Channel-Aware Domain-Adaptive Generative Adversarial Network for Robust Speech Recognition

Chien-Chun Wang*, Li-Wei Chen‡, Cheng-Kang Chou‡, Hung-Shin Lee‡, Berlin Chen*, and Hsin-Min Wang†
* Dept. Computer Science and Information Engineering, National Taiwan Normal University, Taiwan
† Institute of Computer Science, Academia Sinica, Taiwan
‡ United-Link Co., Ltd., Taiwan

***Abstract*—While pre-trained automatic speech recognition (ASR) systems demonstrate impressive performance on matched domains, their performance often degrades when confronted with channel mismatch stemming from unseen recording environments and conditions. To mitigate this issue, we propose a novel channel-aware data simulation method for robust ASR training. Our method harnesses the synergistic power of channel-extractive techniques and generative adversarial networks (GANs). We first train a channel encoder capable of extracting embeddings from arbitrary audio. On top of this, channel embeddings are extracted using a minimal amount of target-domain data and used to guide a GAN-based speech synthesizer. This synthesizer generates speech that faithfully preserves the phonetic content of the input while mimicking the channel characteristics of the target domain. We evaluate our method on the challenging Hakka Across Taiwan (HAT) and Taiwanese Across Taiwan (TAT) corpora, achieving relative character error rate (CER) reductions of 20.02% and 9.64%, respectively, compared to the baselines. These results highlight the efficacy of our channel-aware data simulation method for bridging the gap between source- and target-domain acoustics.**

***Index Terms*—automatic speech recognition, channel compensation, domain adaptation, data simulation**

I. Introduction

Automatic speech recognition (ASR) has become an indispensable technology, powering applications ranging from virtual assistants to transcription services. Recent advancements in deep learning, particularly with architectures such as convolutional neural networks (CNNs) [1], long short-term memory networks (LSTMs) [2], [3], Transformers [4]–[7], and Conformers [8]–[10], have significantly improved ASR accuracy across various conditions. However, these architectures often remain susceptible to performance degradation caused by channel mismatch—a discrepancy in acoustic characteristics between training and test data due to differences in recording equipment.

This vulnerability is particularly evident in scenarios like teleconferencing, where diverse microphones, ranging from professional condenser microphones to built-in webcams, introduce significant variations in signal quality. This mismatch can drastically degrade performance. For instance, as shown in Table I, using $\text{Whisper}_{\text{Tiny}}$ [7] fine-tuned on Condenser data from the Hakka Across Taiwan (HAT) [11] and Taiwanese Across Taiwan (TAT) [12] corpora results in drastically increased character error rates (CERs) when evaluated on other microphone types, highlighting the urgent need for more channel-robust ASR systems.

To address this challenge, researchers have explored domain adaptation techniques [13]–[18] that aim to bridge the gap between training and test distributions. While these techniques have shown promise, they often involve complex training procedures or may not fully exploit the underlying relationship between domains. Recently, data simulation has emerged as an alternative approach [19]–[21], generating synthetic target-domain data from source-domain data to facilitate model adaptation without requiring paired samples. However, existing data simulation approaches primarily focus on mitigating noise-related mismatches and lack the sophistication to effectively tackle channel discrepancies. This highlights a crucial need for novel methods specifically designed to enhance ASR robustness across diverse recording channels.

TABLE I
CERs (%) WITH RESPECT TO VARIOUS TEST CHANNELS.

Test Channels	HAT	TAT
Condenser	**2.43**	**9.39**
Lavalier	2.94	9.56
iPhone	4.10	11.21
Android	4.70	12.76

To overcome the limitations of existing approaches, we introduce CADA-GAN, a novel Channel-Aware Domain-Adaptive Generative Adversarial Network, designed to enhance ASR robustness to channel mismatch. Our method leverages a two-step process: channel embedding extraction and domain-adaptive speech synthesis.

First, a channel encoder is trained to extract detailed channel embeddings from target-domain speech. These embeddings capture the unique acoustic characteristics of the target recording environment. Next, a GAN architecture, guided by the extracted embeddings, learns a source-to-target domain transformation. The generator synthesizes speech that accurately replicates target-domain channel characteristics while preserving the phonetic content of the source speech. The discriminator ensures the authenticity of the generated samples, further refining the domain adaptation process.

