# Phonological Abstraction and Exception Memorization in Noisy Tone Learning

Anonymous ACL submission

## Abstract

Transformer-based speech models have recently been shown to perform well under impoverished and ambiguous supervision, a setting motivated by poverty-of-the-stimulus arguments, yet their learning dynamics remain unclear. In this work, we study this question with `wav2vec2.0` on Mandarin tone classification, using 70 scale–noise configurations, capacity interventions, and perturbation probes to examine the dynamics of abstraction-consistent generalization and exception memorization, yielding representative regimes of *Robust Generalization*, *Abrupt Transition*, and *High-noise Collapse*. We find a consistent abstraction–exception trajectory: when phonological recovery is possible, the model first exhibits phonologically abstraction-consistent behavior and only later fits corrupted labels as empirical exceptions. Capacity analyses show that this trajectory is not merely a large-model artifact: it persists under truncated Transformer encoders, while acoustic rank compression modulates but does not selectively separate abstraction from exception memorization. Finally, acoustic perturbations show different local robustness: abstraction-consistent predictions remain stable under stronger waveform perturbations, whereas fitted exceptions behave as fragile local overrides on top of the broader abstraction-consistent behavior. [1]

## 1 Introduction

Infants acquire spoken language from sparse, variable, and often ambiguous speech, a condition traditionally discussed as the Poverty of the Stimulus (POS) (Chomsky, 1980). Does linguistic knowledge require innate biological predispositions (Chomsky et al., 2023), or can it emerge empirically from sparse and noisy exemplars (Ambridge, 2020; Mahowald et al., 2020)? The recent success of foundation models (OpenAI, 2023) demonstrates the possibility of empirical learning, but it is achieved through massive pretraining corpora and carefully designed post-training pipelines (Guo et al., 2025), still far from the efficiency with which infants acquire language under POS conditions (McCoy and Griffiths, 2025). Given that infants bootstrap linguistic structure from continuous speech input (Saffran et al., 1996), controlled speech POS tasks therefore serve as a promising testbed for both language acquisition theory and the interpretability of transformer architectures through the lens of human cognitive development. Early POS-style speech studies have shown that transformer-based speech models can maintain high recognition accuracy even when supervision is made sparse or noisy (Yang et al., 2025). This performance-level finding naturally raises a deeper question:

*How do transformers trade off phonological rule abstraction and empirical exception memorization when optimized on noisy POS supervision?*

In this work, we instantiate this question with lexical tone classification, a basic suprasegmental phonemic contrast in Mandarin. We use the widely used speech architecture `wav2vec2.0` as the model backbone (Baevski et al., 2020). To simulate POS conditions, we construct a $7 \times 10$ scale–noise grid: seven levels of available training speech and ten levels of syllable-level tone-label replacement. We then conduct a series of experiments to progressively characterize this tradeoff. Rather than treating test accuracy as the only endpoint, we examine the model's learning dynamics, how acoustic capacity modulates these dynamics, and whether abstraction-conforming items and exceptions, despite both being ultimately fit, differ in the organization and robustness of their learned representations. Our key findings are summarized

[1] Full trajectories, perturbation results, and a visualization interface are included in the submitted data; all checkpoints and original experiment manifests will be released upon acceptance.

as follows:

- **(Dynamics)** Across the scale–noise grid, we identify a consistent abstraction–exception trajectory: in recoverable regimes, wav2vec2.0 first recovers abstraction-consistent tone behavior and only later fits corrupted labels as empirical exceptions, often reaching near-perfect training accuracy. This temporal ordering yields three regimes: *Robust Phonological Generalization*, where exception fitting remains largely benign; *Abrupt Transition*, where abstraction-consistent behavior remains recoverable under heavy noise but is later sharply degraded by exception fitting; and *High-noise Collapse*, where training accuracy rises through exception fitting without recoverable abstraction-consistent generalization (Figures 1,2).

- **(Capacity)** We first ask whether the abstraction–exception pattern is simply a consequence of excessive model capacity; its persistence under progressively truncated Transformer encoders suggests otherwise. The low-rank constraint reveals a more specific capacity effect: low-rank channels can still preserve substantial tone structure (e.g., $r = 2$ reaches $79.9\%$ test accuracy), but compression is not selective. Any reduction in exception fitting comes at the cost of also weakening abstraction recovery. Thus, the acoustic cues that support abstraction and those that support memorization remain entangled in the acoustic channel (Table 1,2).

- **(Internalization)** Acoustic perturbation probes show that abstraction-consistent and exception-fitting predictions have different local robustness profiles: abstraction-consistent predictions behave as if they are supported by broader, more robust local neighborhoods, whereas fitted exceptions behave like narrow local neighborhoods around the original waveform. In *Robust Phonological Generalization*, perturbing the original waveform often recovers the corresponding abstraction-consistent label, indicating that exception memorization locally overlaid on broader abstraction-consistent behavior. Thus, although the cues supporting abstraction and exception fitting are entangled in the acoustic channel, their predictions differ in perturbation robustness and local recoverability (Figure 3).

Beyond the representative figures and metrics reported in the paper, we release the full set of epoch-wise trajectories for all 70 scale–noise configurations, the complete perturbation results across SNR levels, and over 2000 saved epoch-wise checkpoints to help readers inspect abstraction–exception dynamics as continuous training processes and to support further work on model-internal mechanisms and representation geometry.

## 2 Background

### 2.1 Poverty of the Stimulus in Speech Acquisition

The Poverty of the Stimulus frames language acquisition as a learnability problem: children acquire systematic linguistic knowledge from input that is finite, noisy, and only partially informative (Chomsky, 1980; Berwick et al., 2011). This observation has motivated nativist accounts, which posit innate constraints on the hypothesis space (Lidz et al., 2003), as well as empiricist and exemplar-based accounts, which argue that structure can emerge from distributional regularities in the input (Saffran et al., 1996; Ambridge, 2020; Perfors et al., 2011). At its core, the POS argument concerns a learnability gap: the observable input appears underdetermined relative to the abstract structure that learners infer (Pullum and Scholz, 2002; Perfors et al., 2011). In typical early acquisition, this gap is confronted primarily through speech rather than presegmented symbolic input, requiring infants to extract structure from continuous acoustic streams (Jusczyk and Aslin, 1995).

Speech perception turns continuous acoustic variation into discrete linguistic categories (Liberman et al., 1957), a mapping central to early phonological acquisition (Kuhl, 2004). Mandarin lexical tone provides a compact case of this mapping: variable pitch trajectories are interpreted as discrete tone categories that distinguish word meanings (Yip, 2002), making tone a useful testbed for speech perception and linguistic cognition (Liang and Du, 2018; Shuai and Gong, 2014).

### 2.2 Neural Models as POS Testbeds

Neural language models provide a controlled testbed for POS-style questions by allowing researchers to manipulate training input, model architecture, and evaluation targets (Warstadt and Bowman, 2022). Recent work (Yedetore et al., 2023; Warstadt et al., 2023) has therefore evaluated models under developmentally motivated data budgets, asking whether linguistic generalizations can

emerge from child-scale or child-directed corpora. These studies show that neural learners can acquire nontrivial linguistic behavior from restricted input (Warstadt and Bowman, 2020), but they also reveal that human-like generalization is not guaranteed and may depend on inductive biases (McCoy et al., 2020; McCoy and Griffiths, 2025). Speech is central to early language acquisition, yet most neural POS studies focus on text-based models because controlled speech simulations pose additional challenges in task formulation, dataset manipulation, and metric design. Initial work shows that transformer-based speech models can maintain strong tone recognition under sparse and noisy supervision (Yang et al., 2025), but how this behavior unfolds over training and how abstractions and exceptions are internalized by the model remain unclear.

### 2.3 Abstraction and Memorization

Transformer-based models have been shown to store idiosyncratic input–label mappings at scale (Mahdavi et al., 2024), with recent work quantifying memorization capacity in attention-based architectures and estimating that GPT-style models can memorize roughly 3.6 bits per parameter (Morris et al., 2025). This capacity has motivated extensive work on the relation between abstraction and memorization (Arpit et al., 2017; Feldman and Zhang, 2020; Panwar et al., 2026; Huang et al., 2025). In particular, bottleneck and compression views suggest that restricting representational capacity may selectively preserve shared abstraction while filtering input-specific noise (Alemi et al., 2016; Yu, 2025). Prior work reports several distinct temporal patterns: simple-to-complex learning, where shared structure precedes noise fitting (Kalimeris et al., 2019; Rahaman et al., 2019); grokking, where generalization emerges only after prolonged overfitting (Power et al., 2022; Nanda et al., 2023); and double descent, where generalization varies non-monotonically with training time or effective capacity (Nakkiran et al., 2021).

