Domain-Specific Adaptation for ASR through Text-Only Fine-Tuning

Betty Kurian
Harman International
Bangalore, India
betty.kurian@harman.com

Abhinav Upadhyay
Harman International
Bangalore, India
abhinav.upadhyay@harman.com

Abhijeet Sengupta
Harman International
Bangalore, India
abhijeet.sengupta@harman.com

Abstract

Speech recognition models often struggle in specialized domains due to the lack of domain-specific paired audio-text data, making it difficult to adapt general-purpose systems to unique terminology and linguistic patterns. In this work, we propose a text-only domain adaptation method for Whisper, fine-tuning only the decoder using domain-relevant text. Our approach introduces trainable cross-attention bias embeddings, extended with a gated mixture-of-experts routing mechanism, enabling the model to encode domain-specific linguistic priors without any audio data. Unlike ASR adaptation methods that require paired audio-text datasets, our approach is lightweight and resource-efficient. We observe up to a 56% relative improvement in word error rate over the baseline. Our findings demonstrate that text-only adaptation is a practical and effective strategy for improving speech recognition in specialized domains with limited or no domain-specific audio.

1 Introduction

Speech recognition technology has advanced significantly in recent years, with applications in virtual assistants, transcription services, and real-time communication systems. These improvements have been driven by supervised learning approaches that rely on paired audio-text datasets to train models capable of mapping language to text Watanabe et al. (2017). Such datasets enable models to learn the complex relationships between speech signals and their textual representations, resulting in robust general-purpose Automatic Speech Recognition (ASR) systems. However, achieving high accuracy in specialized domains remains challenging. Domain-specific ASR systems must address unique linguistic patterns, specialized terminology, and the limited availability of paired audio-text data Bataev et al. (2023). In domains such as healthcare, legal, or scientific research, the limited availability of annotated domain audio constrains the adaptation of general-purpose models, highlighting the need for approaches that reduce reliance on domain-specific audio resources.

To address these challenges, researchers have investigated integrating ASR systems with language models (LMs) through shallow and deep fusion Gulcehre et al. (2015), as well as generating synthetic domain audio using text-to-speech (TTS) systems Huang et al. (2020). Shallow fusion can improve recognition accuracy but requires an external LM during inference, which increases computational cost and latency. Deep fusion incorporates the LM within the ASR training process, but this often demands substantial computational resources and careful tuning to prevent overfitting. TTS-based augmentation provides a way to create domain-specific audio from text, yet the generated speech may contain artifacts and fail to replicate the prosody and acoustic variability of natural speech, limiting its effectiveness for adaptation.

In this work, we propose a text-only domain adaptation method for Whisper, fine-tuning only the decoder using domain-relevant textual corpora. Our approach introduces trainable cross-attention bias embeddings, extended with a gated mixture-of-experts routing mechanism, enabling the model to encode domain-specific linguistic priors without any audio data. This eliminates the dependence on paired domain audio while offering a lightweight and resource-efficient adaptation strategy. We observe up to a 56% relative reduction in word error rate compared to the baseline. These results demonstrate that text-only fine-tuning is a practical and effective approach for improving ASR performance in specialized domains.

The remainder of this paper is organized as follows: Section 2 discusses related work, Section 3 presents the proposed approach, Section 4 details the evaluation and results, and Section 5 concludes

the paper.

2 Related Work

Traditional ASR systems rely heavily on handcrafted features and statistical models, involving multiple stages such as acoustic modeling, phoneme recognition, and language modeling Bell et al. (2020). Recent advances in deep learning, particularly Transformer-based architectures Vaswani (2017), have enabled end-to-end models that map audio directly to text, simplifying the pipeline and achieving state-of-the-art performance. However, adapting these models to new domains remains challenging due to the need for large amounts of labeled audio data, motivating research into more efficient domain adaptation techniques Bell et al. (2020).