Moreover, CADA-GAN requires only a small amount of unpaired target-domain data during training, making it highly practical for real-world scenarios. At inference time, only the generator and channel encoder are utilized, eliminating the need for additional transcriptions. The generator synthesizes abundant training data by randomly pairing source speech with the limited target speech, inheriting the source transcriptions. This data augmentation strategy allows for fine-tuning any ASR model to significantly improve channel robustness.

Unlike previous GAN-based approaches, such as UNA-GAN [21], which adopt a global approach to domain adaptation, CADA-GAN generates customized, domain-specific speech by explicitly conditioning the generation process on channel embeddings extracted from individual target utterances. This fine-grained control enables it to precisely align the synthesized speech with the target channel characteristics. We evaluate CADA-GAN on the challenging HAT and TAT corpora. Our method demonstrates its effectiveness by achieving relative CER reductions of 20.02% and 9.64%, respectively, compared to the baseline trained solely on the source domain.

II. Proposed Method

Fig. 1 illustrates the architecture of our proposed CADA-GAN, which comprises three key components: a generator (G), a discriminator (D), and a channel encoder (E). The process begins with the

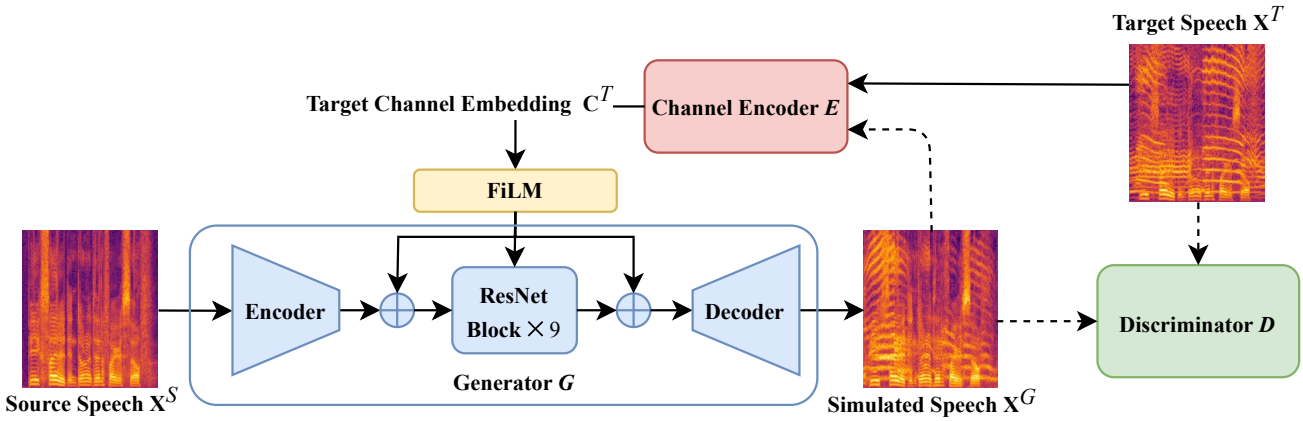


Fig. 1. The architecture of our proposed method, CADA-GAN. The dotted arrows indicate that during the training phase, simulated speech $\mathbf{X}^G$ is used together with target speech $\mathbf{X}^T$ to 1) train the discriminator, and 2) contribute to channel reconstruction. The $\bigoplus$ operator denotes element-wise tensor addition.

channel encoder, which extracts a channel embedding ($\mathbf{C}^T$) from a target-domain spectrogram ($\mathbf{X}^T$). This embedding encapsulates the distinct acoustic characteristics of the target recording environment. The generator then utilizes this embedding alongside a source-domain spectrogram ($\mathbf{X}^S$) to synthesize a simulated spectrogram ($\mathbf{X}^G$) that mimics the target-domain channel characteristics while preserving the phonetic content of the source speech. Finally, the discriminator distinguishes between real target spectrograms and simulated spectrograms, providing feedback to the generator during training.

A. Generator and Discriminator

The generator (G) is designed to transform a source-domain spectrogram ($\mathbf{X}^S$) into a simulated target-domain spectrogram ($\mathbf{X}^G$). It achieves this by first processing the source spectrogram through two 2D downsampling convolutional layers (kernel size: 3×3, stride: 2×2), followed by nine residual blocks to capture deep, hierarchical representations. Each residual block consists of two convolutional layers (kernel size: 3×3, stride: 1×1) and a dropout layer to prevent overfitting. Finally, two transposed convolutional layers (kernel size: 3×3, stride: 2×2) upsample the learned representations to generate the simulated spectrogram.

