Multi-Conditioning & Data Augmentation using Generative Noise Model for Speech Emotion Recognition in Noisy Conditions

Upasana Tiwari, Meet Soni, Rupayan Chakraborty, Ashish Panda, Sunil Kumar Kopparapu

TCS Research and Innovation – Mumbai, INDIA

Paper Code: 5701

ICASSP 2020, Barcelona, Spain

May 04-08, 2020

Introduction

The discipline of automatically recognizing human emotion and affective states from speech, usually referred to as **Speech Emotion Recognition (SER)**.

SER in **clean** scenario has been studied widely over the last two decades [1].

Performance **degrades** when these models are **trained with clean speech** and **tested in realistic environment (mostly unseen noises).**

[1] Bjorn W Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks and ongoing trends", Communications of the ACM, vol. 61, no. 5, pp. 90–99, 2018

SER in Noisy Environment:

Signal Level

Front-end signal processing:

- Voice activity detector (VAD)
- Non-negative matrix factorization (NMF) [2]
- Blind source separation (BSS), etc.

Feature Level

- Feature compensation
- Denoising enhancement, etc. [3]

Model Level

- Model adaptation
- Multi-conditioning, etc.

[2] M. Pandharipande, R.Chakraborty, A. Panda, and S. K. Kopparapu, “An unsupervised frame selection technique for robust emotion recognition in noisy speech”, IEEE EUSIPCO, 2018, pp. 2055–2059.

[3] R. Chakraborty, A. Panda, M. Pandharipande, S. Joshi, and S. K. Kopparapu, “Front-end feature compensation and denoising for noise robust speech emotion recognition”, Proc. Interspeech, pp. 3257–3261, 2019.

Motivation

- **Enhancement techniques** --> seen noises ⬆, unseen noises ⬇
- Most of the **previous works** in SER dealt with noise at signal/feature level only.
- **Multi-conditioning and augmentation**
  - Other speech processing technologies (e.g. ASR): ✓
  - Noise robust SER tasks : ✗
- **Generative Adversarial Network (GAN) based data augmentation** [4,5]:
  - SER tasks : ✓
  - Noise robustness aspect : ✗

We have addressed these gaps in our work.

[4] D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, S. Narayanan, A. Chatziagapi, G. Paraskevopoulos, “Data augmentation using GANs for Speech Emotion Recognition”, Proc. Interspeech 2019, pp. 171–175, 2019.
[5] Lu Yi and Man-Wai Mak, “Adversarial data augmentation network for Speech Emotion Recognition,” Proc. APSIPA, 2019

Proposed Approach

Proposed Approach

- **Noise robust SER** [multi-conditioned + augmented data]
  - **Parametric Generative noise model** that can simulate multiple unseen noise conditions [6]
  - **Wide variety of generated noise** allows the data augmentation that facilitates deep learning system for the SER task.

Hypothesis

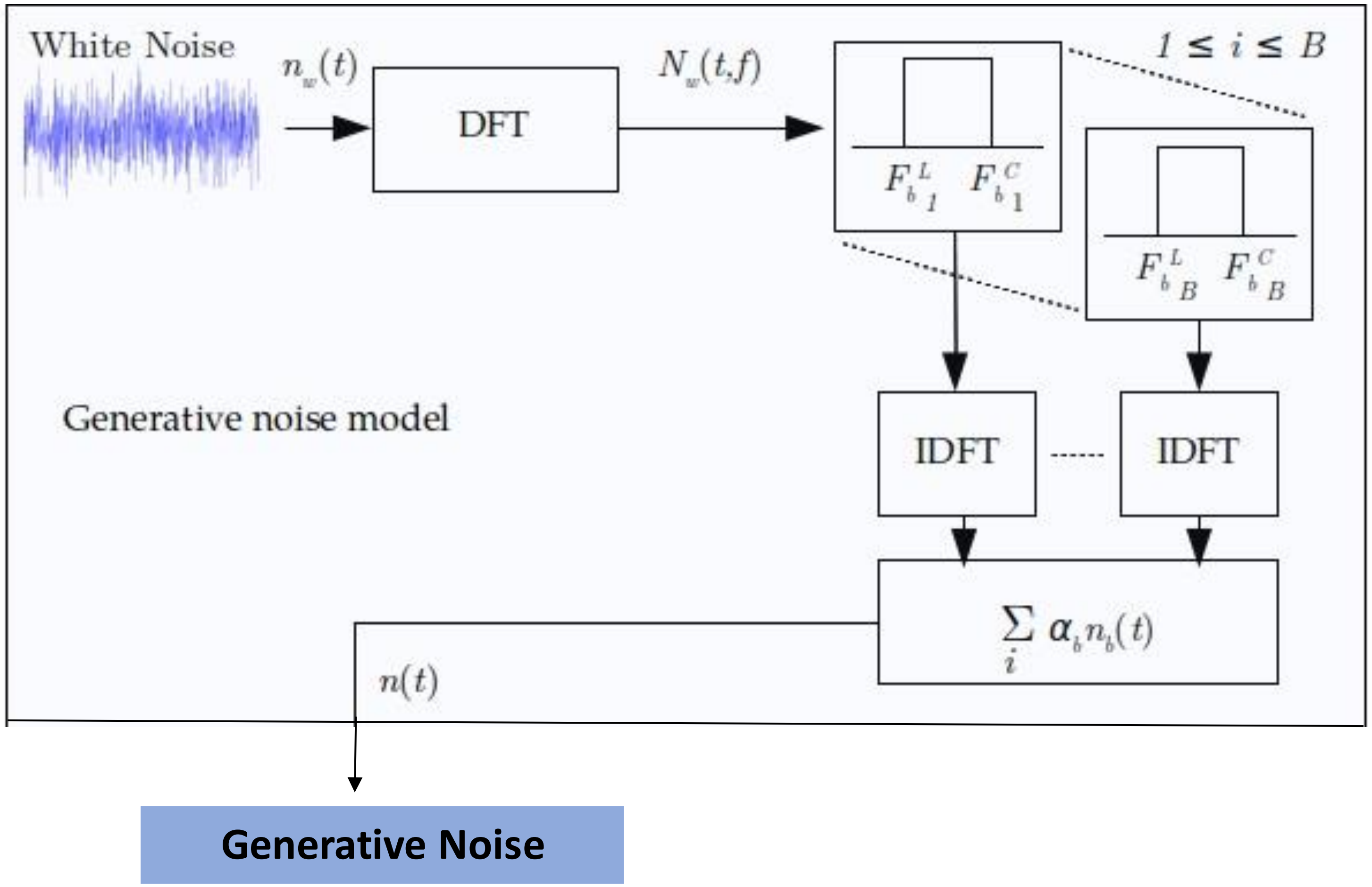
- **Generated noises cover a diverse (and bigger) noise space [difficult to get with the recorded noises].**
- **Expected to generate better emotion models in the realistic applications.**

