Dynamic Model-bank Test-time Adaptation for Automatic Speech Recognition

Yanshuo Wang[1,2], Yanghao Zhou[3], Yukang Lin[4], Haoxing Chen[5], Jin Zhang[6], Wentao Zhu[2], Jie Hong[*7], Xuesong Li[*8,9],
[1]Hong Kong Polytechnic University, [2]Eastern Institute for Advanced Study
[3]Beijing Institute of Technology, [4]Tsinghua University, [5]Ant Group, [6]Shenzhen University
[7]The University of Hong Kong, [8]Australian National University, [9]CSIRO

Abstract

End-to-end automatic speech recognition (ASR) based on deep learning has achieved impressive progress in recent years. However, the performance of ASR foundation model often degrades significantly on out-of-domain data due to real-world domain shifts. Test-Time Adaptation (TTA) methods aim to mitigate this issue by adapting models during inference without access to source data. Despite recent progress, existing ASR TTA methods often struggle with instability under continual and long-term distribution shifts. To alleviate the risk of performance collapse due to error accumulation, we propose Dynamic Model-bank Single-Utterance Test-time Adaptation (DMSUTA), a sustainable continual TTA framework based on adaptive ASR model ensembling. DMSUTA maintains a dynamic model bank, from which a subset of checkpoints is selected for each test sample based on confidence and uncertainty criteria. To preserve both model plasticity and long-term stability, DMSUTA actively manages the bank by filtering out potentially collapsed models. This design allows DMSUTA to continually adapt to evolving domain shifts in ASR test-time scenarios. Experiments on diverse, continuously shifting ASR TTA benchmarks show that DMSUTA consistently outperforms existing continual TTA baselines, demonstrating superior robustness to domain shifts in ASR.

1 Introduction

In recent years, end-to-end automatic speech recognition has made substantial progress, achieving strong performance in in-domain settings where training and test data follow similar distributions (Baevski et al., 2020; Hsu et al., 2021; Radford et al., 2023). However, in real-world scenarios, ASR systems frequently encounter out-of-domain conditions, where data distributions shift due to factors such as background noise, varying recording devices, or unseen acoustic environments (Radford et al., 2023; Chen et al., 2022a). These domain shifts often lead to significant performance degradation, rendering ASR systems unreliable in practice. This challenge becomes even more severe when distribution shifts occur continuously over time, as is common in real-world deployment.

To address this issue, it is essential to develop ASR systems that can dynamically adapt to changing input conditions and maintain robustness across diverse environments. Recent work has explored test-time adaptation to mitigate domain shifts during inference, primarily through objectives such as entropy minimization (Lin et al., 2022) and logit sequence entropy regularization (Kim et al., 2023). These methods have shown effectiveness on out-of-domain noisy speech data. However, they typically assume a fixed target domain, limiting their applicability in scenarios involving continuous and dynamic distribution shifts.

To handle such realistic conditions, recent studies like AWMC (Lee et al., 2023) and DSUTA (Lin et al., 2024) have proposed continual test-time adaptation (CTTA) methods for ASR. AWMC adopts a pseudo-labeling strategy within a mean teacher framework to mitigate model collapse, while DSUTA introduces a fast-slow adaptation scheme with full model resets to maintain stability under domain drift. Although both methods represent significant progress, they also face key limitations: AWMC is evaluated only on a single-domain setup and is susceptible to error accumulation, whereas DSUTA's reset mechanism impairs long-term knowledge retention, potentially causing forgetting during adaptation.

Moreover, both approaches process each test sample independently using a single model state, which restricts their ability to exploit temporal cues from evolving input streams. To address these limitations, we propose a model-bank-based CTTA framework named Dynamic Model-bank Single-

*Corresponding author

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 21843–21853
November 4-9, 2025

Utterance Test-time Adaptation (DMSUTA) that dynamically selects reliable model checkpoints from a diverse model bank, enabling robust and adaptive predictions for each incoming audio sample. We evaluate DMSUTA on ASR CTTA benchmarks (Lin et al., 2024), encompassing a wide range of acoustic variations that simulate realistic, continuously shifting environments. Experimental results demonstrate that DMSUTA consistently outperforms the prior work, highlighting its superior ability to robustly adapt to non-stationary test distributions in ASR.

Our main contributions can be summarized as:

- We propose a dynamic model-bank TTA framework for ASR that exploits relevant data distribution information while minimizing the risk of model collapse in adaptation.
- We introduce dual-criterion with active bank maintenance to stabilize the content of the model bank across different domains and long test data streams.
- We achieve noticeable performance boosts relative to both non-continual and continual baselines in scenarios involving single-domain and time-varying domains.

2 Related Works

2.1 Continual Test Time Adaptation

CTTA requires continuous model updates for target domain data streams, with the core challenge being how to avoid catastrophic forgetting during the adaptation process. The online version of Tent (Wang et al., 2020) offers a feasible approach by minimizing prediction entropy, but it assumes ideal online learning conditions and lacks stability in continual test-time settings. To address this, CoTTA(Wang et al., 2022a) was proposed as a method specifically designed for online continual test-time adaptation. It incorporates a teacher-student framework with a weighted augmentation-averaged mean teacher strategy to enhance model stability. Meanwhile, EATA(Niu et al., 2022) addresses catastrophic forgetting through a sample selection strategy that filters reliable and non-redundant samples.

