Closing the Speech-Text Gap with Limited Audio for Effective Domain Adaptation in LLM-based ASR

Thibault Bañeras-Roux [1,*,**], *Sergio Burdisso* [1,*,**], *Esaú Villatoro-Tello* [1,*,**], *Dairazalia Sánchez-Cortés* [1], *Shiran Liu*[1], *Severin Baroudi*[1,2], *Shashi Kumar* [1,3], *Hasindri Watawana* [1,3], *Manjunath K E*[4], *Kadri Hacioglu*[4], *Petr Motlicek* [1,5], *Andreas Stolcke*[4]

[1] Idiap Research Institute, Switzerland
[2] Laboratoire d'Informatique et des Systèmes, France
[3]EPFL, Switzerland
[4] Uniphore, USA & India
[5] Brno University of Technology, Czech Republic

{thibault.roux, sergio.burdisso, esau.villatoro}@idiap.ch

Abstract

Conventional end-to-end automatic speech recognition (ASR) systems rely on paired speech–text data for domain adaptation. Recent LLM-based ASR architectures connect a speech encoder to a large language model via a projection module, enabling adaptation with text-only data. However, this introduces a modality gap, as the LLM is not exposed to the noisy representations produced by the speech projector. We investigate whether small amounts of speech can mitigate this mismatch. We compare three strategies: text-only adaptation, paired speech–text adaptation, and mixed batching (MB), which combines both. Experiments in in-domain and out-of-domain settings show that even limited speech consistently improves performance. Notably, MB using only 10% of the target-domain (less than 4 hours) speech achieves word error rates comparable to, or better than, conventional ASR fine-tuning with the full dataset, indicating that small amounts of speech provide a strong modality-alignment signal.

Index Terms: speech recognition, domain adaptation, low-resource training.

1. Introduction

Large language models (LLMs) have transformed speech processing, enabling natural voice interactions. A common approach is LLM-based automatic speech recognition (ASR), where a pretrained speech encoder is connected to a pretrained language model through a lightweight projection layer [1, 2, 3, 4, 5, 6, 7]. In particular, SLAM-ASR achieves strong transcription performance without retraining the entire system [8, 9, 10, 11, 12].

Rather than training from scratch, pretrained systems can be adapted to new domains through fine-tuning on domain-specific data, transferring learned representations [13, 14, 15]. However, adaptation to specialized domains such as healthcare or finance remains challenging because, although in-domain text is often abundant, labeled speech is expensive to collect.

A standard adaptation strategy is to fine-tune the LLM using in-domain transcripts only [16, 17]. However, this often creates a modality gap [18, 19, 10, 17, 20]: the model no longer receives representations similar to those produced by the speech projector, weakening speech-text alignment. Existing solutions, such as soft prompting, generally assume a text-only setting [19, 17, 20] and do not investigate whether limited target-domain speech can help preserve cross-modal alignment.

In this work, we study adaptation across Banking, Insurance, Agriculture, and Musical Instruments domains, analyzing the effect of incorporating limited paired speech-text data through a mixed batching strategy. We compare in-domain and cross-domain adaptation settings to derive practical guidance for scenarios where target-domain text is plentiful, but speech resources are scarce.

Our contributions are fourfold: *(i)* we analyze how the proportion of target-domain data in training batches affects the tradeoff between target-domain specialization and source-domain retention in LLM-based ASR; *(ii)* we propose a hybrid adaptation method that incorporates limited paired speech-text data into a predominantly text-based setting, mitigating the modality gap introduced by text-only fine-tuning; *(iii)* we show that the proposed mixed batching strategy reduces catastrophic forgetting while improving target-domain recognition; and *(iv)* we release our source code for reproducibility.[1]

2. Related Work

LLM-Based ASR. Recent work in ASR has shifted toward modular architectures that couple pretrained self-supervised speech encoders such as wav2vec 2.0 [11], HuBERT [12], or WavLM [21], with LLMs such as LLaMA [22], or Vicuna [23]), through trainable projection layers that map continuous acoustic features into the LLM embedding space [1, 24, 25, 26]. In this work, we refer to this trainable projection module as the *speech projector*, and to the overall architecture as an *LLM-based ASR* system. Recent studies demonstrate that such lightweight connectors, ranging from linear projections to Q-Former-style modules, can yield competitive or state-of-the-art WER on benchmarks like LibriSpeech without retraining the entire LLM [1].

