Project 1

Single-Channel VAD Technology & NISA+ Development

Lucia Eve Berger

M. Sci in Software Engineering

Managers: DoYeong Kim, Dushyant Sharma

Outline

Objectives

Pipeline

SCT configurations
Feature extractions
Model Selection

Results

Integration

Let’s play with Acoustic Data/Models!

Objectives

Automatic Acoustic Data Augmentation: using NISA+ to generate SCT Configs

- Conduct a Literature Survey on VAD/ASR
- Experiment with CNN architectures
- More flexible data provision and feature extraction

High Level Progress

Started looking at VAD component

- More robust to reverberation and noise
- Exploiting latest CNN technology and feature extraction

Extending NISA+ with external VAD features

- Refactor code-base
- Feeding in VAD posteriors as features

Developing Scripts for generating SCT configs

- Using distribution of SNR, T60, DRR, SNR (NISA+ output)

VAD Pipeline

Pipeline

Preprocessing

Selection & corruption of dataset

Training

Supervised problem

Testing

Comparing against other baselines

Dataset Selection

Used Speech from TIMIT

Libri (clean) for

train, test and evaluation sets

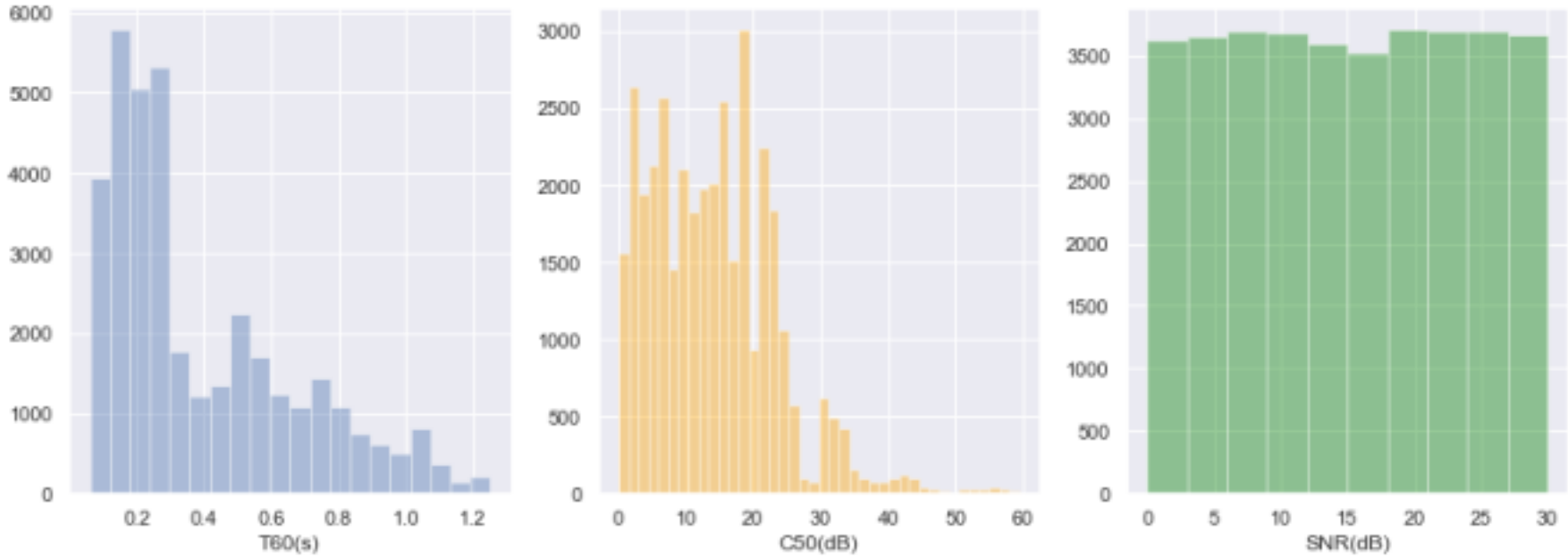
- Data augmentations
  - Corruptions were performed asymmetrically and symmetrically
  - Reverb ambient, babble, music, white and other (domestic, fans)

Music Corruption

- Free Music Archive (FMA)(https://github.com/mdeff/fma)
  - Large genres of music (jazz, hiphop, classic, etc)
  - Code for lookup and querying on desired hours {10 hours for training, unique for testing}
  - On grid for others use (via querying the dataframe)
  - *Libre speech with rap?!*