Building on these foundations, several recent works have further advanced CTTA by improving stability, efficiency, and sample selection mechanisms. ViDA (Liu et al.) introduces a homeostatic adaptation strategy that constrains feature space dynamics to enhance robustness under domain shifts. DSS (Wang et al., 2024c) proposes a confidence-based filtering mechanism to reduce error propagation during continual updates. EcoTTA (Song et al., 2023) improves memory efficiency through self-distilled regularization, maintaining model plasticity without relying on large memory buffers. Additionally, RMT (Döbler et al., 2023) extends consistency-based learning to gradual and continual adaptation scenarios by employing a teacher-student paradigm to stabilize updates. While these approaches offer various advantages, they often require careful trade-offs between domain adaptation and model forgetting or introduce additional training complexity. In contrast, our method employs a dynamic framework that leverages multiple dynamically updated checkpoints, selected based on confidence and uncertainty metrics, to achieve robust adaptation without relying on frequent resets.

2.2 Test Time Adaptation on ASR

Recent advancements in speech foundation models have led to significant improvements in ASR, particularly within the distribution of in-domain training data (Hsu et al., 2021; Chen et al., 2022b; Baevski et al., 2020; Döbler et al., 2023; Ao et al., 2022). However, during edge-side deployment, models frequently encounter diverse and non-stationary acoustic conditions, resulting in degraded performance in real-world applications. Although joint training (Zhang et al., 2021; Wang and Wang, 2016; Fan et al., 2020; Jain et al., 2018), fine-tuning (Wang et al., 2022b; Gong et al., 2023; Wang et al., 2024a), and domain adaptation (Lei et al., 2024; Sim et al., 2024; Fu et al., 2025; Tran et al., 2025; Sun et al., 2017; Wang et al., 2024b; Xiong et al., 2020; Ghorbani and Hansen, 2022) methods have shown strong performance in adapting ASR models to target domains, they are often computationally expensive and typically require access to test label or source domain data. In practice, ASR models deployed on edge devices are often restricted to streaming test data due to privacy and storage constraints, limiting the applicability of these methods.

In this context, TTA has emerged as a promising direction, as it enables ASR models to adapt to target domains directly during inference and without requiring any training data and test labels. Recent studies such as SUTA (Lin et al., 2022) have applied entropy minimization and minimal class con-

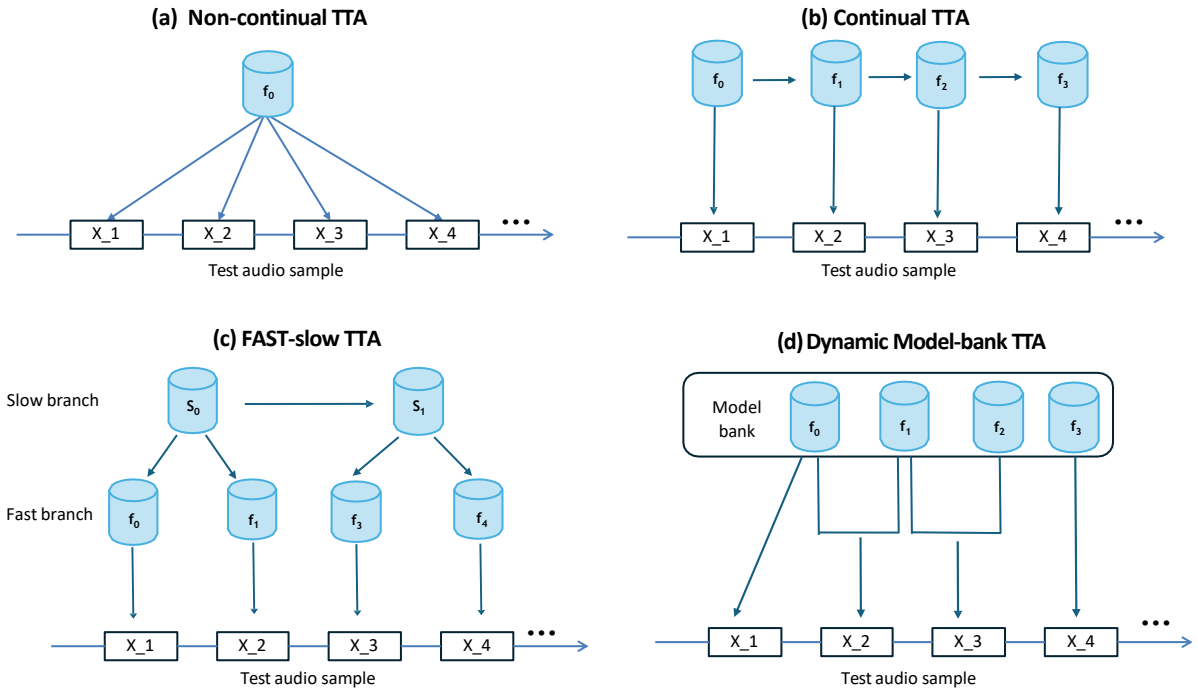

Figure 1: Method comparison between our dynamic TTA framework, DMSUTA, and existing methods. **(a)** The common ASR test-time adaptation approach is where the model adapts the incoming sample independently. **(b)** The continual adaptation of test time, where the model is continuously updated in a sequence as the test sample comes. **(c)** The fast-slow approach, where there are two separate branches responsible for local and meta adaptation. **(d)** Our method, DMSUTA, provides a model bank to perform adaptation. For each test sample, the prediction is an output from a carefully selected bank of members.

fusion techniques to achieve non-autoregressive instance-level adaptation. SGEM (Kim et al., 2023), in contrast, proposes generalized entropy minimization at the sequence level, adapting autoregressive ASR models by minimizing logit sequence entropy. Recent methods also try to incorporate the use of confidence and uncertainty for stable adaptation in ASR (Yoon et al., 2024; Lee et al., 2024; Liu et al., 2024).

