Domain Adaptation

Nicolas Tirel

GreenAI U.P.P.A. x Prof en Poche

13-02-2023

Glossary

ASR	Automatic Speech Recognition
DA	Domain Adaptation
GAN	Generative Artificial Network
TTS	Text To Speech
VC	Voice Conversion

References I

[Hasija et al., 2021] Hasija, T., Kadyan, V., and Guleria, K. (2021).
Out domain data augmentation on punjabi children speech recognition using tacotron.
Journal of Physics: Conference Series, 1950:012044.

[Joshi and Singh, 2022] Joshi, R. and Singh, A. (2022).
A simple baseline for domain adaptation in end to end ASR systems using synthetic data.
In *Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5)*. Association for Computational Linguistics.

References II

[Kaneko and Kameoka, 2017] Kaneko, T. and Kameoka, H. (2017).
Parallel-data-free voice conversion using cycle-consistent adversarial networks.

[Kaneko et al., 2019] Kaneko, T., Kameoka, H., Tanaka, K., and Hojo, N. (2019).
Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion.

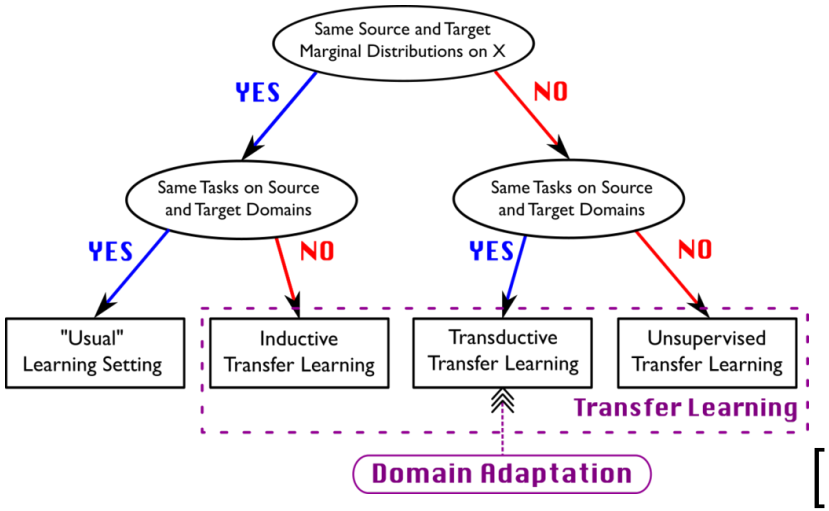
[Redko et al., 2020] Redko, I., Morvant, E., Habrard, A., Sebban, M., and Bennani, Y. (2020).
A survey on domain adaptation theory: learning bounds and theoretical guarantees.

References III

[Shahnawazuddin et al., 2020] Shahnawazuddin, S., Adiga, N., Kumar, K., Poddar, A., and Ahmad, W. (2020).
Voice conversion based data augmentation to improve childrens speech recognition in limited data scenario.

Goal and challenges

Domain Adaptation is a sub-field of transfer learning. We aim at learning from a source data distribution a model on different target data distribution.



[Redko et al., 2020]

