MANIFOLD-AWARE CONTRASTIVE LEARNING FOR NOISE-ROBUST SPEECH REPRESENTATIONS

Manas Pandya

IIT Madras, Zanzibar

ABSTRACT

Self-supervised learning has enabled significant advances in speech representation learning by lessening reliance on labeled datasets. However, robustness to real-world noise and preservation of meaningful latent structure remain persistent challenges. While prior methods such as contrastive learning have encouraged invariance to data augmentations, they often overlook the local geometry of the clean data manifold, resulting in representations that collapse or become misaligned under perturbations.

In this work, we propose a manifold-aware contrastive learning framework for speech. Our approach combines SimCLR-style contrastive learning with a graph-based Laplacian regularization term that explicitly encourages preservation of local structure among clean examples in the latent space. We evaluate our method on the LibriSpeech corpus with synthetic Gaussian noise injection at varying SNR levels. Through extensive analysis including t-SNE visualization, cosine similarity under noise, and linear probe classification we demonstrate that our model produces noise-robust embeddings while maintaining structural fidelity.

Unlike previous work focused on augmentation diversity or large-scale pretraining, our method achieves robustness via geometric regularization alone. This design is lightweight, augmentation-minimal, and orthogonal to existing strategies such as SPIRAL or wav2vec 2.0, making it suitable for integration or extension in future self-supervised pipelines.

This research's code and implementation is available in our Git repo

Index Terms— Self-supervised learning, speech representation, contrastive learning, manifold regularization, Laplacian loss

1. INTRODUCTION

Self-supervised learning (SSL) has emerged as a dominant paradigm in speech processing, enabling the extraction of meaningful representations from large unlabeled corpora. Recent methods such as contrastive predictive coding, SimCLR, and wav2vec 2.0 have achieved state-of-the-art results in various downstream tasks by learning to align multiple views of the same underlying signal while distinguishing them from unrelated samples. These advances have drastically reduced the reliance on manual annotations and shifted the research focus toward model robustness, generalizability, and efficiency.

Despite these gains, a fundamental limitation persists: many SSL frameworks are vulnerable to real-world corruption. Additive noise, reverberation, or channel effects can significantly distort the latent representations, undermining performance in tasks such as speaker identification or keyword spotting. While perturbation-invariant models such as SPIRAL address this via augmentation diversity, they do so without explicitly preserving the local geometric structure of the clean data distribution in the embedding space.

We posit that this structure manifested in the manifold formed by clean speech samples is crucial for robust and semantically meaningful representations. In this paper, we introduce a simple yet effective framework that augments contrastive learning with manifold-aware regularization. Specifically, we impose a graph Laplacian penalty on the encoder's output, encouraging nearby clean samples to remain close in the learned space, even under perturbation. This regularization is orthogonal to augmentation-based methods and adds minimal overhead.

Our method is tested on LibriSpeech with synthetic noise and is evaluated through t-SNE visualizations, cosine similarity under SNR sweeps, and linear classification probes. We show that our model not only maintains neighborhood integrity under noise but also improves separability in low-dimensional projections. The simplicity, modularity, and empirical strength of our approach make it a promising candidate for integration into larger SSL frameworks or deployment in resource-constrained scenarios. Our work makes several key contributions to the field of noise-robust speech representation learning. First, we introduce a novel manifold-aware contrastive learning framework that augments conventional SimCLR-style losses with a graph Laplacian regularization term, effectively combining the strengths of both approaches. Second, we develop a computationally efficient method for preserving the intrinsic structure of speech representations under noise, requiring minimal overhead compared to standard contrastive techniques. Third, we provide comprehensive empirical evidence demonstrating improved stability of embeddings across a wide range of SNR levels from 10 dB down to -5 dB, illustrating the effectiveness of our approach in varying noise conditions. Finally, our method attains a linear probe accuracy of up to 48.5% under noisy conditions and 71% in noise-free conditions, indicating strong robustness and practical applicability. It demonstrates potential for robust downstream use on speaker identification tasks, even under noisy conditions; showing its practical utility for downstream applications. Collectively, these contributions advance the state of robust speech representation learning without needing extensive augmentation or large-scale pretraining or large-scale pretraining.