Domain Adaptation in Low-Resource Scenarios. Adapting ASR systems to specialized domains such as healthcare, legal, and finance remains a practical necessity, as domain-specific terminology and discourse patterns often diverge significantly from general-purpose training corpora [9, 27]. In low-resource

*These authors contributed equally.
**indicates the corresponding author.

[1] https://github.com/idiap/llm-asr-text-only-adaptation

scenarios, the primary bottleneck is the scarcity and cost of collecting in-domain speech data [28, 29], motivating research into text-only domain adaptation. Rather than retraining acoustic models with new audio, these approaches adapt language models or decoding strategies using domain-specific documents such as reports, transcripts, or vocabulary lists. More recent work [10, 30] investigates zero-shot and semi-supervised adaptation, where pretrained ASR models are specialized using textual resources. Collectively, these approaches demonstrate that substantial domain-specific WER reductions can be achieved through efficient use of textual data, offering scalable solutions for low-resource deployment.

However, fine-tuning LLMs on clean text degrades their ability to interpret "noisy" representations generated by upstream speech projectors, leading to misalignment and degraded transcription quality [19]. To address this, prior work has explored strategies such as soft prompting [10], where learned continuous tokens condition the LLM on speech features without rigid tokenization, and soft discretization techniques [4].

3. Mixed Batching Strategy

We consider a domain adaptation setting for LLM-based ASR. First, a base model is trained on source-domain paired speech-text data. This base system learns the initial speech–text alignment and serves as the starting point for all experiments.

We then adapt this pretrained model to a target domain using different fine-tuning strategies. Depending on the setting, the target-domain data may include text-only data, paired speech-text data, or a combination of both. The goal of adaptation is to improve recognition performance on the target domain while preserving performance on the original source domain.

To mitigate the modality gap introduced by text-only adaptation, we build upon the mixed batch composition strategy proposed in a previous study [19]. In that work, the authors integrate target-domain text-only data with source-domain auxiliary data during fine-tuning to preserve speech-text alignment and mitigate catastrophic forgetting while enabling adaptation without target-domain audio, by dividing each batch into a source portion (auxiliary data) and a target-domain portion

In our experiments, we extend this strategy by integrating target-domain audio into the batch composition. More precisely, similar to [19], the source data in the batch is composed of three segments:

(1) **Paired source speech-text** (batch proportion σ_a): the input to the model is a source-domain audio utterance, and the output is its corresponding transcript. This component allows the model to preserve the original audio-text alignment learned during base training.

(2) **Projector-induced noisy tokens** (batch proportion σ_{ta}): the input to the model is a discrete sequence obtained by mapping each speech projection to its nearest tokens in the LLM embedding space, and the output is the corresponding clean transcript. This helps the model to understand discrete text from projected speech representations.

(3) **Synthetically corrupted source transcripts** (batch proportion σ_t): the input to the model is a corrupted transcript generated through random character-level perturbations using `nlpaug`[2] (with the same parameters as in the original paper), and the output is the original clean transcript. This discrete input sequence simulates the irregularities observed in projected speech-token representations without requiring audio.

Following the paired speech-text and synthetically corrupted transcripts, the target portion of the batch consists of:

(1) **Paired target speech-text** (batch proportion τ_a): the input to the model is a target-domain audio utterance, and the output is its corresponding transcript. This component induces the adaptation in real target-domain acoustics and preserves speech-text alignment.

(2) **Corrupted target transcripts** (batch proportion τ_t): the input to the model is a corrupted target-domain transcript generated through the same character-level perturbations as in σ_t, and the output is the corresponding clean transcript. This component supports domain-specific adaptation using text-only supervision.

While τ_t controls domain-specific language modeling, τ_a keeps the adaptation grounded in real target-domain speech and acoustics. From now on, we will refer the strategy described here as Mixed Batch (MB).

4. Method

4.1. LLM-based ASR Architecture

In our experiments, the ASR system follows the SLAM-ASR framework [1]. The architecture consists of three main components: (i) a pretrained speech encoder, (ii) a trainable speech projector, and (iii) an LLM.