However, these TTA methods largely assume static test distributions and do not account for the continuous adaptation required in real-world streaming scenarios. Addressing this gap, AWMC (Lee et al., 2023) introduces pseudo-label refinement and anchor model merging to improve model reliability under continuous adaptation. DSUTA (Lin et al., 2024) proposes a dynamic reset strategy based on a fast-slow TTA framework, enabling stable CTTA across multiple domains and long-duration test streams.

Despite these advances, existing ASR continual test-time adaptation methods still face challenges such as error accumulation and catastrophic forgetting, especially in complex, noisy environments. To tackle this, our work proposes a dynamic model ensembling strategy tailored to evolving acoustic conditions, aiming to enhance the robustness of ASR models in real-world, out-of-domain scenarios.

3 Methodology

The proposed dynamic model-bank TTA framework for ASR, DMSUTA, is proposed in this section. We first compare DMSUTA with existing methods, then we explain the details of the selection strategy and bank operations in DMSUTA.

3.1 TTA Methods for ASR

As shown in Figure 1, the most straightforward approach for adaptation in ASR is non-continual TTA, where the model adapts to test data separately each time. In this case, the original model is reused for every new test sample. However, this limits the generalizability of the adaptation process, as knowledge learned from past samples is not carried forward. On the other hand, continual adaptation does not reset to the pre-trained model each time; instead, it continuously adapts throughout the pro-

cess. However, this increases the risk of model collapse due to error accumulation over time. Recently, Fast-Slow TTA (Lin et al., 2024) has been proposed, utilizing two branches to leverage advantages from both non-continual and continual TTA. The slow branch is updated periodically to gain the meta knowledge. In contrast, the fast branch is initialized from the current slow branch model and updated based on the latest samples to generate predictions. However, for the slow branch to update, test samples from previous intervals must be stored and processed in batches. This imposes strict constraints on test-time adaptation, requiring access to past test samples. In addition, it also employs a dynamic reset strategy to mitigate model collapse during adaptation, which also relies on collecting many past test samples. Those additional data usage may not be ideal for real-world test-time adaptation, as access to previous samples is often restricted due to security concerns.

In this paper, we propose a CTTA strategy in ASR, named DMSUTA, which is built upon an adaptive framework that maintains a dynamic bank of model checkpoints to generate robust predictions tailored to continuously evolving input conditions. By leveraging multiple reliable historical checkpoints from this bank, DMSUTA dynamically selects and updates the most relevant ones for each test instance, while preserving the rest to retain prior knowledge. To manage the checkpoint bank effectively, we design a three-stage mechanism: selection, appending, and pruning. The selection stage identifies relevant checkpoints based on a combination of prediction confidence and uncertainty estimation, where uncertainty is measured by the variance of the output probabilities across multiple augmented views of the same input. The selected subset is then optimized on the current test sample, and when performance improves, the updated model may be added to the bank. To prevent redundancy and model collapse, we remove outdated or degraded checkpoints using historical performance trends and divergence scores. Our motivation is to exploit useful information from the model's historical states to mitigate error accumulation and reduce catastrophic forgetting, both of which are persistent challenges in continual test-time adaptation.

3.2 Dynamic Model-bank SUTA for ASR

As shown in Figure 2, we present the overall framework of our unified test-time adaptation (TTA) framework for speech recognition. Unlike prior TTA approaches for speech recognition that typically rely on a single model to adapt to all incoming samples, our framework continuously maintains a bank of diverse teacher models for prediction. The process begins by initializing a model bank $\{f_1, f_2, \ldots, f_{\text{m}}\}$ with size m, where each member is initialized from the original pretrained checkpoint. Upon the arrival of a test-time sample x, it is forwarded to the most relevant models in the bank that meet the uncertainty and confidence criteria for inference. The final prediction is then generated by the overall mean of the selected members' predictions. After making the prediction, the selected models are updated to better adapt to the current data distribution. Noted that our method also builds upon the SUTA (Lin et al., 2022) framework. It adapts the model parameters f_{pre} for fixed number of steps on each test sample x_t by optimizing the combined loss of entropy minimization $\mathcal{L}_{\text{em}}$ and minimum class confusion $\mathcal{L}_{\text{cc}}$. Entropy minimization loss encourages confident predictions by reducing uncertainty, while the other reduces overlap between different class predictions.

Moreover, to enable the model bank to better accumulate meta-knowledge over time, we also employ an exponential moving average (EMA) strategy during bank updates:

$$\mathbf{W}_{t+1} = (1-\alpha)\hat{\mathbf{W}}_t + \alpha \mathbf{W}_t \tag{1}$$

where $\mathbf{W}_t$ denotes the updated model weights, and $\hat{\mathbf{W}}_t$ denotes the current update value. In general, the overall update mechanism allows multiple models to adapt simultaneously when appropriate, enabling the model bank to accumulate more diverse and reliable knowledge over time. As a result, the framework is more robust to distribution shifts and better suited to diverse test-time conditions. The algorithm is described in Algorithm 1.