A common strategy for domain adaptation is to use text-to-speech (TTS) to synthesize paired speech-text data from target-domain text for fine-tuning ASR models Huang et al. (2020). While effective, this process requires training high-quality multi-speaker TTS models, which is computationally expensive Zheng et al. (2021). To reduce this cost, text-to-spectrogram approaches generate synthetic spectrograms directly from text, removing the need for TTS and audio storage while minimizing the mismatch between synthetic and real audio Bataev et al. (2023). This approach still requires careful training of the spectrogram generator to ensure quality.

Text-only adaptation methods offer a more cost-efficient alternative. These include fine-tuning external language models on target-domain text and integrating them into ASR decoding via shallow fusion Kannan et al. (2018). Contextual biasing methods embed domain-specific phrases to improve recognition of rare terms Aleksic et al. (2015); Chang et al. (2023), while prompt-based techniques condition the ASR model on additional domain cues to guide transcription Suh et al. (2024). Another approach uses pseudo-audio embeddings as prompts for fine-tuning Ma et al. (2024), allowing adaptation without paired data. Tran et al. Tran et al. (2025) propose DAS, a domain adaptation framework that generates domain-specific synthetic speech from LLM-produced text and fine-tunes Whisper with LoRA adapters.

Our approach adapts the ASR model to new domains using text-only fine-tuning, without relying on synthetic audio generation, prompt-based conditioning, or external rescoring. This design reduces computational cost, lowers latency, and simplifies deployment, while enhancing recognition of domain-specific vocabulary. Our method directly adapts the model's language understanding capabilities using only textual data, making it practical for scenarios where domain audio is limited or unavailable.

3 Approach

Our approach builds on Whisper, a state-of-the-art ASR model developed by OpenAI Radford et al. (2023). Whisper employs a Transformer-based encoder-decoder architecture, where the encoder processes audio inputs into latent representations, and the decoder generates transcriptions by attending to both the encoder's output and prior textual context. The encoder captures acoustic features such as phonemes, pitch, and rhythm, while the decoder aligns these features with linguistic patterns to produce accurate transcriptions. Whisper's decoder accepts input sequences, enabling the model to incorporate textual descriptions or prompts as part of the input. This feature allows Whisper to condition its transcription generation on additional context, such as domain-specific instructions or metadata. We leverage this capability for domain adaptation by modifying the architecture to focus only on the decoder, bypassing the encoder. This enables adaptation using text-only data without requiring paired audio-text inputs.

The encoder in Whisper generates contextualized representations of the input audio, which are passed to the decoder for processing via the cross-attention mechanism. During cross-attention, the decoder queries the encoder outputs using keys (K) and values (V), where K represents the contextualized embeddings generated by the encoder, and V serves as the basis for computing attention-weighted outputs that guide the decoder's predictions. We freeze the encoder during training, but the decoder still requires valid K and V representations for the cross-attention mechanism to function correctly, even though the encoder's outputs are no longer updated. To address this, we replace the encoder's output in the cross-attention mechanism with trainable biases B. The bias embeddings B denoted as $R^{N \times d}$, where N is the bias sequence length (representing the number of tokens) and d is the embedding dimension, which matches the output dimension of the frozen encoder. These biases