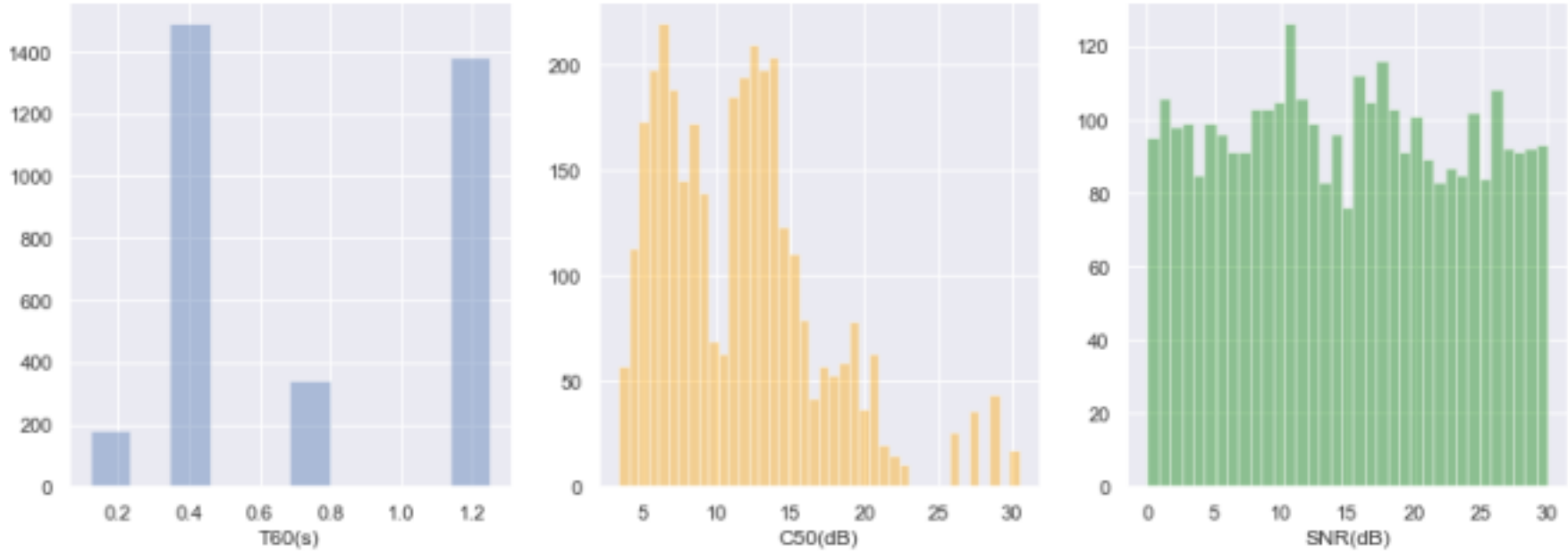
Training Data

- Reverb Files: 22,300 & Noise Sources: 260 unique
- Reverb Ambient: 0.08%, Reverb Babble: 0.36 %, Reverb Music (FMA dataset): 0.62% (10 hours), Reverb White: generated, Reverb Other (fans, domestic): 0.06%



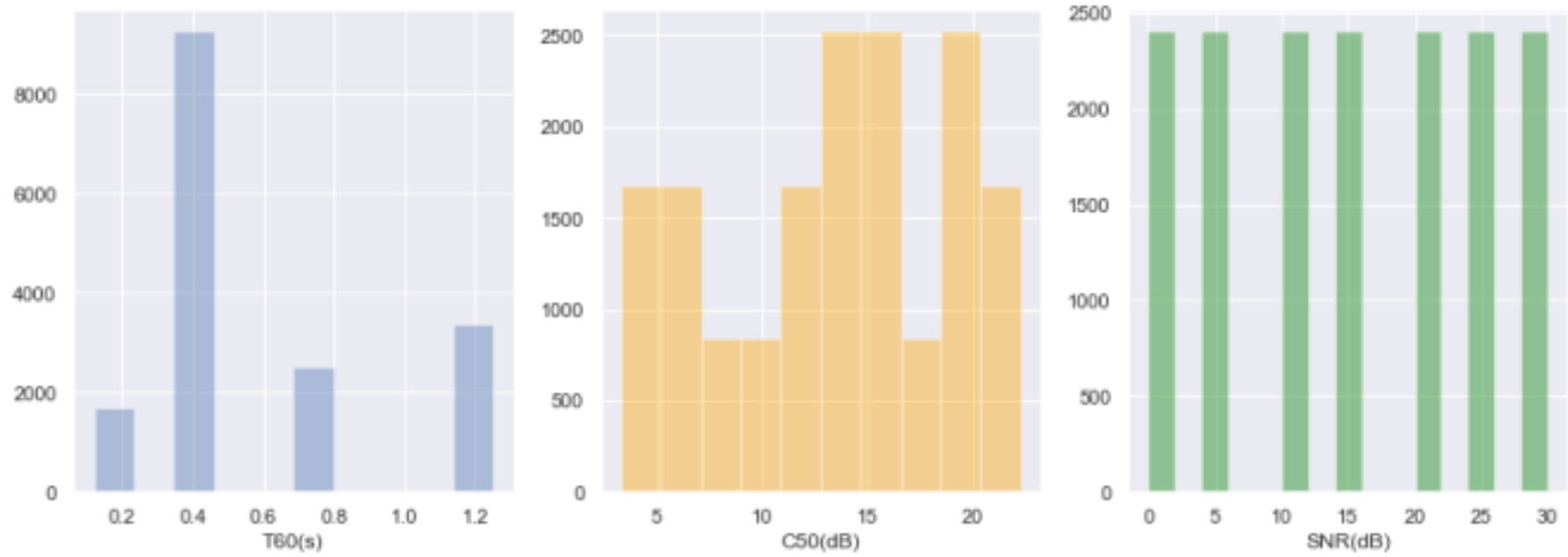
Asymmetric Corrupted Test Set

- Reverb Files: **224,** Noise Sources: 787 unique (not in the train)
- Reverb Ambient: 0.24%, Reverb Babble: 0.25%, Reverb Music (FMA dataset): 0.19% Reverb White: generated, Reverb Other (fans, domestic): 0.28%