3.3 Selection Strategy

To determine which checkpoints are suitable for given test-time samples, we design a selection strategy that jointly considers prediction confidence and uncertainty. This dual-criteria approach helps identify reliable models for inference and update, while effectively filtering out unstable and overfit ones. High confidence typically indicates that the model is likely to produce a correct prediction. Therefore, we select only models from the bank with high confidence values. The selection process is formalized

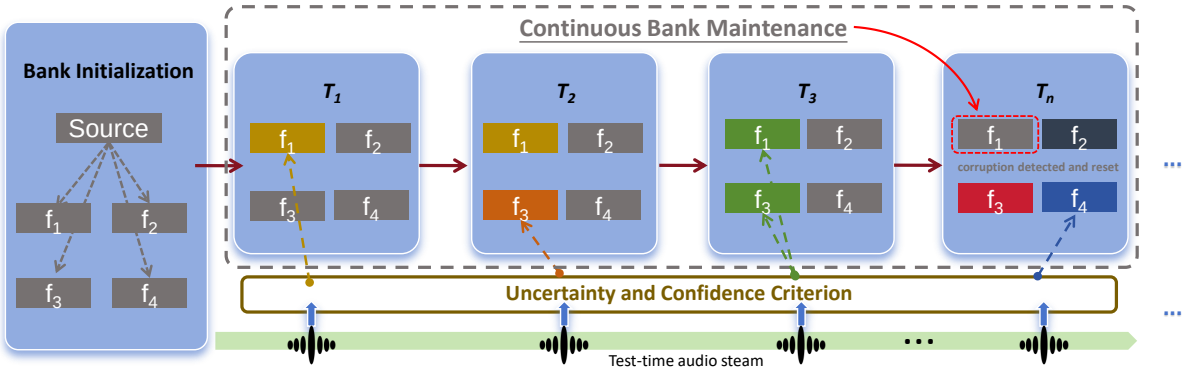

Figure 2: Overall framework for our method. We begin by initializing a model bank composed of pretrained source checkpoints. Since all models are initially unadapted, we avoid activating all of them at the beginning. Instead, we only activate one of them for update, and when no existing checkpoint is found suitable for the coming test samples, a new checkpoint is then activated and adapted. During the adaptation process, if suitable updates are available for activated checkpoints, they are applied to the corresponding ones. We also continuously monitor the status of each model in the bank. If a model remains inactive or exceeds its update threshold, it is pruned and replaced with a fresh source checkpoint to prevent potential collapse.

as follows:

$$\mathcal{S}_c(x) = \{f_i \mid \text{Conf}_i(x) \geq \gamma_c\} \tag{2}$$

where $\text{Conf}_i(x)$ is the confidence score of model f_i on input sample x, and γ_c represents the confidence threshold score among all models. However, in practice, models can sometimes become overconfident, especially when facing out-of-distribution samples, leading to unreliable predictions despite high confidence scores. To mitigate this issue, we introduce an additional criterion based on uncertainty, which is measured by the variance of predictions across multiple augmentations of the input. We apply n random augmentations to the audio sample x, and compute the variance of the model's predictions across these augmented versions. If the prediction variance is relatively low, the model is considered reliable. On the other hand, the model may be uncertain and unstable if the variance is high, and thus, we exclude such models from contributing to the final prediction. The selection criterion for uncertainty is defined as:

$$\mathcal{S}_v(x) = \{f_i \mid \text{Var}_i(x) \leq \tau_v\} \tag{3}$$

where $\text{Var}_i(x)$ is the variance of the model predictions in the n augmented versions of the input x, and τ_v represents the lowest variance allowed threshold for all models. Finally, the overall selection criterion is defined as follows:

$$\mathcal{S}(x) = \mathcal{S}_c(x) \cap \mathcal{S}_v(x) \tag{4}$$

where $\mathcal{S}(x)$ represents the full selection criterion, formed by combining the confidence-based and variance-based components through intersection, which allows only sufficiently certain predictions to be selected.

3.4 Bank Maintenance

It has been observed that the TTA methods are prone to collapse, a phenomenon in which the model gradually diverges from its original knowledge due to accumulated errors during self-supervised updates. This typically occurs when the model repeatedly adapts to unlabeled samples via self-learning, leading to degraded performance or even complete failure. Previous work (Lin et al., 2024) addresses this by completely resetting the adapted model once a distribution shift is detected. However, such full resets result in a complete loss of the model's learned knowledge.

To mitigate this issue, we introduce a novel dropout mechanism for checkpoint management within a model bank. Specifically, we design a population-based reset strategy that monitors each model using two criteria: the number of updates it has undergone and how long it has been inactive. If a model exceeds a threshold for update count or becomes inactive over time, we consider it unreliable and reinitialize it from the original pretrained checkpoint.

$$\mathcal{F}_{\text{reset}} = \{f_i \mid C(f_i)) > \epsilon_c \vee A(f_i)) < \epsilon_a\} \tag{5}$$

where $C(f_i)$ denotes the count number of updates applied to model f_i, and $A(f_i)$ indicates its recent activity level (how long it has not been updated).

