Rethinking Continual Learning for Speech and Audio: A Representation-Centric Taxonomy and Open Problems

Anonymous ACL submission

Abstract

Speech and audio systems operate in inherently non-stationary environments, yet continual learning (CL) research in this domain, especially in the foundation model era, remains fragmented that fail to account for the coupled, geometry-sensitive nature of acoustic representations. Modern speech foundation models operate over highly entangled, continuous representations that jointly encode linguistic, speaker, and paralinguistic factors within a shared latent space. CL is therefore fundamentally about preserving and evolving shared representation structure rather than retaining isolated task knowledge. In this work, we revisit CL for speech from a representation-centered perspective, and introduce a new taxonomy that organizes CL according to how underlying representation geometry evolves under non-stationary acoustic conditions. We further identify key mismatches between current CL assumptions and speech foundation model behavior, and finally outline a set of open challenges and future research directions.

1 Introduction

The real world is inherently continuous and non-stationary. Acoustic environments evolve, speakers age, languages and accents shift, and novel sound events emerge over time. However, most speech systems are still trained once on static datasets and deployed under the assumption that the underlying data distribution remains stable. This mismatch between the dynamic nature of acoustic experience and the static paradigm of model training is precisely the setting that continual learning (CL) aims to address (De Lange et al., 2021). Yet existing CL methodologies were largely developed under assumptions that align poorly with speech and audio data: discrete task boundaries, stationary input distributions, and relatively disentangled representations learned from small-scale models trained from scratch. As a result, many established CL techniques encounter fundamental limitations when applied to modern speech systems.

Speech presents challenges that differ fundamentally from vision and text due to both the nature of the acoustic signal and the structure of its learned representations. *At the signal level*, vision and text operate on relatively stable and explicit units: pixels are spatially localized and tokens are discretized symbolic units with predefined boundaries. In contrast, speech is a continuous temporal waveform whose meaningful structures, such as phonemes, words, speaker turns, and prosodic events, are not explicitly given but must be inferred by the model itself. *At the representation level*, speech embeddings simultaneously encode multiple highly entangled factors, including linguistic content, speaker identity, emotion, accent, recording conditions, and paralinguistic information. Unlike many vision or NLP settings where task-relevant attributes can often be partially separated or localized, these factors in speech are deeply coupled within a shared latent geometry. Consequently, adapting the model for one objective frequently perturbs representations required for others (Chen et al., 2025). For example, continual adaptation for emotion recognition may unintentionally distort phonetic structures important for automatic speech recognition (ASR) or speaker characteristics required for speaker verification. Continual learning in speech therefore cannot be framed simply as preserving task-specific knowledge; rather, it requires maintaining the stability of a shared and continuously evolving representational geometry across competing objectives and acoustic conditions.

These challenges become even more pronounced in the era of speech and audio foundation models. Recent years have witnessed a rapid shift from task-specific architectures toward large-scale pretrained models such as *wav2vec 2.0* (Baevski et al., 2020), *HuBERT* (Hsu et al., 2021), *Whisper* (Radford et al., 2022), and recently LALMs (Yang

et al., 2025). Trained on massive and heterogeneous audio corpora, these models learn highly generalizable latent representations that support a broad spectrum of capabilities, including ASR, speaker verification, emotion recognition, audio captioning, spoken dialogue, and multimodal reasoning. Consequently, continual learning is no longer simply a problem of sequentially acquiring new tasks. Instead, it becomes a problem of continuously adapting a shared pretrained representation while preserving the acoustic and linguistic structures that underpin diverse downstream capabilities. Under this setting, catastrophic forgetting can manifest not only as performance degradation on previously learned tasks, but also as gradual corruption of the latent representational geometry itself. A model may maintain benchmark-level accuracy while its internal structure progressively deteriorates, for example through reduced phonetic separability, compressed speaker manifolds, or degradation of paralinguistic organization. Such representational shifts are particularly problematic in speech because multiple capabilities depend on the same shared embedding space.