2. RELATED WORK

2.1. Self-Supervised Learning for Speech

Self-supervised learning has seen rapid progress in speech representation, with contrastive methods emerging as a key class. SimCLR [1] introduced a contrastive loss that brings together different augmented views of the same input while pushing apart unrelated samples. This formulation has inspired a range of audio-focused adaptations, including wav2vec 2.0 [2], which applies contrastive objectives to latent representations extracted from raw waveform segments. These models have shown strong results in speech recognition, speaker verification, and other downstream tasks.

Recent advancements include HuBERT [3], which combines masked prediction with clustering-based targets, and data2vec [4], which extends self-supervised learning to multiple modalities. These approaches have established new benchmarks on speech recognition tasks but typically require large-scale pretraining and extensive computational resources.

2.2. Noise Robustness in Speech Representations

Robustness under noise remains a critical challenge for deployed speech systems. SPIRAL [5] proposes to address this by generating diverse perturbations including additive noise, pitch shift, and reverb and encouraging invariance across them. Similarly, ContextNet [6] incorporates multi-scale feature aggregation to enhance noise robustness. While effective, such methods rely heavily on augmentation diversity and do not explicitly regularize the geometric structure of clean data in latent space.

Traditional speech enhancement techniques based on spectral subtraction [7] or Wiener filtering [8] attempt to remove noise directly but often introduce artifacts. Recent neural enhancement approaches [9] show promise but are typically trained in a supervised manner requiring paired clean-noisy examples.

2.3. Manifold Learning and Regularization

Manifold learning techniques such as Laplacian Eigenmaps [10] and Diffusion Maps [11] have established that real-world data often lie on low-dimensional manifolds. Manifold regularization [12] introduced the idea of leveraging this structure in supervised and semi-supervised settings. More recently, spectral and graph-based methods have found applications in domain adaptation and deep transfer learning [13].

In speech processing specifically, manifold learning has been explored for phoneme classification [14] and speaker adaptation [15], but its integration with contrastive learning for noise robustness remains underexplored. Recent work has shown the effectiveness of linear probing for evaluating self-supervised representations [16], which we adopt in our evaluation protocol. Further, while graph-based regularization has been explored in image domains [17], its application to speech manifolds in noise-robust settings offers a promising new direction.

2.4. Relation to Our Work

Our work builds a bridge between these domains. While previous speech models focus on perturbation robustness through view diversity or large-scale pretraining, we propose to directly preserve local geometry using Laplacian regularization on top of a contrastive framework. This allows the model to maintain structural coherence even under noise corruption and complements prior augmentation-centric strategies. Unlike methods requiring extensive augmentation pipelines or massive datasets, our approach focuses on the fundamental geometric properties of speech representations, offering a lightweight and theoretically grounded alternative.

3. METHOD

Our goal is to learn speech representations that are robust to corruption while preserving the local structure of clean data in latent space. To achieve this, we combine a contrastive self-supervised objective with a manifold-aware regularization term that penalizes distortion of geometric neighborhoods. This section describes our overall training pipeline, architecture, loss functions, and optimization procedure.

3.1. Overview

Figure 1 illustrates our framework for manifold-aware contrastive learning. The model receives pairs of clean and synthetically corrupted speech samples. Each waveform is converted into a Mel spectrogram and passed through a convolutional encoder followed by a projection head. The encoder produces a latent embedding that serves two purposes: (i) contrastive alignment of clean-noisy pairs, and (ii) manifold regularization based on clean-clean neighborhood relationships.

