Interspeech 2024
1-5 September 2024, Kos, Greece

LI-TTA: Language Informed Test-Time Adaptation for Automatic Speech Recognition

Eunseop Yoon[1*], *Hee Suk Yoon*[1*], *John Harvill*[2], *Mark Hasegawa-Johnson*[2], *Chang D. Yoo*[1†]

[1]Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea
[2]University of Illinois at Urbana-Champaign, USA

{esyoon97, hskyoon, cd_yoo}@kaist.ac.kr, {harvill2, jhasegaw}@illinois.edu

Abstract

Test-Time Adaptation (TTA) has emerged as a crucial solution to the domain shift challenge, wherein the target environment diverges from the original training environment. A prime exemplification is TTA for Automatic Speech Recognition (ASR), which enhances model performance by leveraging output prediction entropy minimization as a self-supervision signal. However, a key limitation of this self-supervision lies in its primary focus on acoustic features, with minimal attention to the linguistic properties of the input. To address this gap, we propose Language Informed Test-Time Adaptation (LI-TTA), which incorporates linguistic insights during TTA for ASR. LI-TTA integrates corrections from an external language model to merge linguistic with acoustic information by minimizing the CTC loss from the correction alongside the standard TTA loss. With extensive experiments, we show that LI-TTA effectively improves the performance of TTA for ASR in various distribution shift situations.

Index Terms: automatic speech recognition, test-time adaptation, linguistic

1. Introduction

Automatic Speech Recognition (ASR) systems have seen remarkable advancements through the development of pre-trained models. These pre-trained models have significantly improved ASR performance by leveraging vast amounts of data to capture complex patterns in speech. However, in real-world applications, ASR models frequently encounter inputs with distributions that differ from their training datasets, leading to performance degradation. This issue is exacerbated by the practical impossibility and computational expense of continuously collecting and training on new data for every possible distribution shift encountered in real-world scenarios.

To address this challenge, Test-Time Adaptation (TTA) has emerged as a promising solution, drawing considerable attention for its ability to adapt models at inference time without the need for additional training data. Drawing inspiration from Test-Time Adaptation (TTA) strategies employed in computer vision [1, 2, 3, 4, 5], recent methods [6, 7] have demonstrated noteworthy effectiveness in adapting pre-trained ASR models [8, 9, 10] by utilizing entropy minimization as a self-supervision signal.

A significant limitation of the self-supervision signals in ASR models is their excessive focus on acoustic features, frequently neglecting the linguistic aspects of spoken language. While external language models are commonly employed during the beam search decoding phase to process logits from ASR models, no feedback from these language models can backpropagate through the ASR system to aid the TTA process. For example, Figure 1 illustrates failure cases of traditional TTA methods where the absence of linguistic feedback leads to two main issues: phonetically similar but contextually inappropriate predictions remain uncorrected (Figure 1-(a)), and in some instances, initially correct predictions are erroneously altered to phonetically similar words that do not suit the context of the sentence (Figure 1-(b)). These examples underscore the necessity for integrating linguistic information into the TTA process to enhance both the accuracy and contextual relevance of ASR predictions.



Figure 1: ***Examples of failure cases with traditional TTA methods in ASR due to the absence of linguistic feedback.*** *(a) Predictions with phonetically similar words persist incorrectly, misaligning with the sentence context. (b) Additionally, some accurate predictions are incorrectly replaced with contextually unfit words of similar phonetic structure.*

In response to this challenge, this paper presents Language Informed Test-Time Adaptation (LI-TTA) which integrates linguistic guidance into the TTA process for ASR systems. By employing corrections derived from an external instructed-tuned language model, LI-TTA effectively combines linguistic and acoustic information. This integration is achieved through the joint minimization of the Connectionist Temporal Classification (CTC) [11] loss, resulting from these corrections, and the conventional TTA loss. Our comprehensive evaluations across several domain shift benchmarks demonstrate the efficacy of LI-TTA: it not only enhances the precision of ASR-generated transcriptions, as evidenced by lower word error rates but also ensures their contextual fidelity, as indicated by reduced perplexity scores.

