t) = \Delta_t(p^*), \qquad p^* = \arg\max_p |\Delta_t(p)|.$$

A token is considered biased if $|\lambda(t)| \geq 0.015$, where the sign indicates the direction of the bias. These token–phone embeddings capture how discrete tokens reflect phonetic variation and enable comparison across token sets in a shared phoneme space.

5. Experimental Setup

5.1. Token sets

For each utterance in the datasets we extract discrete speech token sequences at different stages of the spoken language model pipeline:

`tokenizer` tokens are obtained by passing the utterances to the respective tokenizer.

`zero-shot` tokens are generated by the SLM via zero-shot synthesis from the utterance text, using the original utterance audio as the reference signal (same-audio-conditioning).

`text-only` tokens are generated by the SLM from the utterance text only without reference audio.

Out of these three sets only `tokenizer` tokens afford the creation of an aligned phoneme sequence as it is derived from audio directly whereas the others are LM generations.

5.2. Attribute labels

The used datasets inherently provided gender labels for the utterances. To expand the analysis to other labels, we used audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim [1] to automatically label the utterances with valence, arousal and dominance scores.

6. Results

6.1. Probing of Gender

We present the results of the three probing methods on LibriTTS-R (Zen et al., 2019) and VCTK (Veaux et al., 2019). We analyze which tokens carry gender-specific information, how this differs between tokenizer and SLM outputs, and how robust the associations are across datasets.

[1] https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim

6.1.1. Distributional divergence analysis

CosyVoice2 Tokenizer and reference-conditioned SLM tokens show clear gender separation relative to their shuffled baselines. In LibriTTS-R, `zero-shot` shows stronger separation than `tokenizer` (0.0614 vs. 0.0385). In VCTK, both remain far above shuffle (`tokenizer`: 0.1183/0.0011; `zero-shot`: 0.1080/0.0009), while `text-only` shows minimal divergence (0.0013 vs. 0.0012). Several recurring token IDs, notably 1950, 2031, 4137, 4299, 2112, and 3975, reappear across datasets and across `tokenizer` and `zero-shot`, indicating stable encoding of speaker-dependent cues.

SparkTTS Tokenizer and reference-conditioned SLM tokens also show gender separation relative to their shuffled baselines. In LibriTTS-R, `zero-shot` again shows stronger separation than `tokenizer` (0.0162 vs. 0.0079). In VCTK, both remain above shuffle (`tokenizer`: 0.0107/0.0010; `zero-shot`: 0.0227/0.0015), while `text-only` remains at baseline (0.0009 vs. 0.0009). Recurring token IDs are less consistent: in LibriTTS-R, female-associated tokens such as 908 and 791 recur across `tokenizer` and `zero-shot`, while in VCTK the clearest repeated markers are 6182 on the male side and 857 on the female side.

Cross-model comparison Both models show distributional gender separation for `tokenizer` and `zero-shot`, while `text-only` remains at or near the shuffled baseline, indicating that the signal is primarily introduced by speaker conditioning rather than text. The effect is stronger and more stage-stable in CosyVoice2, which also shows more consistent recurring token IDs across datasets and token sources. SparkTTS exhibits the same overall pattern, but with weaker and less stable markers.

6.1.2. Token-based classifier probes

CosyVoice2 Gender is highly retrievable from `tokenizer` and `zero-shot` tokens in CosyVoice2 (Table 1), with balanced accuracy above 0.9 for BOW, SHARE, and SET on both datasets. The normalized SHARE weights are also highly similar between `tokenizer` and `zero-shot` (cosine 0.8940 on LibriTTS-R and 0.9625 on VCTK), indicating that the gender-related token associations remain stable between the tokenizer and the speaker-conditioned SLM. On VCTK, `text-only` drops to chance (0.50). The most predictive tokens are also stable across datasets and stages: 1950 and 4137 act as strong male indicators, while 4299 and 2112 consistently mark female speech.