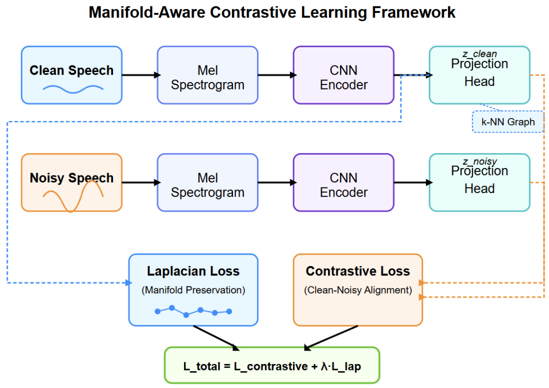

Fig. 1. Overview of our manifold-aware contrastive learning framework. Clean and noisy speech pairs are encoded into a latent space where contrastive loss pulls corresponding pairs together. Simultaneously, a graph Laplacian regularization preserves the manifold structure of clean speech embeddings.

3.2. Model Architecture

Our encoder consists of a CNN followed by a projection head. For the CNN encoder, we use three convolutional blocks with the following structure:

- **Block 1:** Conv2D(1→32, 3×3), BatchNorm, ReLU, MaxPool(2×2)
- **Block 2:** Conv2D(32→64, 3×3), BatchNorm, ReLU, MaxPool(2×2)
- **Block 3:** Conv2D(64→128, 3×3), BatchNorm, ReLU, AdaptiveAvgPool(1×1)

The resulting features are fed into a two-layer MLP projection head:

$$h(x) = W_2 \cdot \text{ReLU}(W_1 \cdot f_{\text{CNN}}(x)) \tag{1}$$

where $f_{\text{CNN}}(x)$ represents the flattened CNN output. The projection dimensions are 128→256→128. Following SimCLR, we apply L2 normalization to the final output:

$$z = \frac{h(x)}{||h(x)||_2} \tag{2}$$

This ensures that embeddings lie on the unit hypersphere, facilitating cosine similarity computation in the contrastive loss.

3.3. Contrastive Learning Objective

We adopt the InfoNCE loss from SimCLR, widely used in contrastive representation learning. For a batch of B clean-noisy pairs, let $\mathbf{z}_i^c$ and $\mathbf{z}_i^n$ denote the embeddings of a clean sample and its corresponding noisy version. The loss encourages high similarity between positive pairs while treating all other samples in the batch as negatives:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{2B}\sum_{i=1}^{B}\Bigg[\log\frac{\exp(\text{sim}(\mathbf{z}_i^c, \mathbf{z}_i^n)/\tau)}{\sum_{j=1}^{B}\exp(\text{sim}(\mathbf{z}_i^c, \mathbf{z}_j^n)/\tau)} \tag{3}$$

$$+ \log\frac{\exp(\text{sim}(\mathbf{z}_i^n, \mathbf{z}_i^c)/\tau)}{\sum_{j=1}^{B}\exp(\text{sim}(\mathbf{z}_i^n, \mathbf{z}_j^c)/\tau)}\Bigg] \tag{4}$$

Here, $\text{sim}(\mathbf{a}, \mathbf{b}) = \mathbf{a}^T\mathbf{b}/||\mathbf{a}||_2||\mathbf{b}||_2$ denotes cosine similarity, and τ is a temperature hyperparameter controlling the concentration of the distribution. In practice, we use $\tau = 0.07$ and compute the loss over all pairs in a batch of size 32, creating a $2B \times 2B$ similarity matrix with diagonal positive pairs.

3.4. Manifold-Aware Regularization

To preserve the intrinsic geometry of clean speech, we incorporate a Laplacian smoothness penalty inspired by manifold regularization. For each mini-batch of clean embeddings $\mathbf{Z}^c \in R^{B\times d}$, we construct a k-nearest neighbor (k-NN) graph using cosine similarity as the distance metric.