Given a speech segment X^S, the speech encoder produces a sequence of acoustic representations $H^S = [h_1^S, \dots, h_T^S]$ over T time frames. To better match the token-level processing of the LLM, this sequence is temporally downsampled by a factor k, resulting in $Z^S = [z_1^S, \dots, z_N^S]$ with $N = T//k$.

The downsampled representations are then passed through a linear speech projector, which maps them into the LLM embedding space, producing E^S. These projected embeddings are directly consumed by the LLM, together with a task-specific prompt that conditions the model for speech recognition. The LLM then autoregressively generates the output transcript.

In our setup, we first train a base model using paired speech-text data from the source corpus (see Section 4.2). During this stage, the speech encoder is kept frozen, while the speech projector is fine-tuned to learn the alignment between acoustic representations and the LLM embedding space. This results in a base ASR system adapted to the source domain.

The base model is subsequently adapted to target domains using the various strategies described, by fine-tuning only LoRA adapters [31]. During adaptation, the speech encoder and the speech projector are frozen, and only the LoRA modules inserted into the LLM are updated. This design enables parameter-efficient specialization to the target domain while preserving the previously learned speech–text alignment.

For both base training and adaptation, models are trained for 5 epochs with a batch size of 10. The speech encoder is `WavLM-Large` [21], and the LLM is `Llama-3.2-3B-Instruct` [22].

All experiments use a fixed instruction-style prompt. Specifically, the LLM input follows the template:

```
<s>USER: Transcribe speech to text.
Speech:{speech}
ASSISTANT: {transcript}</s>
```

where `{speech}` corresponds to the projected speech embeddings. This prompt conditions the LLM to perform transcription in an instruction-following setup consistent with its pretraining.

[2] `https://nlpaug.readthedocs.io/en/latest/augmenter/char/random.html`

Table 1: *Datasets used for training, validation, and testing in our experiments. From the DefinedAI corpus, we use Banking (B), Insurance (I) and Healthcare (H) partitions. From the SlideSpeech corpus, we use Agriculture (Ag) and Musical Instruments (MI). We provide the total number of utterances (#Utts) and the duration (Hrs) of each partition.*

Dataset	Domain	Train		Dev		Test	
		#Utts	Hrs	#Utts	Hrs	#Utts	Hrs
Source							
DefinedAI	B/I/H	17,398	38h10	1,008	2h12	2,009	4h21
Target							
DefinedAI	B	26,704	36h20	1,625	2h08	3,111	4h16
SlideSpeech	Ag	30,498	29h20	1,679	1h43	3,470	3h26
SlideSpeech	MI	9,981	8h30	852	0h44	984	0h54

4.2. Datasets and Domain Configuration

Our experiments leverage two corpora, DefinedAI[3] and SlideSpeech [32]. The source-domain data comes from DefinedAI, encompassing Banking (B), Insurance (I), and Healthcare (H) subsets. This dataset is used to pretrain the projector and serve as auxiliary data during LLM adaptation.

The target-domain data consists of specific subsets from both corpora. From DefinedAI, we select the Banking partition to train and evaluate in-domain adaptation. SlideSpeech contributes data from the Agriculture (Ag) and Musical Instruments (MI) domains, representing out-of-distribution scenarios with limited labeled audio. Table 1 summarizes the statistics for each dataset, including the number of utterances and total audio duration for training, development, and test splits.

5. Results

Before conducting the adaptation experiments, we first train a base ASR model exclusively on the source-domain data from the DefinedAI corpus, including the Banking, Insurance, and Healthcare partitions. This base model serves as the starting point for all subsequent experiments. We then perform domain adaptation on different target domains: (i) the in-domain DefinedAI Banking set, and (ii) out-of-distribution domains from the SlideSpeech corpus, namely Agriculture (Ag) and Musical Instruments (MI). All adaptation strategies described in the following sections are applied to this same base model to ensure controlled comparisons across domains and training settings.

5.1. Target Data Proportion in Text-Only Adaptation

We investigate the effect of varying the proportion of target text data per batch (τ_t) in the text-only adaptation setting. No audio is included ($\tau_a = 0$); the remaining portion of a batch is split uniformly among the various types of source-domain text data.