Dataset	Token Set	Top Tokens (M/F)	Bal. Acc. (B/S/Set)
LibriTTS-R	`tokenizer`	1950, 4137, 2031 / 4299, 6374, 2112	0.917 / 0.909 / 0.920
LibriTTS-R	`zero-shot`	1950, 4137, 3651 / 4299, 2112, 3975	0.927 / 0.932 / 0.929
VCTK	`tokenizer`	1950, 4137, 2031 / 4299, 3975, 2112	0.977 / 0.950 / 0.972
VCTK	`zero-shot`	1950, 4137, 2031 / 4299, 2112, 3975	0.951 / 0.948 / 0.950

Table 1: Top gender-predictive tokens (male / female) by classifier weights (SHARE) with corresponding balanced accuracies for BOW, SHARE, and SET on CosyVoice2. Unique colors indicate recurring tokens across analyses.

Dataset	Token Set	Top Tokens (M/F)	Bal. Acc. (B/S/Set)
LibriTTS-R	`tokenizer`	528, 3676, 4690 / 7663, 7516, 248	0.842 / 0.717 / 0.842
LibriTTS-R	`zero-shot`	7346, 8095, 6905 / 908, 6689, 5610	0.892 / 0.842 / 0.894
VCTK	`tokenizer`	*no stable markers*	0.789 / 0.500 / 0.790
VCTK	`zero-shot`	*no stable markers*	0.857 / 0.529 / 0.856

Table 2: Top gender-predictive tokens (male / female) by classifier weights (SHARE) with corresponding balanced accuracies for BOW, SHARE, and SET on SparkTTS.

SparkTTS Gender is still retrievable from `tokenizer` and `zero-shot` tokens in SparkTTS (Table 2), but less uniformly across probe types. BOW and SET remain above chance on both datasets, whereas SHARE is lower on LibriTTS-R (0.717 for `tokenizer`, 0.842 for `zero-shot`) and at or near chance on VCTK (0.500 and 0.529). The normalized SHARE weights between `tokenizer` and `zero-shot` are also only moderately similar, with cosine similarities of 0.7745 on LibriTTS-R and 0.5223 on VCTK. On VCTK, `text-only` again drops to chance (0.49–0.50). Stable predictive tokens are sparse: in LibriTTS-R there is limited overlap between `tokenizer` and `zero-shot`, such as token 300 on the male side and 908 on the female side, while in VCTK there are no stable markers.

Cross-model comparison The classifier probe sharpens the difference between the two models. In CosyVoice2, gender is robustly recoverable across all three feature sets, the top predictive tokens recur across datasets and stages, and the SHARE weights remain highly stable between `tokenizer` and `zero-shot`. In SparkTTS, gender is still recoverable, but mainly through BOW and SET; SHARE weakens substantially, especially on VCTK, and predictive tokens are much less stable across stages and datasets. In both models, `text-only` falls to chance, again indicating that the classifier signal depends on speaker conditioning rather than text alone.

6.1.3. Attribute-conditioned representation analysis

For this analysis, we focus on `tokenizer` tokens, since these are aligned to phoneme sequences. The probe again shows gender-dependent phonetic divergence, but now at the level of token–phoneme realizations.

CosyVoice2 In CosyVoice2, the strongest biased tokens are not simply the same recurring markers from the divergence and classifier probes. In LibriTTS-R, the top male-biased tokens (2168, 4761, 1260, 6485, 558) are associated with phones such as /AO1/, /Z/, /DH/, and /AA1/, while the top female-biased tokens (1546, 1547, 6105, 1555, 4146) are predominantly associated with /sp/ (silence/non-speech). In VCTK, the strongest male- and female-biased tokens are likewise dominated by /sp/ (male: 2939, 3749, 3668; female: 1616, 1535, 875).

