석사학위논문
Master's Thesis

문장 수준의 일반화된 엔트로피 최소화를 통한 음성 인식 모델에 대한 테스트타임 적응

Test-Time Adaptation for Automatic Speech Recognition via Sequential-Level Generalized Entropy Minimization

2024

김창훈 (金 昶 勳 Kim, Changhun)

한국과학기술원

Korea Advanced Institute of Science and Technology

석사학위논문

문장 수준의 일반화된 엔트로피 최소화를 통한 음성 인식 모델에 대한 테스트타임 적응

2024

김창훈

한국과학기술원

김재철AI대학원

문장 수준의 일반화된 엔트로피 최소화를 통한 음성 인식 모델에 대한 테스트타임 적응

김 창 훈

위 논문은 한국과학기술원 석사학위논문으로
학위논문 심사위원회의 심사를 통과하였음

2023년 12월 18일

심사위원장 양 은 호 (인)

심 사 위 원 신 진 우 (인)

심 사 위 원 황 성 주 (인)

Test-Time Adaptation for Automatic Speech Recognition via Sequential-Level Generalized Entropy Minimization

Changhun Kim

Advisor: Eunho Yang

A dissertation submitted to the faculty of
Korea Advanced Institute of Science and Technology in
partial fulfillment of the requirements for the degree of
Master of Science in AI

Daejeon, Korea
December 18, 2023

Approved by

Eunho Yang
Professor in Kim Jaechul Graduate School of AI

The study was conducted in accordance with Code of Research Ethics[1].

[1] Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

MAI 김창훈. 문장 수준의 일반화된 엔트로피 최소화를 통한 음성 인식 모델에 대한 테스트타임 적응. 김재철AI대학원. 2024년. 22+iv 쪽. 지도교수: 양은호. (영문 논문)
Changhun Kim. Test-Time Adaptation for Automatic Speech Recognition via Sequential-Level Generalized Entropy Minimization. Kim Jaechul Graduate School of AI. 2024. 22+iv pages. Advisor: Eunho Yang. (Text in English)

초록

음성 인식 모델은 실제 배포 환경에서 데이터 분포 변화에 빈번히 노출되며, 이에 따라 모델은 부정확한 예측을 하게 된다. 이러한 문제를 해결하기 위해 이미 학습된 음성 인식 모델을 학습 데이터에 접근하지 않은 채 레이블이 없는 테스트 데이터에 적응시키기 위한 테스트타임 적응 방법이 최근에 제안되었다. 이 방법은 상당한 성능 향상을 이루었지만, 단순한 탐욕적 디코딩에만 의존하고 각각의 타임 스텝에서 독립적으로 모델의 적응을 진행한다. 이러한 테스트타임 적응 방법은 음성 인식 모델 출력의 순차적인 특성을 고려할 때 모델의 전체 출력인 문장 수준에서는 최적이 아닐 수 있다. 이러한 문제에 주목하여 본 논문에서는 일반적인 음성 인식 모델에 적용할 수 있는 테스트타임 적응 프레임워크인 문장 수준의 일반화된 엔트로피 최소화 (Sequential-level Generalized Entropy Minimization; SGEM)를 소개한다. 순차적 출력을 고려하기 위해 SGEM은 먼저 빔 서치를 이용하여 후보 출력 로짓을 탐색한 뒤 가장 가능도가 높은 출력 로짓을 선택한다. 선택된 출력 로짓을 바탕으로 SGEM은 일반화된 엔트로피 최소화와 네거티브 샘플링을 비지도 목적 함수로 사용하여 모델의 파라미터를 학습시킨다. 광범위한 실험을 통해 SGEM은 다양한 분포 변화하에서 세 가지 주요 음성 인식 모델에 대해 최고의 성능을 입증한다.

핵심낱말 기계 학습, 음성 인식, 분포 변화 강건성, 테스트타임 적응, 빔 서치, 엔트로피 최소화, 네거티브 샘플링

Abstract

In real-world scenarios, automatic speech recognition (ASR) models often encounter data distribution shifts, leading to erroneous predictions. To tackle this issue, a recent test-time adaptation (TTA) method has been proposed to adapt the pre-trained ASR model to the unlabeled target domain without source data. Despite decent performance gain, this approach relies solely on naive greedy decoding and performs adaptation across timesteps at the frame level, which may not be optimal given the sequential nature of model outputs. Motivated by this limitation, this thesis introduces a novel Sequential-level Generalized Entropy Minimization (SGEM) framework for general ASR models. To handle sequential output, SGEM first exploits beam search to explore candidate output logits and selects the most plausible one. Then, it utilizes generalized entropy minimization and negative sampling as effective unsupervised objectives to adapt the model. Through extensive experiments, SGEM verifies its state-of-the-art performance across three mainstream ASR models under various distribution shifts.

Keywords machine learning, automatic speech recognition, distribution shift robustness, test-time adaptation, beam search, entropy minimization, negative sampling

Contents

Contents ii
List of Tables iii
List of Figures iv

Chapter 1. Introduction 1

Chapter 2. Related Works 3
2.1 Automatic Speech Recognition 3
2.2 Domain Adaptation and Generalization for ASR Models 3
2.3 Test-Time Adaptation 3

Chapter 3. Proposed Method: SGEM 5
3.1 Test-Time Adaptation Setup for ASR Models 5
3.2 Beam Search-Based Logit Acquisition 5
3.3 Generalized Entropy Minimization 6
3.4 Negative Sampling 7
3.5 Overall Framework 7

Chapter 4. Experiments 9
4.1 Experimental Setup 9
4.2 Main Results 10
4.3 Evaluation under Challenging yet Realistic Scenarios 11
4.4 Furthering Performance Improvements 12
4.5 Ablation Study 14

Chapter 5. Concluding Remarks 15
5.1 Conclusion 15
5.2 Limitations 15
5.3 Future Works 15

Bibliography 17

Acknowledgments 21

Curriculum Vitae 22

List of Tables

4.1 Word Error Rate (%) of three mainstream ASR models evaluated on 12 diverse datasets exhibiting various distribution shifts. The results are obtained using greedy decoding during the inference phase. 10

4.2 Word Error Rate (%) of the CTC-based model evaluated on 12 diverse datasets exhibiting various distribution shifts. The results are derived using beam search decoding, supplemented by an external language model during the inference phase. 10

4.3 Word Error Rate (%) of the CTC-based model on six non-native English speech corpora. 11

4.4 Word Error Rate (%) of the CTC-based model on 8 diverse datasets, considering additional SNRs of 0dB and -10dB, introducing more severe background noises. 12

4.5 Ablation study investigating the core components of SGEM, namely beam search-based logit acquisition (BS), generalized entropy minimization (GEM), and negative sampling (NS), conducted across three mainstream ASR models on the CH dataset. 14

List of Figures

3.1 The overall pipeline of the proposed SGEM framework. Given a single utterance x in the target domain, we first acquire logits based on the most plausible beam search output. Then, we utilize generalized entropy minimization and negative sampling as unsupervised objectives with x itself to adapt the ASR model $f(\cdot|\theta)$ pre-trained on the source domain. . 6

4.1 Word Error Rate Reduction (%) of the CTC-based model across various utterance lengths on the CH dataset. 11

4.2 Word Error Rate (%) of SUTA and SGEM for the CTC-based model on the CH dataset under batch sizes of 4, 16, and 64. 12

4.3 Word Error Rate Difference (%p) of SUTA and SGEM calculated with the unadapted CTC-based model. The evaluation is conducted across 10 datasets with diverse distribution shifts, utilizing a fixed batch size of 16. 13

4.4 Word Error Rate (%) of SGEM combined with KLD for the CTC-based model, showcasing the impact of various regularization weight ρ on the CH dataset. 13

Chapter 1. Introduction

While deep neural networks have achieved remarkable progress in a broad range of areas, such as computer vision [27, 2], natural language processing [55, 52], and speech processing [48, 4], these models are known to be susceptible to data distribution shifts [53, 18]. This so-called *domain shift problem*[1] readily occurs in automatic speech recognition (ASR) models, which imposes challenges in deploying these models to real-world applications. For example, utterances of unseen speakers/words not exposed during training or utterances with accidental background noise can be given at test time.