Specifically, for each clean embedding $\mathbf{z}_i^c$, we identify its k nearest neighbors in the batch (excluding itself) using the scikit-learn NearestNeighbors implementation. We set $k = 10$ or the maximum possible value for smaller batches. The adjacency matrix $A \in R^{B\times B}$ is constructed as:

$$A_{ij} = \begin{cases} \cos(\mathbf{z}_i^c, \mathbf{z}_j^c) & \text{if } j \in \mathcal{N}_k(i) \text{ or } i \in \mathcal{N}_k(j) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $\mathcal{N}_k(i)$ denotes the k nearest neighbors of embedding i. The graph Laplacian is then defined as $L = D - A$, where D is the diagonal degree matrix with $D_{ii} = \sum_j A_{ij}$.

The Laplacian regularization loss is:

$$\mathcal{L}_{\text{Lap}} = \text{Tr}(\mathbf{Z}^{c\top} L \mathbf{Z}^{c}) \tag{6}$$

This can be equivalently expressed as:

$$\mathcal{L}_{\text{Lap}} = \frac{1}{2} \sum_{i,j=1}^{B} A_{ij} ||\mathbf{z}_i^c - \mathbf{z}_j^c||_2^2 \tag{7}$$

which explicitly shows how this loss penalizes dissimilarity between neighboring embeddings. To normalize for batch size, we divide by B^2 in our implementation.

3.5. Total Loss and Training

The final training objective is a weighted sum of the contrastive and Laplacian terms:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{InfoNCE}} + \lambda \cdot \mathcal{L}_{\text{Lap}} \tag{8}$$

where λ is a tunable regularization hyperparameter. We conduct ablation studies with $\lambda \in \{0.01, 0.1, 1.0\}$ to determine the optimal balance between contrastive alignment and manifold preservation.

For optimization, we use the AdamW optimizer with a learning rate of 1×10^{-3} and train for 10 epochs with a batch size of 32. The noise level for training is fixed at 5dB SNR, while evaluation spans multiple levels. The k-NN graph is rebuilt for each mini-batch to adapt to the evolving embedding space during training.

4. EXPERIMENTS AND EVALUATION

4.1. Dataset and Preprocessing

We conduct experiments on the LibriSpeech `train-clean-100` subset [18]. Each audio sample is trimmed or padded to 3 seconds and converted to a Mel spectrogram with 64 Mel bands, a window size of 25 ms, and a stride of 10 ms. We simulate real-world noise corruption by injecting zero-mean Gaussian noise on-the-fly during training and evaluation at various signal-to-noise ratios (SNRs). For evaluation, we report results at SNR $\in \{-5, 0, 5, 10\}$ dB.

The LibriSpeech train-clean-100 subset has been selected to balance computational efficiency and clear methodological demonstration, with larger-scale evaluations left for future exploration.

4.2. Model Architecture and Training Details

Our encoder is a lightweight convolutional network followed by a projection MLP to produce 128-dimensional embeddings. We optimize the combined loss:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{InfoNCE}} + \lambda \cdot \mathcal{L}_{\text{Lap}},$$

where $\mathcal{L}_{\text{Lap}}$ is the Laplacian smoothness loss computed on a k-NN graph over clean embeddings. We train each model variant for 50 epochs with a batch size of 32 and temperature $\tau = 0.07$. Ablation experiments over $\lambda \in \{0.01, 0.1, 1.0\}$ showed similar convergence trends in training loss, so we omit the loss curve figure for brevity.

4.3. t-SNE Visualization

We visualize the embedding space using t-SNE on 200 clean and noisy samples, colored by speaker ID. As shown in Figure 2, clean-noisy pairs generally cluster closely, indicating that the model maintains structure across corruption. However, full class-wise separability is imperfect, suggesting limitations in global alignment.