2. Related Work

2.1. Distribution shift on Automatic Speech Recognition

Recent advancements in Automatic Speech Recognition (ASR) models have achieved surprising performance on in-distribution

* Equal Contribution
† Corresponding Author

10.21437/Interspeech.2024-1829

training data, primarily attributed to the methodology of self-supervised pre-training followed by supervised fine-tuning [12, 8, 9, 13]. During the pre-training phase, models comprehensively understand acoustic features from audio inputs without necessitating labeled data. This foundational knowledge facilitates the subsequent fine-tuning stage, wherein models are explicitly trained to correlate audio inputs with their corresponding transcription texts. Despite these advancements, the inherent variability of speech recognition—attributable to diverse contexts, speaker characteristics, and background noises—presents significant challenges. Models frequently encounter samples from distributions unseen during training, leading to degraded performance in real-world applications.

Addressing the issue of distribution shift, [14, 15, 16, 17] proposed a strategy of fine-tuning ASR models using target domain training data. Although this approach demonstrates effectiveness in aligning the model with the target domain, it is computationally inefficient. In response, [18] introduced a prompt tuning approach to adapt the model to the target domain, which not only enhances model performance in new domains but also preserves efficacy in the source domain. A commonality among these strategies is their reliance on domain-specific training data and the necessity of a training process to achieve improvements in the target domain.

In light of these limitations, the focus has recently shifted towards test-time adaptation to optimize ASR models for target domains without the need for extensive training data or following training procedures. This approach necessitates only minimal adaptation steps during inference to achieve significant improvements [6, 7].

2.2. Test-time adaptation

Test-Time Adaptation (TTA) enhances deep learning model performance in new domains by adapting the model during inference without requiring labeled training data. In the field of computer vision, strategies such as entropy minimization [1], batch normalization [19], and data augmentation [3] have been proposed, demonstrating effectiveness in image classification. Recent research has also extended TTA to vision-language models, highlighting the approach's versatility and potential across different domains [4, 5]. Despite the growing attention of TTA research in computer vision, its application within the speech domain remains comparatively sparse.

To address this gap, [6] pioneered the application of TTA to Automatic Speech Recognition (ASR), utilizing a single-utterance adaptation method. Further enhancing this approach, [7] introduced a beam search strategy to improve performance during adaptation. Nevertheless, the principal drawback of these methods is their overwhelming emphasis on acoustic aspects, largely overlooking the linguistic dimensions of the input. In response to this limitation, we introduce Language Informed Test-Time Adaptation (LI-TTA), a method that integrates linguistic attributes into the TTA framework for ASR.

3. Method

In this section, we present Language Informed Test-Time Adaptation (LI-TTA), a novel approach that enhances Test-Time Adaptation (TTA) for Automatic Speech Recognition (ASR) by integrating linguistic insights from an external language model. We begin with Section 3.1, where the setup for TTA in ASR contexts is outlined. In Section 3.2, we show that the existing TTA framework fails to uphold linguistic knowledge, although