[6] M. Soni, S. Joshi and A. Panda, "Generative noise modeling and channel simulation for robust speech recognition in unseen conditions", Proc. Interspeech, pp. 441–445, 2019.

Multi-Conditioning for Robust Speech Emotion Recognition

Generative Noise Model [6]

- $n_w(t)$ = white noise signal
- $N_w(t,f)$ = STFT magnitude of $n_w(t)$
- F^L_b and F^C_b are lower and central frequency of b^{th} filter of Mel-Filter Bank
- B = 24, total number of bands
- $n_b(t)$ = band-limited signal
- α = weights of noise bases ($B+1$)



Note : **$n_b(t)$ in time domain are linearly combined with different values of α from [0.1,1] with steps of 0.1.**

[6] M. Soni et al., "Generative noise modeling and channel simulation for robust speech recognition in unseen conditions", Proc. Interspeech, 2019.

Generating Noise : Noise Bases

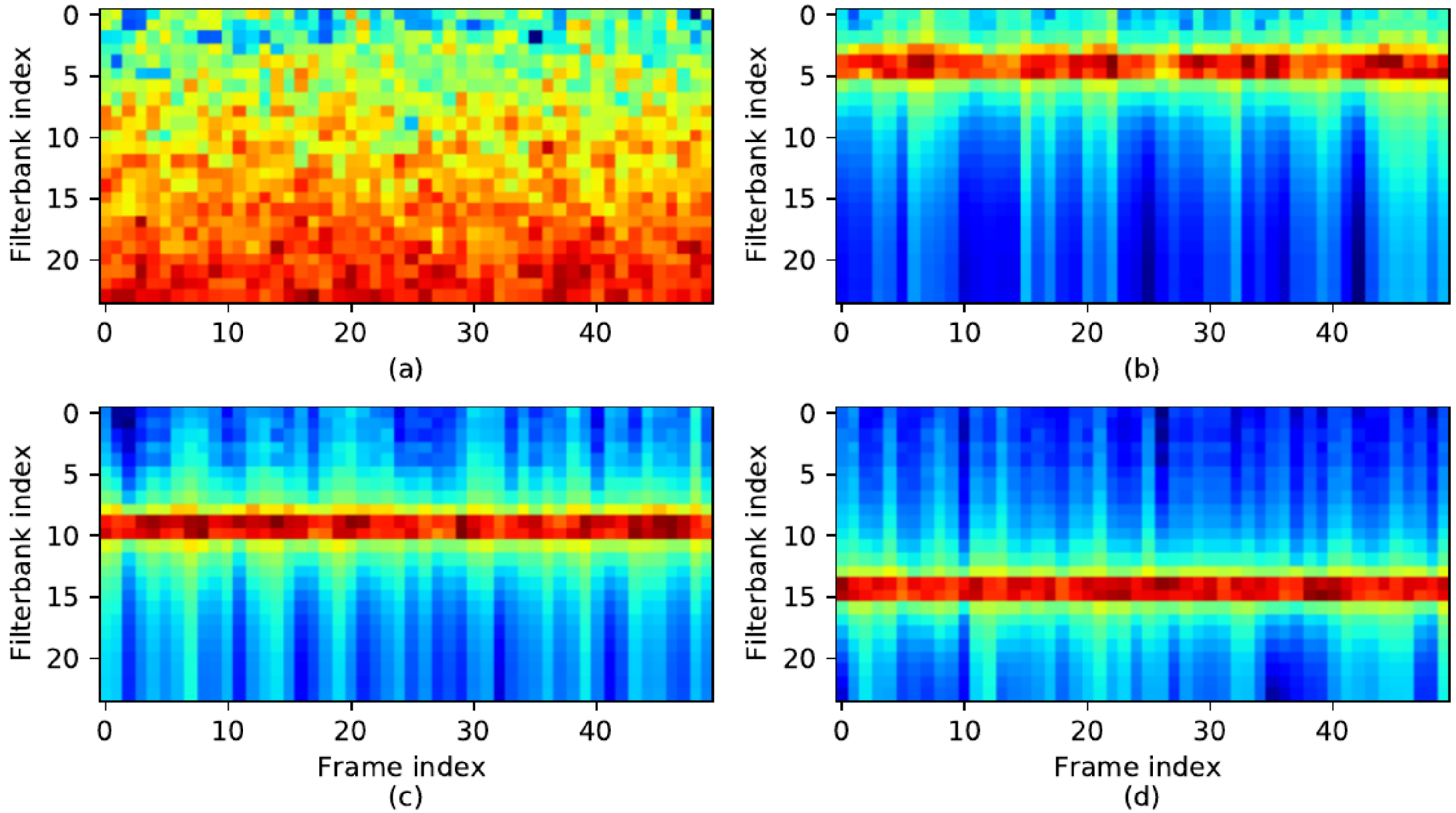


Figure : log-MFBEs of (a) white noise signal, (b) – (d) filtered white noise using 5th, 10th and 15th filter in Mel-filterbank

Generating Noise : Noise Signal

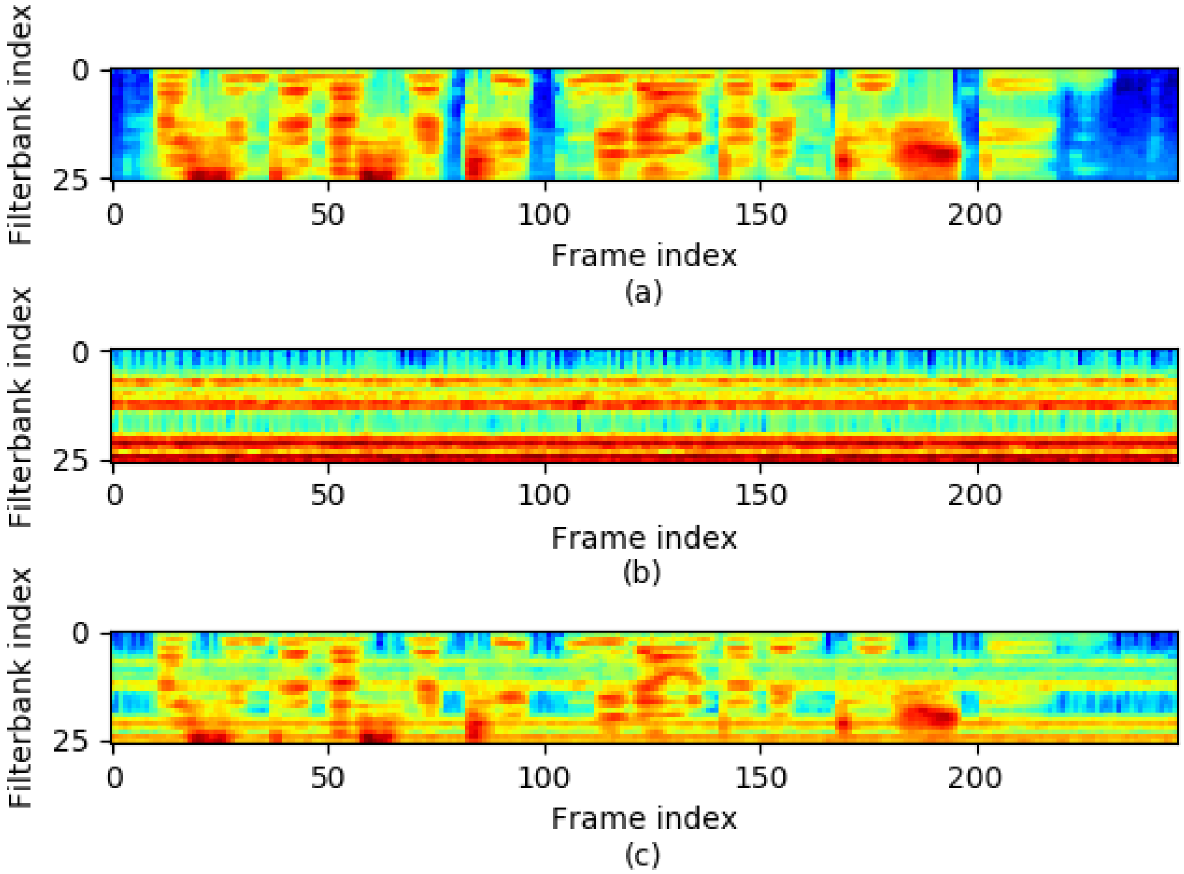
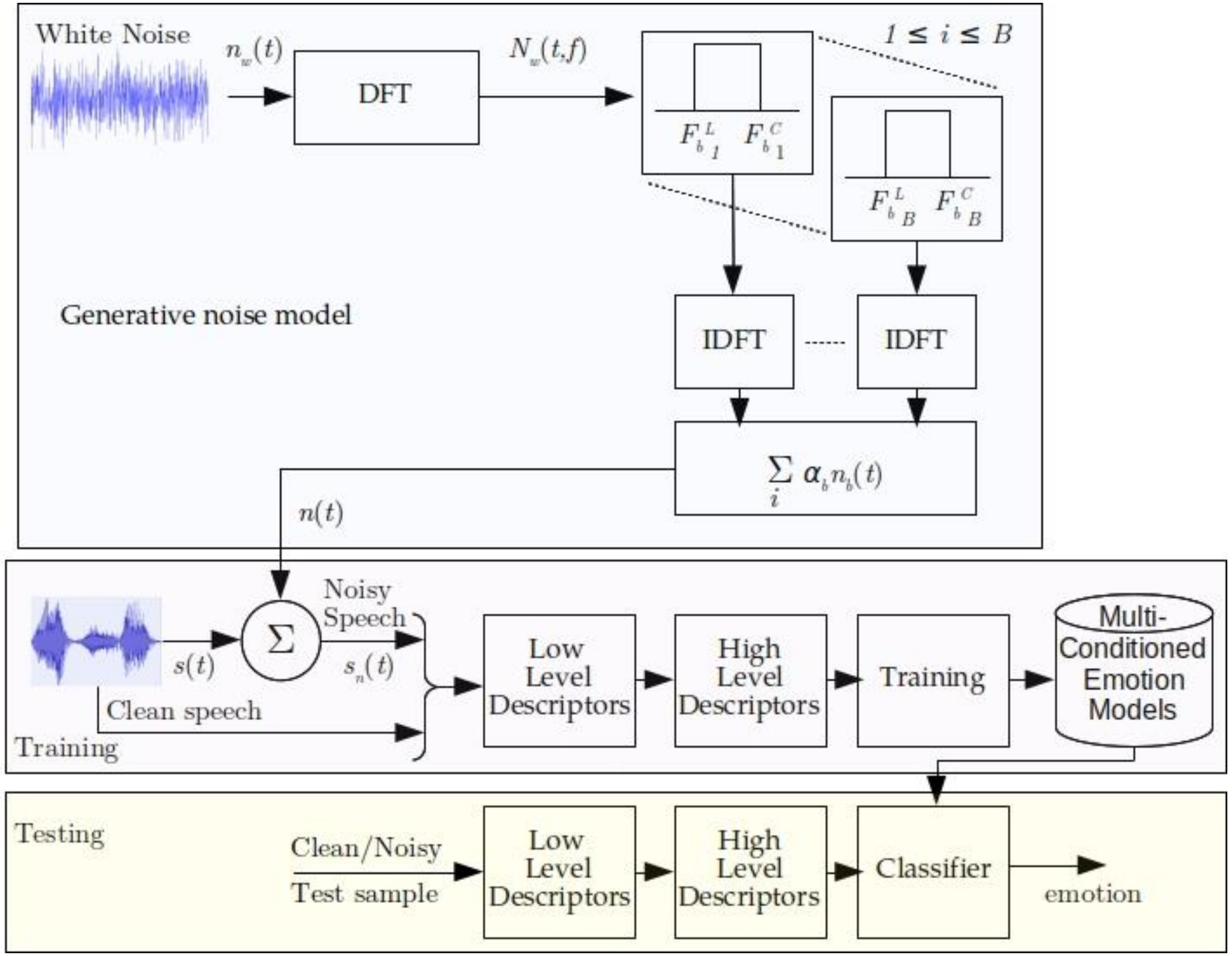