Existing adaptation strategies only partially address this challenge. Parameter-efficient tuning approaches, including LoRA (Hu et al., 2022), adapters (Selvaraj et al., 2023), and prompt-based adaptation (Cui et al., 2025), reduce the extent of destructive full-model updates, but they do not eliminate representational drift and may introduce new forms of interference across tasks, domains, and adaptation stages. As foundation models continue to expand in scale and capability, the central question of continual learning for speech shifts from *how to prevent forgetting previously learned tasks* to *how to continuously adapt foundation-scale speech representations while preserving the latent structures that support broad acoustic and linguistic generalization*.

We argue that the combination of speech-specific representational complexity and foundation-scale pretrained models fundamentally reshapes the continual learning problem beyond the conventional CL settings. This paper presents a representation-centered perspective on CL for speech and audio in the foundation model era. We introduce a taxonomy of speech-centric continual learning scenarios, analyze how existing CL paradigms fail under the dynamics of large-scale speech representations and adaptation, and outline future research directions centered on preserving representational geometry,

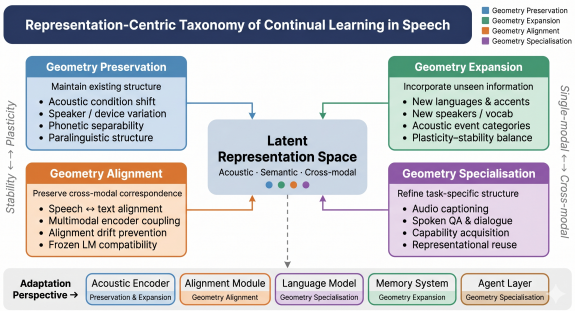

Figure 1: A representation-centric taxonomy of continual learning in speech. The bottom strip provides a complementary adaptation perspective, mapping each evolution type to the model.

adapting entangled latent spaces, and developing evaluation protocols that reflect the behavior and degradation modes of modern speech foundation models and large audio language models (LALMs).

2 A Representation-Centric Taxonomy of Continual Learning in Speech

Classical CL taxonomies (De Lange et al., 2021) typically organize learning scenarios according to shifts in tasks, domains, or label spaces, leading to well-established settings such as task-incremental, domain-incremental, and class-incremental learning. While effective in controlled benchmarks, these formulations provide only a partial view of continual learning in modern speech systems. We therefore propose a representation-centric taxonomy of continual learning in speech and audio, characterizing it by how the underlying representation evolves over time and which structural properties must be preserved during adaptation. We identify four forms of representational evolution: geometry preservation, expansion, alignment, and specialization, as shown in Figure 1.

Geometry Preservation. Geometry preservation refers to settings where the primary objective is to maintain existing representational structure under distributional shift. In real-world deployment, speech systems encounter continuously changing acoustic conditions such as new speakers, recording devices, noise profiles, and channel effects. While adaptation improves robustness, repeated updates can gradually distort previously learned latent geometry, leading to effects such as reduced phonetic separability, collapsed speaker manifolds, or weakened paralinguistic structure. The goal of geometry preservation is therefore to constrain adaptation such that existing structure remains stable

while allowing limited adjustment to new input distributions.

Geometry Expansion. Geometry expansion describes scenarios in which the representation must incorporate previously unseen information while preserving compatibility with existing structure. This includes new languages, accents, vocabularies, speakers, and acoustic events. The central challenge is balancing plasticity and stability within a shared latent space: new information must be embedded without overwriting or fragmenting previously established organization. For example, multilingual extension requires integrating new phonetic systems without degrading separability of existing languages, while speaker expansion requires adding new identity structure without collapsing previously learned distinctions.

Geometry Alignment. Geometry alignment refers to CL settings in which relationships between multiple representation spaces must be preserved or updated consistently. This is particularly important in speech foundation models that integrate acoustic encoders with language models, multimodal modules, or external memory systems. Continual adaptation may introduce alignment drift, where mappings between speech representations and textual or multimodal spaces degrade even when individual modalities remain stable. For instance, updating a speech encoder may break its correspondence with a frozen language model, leading to degraded speech-to-text performance.