To tackle this domain shift problem for ASR models, many prior works have been suggested, including data augmentation [29], feature alignment [28], domain adversarial learning [49, 50], and knowledge distillation [39]. These approaches primarily address this issue under the unsupervised domain adaptation (UDA) setting, where the source models are jointly trained using both labeled source data and unlabeled target data, aiming to adapt to unlabeled target domains. However, this UDA setting has several impractical assumptions in real-world scenarios. First, it assumes that the source data is accessible, which might be unavailable due to privacy/storage issues. Second, a pile of target data has to be collected *in advance*. This is also unrealistic as it requires substantial resources. Even worse, it restricts the generalization capacity of the model only to the pre-collected target data, although the target distribution can change arbitrarily at test time. In parallel, several speaker adaptation methods [47, 60, 62] have demonstrated satisfactory adaptation performance on variation in speakers during inference. However, their restricted focus on speaker changes and reliance on prior knowledge of test-time speakers also impose limitations in effectively addressing arbitrary distribution shifts during test time.

Inspired by these limitations, SUTA [36] was first proposed to address arbitrary distribution shifts under a more realistic test-time adaptation (TTA) setting for ASR models as in other domains [51, 56, 37, 15, 63, 12, 57, 6, 22, 42]. Given an off-the-shelf ASR model pre-trained on the source domain, TTA methods aim to adapt the model on the fly using unlabeled instances from the target domain in test time without access to source data. Inheriting the ideas of the TTA approaches in the computer vision domain [56, 15, 63], SUTA shows decent performance in a single-utterance TTA setting for the CTC-based ASR model [4].

However, directly adopting this approach to advanced ASR models [8, 44] could not be optimal as SUTA was developed with the CTC-based models [25] in mind. This is because, unlike the CTC-based models, which generate each output token independently in a greedy manner, advanced ASR models are designed to work in an autoregressive manner or typically utilize beam search decoding with an external language model during the test phase. This indicates that output logits acquired by greedy decoding may not adequately capture the output distribution, and naively adapting with these logits at the frame level as in Lin et al. [36] can cause undesirable behavior at the sequential level for general ASR models.

Motivated by this, we propose the Sequential-level Generalized Entropy Minimization (SGEM) framework to achieve effective TTA for general ASR models. To this end, SGEM first explores candidate output logits and selects the most plausible one using beam search to leverage the sequential nature of the output. Then, SGEM leverages generalized entropy minimization and negative sampling as unsupervised losses to adapt the model at the sequential level. We validate SGEM for three representative ASR models on various datasets with different distribution shifts and demonstrate that SGEM achieves state-of-the-art

[1] In this thesis, we use the terms *domain shift* and *distribution shift* interchangeably.

performance in most settings. To the best of our knowledge, this is the first work suggesting the TTA method for general ASR models. In summary, our contribution is threefold:

- Inspired by the suboptimality in the existing TTA method for ASR models at the sequential level, we introduce SGEM tailored for general ASR models. SGEM is a carefully designed framework considering the sequential nature of ASR output, utilizing beam search-based logit acquisition.

- Furthermore, we suggest two effective modality-agnostic unsupervised objectives–generalized entropy minimization and negative sampling–for the first time in the context of test-time adaptation. These objectives are capable of orthogonal integration with the proposed logit acquisition strategy.

- Through comprehensive experiments, we empirically demonstrate that SGEM achieves state-of-the-art performance across three mainstream ASR models, encompassing diverse datasets with various distribution shifts.

Chapter 2. Related Works

2.1 Automatic Speech Recognition

In recent years, automatic speech recognition (ASR) has witnessed significant advancements, driven by the convergence of deep learning techniques [10, 26] and the availability of large-scale datasets [3, 17, 4]. Connectionist temporal classification (CTC)-based models prioritize simplicity [25, 10, 32, 4, 26], ensuring computational efficiency despite limitations in handling fine-grained alignment information. Attention-based models [14, 10, 44] dynamically weigh input frames for adaptive decoding, enhancing adaptation to varying input lengths and complex alignments. In online ASR, Transducers [24, 8] offer continuous and incremental decoding for low-latency applications like live transcription and interactive voice systems. Moreover, a growing focus on self-supervised learning [4] utilizes unlabeled data for pre-training, and the exploration of multi-modal ASR [9, 11] and multi-lingual ASR [59] continues to broaden the horizons of ASR research. Although various research efforts are actively advancing ASR models in multiple directions, these models inherently experience performance degradation in situations characterized by a distribution mismatch between the training set and the test set.

2.2 Domain Adaptation and Generalization for ASR Models

With the rapid evolution of automatic speech recognition models and their susceptibility to data distribution shifts, recent efforts have been directed toward enhancing the generalization capabilities of models trained on the source domain. A prominent subset of these methods pertains to unsupervised domain adaptation (UDA) techniques [29, 49, 50, 39, 28]. UDA methods aim to attain robust performance on a target domain by leveraging both labeled source data and unlabeled target data during the training phase. Strategies within this category encompass explicit feature alignment [28], domain adversarial learning [49, 50], and knowledge distillation [39]. Another significant category is domain generalization [1, 41, 58], which seeks to enhance models' generalization performance across multiple target domains. These strategies primarily encompass data augmentation [1], large-scale training [41], and margin-based training [58] to augment the model's generalization capabilities. Despite demonstrating commendable performance gains, these methods have limitations, including their inability to operate when source data is inaccessible due to privacy or storage concerns and a lack of adaptability to tailor the model to the current test data.

2.3 Test-Time Adaptation

Test-time adaptation (TTA) [33, 34, 56, 51, 37, 63, 7, 20, 40, 42] represents an evolving paradigm aimed at addressing the limitations of unsupervised domain adaptation and domain generalization strategies. These methods focus on adapting the pre-trained model from the source domain to an arbitrary target domain on the fly, utilizing unlabeled target data exclusively, without requiring source data. They can be categorized into three groups: fully test-time adaptation, test-time training, and batch-norm statistics calibration. Fully test-time adaptation [33, 34, 56, 63, 42] involves unsupervised training, relying solely on the model's predictions without modifying the training process. Test-time training [51, 37] involves training the model with both the main objective and an unsupervised objective. The unsupervised objective is then utilized for model adaptation during the inference phase. Batch-norm statistics calibration methods [40, 61, 35, 20] update the statistics of batch normalization layers (*i.e.*, mean and standard deviation) using test instances to alleviate distributional mismatches between source and target data. Recently, SUTA [36] proposed fully test-time adaptation for ASR models, demonstrating decent performance in a single-utterance TTA setting for the CTC-based ASR model [4]. However, applying this approach directly to advanced ASR models [8, 44] may not be optimal. This is because logits acquired by greedy decoding may not sufficiently capture the output distribution, and naive adaptation at the frame level can lead to undesirable behavior at the sequential level for general ASR models.

Chapter 3. Proposed Method: SGEM

This chapter introduces Sequential-level Generalized Entropy Minimization (SGEM), an effective TTA framework for general ASR models. To this end, we describe the test-time adaptation setup for ASR models in Section 3.1 and illustrate the core strategies of SGEM from Section 3.2 to Section 3.5. Figure 3.1 depicts the overall pipeline of the proposed method.

