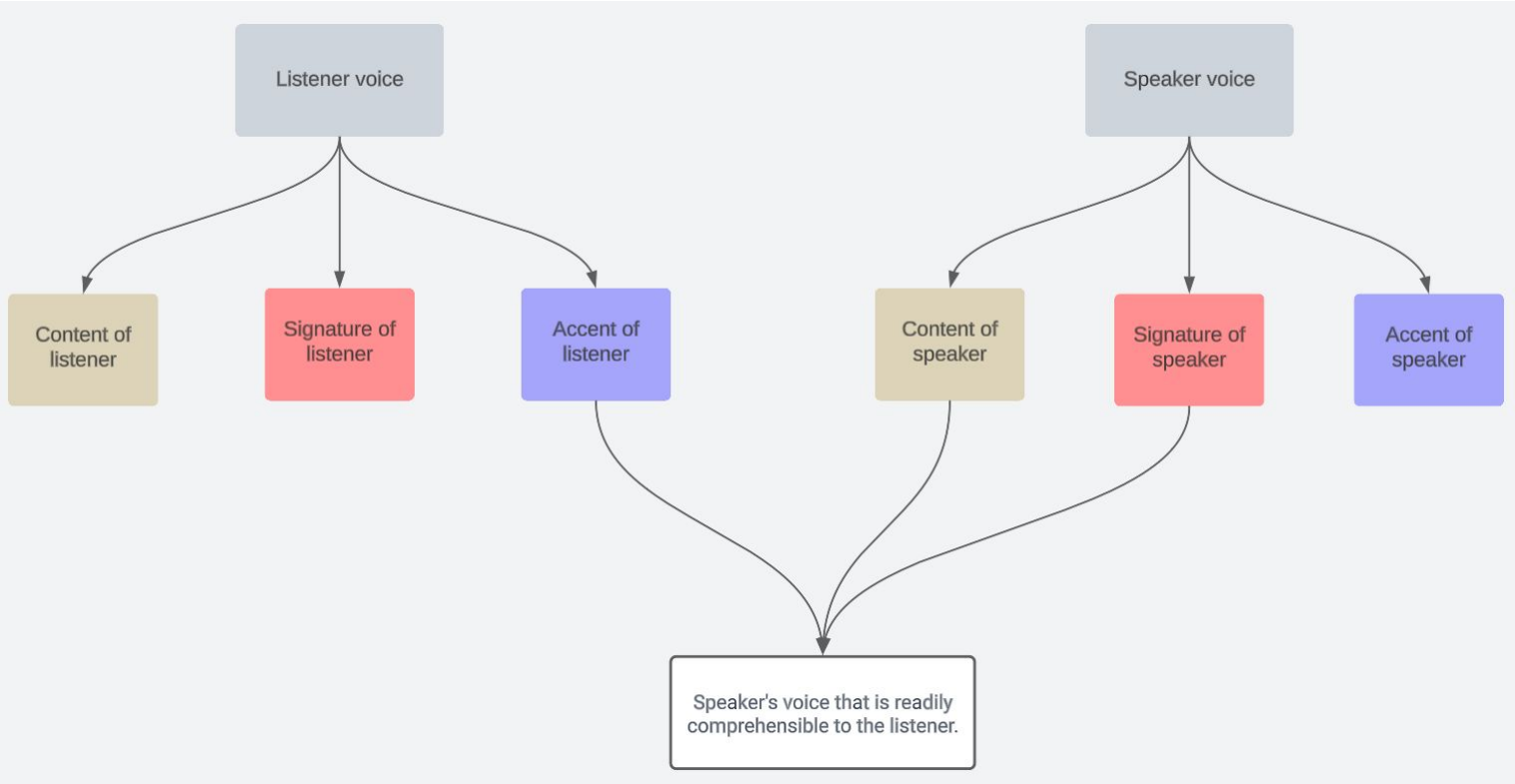
Accent Transfer Techniques for Speech Synthesis