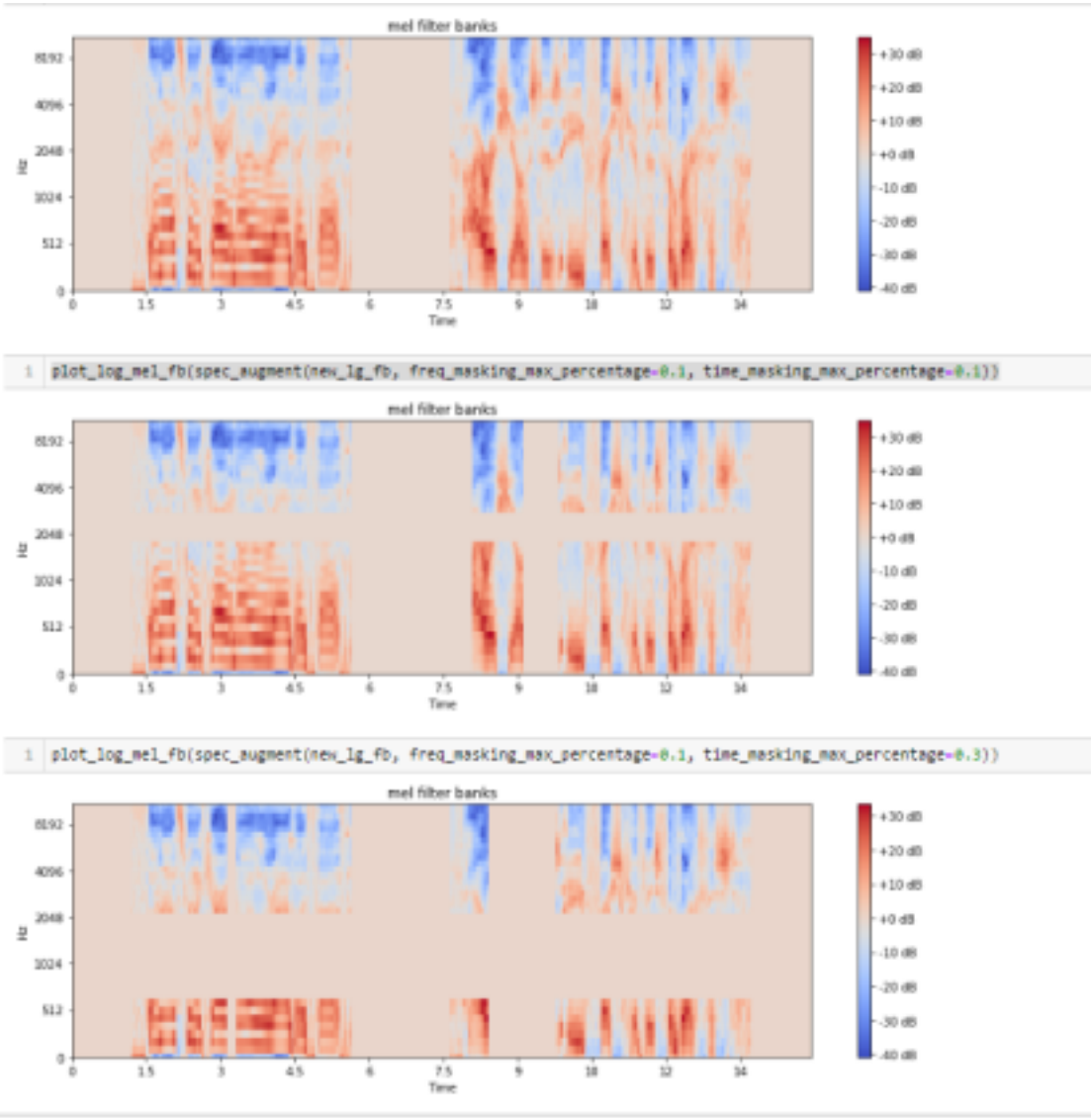
Symmetrically Corrupted Test Set

- Reverb Files: **20,** Noise Sources: 4 unique (not in the train), plus white generated (one per type)
  - English, Japanese, French