3.1 Test-Time Adaptation Setup for ASR Models

We briefly explain general ASR models before formulating test-time adaptation (TTA). Let $f(\cdot|\theta)$ be an ASR model trained on the labeled source domain $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_i$ in pairs of speech x_i^s and text y_i^s, which takes a raw waveform x and returns output logits $f(x|\theta) \in \mathbb{R}^{L \times C}$ for each timestep. Here, L is the number of timesteps, C is the number of vocabulary classes, and θ is the model parameter. It models the log joint probability $\log p(\boldsymbol{y}|x, \theta)$ of a candidate transcript

$$\boldsymbol{y} = (y_1, \cdots, y_L),$$

in an *autoregressive manner* as follows:

$$\begin{aligned}\log p(\boldsymbol{y}|x,\theta) &:= \log p_{\mathrm{AM}}(\boldsymbol{y}|x,\theta) + \lambda_{\mathrm{LM}} \log p_{\mathrm{LM}}(\boldsymbol{y}) + Z \\ &= \sum_{i=1}^{L} \log p_{\mathrm{AM}}(y_i|y_{<i}, x, \theta) + \lambda_{\mathrm{LM}} \log p_{\mathrm{LM}}(y_i|y_{<i}) + Z,\end{aligned}$$

where $y_i \in \{1, \cdots, C\}$, $p_{\mathrm{AM}}(\boldsymbol{y}|x,\theta)$ is the joint probability given by the model output $f(x|\theta)$, $p_{\mathrm{LM}}(\boldsymbol{y})$ is the joint probability of an autoregressive language model (LM), λ_{LM} is a hyperparameter to control the effect of the LM, and Z is a normalizing constant. LM aims to boost the ASR model to generate more faithful sentences. ASR decoding strategies approximate the optimal solution

$$\boldsymbol{y}^* = \arg\max_{\boldsymbol{y}} \log p(\boldsymbol{y}|x, \theta).$$

TTA methods for an ASR model $f(\cdot|\theta)$ aim to adapt the model to the unlabeled target speech domain $\mathcal{D}_t = \{x_i^t\}_i$ without access to $\mathcal{D}_s$. Specifically, we consider a *single-utterance TTA setting* where we fine-tune the ASR model $f(\cdot|\theta)$ for each utterance $x_i^t \in \mathcal{D}_t$ to get more precise output logits $\log p(\boldsymbol{y}|x_i^t, \theta)$ with unsupervised objectives using only x_i^t itself. This single-utterance TTA setting is considerably pragmatic regarding low latency without presuming that the test instances are independent and identically distributed [36, 57, 42].

3.2 Beam Search-Based Logit Acquisition

An existing TTA method for ASR models [36] exploits the greedy decoding strategy without leveraging the external LM (*i.e.*, $\lambda_{\mathrm{LM}} = 0$) to get output logits for all timesteps. Also, it utilizes the logits with unsupervised objectives such as entropy minimization [23] and minimum class confusion [30] as if they are appropriate logits to adapt the ASR model. However, naively using greedy decoding is proven defective [16] and can mislead the model to be adapted on the wrong labels. Furthermore, this frame-level greedy

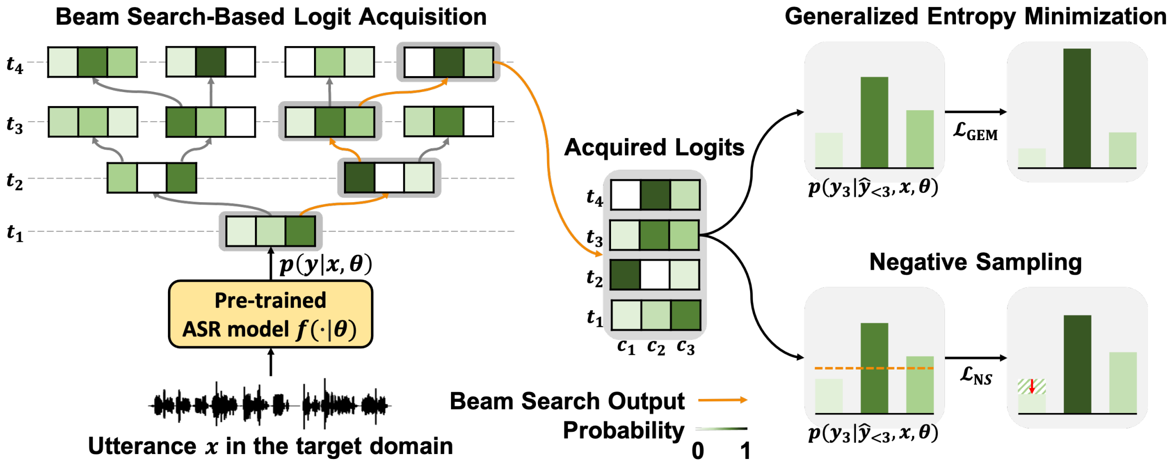
Beam Search-Based Logit Acquisition
t_4
t_3
t_2
t_1
$p(y|x,\theta)$
Pre-trained ASR model $f(\cdot|\theta)$
Utterance x in the target domain
Acquired Logits
c_1 c_2 c_3
Beam Search Output
Probability
0 1
Generalized Entropy Minimization
$\mathcal{L}_{\text{GEM}}$
$p(y_3|\hat{y}_{<3},x,\theta)$
Negative Sampling
$\mathcal{L}_{\text{NS}}$
$p(y_3|\hat{y}_{<3},x,\theta)$

Figure 3.1: The overall pipeline of the proposed SGEM framework. Given a single utterance x in the target domain, we first acquire logits based on the most plausible beam search output. Then, we utilize generalized entropy minimization and negative sampling as unsupervised objectives with x itself to adapt the ASR model $f(\cdot|\theta)$ pre-trained on the source domain.

adaptation might be sub-optimal on the sequential output since it only considers the joint probability of a sequence myopically over timesteps.

To overcome these challenges, we exploit a novel logit acquisition strategy based on more systematic beam search decoding. Given a beam width B, we find the most plausible output sequence

$$\hat{\boldsymbol{y}} = (\hat{y}_1, \cdots, \hat{y}_L), \tag{3.1}$$

which approximates $\boldsymbol{y}^*$ using beam search [16]. Note that we do not hold the logits of beam candidates in this step to reduce memory consumption. Instead, the estimated sequence $\hat{\boldsymbol{y}}$ is passed to the model again to acquire the i-th logit

$$\boldsymbol{o}_i = (o_{i1}, \cdots, o_{iC}) \in \mathbb{R}^C$$

for all $i \in \{1, \cdots, L\}$, where $o_{ij} = \log p(y_i = j|\hat{y}_{<i}, x, \theta)$. Our intuition behind the beam search-based logit acquisition is that *considering the logits obtained from beam search for adaptation naturally aligned with how ASR models decode sentences*, *i.e.*, our approach can tailor the model to adapt toward the actual yet accurate sentences generated by the ASR models.