Figure 2: Domain Adaptation position

It can be supervised, semi-supervised and unsupervised

Techniques for Domain Adaptation

- Divergence based DA
- Adversarial based DA
- Reconstruction based DA

Comparison with classical classifier

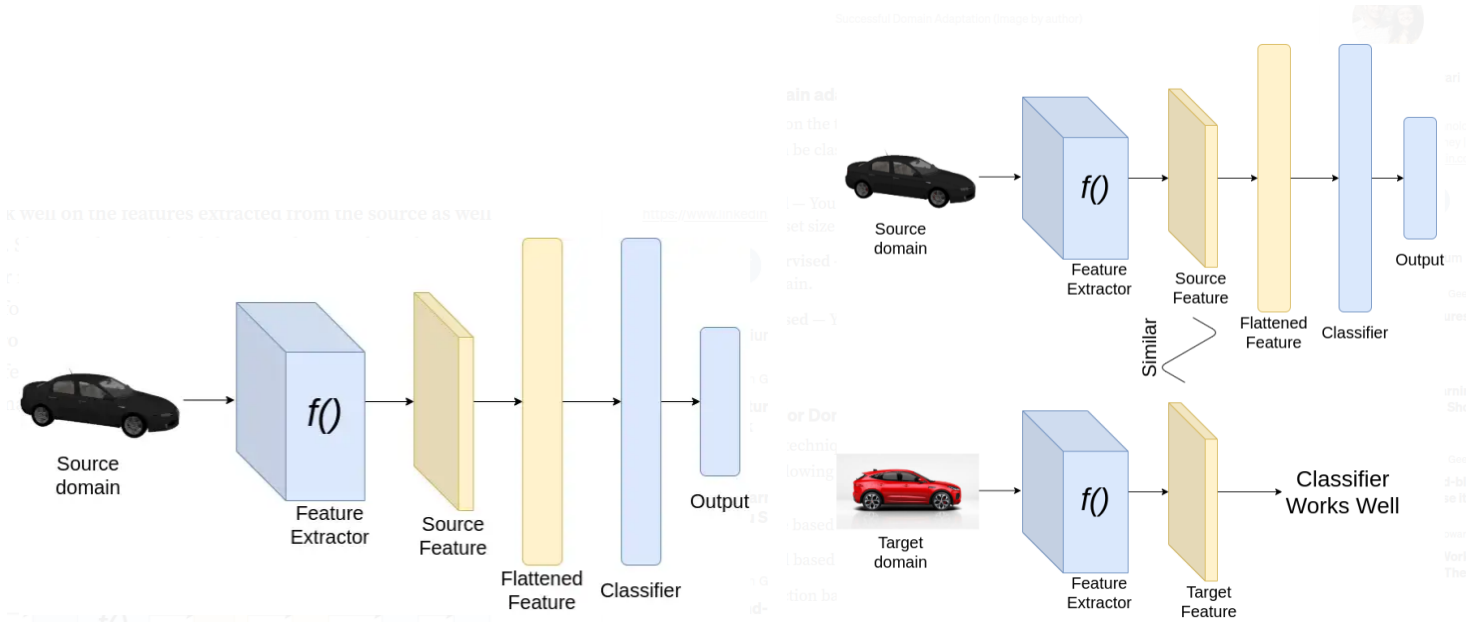


Figure 3: Goal for Domain Adaptation

Divergence based DA

Common divergence based criterion :
Contrastive Domain Discrepancy, Correlation Alignment, Maximum Mean Discrepancy (MMD), Wasserstein etc.

