SALSA: Speech Aware LLM Adaptation via Learned Steering Activation Vectors

Anonymous ACL submission

Abstract

Speech-aware large language models often generalize poorly to out-of-domain settings. We propose SALSA (Speech-Aware LLM Adaptation via Learned Steering Activations), a lightweight adaptation method that learns layer-wise steering vectors. Unlike commonly used steering approaches that rely on contrastive activation differences, SALSA directly optimizes steering vectors using a supervised objective. Across children's speech, multilingual speech, and Mandarin-English code-switching benchmarks, SALSA substantially improves performance over zero-shot inference and speech in-context learning baselines, achieving up to 46.8% relative improvements over zero-shot. Analysis further demonstrates that steering the encoder, particularly the later layers, is more effective than steering the LLM backbone. These findings suggest that steering improves downstream ASR performance by adapting higher-level acoustic and phonetic representations to better align with the pretrained language model representation space, rather than by modifying the decoder itself.

1 Introduction

Speech-aware large language models (SALLMs) have emerged as a powerful paradigm for speech processing, leveraging linguistic knowledge encoded in pretrained large language models (LLMs) to advance performance across a broad range of speech tasks (Peng et al., 2025). Despite their strong performance on high-resource languages, they generalize poorly to out-of-domain scenarios, and this occurs even when the underlying model components have already been exposed to the target languages or tasks during pretraining. This suggests that adaptation in SALLMs is not solely a problem of acquiring new linguistic knowledge, but also one of aligning pretrained acoustic representations with the language model decoding space for downstream ASR tasks. Fine-tuning can address these gaps, but is computationally expensive and data-intensive. Parameter-efficient alternatives such as LoRA (Hu et al., 2021) reduce this cost but still require gradient updates to model weights. In-context learning (ICL) (Brown et al., 2020) is training-free, but is constrained by the high acoustic variability of speech. Unlike text, speech encodes speaker identity and phonetic properties alongside semantics, making it difficult to identify informative demonstration examples (Zheng et al., 2025).

A complementary method is activation steering: directly perturbing a model's internal representations at inference time to shift its behavior toward a target domain, without modifying its weights. Steering has been shown to be effective in language models (Im and Li, 2025) and has recently started to be explored for speech (Sun et al., 2026; Feng et al., 2025). Existing approaches for the speech modality rely on contrastive activation differences, where steering vectors are derived from paired contrastive examples. For automatic speech recognition (ASR), natural paired examples are difficult to obtain, unlike for semantic tasks. ASR requires a direct correspondence between the input audio and output transcript, and the high acoustic variability of speech makes contrastive pairs noisy and hard to construct. This limits the applicability of existing steering methods to exactly the low-resource, out-of-domain settings where efficient adaptation is most needed.

In this paper, we propose SALSA (**S**peech-aware LLM **A**daptation via **L**earned **S**teering **A**ctivations), a lightweight adaptation method that directly learns layer-wise steering vectors for speech-language models without requiring paired contrastive examples. By training steering vectors, SALSA can learn around acoustic variability and adapt intermediate encoder representations to better align with the pretrained language model representation space. Our contributions are as follows:

- We propose SALSA, a lightweight steering-based adaptation method that learns layer-wise encoder steering vectors for pretrained SALLMs while keeping all backbone parameters frozen.

- We demonstrate that simple learned steering vectors are highly effective for out-of-domain ASR settings, including children's speech, multilingual speech, and code-switched speech, substantially improving over zero-shot and speech in-context learning baselines without requiring paired contrastive examples.

- We analyze steering behavior across training scale, encoder depth, and intervention location, showing that adaptation is most effective in higher-level encoder representations, while steering the language model backbone provides substantially smaller gains than encoder steering.

2 Background

SALLMs, such as SALMONN (Tang et al., 2024), WavLLM (Hu et al., 2024), the Qwen-Audio family (Chu et al., 2023, 2024b), and Granite-speech (Saon et al., 2025), are typically constructed by coupling a pretrained speech encoder with an LLM backbone, where only a lightweight projection layer is trained to bridge the two modalities. These models support a broad range of tasks, from traditional speech processing tasks such as ASR to more open-ended tasks, including conversational speech and question answering. However, relative to speech-native models such as Whisper (Radford et al., 2023), SALLMs are generally trained on considerably less speech data and tend to focus on a limited set of predominantly European languages. As a result, SALLMs are prone to hallucination and generalize poorly to out-of-domain data.

2.1 Adaptation Methods for SALLMs

Adaptation through full fine-tuning or parameter-efficient methods such as LoRA (Hu et al., 2021) can address out-of-domain limitations, but such approaches require modifying model parameters and may be impractical in low-resource or computationally-constrained settings. More recently, reinforcement learning (RL)-based post-training methods have also been explored for improving speech-language models, although these approaches similarly require iterative model updates and large-scale optimization (Guo et al., 2025; Rafailov et al., 2023; Elmakies et al., 2026). A lighter-weight alternative is speech in-context learning (SICL) (Wang et al., 2024), which was originally proposed for Whisper and has since been extended to SALLMs (Omnilingual et al., 2025; Abouelenin et al., 2025; Roll et al., 2025). Zheng et al. (2026) further demonstrates that selecting semantically similar in-context examples substantially improves performance across a range of challenging conditions, including children's speech, accented speech, and multilingual settings. However, unlike text, speech contains substantial variability in speaker identity, prosody, and acoustic conditions, making it difficult to retrieve informative demonstrations across out-of-domain settings.