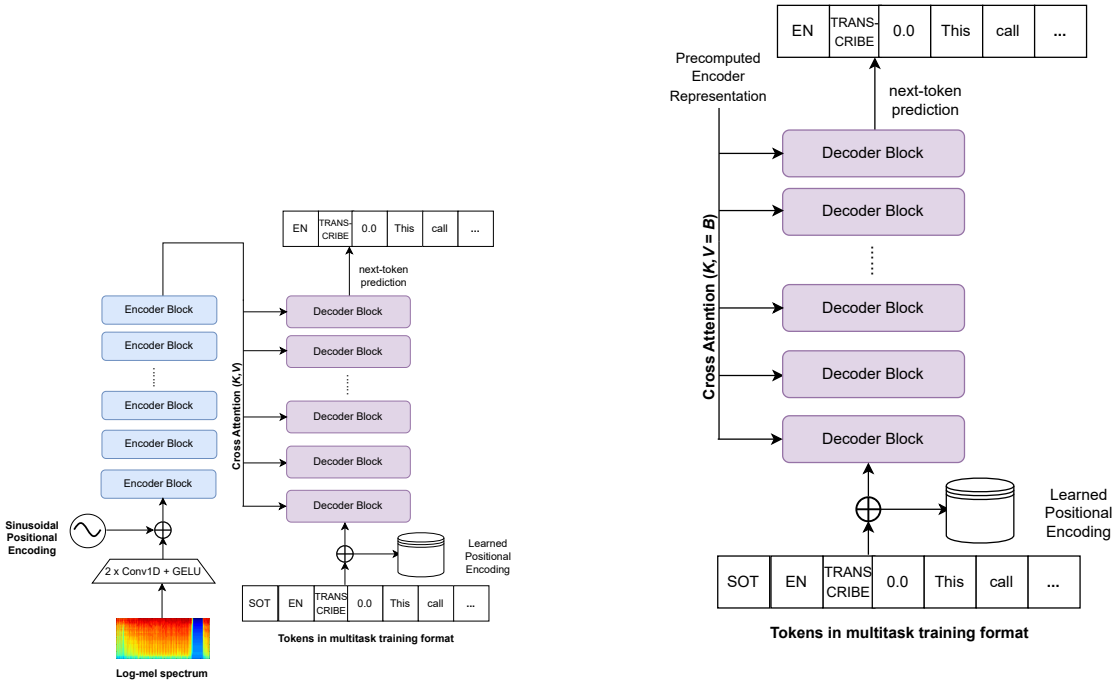

Figure 1: Figure (a) shows the Whisper-base encoder-decoder architecture. Figure (b) shows the modified architecture with domain-specific bias adapters for text-only adaptation (our approach), where multiple expert bias matrices are introduced into the decoder to incorporate domain-specific linguistic priors and guide transcription without using paired audio.

serve as trainable substitutes for the encoder's representations, allowing the decoder to focus entirely on linguistic patterns while maintaining structural compatibility with the original architecture Suh et al. (2024).

We compute cross-attention as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

We replace the encoder outputs with the bias embeddings as K and V in the decoder's cross-attention layers. Substituting K and V with B, the cross-attention becomes:

$$\text{Cross-Attention} = \text{Attention}(Q, \mathbf{B}, \mathbf{B}) \quad (2)$$

To further refine the contribution of bias embeddings, we introduce a $\tanh$ gating mechanism:

$$G = \tanh(W_g \cdot B) \quad (3)$$

where W_g represents a learnable weight matrix that modulates bias embeddings.

We initialize the bias embeddings B using precomputed representations from the pretrained Whisper encoder.

$$B = E_{pretrained} \quad (4)$$

where $E_{pretrained}$ represents the fixed output of the pretrained Whisper encoder. These bias embeddings are then made trainable during fine-tuning, allowing the model to adapt them for domain-specific text representations. This initialization ensures that the model starts with relevant embeddings while retaining the flexibility to refine them through backpropagation. The trainable biases introduced into the cross-attention layers implicitly capture domain-relevant features during training, allowing the decoder to operate effectively in a text-only setting.

Gated Routing for Multi-Domain Adaptation: To handle multiple domain subspaces, we extend the bias embedding design into a *mixture-of-experts* (MoE) framework. Instead of a single bias matrix B, we maintain a set of M expert bias matrices $\{\mathbf{B}_m\}_{m=1}^{M}$, each representing domain-specific linguistic priors. A lightweight routing network com-

putes mixture weights $\pi \in R^M$ conditioned on the current decoding context:

$$\pi = \text{softmax}(W_r, \phi(y_{<t})), \tag{5}$$

where $\phi(y_{<t})$ encodes the partial transcription history and W_r is a learnable projection. The aggregated bias is then:

$$\mathbf{B}^* = \sum_{m=1}^{M} \pi_m, \mathbf{B}_m. \tag{6}$$

This $\mathbf{B}^*$ replaces B in the cross-attention mechanism, enabling the decoder to dynamically route attention toward the most relevant domain priors. This structure not only improves adaptation to diverse subdomains but also retains efficiency, as only the small bias matrices and routing parameters are updated during training.