SparkTTS In SparkTTS, the strongest biased tokens are dominated even more consistently by /sp/. In LibriTTS-R, both the top male-biased and top female-biased tokens are predominantly associated with /sp/. The same pattern holds in VCTK, where male-biased and female-biased tokens are again dominated by /sp/.

Cross-model comparison Both models still encode speaker-related information at the token level, but in the attribute-conditioned probe the effect looks less like a small set of stable segmental gender markers and more like differences in token–phoneme realizations across groups. In CosyVoice2, this includes some non-pause segmental associations in LibriTTS-R, but becomes strongly /sp/-dominated in VCTK, suggesting an important pause/prosodic component. In SparkTTS, the /sp/ dominance holds across both datasets, pointing even more clearly to pause/prosodic token realizations. This is consistent with the weaker and less stable gender signal observed for SparkTTS in the divergence and classifier probes.

Model	Dataset	Valence	Arousal	Dominance
CosyVoice2	LibriTTS-R	0.514 / 0.504 / –	0.557 / 0.540 / –	0.525 / 0.545 / –
SparkTTS	LibriTTS-R	0.521 / 0.527 / –	0.522 / 0.516 / –	0.522 / 0.519 / –
CosyVoice2	VCTK	0.556 / 0.541 / 0.513	0.536 / 0.562 / 0.521	0.635 / 0.637 / 0.549
SparkTTS	VCTK	0.541 / 0.589 / 0.505	0.547 / 0.582 / 0.524	0.688 / 0.587 / 0.521

Table 3: Best balanced accuracy across BOW/SHARE/SET for exploratory probes of model-predicted VAD. Values are reported as `tokenizer / zero-shot / text-only`.

6.2. Probing of Valence, Arousal and Dominance (VAD)

We perform an analogous but exploratory probe for model-predicted valence, arousal, and dominance. For each dimension, we partition each dataset into equally sized low and high groups using the 0.2 and 0.8 quantiles and train the classifiers to predict these groups. Since the score distributions differ substantially between LibriTTS-R and VCTK, we interpret the results within each dataset rather than directly across corpora.

Across both models, VAD is encoded much more weakly and less consistently than gender. For `tokenizer` and `zero-shot`, distributional divergence is typically above shuffle, but classifier performance mostly stays near chance, especially on LibriTTS-R. The clearest exception is dominance on VCTK: CosyVoice2 reaches 0.635/0.637 for `tokenizer`/`zero-shot`, and SparkTTS reaches 0.688/0.587. Valence and arousal are weaker, with best balanced accuracy generally between 0.50 and 0.59. In most VCTK settings, `text-only` remains close to chance, although CosyVoice2 dominance retains a small residual signal (0.549).

The VAD signal is also less stage-stable than gender. The cosine similarity between SHARE coefficients for `tokenizer` and `zero-shot` ranges from 0.411 to 0.695 in CosyVoice2 and from 0.231 to 0.516 in SparkTTS, suggesting that the recovered signal is not organized as a robust recurring token-identity code. The attribute-conditioned analyses further support a largely prosodic interpretation: in VCTK and throughout SparkTTS, the strongest high/low-biased tokens are predominantly associated with /sp/, while CosyVoice2 on LibriTTS-R shows somewhat more segmental patterns in some dimensions. Overall the signal is weak, model- and corpus-dependent, and may partly reflect speaker-related or labeling confounds rather than a stable affective representation.

7. Analysis of Gender Encoding

Section 6.1 revealed systematic gender signals in the tokens of both models. Because the probes provide correlational rather than causal evidence, we conduct additional controlled analyses to test whether these effects remain robust under targeted ablations.

7.1. Summary of cross-model differences

Across the three probes, four differences stand out. First, CosyVoice2 exhibits a stronger and more stage-stable token-level gender code: gender remains robust under both count/presence and normalized-share probes, and recurring markers persist across datasets and stages. Second, SparkTTS shows a more probe-sensitive pattern: gender is still recoverable, mainly through BOW and SET, while SHARE is weaker and cross-stage consistency lower. Third, `zero-shot` tends to be stronger than `tokenizer` in SparkTTS, whereas the two stages are more similar in CosyVoice2. Fourth, the attribute-conditioned analysis suggests different encoding forms: SparkTTS is dominated by /sp/ and appears more pause/prosody-driven, while CosyVoice2 retains clearer non-pause segmental differences, especially in LibriTTS-R.