The discriminator (D) plays a crucial role in ensuring the authenticity of the generated spectrograms. It distinguishes between real target spectrograms ($\mathbf{X}^T$) and simulated ones using five 2D convolutional layers (kernel size: 4×4) with Leaky ReLU activation functions. The stride is set to 2×2 for the first three layers and 1×1 for the last two, gradually increasing the receptive field. The adversarial loss employed during training is defined as follows:

$$\begin{aligned}\mathcal{L}_{adv}(G,D,\mathbf{X}^T,\mathbf{X}^S,\mathbf{C}^T) &= \mathbb{E}_{\mathbf{x}\sim\mathbf{X}^T}\left[\log D(\mathbf{x})\right] \\ &+ \mathbb{E}_{\mathbf{x}\sim\mathbf{X}^S,\mathbf{c}\sim\mathbf{C}^T}\left[\log(1-D(G(\mathbf{x},\mathbf{c})))\right].\end{aligned} \tag{1}$$

This adversarial loss encourages the generator to produce spectrograms that closely resemble real target spectrograms, while the discriminator learns to identify subtle differences that distinguish real from simulated data. This adversarial training process compels the generator to continuously improve its ability to generate realistic and domain-specific speech.

B. Channel Encoder

Drawing inspiration from recent advancements in "aware" techniques [17], [22], [23], we introduce a dedicated channel encoder (E) to extract channel embeddings ($\mathbf{C}^T$) from the final layer of a pre-trained model. Unlike conventional approaches that directly utilize raw spectrograms as channel information, our channel encoder focuses on capturing high-level, discriminative channel characteristics.

Specifically, our channel encoder leverages a MFA-Conformer model [24], which is pre-trained on the HAT corpus [11]. The training data consists of recordings from speakers uttering identical content using different microphones at the same time, ensuring that the model is unaffected by speech content or speaker identity. By excluding both the source and target channels used in the main experiment, we enhance the model's ability to classify various channels based purely on their acoustic properties. This strategy enables the channel encoder to effectively capture detailed channel characteristics without incorporating phonetic information, leading to more robust and generalizable channel embeddings.

The channel embeddings are integrated to the generator using Feature-wise Linear Modulation (FiLM) [25]. The embeddings undergoes separate linear transformations to produce weights and biases. These are used to modulate the output features from specific layers in the generator. To further ensure that the generated spectrograms accurately reflect the target-domain channel characteristics, we introduce a channel reconstruction loss:

$$\mathcal{L}_{ch}(G,\mathbf{X}^S,\mathbf{C}^T) = \mathbb{E}_{\mathbf{x}\sim\mathbf{X}^S,\mathbf{c}\sim\mathbf{C}^T}\left[\|\mathbf{c}-E(G(\mathbf{x},\mathbf{c}))\|_1\right]. \tag{2}$$

This loss function encourages the generator to synthesize spectrograms that, when processed by the channel encoder, yield embeddings highly similar to the original target channel embeddings. This reinforces the channel awareness of the generator, ensuring that the generated speech accurately captures the subtle nuances of the target recording environment.

C. Patch-wise Contrastive Learning

To maintain linguistic consistency between the simulated and source speech, we apply patch-wise contrastive learning [26]. This approach maximizes mutual information, particularly shared speech content, between source and simulated spectrograms. Specifically, we utilize the generator to extract deep features from both spectrograms. A small patch from the simulated representation serves as the "query", with the corresponding patch from the source as the "positive" sample. We select 256 patches from the source as "negative" samples. These patches are projected into a lower-dimensional space using two linear layers with 256 units each and ReLU activation. The contrastive loss, computed across five generator layers, measures the cross-entropy loss between the "query" patch and both positive

and negative patches. This encourages high similarity between corresponding patches in the source and simulated spectrograms, while distinguishing them from random patches. The loss is defined as:

$$\mathcal{L}_{pcl}(G, \mathbf{X}^S) = \sum_{l=1}^{L} \sum_{i=1}^{I} -\log \frac{e^{\left(\hat{z}_l^i \cdot z_l^i / \tau\right)}}{e^{\left(\hat{z}_l^i \cdot z_l^i / \tau\right)} + \sum_{j=1}^{J} e^{\left(\hat{z}_l^i \cdot z_l^j / \tau\right)}}, \tag{3}$$

where z_l^i represents the i^{th} positive patch from source representations at the l^{th} layer of the generator, $\hat{z}_l^i$ denotes the corresponding patch from simulated representations, and z_l^j refers to the j^{th} negative patch from simulated representations at the same layer. The temperature parameter τ regulates the contrastive learning process. This loss function is applied to both source ($\mathcal{L}_{pcl}(G, \mathbf{X}^S)$) and target ($\mathcal{L}_{pcl}(G, \mathbf{X}^T)$) spectrograms to maintain consistent speech content and minimize unnecessary changes.