Method	AA	AC	BA	CM	GS	MU	NB	SD	TP	VC
Source model	40.6	27.7	66.9	49.7	75.6	51.4	120.1	19.4	25.8	49.7
Non-continual										
SGEM (Kim et al., 2023)	30.9	17.8	54.5	39.2	56.3	39.2	113.0	14.9	17.5	40.3
SUTA (Lin et al., 2022)	30.6	17.4	53.7	38.7	54.5	39.0	112.3	15.0	17.4	39.3
Continual										
SUTA (Lin et al., 2022)	39.8	22.6	63.4	53.4	58.4	54.7	68.1	23.2	23.0	50.9
AWMC (Lee et al., 2023)	31.6	18.0	61.6	37.7	48.5	36.2	131.9	17.0	18.0	36.1
Fast-slow										
DSUTA (Lin et al., 2024)	25.9	15.4	**33.2**	33.5	37.0	28.4	36.3	15.5	15.6	29.9
Dynamic Model-bank										
DMSUTA (Ours)	**25.2**	**14.9**	34.1	**32.5**	**35.8**	**28.3**	**35.8**	**14.6**	**15.3**	**29.8**

Table 1: Word Error Rate (WER) of different ASR TTA methods on LS-C with 10 distinct domains. Reported WER is averaged per noise type.

The thresholds ϵ_c and ϵ_a control the maximum tolerated update count and minimum acceptable activity, respectively. As a result, our model bank remains robust and healthy over time. In addition, there is also a possibility that no checkpoints satisfy the selection criterion. In such cases, we replace the most updated models with newly initialized ones to maintain the diversity and adaptability of the model bank. By periodically refreshing these models, we can better handle distribution shifts and improve overall performance during test-time adaptation.

4 Experiments

4.1 Benchmark

In this work, we follow the test-time adaptation setup proposed in (Lin et al., 2024) and evaluate our method on two types of benchmarks based on the LibriSpeech test set (Panayotov et al., 2015).

Single-domain Simulated Noisy Data.

Following (Kim et al., 2023), we utilize the Corrupted LibriSpeech (LS-C) Dataset, which is created by adding background noises from the MS-SNSD dataset (Reddy et al., 2019) to the original LibriSpeech test set. Specifically, background noises are added to the test set to simulate challenging acoustic environments. The added noise types include air conditioner (AC), airport announcement (AA), babble (BA), copy machine (CM), munching (MU), neighbors (NB), shutting door (SD), typing (TP), vacuum cleaner (VC), and Gaussian noise (GS) noise—resulting in a total of ten different noise types. The Signal-to–Noise Ratio (SNR) is set to 5 dB to ensure a consistent level of corruption across samples.

Multi-domain Time-varying Data. Following (Lin et al., 2024), we also utilize three streams created by concatenating samples with different noise types to evaluate our method.

- **MD-Easy:** Contains 2,500 samples with relatively easy noise types. The noise order is: AC $\rightarrow$ CM $\rightarrow$ TP $\rightarrow$ AA $\rightarrow$ SD.
- **MD-Hard:** Contains 2,500 samples with relatively challenging noise types. The noise order is: GS $\rightarrow$ MU $\rightarrow$ VC $\rightarrow$ BA $\rightarrow$ NB.
- **MD-Long:** A long sequence of 10,000 samples generated by randomly sampling from the 10 available noise types. Each sampled segment has a random length between 20 and 500 samples, and sampling continues until the full length is reached.

Algorithm 1 Dynamic Model-bank Test-Time Adaptation

Input: Data stream $\{x_t\}_{t=1}^{T}$, model bank $\mathcal{F} = \{f_1, \ldots, f_m\}$, thresholds γ_c, τ_v, update limit ϵ_c, inactivity limit ϵ_a
Output: Predictions $\{\hat{y}_t\}_{t=1}^{T}$

Initialize each $f_i \in \mathcal{F}$ with pretrained weights
for $t = 1$ to T **do**
  compute confidence $\text{Conf}_i(x_t)$;
  compute variance $\text{Var}_i(x_t)$ for each f_i;
  check selection criterion in Equation (3);
  for $f_i \in \mathcal{S}(x_t)$ **do**
    adapt parameters w.r.t. $\mathcal{L}_{\text{em}} + \mathcal{L}_{\text{cc}}$;
    compute checkpoint prediction $\hat{y}_{t_i}^{T}$;
    EMA update in Equation (1);
  end for
  compute ensemble prediction $\{\hat{y}_t\}_{t=1}^{T}$
  for $f_i \in \mathcal{F}$ **do**
    check maintenance in Equation (5);
    for $f_i \in \mathcal{F}_{\text{reset}}$ **do**
      apply reset operation;
    end for
  end for
end for
return $\{\hat{y}_t\}_{t=1}^{T}$

Method	Easy	Hard	Long
Source model	32.7	74.6	61.0
Non-continual			
SUTA	24.0	60.4	53.3
SGEM	25.0	61.0	53.4
Continual			
SUTA	37.3	83.6	100.3
AWMC	25.8	66.1	60.6
Fast-slow			
DSUTA	24.0	45.6	43.2
- Dynamic reset	22.7	39.8	35.8
- Fixed reset	22.8	49.4	45.2
Dynamic Model-bank			
DMSUTA (Ours)	**22.0**	**39.2**	**35.0**

Table 2: WER comparison across MD-Easy, MD-Hard, and MD-Long benchmarks for different baselines.