SignalMinded

Overview

- Problem and Motivation
- Existing Work
- Key Contributions
- Aligning Work with Course
- Proposed Model
- Transfer Learning: Feature Extraction and Fine Tuning
- Preprocessing for Audio Synchronization
- Hyperparameter Tuning
- Dataset and Augmentations
- Results

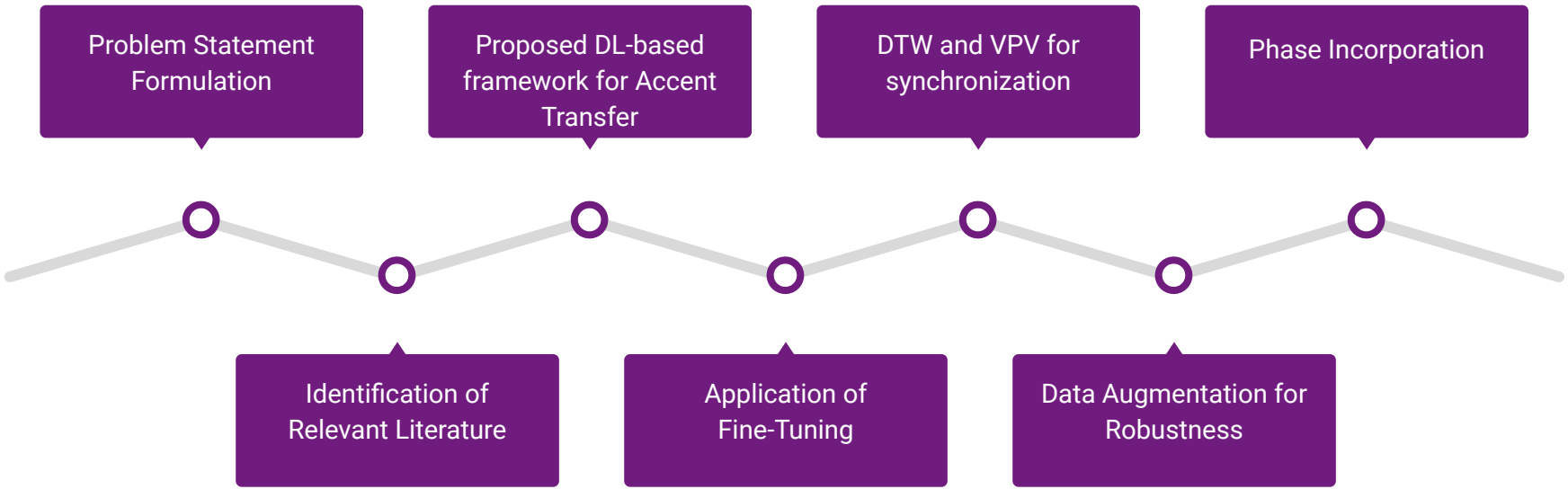
Problem and Motivation



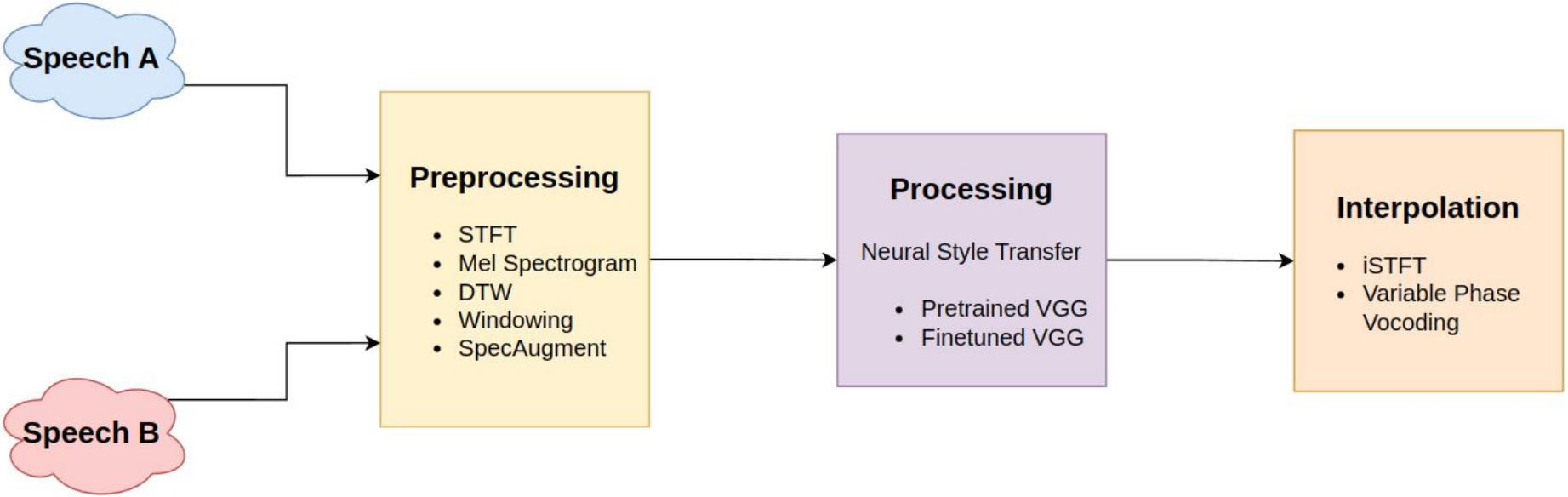
Existing Work

- Methods like region-Based artistic rendering, stroke-based artistic rendering have poor style transfer, lack in generalization etc.*[1]*