Geometry Specialization. Geometry specialization captures settings where the representation is adapted to support new or refined capabilities built on a shared foundation model. These include audio captioning, spoken question answering, and agentic dialogue. Adaptation typically reshapes or reweights existing regions of the representation space to emphasize task-relevant structure, improving performance on new objectives but potentially interfering with previously learned capabilities that rely on overlapping representations. This highlights the tension between capability acquisition and representational reuse in foundation models.

Adaptation Perspective. While the taxonomy above characterizes what changes in the representation space, CL behavior is also determined by where these changes are induced within the model. In modern speech foundation models, adaptation occurs across multiple components, including *acoustic encoders, alignment modules, language models, memory systems, and agentic components*. These layers provide a complementary mechanistic view, as the same form of representational evolution may arise from different adaptation sites, leading to distinct interference patterns and forgetting behaviors.

Acoustic encoder adaptation affects low-level acoustic geometry and is most closely associated with geometry preservation and expansion, as it directly shapes phonetic structure, speaker identity, and environmental robustness. Alignment-layer adaptation governs cross-modal correspondence and is central to geometry alignment, ensuring stable mappings between speech and textual or multimodal representations. Language-model adaptation influences higher-level semantic reasoning and is often linked to geometry specialization. Memory systems support incremental knowledge accumulation and user-specific information, making them closely related to geometry expansion. In contrast, agent-level adaptation operates at the level of behavioral policy and interaction, corresponding primarily to geometry specialization through changes in planning, tool use, and decision-making.

Discussion. This taxonomy reframes CL in speech from *the interaction between evolving representations and the architectural layers through which adaptation is performed*. Representation evolution and adaptation location are only partially coupled: the same form of geometric change can be induced by updates at different layers, and updates to a single layer may simultaneously impact multiple forms of representational geometry. These forms of evolution are not mutually exclusive and often co-occur in real systems, providing a unified perspective that connects representation dynamics with system-level adaptation mechanisms in modern speech foundation models. Detail discussion of existing CL methods, please refer to the Appendix.

3 Mitigation Strategies: A Brief Overview

Existing CL approaches for speech can be categorized into three fundamental mechanisms: **replay**, **regularization**, and **architectural isolation**. While each has demonstrated effectiveness in constrained settings, all three encounter a common bottleneck in modern speech foundation models: the coupled nature of acoustic representations means that neither data rehearsal, parameter-level penalties, nor module isolation can fully prevent geo-

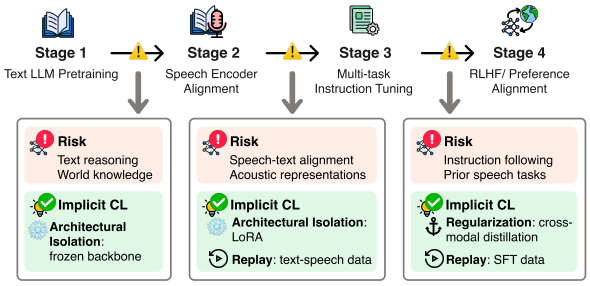

Figure 2: Decoding Speech LLM Post-Training as an Implicit Multimodal Continual Learning Pipeline. The 4-stage development process (from text-only pretraining to preference alignment).

metric distortion of the shared latent space during sequential adaptation. This representational entanglement, rather than any individual method's limitation, constitutes the primary obstacle to continual learning at foundation-model scale. Extended analysis of each strategy is provided in Appendix A.

4 LALMs Post-Training as Implicit CL

When the representation-centric perspective introduced above is applied to the LLM era, it indicates an underexplored reality: much of the current LALM engineering practice already implements implicit continual learning. The standard multi-stage post-training pipeline is, in fact, a sequence of complex, cross-modality continual learning problems. So we can map standard LALM recipes directly to their underlying CL mechanisms as illustrated in Figure 2:

Stage 1 → 2 (Speech Encoder Alignment): When transitioning from a pretrained text LLM to cross-modal alignment, the model is at risk of forgetting its existed text reasoning and world knowledge. To avoid this, researchers almost universally rely on freezing the text backbone while training a speech encoder, which is fundamentally architectural isolation (Hsiao et al., 2025).