## 3 Experiment Design

### 3.1 Task Formulation

Following the POS speech-learning framework of Yang et al. (2025), we focus on Mandarin tone classification. Lexical tone is a suprasegmental phonemic contrast: pitch contours distinguish word meanings while interacting with segmental context (Yip, 2002), which makes tone recognition a compact but linguistically meaningful testbed for studying phonological abstraction.

**Noisy sparse targets.** We simulate POS conditions along two axes: training scale and label noise. Training scale controls the amount of speech available for supervision, while label noise controls the proportion of syllable-level tone labels randomly replaced. Specifically, each syllable's tone label is replaced with a uniformly sampled label from $\{\mathrm{T1}, \ldots, \mathrm{T5}\}$ with probability $p$.

**Masked optimization objective.** For the $i$-th utterance, let $\mathbf{X}_i$ denote the input speech sequence with $K_i$ acoustic frames, $\mathbf{Y}_i^\star = (y_{i,1}^\star, \ldots, y_{i,K_i}^\star)$ denote the ground-truth label sequence for utterance $i$, and $\tilde{\mathbf{Y}}_i = (\tilde{y}_{i,1}, \ldots, \tilde{y}_{i,K_i})$ denote the observed noisy sequence. Following the central-frame labeling scheme (Yang et al., 2025; Yuan et al., 2023), only the center frame of each syllable is assigned a tone label, while all non-target frames are assigned the null label 0. We therefore optimize cross-entropy only on valid tone targets by masking out 0 frames:

$$\mathcal{L}(\theta) = -\frac{1}{|\mathcal{V}|} \sum_{i=1}^{N} \sum_{c=1}^{K_i} \mathbb{I}(\tilde{y}_{i,c} \neq 0) \log q_\theta(\tilde{y}_{i,c} \,|\, \mathbf{X}_i),$$

where $\mathcal{V} = \{(i,c) : \tilde{y}_{i,c} \neq 0\}$ denotes the set of valid tone-target frames.

### 3.2 Learning Trajectory Probes

To characterize model performance across a broad continuum of stimulus poverty, we first report test accuracy of the checkpoint with the lowest validation loss under each scale–noise condition, which measures generalization to unseen speech. To further reveal how phonological rule abstraction and empirical exception memorization evolve during fine-tuning, we therefore save epoch-wise checkpoints and evaluate each checkpoint with three trajectory probes. Let $\hat{\mathbf{Y}}_i^{(t)} = (\hat{y}_{i,1}^{(t)}, \ldots, \hat{y}_{i,K_i}^{(t)})$ denote the prediction made by the checkpoint at epoch $t$. We define $\mathcal{V}_{\text{train}}$ and $\mathcal{V}_{\text{test}}$ as the sets of valid tone-target frames in the training and test splits, respectively, and define $\mathcal{E}_{\text{train}} \subseteq \mathcal{V}_{\text{train}}$ as the subset of training positions whose labels are corrupted.

- **Exception Memorization.** The fraction of corrupted training positions for which the model predicts the observed corrupted label:

$$\mathrm{Acc}_{\mathrm{EM}} = \frac{1}{|\mathcal{E}_{\text{train}}|} \sum_{(i,c)\in\mathcal{E}_{\text{train}}} \mathbb{I}\Big(\hat{y}_{i,c}^{(t)} = \tilde{y}_{i,c}\Big).$$

- **Self-Correction.** The fraction of the same corrupted positions for which the model predicts the corresponding ground-truth tone label:

$$\mathrm{Acc}_{\mathrm{SC}} = \frac{1}{|\mathcal{E}_{\mathrm{train}}|} \sum_{(i,c)\in\mathcal{E}_{\mathrm{train}}} \mathbb{I}\Big(\hat{y}_{i,c}^{(t)} = y_{i,c}^{*}\Big).$$

- **Test Accuracy.** Accuracy on held-out test speech against ground-truth tone labels:

$$\mathrm{Acc}_{\mathrm{GEN}} = \frac{1}{|\mathcal{V}_{\mathrm{test}}|} \sum_{(i,c)\in\mathcal{V}_{\mathrm{test}}} \mathbb{I}\Big(\hat{y}_{i,c}^{(t)} = y_{i,c}^{*}\Big).$$

### 3.3 Model Capacity Interventions

To test whether the observed abstraction–exception dynamics are simply a consequence of excessive model capacity, we introduce capacity interventions on the acoustic representation passed into the Transformer encoder. In the original `wav2vec2.0` pipeline, the CNN feature encoder outputs 512-dimensional acoustic features, which are mapped by the standard feature projection to the Transformer hidden dimension $D = 768$. We keep this original projection unchanged and insert an additional low-rank affine constraint immediately after it. Specifically, let $u_t \in \mathbb{R}^{768}$ denote the output of the standard feature projection at time $t$. Our intervention maps

$$z_t = A_r u_t, \qquad A_r \in \mathbb{R}^{r\times 768},$$
$$u_t' = C_r z_t, \qquad C_r \in \mathbb{R}^{768\times r}.$$

We vary $r \ll 768$ to control the rank of the acoustic representation passed into the Transformer encoder, while keeping the encoder input dimension fixed at 768. We further validate the capacity analysis using smaller/truncated `wav2vec2.0` base encoders and additional low-rank interventions between the Transformer encoder and classifier.

### 3.4 Acoustic Perturbation Probes

By construction, ground-truth tone labels instantiate phonological abstraction, whereas corrupted training labels instantiate empirical exceptions. The model, however, only sees the resulting noisy POS objective. Early abstraction recovery may therefore reflect a population-level regularity (Natarajan et al., 2013) in the supervision, while later exception fitting reflects optimization on fixed corrupted labels. We therefore ask whether abstraction-consistent predictions and memorized exceptions are internalized in the same way.

To probe this distinction, we apply local acoustic perturbations and measure two complementary properties: *perturbation robustness*, whether a prediction is retained under label-preserving waveform changes, and *local recoverability*, whether a broken exception prediction returns to the corresponding abstraction-consistent label. Specifically, we perturb the original training speech $\mathbf{X}_i$ by adding Gaussian noise,

$$\mathrm{Aug}_s(\mathbf{X}_i) = \mathbf{X}_i + \epsilon_s,$$

where $\epsilon_s$ is scaled to achieve signal-to-noise ratio $s$ over the full utterance waveform; lower $s$ indicates stronger perturbation. This utterance-level SNR treats any pre-existing recording noise as part of the input signal. Let $\hat{y}_{i,c}^{(s)}$ denote the model prediction on $\mathrm{Aug}_s(\mathbf{X}_i)$, and let $\mathcal{C}_{\mathrm{train}} = \mathcal{V}_{\mathrm{train}} \setminus \mathcal{E}_{\mathrm{train}}$ denote uncorrupted training positions. We evaluate perturbation behavior with three probes:

- **Exception Retention.** The fraction of corrupted training positions for which the model retains the observed corrupted label after perturbation:

$$\mathrm{Acc}_{\mathrm{ER}}(s) = \frac{1}{|\mathcal{E}_{\mathrm{train}}|} \sum_{(i,c)\in\mathcal{E}_{\mathrm{train}}} \mathbb{I}\Big(\hat{y}_{i,c}^{(s)} = \tilde{y}_{i,c}\Big).$$

- **Exception Correction.** The fraction of the same corrupted positions for which the model returns to the corresponding ground-truth tone label after perturbation:

$$\mathrm{Acc}_{\mathrm{EC}}(s) = \frac{1}{|\mathcal{E}_{\mathrm{train}}|} \sum_{(i,c)\in\mathcal{E}_{\mathrm{train}}} \mathbb{I}\Big(\hat{y}_{i,c}^{(s)} = y_{i,c}^{*}\Big).$$

- **Rule Retention.** The fraction of uncorrupted training positions for which the model preserves the corresponding ground-truth tone prediction after perturbation:

$$\mathrm{Acc}_{\mathrm{RR}}(s) = \frac{1}{|\mathcal{C}_{\mathrm{train}}|} \sum_{(i,c)\in\mathcal{C}_{\mathrm{train}}} \mathbb{I}\Big(\hat{y}_{i,c}^{(s)} = y_{i,c}^{*}\Big).$$

## 4 Experimental Setup

**Data.** We conduct experiments on AISHELL-1 (Bu et al., 2017), a standard Mandarin ASR benchmark with approximately 179 hours of read speech split into 151/18/10 hours for training, validation, and testing. We use an HTK-based HMM-GMM forced aligner (Young et al., 2002)[2] to map

[2] https://htk.eng.cam.ac.uk/

word transcriptions to Pinyin initials and tonal finals. Following the central-frame labeling scheme, each aligned segment contributes only the acoustic frame nearest its temporal center under the 20 ms output stride; all other frames are assigned `0` and masked out of the cross-entropy loss. All scale, noise, and perturbation interventions are restricted to training examples; model selection and standard evaluation use the same held-out validation and test splits. Corpus statistics and alignment details are provided in Appendix A.