During inference, real audio input is available, and the encoder is reintroduced to generate contextual representations. However, the decoder is trained with bias embeddings, creating a potential mismatch between the learned adaptation and the actual encoder output. To ensure a smooth transition while preserving domain-specific knowledge, we integrate the learned biases with the encoder's output through a linear interpolation.

Given the encoder-generated key and value matrices during inference, we modify the cross-attention mechanism as follows:

$$K' = \alpha K + (1-\alpha)B^*, \quad V' = \alpha V + (1-\alpha)B^* \tag{7}$$

where K, V are the encoder's outputs derived from the audio input, $\mathbf{B}^*$ is the aggregated bias embedding from the routing network, and α is a weight that balances the contribution of the encoder and bias embeddings.

The interpolation ensures that the decoder receives both domain-specific cues (from B) and actual acoustic representations (from K, V). We consider the value of α as 0.5.

Loss Function: To improve Whisper's performance on domain-specific transcription tasks, we explore alternative loss functions beyond standard cross-entropy. Specifically, we incorporate two loss functions:

1. **Kullback-Leibler (KL) Divergence**: This loss function measures the divergence between two probability distributions, guiding the model towards a better approximation of the true transcription distribution. Minimizing this divergence improves the fluency and accuracy of generated transcriptions.

2. **Bregman Divergence-Inspired Loss**: This loss function prioritizes correct predictions of domain-specific terms (e.g., technical jargon, medical terminology) by assigning higher penalties to errors involving critical domain-specific words.

The combined loss function as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{KL}} \cdot \text{KL}(P_{\text{true}} || P_{\text{pred}}) + \lambda_{\text{BD}} \cdot \sum_{i=1}^{n} \delta_i \cdot I(w_i \in \mathcal{D}) \tag{8}$$

where $\text{KL}(P_{\text{true}} || P_{\text{pred}})$ is the Kullback-Leibler Divergence between the true transcription distribution P_{true} and the predicted transcription distribution P_{pred}, λ_{KL} is a hyperparameter that controls the weight of the KL divergence term, δ_i is a penalty factor for incorrect predictions of domain-specific words, $I(w_i \in \mathcal{D})$ is an indicator function that is 1 if the word w_i belongs to the domain-specific vocabulary $\mathcal{D}$, and 0 otherwise, and λ_{BD} is a hyperparameter controlling the weight of the Bregman Divergence-Inspired loss term. In this work, we construct the domain-specific vocabulary D using public data sources and private data, including domain-specific reports.

4 Evaluation and Results

We utilize the Whisper-base English model as the foundation for domain-specific adaptation in our experiments. The domain-specific text data is normalized and tokenized to ensure compatibility with the Whisper tokenizer. We measure performance using Word Error Rate (WER) against the baseline to assess the effectiveness of text-only fine-tuning in improving recognition accuracy in specialized domains.

4.1 Dataset

We evaluate the model on three domain-specific datasets:

Earnings Call: The dataset contains quarterly earnings conference calls from S&P 500 companies in 2017 Qin and Yang (2019). It includes domain-specific financial discussions from corporate meetings. For text-only fine-tuning, we use

Table 1: Performance of our model with existing baselines

Methods	**Earnings Call**	**OCW2**	**MedReport**
Whisper-base Radford et al. (2022)	32.9	33.5	32
Context Pertubation Suh et al. (2024)	15.15	**9.79**	NA
Ours	**14.24**	17.4	**18**

11,736 text files for training and 1,678 audio files for testing, with total audio durations of approximately 30 hours and 5 hours, respectively.