D. Training Objective and Adaptation Process

The training objective is to optimize GAN using a comprehensive loss function that includes the adversarial loss, patch-wise contrastive learning losses for both source and target spectrograms, and the channel reconstruction loss. The total loss function is defined as:

$$\begin{aligned} \mathcal{L}_{total} =& \mathcal{L}_{adv}(G, D, \mathbf{X}^T, \mathbf{X}^S, \mathbf{C}^T) + \mathcal{L}_{pcl}(G, \mathbf{X}^S) \\ &+ \mathcal{L}_{pcl}(G, \mathbf{X}^T) + \lambda_{ch}\mathcal{L}_{ch}(G, \mathbf{X}^S, \mathbf{C}^T), \end{aligned} \tag{4}$$

where λ_{ch} weights the channel reconstruction loss.

To address limited target-domain data, an equal amount of speech is randomly sampled from the source domain. The model is trained on this unpaired dataset, optimizing with the total loss function $\mathcal{L}_{total}$ as specified in (4). After training, the generator serves as a domain converter $F^{S\sim T}$, transforming $\mathbf{X}^S$ to $\mathbf{X}^T$, with the pre-trained channel encoder aiding in channel simulation. This simulation uses plentiful source speech and randomly selected target speech from training. The augmented data enhances the fine-tuning of ASR models without requiring additional transcriptions.

III. Experiments

We conducted extensive experiments on two benchmark datasets to evaluate the efficacy of our proposed CADA-GAN method for domain-adaptive ASR.

HAT [11]: The HAT corpus comprises hundred thousands sets of recordings, where each set was uttered by the same speaker with identical speech content using eight different microphones, reflecting diverse recording conditions. The recorders include an **iPhone**, an **Android** phone, a **Webcam**, a professional **Condenser** microphone, a **Lavalier** microphone, a cheap PC microphone (**PC-Mic**), and an X-Y stereo microphone (**ZOOM-X** and **ZOOM-Y**). The dataset contains 97,385 training sets (779,080 utterances in total) and 4,559 test sets (36,472 utterances in total). We selected recordings from **Condenser** as the source domain and those from **Webcam** as the target domain due to their significant acoustic mismatch. To train our GAN, we randomly sampled 40 utterances from each domain, demonstrating the method's effectiveness with limited target-domain data.

TAT [12]: To further validate that our channel encoder does not inadvertently capture phonetic information, we conducted additional experiments on the TAT corpus. This dataset is similar to HAT but excludes recordings from **Webcam** and **PC-Mic**. We used **Condenser** recordings as the source domain and **Android** recordings as the target domain for this evaluation. There is no information indicating that HAT and TAT use the same type and brand of devices.

TABLE II
CERs (%) and their relative CER reductions (Rel. %) on HAT and TAT with respect to various methods.

Model	HAT		TAT	
	CER	**Rel.**	**CER**	**Rel.**
Vanilla ASR	10.24	-	12.76	-
UNA-GAN [21]	9.76	4.69	11.82	7.37
CADA-GAN	**8.19**	**20.02**	**11.53**	**9.64**
Topline ASR	3.88	62.11	10.30	19.28

A. Backbone ASR Models

We employ Whisper [7], an ASR system developed by OpenAI, as our downstream evaluation model. Whisper, based on a large-scale Transformer architecture, is trained on a massive dataset of 680,000 hours of multilingual speech data, showcasing robust performance across various languages and acoustic conditions. Given the resource constraints of edge devices, we specifically chose Whisper$_{\text{Tiny}}$ as our downstream ASR model. This lightweight version maintains the impressive performance of Whisper while being optimized for deployment on devices with limited computational resources.