- NST can be performed via Image-Iteration Based Style Transfer, Model-Iteration Based Style Transfer
- *Gatys et al.[2]* first introduced NST algorithm using VGG19 where he proposed that the content and style can be represented by internal layers
- *Chen et al.[3]* produced cartoonised real images using GAN where generator input are mixture of content image and style images and the discriminator guesses whether it’s a real or cartoon image

Key Contributions



Aligning Work with Course

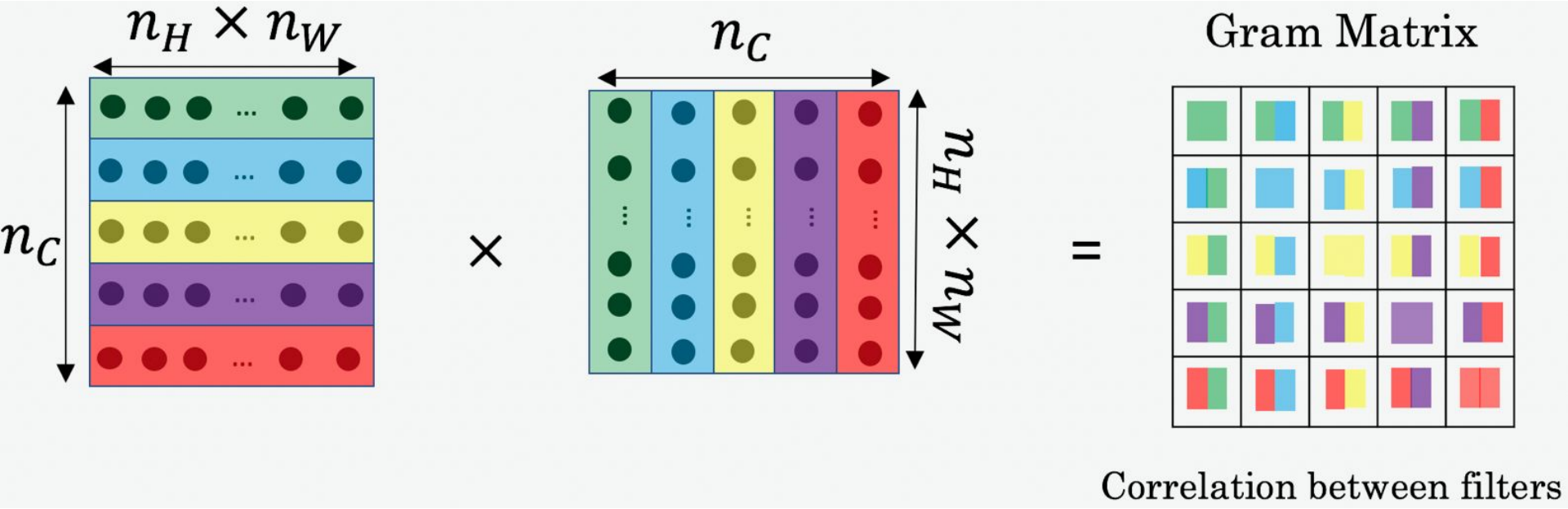


Proposed Model

- Input Audios -> Spectrograms -> Image
- Neural Style Transfer
- Inverting the spectrogram