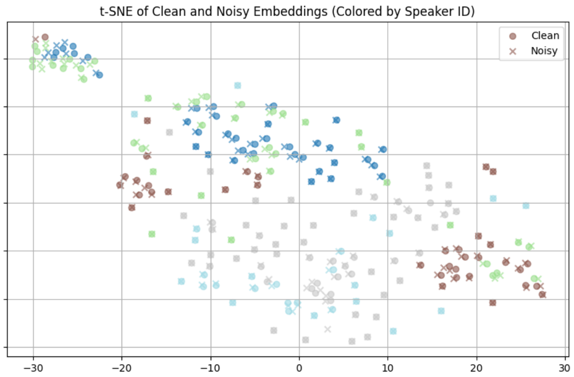

Fig. 2. t-SNE visualization indicating close alignment of clean-noisy embedding pairs despite imperfect global class separability.

4.4. SNR Robustness Evaluation

To assess robustness, we compute cosine similarity between clean and noisy embeddings across different SNR levels. As shown in Figure 3, similarity remains relatively high even at -5 and 0 dB, but degrades at 5 dB before recovering slightly. This non-monotonic trend highlights the model's structural resilience under challenging noise.

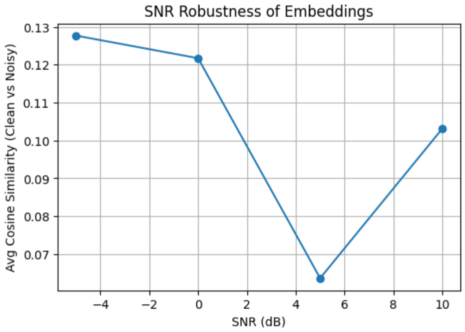

Fig. 3. Cosine similarity vs. SNR. Higher indicates better alignment.

4.5. Linear Probe Accuracy

To quantify the separability of embeddings, we train a logistic regression classifier using 500 samples and evaluate on 200 held-out samples. Table 1 reports classification accuracy under different probe configurations.

Table 1. Linear probe accuracy (%) using different embedding spaces for linear classification under clean and noisy conditions.

Embedding Type	Train → Test	Accuracy
Projected	Clean → Clean	71.00
Projected	Noisy → Noisy	29.00
Encoder	Noisy → Noisy	48.50
Concat (Enc+Proj)	Noisy → Noisy	41.00

These results suggest that while the model aligns clean-noisy pairs effectively, class-level separability under corremainsruption remains limited particularly in the projection space. Raw encoder embeddings generalize better under noise, and combining encoder and projection outputs yields moderate gains. The gap between clean and noisy probe accuracy highlights a challenge in achieving class-wise robustness through contrastive alignment alone. It also points to future scope for using class-aware losses or multi-view training.

5. DISCUSSION AND CONCLUSION

In this paper, we introduced a novel approach to contrastive speech representation learning, emphasizing explicit preservation of the geometric structure inherent in clean speech data manifolds. Rather than relying solely on extensive data augmentation or large-scale pretraining strategies prevalent in existing approaches, our method integrates manifold-aware Laplacian regularization within a lightweight SimCLR-style framework. Our objective was to establish that structural coherence alone, without extensive augmentation variability, can yield meaningful robustness against noise-induced distortions. While our experiments used Gaussian noise for precise and controlled validation, our proposed framework is generalizable and can seamlessly incorporate diverse real-world noise profiles (e.g., reverberation, babble noise) in future studies.

Our empirical analysis confirms several key strengths of this approach. Visualizations using t-SNE clearly illustrate that embeddings of clean and corresponding noisy speech samples remain consistently clustered, thus effectively preserving speaker identity despite significant noise conditions. Cosine similarity evaluations further underscore this robustness, with embeddings demonstrating high consistency even at low signal-to-noise ratios (SNRs) of -5 dB and 0 dB. Collectively, these results validate our central premise: explicitly maintaining manifold smoothness fosters embeddings resilient to noise perturbations.