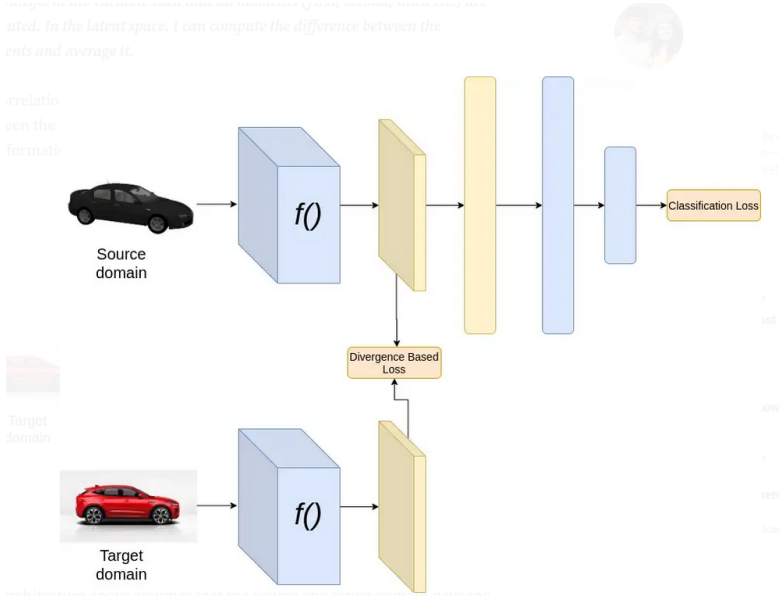


Figure 4: Double loss for divergence : classification and divergence-based loss

Adversarial based DA

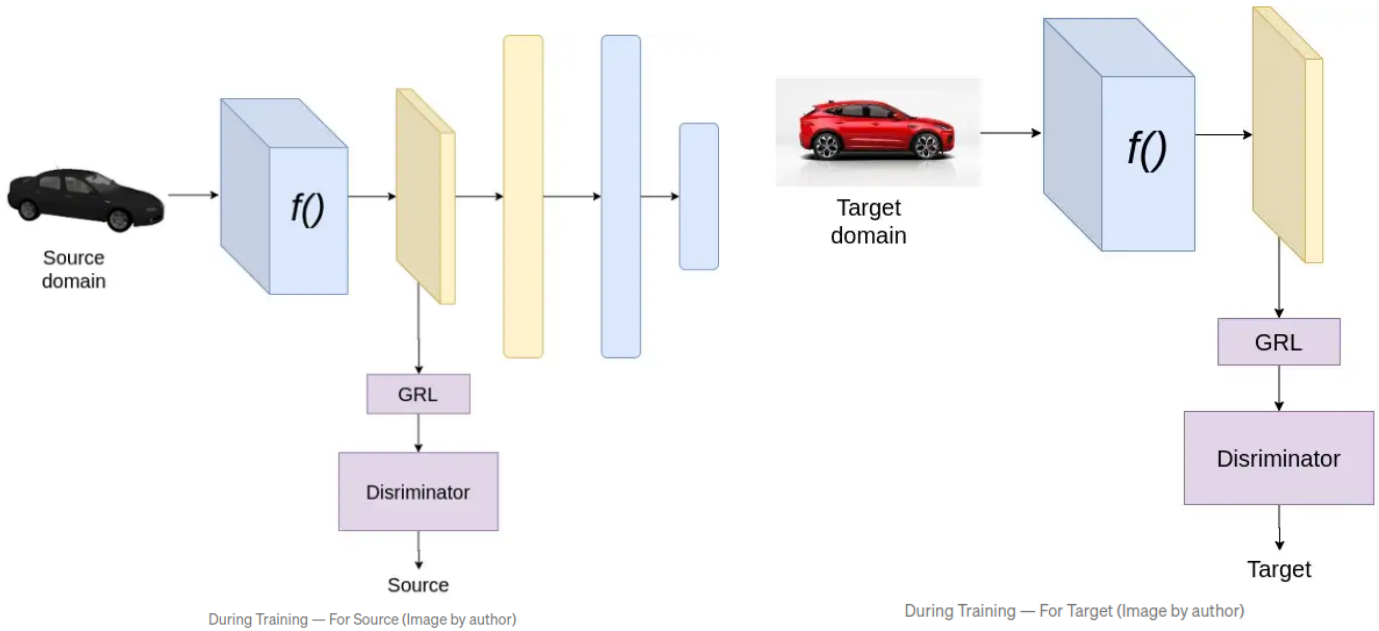


Figure 5: Training for source and target

Reconstruction based DA

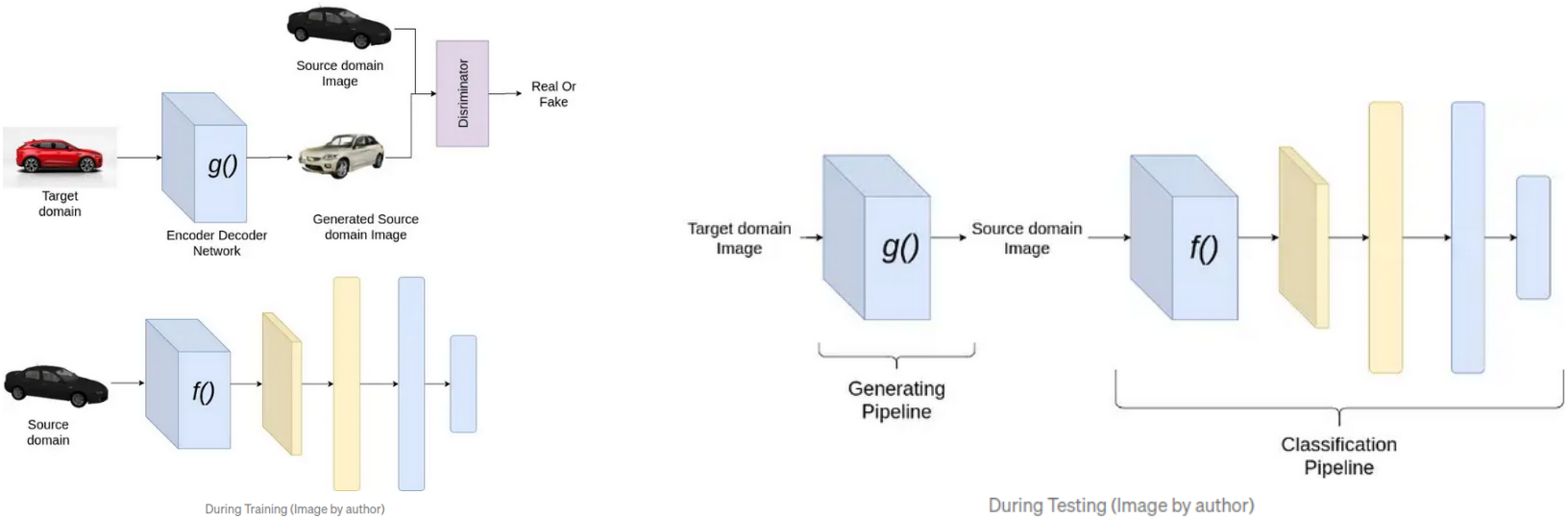


Figure 6: Train and testing

Simple Baseline with Synthetic Data

Improve ASR with only text using TTS [Joshi and Singh, 2022]