Neural Style Transfer: An Overview

- Optimizing the target image using content and style loss

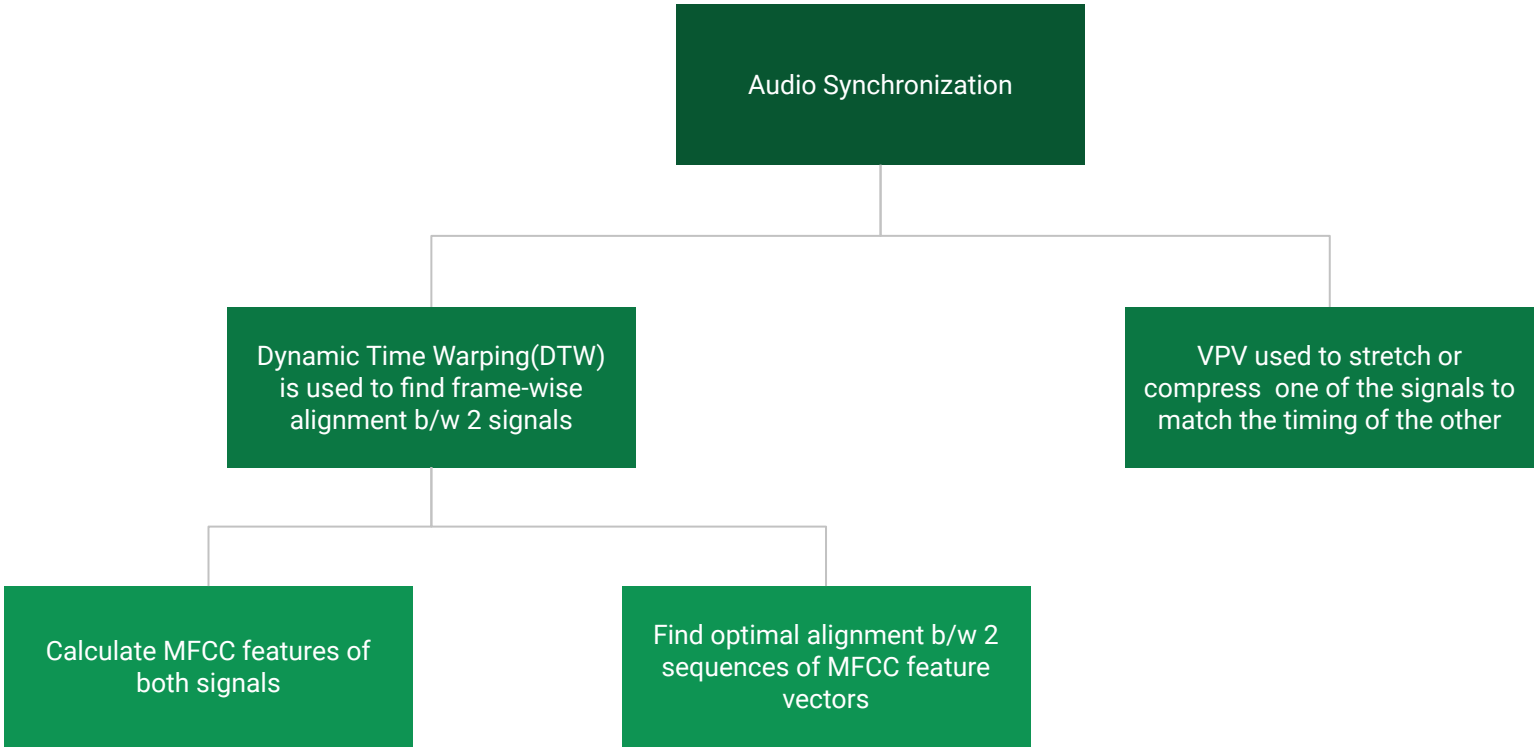


Transfer Learning using Fine Tuning

- Use pre-trained neural network models to solve similar problems with new datasets
- Pre-trained models are trained on large datasets like ImageNet
- Leverage knowledge learned by the pre-trained model
- Reduce computation cost and training time on a new dataset.

- Freeze the entire model and add a new classification layer
- Train for few epochs to make the last layer sensible
- Next, all the layers are unfrozen and trained together with new layer to fine-tune the entire model
- Allows model to learn task-specific features while retaining the previously learned features
- Fine-tuning is particularly useful when working with limited resources or small datasets

Preprocessing for Audio file Synchronization



Hyperparameter Tuning

- 4 combinations of the content-style weight, almost same
- Content layer: Last Conv layer worked the best, any other layer was almost same, but a bit more noisy
- Style Layer
  - For first layer -> Accent wasn't getting transferred even a bit
  - For last layer -> Low quality audio
  - Equal weighting on 5 layers -> best result
- Incorporation of phase didn't lead to improved results

Dataset and Data Augmentation

Dataset	Description	Intuition
Speech Accent Dataset with clipping	200 accents consisting of audio clips of 5 seconds	Long clips but less in number so clipping was done
Arctic Dataset	18 accents consisting of audio clips	—
Mel Spectrogram Arctic Dataset	18 accents consisting Mel Spectrograms of audio clips	Mel Scale is consistent with the way humans perceive sound
Mel Spectrogram with SpecAugment Arctic Dataset	18 accents consisting Mel Spectrograms of audio clips transformed using SpecAugment	SpecAugment transforms spectrograms using randomized time warping, frequency masking, time masking to create variations in the data