To comprehensively evaluate the impact of our proposed method, we trained four variants of the ASR model:

1) **Vanilla ASR (Baseline)**: The Whisper$_{\text{Tiny}}$ fine-tuned solely on the source-domain data without any channel compensation.
2) **UNA-GAN**: The Whisper$_{\text{Tiny}}$ fine-tuned using a dataset generated by UNA-GAN [21], representing a recent baseline approach.
3) **CADA-GAN**: The Whisper$_{\text{Tiny}}$ fine-tuned with the augmented dataset generated by CADA-GAN, incorporating both source and simulated target-domain data.
4) **Topline ASR**: The Whisper$_{\text{Tiny}}$ fine-tuned directly on the target-domain data, simulating an ideal scenario with abundant labeled target-domain data.

B. Configuration

To ensure effective channel information extraction, we segmented the input speech spectrograms into frames of 129×128 dimensions. We trained the CADA-GAN model for 400 epochs, optimizing the balance between adversarial training and channel reconstruction with a weighting factor of $\lambda_{ch} = 0.5$. The Adam optimizer [27] was employed, with an initial learning rate of 0.0002, for stable training. For domain adaptation, the Whisper$_{\text{Tiny}}$ model was fine-tuned for 10 epochs using a learning rate of 0.0001.

IV. Results and Discussion

We present the results of our experiments, comparing the performance of CADA-GAN with the baseline models on both HAT and TAT corpora. Additionally, ablation studies were conducted to analyze the contribution of individual components in our method.

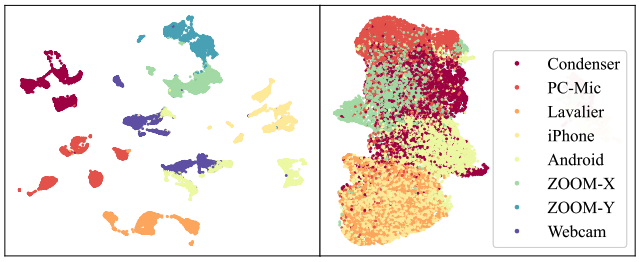
A. Main Results on HAT and TAT

Table II presents the CERs achieved by all ASR models on the HAT and TAT corpora. CADA-GAN demonstrates substantial improvements over the baseline approaches, achieving a remarkable 20.02% relative CER reduction on the HAT corpus and a 9.64% relative CER reduction on the TAT corpus compared to the Vanilla ASR model. These results highlight the effectiveness of incorporating the pre-trained channel encoder within the domain adaptation framework.

Furthermore, the consistent performance gains observed on both HAT and TAT, despite the channel encoder being trained solely on the HAT corpus, underscores its ability to capture and leverage channel-specific features effectively while remaining agnostic to phonetic

TABLE III
CERs (%) ON HAT AND TAT WITH RESPECT TO ABLATION STUDIES.

Model	HAT	TAT
CADA-GAN	**8.19**	**11.24**
- $\mathcal{L}_{ch}$	8.77	11.59
- Embeddings	9.05	11.65



(a) HAT (b) TAT

Fig. 2. The UMAP visualization of channel embeddings extracted from eight channel types in the HAT corpus and six channel types in the TAT corpus.

TABLE IV
MOSs OF SIMULATED SPEECH ON HAT AND TAT.

Model	HAT	TAT
UNA-GAN [21]	2.90 ± 0.75	2.55 ± 1.11
CADA-GAN	**4.06 ± 0.71**	**3.09 ± 1.06**

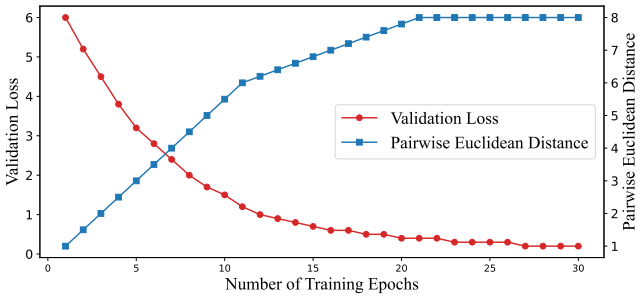


Fig. 3. Validation loss of our channel encoder on the HAT corpus, alongside the average pairwise Euclidean distance between channel embeddings, with respect to the number of training epochs.

information. This crucial capability enables the model to generalize well across diverse languages and unseen channel conditions.