As shown in Figure 1, performance follows a clear tradeoff pattern. Increasing τ_t initially improves recognition accuracy, reaching optimal performance around $\tau_t = 50\%$. Beyond this point, performance degrades, reflecting a balance between: (1) Low τ_t: excessive reliance on auxiliary data, which may be repeatedly sampled and lead to overfitting. (2) High τ_t: insufficient exposure to auxiliary data, potentially causing underfitting and partial catastrophic forgetting.

Based on these results, we adopted $\tau = \tau_t + \tau_a = 50\%$ in all experiments and σ values being uniformly distributed, as

[3] https://defined.ai/

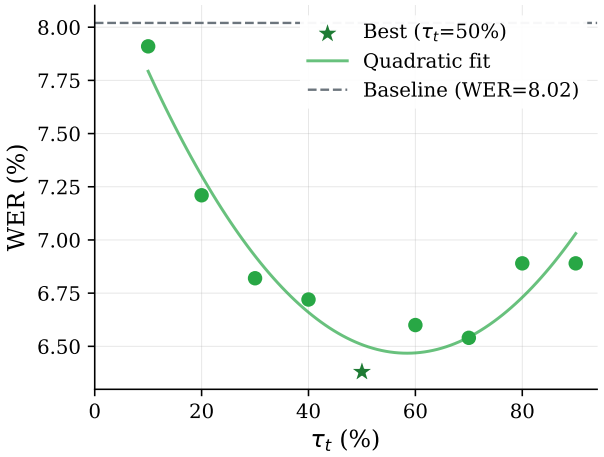


Figure 1: *Effect of different target text proportions (τ_t) on mixed-batch adaptation on DefinedAI Banking.*

that seems to provide the best performance.

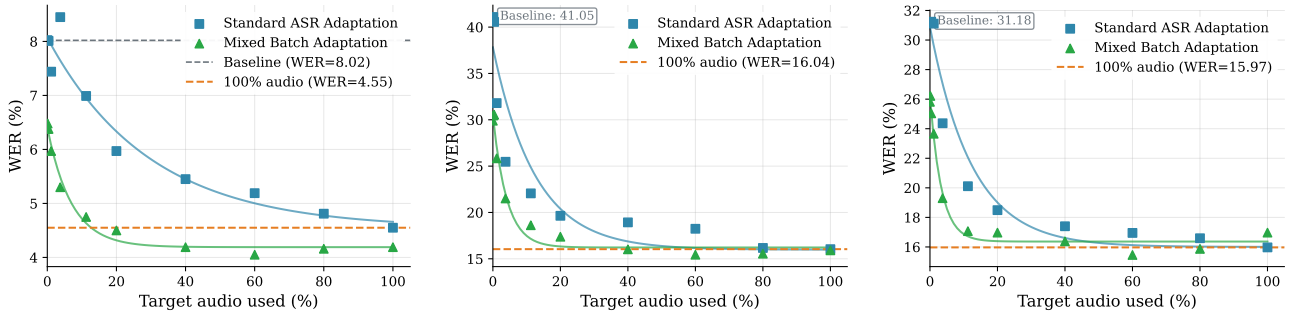
5.2. Paired Speech-Text vs. Mixed-Batch Adaptation

We now look at whether text-only adaptation combined with a small amount of audio can reduce the modality gap between text-only and paired speech-text training. Both in previous studies [19, 8, 17] and in our current setup, we observe an important performance difference between the text-only adapted model and the speech-text adapted model. In particular, in our experience, paired speech-text adaptation outperforms text-only adaptation by 1.83% in absolute WER ($\approx$ 29% relative improvement) on the Banking domain (4.55% vs. 6.38%), highlighting a substantial modality gap. We therefore investigate whether incorporating a small quantity of target-domain audio during adaptation can effectively bridge this gap, and if the system can get close to the best-case scenario where we have full audio. More specifically, we compare:

- **Speech-text adaptation**, or *standard ASR adaptation*, where the model is fine-tuned exclusively on various amount of labeled speech with the associated transcript.
- **Mixed-batch adaptation**, where text-based adaptation is complemented with a small quantity of audio speech with associated transcript (*i.e.*, text-only and paired speech-text together comprise the full training set).