**Models.** For the scale–noise trajectory experiments and acoustic perturbation probes, we use `wav2vec2.0` large pretrained on 960 hours of LibriSpeech audio[3] as the backbone. For the capacity-intervention experiments, including encoder truncation and acoustic rank constraints, we use `wav2vec2.0` Base[4]. The rank constraint is inserted after the feature projection and before the Transformer encoder. Contextual representations are passed through a linear prediction head to produce logits over the label vocabulary. We fine-tune with Fairseq (Ott et al., 2019) using a two-stage schedule: all pretrained `wav2vec2.0` parameters are frozen during the first 10% of updates to warm up the prediction head, after which the Transformer encoder is fine-tuned while the convolutional feature extractor remains frozen. Accuracy is computed at aligned non-`0` frame positions against the reference labels. Experiments are run on NVIDIA RTX 4090 GPUs, and all reported numbers are single-run percentages rounded to one decimal place. Model and training details are provided in Appendix B.

## 5 Experiments

Our experiments proceed from global recoverability conditions, to epoch-wise dynamics, and finally to local perturbation robustness. We first map when abstraction-consistent performance is recoverable under scale, noise, backbone, and capacity interventions, and then use epoch-wise checkpoints to trace the abstraction-exception trajectories. Finally, given abstraction-consistent predictions and fitted exceptions, we ask whether the two show the same locality and perturbation robustness under acoustic perturbation. We organize this progression around two questions:

- **RQ1:** Across POS-style environments with different scale, noise, backbone, and capacity interventions, when is abstraction-consistent performance recoverable, and how does training move from abstraction recovery to exception fitting?
- **RQ2:** Do abstraction-consistent predictions and fitted exceptions have the same local structure and robustness under acoustic perturbation?

### 5.1 Scale–noise Performance Landscape

We construct a two-dimensional scale–noise landscape by jointly varying training scale and nominal label-replacement ratio, rather than varying either factor in isolation (Yang et al., 2025). The training scales are $\{1\%, 2\%, 3\%, 5\%, 10\%, 50\%, 100\%\}$ of the full training corpus, and the replacement ratios range from $0\%$ to $90\%$ in increments of $10\%$. For each configuration, we fine-tune a `wav2vec2.0` large model and save checkpoints after every epoch. We select the checkpoint with the lowest validation loss and report its test accuracy. Figure 1 summarizes the grid-level behavior: the x-axis denotes training scale, the y-axis denotes nominal label-replacement ratio, and color intensity indicates test accuracy.

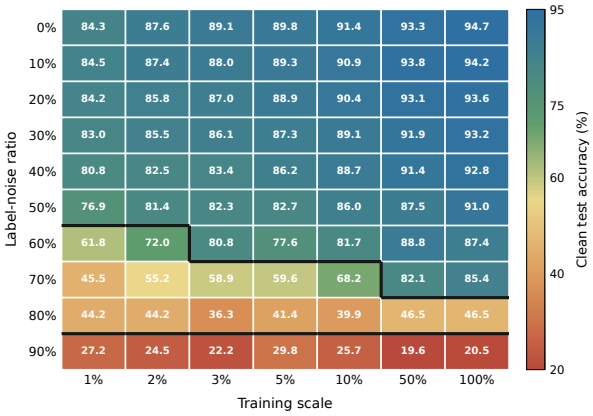


Figure 1: Test accuracy at the best-validation checkpoint across training scale and label-noise ratio. The landscape is partitioned into three empirical regimes: Robust Phonological Generalization ($\geq 75\%$), Abrupt Transition ($30\%$–$75\%$), and High-noise Collapse ($< 30\%$). These thresholds are used for descriptive grouping rather than theoretical phase boundaries.

**Three recoverability regimes.** Figure 1 shows that abstraction-consistent test performance does not degrade uniformly with label replacement. Instead, the scale–noise landscape separates into three recoverability regimes: *(i) Robust Phonological Generalization.* The upper landscape forms

[3] https://dl.fbaipublicfiles.com/fairseq/wav2vec/libri960_big.pt

[4] https://dl.fbaipublicfiles.com/fairseq/wav2vec/wav2vec_small.pt

a broad high-accuracy plateau: across all training scales, wav2vec2.0 maintains above 75% test accuracy up to 70% nominal label replacement, indicating robust recovery of abstraction-consistent tone performance under sparse and noisy supervision. Additional speech exposure expands this recoverable region: while the 1%-scale model reaches 80.8% accuracy at 40% replacement, the 100%-scale model still reaches 85.4% at 70% replacement. *(ii) Abrupt Transition.* Once corrupted supervision becomes dominant, the high-accuracy plateau gives way to a sharp transition band. Increasing training scale delays this boundary but does not remove it: at 100% scale, accuracy still drops from 85.4% at 70% replacement to 46.5% at 80%, marking a recoverability limit for abstraction-consistent tone generalization. *(iii) High-noise Collapse*. At 90% label noise, test accuracy falls near chance across all scales, indicating that abstraction-consistent test behavior is no longer recoverable.

| Run | Final Train | Last Test | Best-Val Test |
|---|---|---|---|
| 12L, 50%, 50% | 100.0 | 56.6 | 88.2 |
| 6L, 50%, 50% | 100.0 | 58.7 | 86.9 |
| 4L, 50%, 50% | 100.0 | 56.1 | 88.1 |
| 2L, 50%, 50% | 100.0 | 55.9 | 83.1 |
| 2L, 100%, 50% | 100.0 | 54.4 | 86.3 |
| 2L, 100%, 90% | 100.0 | 25.8 | 22.2 |

Table 1: Depth-control results. Each run is denoted by encoder depth, training scale, and label-noise ratio; Best-Val Test is selected by validation loss.

**Backbone and capacity checks.** The landscape above is obtained with wav2vec2.0 large. We therefore ask whether the three-regime pattern is specific to this backbone or simply reflects excessive model capacity. As a backbone check, we repeat the analysis with HuBERT at 10% training scale and observe the same qualitative pattern (Appendix C). We then test capacity in two ways: reducing Transformer depth with wav2vec2.0 base variants and restricting the pre-Transformer acoustic channel with a trainable $768 \rightarrow r \rightarrow 768$ bottleneck. Table 1 shows that depth reduction does not remove the pattern: even a two-layer encoder reaches 83.1% and 86.3% best-validation test accuracy in the 50%-replacement runs, while all depth-control runs reach 100.0% final training accuracy. Table 2 further shows that abstraction-consistent tone performance remains partly recoverable even under severe acoustic rank compression, with $r = 2$ still reaching 67.6% at 70% replacement. Together, these checks suggest that the observed scale–noise regimes are not merely a wav2vec2.0 large artifact.

| Rank $r$ | 50% | 60% | 70% | 80% |
|---|---|---|---|---|
| 2 | 79.9 | 76.1 | 67.6 | 46.6 |
| 4 | 82.3 | 78.0 | 71.7 | 46.5 |
| 8 | 84.6 | 79.8 | 75.9 | 43.1 |
| 32 | 88.0 | 85.1 | 76.5 | 38.4 |
| 64 | 87.6 | 85.3 | 82.5 | 36.4 |

Table 2: Best-validation test accuracy under acoustic rank interventions at 50% training scale. Columns denote label-noise ratios.

**Rank compression modulates both abstraction and exception.** The rank results further show that acoustic capacity does not act as a simple monotonic recovery knob. In the recoverable 50%–70% replacement regimes, increasing rank generally improves best-validation accuracy: for example, at 70% replacement, performance rises from 67.6% at $r = 2$ to 82.5% at $r = 64$. This indicates that stronger abstraction recovery still benefits from additional acoustic information. At 80% replacement, however, the trend reverses: ranks 2 and 4 reach 46.6% and 46.5%, while ranks 32 and 64 fall to 38.4% and 36.4%. This suggests that near the transition boundary, additional acoustic capacity may also make the model more susceptible to following corrupted supervision. Thus, rank compression limits acoustic information rather than memorization alone: it can partly blunt harmful fitting pressure in the highest-noise setting, but it also weakens abstraction recovery in recoverable regimes. Capacity therefore modulates the recovery–fitting tradeoff rather than cleanly separating abstraction generalization from exception memorization. A post-encoder capacity constraint ablation shows the same pattern when the rank constraint is moved after the encoder (Appendix D).