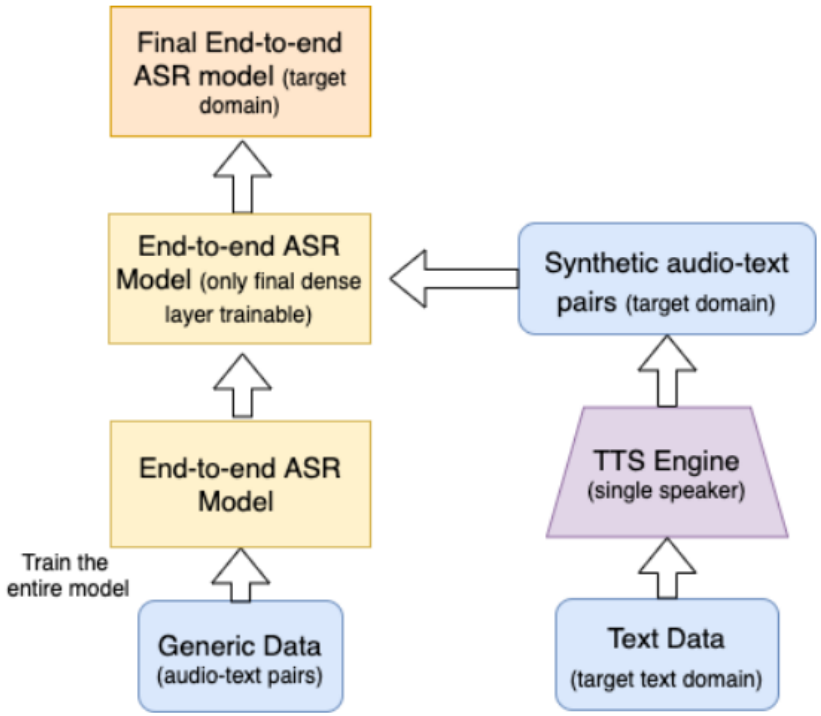


Figure 7: Using synthetic data for dataset

Model	Test WER	Test WER + LM Rescoring	N-Best WER
LAS-Gen	25.31	22.18	13.71
LAS-Dense	16.25	15.55	7.6
LAS-Decoder	13.65	**13.36**	5.82
CTC-Gen	31.84	25.58	13.83
CTC-Dense	20.32	17.66	8.24

Table 1: Word Error Rate(WER) for different model variations using Voice Search Domain. The N-Best WER indicates the best WER in the top N=10 beams.