3.3 Generalized Entropy Minimization

While entropy minimization achieves decent performance on domain adaptation tasks [56, 15, 63, 31] by reducing the uncertainty of predictions, we can further improve this objective by adopting its generalized version, Rényi entropy, and searching for more effective hyperparameter. For a discrete random variable X, which takes values in $\{1, \cdots, C\}$, Rényi entropy $\mathrm{H}_\alpha(X)$ of order α with $\alpha \in (0,1) \cup (1, \infty)$ is defined as follows:

$$\mathrm{H}_\alpha(X) = \frac{1}{1-\alpha} \log \Big(\sum_{j=1}^{C} \mathbb{P}(X = j)^\alpha \Big).$$

When $\alpha \to 1$ and $\alpha \to \infty$, $\mathrm{H}_\alpha(X)$ becomes Shannon entropy, and cross-entropy with a pseudo-label $\arg\max_j \mathbb{P}(X = j)$, respectively. For a single-utterance TTA setting, we hypothesize that there exists an optimal $\alpha \in (0,1) \cup (1,\infty)$ and define the generalized entropy loss as follows:

$$\mathcal{L}_{\text{GEM}} = \frac{1}{L}\sum_{i=1}^{L} \frac{1}{1-\alpha} \log \Big(\sum_{j=1}^{C} p_{ij}^{\alpha} \Big), \tag{3.2}$$

where

$$p_{ij} = \frac{\exp(o_{ij}/T)}{\sum_{j'=1}^{C} \exp(o_{ij'}/T)},$$

and T is a temperature hyperparameter for preventing *vanishing gradient.* As the blank token dominates all timesteps, we ignore timesteps with the highest probability of blank token among all classes to alleviate the class imbalance problem as in Lin et al. [36].

3.4 Negative Sampling

We further exploit negative sampling loss, originally adopted for semi-supervised learning (SSL) in Chen et al. [13]. Negative sampling loss penalizes the probabilities of low-confident classes, and Chen et al. [13] have shown that adding it can further boost the performance of existing SSL algorithms. It can be derived from the standard cross-entropy loss as follows. Given L labeled samples $\{(x_i, y_i)\}_{i=1}^{L}$, the standard cross-entropy loss is defined as

$$\mathcal{L}_{\text{CE}} = -\frac{1}{L}\sum_{i=1}^{L}\sum_{j=1}^{C} \mathbb{1}_{[j=y_i]} \log p_{ij}.$$

Here, please note that

$$\sum_{j=1}^{C} \mathbb{1}_{[j=y_i]} \log p_{ij} = \log \Big(1 - \sum_{j \neq y_i} p_{ij} \Big).$$

Since we do not know the ground truth label y_i for each x_i in the unlabeled target domain, we approximate $\mathcal{L}_{\text{CE}}$ with the negative sampling loss $\mathcal{L}_{\text{NS}}$ defined as follows:

$$\mathcal{L}_{\text{NS}} = -\frac{1}{L}\sum_{i=1}^{L} \log \Big(1 - \sum_{j=1}^{C} \mathbb{1}_{[p'_{ij} < \tau]} p_{ij} \Big), \tag{3.3}$$

where

$$p_{ij} = \frac{\exp(o_{ij}/T)}{\sum_{j'=1}^{C} \exp(o_{ij'}/T)},\ p'_{ij} = \frac{\exp(o_{ij})}{\sum_{j'=1}^{C} \exp(o_{ij'})},$$

with a temperature hyperparameter T to avoid vanishing gradient, and $\mathbb{1}$ is an indicator function. j-th class of x_i is considered as a negative class when the probability p'_{ij} is less than a threshold τ. Without any modification, Equation (3.3) can be interpreted in a single-utterance TTA setting as penalizing probabilities of negative classes at every timestep for a sequential output of length L.

3.5 Overall Framework

Our entire unsupervised objective is the weighted sum of the generalized entropy loss and the negative sampling loss as follows:

$$\mathcal{L} = \mathcal{L}_{\text{GEM}} + \lambda_{\text{NS}} \mathcal{L}_{\text{NS}}, \tag{3.4}$$

where λ_{NS} is negative sampling weight for balancing two losses. For each utterance, we adapt the model for N iterations in an *episodic manner* where we newly reset the model to the pre-trained one to preserve the knowledge from the source domain. The detailed procedure of SGEM is minutely described in Algorithm 1.

Algorithm 1 SGEM

```
**Input:** $f(\cdot|\theta)$: pre-trained ASR model on the source domain, $x$: an utterance in the target domain
**Parameters:** Opt: optimizer, Lrs: learning rate scheduler, $\eta_i$: initial learning rate, $\eta_f$: final learning rate
**Output:** Transcription of $f(x|\theta_N)$
$\theta_0,\ \eta_0 \leftarrow \theta,\ \eta$
**for** $n = 1$ to $N$ **do**
    **for** $i = 1$ to $L$ **do**
        $\hat{y}_i \leftarrow i$-th token of the beam search output $\hat{\boldsymbol{y}}$
    **for** $i = 1$ to $L$ and $j = 1$ to $C$ **do**
        $o_{ij} \leftarrow \log p(y_i = j|\hat{y}_{<i}, x, \theta)$
    $p_{ij},\ p'_{ij} \leftarrow \frac{\exp(o_{ij}/T)}{\sum_{j'=1}^{C}\exp(o_{ij'}/T)},\ \frac{\exp(o_{ij})}{\sum_{j'=1}^{C}\exp(o_{ij'})}$
    $\mathcal{L}_{\text{GEM}} \leftarrow \frac{1}{L}\sum_{i=1}^{L}\frac{1}{1-\alpha}\log\left(\sum_{j=1}^{C}p_{ij}^{\alpha}\right)$
    $\mathcal{L}_{\text{NS}} \leftarrow -\frac{1}{L}\sum_{i=1}^{L}\log\left(1-\sum_{j=1}^{C}\mathbb{1}_{[p'_{ij}<\tau]}p_{ij}\right)$
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{GEM}} + \lambda_{\text{NS}}\mathcal{L}_{\text{NS}}$
    $\theta_n,\ \eta_n \leftarrow \text{Opt}(\mathcal{L}, \theta_{n-1}, \eta_{n-1}),\ \text{Lrs}(n, \eta_i, \eta_f)$
```

Chapter 4. Experiments

In this chapter, we thoroughly demonstrate the empirical efficacy of SGEM. To achieve this, we first outline our experimental setup in Section 4.1 and present the main results in Section 4.2. Following that, we delve into a deeper analysis of SGEM under challenging yet realistic scenarios in Section 4.3 and validate its enhanced performance in situations with larger batch sizes and its synergy with KLD in Section 4.4. Finally, we verify each component of the proposed method through an ablation study in Section 4.5.

4.1 Experimental Setup

Source ASR models. To verify the efficacy of SGEM, we evaluate it on three mainstream ASR architectures: the CTC-based model [25], Conformer [26], and Transducer [24]. More specifically, for the CTC-based model, we use wav2vec 2.0 [4] [1] trained on the LibriSpeech dataset [43]. For Conformer, we exploit Conformer-CTC [26] [2] trained on the LibriSpeech dataset. For Transducer, we adopt Conformer-Transducer [8] [3] trained on a composite NeMO ASRSET dataset, including the LibriSpeech dataset. We utilize the external 4-gram language model [4] for the CTC-based model and Conformer.

Datasets. We assess the performance of SGEM on multiple datasets under various domain shift settings. To test SGEM under unseen speakers/words, we use the test set of four datasets: CHiME-3 (CH) [5], TED-LIUM 2 (TD) [46], Common Voice (CV) [3], and Valentini (VA) [54]. In addition, we validate SGEM under accident background noise by injecting the following eight types of noises to each utterance of in-domain LibriSpeech test-other dataset [43]: air conditioner (AC), airport announcement (AA), babble (BA), copy machine (CM), munching (MU), neighbors (NB), shutting door (SD), and typing (TP) with Signal-to-Noise Ratio (SNR) = 10dB. For each noise type, we randomly select one noise sample from the MS-SNSD noise test set [45]. We also evaluate SGEM on L2-Arctic [64], non-native English speech corpora, to verify SGEM under extreme pronunciation/accent shifts. Specifically, we select one speaker for each first language.