The following analyses test these differences more directly. We use shared-text subsets to separate text-invariant token-level encoding from content-sensitive effects, silence removal to measure the contribution of pause-aligned positions, a small generation probe to assess downstream influence on synthesized speech, and architectural considerations to relate the probe patterns back to the two model pipelines.

7.2. Shared-text control

To control for content confounds, we rerun the probes on shared-text subsets, i.e. settings with matched text. Table 4 reports the resulting balanced accuracies. CosyVoice2 LibriTTS-R is included for completeness but should be interpreted cautiously, since only 314 matched examples remain after balancing.

On VCTK, text balancing leaves CosyVoice2 almost unchanged, indicating that its token-level gender signal is largely text-invariant.In SparkTTS, balancing improves performance substantially, especially for SHARE (`tokenizer`: 0.500 $\rightarrow$ 0.806; `zero-shot`: 0.529 $\rightarrow$ 0.703), suggesting that lexical variability in the original split masked part of a genuine normalized-share gender code.

Model	Dataset	Entries	tokenizer (B/S/Set)	zero-shot (B/S/Set)	text-only (B/S/Set)
CosyVoice2	VCTK	24,332	0.977 / 0.946 / 0.969	0.948 / 0.950 / 0.947	0.497 / 0.493 / 0.503
CosyVoice2	LibriTTS-R†	314	0.800 / 0.733 / 0.717	0.817 / 0.800 / 0.867	–
SparkTTS	VCTK	24,896	0.816 / 0.806 / 0.816	0.897 / 0.703 / 0.897	0.506 / 0.500 / 0.504
SparkTTS	LibriTTS-R	2,476	0.589 / 0.574 / 0.587	0.637 / 0.627 / 0.622	–

Table 4: Gender probe balanced accuracy on shared-text subsets using BOW/SHARE/SET features. †CosyVoice2 LibriTTS-R uses only 314 matched examples and should therefore be interpreted with caution.

On LibriTTS-R, the shared-text subset is much smaller. SparkTTS drops sharply to 0.589/0.574 for `tokenizer` and 0.637/0.627 for `zero-shot`, indicating that the original LibriTTS-R signal is less text-invariant and more entangled with content.

7.2.1. Silence-removal control

The attribute-conditioned analysis suggested that many of the strongest gender-biased token–phone pairs were dominated by /sp/. To test whether pause-aligned positions carry the main signal or only one component of it, we perform two ablations. First, for the attribute-conditioned probe, we remove pause phones from the token–phone pairs. Second, for the divergence and classifier probes, we remove from `tokenizer` all positions aligned to `sil/sp/spn` and recompute the probes.

Without pause phones, CosyVoice2 still shows a substantial non-pause gender signal. In LibriTTS-R, the strongest male-biased tokens remain segmental (/AO1/, /Z/, /DH/, /AA1/), while the female-biased side shifts from pause-dominated tokens to /N/, /ER0/, /K/, and /HH/. In VCTK, the original /sp/-dominated pattern is weakened, but residual non-pause biases remain around /HH/, /AH0/, /IY1/ on the male side and /DH/, /L/, /K/ on the female side. SparkTTS is much more pause-sensitive. In LibriTTS-R, the remaining non-pause effects are mainly around /S/ and /AH0/ on the male side and /D/ and /N/ on the female side; in VCTK they are similarly small and concentrate around /AH0/, /S/, /N/ versus /T/, /AH0/, /N/.