Model	Test WER	Test WER + LM Rescoring	N-Best WER
LAS-Gen	39.42	31.62	25.35
LAS-Dense	22.57	16.38	11.01
LAS-Decoder	18.96	**12.54**	8.17
CTC-Gen	31.08	22.81	19.74
CTC-Dense	22.43	15.42	12.15

Table 2: Word Error Rate(WER) for different model variations using Address Domain. The N-Best WER indicates the best WER in the top N=10 beams.

Figure 8: Improved results

Punjabi children recognition

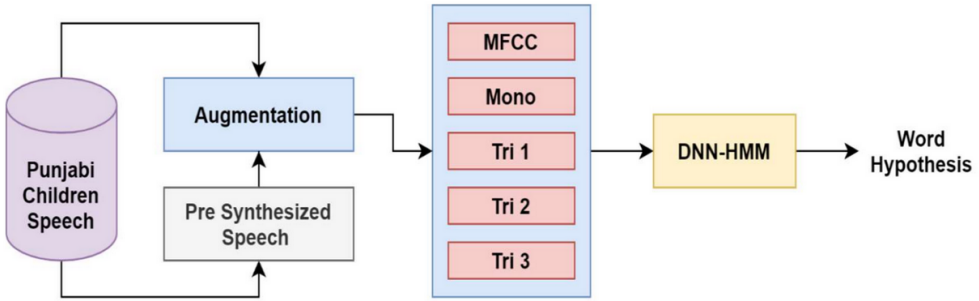


Figure 1. Block Diagram of the ASR system implemented by Augmentation of Original Children Speech and Pre synthesized Speech

Figure 9: Synthetic data using Tacotron synthetis[Hasija et al., 2021]

Voice Conversion

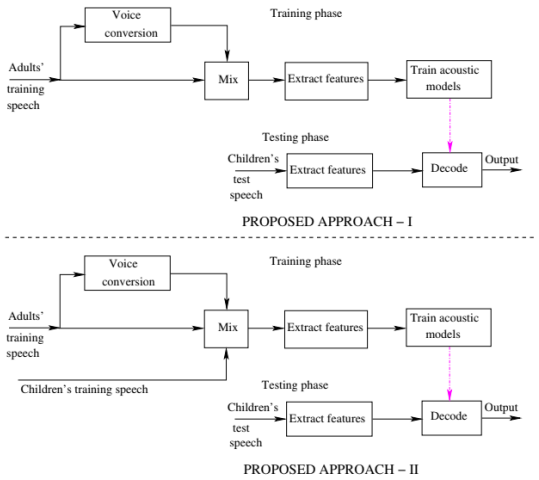


Figure 1: *Proposed schemes for improving children's ASR exploiting voice-conversion-based out-of-domain data augmentation.*

Further, strided convolutional neural networks (CNN) are used

Figure 10: Voice conversion (VC) is a technique for transforming the non/para-linguistic information of given speech while preserving the linguistic information[Shahnawazuddin et al., 2020]

Goal with Prof en Poche

- voix d'enfants -> voix d'adultes pour inférence sur modèle coqui adulte
- voix d'adultes -> voix d'enfants pour augmentation de données pour l'entraînement du modèle enfant