We fine-tune the base model using different fractions of each corpus, including setups with very limited speech data. This allows us to study how the model adapts under low-resource conditions, where the available audio alone is insufficient to achieve strong performance, as well as under scenarios with progressively more adaptation speech.

Figure 2 shows that mixed-batch adaptation consistently outperforms paired speech-text training. Adding speech to a text-adapted model leads to larger gains than training on the same amount of paired speech-text. With only 10% of the adaptation corpus containing speech (corresponding to 3h37, 2h55, and 50m for Banking, Agriculture, and Musical Instruments, respectively), we obtain performance comparable or better than using the full speech corpus. This suggests that text adaptation can improve an LLM's domain-specific abilities, while additional speech helps align audio representations with the embedding space of the LLM, thereby reducing the modality gap.



(a) *Banking* (b) *Agriculture* (c) *Musical Instruments*

Figure 2: *Performance of the base ASR and adapted models. The figure compares standard ASR adaptation using only paired speech–text data with mixed-batch adaptation combining paired speech–text and text-only samples. The x-axis indicates the proportion of speech in the adaptation data, ranging from 0% (text-only) to 100% (paired speech–text).*

Across all target domains, we observe that mixed-batch adaptation achieves its peak performance when approximately 60% of the corpus consists of paired speech-text examples, after which the gains slightly decrease. We hypothesize that this behavior arises from the presence of synthetically generated noisy transcripts in the batch, which introduce additional diversity in the inputs, which seems to help the model learn more robust mappings between speech and text representations. It seems that mixed batching not only leverages the benefits of target-domain text adaptation, but could improve generalization under low-resource conditions.

However, as the amount of available speech increases, the performance difference between the two strategies gradually diminishes. When sufficient speech data is available, both approaches converge toward similar recognition accuracy (with respect to mixed-batch adaptation and paired speech-text adaptation: **4.19%** vs. **4.55%** for Banking, **15.91%** vs. **16.04%** for Agriculture, **16.97%** vs. **15.97%** for Musical Instruments), indicating that audio supervision alone eventually becomes sufficient. For in-domain adaptation (Banking), mixed-batch adaptation still shows a slight advantage, likely due to the contribution of auxiliary data drawn from the same domain.

5.3. Analysis of Catastrophic Forgetting

To assess the impact of adaptation on previously learned knowledge, we evaluate all systems on both source-domain and target-domain test sets. In this analysis, the source domain corresponds to the DefinedAI (B/I/H) data used for base training, while the target domain is the DefinedAI Banking subset. This setup allows us to quantify potential catastrophic forgetting.

As shown in Figure 3, the base model achieves a WER of 12.8% on the source domain and 8.0% on the target domain. This indicates that, in our setup, the target test set is inherently easier than the source data for speech recognition. When adapting using paired speech-text (with 20% of target audio), we observe a clear degradation in the source domain, even though the adaptation remains in-domain. While recognition improves on the target domain, the substantial accuracy drop in the source domain is evidence of catastrophic forgetting. In contrast, text-only and mixed-batch adaptation result in an improvement in target-domain performance while limiting degradation on the source domain.

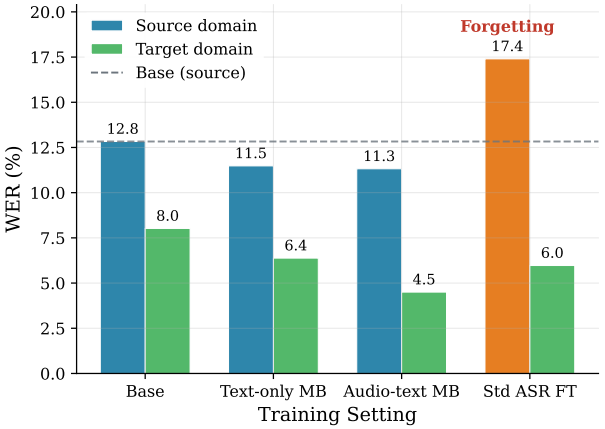


Figure 3: *Performances of base system and adapted model with text-only, paired speech-text and mixed-batch adaptation. For both audio-adaptation settings, the share of speech data is 20%.*

6. Conclusions

we have investigated domain adaptation for LLM-based ASR in settings where target-domain speech is limited, while textual data is abundant. We identified the modality mismatch introduced by text-only fine-tuning as a key factor behind performance degradation and proposed a hybrid batching strategy that mixes speech-text pairs with text to address this issue.