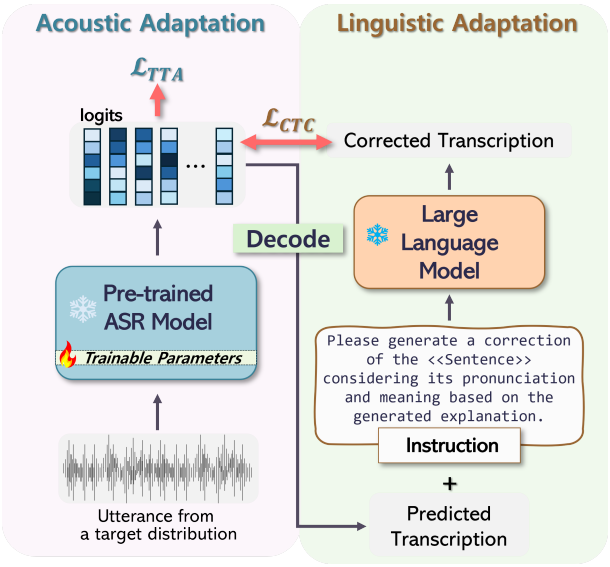


Figure 2: ***An overview of our proposed Language Informed Test-Time Adaptation (LI-TTA).*** *LI-TTA integrates corrections from an external instruction-tuned language model to merge linguistic with acoustic information by minimizing the CTC loss from the correction alongside the standard TTA loss.*

using a language model during the decoding phase. Following this, in Section 3.3, we provide a comprehensive explanation of our proposed LI-TTA.

3.1. Preliminary

General Automatic Speech Recognition (ASR) Consider an ASR model, denoted by $f(\cdot|\theta)$, which is trained using labeled data from a source domain, represented as $\mathcal{D}_{\text{source}} = \{(x_i^{\text{source}}, \mathbf{y}_i^{\text{source}})\}_{i=1}^{N}$. This model processes a raw audio waveform x, producing logits $f(x|\theta) \in \mathbb{R}^{L\times C}$ across L timesteps for a vocabulary $\mathcal{V}$ of size C, where θ denotes the model's parameters. The model aims to calculate the joint log probability, $\log p(\mathbf{y}|x,\theta)$, for a potential transcription $\mathbf{y} = (y_{(1)}, \ldots, y_{(L)})$ using an autoregressive approach:

$$\log p(\mathbf{y}|x,\theta) = \log p_{\text{AM}}(\mathbf{y}|x,\theta) + \lambda \log p_{\text{LM}}(\mathbf{y}) \quad (1)$$
$$= \sum_{t=1}^{L} [\log p_{\text{AM}}(y_{(t)}|y_{(<t)}, x, \theta) + \lambda \log p_{\text{LM}}(y_{(t)}|y_{(<t)})],$$

where $y_{(t)} \in \mathcal{V}$. Here, $p_{\text{AM}}(\mathbf{y}|x,\theta)$ refers to the probability as predicted by the ASR model for the input x, whereas $p_{\text{LM}}(\mathbf{y})$ represents the likelihood according to an autoregressive language model (LM). The parameter λ serves to balance the influence of the LM. *It is important to note that this LM is utilized solely during the decoding stage of inference to guide the selection of the optimal solution, denoted as* $\mathbf{y}^*$.

Test-Time Adaptation (TTA) for ASR In the context of TTA for ASR, the objective is to tailor the model $f(\cdot|\theta)$ to operate optimally within the domain of unlabeled target speech, denoted by $\mathcal{D}_{\text{target}} = \{x_i^{\text{target}}\}_{i=1}^{M}$. This adjustment process particularly focuses on a scenario termed as the *single-utterance TTA* approach [6, 7], where individual utterance x_i^{target} from $\mathcal{D}_{\text{target}}$ is used to fine-tune $f(\cdot|\theta)$. The aim is to refine the model's predictions, achieving enhanced prediction logits $\log p(\mathbf{y}|x_i^{\text{target}}, \theta)$

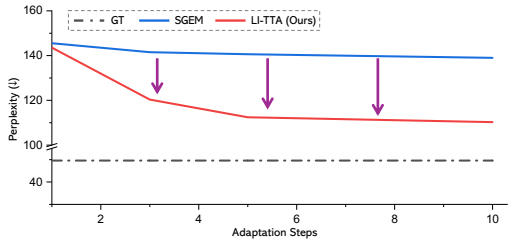


Figure 3: ***The trend of perplexity (PPL) with respect to the TTA adaptation steps***. *Traditional TTA approach (SGEM [7]) shows only marginal reductions in perplexity due to the absence of linguistic feedback during the adaptation process. Conversely, our proposed LI-TTA demonstrates a significant decrease in perplexity, benefiting from the integration of linguistic insights.*

by entropy minimization. Since p_{LM} is independent of θ, the adaptation process is influenced solely by p_{AM}, highlighting the model's reliance on acoustic modeling adjustments during TTA.