2.2 Steering

Representation steering methods aim to control model behavior by intervening directly in the hidden representations of a neural network during inference. Given a model with hidden representation $h_l \in \mathbb{R}^d$ at layer l, steering methods modify the forward pass by applying an intervention function

$$\tilde{h}_l = f(h_l, v, \alpha),$$

where v denotes a steering direction or transformation and α controls the intervention strength. The modified representation $\tilde{h}_l$ is then propagated through the remainder of the network to influence model outputs.

Many steering methods use additive interventions of the form

$$\tilde{h}_l = h_l + \alpha v,$$

where v corresponds to a direction in representation space associated with a desired behavior or attribute. Early work demonstrated that meaningful behavioral directions can often be extracted from contrastive examples. (Rimsky et al., 2024) computes steering vectors from differences between hidden activations of paired prompts exhibiting opposite behaviors. Given activations h_l^+ and h_l^- corresponding to positive and negative examples, respectively, a steering vector can be computed as

$$v = \frac{1}{N} \sum_{i=1}^{N} \left(h_{l,i}^+ - h_{l,i}^- \right).$$

This contrastive approach has been extended through PCA-based steering methods (Liu et al.,

2024; Zou et al., 2023), which estimate steering vectors from the principal components of contrastive activation differences. By identifying directions that explain the largest variance across contrastive pairs, PCA-based methods aim to isolate more robust and semantically meaningful steering directions. Approaches like Inference-Time Intervention (Li et al., 2023b), representation extraction methods (Subramani et al., 2022), personalized steering techniques (Cao et al., 2024), and Representation Finetuning (Wu et al., 2024) learn steering transformations directly from data. Rather than using fixed additive vectors, these methods parameterize interventions as trainable functions,

$$\tilde{h}_l = h_l + g_\theta(h_l),$$

where g_θ may consist of lightweight modules trained to modify intermediate representations. Compared to the directly calculated steering vectors, learned interventions can capture more complex transformations.

While most prior steering work has focused on language models, steering has recently begun to expand into multimodal (Sivakumar et al., 2025; Peng et al., 2024; Parekh et al., 2026) and speech settings. In speech-language models, activation steering has been explored to improve modality alignment across speech tasks (Feng et al., 2025) and to enhance robustness in out-of-distribution scenarios (Sun et al., 2026). In these settings, steering interventions are typically applied to encoder representations prior to projection into the language model, with the goal of better aligning acoustic representations with the LLM token space.

3 Steering Vector Training

Prior steering methods commonly construct steering vectors from contrastive activation differences between paired examples, as demonstrated in Section 2.2. This paradigm has also recently been explored in speech-language models, where Sun et al. (2026) construct steering directions from contrastive activation differences to improve ASR on accented speech. However, this formulation is poorly suited to speech tasks. Paired corpora are scarce in speech, particularly in out-of-domain settings, and the acoustic variability inherent to speech makes clean contrastive signals difficult to extract. Furthermore, unlike many semantic language generation tasks, ASR requires preserving a precise correspondence between acoustic inputs and textual outputs, leaving limited tolerance for semantic shifts that contrastive steering may produce.

Instead of extracting steering directions from activation differences, we directly optimize steering vectors using a supervised task objective. This removes the need for paired contrastive examples and learns interventions end-to-end in the representation space.

Let f_θ denote a frozen speech-language model with hidden representation $\mathbf{h}_l \in \mathbb{R}^d$ at layer l. We learn a set of layer-wise steering vectors $\mathcal{V} = \{\mathbf{v}_l\}_{l=1}^{L}$ where $\mathbf{v}_l \in \mathbb{R}^d$ and L is the number of steered layers in the speech encoder. Following prior additive steering approaches, intermediate representations are modified via

$$\tilde{\mathbf{h}}_l = \mathbf{h}_l + \mathbf{v}_l.$$

During training, all parameters θ remain frozen and only the steering vectors $\mathcal{V}$ are optimized.

Steering Mechanism. To stabilize steering, we apply a norm-preserving update at each steered layer:

$$\tilde{\mathbf{h}}_l = \frac{\mathbf{h}_l + \mathbf{v}_l}{\|\mathbf{h}_l + \mathbf{v}_l\|} \cdot \|\mathbf{h}_l\|. \tag{1}$$

This renormalization preserves the original activation magnitude while modifying only the representation direction.

Training Objective. Let $p_{\theta,\mathcal{V}}$ denote the next-token distribution induced by the frozen model under steering intervention $\mathcal{V}$. Given an audio input x and reference transcription $y^\star$, we optimize the steering vectors using the autoregressive cross-entropy objective

$$\mathcal{L}(\mathcal{V}) = -\mathbb{E}_{(x,y^\star)}\left[\sum_{t=1}^{|y^\star|} \log p_{\theta,\mathcal{V}}\left(y_t^\star \mid x, y_{<t}^\star\right)\right].$$

Since all backbone parameters remain frozen, optimization occurs entirely through representation-level interventions.

Optimization Details. All configurations share the same training setup. We optimize steering vectors using AdamW (Loshchilov and Hutter, 2019) with learning rate $\eta \in \{10^{-4}, 5\times10^{-4}\}$, batch size of 1, and gradient clipping with maximum norm 1.0. Training proceeds for up to 20 epochs with early stopping (patience 3) based on validation WER.

Inference. At inference time, the learned steering vectors $\mathcal{V}$ are injected into the selected layers using Eq. 1. Encoder-only configurations apply steering to all audio encoder hidden states, while decoder-only configurations apply steering only to the audio-conditioned hidden representations within the language model backbone. Joint steering applies interventions to both modules simultaneously.