Implementation details. Since the TTA setting has no validation set, we optimize hyperparameters on the CH dataset for each model and apply them to the other datasets. The best settings are as follows. For all models, we use AdamW optimizer [38] and cosine annealing learning rate scheduler with η_i and η_f for initial and final learning rates, respectively, and set $(N, T, \tau) = (10, 2.5, 0.4/C)$ with vocabulary size C. We only train feature extractors for the CTC-based model and encoders for the others. Furthermore, we set $(\eta_i, \eta_f, B, \lambda_{\text{LM}}, \alpha, \lambda_{\text{NS}}) = (4{\cdot}10^{-5}, 2{\cdot}10^{-5}, 5, 0.3, 1.5, 1)$ for the CTC-based model, $(4{\cdot}10^{-5}, 2{\cdot}10^{-5}, 5, 0.3, 1.25, 2)$ for Conformer, and $(4{\cdot}10^{-6}, 2{\cdot}10^{-6}, 3, 0, 1.25, 0.5)$ for Transducer. All experiments are conducted on Nvidia TITAN Xp and Nvidia GeForce RTX 3090. Adaptation takes

[1] `https://huggingface.co/facebook/wav2vec2-base-960h`

[2] `https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_conformer_ctc_small_ls`

[3] `https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_conformer_transducer_small`

[4] `https://huggingface.co/patrickvonplaten/wav2vec2-base-100h-with-lm`

Table 4.1: Word Error Rate (%) of three mainstream ASR models evaluated on 12 diverse datasets exhibiting various distribution shifts. The results are obtained using greedy decoding during the inference phase.

	Setting	**CH**	**TD**	**CV**	**VA**	**AC**	**AA**	**BA**	**CM**	**MU**	**NB**	**SD**	**TP**	**Avg.**
CTC-based model	**Unadapted**	31.2	13.2	36.9	14.5	28.1	40.9	66.6	49.8	50.4	119.2	19.2	26.2	41.4
	SUTA	25.0	**12.0**	31.4	11.8	17.7	31.3	55.2	39.4	39.7	113.0	15.0	17.8	34.1
	SGEM	**24.7**	**12.0**	**31.1**	**11.6**	**17.3**	**30.7**	**53.1**	**38.5**	**38.6**	**110.5**	**14.8**	**17.5**	**33.4**
Conformer	**Unadapted**	28.7	15.1	36.8	17.4	18.8	44.8	74.3	45.7	56.0	122.1	20.8	36.9	43.1
	SUTA	25.2	13.4	32.4	14.7	14.5	39.8	73.3	**38.4**	48.7	125.5	**16.4**	**28.8**	39.3
	SGEM	**24.5**	**13.3**	**31.6**	**14.6**	**14.4**	**38.5**	**70.4**	38.7	**48.5**	**120.9**	16.8	28.9	**38.4**
Transducer	**Unadapted**	11.8	7.2	12.9	6.5	14.1	20.4	31.0	29.7	31.3	**74.6**	12.7	16.2	22.4
	SUTA	10.3	6.8	12.1	5.5	12.0	18.5	28.3	26.7	28.7	**74.6**	11.7	14.7	20.8
	SGEM	**9.9**	**6.6**	**12.0**	**5.2**	**11.6**	**18.0**	**27.5**	**26.0**	**28.0**	76.5	**11.5**	**14.3**	**20.6**

Table 4.2: Word Error Rate (%) of the CTC-based model evaluated on 12 diverse datasets exhibiting various distribution shifts. The results are derived using beam search decoding, supplemented by an external language model during the inference phase.

Setting	**CH**	**TD**	**CV**	**VA**	**AC**	**AA**	**BA**	**CM**	**MU**	**NB**	**SD**	**TP**	**Avg.**
Unadapted	29.5	12.2	36.9	13.0	26.1	38.6	58.9	48.9	49.0	**91.6**	17.4	23.7	37.2
SUTA	**24.1**	**11.6**	31.5	11.4	16.8	30.3	53.2	38.1	38.6	107.9	14.1	**16.9**	32.9
SGEM	**24.1**	11.7	**31.1**	**11.1**	**16.5**	**29.8**	**51.6**	**37.7**	**37.7**	106.7	**14.0**	**16.9**	**32.4**

about 0.771 seconds for a 1-second utterance averaged over three models.[5]

4.2 Main Results

We evaluate the TTA performance of three prominent ASR models–CTC-based model, Conformer, and Transducer–across 12 datasets with diverse distribution shifts. In Table 4.1, we present the word error rate (WER) of ASR model outputs generated using the greedy search decoding method, following the evaluation protocol established in the prior study by Lin et al. [36]. Additionally, Table 4.2 illustrates the TTA performance for CTC-based models using beam search decoding with an external language model. For both decoding methods, ASR models integrated with SGEM consistently demonstrate improved accuracy in recognizing target utterances, resulting in an average word error rate reduction of 15.6%. An exception is noted in two cases on NB, where the performance without adaptation outperforms when using beam search decoding. Moreover, SGEM exhibits superior performance over SUTA in terms of average WER across all 12 datasets for each of the three model architectures (CTC-based model: (greedy) 34.1% → 33.4%, (beam search) 32.9% → 32.4%; Conformer: 39.3% → 38.4%; Transducer: 20.8% → 20.6%). This underlines the effectiveness of our unsupervised objectives as well as the logit acquisition method for adapting sequential language outputs, irrespective of the decoding strategy.

[5] Our code is available at `https://github.com/drumpt/SGEM`.

4.3 Evaluation under Challenging yet Realistic Scenarios

Non-native English speech corpora. To show the usability of SGEM at various distribution shifts, we further analyze SGEM on six different non-native English speech corpora. As shown in Table 4.3, SGEM achieves the best results for all corpora, outperforming the baseline. This implies the adaptability of SGEM under extreme pronunciation/accent shifts, demonstrating its versatility in practical situations with severe speaker shifts, such as globally used online ASR systems.

Table 4.3: Word Error Rate (%) of the CTC-based model on six non-native English speech corpora.

Setting	Unadapted	SUTA	SGEM
Arabic	32.5	27.1	**26.5**
Mandarin	28.5	23.3	**23.1**
Hindi	15.7	12.5	**12.3**
Korean	23.3	19.7	**19.5**
Spanish	35.7	29.8	**29.3**
Vietnamese	18.5	15.7	**15.4**
Average	25.7	21.4	**21.0**

Data deficient condition. It is commonly known that TTA methods fail under the data deficient condition where the number of test instances in each batch is limited [42, 19]. This still holds in the single-utterance TTA setting for ASR models, where the length of utterance is short, so the number of output timesteps is insufficient. To validate SGEM under this harsh condition, we split the CH dataset according to utterance length and evaluate SGEM with the CTC-based model on each split. As depicted in Figure 4.1, SGEM performs best in every length interval. In addition, it is worth noting that SGEM significantly outperforms the baseline for extremely short utterances of less than 2 seconds, showing the superiority of our method in real-world scenarios where short utterances are prevalent and negligible latency is required.

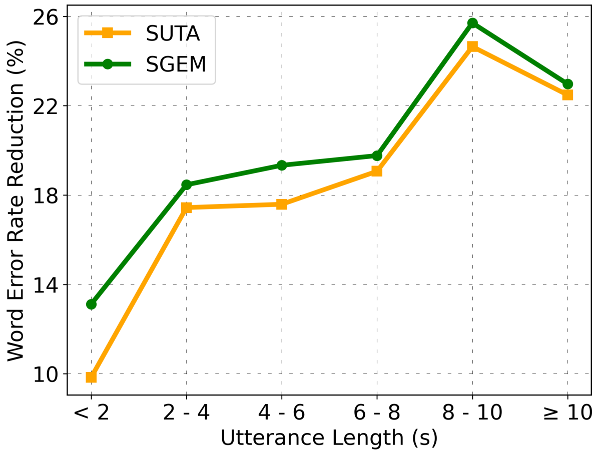

Figure 4.1: Word Error Rate Reduction (%) of the CTC-based model across various utterance lengths on the CH dataset.