3.2. Lack of Linguistic Supervision of Existing TTA

As discussed in Section 3.1, *while external language models play a crucial role during the beam search decoding phase of ASR models by processing logits, they are unable to provide feedback that can backpropagate through the ASR system to enhance the TTA process.* This constraint is emphasized in Figure 3, where sentence semantic coherence and grammatical accuracy are evaluated using perplexity (PPL) scores from an advanced external LM (GPT-2 [20]). Lower PPL scores reflect the greater confidence of LM in the grammatical and semantic integrity of the sentence, whereas higher scores indicate potential issues in terms of the coherence and precision of the sentence. Our observations reveal that increasing the number of adaptation steps does not necessarily lead to a reduction in PPL scores for the previous TTA method (SGEM [7]). This leads us to propose LI-TTA in the subsequent section, aiming to provide more effective linguistic feedback during TTA, thereby achieving PPL scores that closely align with the ground truth (GT).

3.3. LI-TTA: Language Informed TTA for ASR

This section presents LI-TTA, depicted in Figure 2, which leverages an external instruction-tuned language model to refine ASR model predictions within TTA framework through CTC loss minimization. In this process, an ASR model represented as $f(\cdot|\theta)$, generates an initial prediction $\mathbf{y}$ and decoded transcription $\mathbf{y}'$. Subsequently, this transcription is fed to an external instruction-tuned LM ($\text{LM}_{\text{external}}$)[1], which, based on a specific set of instructions, produces a corrected version $\tilde{\mathbf{y}}$:

$$\tilde{\mathbf{y}} = \text{LM}_{\text{external}}(\text{Instruction}, \mathbf{y}'). \tag{2}$$

Given that the original prediction $\mathbf{y}$ and the corrected transcription $\tilde{\mathbf{y}}$ might not align perfectly, we employ the CTC loss [11] as a means of linguistic feedback to enhance the ASR model. The finetuning of $f(\cdot|\theta)$ is thus guided by a composite loss function:

$$\mathcal{L} = \mathcal{L}_{\text{TTA}} + \lambda_{\text{LI}} \cdot \mathcal{L}_{\text{CTC}}(\tilde{\mathbf{y}}, \mathbf{y}), \tag{3}$$

where $\mathcal{L}_{\text{TTA}}$ represents the original TTA loss [7] and λ_{LI} is a weighting factor that adjusts the influence of the linguistic feedback on the overall loss. Such an approach allows linguistic information to guide the TTA process, thereby allowing the predicted transcription to have a better coherent sentence structure.

[1]Note that this language model is different from the one used during the beam search decoding phase explained in Section 3.1.

4. Experimental Setting

Model For the evaluation of LI-TTA we use the wav2vec 2.0 model[2] [8], which was trained on the LibriSpeech dataset [21]. For the language model used during the beam search decoding phase, we utilize the external 4-gram language model[3]. For the external instruction-tuned language model ($\text{LM}_{\text{external}}$), we use OpenChat 3.5[4].

Datasets To assess LI-TTA's adaptability across different domain shifts, we conducted tests on a variety of datasets. For evaluating performance against unseen speakers and vocabulary, we use the test set from TED-LIUM 2 (TD) [22], Common Voice (CV), and [23]. Furthermore, to examine LI-TTA's robustness against background noise, we introduced eight noise conditions (air conditioner (AC), airport announcement (AA), babble (BA), copy machine (CM), munching (MU), neighbors (NB), shutting door (SD), and typing (TP)) at an SNR of 10dB into the LibriSpeech test-other dataset [21]. These noise samples were drawn from the MS-SNSD noise dataset [24]. Lastly, the L2-Artic corpus [25], consisting of non-native English speakers, served to test LI-TTA's efficacy in handling pronounced accent and pronunciation variations, with a focus on a single speaker from each first language category. To underscore the effectiveness of LI-TTA in ensuring that transcriptions accurately reflect the spoken content, our evaluation prioritizes test sets where the ground truth (GT) perplexity, as measured by GPT-2, falls below 70. This criterion helps to selectively focus on those datasets that present meaningful sentence contexts, efficiently filtering out test sets characterized by a lack of coherent sentence structure.