4 Experiments

4.1 Datasets

Children's Speech Children's speech differs systematically from adult speech in its acoustic and linguistic properties and is severely underrepresented in the training data of SALLMs. This makes children's speech a compelling scenario to evaluate whether steering can redirect models toward a structurally distinct and underrepresented population. The Redmond Speech Recall (RSR) Dataset (Redmond et al., 2019; Preza et al., 2026) is a corpus of children's speech (grades K-3) collected during a standardized screening task for developmental language disorder (DLD). The My Science Tutor (MyST) corpus (Pradhan et al., 2024) consists of approximately 230 hours of conversational American English speech from elementary school children (grades 3-5) interacting with a virtual science tutor. The OGI Kids' Speech Corpus (Shobaki et al., 2000) contains read and spontaneous speech from children ranging from grades K through 10.

Multilingual Speech CommonVoice 25.0 (Ardila et al., 2020) is a massively crowd-sourced, multilingual speech corpus. We evaluate on two linguistically distinct subsets: Russian, a high-resource Slavic language, and Twi, a low-resource Kwa language spoken primarily in Ghana. Both Qwen2-Audio-7B-Instruct's LLM backbone and its speech encoder have been exposed to Russian separately during pre-training, whereas neither model has seen Twi. This pairing allows us to contrast adaptation to a language seen during pre-training against transfer to a language entirely absent from both models' training, letting us assess the extent to which pre-training language coverage can compensate for cross-modal representation gaps. SEAME (Lyu et al., 2010) is a conversational Mandarin-English code-switching corpus recorded from Singaporean and Malaysian speakers. Because both modules have seen Mandarin and English during training, SEAME tests a complementary capability: handling intra-utterance code-switching between two individually known languages, rather than transfer to an unseen one.

Together, these corpora span three distinct multilingual conditions: pretraining coverage (Russian), complete absence of training exposure (Twi), and code-switching between known languages (SEAME). This enables a structured analysis of multilingual representation within the models investigated.

4.2 Models

Qwen2-Audio-7B-Instruct Qwen2-Audio-7B-Instruct (Chu et al., 2024a) is an open-weight SALLM that integrates a Whisper-based speech encoder with the pretrained QwenLM. The model supports two interaction modes: voice chat, where it responds directly to spoken input, and audio analysis, where the model conditions on both an audio signal and a text instruction. Qwen2-Audio-7B-Instruct supports 5 major European languages as well as Chinese, Cantonese, and Japanese. It is also instruction-tuned to follow natural language prompts about audio content.

Granite-Speech-3.3-8B Granite-Speech-3.3-8B (Saon et al., 2025) is a SALLM whose architecture contains a speech encoder built from 10 conformer blocks trained with Connectionist Temporal Classification loss (Graves et al., 2006) on ASR-focused data, a two-layer Q-Former (Li et al., 2023a) projector that maps audio embeddings into the input space of the Granite-3.3-8B-Instruct LLM, with LoRA adapters (rank 64) applied to the query and value projection matrices. The model supports 5 major European languages, as well as English-to-Japanese and English-to-Mandarin translation.

4.3 Baseline

We compare against TICL (Zheng et al., 2026), an in-context learning adaptation method for speech-language models that conditions generation on the most semantically similar retrieved demonstrations from a candidate set. In experiments varying the amount of training data, the corresponding training split is additionally used as the retrieval candidate set for TICL. For Qwen2-Audio-7B-Instruct, retrieved demonstrations are prepended directly to the prompt following the standard TICL formulation. Granite-Speech-3.3-7B instead requires

Model	n	System	Children's Speech			SEAME	
			MyST	OGI	RSR	dev-man	dev-sge
Granite-Speech-3.3-8B	–	zero-shot	27.14	28.11	27.94	88.02	72.63
	500	TICL	$33.81_{\pm 2.16}$	$24.91_{\pm 1.77}$	$42.70_{\pm 9.43}$	$365.61_{\pm 25.31}$	$248.86_{\pm 10.11}$
		SALSA	$24.56_{\pm 0.03}$	$14.95_{\pm 1.13}$	$17.13_{\pm 0.43}$	$\mathbf{87.22_{\pm 0.26}}$	$\mathbf{73.49_{\pm 1.17}}$
	2000	TICL	$31.78_{\pm 5.34}$	$17.35_{\pm 0.20}$	$38.85_{\pm 3.85}$	$336.97_{\pm 33.02}$	$224.39_{\pm 11.39}$
		SALSA	$\mathbf{24.10_{\pm 0.07}}$	$\mathbf{12.28_{\pm 0.37}}$	$\mathbf{14.46_{\pm 0.39}}$	$87.40_{\pm 0.50}$	$73.88_{\pm 1.57}$
Qwen2-Audio-7B-Instruct	–	zero-shot	30.51	20.51	28.60	88.46	73.05
	500	TICL	$\mathbf{21.97_{\pm 0.12}}$	$16.14_{\pm 0.66}$	$30.29_{\pm 6.48}$	$188.75_{\pm 18.93}$	$174.46_{\pm 17.76}$
		SALSA	$24.92_{\pm 0.40}$	$14.69_{\pm 1.32}$	$18.89_{\pm 0.24}$	$52.87_{\pm 1.61}$	$44.10_{\pm 2.30}$
	2000	TICL	$22.09_{\pm 0.23}$	$14.40_{\pm 9.54}$	$28.19_{\pm 2.26}$	$173.59_{\pm 4.36}$	$159.12_{\pm 11.32}$
		SALSA	$38.68_{\pm 8.24}$	$\mathbf{12.07_{\pm 0.45}}$	$\mathbf{15.84_{\pm 0.33}}$	$\mathbf{47.03_{\pm 1.67}}$	$\mathbf{40.12_{\pm 1.42}}$

Table 1: WER (%) (↓) on Children's Speech datasets and MER (%) (↓) on SEAME for Qwen2-Audio-7B-Instruct and Granite-Speech-3.3-8B. Results are mean ± std across five random seeds.

demonstrations to be formatted as dialogue history turns due to its input structure. Retrieved examples are therefore inserted as prior user-assistant interactions before the target query.