Figure : log-MFBEs of (a) clean signal from EmoDB, (b) noise signal generated using parametric Generative model, (c) noisy signal after adding noise with 15 dB SNR

Multi conditioning data augmentation using Generative noise model



Multi-conditioning using Noises

Generative Noises

Noisex-92 [7]

SetC (Training)	SetA (seen) - Training	SetB (unseen) - Testing
5 types (generated using different values of α from [0.1,1] with steps of 0.1)	1: Voice Babble	1: Bucaneer
	2: Factory Noise	2: Machine Gun
	3: HF radio channel	3: Destroyer_ops
	4: F-16 fighter jets	4: Destroyer_engine
	5: Volvo 340	

Both setA and setB noises are sufficiently distinct in characteristics

Note : α = weights of noise bases of Generative noise model

[7] “Noisex-92”, http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html.

Data Augmentation

Database

Berlin Emotional Database (Emo-DB) [8]-

- 535 acted utterances recorded in fairly clean environment.
- Eliciting 7 emotion categories.

Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [9]-

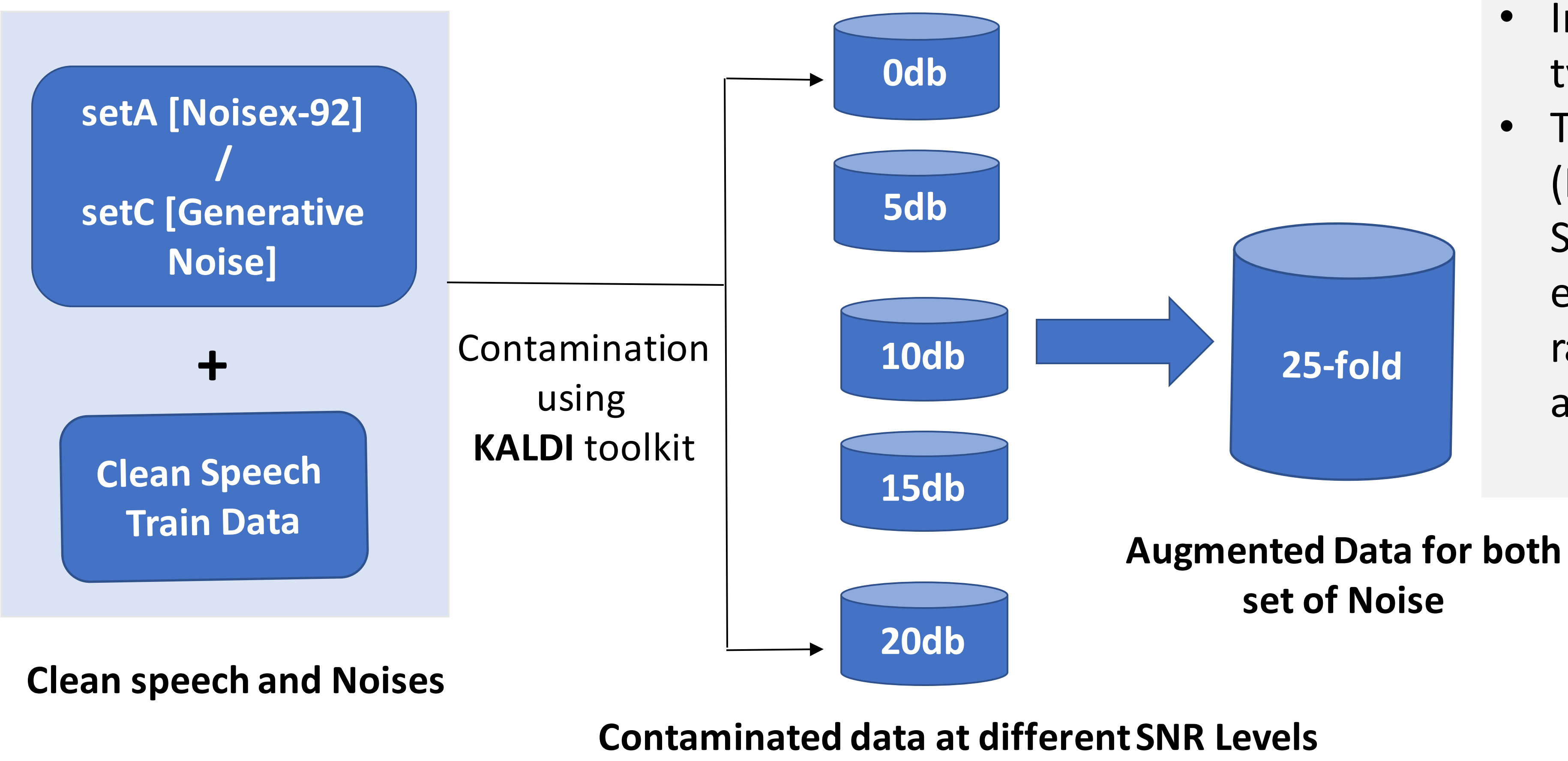
1. **Scripted Recording** - participants have rehearsed the memorized script
2. **Improvised Recording**-participants have improvised some hypothetical situations

We experimented on Scripted + Improvised samples from 4 emotion categories (i.e. Happy, Anger, Neutral, Sad)

[8] F. Burkhardt, A. Paeschke, M.A. Rolfes, W.F. Sendlmeier and B.Weiss, "A Database of German Emotional Speech", in Proc.Intersp eech, 2005.
[9] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S.Lee, and S.S. Narayanan, "IEMOCAP: Interactiveemotional dyadic motion capture database", Language resources and evaluation, vol. 42, no. 4, pp. 335, 2008.

Data Augmentation

- 80% - > Train Data, 20% -> Test Data



To prevent Overfitting:

- Initially both set has 5 types of Noises
- To get optimal combination (Number of noise types, SNR levels) , we experimented with possible random choices of Noises and SNR

Data Augmentation

- **Random selection of number of noise types and SNR levels**
  - **EmoDB**
    - fold1, fold4, fold8, fold16 and fold25 -> 1, 4, 8, 16 and 25 times of train data
  - **IEMOCAP**
    - fold1, fold2, fold3 and fold4 -> 1, 2, 3 and 4 times of train data
- **Reason :**

$\text{IEMOCAP}_{\text{number of samples}} > \text{EmoDB}_{\text{number of samples}}$

Experimental Setup

I. Acoustic Features

- 6552-dimensional feature vector using “emo-large" configuration file of openSMILE toolkit [8].