Varying Window Sizes Arctic Dataset	18 accents consisting Spectrograms of audio clips with 4 window sizes	Varying window sizes allows to overcome time-frequency tradeoff in spectrogram

Simulations

- Transfer learning on all datasets mentioned earlier
- Neural style transfer experiment setup

Content Weight	Style Weight
1e-2	1
1e-2	1e2
1e-2	1e4
1e-2	1e6

- Final simulation was run by varying the above configuration (4):
  - Each parameter set from transfer learning and also original parameter set (5 + 1 = 6)
  - With / without DTW (2)
- This gives us 4 * 6 * 2 = 48 different simulations

Results

Classifier	Correctly Classified(Yes/No) (with DTW)	Correctly Classified(Yes/No) (without DTW)
Pre-trained	Yes	Yes
Fine-tuned with MelDataset	No	Yes
Fine-tuned with SpecAugment MelDataset	No	No

Results

FinetunedMelWithDTW

American Content

Indian Style

FinetunedMelWithoutDTW

Indian Warp

PretrainedWithDTW

PretrainedWithoutDTW

Conclusions

- The output audios obtained from fine-tuned models seemed to be a bit smoother compared to the pretrained model.
- The outputs of the fine-tuned model using Mel Spectogram performed to be the best (qualitatively better than the pre-trained model)
- The outputs of the other fine-tuned models were not classified correctly.
- In all the cases the output of the models was noisy.
- The outputs generated with DTW as preprocessing step were found to be noisier.

References

[1] Qiang Cai, Mengxu Ma, Chen Wang, Haisheng Li, “Image neural style transfer: A review”

[2] L. A. Gatys, A. S. Ecker and M. Bethge, "Image Style Transfer Using Convolutional Neural Networks," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2414-2423, doi: 10.1109/CVPR.2016.265.

[3] Y. Chen, Y. -K. Lai and Y. -J. Liu, "CartoonGAN: Generative Adversarial Networks for Photo Cartoonization," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 9465-9474, doi: 10.1109/CVPR.2018.00986.