4.4 Evaluation

We evaluate transcription quality using Word Error Rate (WER) and Mixed Error Rate (MER). WER is the standard metric for automatic speech recognition (ASR), while MER is commonly used for Mandarin-English code-switched ASR.

Given the number of substitutions S, deletions D, insertions I, and reference tokens N, both metrics are computed as

$$\text{ErrorRate} = \frac{S + D + I}{N}.$$

The metrics differ in how the reference and predicted transcriptions are tokenized prior to alignment. For WER, all text is tokenized at the word level. For MER, English speech is tokenized at the word level while Mandarin speech is tokenized at the character level, following standard evaluation protocols for Mandarin-English code-switched ASR. This mixed tokenization better reflects transcription quality across languages with different linguistic granularity.

All experiments use greedy autoregressive decoding, where the highest-probability token is selected at each generation step without beam search. WER and MER were computed using the jiwer package (Morris et al., 2004) with standard text normalization and tokenization settings. Performance is reported as the mean and standard deviation across five independently sampled training subsets of size n, with random seeds 42-46 used for reproducibility.

4.5 Results

Table 1 reports performance on the Children's Speech datasets and SEAME using training sets of size $n \in \{500, 2000\}$ for both Qwen2-Audio-7B-Instruct and Granite-Speech-3.3-8B. Figure 1 further shows scaling trends across a wider range of training sizes for Qwen2-Audio-7B-Instruct.

Qwen2-Audio-7B-Instruct. SALSA substantially improves performance over zero-shot prompting on OGI, RSR, and SEAME. On RSR, SALSA achieves a relative improvement of 44.6% over zero-shot performance at $n = 2000$, outperforming TICL by 43.8%. Similarly, on OGI, SALSA improves over zero-shot by 42.3%, while remaining competitive with TICL, differing by approximately 2 absolute WER points. These results suggest that encoder steering effectively adapts acoustic representations to better align with the decoder's pretrained phonetic and linguistic representations.

In contrast, MyST exhibits different scaling behavior. While SALSA improves over zero-shot performance at smaller training sizes, performance degrades as additional steering data is introduced, as shown in Figure 1. At $n = 1000$, TICL outperforms SALSA on MyST, suggesting that steering may be less effective for this dataset or more sensitive to overfitting. This may stem from the substantial intra- and inter-speaker variability in MyST, which may introduce inconsistent or noisy adaptation signals into the learned steering directions.

On SEAME, SALSA produces large gains on both evaluation splits, achieving up to a 46.8% relative improvement on dev-man and 45.1% on dev-sge compared to zero-shot prompting. Performance improvements begin to saturate beyond

2000 training samples. In contrast, TICL exhibits severe instability on SEAME, increasing MER by more than 100% relative to the zero-shot baseline. This behavior may stem from interference between retrieved multilingual demonstrations and the model's existing multilingual representations. In comparison, encoder steering appears to more effectively adapt the acoustic representations to the multilingual LLM backbone without disrupting the pretrained alignment between Mandarin and English representations.

Table 2 reports results on CommonVoice Russian and Twi. Across both datasets, SALSA consistently improves over zero-shot prompting, while TICL often provides little benefit or substantially degrades performance. On Russian, even 200 steering examples dramatically improve performance, suggesting that steering can effectively adapt the model for ASR in a language already seen during pretraining of both the encoder and LLM backbone. These results suggest that the primary challenge may not be language acquisition itself, but rather, aligning the encoder representations with the decoder's transcription space for the downstream ASR task. On Twi, SALSA provides modest gains over zero-shot inference, although the limited amount of available training data makes it difficult to determine whether additional scaling would further improve performance.

Granite-Speech-3.3-8B. SALSA consistently improves over zero-shot inference on Granite-Speech-3.3-8B across the Children's Speech datasets. The largest gains are observed on OGI and RSR, where SALSA achieves relative WER reductions of 56.3% and 48.2%, respectively, at $n = 2000$. SALSA also substantially outperforms TICL on these datasets, particularly on RSR, where TICL degrades performance relative to the zero-shot baseline.

Granite-Speech-3.3-8B additionally exhibits more stable scaling behavior than Qwen2-Audio-7B-Instruct as the number of steering examples increases from 500 to 2000. On MyST, SALSA achieves consistent improvements over zero-shot prompting, reducing WER from 27.14 to 24.10 at $n = 2000$. This contrasts with the degradation observed for Qwen2-Audio-7B-Instruct on the same dataset and may suggest that Granite-Speech-3.3-8B is more robust to the substantial intra- and inter-speaker variability present in MyST. On SEAME, steering provides limited benefit for Granite-Speech-3.3-8B. This may stem from differences in the model's pretraining objectives. Although Granite-Speech supports English-to-Mandarin translation, the speech encoder was not explicitly optimized for Mandarin ASR, potentially resulting in weaker Mandarin acoustic representations for code-switched recognition. SALSA remains close to the zero-shot baseline on both dev-man and dev-sge, while TICL substantially degrades performance, increasing MER by more than 200% relative to zero-shot. These results suggest that although steering transfers effectively to children's speech, adaptation to code-switching is more challenging and may depend strongly on the pre-training of the underlying model components.