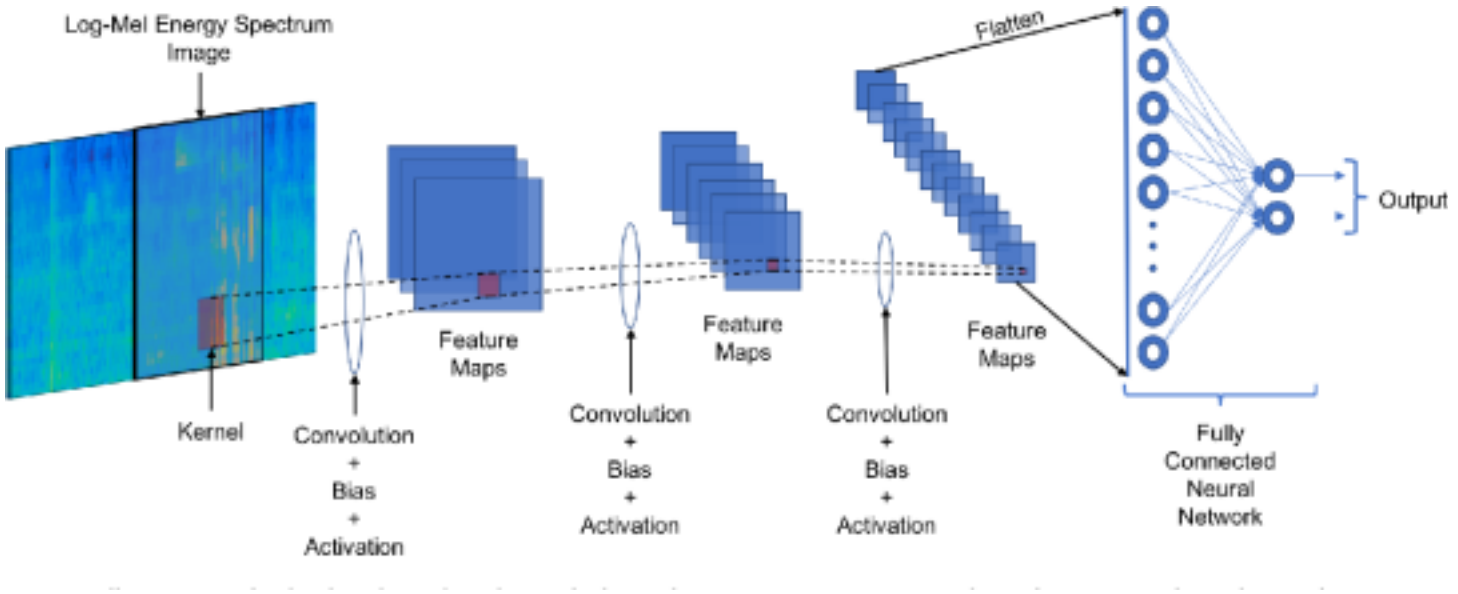
Feature Extractions

- MFCCs (20)
- Log Mel Filter-banks (40/80)
  - Spec Augmentation
    - Different frequency components
- PASE
  - Problem agnostic speech features
  - waveform based with out of box, no re-training
  - “derive useful speech representations by employing a self-supervised encoder-discriminator approach”
  - https://arxiv.org/abs/1904.03416



Model Selection

- Robust **frame-wise** voice activity detector with low False Alarm Rate
- Experimented several CNNs
  - Capture the “patterns” of spectrograms
- SwishNET:
  - **Swish refers to the swish activation functions**
  - x * sigmoid(β * x)
  - *** use with deeper models** (vanishing gradient! Case of 0! Relu :/)



https://arxiv.org/abs/1812.00149
https://www.utdallas.edu/ssprl/files/CNN-VAD-IEEE-Access.pdf

Swish-Net Architecture

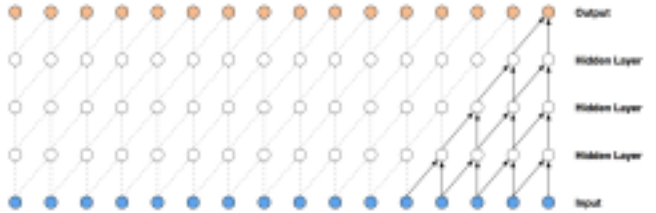


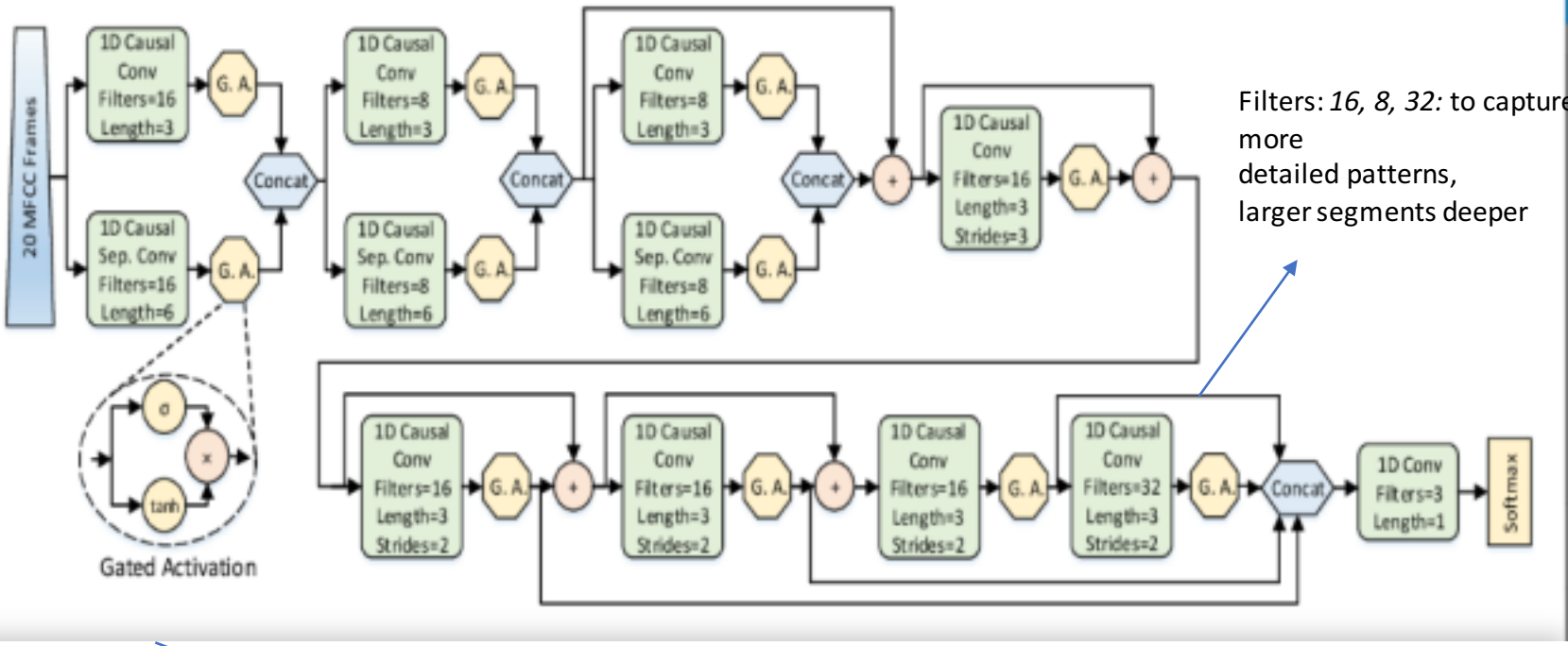
Figure 2: Visualization of a stack of causal convolutional layers.

