An Investigation of End-to-End Models for Robust Speech Recognition

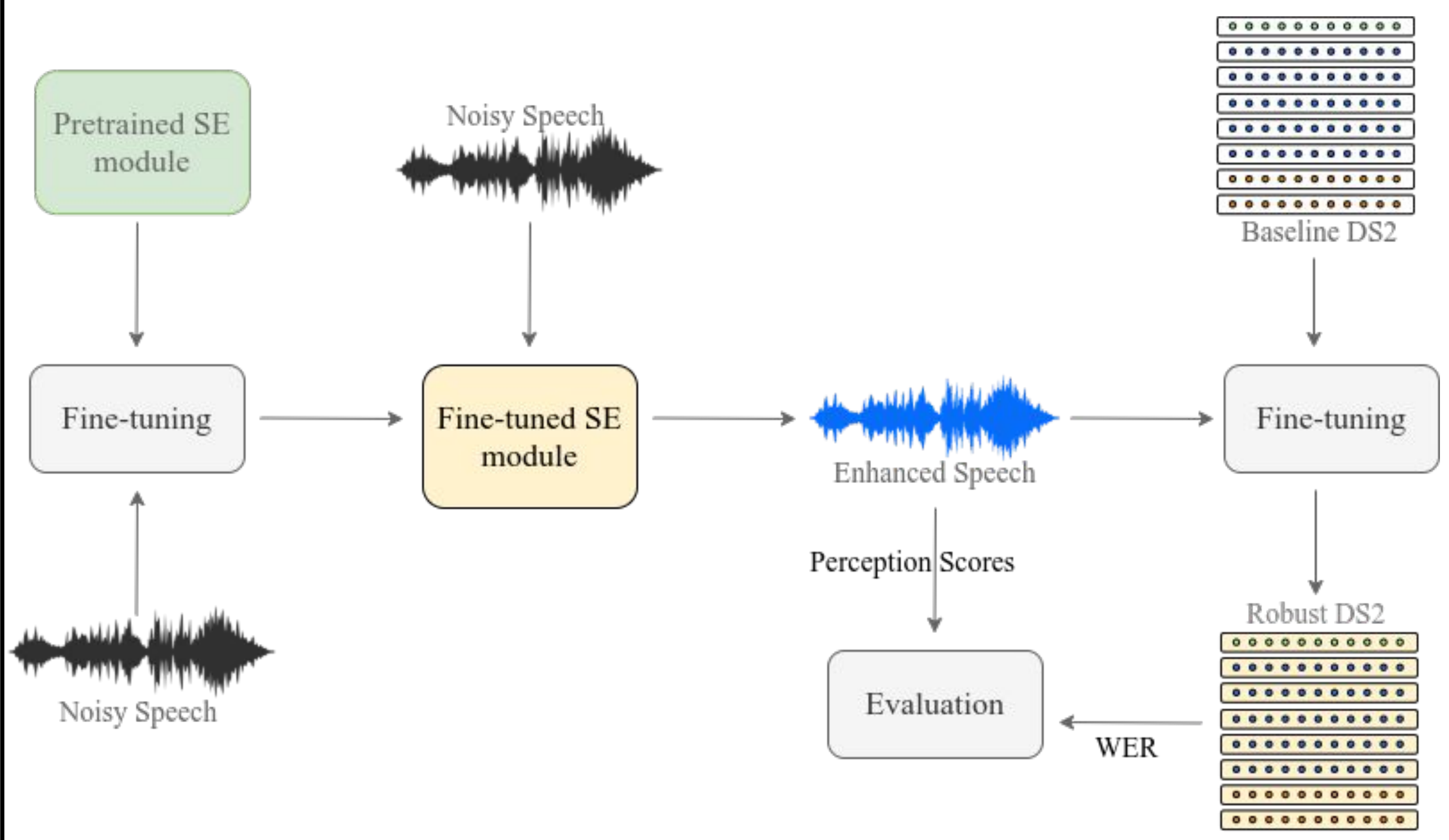
Paper ID: 5280

Archiki Prasad*, Preethi Jyothi*, and Rajbabu Velmurugan*
*Indian Institute of Technology Bombay, India | ✉ archikiprasad@gmail.com