OCW2: The OCW2 dataset from MIT OpenCourseWare covers a range of academic lectures Suh et al. (2024). It contains 24,123 text files for training and 3,447 audio files for testing, corresponding to approximately 40 and 10 hours of audio, respectively.

MedReport: The medical domain often lacks large paired audio-text datasets, but has abundant domain-specific text. We curate a set of medical sentences, including drug and medicine names, from pharmaceutical company annual reports and industry publications. This dataset contains 4,000 text files for training and 1,000 audio files for testing, with total audio durations of approximately 10 and 3 hours, respectively.

4.2 Experimental Setup

Training is conducted on an NVIDIA Tesla V100 GPU for a maximum of 1,000 steps, with 200 warmup steps to gradually increase the learning rate. Data loading uses 16 workers, and evaluation is performed every 50 steps. Logging occurs every 10 steps, and model checkpoints are saved every 50 steps. Intermediate evaluations are skipped to accelerate training iterations.

4.3 Results and Discussion

We evaluate our approach using Whisper-base as the base model and compare it with two baselines. The first is the pre-trained Whisper-base model Radford et al. (2022). The second is a domain-adapted ASR model trained on paired audio-text data and prompted with LLM-generated descriptions combined with context perturbation Suh et al. (2024).

Model performance is measured using WER. Table 1 presents the overall results, showing that our model achieves notable improvements over both baselines. These gains confirm that text-only fine-tuning enhances recognition accuracy in specialized domains. Table 2 provides example transcriptions for the Earnings Call, OCW2, and MedReport datasets, comparing Whisper-base and our adapted model. After fine-tuning, our model more accurately recognizes domain-specific vocabulary, leading to better transcription quality. In the examples, bold text denotes correct outputs matching the ground truth (previously misclassified by Whisper-base), while underlined text indicates incorrect outputs that differ from the ground truth.

Earnings Call: Table 2 compares transcriptions among the ground truth, Whisper-base, and our proposed model (“Ours") for the Earnings Call dataset. In the first example, Whisper-base misinterprets “in fact" as “affect" and “net interest expense" as “that interest expense" while also transcribing “debt" as “dad." The proposed model restores these key financial terms correctly. In the second example, Whisper-base introduces disfluencies such as “you know," misrecognizes “build" as “bill" and distorts “UK" into “u.k.a." In the third example, Whisper-base produces an entirely altered phrase, “many ways when I’m very much focused" deviating significantly from the ground truth, whereas the proposed model correctly outputs “we’re now" preserving the intended meaning.

As shown in Table 1, the context perturbation approach improves accuracy, reducing WER to 15.15 on the Earnings Call dataset compared with Whisper-base’s 32.9. The proposed method achieves the best performance, lowering WER further to 14.24, corresponding to a 19% relative reduction over Whisper-base and a 0.9% reduction compared with the context perturbation approach.

OCW2: Table 2 shows that, for the OCW2 dataset, our model correctly retains technical terms such as “fetch and decode" instead of Whisper-base’s “FETCH-ND code" and produces grammatically accurate phrases such as “generator is going to burn out" instead of “generators gonna burn out". For complex scientific content, the proposed model maintains coherence better than Whisper-base, although minor errors persist, such as “phalmus" for “thalamus" (an improvement over Whisper-base’s “phalm") and the introduction of “cells that goes" instead of “cells that go".

Table 2: Comparison of the transcription output obtained from Whisper-base and Ours with the ground truth. Bold text represents responses that are correct according to the ground truth, but were misclassified by the Whisper-base model. Incorrect responses are underlined.