- **8 1D Convolutions Layers**
- **Inspired by WaveNet, Inception Models * simpler**
  - *Across temporal domain feature extractions as channels*
  - *Divide and conquer {width}*
    - "parallel branches to capture features at different scales." Separable depth-wise convolutions
    - Convolutions that have *"causal* filters"* sliding over the data with reference.

    {t-*stride, t, t+stride}*
    - "Causal" can refer to a filter at time *step t can only see inputs that are no later than t+1 {not exactly here}*
    - Concatenate!

Different than auto-generative models *linear, no dilation =>

Keep increasing size of ***receptive field****: {non-exponentially} calculation of voice/vs. non-voice*



Filters: *16, 8, 32:* to capture more detailed patterns, larger segments deeper

Activation Functions: Gated activations: "act as memory gates" for previous features => ½ feature maps each layer

Swish-Net Training

- Regularization:
  - some dropout added between layers
- **FASTER THAN 2D!**
  - **1D convolutions, some skip connections between layers**
  - **Concatenation of feature maps**:
  - Branches with different filter sizes: merged
- Adam optimizer
- Batch size tuning
- Training on GPU on NRG1
  - ~45min per epoch, **very fast! : ~1 day per training, 1 hour to test with parallelization**

Evaluation

Classification Evaluation

F1 Score: 2*((precision*recall)/(precision + recall))

Speech Hit Rate (Recall): true positive/(true positive + false negative)

False Alarm Rate: false positive/(false positive + true negative)

	SPEECH PREDICTED [1]	NOISE PREDICTED [0]
SPEECH ACTUAL [1]	TRUE POSITIVE	FALSE NEGATIVE
NOISE ACTUAL [0]	FALSE POSITIVE	TRUE NEGATIVE

- per corruption-type results
  - level of SNR, T60
- per language on the symmetric dataset
  - French, English, Japanese

Baselines

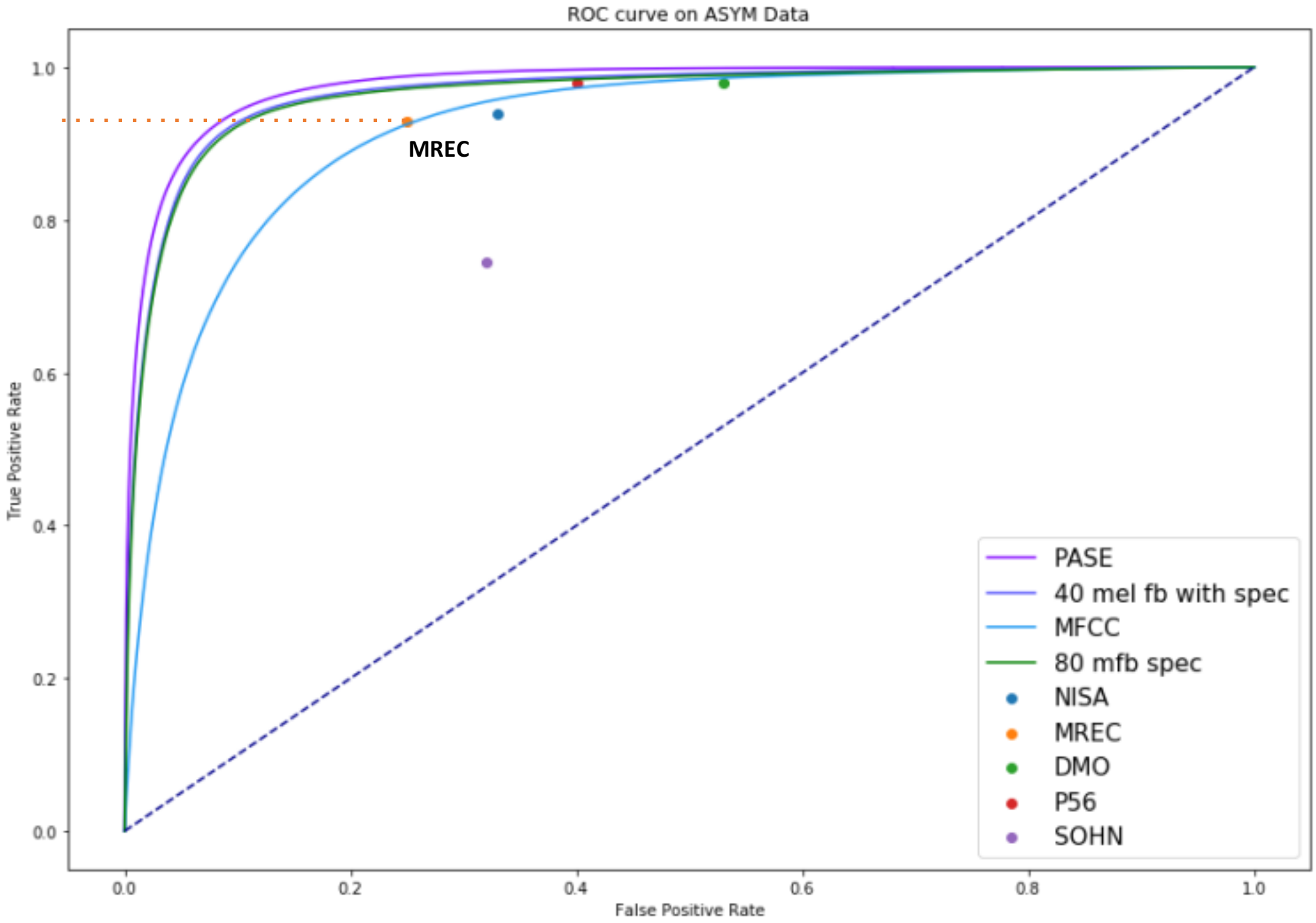
Chose models from Literature Review and Inhouse selections.