Figure 11: Two scenarios to recognize children voice in mathia application

CycleGan-VC

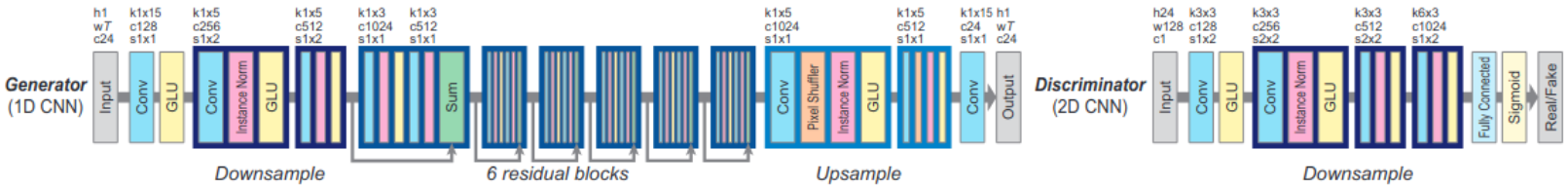


Fig. 2. Network architectures of generator and discriminator. In input or output layer, h, w, and c represent height, width, and number of channels, respectively. In each convolutional layer, k, c, and s denote kernel size, number of channels, and stride size, respectively. Since generator is fully convolutional [42], it can take input of arbitrary length T.

Figure 12: A CycleGAN learns forward and inverse mappings simultaneously using adversarial and cycle-consistency losses.[Kaneko and Kameoka, 2017]

CycleGan-VC2

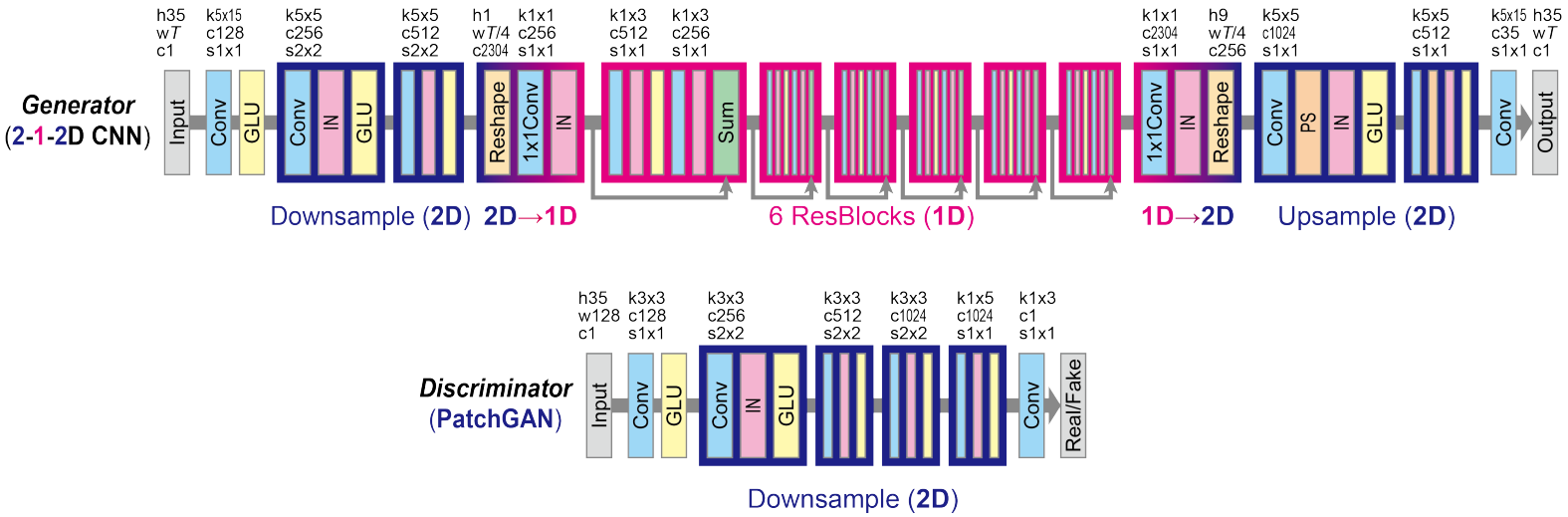


Figure 13: New architecture[Kaneko et al., 2019]

Thanks!