	Ground Truth	**Whisper-base**	**Ours**
Earnings Call	most of that in fact almost all of that was net interest expense on our automotive debt	most of that affect almost all of that was that interest expense on our automotive dad	most of that, **in fact** almost all of that, was **net** interest expense on our automotive **debt**.
	no question about it because customers were trying to decide do they want to build their next datacenter in the uk or should they be building that datacenter someplace else in europe	no question about it because customers were trying to decide that they want to you know bill that next data center in the u.k.a. or should they be building that data center someplace else in europe	no question about it because customers were trying to decide that they want to build that next data center in the **UK** or should they be building that data center someplace else in Europe.
	we're now very much focused on operating effectively in a warehouse delivered model which we think we can do because we do it across our other businesses	many ways when I'm very much focused on operating effectively in a warehouse delivered model which we think we can do because we do it across our other businesses.	**we're now** very much focused on operating effectively in a warehouse delivered model which we think we can do because we do it across our other businesses.
OCW2	generator is going to burn out in let's say 10 or 20 years	generators gonna burn out in let's say 10 or 20 years.	generator **is going to** burn out in let's say 10 or 20 years.
	and the fetch and decode stages implement optimizations	and the FETCH-ND code stages implement optimizations	and the FETCH **and Decode** stages implement optimizations
	it's the ventral the posterior part of the ventral nucleus the thalamus. and that's where we find the cells that goes to the neocortex as i show there	It's the ventral posterior part of the ventral nucleus of the phalm. And that's where we find the cells that go to the neocortex as I show there.	It's the ventral posterior part of the ventral nucleus of the phalmus. And that's where we find the cells that **goes** to the neocortex as I show there.
MedReport	Paracetamol is one of the most commonly used medications for pain relief and fever reduction.	Parasetimal is one of the most commonly used medications for pain relief and fever reduction.	**Paracetamol** is one of the most commonly used medications for pain relief and fever reduction.
	Azee is a commonly prescribed antibiotic to treat bacterial infections. It contains azithromycin, which is effective against respiratory and skin infections.	A Z is a commonly prescribed antibiotic to treat bacterial infections. It contains azithromycin, which is effective against respiratory and skin infections.	**Azee** is a commonly prescribed antibiotic to treat bacterial infections. It contains azithromycin, which is effective against respiratory and skin infections.
	Broncol is a bronchodilator that helps manage respiratory conditions like asthma. Cipla's broncol is effective in relieving broncospasm and improving breathing.	Bronkel is a bronco dilator that helps manage respiratory conditions like asthma. Sipla's bronkel is effective in relieving bronchospasm and improving breathing.	Broncal is a **bronchodilator** that helps manage respiratory conditions like asthma. **Cipla's** broncal is effective in relieving broncospasm and improving breathing.

As reported in Table 1, Whisper-base has a high WER of 33.5. The context perturbation method achieves the lowest WER of 9.79, reflecting strong adaptation when trained with audio data and domain-specific prompts. Our method achieves a WER of 17.4, representing a 16% relative reduction compared to Whisper-base, but not matching the performance of context perturbation due to its reliance on text-only fine-tuning.

The relatively higher WER on the OCW2 dataset arises from the nature of our text-only adaptation strategy. Unlike the context perturbation method,

which leverages paired audio–text data and domain audio cues to align acoustic and lexical variations, our approach operates purely on text, without exposure to acoustic or prosodic features present in lecture recordings. OCW2 includes substantial variability in speaker style, pacing, and background conditions, which purely textual fine-tuning cannot capture. Despite this limitation, our model achieves consistent improvements over the base Whisper model, demonstrating that linguistic adaptation alone can transfer domain knowledge effectively even in acoustically complex settings. Future extensions could integrate lightweight audio-conditioned adapters or multi-modal alignment losses to further close this gap.

MedReport: For the MedReport dataset, our model shows substantial improvements in recognizing medical terminology. It accurately transcribes “Paracetamol" instead of “Parasetimal" and “Azee" instead of “A Z" which are critical distinctions in medical transcription. Some errors remain, such as “Broncol" being transcribed as “Broncal" but these are less severe than Whisper-base’s phonetic distortions. For example, in a case where “Broncol" appears with additional context about Cipla’s product, the proposed model correctly restores the term. As shown in Table 1, WER improves from 32% to 18% after fine-tuning, corresponding to a 14% relative reduction over Whisper-base.