Table 5 confirms the same pattern. CosyVoice2 retains a strong non-pause gender signal: on LibriTTS-R, filtering changes the `tokenizer` probes only marginally, and on VCTK the signal weakens but remains far above chance across all three classifier variants. SparkTTS is more fragile. On LibriTTS-R, removing silence-aligned positions leaves JSD, BOW, and SET almost unchanged but collapses SHARE to chance, indicating that the normalized token-share signal was disproportionately carried by pause-aligned positions. On VCTK, removing silence slightly improves JSD, BOW, and SET while leaving SHARE at chance, suggesting that the surviving signal is primarily count/presence-based rather than a stable normalized token mix.

Taken together, the silence-removal control shows that pause/prosodic positions contribute in both models, but much more strongly in SparkTTS. CosyVoice2 retains a clearer non-pause segmental gender code after silence removal, whereas SparkTTS depends more on pause-aligned positions or on weaker, more diffuse non-silence cues.

7.3. Generation experiment

Both model pipelines (see 3.1, 3.2) regenerate audio from a token sequence and a fixed-length global embedding (`ref`) capturing global speech properties. Since these are highly relevant for gender, we test the downstream influence of each component in a small qualitative probe. For both models, we generate speech in four settings:

1. male token sequence + male `ref` embedding
2. female token sequence + female `ref` embedding
3. male token sequence + female `ref` embedding
4. female token sequence + male `ref` embedding

For each setting and each model, we generate 25 samples by randomly pairing token sequences and `ref` embeddings drawn from male and female speakers. In SparkTTS, samples are synthesized with the BiCodec decoder, whereas in CosyVoice2 generation uses the flow-matching model followed by the vocoder. The resulting audio samples are shuffled and evaluated in a qualitative listening test by an expert listener, who annotates the perceived speaker gender for each sample.

In both models, the matched settings (1) and (2) produce the expected male and female outputs. In CosyVoice2, the perceived gender in the mismatched settings is consistently determined by the `ref` embedding: setting (3) (male tokens + female `ref`) yields speech perceived as female, while setting (4) yields speech perceived as male. All four settings produce high-quality speech. This suggests that although gender information is highly retrievable from CosyVoice2 tokens, the reference embedding rather than the token sequence primarily determines the realized speaker gender during generation.

Model	Dataset	Removed	JSD	BOW	SHARE	SET
CosyVoice2	LibriTTS-R	6.0%	0.0385 → 0.0387	0.917 → 0.907	0.909 → 0.895	0.920 → 0.912
CosyVoice2	VCTK	29.2%	0.1183 → 0.0826	0.977 → 0.941	0.950 → 0.918	0.972 → 0.939
SparkTTS	LibriTTS-R	6.1%	0.0079 → 0.0080	0.842 → 0.838	0.717 → 0.500	0.842 → 0.837
SparkTTS	VCTK	29.1%	0.0107 → 0.0149	0.789 → 0.801	0.500 → 0.500	0.790 → 0.799

Table 5: Effect of removing silence-aligned tokenizer positions. Values show before → after filtering for `tokenizer` only. BOW, SHARE, and SET denote balanced accuracy.

In SparkTTS, the mismatched settings (3) and (4) produce speech that often exhibits a mixture of male and female characteristics.

7.4. Architectural interpretation

The generation experiment shows that SparkTTS tokens contribute to the realized gender in the generated audio, even though the signal is less stable under the SHARE probe than in CosyVoice2. All three probes operate on token identities and unigram statistics and therefore primarily capture local rather than contextual information.

A plausible explanation is therefore not just a difference in how much gender information is present, but in how it is organized. CosyVoice2 behaves like a model with a direct and stage-stable token-level gender code: BOW and SHARE are both strong, `tokenizer` and `zero-shot` remain highly similar, and VCTK changes little under text balancing. SparkTTS instead behaves like a model with a more diffuse, context-sensitive token code: BOW remains strong, but SHARE is weaker, cross-stage consistency is lower, and the shared-text control shows greater sensitivity to lexical composition. The information is distributed less as a stable normalized token mix and more as sparser count/presence and prosody-related cues.