Additional SNRs. To verify SGEM under more challenging distribution shifts, we conduct further experiments on additional SNRs with 0dB and -10dB for the CTC-based model. Evident from Table 4.4, SGEM consistently reduces WER in most cases, even in harsh background noises exhibiting superior performance compared to the baseline on average at each SNR.

Table 4.4: Word Error Rate (%) of the CTC-based model on 8 diverse datasets, considering additional SNRs of 0dB and -10dB, introducing more severe background noises.

SNR	Setting	AC	AA	BA	CM	MU	NB	SD	TP	Avg.
0dB	**Unadapted**	65.0	66.3	94.7	76.6	74.6	142.3	29.7	44.5	74.2
	SUTA	41.3	58.2	88.2	67.4	64.4	143.2	23.7	30.5	64.6
	SGEM	**40.8**	**57.4**	**85.3**	**67.2**	**63.1**	**141.8**	**23.3**	**29.5**	**63.6**
-10dB	**Unadapted**	94.2	83.5	111.2	93.5	90.1	**150.1**	45.0	69.0	92.1
	SUTA	**82.8**	**79.3**	108.8	**89.7**	85.3	151.4	37.6	53.5	86.1
	SGEM	83.8	79.4	**105.3**	90.2	**85.1**	150.9	**37.5**	**52.0**	**85.5**

4.4 Furthering Performance Improvements

Larger batch sizes. We also execute two further experiments on large batch sizes to see whether aggregating multiple utterances to adapt the model on themselves boosts the adaptation performance. First, we measure the performance of TTA methods for the CTC-based model on the CH dataset with batch sizes of 4, 16, and 64, respectively. Illustrated by Figure 4.2, increasing the batch size tends to boost the TTA performance in both methods, and SGEM achieves the best WER of 23.9% when the batch size is 16 while outperforming SUTA in all batch sizes. Second, we report the performance change of TTA methods with a batch size of 16 against the single utterance TTA setting for the CTC-based model on ten different datasets. Noted in Figure 4.3, we find that enlarging the batch size to 16 yields significant performance improvements under most distribution shifts for both methods (SUTA: 36.6% → 33.8%, SGEM: 35.8% → 32.2%).

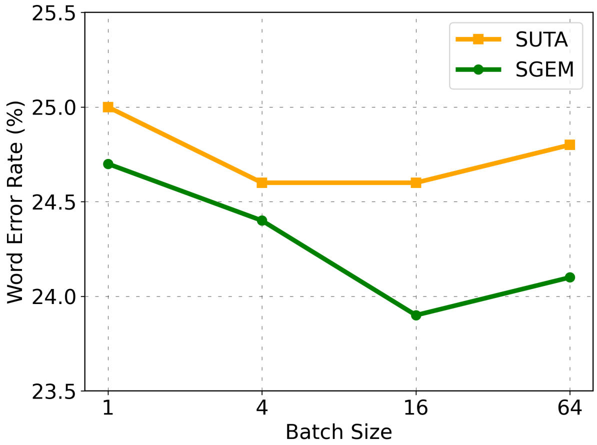

Figure 4.2: Word Error Rate (%) of SUTA and SGEM for the CTC-based model on the CH dataset under batch sizes of 4, 16, and 64.

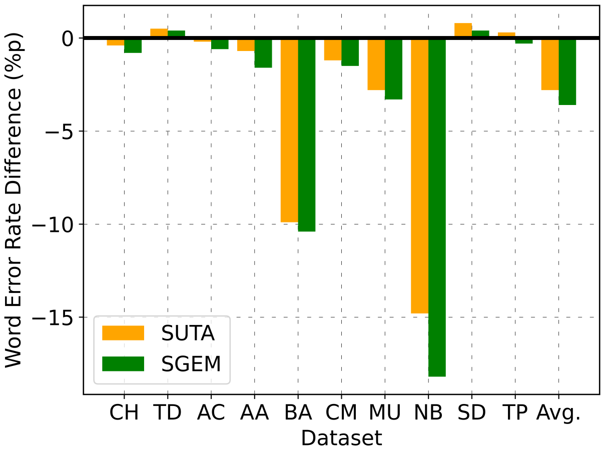

Figure 4.3: Word Error Rate Difference (%p) of SUTA and SGEM calculated with the unadapted CTC-based model. The evaluation is conducted across 10 datasets with diverse distribution shifts, utilizing a fixed batch size of 16.

Combination with Kullback-Leibler Divergence. In Yu et al. [62], the Kullback–Leibler Divergence (KLD) $\mathcal{L}_{\text{KLD}}$ between the output probabilities of the currently adapted model and the source-only pre-trained model is used to maintain the source knowledge. We adopt it to further boost the adaptation performance with a combined loss

$$\mathcal{L} = (1-\rho)\mathcal{L}_{\text{SGEM}} + \rho\mathcal{L}_{\text{KLD}}.$$

Figure 4.4 shows that combining these two losses with a regularization weight $\rho = 1/16$ for the CTC-based model on the CH dataset boosts the performance by applying regularization to preserve the knowledge of the source model.

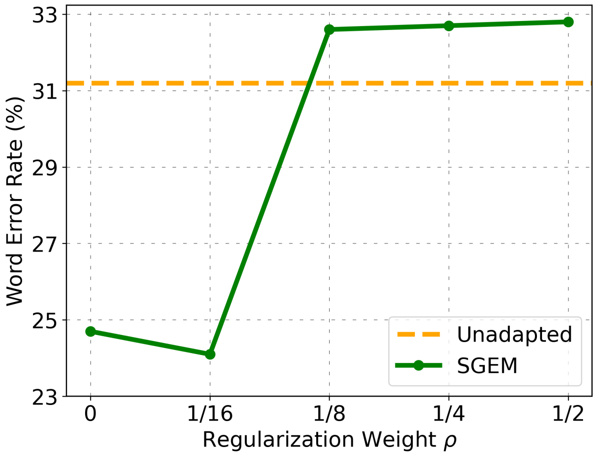

Figure 4.4: Word Error Rate (%) of SGEM combined with KLD for the CTC-based model, showcasing the impact of various regularization weight ρ on the CH dataset.

4.5 Ablation Study

To validate the core components of SGEM, namely beam search-based logit acquisition (BS, Section 3.2), generalized entropy minimization (GEM, Section 3.3), and negative sampling (NS, Section 3.4), we conduct an ablation study for three mainstream ASR models on the CH dataset. Demonstrated by Table 4.5, both generalized entropy minimization and negative sampling alone achieve remarkable performance gains for every model, indicating the efficacy of each component. Meanwhile, substituting greedy search for beam search even with small beam width (for all models) and without external LM (for Transducer) consistently boosts the performance in all cases, showing the effectiveness of beam search-based logit acquisition and implying that additional performance improvement can be expected using larger beam sizes or language model if resources are allowed.

Table 4.5: Ablation study investigating the core components of SGEM, namely beam search-based logit acquisition (BS), generalized entropy minimization (GEM), and negative sampling (NS), conducted across three mainstream ASR models on the CH dataset.