Implementation For the hyperparameters we follow [7] which optimized hyperparameters on CHiME-3 [26]. Specifically, we use the AdamW optimizer [27], coupled with a cosine annealing learning rate scheduler, setting the initial and final learning rates to $4 \cdot 10^{-5}$ and $2 \cdot 10^{-5}$, respectively. The number of TTA adaptation steps is set to 10, applying SGEM [7] to compute $\mathcal{L}_{\text{TTA}}$. The decoding stage hyperparameter λ is set to 0.3. For λ_{LI}, we use an adaptive strategy, setting it as $\frac{\mathcal{L}_{\text{TTA}}}{\mathcal{L}_{\text{TTA}}+\mathcal{L}_{\text{CTC}}}$. The external language model ($\text{LM}_{\text{external}}$) receives instructions formulated as `"Please generate a correction of the <<sentence>> considering the pronunciation and overall context"`. We only train feature extractors of the ASR model. All experiments are conducted on NVIDIA A100 80GB PCIe.

5. Result

5.1. Main Results

We compare the TTA performance of the ASR model across 11 datasets with various domain shifts. The results, detailed in Table 1, include both the word error rate (WER) and perplexity (PPL) metrics for the ASR model's outputs, with PPL scores derived using the GPT-2 [20] model. Notably, the incorporation of LI-TTA into the ASR models consistently yields the lowest

[2]https://huggingface.co/facebook/wav2vec2-base-960h
[3]https://huggingface.co/patrickvonplaten/wav2vec2-base-100h-with-lm
[4]https://huggingface.co/openchat/openchat_3.5

Table 1: *Comparison of TTA performance, measured using word error rate (WER) (↓) and perplexity (PPL) (↓), across 11 datasets with various types of domain shifts. The results are obtained using greedy decoding for inference. Bold represents the best scores.*

Method		TD	CV	AC	AA	BA	CM	MU	NB	SD	TP	VC	Avg.
Unadapted	WER	11.72	26.08	21.17	32.45	60.05	42.38	45.33	108.81	14.07	18.30	51.26	39.24
	PPL	485.66	689.39	413.56	554.97	1101.00	1086.52	942.13	1333.51	212.45	290.10	1490.31	781.78
SUTA [6]	WER	10.53	21.97	13.85	31.53	49.82	33.29	34.61	102.59	10.66	12.01	36.49	32.49
	PPL	358.67	447.86	545.96	512.23	1029.24	625.73	1358.04	1008.84	169.20	127.34	1138.00	665.56
SGEM [7]	WER	10.33	21.31	12.53	32.45	48.20	32.44	33.76	101.0	10.46	11.72	34.11	31.66
	PPL	297.37	471.61	630.62	507.67	1034.70	564.48	1468.38	854.44	128.27	135.23	1284.91	670.70
LI-TTA (ours)	WER	**10.30**	**20.35**	**12.16**	**24.55**	**46.55**	**31.88**	**32.31**	**99.58**	**10.44**	**11.33**	**33.23**	**30.24**
	PPL	**271.18**	**389.76**	**412.74**	**444.65**	**803.42**	**563.37**	**1262.94**	**812.31**	**122.71**	**120.49**	**1098.82**	**572.94**

Table 2: *Comparison of TTA performance, measured using word error rate (WER) (↓) and perplexity (PPL) (↓) on six non-native English speech corpora.*