P56	Energy based VAD system
Sohn	Statistical VAD
NISA	Mean VAD posterior is being computed using a multi-task approach using MREC derived features and 350ms of context
MREC VAD	Feed forward neural network with output smoothing
DMO VAD	Neural Network based system with output smoothing

Test Dataset 1 Results

ASYMETRIC DATA SET: Method	F1 Score	Speech Hit Rate (Recall)	False Alarm Rate
NISA+ R1.3	0.76	0.94	0.38
P.56	0.79	0.98	0.4
SOHN	0.75	0.93	0.39
MREC VAD (utt. det.)	**0.83**	**0.93**	**0.25**
DMO	0.69	0.98	0.53
SNET 20_MFCC	0.83	0.77	0.12
SNET 40_MelFB	0.85	0.87	0.17
SNET 40_MelFB+SpecAug	**0.89**	**0.95**	**0.1**
SNET 80_MelFB+SpecAug	**0.89**	**0.95**	**0.11**
SNET PASE	0.9	**0.9**	**0.06**

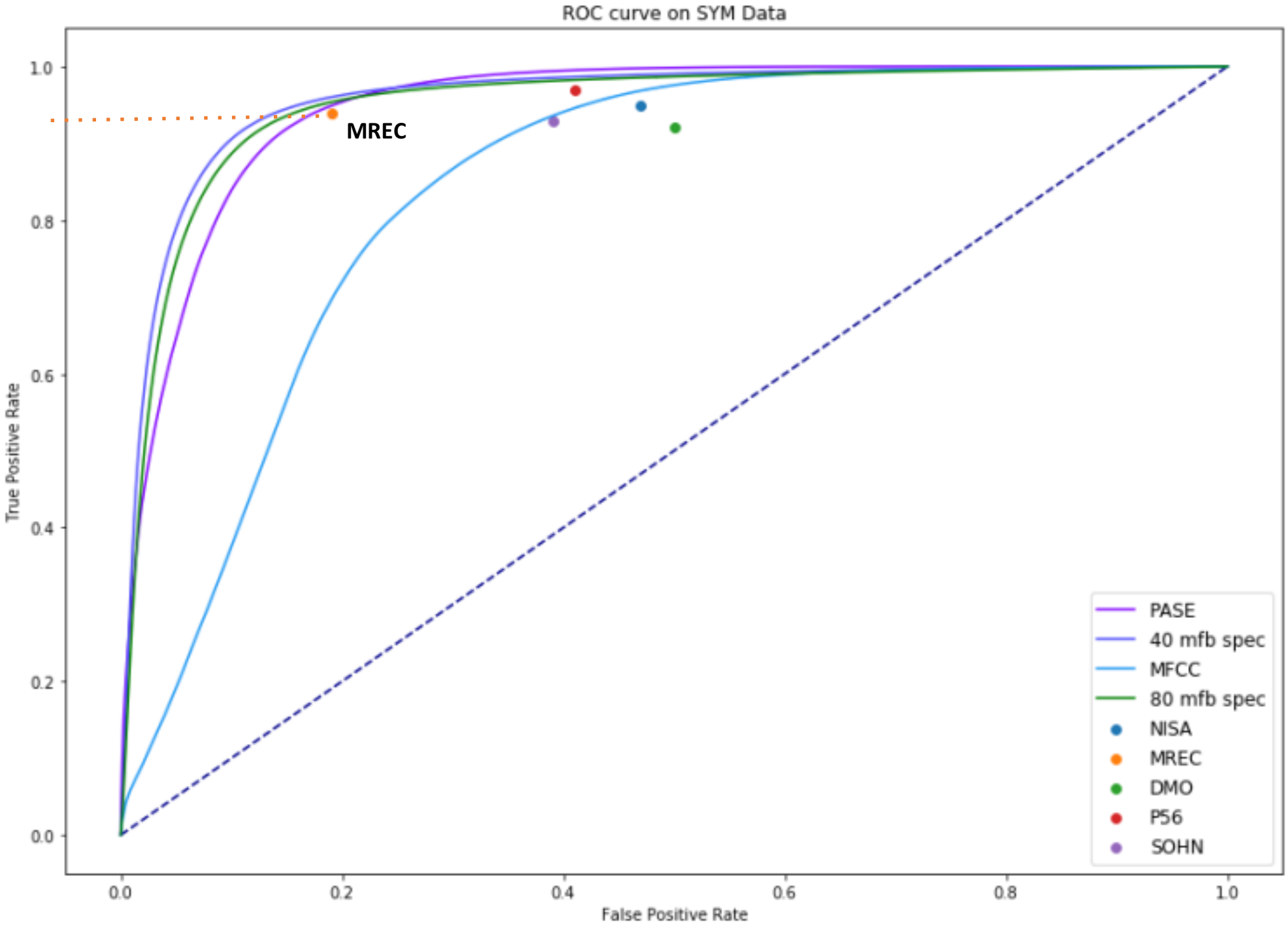
ASYM ROC Curve



Test Dataset 2 Results

SYMETRIC DATA SET: Method	F1 Score	Speech Hit Rate (Recall)	False Alarm Rate
NISA+ R1.3	0.72	0.95	0.47
P.56	0.83	0.97	0.41
SOHN	0.75	0.93	0.39
MREC VAD (utt. det.)	**0.87**	**0.94**	**0.19**
DMO	0.69	0.92	0.5
SNET 20_MFCC	0.76	0.92	0.37
SNET 40_MelFB	0.85	0.81	0.17
SNET 40_MelFB+SpecAug	**0.9**	**0.94**	**0.13**
SNET 80_MelFB+SpecAug	0.89	**0.93**	**0.14**
SNET PASE	**0.88**	**0.92**	**0.15**