### 5.2 Epoch-wise Learning Trajectories

The scale–noise landscape shows where abstraction-consistent performance is recoverable, but not how recovery and fitting unfold during training. We therefore track all 70 configurations across epochs with three probes: Generalization, Self-Correction, and Exception Memorization. Representative trajectories from the three landscape regimes are shown in Figure 2, with the full set provided in the submitted data. Across recoverable settings, the trajectories reveal a consistent temporal ordering: abstraction-consistent generalization

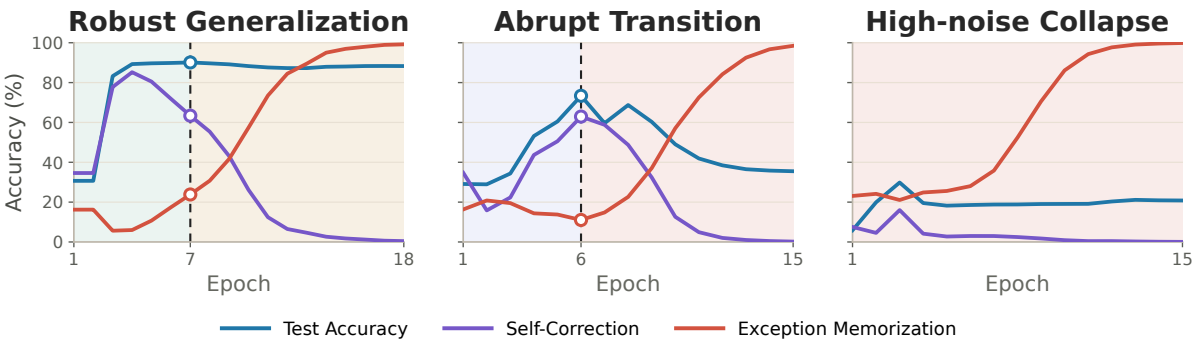


Figure 2: Epoch-wise trajectories at 5% training scale under 10%, 70%, and 90% label noise, corresponding to three representative regimes: *Robust Generalization*, *Abrupt Transition*, and *High-noise Collapse*. Curves show Generalization, Self-Correction, and Exception Memorization; shading marks the associated learning regimes.

and self-correction emerge first, followed by later exception memorization. The three regimes differ in whether this recovered abstraction-consistent behavior remains stable, is later degraded, or never becomes stable.

**Robust Generalization Regime.** In the robust regime, abstraction recovery appears early and remains stable throughout training. Test accuracy changes little from the peak to the final epoch ($90.1\% \rightarrow 88.3\%$), even though Exception Memorization rises later to near-perfect levels. This indicates that late corrupted-label fitting does not necessarily disrupt the recovered phonological abstraction: in this regime, exceptions are fitted on top of a stable abstraction-consistent mapping rather than replacing it.

**Abrupt Transition Regime.** In the transition regime, the model still enters an early recovery phase, but this phase is not stable. The representative run reaches 73.4% test accuracy at the peak epoch under 70% replacement, showing that substantial abstraction-consistent generalization is still recoverable under heavy corruption. Continued training then sharply reverses this state: by the final epoch, test accuracy drops by 37.9 points and Self-Correction drops by 62.7 points. This marks a temporal shift from abstraction recovery toward exception fitting, where fitting the corrupted labels begins to overwrite the earlier behavior.

**High-noise Collapse Regime.** In the collapse regime, the abstraction-consistent phase never becomes reliable: under 90% nominal replacement, peak test accuracy reaches only 29.8%. Exception Memorization nevertheless rises to nearly 100%, showing that the model can fit corrupted training labels even when clean generalization remains near chance. Thus, collapse reflects direct corrupted-label fitting rather than the degradation of a previously stable abstraction.

### 5.3 Acoustic Perturbation for Abstraction and Exceptions

Capacity and trajectory analyses show that abstraction-consistent behavior and exception fitting are coupled at the level of acoustic capacity, but they do not reveal whether the corresponding predictions have the same local perturbation profile. We therefore apply the acoustic perturbation probes to compare perturbation robustness and local recoverability: if abstraction-consistent predictions rely on broader local neighborhoods, they should remain stable under perturbation, whereas local exception memories should be easier to break or shift toward the corresponding abstraction-consistent label. We add Gaussian noise at SNR levels $s \in \{50, 45, 40, 30, 25, 20, 15, 10, 8, 6, 4, 2, 0\}$ dB and evaluate Exception Retention, Exception Correction, and Rule Retention on both best-validation and last checkpoints. We report two representative settings where generalization succeeds: 10% noise for *Robust Phonological Generalization*, and 70% noise for *Abrupt Transition*. The 90%-noise High-noise Collapse case and three additional perturbation types are reported in Appendix E, showing the same pattern discussed in the main text.

**Abstractions are more perturbation-stable than exceptions.** Figure 3 shows a sharp robustness contrast between abstraction-consistent predictions and fitted exceptions. In the 10%-noise regime, the last checkpoint retains the observed corrupted label almost perfectly under weak perturbation (98.7% retention at 50 dB), but this retention drops to 34.1% by 30 dB. In contrast, rule retention remains high under the same perturbation strength (96.3% at 30 dB and 90.0% at 20 dB), falling below 75%

only under much stronger noise. Thus, abstraction-consistent predictions are not simply another form of memorization: they behave as if supported by broader and more stable local neighborhoods under perturbation, whereas fitted exceptions depend on narrower local acoustic neighborhoods.

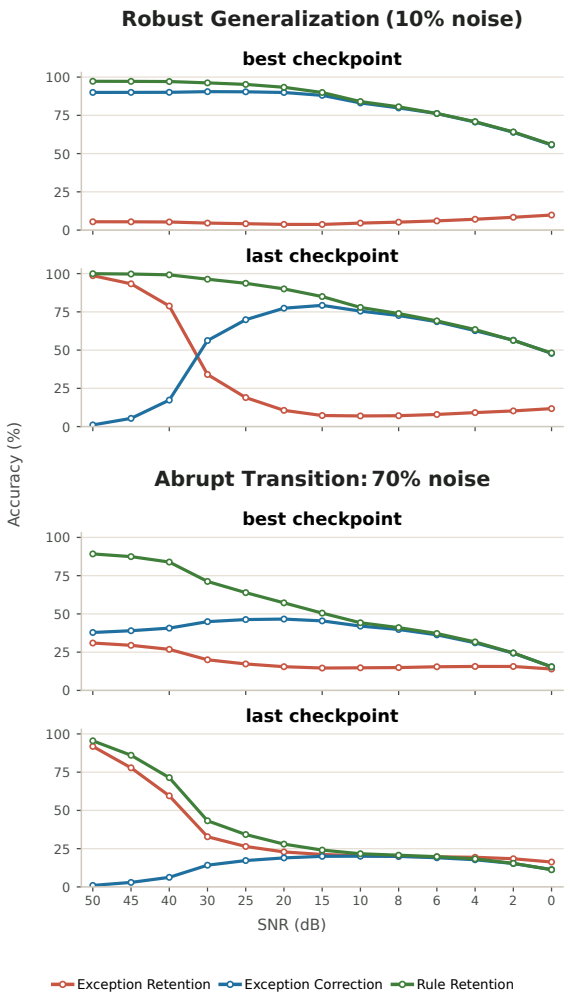


Figure 3: Acoustic perturbation probes at 100% training scale for *Robust Phonological Generalization* (10% label noise) and *Abrupt Transition* (70% label noise). For each regime, we compare the best-validation and last checkpoints under increasing Gaussian perturbation strength. Lower SNR indicates stronger perturbation.

**Perturbation exposes fragile exception fitting.** The 10%-noise regime further shows that fitted exceptions appear is locally overlaid on broader abstraction-consistent behavior. At the best-validation checkpoint, the model is abstraction-dominant: Exception Correction stays around 90% under weak and moderate perturbations, while Exception Retention remains below 6%. At the last checkpoint, the model has fitted the observed corrupted labels on the original waveforms, but these memories are fragile: as Exception Retention drops from 98.7% at 50 dB to 34.1% at 30 dB and 10.6% at 20 dB, Exception Correction rises from 1.1% to 56.2% and 77.4%, respectively. This pattern indicates that perturbation can move a waveform away from an exception-fitting neighborhood and recover the abstraction-consistent label.

**Heavy noise weakens local abstraction recoverability.** The 70%-noise regime reveals a different perturbation-recoverability profile. At the best-validation checkpoint, the model still shows partial abstraction recovery: Exception Correction reaches 37.9% at 50 dB and 44.9% at 30 dB, while Rule Retention remains 71.2% at 30 dB. At the last checkpoint, however, weak perturbation strongly preserves the observed corrupted labels (91.8% Exception Retention at 50 dB), but breaking this retention no longer reliably recovers the abstraction-consistent label. At 30 dB, Exception Retention falls to 32.8%, yet Exception Correction reaches only 14.2%, and Rule Retention drops to 43.2%. Thus, late-stage exception fitting under heavy noise does not merely add local islands on top of a stable abstraction; it weakens the local recoverability of the abstraction-consistent label, so perturbation leaves the exception without reliably returning to the abstraction-consistent label.