Method	Easy	Hard	Long
w/o conf	22.4	39.6	36.0
w/o var	22.3	41.4	40.5
w both	**22.0**	**39.2**	**35.0**

Table 3: Performance comparison (WER) on multi-domain benchmarks with different update criteria.

4.2 Implementations

For all benchmarks, we use the pretrained wav2vec 2.0 Base model (Baevski et al., 2020), fine-tuned on the 960 hours LibriSpeech training set, as the source ASR model, consistent with prior work (Lin et al., 2024). We adopt SUTA (Lin et al., 2022) as the adaptation base and follow the continual adaptation framework proposed in DSUTA (Lin et al., 2024). All baseline implementations and hyperparameter settings follow the continual ASR TTA benchmark (Lin et al., 2024). We compare our method with several baselines, including continual, non-continual, and fast-slow adaptation approaches. Specifically, we evaluate against AWMC (Lee et al., 2023), SGEM (Kim et al., 2023), SUTA (Lin et al., 2022), and DSUTA (Lin et al., 2024). To support the selection mechanism based on prediction variance, we apply data augmentation to each test utterance. The augmentations include volume perturbation, light time-stretching, time shift, pitch shift, and waveform distortion. These augmentations help estimate the variance uncertainty of model predictions, which is used to guide reliable checkpoint selection during adaptation. The exponential moving average is set to 0.94. The variance and confidence thresholds are adaptively set to 6 and 2 multiples of the source model confidence and variance.

4.3 Results

Table 1 shows a comparison of word error rates (WER) across various test-time adaptation (TTA) methods on the LS-C dataset, which encompasses 10 distinct noise types. The results highlight significant variability in model robustness across different audio domains. Specifically, continual learning approaches such as SUTA and AWMC exhibit performance degradation over time, likely due to the accumulation of errors during adaptation. In contrast, non-continual methods like the original SUTA minimize the effect of model collapse but demonstrate limited adaptability in challenging domains since there is no leverage learned from past samples. For instance, in the NB domain, the non-continual SUTA shows a notably high error rate compared to its continual version.

The fast-slow method, DSUTA, integrates the

Method	Easy	Hard	Long
w/o PM	22.3	42.5	44.5
with PM	**22.0**	**39.2**	**35.0**

Table 4: Performance comparison (WER) on multi-domain benchmarks with and without bank maintenance.

strengths of both continual and non-continual approaches by employing dual adaptation branches and a dynamic reset strategy. This design enhances robustness against domain shifts. However, its dynamic strategy depends on our proposed method, further advancing this by leveraging a diverse model bank, resulting in superior performance. Notably, in the NB domain, where previous methods struggled with high error rates, our approach achieves a 1% absolute improvement over SUTA and matches DSUTA's performance. Even in domains like SD, characterized by lower WERs, our method consistently outperforms both the original SUTA and DSUTA, underscoring the efficacy of a diversified model in enhancing adaptability and resilience across varying audio domains. Moreover, Table 2 further shows that our method achieve a more stable adaption in a series of more complicated cases where domains are mixed .

4.4 Ablation

4.4.1 Selection Strategy

Here, we investigate why both confidence and variance are adopted as dual criteria for guiding model updates. As shown in Table 3, removing either the confidence or the variance threshold results in certain performance degradation. Without the confidence threshold, the selected checkpoints may yield low-confidence predictions for the test sample, which are less suitable and potentially noisy for adaptation.

Meanwhile, although confidence-based selection helps reduce pseudo-label error rates, the poor calibration of neural networks can still produce incorrect predictions with deceptively high confidence. By introducing a variance threshold, we further filter out unreliable samples, ensuring that only stable and trustworthy predictions contribute to the adaptation process.

4.4.2 Selection threshold

The introduced hyperparameters, confidence and variance, are set based on the original confidence and variance levels. Instead of using fixed values, we take the source confidence and variance of each test utterance as reference points to avoid relying on static thresholds, since the model may face vastly different adaptation difficulties across domains. The original confidence and variance effectively reflect this variation, and we leverage them to determine the thresholds dynamically. As shown in the table 5 and table 6, both the confidence and variance thresholds are not highly sensitive.

Confidence threshold	AA
$\gamma_c > 1.0$	25.35
$\gamma_c > 1.05$	25.32
$\gamma_c > 1.1$	25.25
$\gamma_c > 1.15$	25.31
$\gamma_c > 1.2$	25.22
$\gamma_c > 1.25$	25.33
$\gamma_c > 1.3$	25.38

Table 5: Performance under different confidence thresholds in AA domain

Variance threshold	AA
$\tau_v < 2.0$	25.41
$\tau_v < 3.0$	25.35
$\tau_v < 4.0$	25.32
$\tau_v < 5.0$	25.24
$\tau_v < 6.0$	25.22
$\tau_v < 7.0$	25.66

Table 6: Performance under different variance thresholds in AA domain

4.4.3 Model Bank Size

We also investigate the impact of the model bank size used in our method, as shown in Figure 3. The experiments are conducted on the LS-C dataset within the AA domain, where the bank size ranges from 2 to 10. We observe that the error rate slightly decreases as the bank size increases. When the bank size reaches 6, the model achieves the lowest word error rate. This trend is expected—larger bank sizes generally lead to lower error rates. However, increasing the bank size beyond 6 does not yield further improvements. This is likely because the AA domain represents a relatively simple and homogeneous acoustic environment. In such cases, the diversity of test samples is limited, and a small number of well-adapted models is sufficient to