B. Ablation Studies

To delve deeper into the contribution of each component within CADA-GAN, we conducted ablation studies, summarized in Table III. Removing the channel reconstruction loss (- $\mathcal{L}_{ch}$) during training resulted in a marginal performance decline, suggesting that while this loss aids in maintaining channel fidelity, its overall impact on speech recognition accuracy is relatively small.

Conversely, omitting the channel embeddings (- Embeddings) during the generation process led to a significant drop in performance. This substantial decrease emphasizes the critical role channel embeddings play in accurately capturing and transferring channel-specific characteristics from the target domain, even though our channel encoder was trained solely on the HAT corpus. These embeddings are essential for enhancing the model's robustness and enabling accurate speech recognition across different channel conditions.

C. UMAP Visualization of Channel Embeddings

To gain further insights into the workings of CADA-GAN, we visualized the learned channel embeddings and evaluate the perceptual quality of the generated speech. Uniform Manifold Approximation and Projection (UMAP) [28] was employed to visualize the channel embeddings extracted from the HAT and TAT corpora, as shown in Fig. 2. In Fig. 2 (a), a clear separation between different channel types in the HAT corpus is observed. This distinct clustering demonstrates the effectiveness of our pre-trained channel encoder in capturing unique acoustic characteristics associated with each microphone. While the separation is less pronounced in Fig. 2 (b) for the TAT corpus (where the encoder was not specifically trained), similar channel types are still effectively grouped together. This observation highlights the generalization ability of our channel encoder across different languages, reinforcing its capacity to learn channel-specific features rather than language-dependent patterns.

D. MOS Evaluation on Simulated Data

To assess the perceptual realism of the generated speech, we conducted a Mean Opinion Score (MOS) evaluation focused specifically on channel characteristics. Ten participants rated the similarity of the perceived recording channel between generated audio samples and target-channel reference recordings on a scale of 1 to 5, with 5 representing the highest similarity. Table IV summarizes the MOS scores for each method. CADA-GAN achieves a significantly higher MOS compared to UNA-GAN, indicating that our method generates speech that more closely resembles the target domain in terms of channel characteristics, despite our channel encoder not being explicitly trained on the TAT corpus. Moreover, the smaller standard deviation observed for CADA-GAN suggests a higher consistency in the perceptual quality of the generated speech.

E. Analysis of Validation Loss and Embedding Distance

To quantitatively assess the channel discrimination capability of our encoder, we analyzed the evolution of both the validation loss and the average pairwise Euclidean distance between channel embeddings during training. The pairwise distance, computed using the validation set, was averaged between embeddings of the same utterance set (same speaker and content) recorded across different channels. As shown in Fig. 3, the validation loss (red curve) decreases steadily over 30 epochs, while the pairwise distance (blue curve) increases, indicating the encoder is effectively learning distinct channel characteristics. This positive trend supports the critical role of the channel encoder in the success of CADA-GAN.

V. CONCLUSION AND FUTURE WORK

This study[1] presents CADA-GAN, a novel method for channel compensation in robust ASR. By integrating a channel encoder with a GAN architecture, CADA-GAN effectively addresses channel mismatch, improving ASR generalization to unseen conditions. Experiments on the HAT and TAT corpora demonstrate that CADA-GAN significantly outperforms strong baselines, achieving substantial CER reductions and higher MOS scores. These results underscore the efficacy of our method in improving both the accuracy and perceptual quality of speech recognition across diverse channel environments.

Future work will involve further validating the effectiveness of CADA-GAN with more advanced ASR models like $\text{Whisper}_{\text{Large}}$ and extending the evaluation to a wider range of challenging datasets. Additionally, we aim to explore integrating CADA-GAN with other domain adaptation techniques to address multiple sources of variability in speech data.

[1]Code: https://github.com/JethroWangSir/CADA-GAN/.

REFERENCES

[1] J. Pan, J. Shapiro, J. Wohlwend, K. J. Han, T. Lei, and T. Ma, "ASAPP-ASR: Multistream CNN and self-attentive SRU for SOTA speech recognition," in *Proc. Interspeech*, 2020.

[2] D. S. Park, Y. Zhang, Y. Jia, W. Han, C.-C. Chiu, B. Li, Y. Wu, and Q. V. Le, "Improved noisy student training for automatic speech recognition," in *Proc. Interspeech*, 2020.

[3] A. Zeyer, A. Merboldt, W. Michel, R. Schlüter, and H. Ney, "Librispeech transducer model with internal language model prior correction," in *Proc. Interspeech*, 2021.