## 6 Conclusion

In this work, we study how pretrained speech transformers such as `wav2vec2.0` learn from sparse and noisy tone supervision. Across scale–noise, capacity, and perturbation analyses, we find a consistent trajectory: recoverable models first acquire shared phonological regularities and only later fit sample-specific corrupted labels as exceptions. Perturbation probes further show that abstraction-consistent predictions occupy broader local neighborhoods, whereas fitted exceptions remain fragile around the original waveform. Further work based on the released checkpoints to examine the internal representations underlying these dynamics is especially promising.

## 7 Ethics and AI Assistance

The released checkpoints could potentially be misused for adversarial robustness research on speech models (Qin et al., 2019). We used ChatGPT for sentence-level language refinement and code assistance. All scripts, experimental results, and revisions were carefully reviewed and verified.

## Limitations

Due to the workload, we use tone classification and random label replacement as compact proxies for abstraction and memorization. In the future, we plan to include broader hierarchical abstraction, such as prosodic phrase structure (Nespor and Vogel, 2007), as well as linguistically grounded exceptions, such as morphophonological alternations (Mohanan, 1982). We only analyze behavior-level metrics here. A promising next step is to use the released epoch-wise checkpoints for internal analyses. Finally, the proposed abstraction–exception perspective could be tested systematically in other modern architectures like diffusion models (Ho et al., 2020), larger data scales, and additional modalities.

## References


Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*.

Ben Ambridge. 2020. Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, 40(5-6):509–559.

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, Yoshua Bengio, and Simon Lacoste-Julien. 2017. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Robert C Berwick, Paul Pietroski, Beracah Yankama, and Noam Chomsky. 2011. Poverty of the stimulus revisited. *Cognitive science*, 35(7):1207–1242.

Hui Bu, Jiayu Du, X. Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *Proceedings of O-COCOSDA 2017*.

Noam Chomsky. 1980. Rules and representations. *Behavioral and brain sciences*, 3(1):1–15.

Noam Chomsky, Ian Roberts, and Jeffrey Watumull. 2023. Noam chomsky: The false promise of chatgpt. *The New York Times*, 8(4):177–179.

Vitaly Feldman and Chiyuan Zhang. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in neural information processing systems*, 33:2881–2891.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Ko-Wei Huang, Yi-Fu Fu, Ching-Yu Tsai, Yu-Chieh Tu, Tzu-ling Cheng, Cheng-Yu Lin, Yi-Ting Yang, Heng-Yi Liu, Keng-Te Liao, Da-Cheng Juan, and Shou-De Lin. 2025. Neuron-level differentiation of memorization and generalization in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16077–16091.

Peter W Jusczyk and Richard N Aslin. 1995. Infants' detection of the sound patterns of words in fluent speech. *Cognitive psychology*, 29(1):1–23.

Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. 2019. Sgd on neural networks learns functions of increasing complexity. *Advances in neural information processing systems*, 32.

Patricia K Kuhl. 2004. Early language acquisition: Cracking the speech code. *Nature reviews neuroscience*, 5(11):831–843.

Baishen Liang and Yi Du. 2018. The functional neuroanatomy of lexical tone perception: an activation likelihood estimation meta-analysis. *Frontiers in neuroscience*, 12:495.

Alvin M Liberman, Katherine Safford Harris, Howard S Hoffman, and Belver C Griffith. 1957. The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology*, 54(5):358.

Jeffrey Lidz, Sandra Waxman, and Jennifer Freedman. 2003. What infants know about syntax but couldn't have learned: Experimental evidence for syntactic structure at 18 months. *Cognition*, 89(3):295–303.

Zeyi Lin, Shaohong Chen, Kang Li, Qiushan Jiang, Zirui Cai, Kaifang Ji, and The SwanLab Team. 2023. SwanLab.

Sadegh Mahdavi, Renjie Liao, and Christos Thrampoulidis. 2024. Memorization capacity of multi-head attention in transformers. In *International Conference on Learning Representations*, volume 2024, pages 49942–49967.

Kyle Mahowald, George Kachergis, and Michael C Frank. 2020. What counts as an exemplar model, anyway? a commentary on ambridge (2020). *First Language*, 40(5-6):608–611.

R Thomas McCoy, Robert Frank, and Tal Linzen. 2020. Does syntax need to grow on trees? sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8:125–140.

R. Thomas McCoy and Thomas L. Griffiths. 2025. Modeling rapid language learning by distilling bayesian priors into artificial neural networks. *Nature communications*, 16(1):4676.

Karuvannur Puthanveettil Mohanan. 1982. *Lexical phonology*. Ph.D. thesis, Massachusetts Institute of Technology.

John Xavier Morris, Chawin Sitawarin, Narine Kokhlikyan, Chuan Guo, G Edward Suh, Alexander M Rush, Kamalika Chaudhuri, and Saeed Mahloujifar. 2025. How much can language models memorize?

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. 2021. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. 2013. Learning with noisy labels. *Advances in neural information processing systems*, 26.

Marina Nespor and Irene Vogel. 2007. *Prosodic phonology: with a new foreword*, volume 28. Walter de Gruyter.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (Demonstrations)*, pages 48–53.

Madhur Panwar, Gail Weiss, Navin Goyal, and Antoine Bosselut. 2026. For better or for worse, transformers seek patterns for memorization. *Advances in Neural Information Processing Systems*, 38:104001–104031.

Amy Perfors, Joshua B Tenenbaum, and Terry Regier. 2011. The learnability of abstract syntactic principles. *Cognition*, 118(3):306–338.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*.

Geoffrey K Pullum and Barbara C Scholz. 2002. Empirical assessment of stimulus poverty arguments. *The linguistic review*, 19(1-2):9–50.

Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. 2019. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *International conference on machine learning*, pages 5231–5240. PMLR.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. 2019. On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR.

Jenny R Saffran, Richard N Aslin, and Elissa L Newport. 1996. Statistical learning by 8-month-old infants. *science*, 274(5294):1926–1928.

Lan Shuai and Tao Gong. 2014. Temporal relation between top-down and bottom-up processing in lexical tone perception. *Frontiers in behavioral neuroscience*, 8:97.

Alex Warstadt and Samuel R Bowman. 2020. Can neural networks acquire a structural bias from raw linguistic data? *arXiv preprint arXiv:2007.06761*.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language*, pages 17–60. CRC Press.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjape, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the babylm challenge at the 27th conference on computational natural language learning*, pages 1–34.

Yi Yang, Yiming Wang, and Jiahong Yuan. 2025. Transformer-based speech model learns well as infants and encodes abstractions through exemplars in the poverty of the stimulus environment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7881–7890.

Aditya Yedetore, Tal Linzen, Robert Frank, and R Thomas McCoy. 2023. How poor is the stimulus? evaluating hierarchical generalization in neural networks trained on child-directed speech. *arXiv preprint arXiv:2301.11462*.

Moira Jean Winsland Yip. 2002. *Tone*. Cambridge University Press.

S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. 2002. The htk book. *Cambridge university engineering department*, 3(175):12.

Fangyuan Yu. 2025. Memorization-compression cycles improve generalization. *arXiv preprint arXiv:2505.08727*.

Jiahong Yuan, Xingyu Cai, and Kenneth Church. 2023. Improved contextualized speech representations for tonal analysis. In *INTERSPEECH*, pages 4513–4517.

# A Data

## A.1 AISHELL-1 Corpus

AISHELL-1 (Bu et al., 2017) is a public Mandarin read-speech corpus recorded in quiet indoor environments, containing 400 speakers from different accent regions in China, 16 kHz audio, and manually verified transcriptions. We use the standard train/dev/test partitions. We do not collect new human-subject data, attempt speaker identification, or release additional personal metadata; all derived artifacts use only corpus-provided IDs, transcripts, and aligned tone labels, and are intended for research use under the original corpus access conditions.

| Split | Speakers | Utterances | Hours | Targets |
|---|---|---|---|---|
| Train | 340 | 120,098 | 150.85 | 1,986,313 |
| Dev | 40 | 14,326 | 18.09 | 235,764 |
| Test | 20 | 7,176 | 10.03 | 122,495 |
| Total | 400 | 141,600 | 178.97 | 2,344,572 |

Table 3: AISHELL-1 statistics. Targets denote aligned non-`0` central-frame labels used for training or evaluation.

The label set contains the four canonical Mandarin lexical tones (`T1`–`T4`), the neutral tone (`T5`), and aligned silence or pause intervals (`sil`).

| Label | Train | Dev | Test | Total |
|---|---|---|---|---|
| T1 | 377,006<br>19.0% | 43,963<br>18.6% | 23,042<br>18.8% | 444,011<br>18.9% |
| T2 | 376,983<br>19.0% | 44,828<br>19.0% | 22,761<br>18.6% | 444,572<br>19.0% |
| T3 | 249,607<br>12.6% | 29,726<br>12.6% | 15,037<br>12.3% | 294,370<br>12.6% |
| T4 | 611,915<br>30.8% | 73,008<br>31.0% | 37,264<br>30.4% | 722,187<br>30.8% |
| T5 | 79,029<br>4.0% | 9,540<br>4.0% | 4,551<br>3.7% | 93,120<br>4.0% |
| sil | 291,773<br>14.7% | 34,699<br>14.7% | 19,840<br>16.2% | 346,312<br>14.8% |

Table 4: Distribution of aligned non-`0` central-frame targets. Percentages are computed within each split.