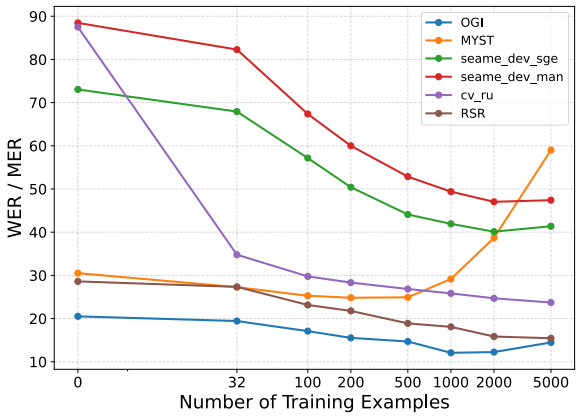


Figure 1: Scaling behavior of encoder steering on Qwen2-Audio-7B-Instruct across training set sizes. SALSA consistently improves performance on OGI, RSR, and SEAME, while MyST exhibits degradation at larger training sizes.

Dataset	n	Zero-shot	TICL	SALSA
cv-tw	200	108.36	$106.32_{\pm 0.00}$	$\mathbf{87.54}_{\pm 2.10}$
cv-ru	200	87.50	$122.07_{\pm 3.21}$	$\mathbf{28.33}_{\pm 0.49}$
	500	87.50	$121.90_{\pm 3.79}$	$\mathbf{26.84}_{\pm 0.34}$
	2000	87.50	$119.29_{\pm 1.28}$	$\mathbf{24.65}_{\pm 0.07}$

Table 2: WER (%)(↓) on CommonVoice datasets. SALSA mean ± std over 5 seeds. Best system per row in **bold**.

5 Analysis

To better understand the behavior of learned steering interventions, we analyze the effects of training set size, steering location within the encoder, and the choice of model module being steered. Prior work on speech representation learning has shown that encoder layers capture progressively different

levels of abstraction, ranging from low-level acoustic information in earlier layers to higher-level phonetic and linguistic information in later layers (Langedijk et al., 2024; A Shams et al., 2024). We therefore investigate whether steering effectiveness depends on where interventions are applied within the encoder.

5.1 Effect of Training Set Size

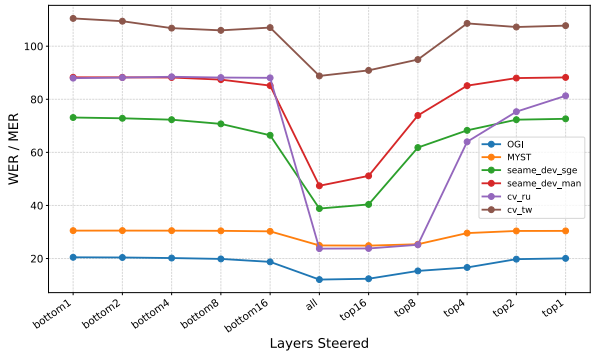
Figure 1 shows the scaling trends for SALSA across varying training set sizes on Qwen2-Audio-7B-Instruct. Across OGI, RSR, and SEAME, increasing the number of steering examples consistently improves performance, with the largest gains typically occurring between 32 and 500 training examples. Beyond approximately 2000 examples, improvements begin to saturate, suggesting diminishing returns from additional steering data. The strongest scaling behavior is observed on RSR and SEAME. On RSR, WER decreases steadily as additional steering examples are introduced, indicating that learned steering vectors can effectively adapt encoder representations to children's speech despite the substantial acoustic mismatch between adult and child speech. Similarly, both SEAME splits show large improvements with increasing training data, suggesting that steering can effectively adapt multilingual acoustic representations for code-switched ASR without modifying the underlying model parameters.

As noted before, MyST exhibits substantially different behavior. While small amounts of steering data initially improve performance, WER increases at larger training sizes. This suggests that steering may be more sensitive to highly heterogeneous speech distributions containing substantial speaker and acoustic variability. Because steering vectors are shared across all utterances, increasing the amount of highly diverse adaptation data may produce less coherent representation shifts.

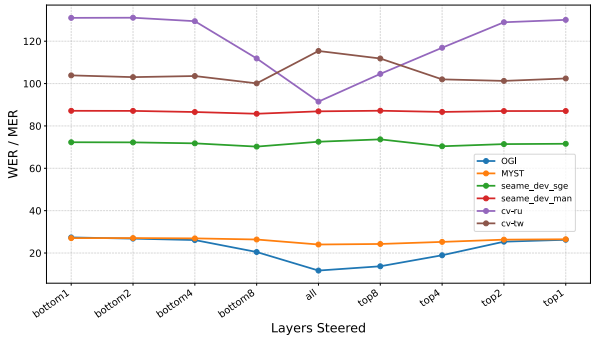
Overall, these results suggest that encoder steering is highly data-efficient, with substantial improvements achievable using only a few hundred adaptation examples. This is particularly important for low-resource and out-of-domain speech settings, where collecting large supervised corpora may be impractical.

5.2 Module-Level Steering Analysis

Figure 3 compares steering applied to the speech encoder, the LLM backbone, and both modules jointly on RSR using Qwen2-Audio-7B-Instruct.



(a) Qwen2Audio-7B-Instruct



(b) Granite-Speech–8B

Figure 2: Performance as a function of encoder steering location. Steering higher encoder layers consistently outperforms steering lower layers across most datasets and models.

Across nearly all training set sizes, encoder steering substantially outperforms decoder-only steering. Decoder steering exhibits only modest improvements over the zero-shot baseline and plateaus quickly as additional training data is introduced. In contrast, encoder steering continues to improve as more steering examples are added, ultimately achieving the best overall performance.