[4] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020.

[5] M. A. Haidar, C. Xing, and M. Rezagholizadeh, "Transformer-based ASR incorporating time-reduction layer and fine-tuning with self-knowledge distillation," 2021, arXiv:2103.09903.

[6] Q. Xu, A. Baevski, T. Likhomanenko, P. Tomasello, A. Conneau, R. Collobert, G. Synnaeve, and M. Auli, "Self-training and pre-training are complementary for speech recognition," in *Proc. ICASSP*, 2021.

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022, arXiv:2212.04356.

[8] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020.

[9] W. Chan, D. Park, C. Lee, Y. Zhang, Q. Le, and M. Norouzi, "SpeechStew: Simply mix all available speech recognition data to train one large neural network," 2021, arXiv:2104.02133.

[10] S. Kim, A. Gholami, A. Shaw, N. Lee, K. Mangalam, J. Malik, M. W. Mahoney, and K. Keutzer, "Squeezeformer: An efficient transformer for automatic speech recognition," in *Proc. NeurIPS*, 2022.

[11] Y.-F. Liao, S.-H. Hwang, Y.-S. Chen, H.-C. Lai, Y.-H. Chung, L.-T. Shen, Y.-C. Huang, C.-J. Huang, H. W. Han, L.-W. Chen, P.-C. Su, and C.-S. Huang, "Taiwanese Hakka across Taiwan corpus and Formosa speech recognition challenge 2023 - Hakka ASR," in *Proc. O-COCOSDA*, 2023.

[12] Y.-F. Liao, J. S. Tsay, P. Kang, H.-L. Khoo, L.-K. Tan, L.-C. Chang, U.-G. Iunn, H.-L. Su, T.-G. Thiann, H.-K. Tiun, and S.-L. Liao, "Taiwanese across Taiwan corpus and its applications," in *Proc. O-COCOSDA*, 2022.

[13] Y.-T. Hsu, Z. Zhu, C.-T. Wang, S.-H. Fang, F. Rudzicz, and Y. Tsao, "Robustness against the channel effect in pathological voice detection," in *Proc. NeurIPS*, 2018.

[14] S. Mun and S. Shon, "Domain mismatch robust acoustic scene classification using channel information conversion," in *Proc. ICASSP*, 2019.

[15] P. Li, G. Li, J. Han, T. Zhi, and D. Wang, "Channel mismatch speaker verification based on deep learning and PLDA," *Journal of Physics: Conference Series*, vol. 1682, no. 1, 2020.

[16] F.-L. Wang, H.-S. Lee, Y. Tsao, and H.-M. Wang, "Disentangling the impacts of language and channel variability on speech separation networks," in *Proc. Interspeech*, 2022.

[17] F.-L. Wang, Y.-F. Cheng, H.-S. Lee, Y. Tsao, and H.-M. Wang, "CasNet: Investigating channel robustness for speech separation," in *Proc. ICASSP*, 2023.

[18] W. Yang, J. Wei, W. Lu, L. Li, and X. Lu, "Robust channel learning for large-scale radio speaker verification," 2024, arXiv:2406.10956.

[19] H. Hu, T. Tan, and Y. Qian, "Generative adversarial networks based data augmentation for noise robust speech recognition," in *Proc. ICASSP*, 2018.

[20] C. Chen, N. Hou, Y. Hu, S. Shirol, and E. S. Chng, "Noise-robust speech recognition with 10 minutes unparalleled in-domain data," in *Proc. ICASSP*, 2022.

[21] C. Chen, Y. Hu, H. Zou, L. Sun, and E. S. Chng, "Unsupervised noise adaptation using data simulation," in *Proc. ICASSP*, 2023.

[22] Y. A. Li, A. Zare, and N. Mesgarani, "StarGANv2-VC: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion," in *Proc. Interspeech*, 2021.

[23] Y. Hu, C. Chen, R. Li, Q. Zhu, and E. S. Chng, "Noise-aware speech enhancement using diffusion probabilistic model," in *Proc. Interspeech*, 2024.

[24] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H.-y. Lee, and H. Meng, "MFA-Conformer: Multi-scale feature aggregation conformer for automatic speaker verification," in *Proc. Interspeech*, 2022.

[25] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI*, 2018.

[26] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Proc. ECCV*, 2020.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[28] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018, arXiv:1802.03426.