Across domains, Whisper-base shows frequent structural inconsistencies and misrecognitions that distort meaning. Our model, which relies solely on text-based domain adaptation, produces more accurate and readable domain-specific transcriptions but occasionally hallucinates, especially on short audio segments where insufficient context leads to completions based on statistical likelihood rather than actual input. For example, in OCW2, *“and you can restore this activity. you have a question the intermediate stuff where it’s reduced but not yet denatured how do you”* was transcribed as *“and you can restore this activity. Do you have a question? Yes, so in the intermediate stuff where it’s reduced, but not yet, do you make sure how to use it?”*. Similarly, in Earnings Call, *“the americas were up in midsingle digits with strength in the united states”* became *“America’s Rop in midsingle digits for strengthening audit states.”*. These errors are rare in Earnings Call and mostly substitutions, while OCW2 shows added explanatory phrases. We mitigate hallucinations by appending silence to short segments and applying prompt constraints, improving consistency without requiring audio-text alignment. Overall, the improvements over Whisper-base demonstrate that text-only adaptation can achieve strong domain-specific performance while keeping computational costs low.

5 Conclusion

In this work, we present a text-only adaptation method for domain-specific speech recognition by fine-tuning the decoder of the Whisper model. The encoder’s output is replaced with trainable biases, allowing the model to capture domain-specific linguistic patterns without requiring paired audio-text data. The proposed method shows substantial improvements in transcription accuracy, particularly for specialized vocabularies, while maintaining computational efficiency. This demonstrates the practicality of our approach for domain adaptation in settings with limited audio resources. Future work explores integrating fine-tuned small language models (SLMs) with additional modalities, such as video, to further enhance domain-specific recognition performance.

References

Petar S Aleksic, Mohammadreza Ghodsi, Assaf Hurwitz Michaely, Cyril Allauzen, Keith B Hall, Brian Roark, David Rybach, and Pedro J Moreno. 2015. Bringing contextual information to google speech recognition. In *Interspeech*, pages 468–472.

Vladimir Bataev, Roman Korostik, Evgeny Shabalin, Vitaly Lavrukhin, and Boris Ginsburg. 2023. Text-only domain adaptation for end-to-end asr using integrated text-to-mel-spectrogram generator. *arXiv preprint arXiv:2302.14036*.

Peter Bell, Joachim Fainberg, Ondrej Klejch, Jinyu Li, Steve Renals, and Pawel Swietojanski. 2020. Adaptation algorithms for neural network-based speech recognition: An overview. *IEEE Open Journal of Signal Processing*, 2:33–66.

Shuo-Yiin Chang, Chao Zhang, Tara N Sainath, Bo Li, and Trevor Strohman. 2023. Context-aware end-to-end asr using self-attentive embedding and tensor fusion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Yan Huang, Jinyu Li, Lei He, Wenning Wei, William Gale, and Yifan Gong. 2020. Rapid rnn-t adaptation

using personalized speech synthesis and neural language generator. In *Interspeech*, pages 1256–1260.

Anjuli Kannan, Yonghui Wu, Patrick Nguyen, Tara N Sainath, Zhijeng Chen, and Rohit Prabhavalkar. 2018. An analysis of incorporating an external language model into a sequence-to-sequence model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5828. IEEE.

Yingyi Ma, Zhe Liu, and Ozlem Kalinli. 2024. Effective text adaptation for llm-based asr through soft prompt fine-tuning. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 64–69. IEEE.

Yu Qin and Yi Yang. 2019. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 390–401.

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

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.

Jiwon Suh, Injae Na, and Woohwan Jung. 2024. Improving domain-specific asr with llm-generated contextual descriptions. *arXiv preprint arXiv:2407.17874*.

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. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. 2017. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Xianrui Zheng, Yulan Liu, Deniz Gunceler, and Daniel Willett. 2021. Using synthetic audio to improve the recognition of out-of-vocabulary words in end-to-end asr systems. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5674–5678. IEEE.