Interestingly, jointly steering both the encoder and decoder does not outperform encoder-only steering. Although joint steering initially matches encoder-only performance at smaller training sizes, the two begin to diverge as additional steering data is introduced. This suggests that modifying decoder representations may partially interfere with the pretrained linguistic representations of the language model backbone. These results suggest that, when both the encoder and language model backbone have prior exposure to the target languages during pretraining, adaptation is more effectively achieved by modifying acoustic representations before projection into the language model rather than intervening directly within the decoder.

5.3 Layer-Level Steering Analysis

Figure 2 compares performance when steering different layers of the encoder. We evaluate steering applied to progressively larger subsets of lower and upper encoder layers.

For Qwen2-Audio-7B-Instruct, steering higher encoder layers consistently produces larger improvements than steering lower layers alone. On both RSR and OGI, performance approaches that of full encoder steering as progressively more upper encoder layers are steered, whereas steering only the earliest layers provides limited benefit. Similar trends are observed on SEAME and Russian. These results suggest that later encoder representations contain more task-relevant phonetic and linguistic information for adaptation.

Granite-Speech-3.3-8B exhibits a similar but less pronounced pattern. Steering higher encoder layers again provides the strongest improvements on OGI and RSR, although performance varies less dramatically across steering locations compared to Qwen2-Audio-7B-Instruct. This difference may stem from variations in encoder pretraining objectives and architectural design between the two models. However, Twi shows the opposite trend, with lower-layer steering outperforming upper-layer steering. Since Twi is absent from the pretraining data of both the encoder and the language model backbone, this may indicate that adaptation relies more heavily on modifying lower-level acoustic representations when higher-level linguistic representations are not already well established.

More broadly, the degradation observed when steering only lower encoder layers suggests that low-level acoustic representations may already be sufficiently aligned in many settings, while the primary adaptation benefits arise from modifying higher-level phonetic and linguistic representations closer to the projection interface with the language model. Taken together, these results support the hypothesis that encoder steering primarily operates by improving alignment between higher-level speech representations and the pretrained language model decoding space, rather than by substantially altering low-level acoustic processing. These findings are also consistent with prior observations from speech activation steering (Sun et al., 2026).

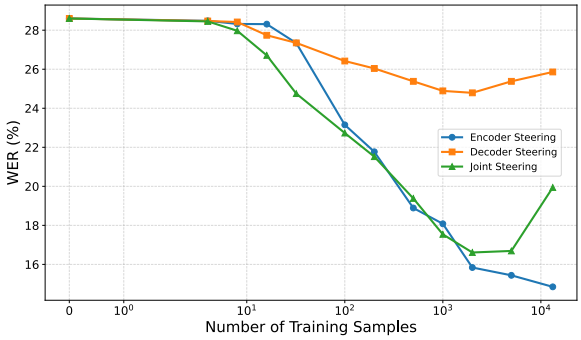


Figure 3: Effect of steering module location on RSR for Qwen2-Audio-7B-Instruct.

6 Conclusion

We introduced SALSA, a lightweight steering-based adaptation method for SALLMs that directly learns layer-wise steering vectors without requiring paired contrastive examples or modifying backbone model weights.

Across children's speech, multilingual speech, and Mandarin-English code-switched speech benchmarks, SALSA substantially improves performance over zero-shot inference and speech in-context learning baselines, demonstrating that simple learned additive interventions are highly effective for out-of-domain ASR settings. By analyzing steering location, encoder depth, and training scale, we further show that adaptation is driven primarily by higher-level encoder representations, whereas steering the language model backbone provides comparatively limited benefit. These findings suggest that adaptation in SALLMs primarily operates through modifying acoustic and phonetic representations before projection into the language model, rather than through interventions within the decoder itself. Our results demonstrate that lightweight representation-level interventions provide an effective and scalable mechanism for adapting pretrained SALLMs to low-resource and out-of-domain speech conditions. Future work may explore input-dependent steering methods that dynamically adapt steering interventions based on the acoustic characteristics of individual utterances, as well as extending steering-based adaptation to other speech settings.

7 Limitations

The scope of this work is limited to automatic speech recognition using two speech-aware large language models. The effectiveness and location of

steering interventions may vary across different architectures and training objectives. Since SALSA learns a single shared set of steering vectors for all utterances within a dataset, highly heterogeneous speech distributions may therefore require more input-dependent or speaker-specific interventions. Our experiments focus exclusively on automatic speech recognition tasks and a limited set of languages and speech conditions. Broader evaluation across additional tasks, languages, and speaker populations is needed to better understand the generalization and fairness implications of representation-level adaptation methods.

8 Ethical Considerations

This work studies adaptation methods for speech-aware large language models in multilingual and out-of-domain speech settings. Although SALSA improves performance for several underrepresented speech conditions, speech recognition systems remain susceptible to demographic and linguistic biases. Steering is substantially more effective for languages already represented during pretraining than for completely unseen languages, highlighting the risk that adaptation methods may reinforce existing imbalances in multilingual speech resources. All pretrained models and datasets used in this work were used in accordance with their respective licenses, access agreements, and terms of use. Some datasets used in this work contain annotations or transcriptions with restricted redistribution terms.

References

Erfan A Shams, Iona Gessinger, and Julie Carson-Berndsen. 2024. Uncovering syllable constituents in the self-attention-based speech representations of whisper. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 238–247, Miami, Florida, US. Association for Computational Linguistics.

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

R. Ardila, M. Branson, K. Davis, and 1 others. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. 2024. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. *Advances in Neural Information Processing Systems*, 37:49519–49551.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024a. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yunfei Chu, Jin Xu, Qian Yang, and 1 others. 2024b. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *Preprint*, arXiv:2311.07919.