Stage 2 → 3 (Multi-task Instruction Tuning): As the model learns to follow diverse audio instructions, the speech-text alignment and acoustic representations established in Stage 2 are at risk. To anchor the embedding space, the community typically mixes text and speech instruction data which acting as replay(Chu et al., 2024) while using LoRA and adapters to constrain the update space as architectural isolation(Xu et al., 2024).

Stage 3 → 4 (RLHF / Preference Alignment): In the final alignment phase, the model is at risk of losing its prior instruction like following behavior and legacy speech tasks (e.g., ASR, TTS). To prevent this, engineers often inject cross-modal distillation to explicitly transfer prior competence (Wang et al., 2025a) alongside data replay. Recent theoretical insights propose that the very use of on-policy Reinforcement Learning (RL) in this stage also acts as a fundamental implicit CL strategy (Shenfeld et al.). On-policy RL biases model updates toward KL-minimal solutions relative to the base policy. By minimizing this distributional shift, RL naturally mitigates catastrophic forgetting and preserves prior multimodal capabilities significantly better than standard offline supervised fine-tuning.

When reviewing these transitions, a clear trend appears: LALM post-training is the only area where hybrid methods are the norm rather than the exception. Since the model must maintain multiple abilities at once (such as text reasoning, speech-text alignment, prior speech tasks, and instruction-following behavior), no single continual learning approach is enough. The field is increasingly relying on combined methods in practice. For instance, maintaining performance often requires keeping text backbones frozen, replaying text data, and using cross-modal distillation at the same time (Hsiao et al., 2025; Cuervo et al., 2025; Wang et al., 2025b, 2026). This practical consensus has formed even without a complete theoretical explanation.

5 Open Problems and Future Directions

The transition from mitigating task-specific forgetting to preserving continuous representational geometry presents a set of unresolved challenges.

The transition from mitigating task-specific forgetting to preserving continuous representational geometry presents a set of unresolved challenges.

A primary constraint in speech is the privacy of biometric voice data, which precludes raw-audio replay at scale. A promising direction is internal generative pseudo-replay (Frascaroli et al., 2024): LALMs can self-generate modality-aligned synthetic samples directly from their latent space, bypassing raw data storage while preserving acoustic manifold geometry during continual updates.

Real-world speech systems also frequently encounter missing modalitie yet existing multimodal CL methods assume full modality availability (Huang et al., 2025). Future work should model this modality gap directly within the shared embedding space, using dynamic routing or masking mechanisms to protect cross-modal alignment from feature drift under incomplete multimodal updates.

Limitations. While this work establishes a representation-centric taxonomy and identifies key open problems for continual learning in speech and audio, the current treatment remains primarily conceptual and analytical. The four proposed geometry categories—preservation, expansion, alignment, and specialisation—are grounded in theoretical reasoning and supported by existing empirical observations from the literature; however, we do not yet provide a unified quantitative benchmark that operationalises each category with standardised metrics for representational geometry change (e.g., intrinsic dimensionality shift, inter-class separability drift, or cross-modal alignment degradation). Constructing such a benchmark requires careful curation of evaluation protocols across diverse speech foundation models and task sequences, which we identify as a concrete and high-priority direction for future work. We therefore view the present taxonomy as a principled conceptual foundation upon which rigorous empirical validation and benchmark development can be built in subsequent work.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Heng-Jui Chang, Hung-yi Lee, and Lin-Shan Lee. 2021. Towards Lifelong Learning of End-to-end ASR. *Interspeech*, pages 2551–2555.

Xinyi Chen, Xi Chen, Zhenyu Weng, and Yang Xiao. 2025. AFT: An exemplar-free class incremental learning method for environmental sound classification. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

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

Santiago Cuervo, Skyler Seto, Maureen de Seyssel, R. Bai, Zijin Gu, Tatiana Likhomanenko, N. Jaitly, and Zakaria Aldeneh. 2025. Closing the Gap Between Text and Speech Understanding in LLMs. *ArXiv*, abs/2510.13632.

Yawen Cui, Li Liu, Zitong Yu, Guanjie Huang, and Xiaopeng Hong. 2025. Few-shot audio-visual class-incremental learning with temporal prompting and regularization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(15):16118–16126.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385.