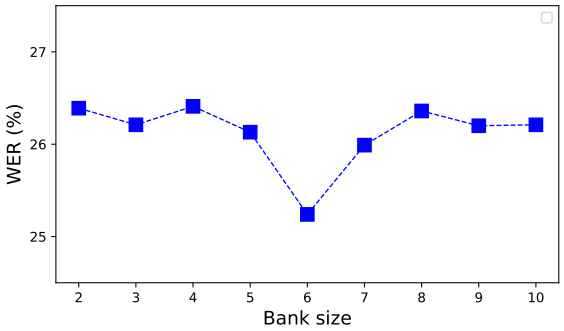

Figure 3: WER (%) performance within the AA domain of the LS-C dataset, with respect to varying bank size.

Method	AA (LS-C)		ASR-SCEChilSC	
	Time (s)	WER (%)	Time (s)	WER (%)
Source	115	40.6	76	60.1
SUTA	803	30.9	582	53.3
CSUTA	142	39.8	135	197.5
DSUTA	610	25.9	502	49.3
DMSUTA	735	**25.2**	560	**48.4**

Table 7: Comparison of adaptation time (T) and WER (W) on AA (LS-C) and ASR-SCEChilSC benchmarks.

cover the variation within the domain. Adding more checkpoints may lead to redundancy rather than complementary benefits.

4.4.4 Bank Maintenance

Here, we illustrate the effectiveness of bank maintenance. As shown in Table 4, under multi-domain conditions where the difficulty level is either hard or long, there is a noticeable performance drop when bank maintenance is not applied. Even in cases such as MD-easy, where model collapse is less severe, applying the maintenance module still leads to slight improvements.

4.4.5 Complexity

Table 7 presents the inference time of DMSUTA compared to other SUTA methods on the ASR-SCEChilSC dataset (2266 utterances) and the AA domain of corrupted LibriSpeech (2893 utterances). From the table, we observe that our DMSUTA is actually faster than the baseline SUTA and only slightly slower than DSUTA. Specifically, SUTA requires reloading the source model for predictions at every step, as it is a type of non-continual TTA, leading to significantly longer loading times. In contrast, our method caches the model after the initial load and reloads it only when necessary, significantly reducing loading time and avoiding unnecessary computational overhead. The clear margin of improvement in accuracy achieved by DMSUTA provides justification for the slightly increased complexity.

5 Conclusion

In this paper, we propose the adaptive dynamic TTA, a novel test-time adaptation strategy for ASR that improves adaptability without relying on past test samples. Our method introduces a dynamic model bank, which adaptively selects and updates a subset of reliable checkpoints for each incoming sample. By leveraging the confidence and variance of predictions, our method identifies suitable models for adaptation while filtering out unreliable cases. DMSUTA also supports continuous bank maintenance, enabling the bank to evolve over time to handle distribution shifts without incurring model collapse. Extensive experiments across multiple benchmarks validate the effectiveness and robustness of our approach over prior non-continual, continual, and fast-slow TTA methods.

6 Limitation

The primary limitations of this paper are as follows:
Realistic Domain Shifts. In this study, we evaluate our framework using synthetic domain shifts created by adding background noise to clean speech. While this setup enables controlled experimentation and reproducibility, it does not fully reflect the complexity and unpredictability of real-world acoustic environments. Realistic domain shifts—such as those caused by spontaneous speech, overlapping speakers, real-world background noise, reverberation, and recording artifacts—remain unexplored in our current experiments. We leave the evaluation of our method on real noisy datasets and in-the-wild test conditions as an important direction for future research.
Model Generalization across Backbones. Our current study focuses on a specific end-to-end ASR backbone: Wav2vec 2.0 (Baevski et al., 2020). However, modern ASR systems are increasingly built on diverse foundation models, including encoder-decoder architectures, transducer-based models, and large-scale pretrained speech-language models. The generalizability of our method across different ASR architectures remains unexplored. Extending and validating the proposed framework on a broader range of speech foundation models is a promising direction for future work.

References

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. 2022. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738.

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.

Chen Chen, Nana Hou, Yuchen Hu, Shashank Shirol, and Eng Siong Chng. 2022a. Noise-robust speech recognition with 10 minutes unparalleled in-domain data. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4298–4302. IEEE.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022b. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Mario Döbler, Robert A Marsden, and Bin Yang. 2023. Robust mean teacher for continual and gradual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7704–7714.

Cunhang Fan, Jiangyan Yi, Jianhua Tao, Zhengkun Tian, Bin Liu, and Zhengqi Wen. 2020. Gated recurrent fusion with joint training framework for robust end-to-end speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:198–209.

Li Fu, Shanyong Yu, Siqi Li, Lu Fan, Youzheng Wu, and Xiaodong He. 2025. Ume: Upcycling mixture-of-experts for scalable and efficient automatic speech recognition. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Shahram Ghorbani and John HL Hansen. 2022. Domain expansion for end-to-end speech recognition: Applications for accent/dialect speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:762–774.

Yuan Gong, Sameer Khurana, Leonid Karlinsky, and James Glass. 2023. Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers. In *Proc. Interspeech 2023*, pages 2798–2802.