II. Model Training

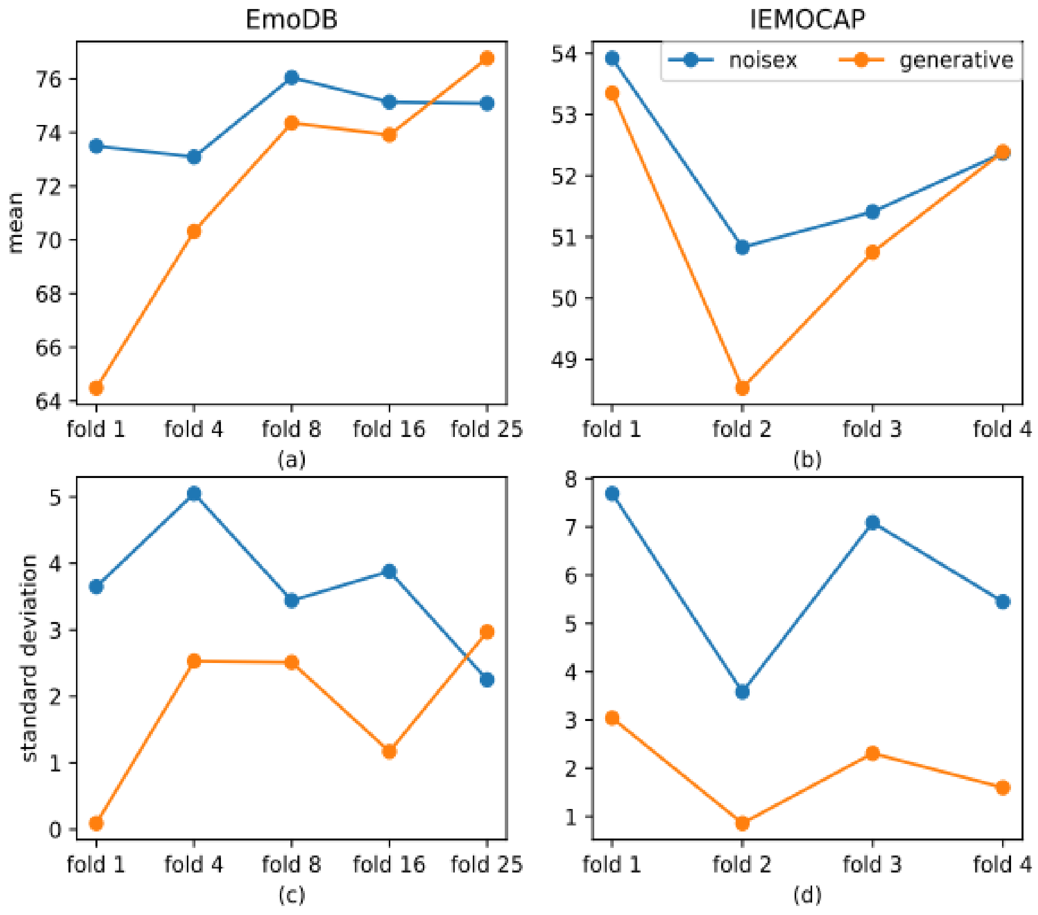
- **Model_1** : clean baseline for both databases by training the model only on clean samples.
- **Model_2** : trained with **Clean + [Augmented speech with Noisex-92]**
- **Model_3** : trained with **Clean + [Augmented speech with Generative model]**

We trained Deep Neural Network (DNN) with sigmoid activation in hidden layers and softmax activation in output layer.