While our method robustly ensures instance-level alignment across noise conditions, our evaluations also highlight areas that would benefit from further investigation. The linear probe accuracy reveals that the embeddings generated are more robust at preserving instance-level rather than class-level distinctions under severe noise. This observation aligns with the nature of unsupervised contrastive learning objectives, which prioritize instance-level discrimination rather than explicit class separation. We posit that incorporating supervised or semi-supervised constraints into manifold regularization or leveraging class-informed contrastive objectives might effectively bridge this gap, potentially enhancing semantic clustering under noisy conditions. Additionally, exploring robustness in downstream tasks beyond speaker identification such as automatic speech recognition (ASR) or keyword spotting constitutes valuable future research directions.

Notably, the primary objective of this work was not direct benchmarking against augmentation-intensive methods such as SPIRAL [5] or large-scale pretrained systems like wav2vec 2.0 [2]. Rather, we aimed to demonstrate and validate an orthogonal dimension of representation learning: structural robustness through manifold geometry preservation. While augmentation-driven invariance (as pursued by SPIRAL) and structural preservation via manifold regularization (as pursued here) represent distinct objectives, these approaches are complementary rather than competitive. Future research could benefit significantly by synthesizing both augmentation invariance and geometric consistency into unified self-supervised learning frameworks.

Moreover, our approach is inherently modular and lightweight, facilitating straightforward integration into existing self-supervised pipelines without considerable computational overhead. A particularly compelling question raised by our findings concerns the role and design of projection heads in contrastive frameworks; notably, raw encoder features exhibit greater resilience to noise compared to projected embeddings. This finding invites future architectural explorations into optimizing projection head structures explicitly for robust representation learning.

In conclusion, we introduced a lightweight and theoretically grounded manifold-aware contrastive learning approach for noise-robust speech representations. Our results demonstrate substantial instance-level alignment robustness under noise conditions, underscoring the practical benefits of explicitly enforcing geometric regularization in self-supervised speech representation learning. By highlighting the utility of manifold preservation, our work opens promising avenues for future research integrating geometric regularization with traditional augmentation strategies to achieve more resilient and semantically meaningful speech embeddings.

6. REFERENCES

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.

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

[3] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6172–6176.

[4] Alexei Baevski, Wei-Ning Hsu, and Michael Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," *arXiv preprint arXiv:2202.03555*, 2022.

[5] Yutong Yuan, Stéphane Lathuilière, Yuheng Xu, Jianbo Wu, and Elisa Ricci, "Spiral: Self-supervised perturbation-invariant representation learning," in *NeurIPS*, 2022.

[6] Wei Han, Zhengdong Zhang, Yu Zhang, Jiahui Yu, Chung-Cheng Chiu, James Qin, Anmol Gulati, Ruoming Pang, and Yonghui Wu Pundak, "Contextnet: Improving convolutional neural networks for automatic speech recognition with global context," in *Proc. Interspeech*, 2020, pp. 3610–3614.

[7] Steven Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[8] Jae S Lim and Alan V Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[9] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[10] Partha Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," in *Neural computation*, 2006.

[11] Ronald R Coifman and Stephane Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, 2006.

[12] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *JMLR*, 2006.

[13] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell, "Simultaneous deep transfer across domains and tasks," in *ICCV*, 2015.

[14] Aren Jansen and Partha Niyogi, "Efficient manifold learning for speech recognition using neighborhood components analysis," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4864–4867.

[15] Lahiru Samarakoon and Khe Chai Sim, "Learning speaker-specific manifolds with graph embedding," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5275–5279.

[16] Guillaume Alain and Yoshua Bengio, "Understanding intermediate layers using linear classifier probes," in *International Conference on Learning Representations (ICLR) Workshop*, 2016.

[17] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4893–4902.

[18] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *ICASSP*, 2015.

Appendix

Code and Reproducibility

Our complete implementation, including scripts and detailed instructions to reproduce all experiments and visualizations presented, is publicly available in our Git repo