By integrating auxiliary source data with both corrupted target text and a controlled proportion of target-domain audio, our approach preserves speech-text alignment while enabling effective domain adaptation. Through systematic experiments, we showed that using our hybrid batch method, even with modest amounts of target-domain audio, substantially improves recognition accuracy and mitigates catastrophic forgetting. Notably, combining text data with a small fraction of audio can outperform adaptation using the full audio dataset, highlighting the data efficiency of the proposed strategy.

We also observed a clear tradeoff between adapting to the target domain and preserving performance on the source domain, with respect to the batch composition. Our findings provide practical modeling guidelines for real application scenarios.

7. Acknowledgments

This work was supported by an Idiap Research Institute and Uniphore collaboration project. Part of this work was also supported by EU Horizon 2020 project ELOQUENCE.

8. Generative AI Use Disclosure

Generative AI was used to proofread the paper, fix orthographic and grammatical errors, and shorten the length of some sections.

9. References

[1] Z. Ma, G. Yang, Y. Yang, Z. Gao, J. Wang, Z. Du, F. Yu, Q. Chen, S. Zheng, S. Zhang *et al.*, "An embarrassingly simple approach for LLM with strong ASR capacity," *arXiv preprint arXiv:2402.08846*, 2024.

[2] M. Wang, W. Han, I. Shafran, Z. Wu, C.-C. Chiu, Y. Cao, N. Chen, Y. Zhang, H. Soltau, P. K. Rubenstein *et al.*, "SLM: Bridge the thin gap between speech and text foundation models," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[3] W. Yu, C. Tang, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Connecting speech encoder and large language model for asr," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 637–12 641.

[4] M. Yang, S.-J. Chen, J. Xie, and H. L. John Hansen, "Bridging the modality gap: Softly discretizing audio representation for LLM-based automatic speech recognition," in *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2025, pp. 1–7.

[5] S. Kumar, I. Thorbecke, S. Burdisso, E. Villatoro-Tello, M. KE, K. Hacioğlu, P. Rangappa, P. Motlicek, A. Ganapathiraju, and A. Stolcke, "Performance evaluation of slam-asr: The good, the bad, the ugly, and the way forward," in *2025 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2025, pp. 1–5.

[6] S. Burdisso, E. Villatoro-Tello, S. Kumar, S. Madikeri, A. Carofilis, P. Rangappa, M. K E, K. Hacioglu, P. Motlicek, and A. Stolcke, "Reducing prompt sensitivity in LLM-based speech recognition through learnable projection," in *Proc. IEEE ICASSP*, 2026.

[7] E. Villatoro-Tello, S. Burdisso, S. Kumar, H. Watawana, S. Madikeri, M. K E, J. Prakash, T. Bañeras-Roux, K. Hacioglu, P. Motlicek, and A. Stolcke, "Context projector: Complementary keyword and dialogue context embeddings for LLM-based ASR," in *Proc. Interspeech*, 2026.

[8] L. Xu, H. Xie, S. J. Qin, X. Tao, and F. L. Wang, "Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[9] H. Zheng, L. Shen, A. Tang, Y. Luo, H. Hu, B. Du, Y. Wen, and D. Tao, "Learning from models beyond fine-tuning," *Nature Machine Intelligence*, vol. 7, no. 1, pp. 6–17, 2025.

[10] Y. Li, Y. Wu, J. Li, and S. Liu, "Prompting large language models for zero-shot domain adaptation in speech recognition," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[11] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[12] 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.

[13] R. Joshi and A. Singh, "A simple baseline for domain adaptation in end to end ASRsystems using synthetic data," in *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, 2022, pp. 244–249.

[14] S. Khurana, A. Laurent, and J. Glass, "Magic dust for cross-lingual adaptation of monolingual wav2vec-2.0," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6647–6651.