[8] "openSMILE, audio feature extraction tool by audEERING”, http://www.audeering.com/ research/opensmile

Result and Analysis

Performance in unseen conditions



Observations :

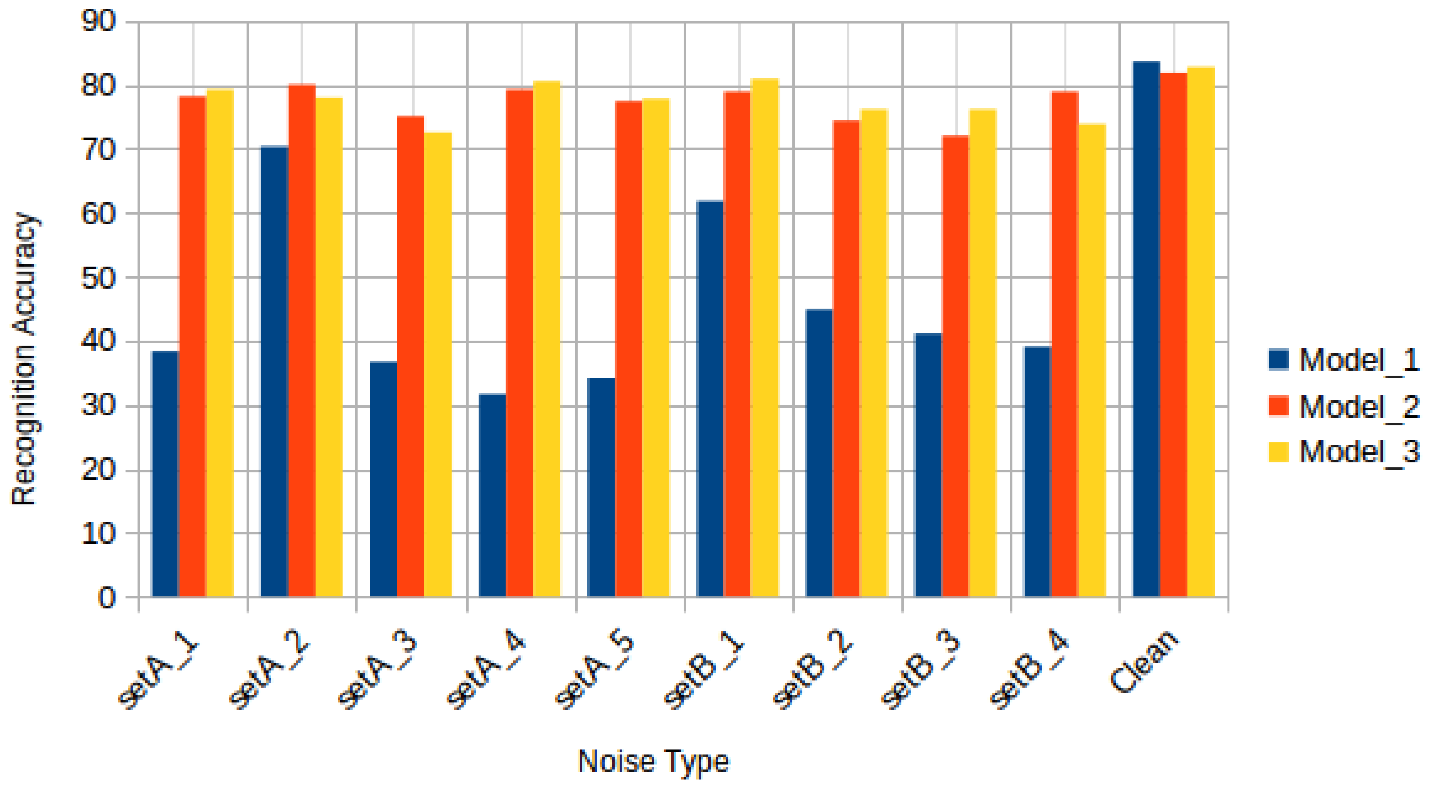
- **NOISEX** noise -> no additional variability after a certain point
- **Generative** noise -> vary with every instance of sampling
- For **EmoDB**, Generative approach outperforms at fold25
- For **IEMOCAP**, the trend suggests that more folds might help