SYM
ROC Curve



Ambient Noise Condition

Method	Type	F1Score	SHR	FAR
SwishNet (40 mb with spec) low T60		0.93	0.98	0.08
SwishNet (40 mb with spec) Ambient high t60		0.94	0.99	0.11
MREC low T60		0.9	0.99	0.19
MREC high T60		0.89	0.98	0.18
DMO low T60		0.77	0.99	0.42
DMO high T60		0.77	1	0.41

Threshold of 0.5

Method	**Type**	**F1Score**	**SHR**	**FAR**
SwishNet (40 mb with spec) low T60		0.94	0.99	0.11
SwishNet (40 mb with spec) Ambient high t60		0.92	0.98	0.15
MREC low T60		0.93	0.99	0.14
MREC high T60		0.93	0.99	0.12
DMO low T60		0.76	1	0.47
DMO high T60		0.71	1	0.54

Music & Babble Noise Condition

ASYMETRIC DATA SET: Method	F1 Score	Speech Hit Rate (Recall)	False Alarm Rate
MREC VAD Babble & Music	0.745	0.935	0.385
DMO Babble & Music	0.69	0.92	0.5
SNET 40_MelFB+SpecAug Babble & Music	0.9	0.94	0.13

SYMETRIC DATA SET: Method	F1 Score	Speech Hit Rate (Recall)	False Alarm Rate
MREC VAD Babble & Music	0.835	0.905	0.21
DMO Babble & Music	0.63	0.905	0.59
SNET 40_MelFB+SpecAug Babble & Music	**0.86**	**0.91**	**0.17**

Narrowband VAD

- To be used with the NTE/VM2T products
  - Created 8KHz SCT configs from 16K setup
    - Added IRS filter: bandpass filter, bandwidth: (400 to 3400 Hz)
  - Retrained SwishNet model using 40 log mel filter banks with spec (3 channels)
- Very similar results observed

Take-Aways & Next Steps

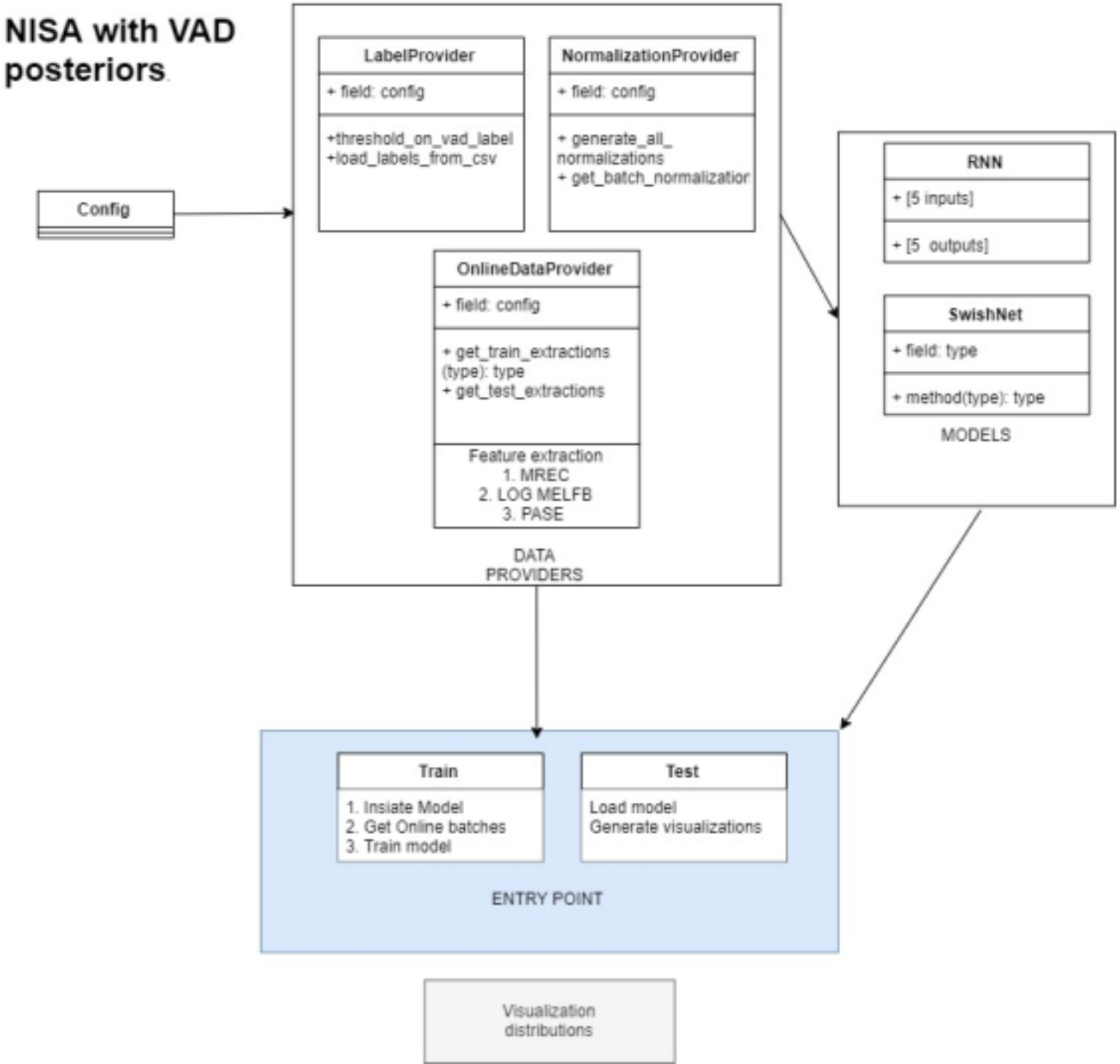
Hands-on exposure to audio data processing, corruption and modelling