## A.2 Forced Alignment Configuration

We use an HTK-based HMM-GMM forced aligner to obtain phone-level boundaries. Audio is represented at 16 kHz. The pronunciation lexicon maps Chinese word transcriptions to Pinyin initials and tone-marked finals. Silence and short pauses are

modeled as separate `sil` intervals and retained as `sil` labels, while all non-target frames are assigned `0` during tone classification.

The aligner uses a context-independent monophone HMM-GMM acoustic model. Each phoneme is modeled without left or right phonetic context, and each emitting HMM state uses a four-component Gaussian mixture model. Alignment is performed against the reference transcription, producing start and end times for each phone segment. For each tone-bearing final, we assign the corresponding tone label `T1`–`T5` to the acoustic frame nearest the segment center under the 20 ms `wav2vec2.0` output stride. All other frames are masked out in the loss.

### A.3 Scale–Noise Configuration Statistics

We first generate training subsets at different scales, and then apply label noise from $0\%$ to $90\%$ in increments of $10\%$ for each scale, yielding $7 \times 10 = 70$ configurations.

**Scale Generation.** We construct each scale by selecting AISHELL-1 training speakers and retaining all of their utterances, rather than subsampling utterances independently. This preserves speaker-level structure. Table 5 summarizes the resulting subset statistics.

| Scale | Speakers | Utterances | Hours | Labels |
|---|---|---|---|---|
| 1% | 3 | 1,083 | 1.51 | 18,592 |
| 2% | 6 | 2,166 | 2.87 | 37,207 |
| 3% | 10 | 3,530 | 4.61 | 59,981 |
| 5% | 17 | 5,987 | 7.53 | 100,743 |
| 10% | 34 | 12,027 | 15.24 | 200,497 |
| 50% | 170 | 60,201 | 75.73 | 995,046 |
| 100% | 340 | 120,098 | 150.85 | 1,986,313 |

Table 5: Statistics of the nested speaker-level training subsets used for scale-controlled experiments. Labels count valid non-`0` central-frame labels.

**Noise Generation.** For each scale-controlled training subset, we generate noisy training sets with nominal replacement ratios from $0\%$ to $90\%$ in increments of $10\%$. Noise is applied only to tone-bearing central-frame labels in $\{\texttt{T1}, \ldots, \texttt{T5}\}$; `sil` labels and all `0` frames are left unchanged. For a nominal replacement ratio $r$, each tone-bearing target in `train.ltr` is selected with probability $r$. If selected, its label is replaced with a uniformly sampled tone label from $\{\texttt{T1}, \ldots, \texttt{T5}\}$. All validation and test manifests are copied from the corresponding clean scale-controlled split. Because `sil` labels are included among aligned non-`0` targets but are never replaced, the nominal replacement ratio differs from the changed-label rate measured over all non-`0` targets.

## B Model and Fine-tuning Details

### B.1 Backbone Configurations

All `wav2vec2.0` variants use the same 7-layer temporal CNN feature encoder with 512 channels and kernel/stride configurations $(10, 5), (3, 2), (3, 2), (3, 2), (3, 2), (2, 2), (2, 2)$. The resulting total stride is 320 samples, corresponding to approximately 20 ms per frame for 16 kHz audio. Table 6 summarizes the Transformer encoder configurations used in our experiments. The base model has 12 Transformer layers with hidden size 768, feed-forward dimension 3072, and 12 attention heads, while the large model has 24 layers with hidden size 1024, feed-forward dimension 4096, and 16 attention heads. For depth-control experiments, we truncate the base encoder by retaining only the first $L \in \{2, 4, 6, 10, 12\}$ Transformer layers; the convolutional feature encoder, feature projection, and prediction head are kept unchanged.

| Model | Layers | Hidden | FFN | Heads | Params |
|---|---|---|---|---|---|
| large | 24 | 1024 | 4096 | 16 | 317.38M |
| base | 12 | 768 | 3072 | 12 | 95.04M |
| base-10L | 10 | 768 | 3072 | 12 | 80.87M |
| base-6L | 6 | 768 | 3072 | 12 | 52.52M |
| base-4L | 4 | 768 | 3072 | 12 | 38.34M |
| base-2L | 2 | 768 | 3072 | 12 | 24.17M |

Table 6: `wav2vec2.0` backbones and truncated base variants. Parameter counts are computed from the Fairseq checkpoints used for fine-tuning.

### B.2 Fine-tuning Pipeline

We fine-tune all models with Fairseq (Ott et al., 2019) using Adam and a linear prediction head over the label vocabulary. Unless otherwise specified, runs use a peak learning rate of $3 \times 10^{-5}$, cosine learning-rate decay, and update frequency 1. The learning-rate warmup and frozen-backbone period are each set to $10\%$ of the total update budget. During the frozen period, the prediction head is warmed up while pretrained `wav2vec2.0` parameters are fixed; afterward, the Transformer encoder is fine-tuned while the convolutional feature encoder remains frozen.

The scale–noise trajectory experiments use `wav2vec2.0` large and save epoch-wise checkpoints for trajectory analysis. The capacity-intervention experiments use `wav2vec2.0` base or its truncated variants. Acoustic rank constraint runs insert a trainable $768 \to r \to 768$ module after the feature projection and before the Transformer encoder; the post-encoder rank constraint ablation instead inserts the constraint between the Transformer encoder and the prediction head.

Compared with Yang et al. (2025), our setup makes two changes. First, scale refers to the amount of speech available for training: a $k\%$ condition uses approximately $k\%$ of AISHELL-1 training speech, selected at the speaker level, rather than retaining full speech while reducing labeled targets. Second, after the initial warmup, we fine-tune the encoder but keep the convolutional feature encoder frozen throughout training; the earlier setup also fine-tuned the convolutional feature encoder.

### B.3 Logging and Checkpointing

Training metrics, hyperparameters, and checkpoint metadata are logged with SwanLab (Lin et al., 2023). We validate once per epoch and select best-validation checkpoints using the development set. For the main trajectory grid, checkpoints are saved after each epoch to support epoch-wise analysis. For capacity interventions used only for endpoint comparisons, we keep the best and last checkpoints unless epoch-wise trajectories are required.

## C Rule–Exception Dynamics with HuBERT

To test whether the scale–noise dynamics are specific to `wav2vec2.0`, we repeat the trajectory analysis with the Fairseq HuBERT base checkpoint[5], pretrained on 960 hours of LibriSpeech. The checkpoint contains a 12-layer Transformer encoder with hidden size 768, feed-forward dimension 3072, 12 attention heads, and 94.70M parameters. Table 7 reports the best-validation test accuracies, and Figure 4 shows the corresponding epoch-wise trajectories.

## D Acoustic Capacity

To test whether the capacity effect depends on where the low-rank constraint is inserted, we run an additional ablation that places a low-rank constraint after the Transformer encoder and before the final classifier. This intervention restricts the contextual representation used for classification, rather than the pre-Transformer acoustic channel used in the main capacity probe. We evaluate this setting on the $50\%$-scale, $50\%$-noise condition using the full test set.

The pattern matches the main capacity analysis. Low-rank constraint still allows strong early generalization, with best-checkpoint test accuracy reaching $83.4\%$–$89.7\%$. However, all ranks eventually fit the corrupted training labels almost perfectly, reaching $99.5\%$–$99.9\%$ final training accuracy, while final test accuracy drops to $49.6\%$–$55.7\%$.

## E Acoustic Perturbation Probes

### E.1 SNR Definition

We control perturbation strength by the signal-to-noise ratio (SNR). For signal power $P_{\text{signal}}$ and noise power $P_{\text{noise}}$, we define

$$\mathrm{SNR}_{\mathrm{dB}} = 10 \log_{10} \frac{P_{\text{signal}}}{P_{\text{noise}}},$$

$$\frac{P_{\text{signal}}}{P_{\text{noise}}} = 10^{\mathrm{SNR}_{\mathrm{dB}}/10},$$

$$\frac{A_{\text{signal}}}{A_{\text{noise}}} = 10^{\mathrm{SNR}_{\mathrm{dB}}/20}.$$

Here, $A_{\text{signal}}/A_{\text{noise}}$ denotes the corresponding RMS amplitude ratio. Lower SNR indicates stronger perturbation; at 0 dB, signal and noise have equal power.