Introduction

- Two approaches for robust adaptation of end-to-end (E2E) ASR systems:
  (i) Front-end Speech Enhancement followed by back-end E2E ASR
  (ii) End-to-End ML-based adaptation for E2E ASR
- **Objective:** Compare these approaches when limited noise samples are available
- **Setup — E2E ASR:** Deep Speech 2[1] pre-trained on clean speech (WER: 10.3)
- **Datasets:** Clean Speech: LibriSpeech dataset (100 hours)
  Noise: Custom dataset with 2 hours in train and test set
  Noise types: 'Babble', 'Airport/Station', 'Car', 'Metro', 'Cafe', 'Traffic', 'AC/Vacuum'

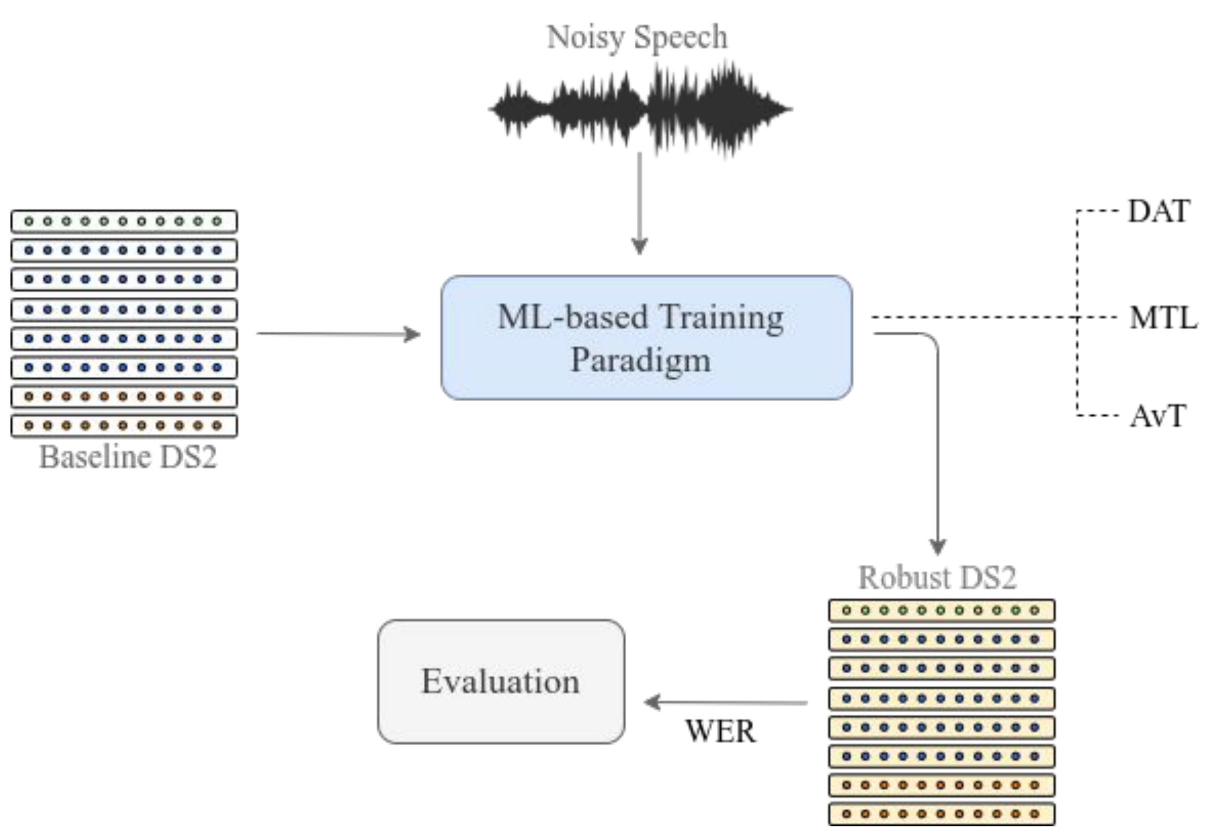
Speech Enhancement (SE) systems w/ back-end ASR