Avishai Elmakies, Hagai Aronowitz, Nimrod Shabtay, Eli Schwartz, Ron Hoory, and Avihu Dekel. 2026. Advancing speech understanding in speech-aware language models with grpo. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 17797–17801. IEEE.

Ruitao Feng, Bixi Zhang, Sheng Liang, and Zheng Yuan. 2025. Steer-moe: Efficient audio-language alignment with a mixture-of-experts steering module. *arXiv preprint arXiv:2510.13558*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proc. International Conference on Machine Learning (ICML)*, pages 369–376.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint*. ArXiv:2106.09685 [cs].

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, and 1 others. 2024. Wavllm: Towards robust and adaptive speech large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4552–4572.

Shawn Im and Sharon Li. 2025. A unified understanding and evaluation of steering methods. *arXiv preprint arXiv:2502.02716*.

Anna Langedijk, Hosein Mohebbi, Gabriele Sarti, Willem Zuidema, and Jaap Jumelet. 2024. Decoderlens: Layerwise interpretation of encoder-decoder transformers. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4764–4780.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Sheng Liu, Haotian Ye, Lei Xing, and James Y. Zou. 2024. In-context vectors: Making in context learning more effective and controllable through latent space steering. In *Forty-first International Conference on Machine Learning*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Dau-Cheng Lyu, Tien Ping Tan, Engsiong Chng, and Haizhou Li. 2010. Seame: a mandarin-english code-switching speech corpus in south-east asia. In *Interspeech*, volume 10, pages 1986–1989.

Andrew Cameron Morris, Viktoria Maier, and Phil D Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Interspeech*, 4-8, page 2004.

ASR Omnilingual, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, and 1 others. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages. *arXiv preprint arXiv:2511.09690*.

Jayneel Parekh, Pegah KHAYATAN, Mustafa Shukor, Arnaud Dapogny, Alasdair Newson, and Matthieu Cord. 2026. Learning to steer: Input-dependent steering for multimodal LLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Jing Peng, Yucheng Wang, Bohan Li, Yiwei Guo, Hankun Wang, Yangui Fang, Yu Xi, Haoyu Li, Xu Li, Ke Zhang, and 1 others. 2025. A survey on speech large language models for understanding.

Yingzhe Peng, Chenduo Hao, Xinting Hu, Jiawei Peng, Xin Geng, and Xu Yang. 2024. Live: Learnable in-context vector for visual question answering. *Advances in Neural Information Processing Systems*, 37:9773–9800.

Sameer Pradhan, Ronald Cole, and Wayne Ward. 2024. My science tutor (MyST)–a large corpus of children's conversational speech. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12040–12045.

T. Preza, D. Liu, C. Miller, A. Alabdallah, J. Xiong, S. Redmond, and P. Hadley. 2026. Novel approaches to language screening: Evaluating measures of effort in a sentence recall task for school-age children. Manuscript under review.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Sean M Redmond, Andrea C Ash, Tyler T Christopulos, and Theresa Pfaff. 2019. Diagnostic accuracy of sentence recall and past tense measures for identifying children's language impairments. *Journal of Speech, Language, and Hearing Research*, 62(7):2438–2454.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522.

Nathan Roll, Calbert Graham, Yuka Tatsumi, Kim Tien Nguyen, Meghan Sumner, and Dan Jurafsky. 2025. In-context learning boosts speech recognition via human-like adaptation to speakers and language varieties. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4412–4426.

George Saon, Avihu Dekel, Alexander Brooks, Tohru Nagano, Abraham Daniels, Aharon Satt, Ashish Mittal, Brian Kingsbury, David Haws, Edmilson Morais, and 1 others. 2025. Granite-speech: Open-source speech-aware llms with strong english asr capabilities. *arXiv preprint arXiv:2505.08699*.

Khaldoun Shobaki, John-Paul Hosom, and Ronald Cole. 2000. The OGI kids' speech corpus and recognizers. pages 258–261.

Anushka Sivakumar, Andrew Zhang, Zaber Hakim, and Chris Thomas. 2025. Steervlm: Robust model control through lightweight activation steering for vision language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23640–23665.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581, Dublin, Ireland. Association for Computational Linguistics.

Jinuo Sun, Yang Xiao, Sung Kyun Chung, Qiuchi Hu, Gongping Huang, Eun-Jung Holden, and Ting Dang. 2026. Activation steering for accent adaptation in speech foundation models. *arXiv preprint arXiv:2603.05813*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, volume 2024, pages 16607–16629.

Siyin Wang, Chao-Han Yang, Ji Wu, and Chao Zhang. 2024. Can whisper perform speech-based in-context learning? In *ICASSP 2024*, pages 13421–13425.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. 2024. Reft: Representation finetuning for language models. *Advances in Neural Information Processing Systems*, 37:63908–63962.

Haolong Zheng, Yekaterina Yegorova, and Mark Hasegawa-Johnson. 2026. Ticl: Text-embedding knn for speech in-context learning unlocks speech recognition abilities of large multimodal models. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 17912–17916. IEEE.

Haolong Zheng, Yekaterina Yegorova, and Mark A. Hasegawa-Johnson. 2025. TICL+: A case study on speech in-context learning for children's speech recognition. In *IEEE ASRU Satellite Workshop-AI for Children's Speech and Language*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Compute and Infrastructure.

Experiments were conducted using NVIDIA A40 GPUs. Qwen2-Audio-7B-Instruct and Granite-Speech-3.3-8B contain approximately 7B and 8B parameters, respectively. Because SALSA trains only lightweight steering vectors while keeping all backbone parameters frozen, training remained computationally efficient relative to full model fine-tuning.