### E.2 Robustness to Perturbation Type

The main perturbation probes use additive Gaussian noise. To test whether the observed rule–exception separation depends on this specific noise distribution, we repeat the same analysis with alternative additive perturbations matched by target SNR. For each waveform $x \in \mathbb{R}^T$, we construct

$$x_s = x + \alpha_s \nu,$$

where $\nu$ is zero-centered and $\alpha_s$ scales the perturbation to the target SNR $s$. All perturbations preserve waveform length and therefore do not change frame-level tone alignment. In all cases, $\alpha_s$ is computed from the RMS ratio between $x$ and $\nu$.

**Uniform white noise.** We sample bounded i.i.d. noise,

$$\nu_t \sim \mathrm{Uniform}(-\sqrt{3}, \sqrt{3}),$$

which has unit variance before SNR scaling.

[5] https://dl.fbaipublicfiles.com/hubert/hubert_base_ls960.pt

| **Noise (%)** | 0 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Best valid epoch** | 7 | 5 | 6 | 5 | 6 | 5 | 5 | 5 | 4 | 13 |
| **Test acc. (%)** | 90.3 | 89.8 | 89.2 | 88.8 | 88.0 | 86.4 | 84.9 | 79.2 | 46.5 | 25.4 |

Table 7: HuBERT best-validation results at $10\%$ training scale across label-noise ratios. Each column reports the selected epoch and its corresponding test accuracy.

| Rank $r$ | Best Epoch | Best Train | Best Test | Final Train | Final Test |
|---|---|---|---|---|---|
| 2 | 7 | 53.2 | 83.4 | 99.5 | 49.6 |
| 8 | 7 | 58.2 | 88.5 | 99.8 | 54.4 |
| 16 | 6 | 56.9 | 89.6 | 99.9 | 55.7 |
| 64 | 6 | 57.1 | 89.7 | 99.9 | 55.1 |

Table 8: Post-encoder capacity intervention ablation at $50\%$ training scale and $50\%$ label noise. The low-rank constraint is inserted between the Transformer encoder and the classifier. "Best" refers to the checkpoint selected by development loss.

| SNR (dB) | $P_{\text{signal}}/P_{\text{noise}}$ | $A_{\text{signal}}/A_{\text{noise}}$ |
|---|---|---|
| 50 | 100,000.00 | 316.23 |
| 45 | 31,622.78 | 177.83 |
| 40 | 10,000.00 | 100.00 |
| 30 | 1,000.00 | 31.62 |
| 25 | 316.23 | 17.78 |
| 20 | 100.00 | 10.00 |
| 15 | 31.62 | 5.62 |
| 10 | 10.00 | 3.16 |
| 8 | 6.31 | 2.51 |
| 6 | 3.98 | 2.00 |
| 4 | 2.51 | 1.58 |
| 2 | 1.58 | 1.26 |
| 0 | 1.00 | 1.00 |

Table 9: Conversion between SNR in decibels and signal-to-noise ratios. Lower SNR corresponds to stronger additive Gaussian noise.

**Laplace white noise.** We sample i.i.d. heavy-tailed noise,

$$\nu_t \sim \text{Laplace}\left(0, \frac{1}{\sqrt{2}}\right),$$

which also has unit variance before SNR scaling.

**Pink Gaussian noise.** We sample white Gaussian noise $w_t \sim \mathcal{N}(0, 1)$, transform it to the frequency domain, and apply $1/\sqrt{f}$ spectral shaping:

$$\hat{\nu}(f) = \frac{\hat{w}(f)}{\sqrt{f + \epsilon}}, \qquad \hat{w}(f) = \mathcal{F}(w),$$

with the DC component set to zero. We then transform back to the time domain and rescale $\nu$ to the target SNR.

As illustrated in Figure 5,6,7, the resulting curves follow the same qualitative pattern as the Gaussian probes across these perturbation types.

### E.3 Perturbations in the High-Noise Collapse Regime

Figure 8 reports the perturbation response at $100\%$ training scale and $90\%$ label noise, corresponding to the high-noise collapse regime. Unlike the robust and transition settings in Figure 3, this condition does not show a recoverable phonological rule even at the best-validation checkpoint.

At the best-validation checkpoint, all three curves remain low, indicating that the model has not formed a stable rule-consistent region near the training waveforms. At the last checkpoint, Exception Retention is high under weak perturbation, showing that the model has fitted the corrupted labels on the original inputs. However, as perturbation strength increases, Exception Retention drops sharply while Exception Correction remains low. Thus, breaking the fitted exception does not recover the corresponding rule-consistent label. This supports the collapse interpretation: under extreme label noise, late fitting produces fragmented local memorization rather than exception islands overlaid on a recoverable phonological rule.

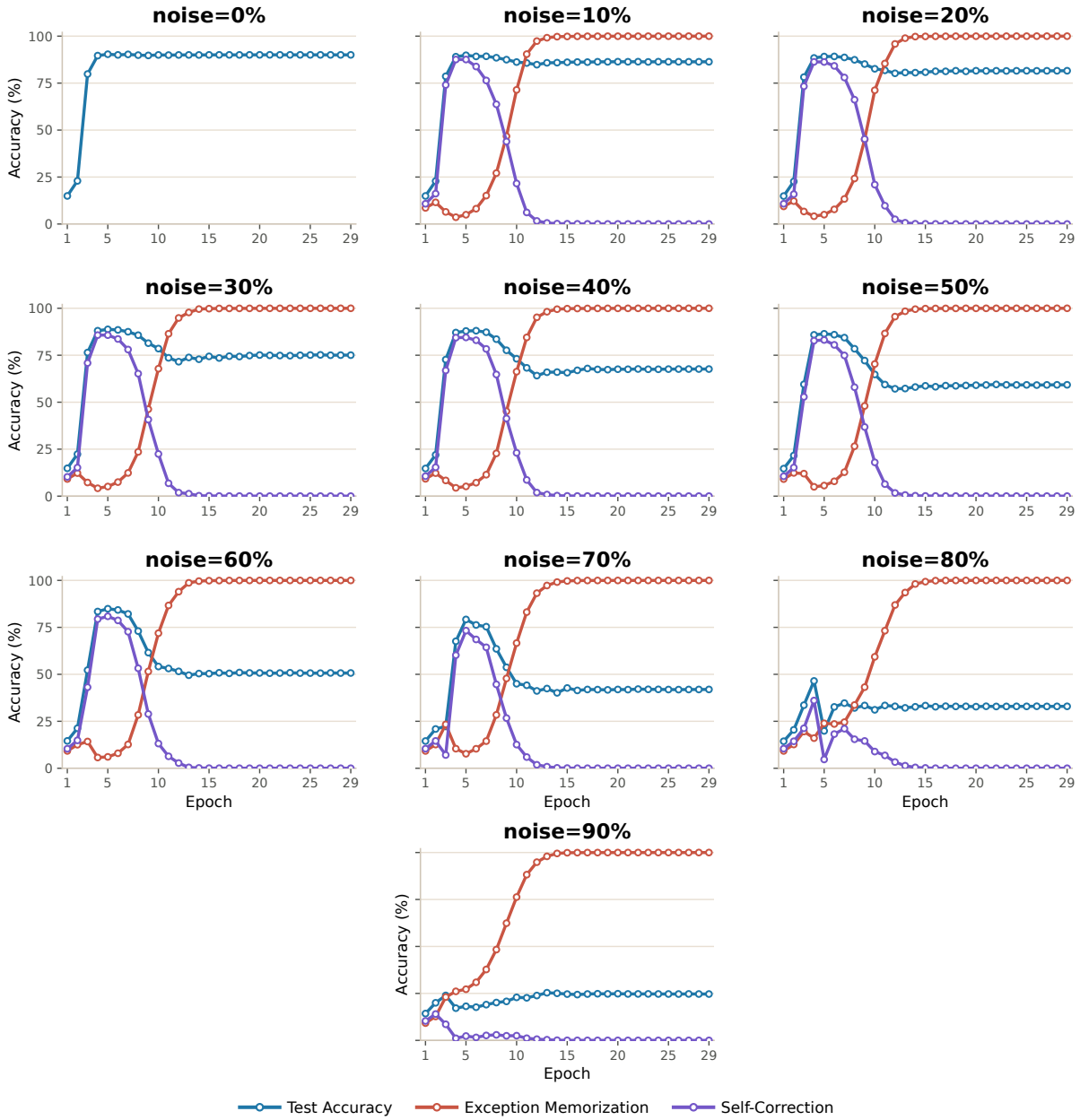


Figure 4: HuBERT epoch-wise trajectories at 10% training scale across label-noise ratios from 0% to 90%. Blue, purple, and red curves denote Generalization, Self-Correction, and Exception Memorization. At 0% label noise, Self-Correction and Exception Memorization are undefined.