Three different SE models:
- **SE-VCAE**[2]
- **DeepXi**[3] **(DeepMMSE)**
- **DEMUCS**[4]

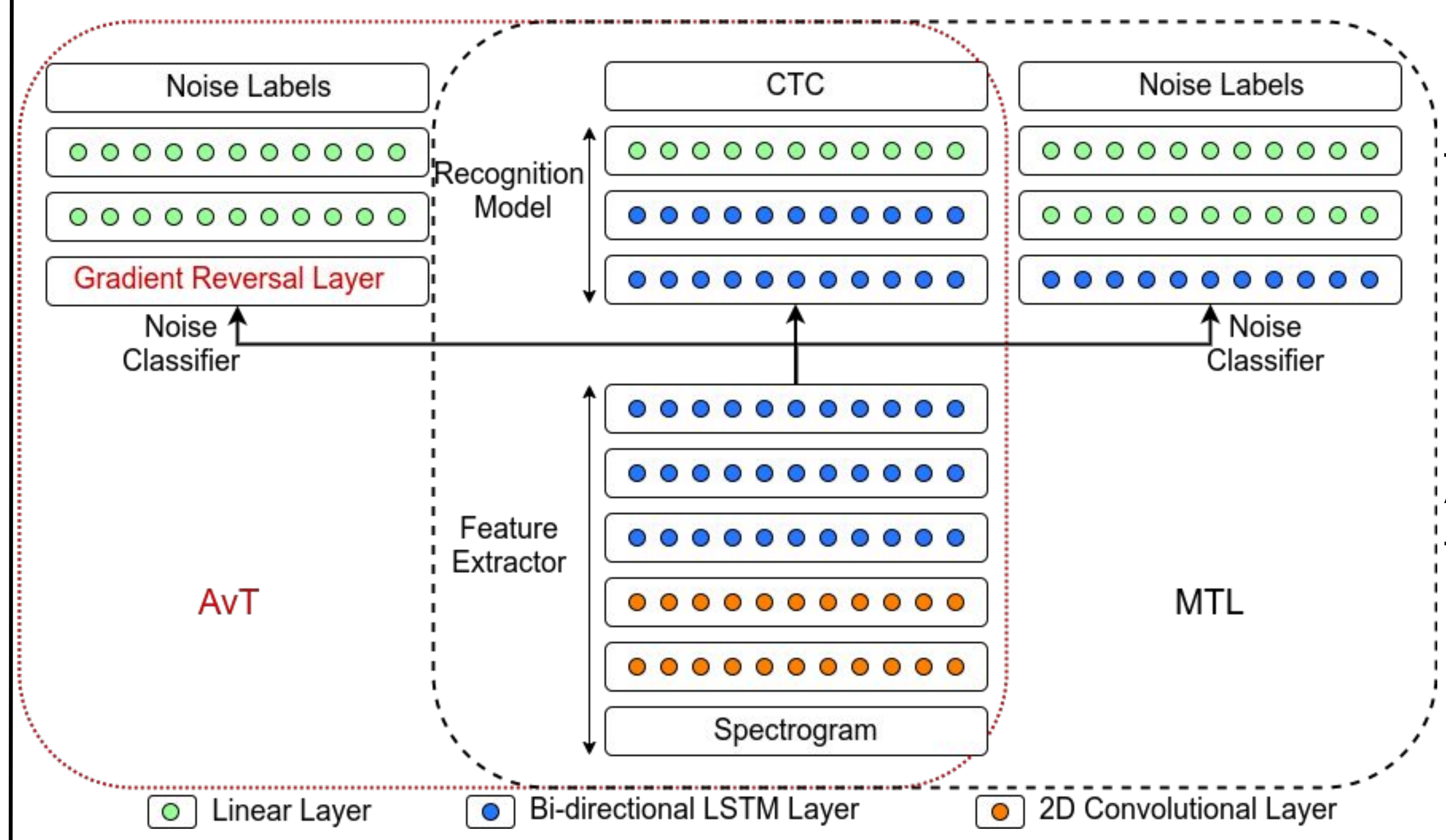
Issues:
1. Scarce noise data requires pretrained SE models
2. Both fine-tuning steps demand extra compute
3. Perception scores do not correlate with WER