Emanuele Frascaroli, Aniello Panariello, Pietro Buzzega, Lorenzo Bonicelli, Angelo Porrello, and Simone Calderara. 2024. Clip with generative latent replay: a strong baseline for incremental learning. *arXiv preprint arXiv:2407.15793*.

Shahram Ghorbani, S. Khorram, and J. Hansen. 2019. Domain Expansion in DNN-Based Acoustic Models for Robust Speech Recognition. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 107–113.

Brady C. Houston and K. Kirchhoff. 2020. Continual Learning for Multi-Dialect Acoustic Models. *Interspeech*, pages 576–580.

Chi-Yuan Hsiao, Ke-Han Lu, Kai-Wei Chang, Chih-Kai Yang, Wei-Chih Chen, and Hung-yi Lee. 2025. Analyzing Mitigation Strategies for Catastrophic Forgetting in End-to-End Training of Spoken Language Models. *ArXiv*, abs/2505.17496.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, and 1 others. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *Preprint*, arXiv:2106.07447.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*.

Linlan Huang, Xusheng Cao, Haori Lu, Yifan Meng, Fei Yang, and Xialei Liu. 2025. Mind the gap: Preserving and compensating for the modality gap in clip-based continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3777–3786.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Chang Li, Kanglei Zhou, and Liyuan Wang. 2026. PACE: Pretrained audio continual learning. *Preprint*, arXiv:2602.03355.

Zhizhong Li and Derek Hoiem. 2016. *Learning Without Forgetting*, pages 614–629. Springer International Publishing.

Wei Liu, Jingyong Hou, Dong Yang, Muyong Cao, and Tan Lee. 2024. A Parameter-efficient Language Extension Framework for Multilingual ASR. *ArXiv*, abs/2406.06329.

Tianyi Peng and Yang Xiao. 2024. Dark Experience for Incremental Keyword Spotting. *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Nithish Muthuchamy Selvaraj, Xiaobao Guo, A. Kong, Bingquan Shen, and A. Kot. 2023. Adapter Incremental Continual Learning of Efficient Audio Spectrogram Transformers. *Interspeech*, pages 909–913.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's Razor: why on-policy reinforcement learning forgets less. In *Non-Euclidean Foundation Models: Advancing AI Beyond Euclidean Frameworks*.

Chaoren Wang, Heng Lu, Xueyao Zhang, Shujie Liu, Yan Lu, Jinyu Li, and Zhizheng Wu. 2026. Closing the Modality Reasoning Gap for Speech Large Language Models. *ArXiv*, abs/2601.05543.

Enzhi Wang, Qicheng Li, Zhiyuan Tang, and Yuhang Jia. 2025a. Cross-modal Knowledge Distillation for Speech Large Language Models. *ArXiv*, abs/2509.14930.

Guitao Wang, Jinming Zhao, Hao Yang, Guilin Qi, Tongtong Wu, and Gholamreza Haffari. 2025b. Continual speech learning with fused speech features. *Preprint*, arXiv:2506.01496.

Yang Xiao and Rohan Kumar Das. 2024a. Ucil: An Unsupervised Class Incremental Learning Approach for Sound Event Detection. *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yang Xiao and Rohan Kumar Das. 2024b. Where's That Voice Coming? Continual Learning for Sound Source Localization. *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Yang Xiao, Eun-Jung Holden, and Ting Dang. 2026a. Adapting where it matters: Depth-aware adaptation for efficient multilingual speech recognition in low-resource languages. In *ACL 2026*.

Yang Xiao, Nana Hou, and E. Chng. 2022a. Rainbow Keywords: Efficient Incremental Learning for Online Spoken Keyword Spotting. *ArXiv*, abs/2203.16361.

Yang Xiao, Xubo Liu, James King, Arshdeep Singh, Eng Siong Chng, Mark D. Plumbley, and Wenwu Wang. 2022b. Continual learning for on-device environmental sound classification. In *DCASE 2022 Workshop on Detection and Classification of Acoustic Scenes and Events*.

Yang Xiao, Aso Mahmudi, Nick Thieberger, E. Ambikairajah, Eun-Jung Holden, and Ting Dang. 2026b. Continual Adaptation for Pacific Indigenous Speech Recognition.