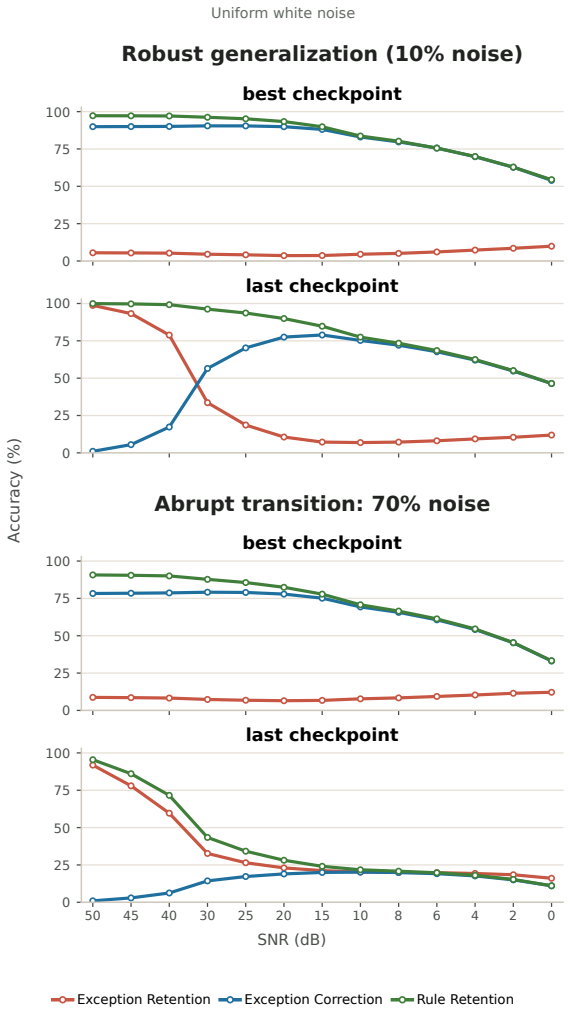


Figure 5: Acoustic perturbation probes at 100% training scale with uniform white noise. The upper two panels show the robust generalization regime with 10% label noise, and the lower two panels show the abrupt transition regime with 70% label noise. For each regime, we compare the best-validation checkpoint and the last checkpoint. Exception Retention measures whether corrupted positions still predict the noisy label; Exception Correction measures whether corrupted positions return to the clean label; Rule Retention measures whether uncorrupted positions remain ground-truth-consistent. Lower SNR indicates stronger perturbation.

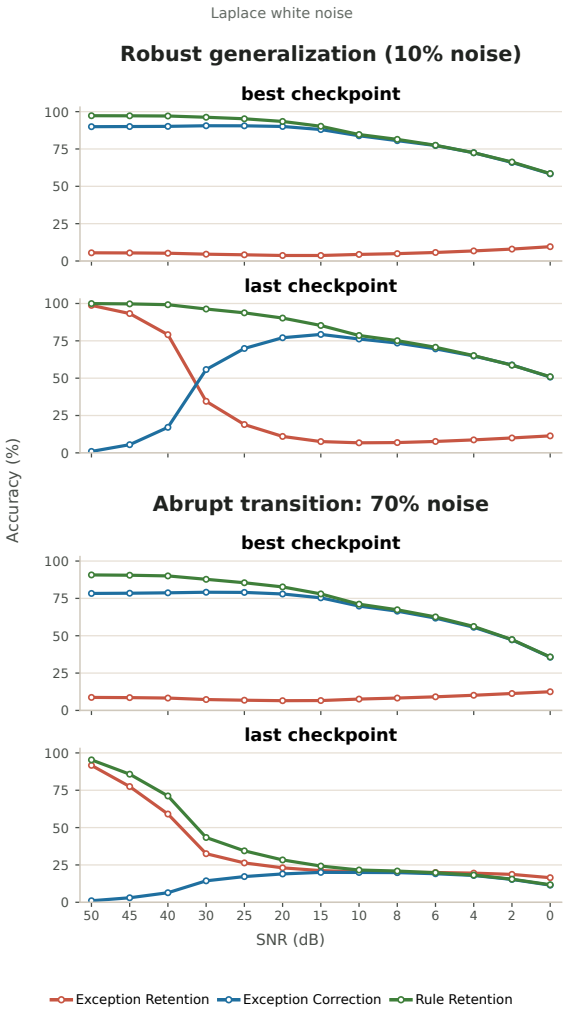


Figure 6: Acoustic perturbation probes at 100% training scale with laplace white noise. The upper two panels show the robust generalization regime with 10% label noise, and the lower two panels show the abrupt transition regime with 70% label noise. For each regime, we compare the best-validation checkpoint and the last checkpoint. Exception Retention measures whether corrupted positions still predict the noisy label; Exception Correction measures whether corrupted positions return to the clean label; Rule Retention measures whether uncorrupted positions remain ground-truth-consistent. Lower SNR indicates stronger perturbation.

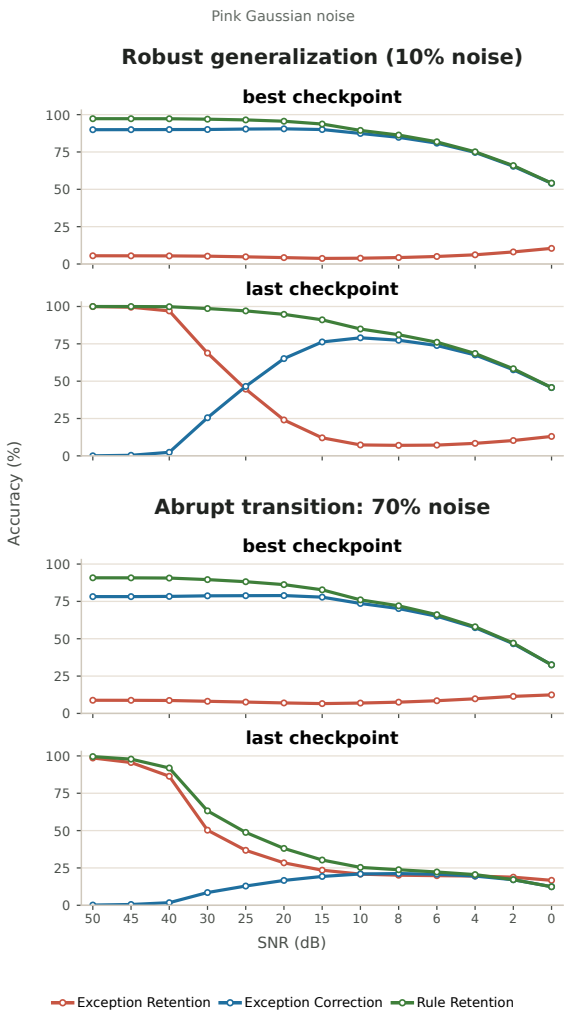


Figure 7: Acoustic perturbation probes at 100% training scale with Pink Gaussian noise. The upper two panels show the robust generalization regime with 10% label noise, and the lower two panels show the abrupt transition regime with 70% label noise. For each regime, we compare the best-validation checkpoint and the last checkpoint. Exception Retention measures whether corrupted positions still predict the noisy label; Exception Correction measures whether corrupted positions return to the clean label; Rule Retention measures whether uncorrupted positions remain ground-truth-consistent. Lower SNR indicates stronger perturbation.

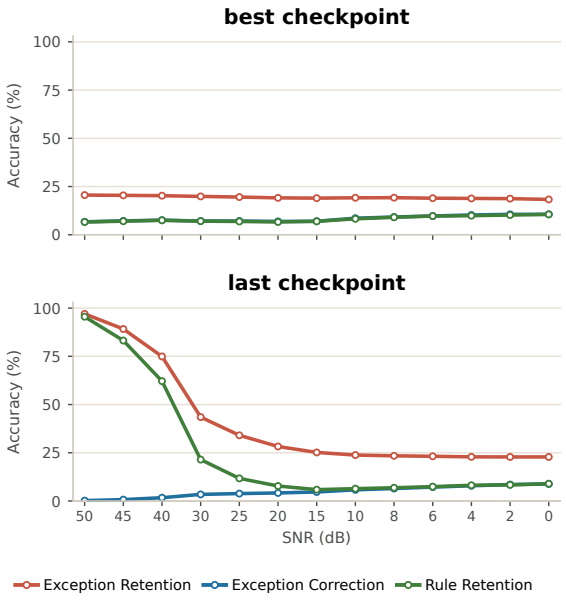


Figure 8: Acoustic perturbation probes at 100% training scale and 90% label noise. We compare the best-validation checkpoint and the last checkpoint in the high-noise collapse regime. Exception Retention measures whether corrupted positions retain the observed corrupted label; Exception Correction measures whether corrupted positions return to the corresponding rule-consistent label; Rule Retention measures whether uncorrupted positions remain rule-consistent. Lower SNR indicates stronger perturbation.