Take Away : Better generalization by Generative noise model

Results (EmoDB)

- **DNN with 3 hidden layers (4K, 2K and 1K neurons), followed by a dropout of 50%**
- **Performance degradation in Model_1 (noisy environment)**
- **Model_2 and Model_3 performs significantly better in noisy test conditions.**

Note : Model_1 -> Clean baseline, **Model_2** -> Clean + [Augmented speech with Noisex-92], **Model_3** -> Clean + [Augmented speech with Generative model]



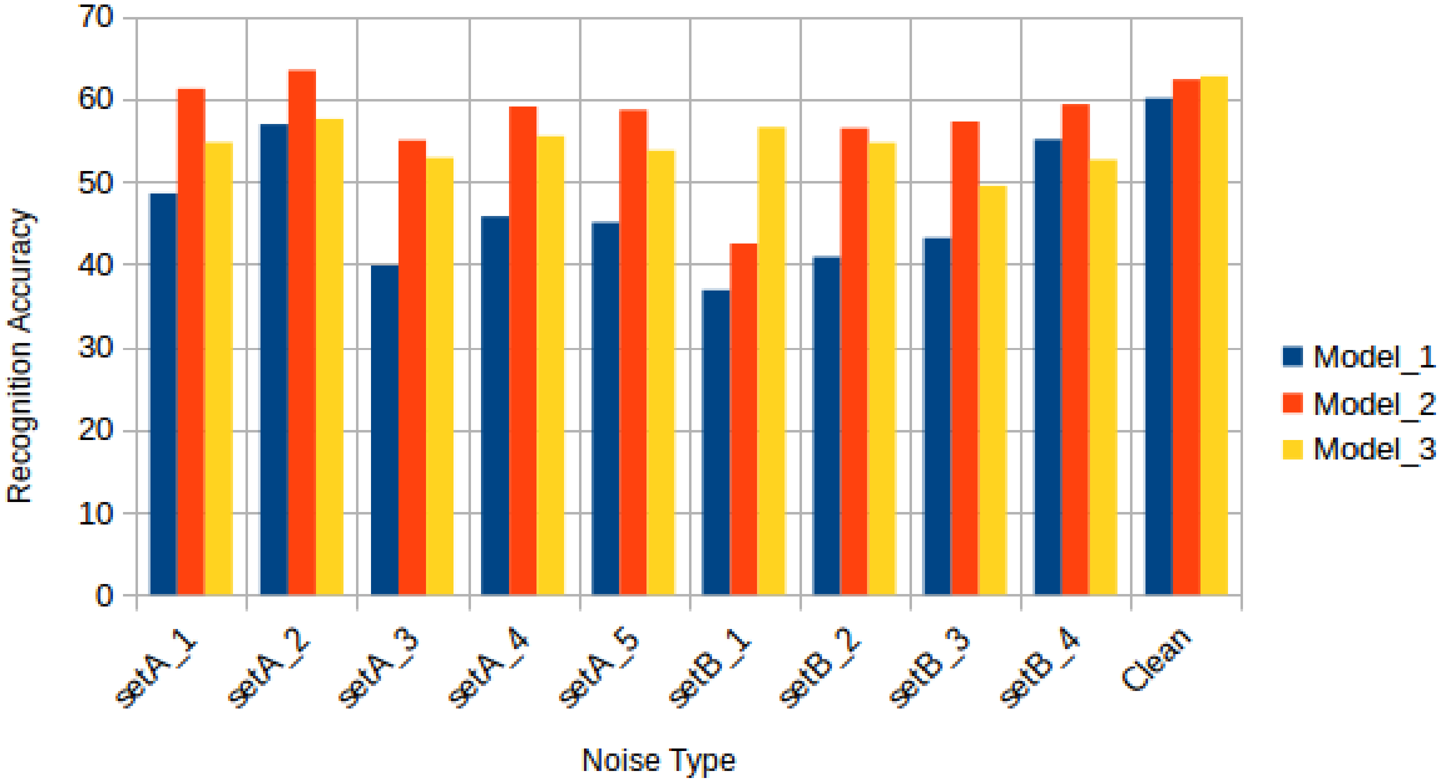
Note :

- **Model_1 and Model_3 : (setA, setB) -> unseen**
- **Model_2 : setA -> seen, set_B -> unseen**

Results (IEMOCAP)