BS	**GEM**	**NS**	**CTC**	**Conformer**	**Transducer**
✗	✗	✗	31.2	28.7	11.8
✗	✓	✗	24.9	24.7	10.0
✗	✗	✓	25.2	25.0	10.1
✗	✓	✓	24.8	24.7	10.0
✓	✗	✓	24.8	24.7	10.1
✓	✓	✓	**24.7**	**24.5**	**9.9**

Chapter 5. Concluding Remarks

5.1 Conclusion

In this thesis, we have proposed SGEM, an effective single-utterance TTA framework designed for general ASR models. SGEM leverages beam search-based logit acquisition and incorporates unsupervised objectives such as generalized entropy minimization and negative sampling to adapt the model at the sequential level. Through comprehensive experiments, we showcased that SGEM achieves state-of-the-art performance across three mainstream ASR models, demonstrating robustness across diverse datasets with various distribution shifts. This also includes challenging yet realistic scenarios with unseen speakers or words during training, utterances with severe background noise, non-native English speech with pronounced accents, data-deficient conditions, and low SNRs. Furthermore, we validated the enhanced performance of SGEM in situations with larger batch sizes and its synergy with KLD. Finally, we conducted an ablation study to evaluate the efficacy of each component of SGEM. SGEM emphasizes the importance of carefully designing speech-specific components for devising effective TTA methods for ASR models.

5.2 Limitations

Although SGEM has led to a significant performance improvement, it also exhibits some limitations. Due to the utilization of beam search, SGEM consumes a considerable amount of time compared to using only greedy search. The computational bottleneck arises in Equation 3.2, where beam search is employed to obtain the most plausible output sequence. This drawback is expected to be alleviated by applying low-bit quantization to the ASR model only during the process outlined in Equation 3.2. This expectation is grounded in the assumption that, unlike output logits, output tokens would remain considerably more invariant even when subjected to quantization. Also, similar to other entropy minimization-based TTA variants [34, 56, 42], as discussed in Zhao et al. [65], SGEM demonstrates sensitivity to hyperparameters such as learning rate and adaptation step across different shift types. Specifically, for the NB dataset, which includes noise from surrounding conversations, Table 4.1, Table 4.2, and Table 4.4 consistently show SGEM's tendency to underperform compared to the unadapted model. We actually found that reducing the number of adaptation steps was observed to alleviate this issue. Addressing hyperparameter sensitivity in TTA methods remains a critical consideration for future research.

5.3 Future Works

Despite the rapid emergence of TTA strategies, they have predominantly been proposed and validated in the field of computer vision, particularly focusing on the image classification task. As observed in this thesis, we believe that TTA methods can be optimized by designing modality-specific components to fully leverage the characteristics of each domain. To this end, we are actively developing TTA methods for various machine learning domains beyond ASR tasks, including deep tabular learning, point cloud recognition, zero-shot classification of vision-language models, time series analysis, and more. Moreover, due to the inherent instability of TTA, which should adapt without knowledge of the ground truth labels, identifying problematic scenarios targeted at TTA instability, as demonstrated in previous works [20, 42, 66, 21], and devising TTA methods tailored to address these challenges, represents an intriguing future research direction. Finally, since test-time adaptation has predominantly focused only on enhancing accuracy, exploring research directions that consider aspects beyond accuracy, such as uncertainty calibration, out-of-distribution detection, and devising TTA methods that seamlessly integrate with these aspects, would be considered a valuable avenue of investigation.

Bibliography

[1] Ashish Alex, Lin Wang, Paolo Gastaldo, and Andrea Cavallaro. Data augmentation for speech separation. *Speech Communication*, 2023.

[2] Lucas Beyer Alexey Dosovitskiy and Alexander Kolesnikov et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[3] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *LREC*, 2020.

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*, 2020.

[5] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In *ASRU*, 2015.

[6] Alexander Bartler, Andre Bühler, Felix Wiewel, Mario Döbler, and Bin Yang. MT3: Meta test-time training for self-supervised test-time adaption. In *AISTATS*, 2022.

[7] Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *CVPR*, 2022.

[8] Maxime Burchi and Valentin Vielzeuf. Efficient conformer: Progressive downsampling and grouped attention for automatic speech recognition. In *ASRU*, 2021.

[9] David M Chan, Shalini Ghosh, Debmalya Chakrabarty, and Björn Hoffmeister. Multi-modal pre-training for automated speech recognition. In *ICASSP*, 2022.

[10] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *ICASSP*, 2016.

[11] Allen Chang, Xiaoyuan Zhu, Aarav Monga, Seoho Ahn, Tejas Srinivasan, and Jesse Thomason. Multimodal speech recognition for language-guided embodied agents. In *INTERSPEECH*, 2023.

[12] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *CVPR*, 2022.

[13] John Chen, Vatsal Shah, and Anastasios Kyrillidis. Negative sampling in semi-supervised learning. In *ICML*, 2020.

[14] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *NIPS*, 2015.

[15] François Fleuret et al. Test time adaptation through perturbation robustness. In *NeurIPS Workshop on Distribution Shifts*, 2021.

[16] Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. In *ACL*, 2017.

[17] Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. In *NeurIPS Track on Datasets and Benchmarks*, 2021.

[18] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.

[19] Jin Gao, Jialing Zhang, Xihui Liu, Trevor Darrell, Evan Shelhamer, and Dequan Wang. Back to the source: Diffusion-driven test-time adaptation. In *ICML Workshop on Dynamic Neural Networks*, 2022.

[20] Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. Note: Robust continual test-time adaptation against temporal correlation. In *NeurIPS*, 2022.

[21] Taesik Gong, Yewon Kim, Taeckyung Lee, Sorn Chottananurak, and Sung-Ju Lee. Sotta: Robust test-time adaptation on noisy data streams. In *NeurIPS*, 2023.

[22] Sachin Goyal, Mingjie Sun, Aditi Raghunathan, and Zico Kolter. Test-time adaptation via conjugate pseudo-labels. In *NeurIPS*, 2022.

[23] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *NIPS*, 2004.

[24] Alex Graves. Sequence transduction with recurrent neural networks. In *ICML Workshop on Representation Learning*, 2012.

[25] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, 2006.

[26] Anmol Gulati and Qin et al. Conformer: Convolution-augmented transformer for speech recognition. In *INTERSPEECH*, 2020.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[28] Wenxin Hou, Jindong Wang, Xu Tan, Tao Qin, and Takahiro Shinozaki. Cross-domain speech recognition with unsupervised character-level distribution matching. In *INTERSPEECH*, 2021.

[29] Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation. In *ASRU*, 2017.

[30] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In *ECCV*, 2020.

[31] Sameer Khurana, Niko Moritz, Takaaki Hori, and Jonathan Le Roux. Unsupervised domain adaptation for speech recognition via uncertainty driven self-training. In *ICASSP*, 2021.

[32] Suyoun Kim, Takaaki Hori, and Shinji Watanabe. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *ICASSP*, 2017.

[33] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*, 2013.

[34] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, 2020.

[35] Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. Ttn: A domain-shift aware batch normalization in test-time adaptation. In *ICLR*, 2023.

[36] Guan-Ting Lin, Shang-Wen Li, and Hung-yi Lee. Listen, adapt, better wer: Source-free single-utterance test-time adaptation for automatic speech recognition. In *INTERSPEECH*, 2022.

[37] Yuejiang Liu, Parth Kothari, Bastien van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. TTT++: When does self-supervised test-time training fail or thrive? In *NeurIPS*, 2021.

[38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2017.

[39] Vimal Manohar, Pegah Ghahremani, Daniel Povey, and Sanjeev Khudanpur. A teacher-student learning approach for unsupervised domain adaptation of sequence-trained asr models. In *SLT*, 2018.

[40] M Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *CVPR*, 2022.

[41] Arun Narayanan, Ananya Misra, Khe Chai Sim, Golan Pundak, Anshuman Tripathi, Mohamed Elfeky, Parisa Haghani, Trevor Strohman, and Michiel Bacchiani. Toward domain-invariant speech recognition via large scale training. In *SLT*, 2018.

[42] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *ICLR*, 2023.

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

[44] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, 2022.

[45] Chandan KA Reddy, Ebrahim Beyrami, Jamie Pool, Ross Cutler, Sriram Srinivasan, and Johannes Gehrke. A scalable noisy speech dataset and online subjective test framework. In *INTERSPEECH*, 2019.