This interpretation is also compatible with the tokenizer architectures. SparkTTS has twice the temporal resolution of CosyVoice2, which may make gender less recoverable from individual tokens. In addition, the CosyVoice2 tokenizer is trained with an ASR supervision objective and is not explicitly encouraged to remove speaker information from the resulting sequence. By contrast, the SparkTTS tokenizer jointly produces the fixed-length global embedding (`ref`) and the semantic token sequence, encouraging factorization with speaker attributes partly assigned to a separate representation bucket. This separation is not perfect: our probes still recover gender from SparkTTS tokens, and the generation experiment confirms that both tokens and `ref` influence gender realization.

8. Conclusion

The framework analyzes token sequences with three complementary probes that reveal how information is encoded in speech tokens at different levels of evidence: (1) distributional evidence, testing whether an attribute is reflected in token usage; (2) recoverability evidence, assessing whether the attribute can be predicted from token sequences; and (3) phonetic realization, examining how the attribute manifests in token–phoneme associations.

Examining this across datasets, models, and pipeline stages affords a systematic comparison and insight into how information is encoded, transformed, and used inside SLM pipelines.

The comparison also illustrates that tokenizers can produce very different information structures and that attributes may remain encoded in token sequences even when the system design intends to separate them (SparkTTS) or when they are not directly used during generation (CosyVoice2).

In many SLM pipelines discrete speech tokens constitute information pathways or even bottlenecks. They should therefore not only be treated as intermediate representations but as critical design elements whose structure determines what information spoken language models can access, preserve, or control. For example, in the CosyVoice2 pipeline, although gender information is clearly encoded in tokens, control of this attribute via SLM prompting appears to be excluded by its design.

The proposed framework enables a simple, interpretable, and systematic study of such phenomena and makes speech token spaces comparable across models. Insights obtained through this approach can help explain observed model behavior and inform the design of future spoken language models. Future work could extend this framework to additional models, tokenizers, and paralinguistic attributes, as well as incorporate causal intervention experiments that directly manipulate token sequences or pipeline components to test their functional role in speech generation

Limitations

While our framework provides an interpretable and model agnostic approach for analyzing discrete speech tokens, several limitations should be acknowledged. First, all analyses are descriptive and correlational, not causal as we identify associations but do not perform any intervention or token level manipulation tests. Consequently, the framework

cannot determine whether specific tokens are actively used by the model to encode or control an attribute. Relatedly, our probes focus on token distributions and sequence level recoverability, but do not directly model sequence structure or token interactions beyond what is captured by the classifier probe.

Second, although the framework is conceptually model agnostic, the experiments in this paper are limited to two models and a restricted set of paralinguistic attributes. A broader survey across additional architectures, tokenizers, and attributes would be required to provide a more comprehensive view and to fully evaluate cross model comparability.

Lastly, evaluating attributes or labels that are not intrinsic to the datasets can introduce additional noise or confounds, especially when external models are used to infer labels.

Ethical Statement

This work analyzes how speaker attributes can be reflected in discrete speech token representations. While our experiments are conducted for analytical purposes, studying the recoverability of attributes from speech representations raises potential concerns. In particular, such analyses could be used to infer demographic or personal characteristics of speakers from intermediate model representations. Our experiments focus on gender as an illustrative attribute and rely on binary labels provided in the underlying datasets, which simplifies a complex and socially nuanced concept.

At the same time, understanding what information speech token representations expose is important for transparency and responsible system design. Analyses of this kind can help identify unintended information leakage, potential biases, or architectural design choices that implicitly preserve or discard certain attributes. Insights from such studies may therefore support the development of spoken language models with clearer control mechanisms and more predictable behavior.

All experiments were conducted on publicly available datasets and models using open-source libraries, which are cited appropriately.

Use of AI Assistants. The authors acknowledge the use of AI assistants solely for correcting grammatical errors, formatting table boundaries, and providing assistance with coding.