[15] J. Huang, O. Kuchaiev, P. O'Neill, V. Lavrukhin, J. Li, A. Flores, G. Kucsko, and B. Ginsburg, "Cross-language transfer learning, continuous learning, and domain adaptation for end-to-end automatic speech recognition," *arXiv preprint arXiv:2005.04290*, 2020.

[16] L. Zheng, X. Wang, Q. Zhao, and T. Li, "Two-stage domain adaptation for LLM-based ASR by decoupling linguistic and acoustic factors," *Applied Sciences*, vol. 16, no. 1, p. 60, 2025.

[17] Y. Fang, J. Peng, X. Li, Y. Xi, C. Zhang, G. Zhong, and K. Yu, "Low-resource domain adaptation for speech LLMs via text-only fine-tuning," *arXiv preprint arXiv:2506.05671*, 2025.

[18] S. Burdisso, E. Villatoro-Tello, T. Bañeras-Roux, S. Kumar, S. Madikeri, M. K E, P. Motlicek, and A. Stolcke, "Avoiding catastrophic forgetting in text-only adaptation of LLM-based ASR via multi-view text denoising," in *Proc. Interspeech*, 2026.

[19] A. Carofilis, S. Burdisso, E. Villatoro-Tello, S. Kumar, K. Hacioglu, S. Madikeri, P. Rangappa, M. K. E, P. Motlicek, S. Venkatesan, and A. Stolcke, "Text-only adaptation in LLM-based ASR through text denoising," 2026. [Online]. Available: https://arxiv.org/abs/2601.20900

[20] Y. Ma, Z. Liu, and O. Kalinli, "Effective text adaptation for LLM-based ASR through soft prompt fine-tuning," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 64–69.

[21] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[22] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The Llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[23] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez *et al.*, "Vicuna: An open-source chatbot impressing GPT-4 with 90% ChatGPT quality," *See https://vicuna. lmsys. org (accessed 14 April 2023)*, vol. 2, no. 3, p. 6, 2023.

[24] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, M. Zejun, and C. Zhang, "SALMONN: Towards Generic Hearing Abilities for Large Language Models," in *The Twelfth International Conference on Learning Representations*, 2024.

[25] J. Wu, Y. Gaur, Z. Chen, L. Zhou, Y. Zhu, T. Wang, J. Li, S. Liu, B. Ren, L. Liu *et al.*, "On decoder-only architecture for speech-to-text and large language model integration," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[26] S. Ghosh, A. Goel, J. Kim, S. Kumar, Z. Kong, S.-g. Lee, C.-H. H. Yang, R. Duraiswami, D. Manocha, R. Valle *et al.*, "Audio Flamingo 3: Advancing audio intelligence with fully open large audio language models," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[27] P. Rangappa, A. Carofilis, J. Prakash, S. Kumar, S. Burdisso, S. Madikeri, E. Villatoro-Tello, B. Sharma, P. Motlicek, K. Hacioglu *et al.*, "Efficient data selection for domain adaptation of ASR using pseudo-labels and multi-stage filtering," in *Proc. Interspeech*, 2025, pp. 4928–4932.

[28] M. Bhattacharjee, P. Motlicek, S. Madikeri, H. Helmke, O. Ohneiser, M. Kleinert, and H. Ehr, "Minimum effort adaptation of automatic speech recognition system in air traffic management," *European Journal of Transport and Infrastructure Research*, vol. 42, no. 4, pp. 133–153, 2024.

[29] S.-I. Ng, L. Xu, I. Siegert, N. Cummins, N. R. Benway, J. Liss, and V. Berisha, "An end-to-end overview of clinical speech AI," *IEEE Transactions on Audio, Speech and Language Processing*, 2026.

[30] D. Hwang, A. Misra, Z. Huo, N. Siddhartha, S. Garg, D. Qiu, K. C. Sim, T. Strohman, F. Beaufays, and Y. He, "Large-scale ASR domain adaptation using self-and semi-supervised learning," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6627–6631.

[31] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[32] H. Wang, F. Yu, X. Shi, Y. Wang, S. Zhang, and M. Li, "Slidespeech: A large scale slide-enriched audio-visual corpus," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 076–11 080.