1. *In progress: Adapt SwishNet CNN to a regression model {estimate NISA parameters?}*

Integration with NISA

Ongoing

NISA with VAD posteriors



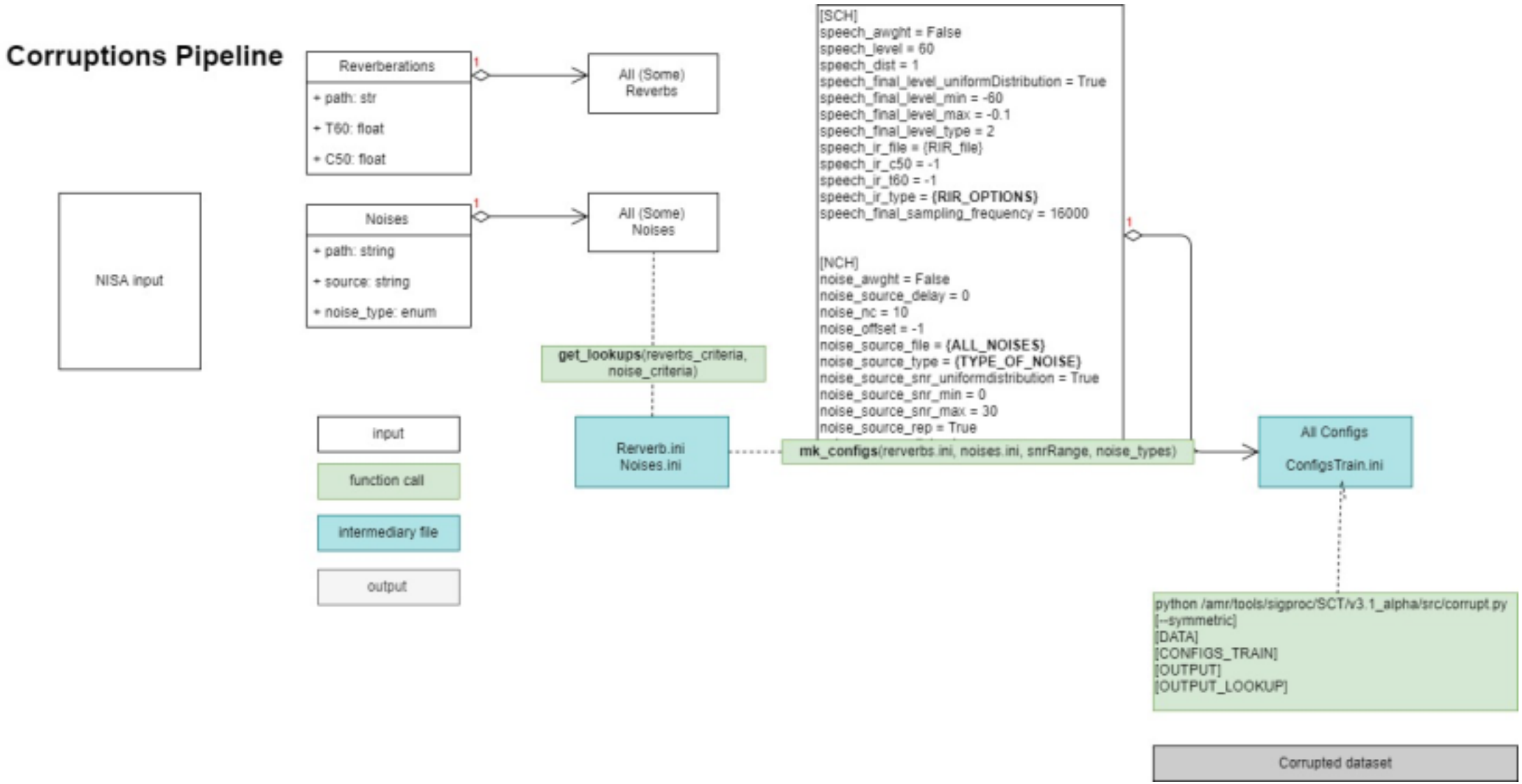
Process

1) Refactor the framework to be Object-Oriented and to use a flexible data provider
2) Determine whether VAD posteriors are helpful?
3) Experiment with different feature extractions with our LSTM
4) Experiment with another model – CNN with GRU gates?

Next Steps:

- Extended internship until April ☺
- VAD:
  - Use an HMM based utterance detection module on top of the posterior generator – with Carl
  - Optimizations? Different activation functions? Filters?
- NISA
  - Test results to come *{training now}*
  - 8k models for NISA+ (needed for the NTE/VM2T products)
  - Add Bit rate estimation to NISA+
  - Architecture experiments
- SCT config generation (in progress) and ASR experiments

Corruptions Pipeline



Questions?

Thank you!

Special thanks to Dushyant & Carl for the mentorship.

Code:

VAD: https://git.labs.nuance.com/lucia.berger/vad_detection

NISA: https://git.labs.nuance.com/dushyant.sharma/NISA

Literature Reviews: https://nuance.jiveon.com/people/lucia.berger/blog

Sources : https://jeddy92.github.io/JEddy92.github.io/ts_seq2seq_conv/