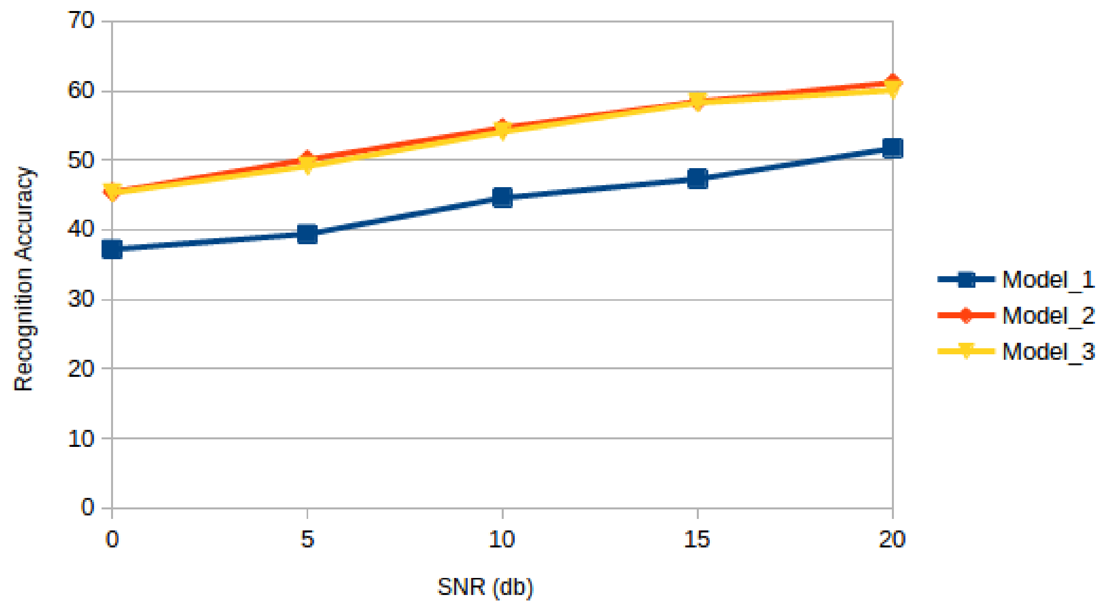
- **DNN with 1 hidden layer (3k neural units)**
- **EmoDB -> Model_3 surpassed the conventional approach**
- **IEMOCAP -> Both are comparable in performance**

Note : Model_1 -> Clean baseline, **Model_2** -> Clean + [Augmented speech with Noisex-92], **Model_3** -> Clean + [Augmented speech with Generative model]

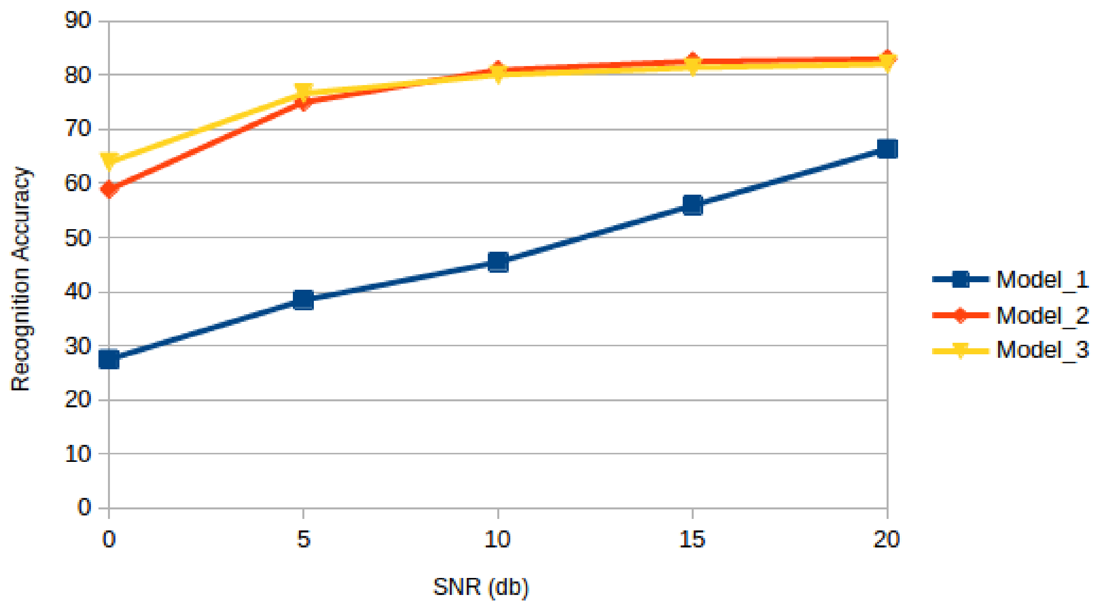


Recognition accuracy at different SNR levels

Note : Model_1 -> Clean baseline, Model_2 -> Clean + [Augmented speech with Noisex-92], Model_3 -> Clean + [Augmented speech with Generative model]



EmoDB



IEMOCAP

Conclusion

- **Noise robust SER**
  - **Multi-conditioning** and **Data augmentation**
  - **Generative noise model**
  - Classification using **deep learning system**

Even with a small database like EmoDB

- Proposed method **imparts robustness** to the SER system in **unseen noise** conditions

- Improved average recognition accuracy for unseen condition
  - **EmoDB - 46.72% to 76.77% (+30.05%)**
  - **IEMOCAP** - **44.01% to 53.35% (+9.34%)**

THANK YOU