Setting		Unadapted	SUTA [6]	SGEM [7]	LI-TTA
Arabic	WER	13.98	12.11	11.93	**10.42**
	PPL	412.75	363.78	358.28	**317.56**
Mandarin	WER	14.28	12.81	12.42	**10.69**
	PPL	219.38	168.14	161.59	**141.78**
Hindi	WER	11.54	10.40	9.86	**9.08**
	PPL	190.16	182.38	177.24	**146.52**
Korean	WER	10.65	8.96	8.21	**7.86**
	PPL	210.96	159.68	182.00	**143.69**
Spanish	WER	19.02	12.68	13.02	**11.73**
	PPL	529.71	202.24	252.68	**159.14**
Vietnamese	WER	34.92	30.83	31.06	**28.80**
	PPL	988.28	733.11	651.78	**591.65**
Average	WER	17.40	14.63	14.41	**13.10**
	PPL	425.21	301.55	297.26	**250.06**

word error rate (WER) of target utterances with an average of 30.24%. Furthermore, LI-TTA also achieves the lowest perplexity (PPL) across all 11 datasets, with an average of 572.94. This indicates the efficacy of our unsupervised objectives, which significantly refine the accuracy of ASR-generated transcriptions—as evidenced by reduced WER—and ensure their contextual fidelity, as demonstrated by lower PPL scores.

5.2. Non-Native English Speech Corpora

To further demonstrate the effectiveness of LI-TTA across domain shifts, we conducted evaluations on six distinct corpora comprising non-native English speech. The results in Table 2 show that LI-TTA consistently surpasses the baselines across all tested corpora in terms of both WER and PPL, highlighting its robustness to significant pronunciation and accent variations.

5.3. Qualitative Results

In Figure 4, we present a qualitative comparison between the conventional TTA method, SGEM [7], and our proposed LI-TTA through specific examples that highlight their prediction differences. These examples illustrate instances where SGEM fails to correctly interpret the context of the utterance, leading to erroneous transcriptions. Conversely, LI-TTA demonstrates its adeptness at navigating such complexities, accurately capturing the intended context, and producing contextually appropriate transcriptions. This distinction highlights the advantage of incorporating linguistic knowledge into the TTA process and underscores the inadequacy of methods like SGEM that depend exclusively on acoustic features.

Ground Truth Transcription	The Russian music **player** the count was her obedient slave
SGEM	The Russian music flayer the count was her obedient slave
LI-TTA (Ours)	The Russian music player the count was her obedient slave
Ground Truth Transcription	He pulled and the **log** crashed down to break his back
SGEM	He pulled and the lok crashed down to break his back
LI-TTA (Ours)	He pulled and the log crashed down to break his back
Ground Truth Transcription	I may manage to freight a **cargo** back as well
SGEM	I may manage to freight a car go back as well
LI-TTA (Ours)	I may manage to freight a cargo back as well
Ground Truth Transcription	She was **sleeping** under his protection as sweetly as a child
SGEM	She was leaping under his protection as sweetly as a child
LI-TTA (Ours)	She was sleeping under his protection as sweetly as a child

Figure 4: ***Examples of failure cases of previous TTA method (SGEM [7]) and the prediction from our proposed LI-TTA with contextually correct transcriptions.***

6. Conclusions

In this study, we tackled the domain shift challenge in ASR by proposing Language Informed Test-Time Adaptation (LI-TTA), a novel method that enriches Test-Time Adaptation (TTA) with linguistic insights. By integrating corrections from an external language model, LI-TTA not only mitigates the limitations of traditional TTA approaches that focus predominantly on acoustic features but also significantly enhances ASR accuracy and contextual correctness. Our comprehensive evaluations across diverse domain shifts show the effectiveness of LI-TTA, evidenced by improved word error rates and perplexity scores.

7. Limitation

In this paper, we show the effectiveness of LI-TTA on a CTC-based ASR model. Future work can extend the validation of LI-TTA's efficacy across diverse ASR architectures, including conformer-based and transducer-based models, to ascertain its adaptability in a broader array of speech recognition systems.