Yang Xiao, Tianyi Peng, Rohan Kumar Das, Yucheng Hu, and Huiping Zhuang. 2025. Analytickws: Towards Exemplar-Free Analytic Class Incremental Learning for Small-footprint Keyword Spotting. *Annual Meeting of the Association for Computational Linguistics*, pages 14147–14158.

Tianyi Xu, Kaixun Huang, Pengcheng Guo, Yu Zhou, Longtao Huang, Hui Xue, and Lei Xie. 2024. Towards Rehearsal-Free Multilingual ASR: A LoRA-based Case Study on Whisper. *ArXiv*, abs/2408.10680.

Chih-Kai Yang, Neo S Ho, and Hung-yi Lee. 2025. Towards holistic evaluation of large audio-language models: A comprehensive survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10155–10181.

Mu Yang, Shaojin Ding, Tianlong Chen, Tong Wang, and Zhangyang Wang. 2021. Towards Lifelong Learning of Multilingual Text-to-Speech Synthesis. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8022–8026.

Kwok Chin Yuen, J. Yip, and Chng Eng Siong. 2025. Continual Learning with Embedding Layer Surgery and Task-wise Beam Search using Whisper. *ArXiv*, abs/2501.07875.

A The Evolution of Mitigation Strategies

This section examines how the existing methods[1] have evolved and why the coupled nature of modern speech representations presents a primary bottleneck for continuous adaptation. We categorize existing approaches into three fundamental mechanisms: replay, regularization, and architectural isolation.

A.1 Replay-Based Methods.

Replay strategies mitigate forgetting by revisiting previous data distributions. In early stages of model development, replaying raw waveforms or acoustic features was standard practice. Interleaving old audio samples with new data anchored the model during the acquisition of new tasks, such as automatic speech recognition (Chang et al., 2021), multilingual speech synthesis (Yang et al., 2021), and audio classification (Xiao et al., 2022b,a; Peng and Xiao, 2024; Xiao and Das, 2024a; Xiao et al., 2025). This approach is effective because directly replaying raw waveforms preserves the full geometric space. However, as models scale, speech data encounters strict privacy constraints and storage limits.

A.2 Regularization-Based Methods.

Regularization methods impose soft constraints on model updates to preserve previously learned knowledge without requiring access to old data. Traditional methods, such as Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) and Learning without Forgetting (LwF) (Li and Hoiem, 2016), were widely applied to early ASR models (Ghorbani et al., 2019; Houston and Kirchhoff, 2020). By computing Fisher information matrices, these approaches identified and penalized changes to weights essential for previous tasks. These approaches can be viewed as indirect mechanisms for geometry preservation, enforcing stability at the parameter level rather than directly constraining the latent representation space. However, this assumption becomes increasingly fragile in modern speech foundation models, where representations are highly entangled across phonetic, speaker, and acoustic factors. In such settings, parameter-level constraints are insufficient to prevent representational drift: acoustic conditions and phonetic structure are jointly encoded, so even small updates can induce global distortions in embedding geometry.

[1] Full references are in our GitHub list.

A.3 Architectural Isolation Methods.

Architectural isolation tackles shared module interference by freezing the main pretrained backbone and updating only task-specific, lightweight parameters. Previously research utilize adapter to constrain the trainable parameters (Xiao and Das, 2024b). With the introduction of foundation models like wav2vec 2.0 (Baevski et al., 2020) and Whisper (Radford et al., 2022), PEFT has become widely adopted. For example, integrating LoRA (Hu et al., 2022) modules into Whisper allows the model to learn new languages sequentially without losing original capabilities (Xu et al., 2024; Liu et al., 2024; Yuen et al., 2025; Xiao et al., 2026b,a). PEFT-based methods can be interpreted as mechanisms for geometry preservation through parameter isolation. However, this assumption breaks down in speech because representational entanglement is not aligned with parameter modularity. Unlike NLP backbones, audio encoders encode strong low-level spectral structure, where speaker identity, phonetics, and paralinguistic cues are jointly embedded within the same continuous acoustic manifold (Li et al., 2026). As a result, isolating updates to bottleneck modules does not isolate their effect on the representation geometry.