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

Abhinav Jain, Minali Upreti, and Preethi Jyothi. 2018. Improved accented speech recognition using accent embeddings and multi-task learning. In *Interspeech*, pages 2454–2458.

Changhun Kim, Joonhyung Park, Hajin Shim, and Eunho Yang. 2023. Sgem: Test-time adaptation for automatic speech recognition via sequential-level generalized entropy minimization. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2023, pages 3367–3371.

Jae-Hong Lee, Do-Hee Kim, and Joon-Hyuk Chang. 2023. Awmc: Online test-time adaptation without mode collapse for continual adaptation. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Jae-Hong Lee, Sang-Eon Lee, Dong-Hyun Kim, DoHee Kim, and Joon-Hyuk Chang. 2024. Online subloop search via uncertainty quantization for efficient test-time adaptation. In *Proc. Interspeech 2024*, pages 2880–2884.

Chengxi Lei, Satwinder Dr Singh, Feng Hou, and Ruili Wang. 2024. Mix-fine-tune: An alternate fine-tuning strategy for domain adaptation and generalization of low-resource asr. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, pages 1–7.

Guan-Ting Lin, Wei-Ping Huang, and Hung-yi Lee. 2024. Continual test-time adaptation for end-to-end speech recognition on noisy speech. *arXiv preprint arXiv:2406.11064*.

Guan-Ting Lin, Shang-Wen Li, and Hung-yi Lee. 2022. Listen, adapt, better wer: Source-free single-utterance test-time adaptation for automatic speech recognition. *arXiv preprint arXiv:2203.14222*.

Hongfu Liu, Hengguan Huang, and Ye Wang. 2024. Advancing test-time adaptation in wild acoustic test settings. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7138–7155.

Jiaming Liu, Senqiao Yang, Peidong Jia, Renrui Zhang, Ming Lu, Yandong Guo, Wei Xue, and Shanghang Zhang. Vida: Homeostatic visual domain adapter for continual test time adaptation. In *The Twelfth International Conference on Learning Representations*.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. 2022. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pages 16888–16905. PMLR.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015*

IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 5206–5210. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Chandan KA Reddy, Ebrahim Beyrami, Jamie Pool, Ross Cutler, Sriram Srinivasan, and Johannes Gehrke. 2019. A scalable noisy speech dataset and online subjective test framework. In *Proc. Interspeech 2019*, pages 1816–1820.

Khe Chai Sim, Zhouyuan Huo, Tsendsuren Munkhdalai, Nikhil Siddhartha, Adam Stooke, Zhong Meng, Bo Li, and Tara Sainath. 2024. A comparison of parameter-efficient asr domain adaptation methods for universal speech and language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6900–6904. IEEE.

Junha Song, Jungsoo Lee, In So Kweon, and Sungha Choi. 2023. Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11920–11929.

Sining Sun, Binbin Zhang, Lei Xie, and Yanning Zhang. 2017. An unsupervised deep domain adaptation approach for robust speech recognition. *Neurocomputing*, 257:79–87.

Minh Tran, Yutong Pang, Debjyoti Paul, Laxmi Pandey, Kevin Jiang, Jinxi Guo, Ke Li, Shun Zhang, Xuedong Zhang, and Xin Lei. 2025. A domain adaptation framework for speech recognition systems with only synthetic data. *arXiv preprint arXiv:2501.12501*.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2020. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*.

Huimeng Wang, Zengrui Jin, Mengzhe Geng, Shujie Hu, Guinan Li, Tianzi Wang, Haoning Xu, and Xunying Liu. 2024a. Enhancing pre-trained asr system fine-tuning for dysarthric speech recognition using adversarial data augmentation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12311–12315. IEEE.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. 2022a. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211.

Yanshuo Wang, Ali Cheraghian, Zeeshan Hayder, Jie Hong, Sameera Ramasinghe, Shafin Rahman, David Ahmedt-Aristizabal, Xuesong Li, Lars Petersson, and Mehrtash Harandi. 2024b. Backpropagation-free network for 3d test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23231–23241.

Yanshuo Wang, Jie Hong, Ali Cheraghian, Shafin Rahman, David Ahmedt-Aristizabal, Lars Petersson, and Mehrtash Harandi. 2024c. Continual test-time domain adaptation via dynamic sample selection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1701–1710.

Yiming Wang, Jinyu Li, Heming Wang, Yao Qian, Chengyi Wang, and Yu Wu. 2022b. Wav2vec-switch: Contrastive learning from original-noisy speech pairs for robust speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7097–7101. IEEE.

Zhong-Qiu Wang and DeLiang Wang. 2016. A joint training framework for robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):796–806.

Feifei Xiong, Jon Barker, Zhengjun Yue, and Heidi Christensen. 2020. Source domain data selection for improved transfer learning targeting dysarthric speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7424–7428. IEEE.

Eunseop Yoon, Hee Suk Yoon, John Harvill, Mark Hasegawa-Johnson, and Chang D Yoo. 2024. Litta: Language informed test-time adaptation for automatic speech recognition. *arXiv preprint arXiv:2408.05769*.

Jicheng Zhang, Yizhou Peng, Van Tung Pham, Haihua Xu, Hao Huang, and Eng Siong Chng. 2021. E2e-based multi-task learning approach to joint speech and accent recognition. In *Proc. Interspeech 2021*, pages 1519–1523.