8. Acknowledgements

This work was partially supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2021-0-01381, Development of Causal AI through Video Understanding and Reinforcement Learning, and Its Applications to Real Environments) and SAMSUNG Research, Samsung Electronics Co.,Ltd.

9. References

[1] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=uXl3bZLkr3c

[2] P. T. S and F. Fleuret, "Test time adaptation through perturbation robustness," in *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021. [Online]. Available: https://openreview.net/forum?id=GbBeI5z86uD

[3] M. Zhang, S. Levine, and C. Finn, "Memo: Test time robustness via adaptation and augmentation," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 38 629–38 642. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/fc28053a08f59fccb48b11f2e31e81c7-Paper-Conference.pdf

[4] S. Manli, N. Weili, H. De-An, Y. Zhiding, G. Tom, A. Anima, and X. Chaowei, "Test-time prompt tuning for zero-shot generalization in vision-language models," in *NeurIPS*, 2022.

[5] H. S. Yoon, E. Yoon, J. T. J. Tee, M. A. Hasegawa-Johnson, Y. Li, and C. D. Yoo, "C-TPT: Calibrated test-time prompt tuning for vision-language models via text feature dispersion," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=jzzEHTBFOT

[6] G.-T. Lin, S.-W. Li, and H.-y. Lee, "Listen, adapt, better wer: Source-free single-utterance test-time adaptation for automatic speech recognition," 2022.

[7] C. Kim, J. Park, H. Shim, and E. Yang, "SGEM: Test-time adaptation for automatic speech recognition via sequential-level generalized entropy minimization," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2023.

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[9] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," 2020.

[10] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[11] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 369–376. [Online]. Available: https://doi.org/10.1145/1143844.1143891

[12] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *Interspeech 2019*, 2019.

[13] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[14] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation," in *2017 IEEE automatic speech recognition and understanding workshop (ASRU)*. IEEE, 2017, pp. 16–23.

[15] W. Hou, J. Wang, X. Tan, T. Qin, and T. Shinozaki, "Cross-domain speech recognition with unsupervised character-level distribution matching," 2021.

[16] G. I. Winata, Z. Lin, and P. Fung, "Learning multilingual meta-embeddings for code-switching named entity recognition," in *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, I. Augenstein, S. Gella, S. Ruder, K. Kann, B. Can, J. Welbl, A. Conneau, X. Ren, and M. Rei, Eds. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 181–186. [Online]. Available: https://aclanthology.org/W19-4320

[17] T. Shibano, X. Zhang, M. T. Li, H. Cho, P. Sullivan, and M. Abdul-Mageed, "Speech technology for everyone: Automatic speech recognition for non-native English," in *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, M. Abbas and A. A. Freihat, Eds. Trento, Italy: Association for Computational Linguistics, 12–13 Nov. 2021, pp. 11–20. [Online]. Available: https://aclanthology.org/2021.icnlsp-1.2

[18] E. Yoon, H. S. Yoon, J. Harvill, M. Hasegawa-Johnson, and C. Yoo, "INTapt: Information-theoretic adversarial prompt tuning for enhanced non-native speech recognition," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 9893–9902. [Online]. Available: https://aclanthology.org/2023.findings-acl.627

[19] H. Lim, B. Kim, J. Choo, and S. Choi, "Ttn: A domain-shift aware batch normalization in test-time adaptation," in *The Eleventh International Conference on Learning Representations*, 2022.

[20] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[22] A. Rousseau, P. Deléglise, and Y. Estève, "Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3935–3939. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1104_Paper.pdf

[23] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222. [Online]. Available: https://aclanthology.org/2020.lrec-1.520

[24] C. K. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, "A scalable noisy speech dataset and online subjective test framework," *Proc. Interspeech 2019*, pp. 1816–1820, 2019.

[25] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-ARCTIC: A Non-native English Speech Corpus," in *Proc. Interspeech 2018*, 2018, pp. 2783–2787.

[26] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 504–511.

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7