B Dataset Statistics

Dataset	Train	Dev	Test
MyST	73425	11732	12744
OGI	50439	5482	16078
RSR	13222	1467	2027
SEAME dev-man	48297	4763	0
SEAME dev-sge	48297	1749	0
cv-ru	26920	10282	10283
cv-tw	213	0	30

Table 3: Dataset split statistics. Values denote the number of utterances in each split.

C AI Usage

AI-based assistants were used for limited editing, wording suggestions, and debugging assistance during experiment development and manuscript preparation. All technical content, experimental design, analysis, and conclusions were verified and finalized by the authors.

D Supplementary Tables

n_{train}	OGI (WER)	MyST (WER)	SEAME Test (MER)	SEAME Val (MER)	CV-RU (WER)	RSR (WER)
32	19.43±0.72	27.27±0.44	67.93±1.32	82.28±1.08	34.78±1.70	27.34±0.88
100	17.09±0.94	25.28±0.32	57.17±1.77	67.37±1.83	29.76±0.78	23.15±0.53
200	15.53±0.94	24.82±0.67	50.41±1.45	60.00±1.74	28.33±0.49	21.77±0.58
500	14.69±1.32	24.92±0.40	44.10±2.30	52.87±1.61	26.84±0.34	18.89±0.24
1000	12.07±0.45	29.15±1.89	41.93±0.50	49.39±0.87	25.83±0.39	18.08±0.35
2000	12.24±0.97	38.68±8.24	40.12±1.42	47.03±1.67	24.69±0.15	15.84±0.33
5000	14.47±2.68	59.01±15.08	41.38±0.24	47.41±0.85	23.72±0.17	15.44±0.42

Table 4: Full numerical values for the results shown in Figure 1.

Mask	$\lvert S\rvert$	OGI (WER)	MyST (WER)	SEAME Test (MER)	SEAME Val (MER)	CV-RU (WER)	CV-TW (WER)
all	32	12.07±0.43	24.93±0.40	38.83±1.28	47.41±0.85	23.73±0.19	88.80±1.57
bottom1	1	20.46±0.04	30.51±0.02	73.13±0.12	88.28±0.15	87.96±0.06	110.47±1.72
bottom2	2	20.39±0.05	30.52±0.02	72.84±0.23	88.26±0.16	88.18±0.30	109.42±1.40
bottom4	4	20.18±0.09	30.50±0.03	72.31±0.40	88.20±0.05	88.48±0.31	106.81±2.71
bottom8	8	19.85±0.17	30.44±0.03	70.73±0.42	87.41±0.34	88.19±0.36	105.97±2.74
bottom16	16	18.74±0.21	30.24±0.04	66.44±0.48	85.16±0.93	88.08±0.41	107.02±2.53
top1	1	20.06±0.09	30.41±0.01	72.64±0.11	88.24±0.05	81.31±1.16	107.75±0.96
top2	2	19.76±0.16	30.39±0.01	72.31±0.14	87.99±0.05	75.33±2.42	107.23±0.84
top4	4	16.63±0.35	29.58±0.15	68.29±0.42	85.13±0.29	63.98±5.05	108.59±7.27
top8	8	15.33±0.56	25.38±0.17	61.81±1.40	73.90±4.50	25.18±0.36	94.97±1.89
top16	16	12.38±0.48	24.86±0.23	40.38±1.52	51.14±0.48	23.79±0.18	90.89±2.21

Table 5: Full numerical values for the Qwen2Audio-7B-Instruct results shown in Figure 2

Mask	$\lvert S\rvert$	OGI (WER)	MyST (WER)	SEAME Test (MER)	SEAME Val (MER)	CV-RU (WER)	CV-TW (WER)
all	16	11.72±0.31	24.02±0.06	72.54±0.71	86.89±0.24	91.48±0.42	115.39±6.24
bottom1	1	27.36±0.37	27.03±0.05	72.30±0.06	87.14±0.04	131.02±0.21	103.87±3.53
bottom2	2	26.79±0.88	27.06±0.09	72.24±0.10	87.10±0.09	131.09±0.24	103.04±3.65
bottom4	4	26.12±0.80	26.91±0.06	71.78±0.25	86.57±0.19	129.45±0.49	103.56±4.31
bottom8	8	20.48±0.82	26.38±0.08	70.24±0.41	85.75±0.34	111.87±1.12	100.10±2.42
top1	1	26.25±0.16	26.49±0.10	71.58±0.17	87.05±0.03	130.08±0.61	102.41±1.08
top2	2	25.35±0.24	26.31±0.10	71.42±0.12	87.03±0.03	128.95±0.52	101.26±0.26
top4	4	18.94±0.49	25.25±0.09	70.40±0.50	86.61±0.16	116.91±1.01	101.99±1.88
top8	8	13.74±0.37	24.27±0.12	73.66±1.21	87.18±0.46	104.57±2.93	111.83±4.45

Table 6: Full numerical values for the Granite-Speech-3.3 results shown in Figure 2

Num Train Samples	Encoder Steering	Decoder Steering	Joint Steering
4	28.48±0.31	28.47±0.06	28.45±0.14
8	28.32±0.19	28.42±0.06	27.97±0.09
16	28.31±0.38	27.74±0.52	26.71±0.37
32	27.34±0.88	27.35±0.25	24.75±0.84
100	23.15±0.53	26.42±0.49	22.73±0.47
200	21.77±0.58	26.04±0.34	21.52±0.46
500	18.89±0.24	25.38±0.60	19.38±0.27
1000	18.08±0.35	24.89±0.41	17.54±0.56
2000	15.84±0.33	24.79	16.61±0.33
5000	15.44	25.38±0.99	16.69±0.32
13222	14.85	25.86	19.94

Table 7: Full numerical values for the results shown in Figure 3.