Code and Data Availability

The code for the probing framework and the full analysis outputs used in this study are publicly available at: https://github.com/DigitalPhonetics/SpeechTokenProbing.

Bibliographical References

Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. 2025. On the landscape of spoken language models: A comprehensive survey.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Heng-Jui Chang, Hongyu Gong, Changhan Wang, James Glass, and Yu-An Chung. 2025. DC-Spin: A Speaker-invariant Speech Tokenizer for Spoken Language Models. In *Interspeech 2025*, pages 5723–5727.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report.

Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. 2019. An Unsupervised Autoregressive Model for Speech Representation Learning. In *Interspeech 2019*, pages 146–150.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. 2024a. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *ArXiv*, abs/2407.05407.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024b. Cosyvoice 2: Scalable streaming speech synthesis with large language models.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue.

Gemini Team, Google DeepMind. 2023. Gemini 1: Unlocking multimodal understanding.

https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf.

Hao-Han Guo, Yao Hu, Kun Liu, Fei-Yu Shen, Xu Tang, Yi-Chen Wu, Feng-Long Xie, Kun Xie, and Kai-Tuo Xu. 2025. Fireredtts: A foundation text-to-speech framework for industry-level generative speech applications.

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

Yichong Leng, Zhifang Guo, Kai Shen, Xu Tan, Zeqian Ju, Yanqing Liu, Yufei Liu, Dongchao Yang, Leying Zhang, Kaitao Song, Lei He, Xiang-Yang Li, Sheng Zhao, Tao Qin, and Jiang Bian. 2023. Prompttts 2: Describing and generating voices with text prompt.

Guan-Ting Lin, Cheng-Han Chiang, and Hung yi Lee. 2024a. Advancing large language models to capture varied speaking styles and respond properly in spoken conversations.

Tzu-Quan Lin, Guan-Ting Lin, Hung-Yi Lee, and Hao Tang. 2024b. Property neurons in self-supervised speech transformers. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 401–408.

Michael McAuliffe, Michael Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. https://montreal-forced-aligner.readthedocs.io. Version 2.2.5, Accessed: 2025-10-24.

Kentaro Onda, Hayato Futami, Yosuke Kashiwagi, Emiru Tsunoo, and Shinji Watanabe. 2026. Phonological tokenizer: Prosody-aware phonetic token via multi-objective fine-tuning with differentiable k-means. *ArXiv*, abs/2601.19781.

OpenAI. 2024. Hello $\mathrm{gpt-4o}$. https://openai.com/index/hello-gpt-4o/. Accessed: 2025-10-24.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Dario Di Palma, Alessandro De Bellis, Giovanni Servedio, Vito Walter Anelli, Fedelucio Narducci, and T. D. Noia. 2025. Llamas have feelings too: Unveiling sentiment and emotion representations in llama models through probing. *ArXiv*, abs/2505.16491.

Samir Sadok, Julien Hauret, and Éric Bavu. 2025. Bringing Interpretability to Neural Audio Codecs. In *Interspeech 2025*, pages 5023–5027.

Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). https://datashare.ed.ac.uk/handle/10283/2950. Released September 2019. Licensed under Creative Commons Attribution 4.0 International.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, Weizhen Bian, Zhen Ye, Sitong Cheng, Ruibin Yuan, Zhixian Zhao, Xinfa Zhu, Jiahao Pan, Liumeng Xue, Pengcheng Zhu, Yunlin Chen, Zhifei Li, Xie Chen, Lei Xie, Yike Guo, and Wei Xue. 2025. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens.

Dan Wells, Hao Tang, and Korin Richmond. 2022. Phonetic Analysis of Self-supervised Representations of English Speech. In *Interspeech 2022*, pages 3583–3587.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. 2023. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024. Speechtokenizer: Unified speech tokenizer for speech language models. In *The Twelfth International Conference on Learning Representations*.