ML-based adaptation of E2E ASR



- **Data Augmentation-based Training (DAT):**
  - **Vanilla DAT:** Same learning rate across all layers of Deep Speech 2
  - **Soft-Freeze DAT:** Differential learning rates (LR): low LR at top, high at bottom
- **Multi-Task Learning (MTL):** Disentangle noise information in the representations
- **Adversarial Training (AvT):** Make the representations invariant to noise

Implementation of MTL and AvT



Noise classifier: 8 labels (7 noise types + label for clean speech)

MTL: Trained on a hybrid loss

$$L_H = \lambda L_{CTC} + \eta(1-\lambda) L_{CE}$$

AvT: Training objective of the feature extractor is:

$$min\ L_{CTC} + max\ L_{CE}$$

(via gradient reversal layer)

Performance Comparison

Method	WER under SNR (in dB)															
	Babble					Airport/Station					Metro					Clean
	0	5	10	15	20	0	5	10	15	20	0	5	10	15	20	
Baseline	104.2	98.3	91.3	79.7	65.0	91.9	84.1	73.7	60.6	50.0	68.4	54.4	46.4	34.9	27.6	10.3
SE-VCAE	85.6	76.4	61.9	54.7	39.7	78.0	68.3	56.8	46.3	39.3	54.0	43.6	38.6	33.0	29.6	15.9
Deep Xi	81.4	69.4	54.0	44.5	31.9	71.4	60.9	46.5	37.8	27.4	44.8	30.5	28.1	20.2	20.5	10.9
DEMUCS	70.3	58.0	41.8	32.3	25.4	58.6	45.5	33.7	25.6	21.5	**35.6**	**24.9**	**22.6**	**17.1**	**15.9**	10.9
Vanilla DAT	80.6	68.1	53.6	41.8	30.3	67.1	55.4	41.9	31.2	24.9	41.8	33.1	27.1	21.9	19.1	10.8
Soft-Freeze DAT	77.4	65.5	52.2	38.5	28.3	64.2	52.9	39.0	29.2	23.7	40.8	30.7	27.0	21.3	18.6	10.9
MTL	71.4	58.8	45.9	35.5	25.8	55.7	46.8	35.3	26.2	**20.7**	38.7	29.2	24.4	20.6	17.3	11.0
AvT	**66.8**	**55.1**	**39.5**	**31.1**	**24.6**	**53.8**	**43.3**	**33.4**	**25.2**	20.9	36.1	26.5	22.6	18.4	17.8	13.1

- Noise type and level of stationarity determines the degree of degradation
- DEMUCS performs the best across SNRs for Metro, followed by MLT and AvT
- AvT performs the best across SNRs for degrading noises like Babble and Airport/Station
- Our approaches (MTL and AvT) perform better than all SE methods other than DEMUCS

Takeaways

- Among speech enhancement, DEMUCS outperforms others on all measures
- AvT is largely the best ML-based technique; however, noise invariance in representations causes degradation in clean speech and high SNR performance
- The best technique for robust adaptation depends on the type of underlying noise

References:
1. Amodei et al. Deep speech 2: End-to-end speech recognition in English and Mandarin. In ICML 2016
2. Braithwaite et al. Speech enhancement with variance constrained autoencoders. In InterSpeech 2019
3. Nicolson et al. Deep xi as a front-end for robust automatic speech recognition. (Arxiv)
4. Defossez et al. Real time speech enhancement in the waveform domain In InterSpeech 2020