[46] Anthony Rousseau, Paul Deléglise, Yannick Esteve, et al. Enhancing the ted-lium corpus with selected data for language modeling and more ted talks. In *LREC*, 2014.

[47] Frank Seide, Gang Li, Xie Chen, and Dong Yu. Feature engineering in context-dependent deep neural networks for conversational speech transcription. In *ASRU*, 2011.

[48] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *ICASSP*, 2018.

[49] Sining Sun, Binbin Zhang, Lei Xie, and Yanning Zhang. An unsupervised deep domain adaptation approach for robust speech recognition. *Neurocomputing*, 2017.

[50] Sining Sun, Ching-Feng Yeh, Mei-Yuh Hwang, Mari Ostendorf, and Lei Xie. Domain adversarial training for accented speech recognition. In *ICASSP*, 2018.

[51] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, 2020.

[52] Nick Ryder Tom B. Brown, Benjamin Mann and Melanie Subbiah et al. Language models are few-shot learners. In *NeurIPS*, 2020.

[53] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017.

[54] Cassia Valentini-Botinhao et al. Noisy speech database for training speech enhancement algorithms and tts models. *University of Edinburgh. School of Informatics. Centre for Speech Technology Research*, 2017.

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[56] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. TENT: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021.

[57] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *CVPR*, 2022.

[58] Xiong Xiao, Jinyu Li, Eng Siong Chng, Haizhou Li, and Chin-Hui Lee. A study on the generalization capability of acoustic models for robust speech recognition. *TASLP*, 2009.

[59] Hemant Yadav and Sunayana Sitaram. A survey of multilingual models for automatic speech recognition. In *LREC*, 2022.

[60] Kaisheng Yao, Dong Yu, Frank Seide, Hang Su, Li Deng, and Yifan Gong. Adaptation of context-dependent deep neural networks for automatic speech recognition. In *SLT*, 2012.

[61] Fuming You, Jingjing Li, and Zhou Zhao. Test-time batch statistics calibration for covariate shift. In *ICLR*, 2022.

[62] Dong Yu, Kaisheng Yao, Hang Su, Gang Li, and Frank Seide. Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In *ICASSP*, 2013.

[63] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In *NeurIPS*, 2022.

[64] Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. L2-arctic: A non-native english speech corpus. In *INTERSPEECH*, 2018.

[65] Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *ICML*, 2023.

[66] Zhi Zhou, Lan-Zhe Guo, Lin-Han Jia, Dingchu Zhang, and Yu-Feng Li. Ods: test-time adaptation in the presence of open-world data shift. In *ICML*, 2023.

Acknowledgment

Thinking back to the rapid progression of the past two years in my master's program, I recognize countless instances for which I am sincerely grateful. Foremost among them is the privilege of being a student under the guidance of Professor Eunho Yang, my master's advisor. Under his guidance, I received comprehensive and enriching support throughout the entire research journey from the selection of a research topic and the formulation of research questions to the development of methods, execution of experiments, intricacies of writing, and more. All of them were truly impossible without his unwavering support. Despite encountering occasional challenges in the research process, he consistently demonstrated patience and never expressed disappointment. His exceptional research skills, unwavering passion for continuous learning, dedication to research meetings, and outstanding personal qualities, characterized by warmth and delicacy, establish him as an ideal and exemplary researcher.

I would like to express my gratitude to Professor Jinwoo Shin and Professor Sung Ju Hwang, who served as members of my master's thesis committee. Their provision of exceptionally constructive and insightful feedback during the defense process has been invaluable. I anticipate that these valuable insights will significantly contribute to the success of my future research endeavors.

I also appreciate the meaningful interactions with all the students I encountered during my master's program at the Machine Learning and Intelligence Laboratory, as their presence significantly contributed to both my personal and academic development. Special acknowledgment goes to my doctoral senior colleagues, Joonhyung Park, Hajin Shim, and June Yong Yang, with whom I collaborated on our research journey. Their effective yet smooth mentoring and collaboration not only substantially improved my research skills but also fueled my ambition to become a doctoral student like them. Additionally, I express my thanks to Taewon Kim, Hyunjin Seo, and Seungyeon Woo, my research collaborators, for the enjoyable research experience. Lastly, I extend my gratitude to my master's peers, Seungyeon Woo, Kyusung Seo, and Sohee Kim, for their mutual support and reliance during challenging periods.

I want to extend my gratitude to my middle school friends, Donggeun Koo, Younghyeok Kim, Jaeuk Jang, and Sangyeong Byun, as well as my high school friends, Jinhyeon Jo, Lael Shin, Jaemin Cho, and Minsung Kim, for their companionship and for making my time enjoyable. I must also express my thanks to all my friends from the acoustic music club Sixline and others, including Jungwoo Kim and Hyunjoon Cho, whom I met during my university years for the same reasons. Furthermore, I extend my gratitude to Matthew Kwon, who guided me from my solitary life in Seongnam to Bundang Woori Church. I would like to express my thanks to everyone I encountered at Bundang Woori Church, especially the members of Matthew's middle group and the praise team. I also express my sincere gratitude to everyone who prayed for me, even though I may not be aware of all of them. I am grateful to you, even though your name was not mentioned, so please do not be too disappointed.

Last but certainly not least, I want to extend words of love to my mother Yihyang Seo, and my father Woan Soo Kim. They have nurtured me up to this point, loved me immensely, always believed in me, supported me, and have consistently been on my side. Finally, I express my gratitude to Jesus Christ, who saved me, became the master of my life, and always guides me.

Curriculum Vitae

Name	:	Changhun Kim
Date of Birth	:	March 20, 1998
E-mail	:	changhun.kim@kaist.ac.kr
Links	:	Homepage, Google Scholar, GitHub, LinkedIn, X

Education

2022. 03. – 2024. 02.	M.S. in Artificial Intelligence, Korea Advanced Institute of Science and Technology (KAIST)
2017. 03. – 2022. 02.	B.S. in Computer Science and Mathematics, Korea Advanced Institute of Science and Technology (KAIST)
2014. 03. – 2017. 02.	High School Diploma, Busan Science High School

Experience

2023. 11. – 2024. 02.	Machine Learning Researcher Intern, Medical AI Division, AITRICS
2021. 06. – 2022. 02.	Research Intern, Machine Learning and Intelligence Laboratory, KAIST
2020. 09. – 2021. 02.	Machine Learning Engineer Intern, MLOps Squad, DeepNatural AI
2019. 10. – 2020. 08.	Research Intern, Vehicular Intelligence Laboratory, KAIST
2019. 06. – 2019. 08.	Data Engineer Intern, Big Data Center, Netmarble

Publications (*: Equal contribution)

1. **Changhun Kim***, Taewon Kim*, and Eunho Yang. Test-Time Adaptation for Time Series Analysis. In Preparation.
2. Taewon Kim*, **Changhun Kim***, Gyeongman Kim, and Eunho Yang. Test-Time Adaptation for Vision-Language Models. In Preparation.
3. Hajin Shim*, **Changhun Kim***, and Eunho Yang. CloudFixer: Test-Time Adaptation for 3D Point Clouds via Diffusion-Guided Domain Translation. Under Review.
4. **Changhun Kim***, Taewon Kim*, Seungyeon Woo, June Yong Yang, and Eunho Yang. AdapTable: Test-Time Adaptation for Tabular Data via Shift-Aware Uncertainty Calibrator and Label Distribution Handler. Under Review.
5. **Changhun Kim**, Joonhyung Park, Hajin Shim, and Eunho Yang. SGEM: Test-Time Adaptation for Automatic Speech Recognition via Sequential-Level Generalized Entropy Minimization. In *INTERSPEECH*, 2023. **(Oral Presentation, 348/2293=15.18%)**