Frame-based Overlapping Speech Detection using Convolutional Neural Networks

Midia Yousefi, John H.L. Hansen

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
Department of Electrical Engineering
University of Texas at Dallas
Richardson, Texas 75083-0688, U.S.A.

ICASSP 2020

May 4-8, 2019 Barcelona, Spain

Introduction

- Spontaneous conversations such as meetings, debates, and telephone conversations tend to contain overlapping speech, i.e., time segments where more than one speaker is active.

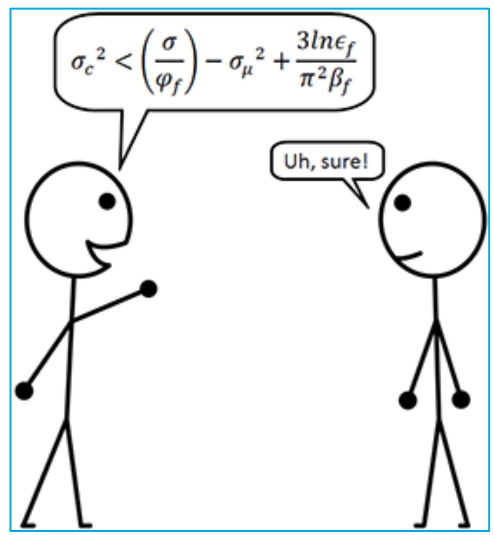
Applications affected by overlapping speech:

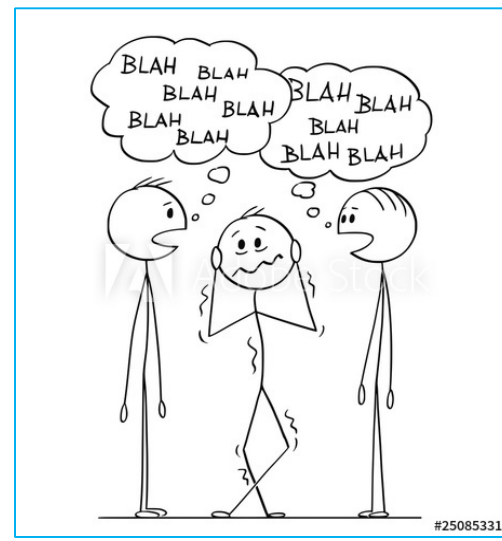
- Digital hearing aids

- Automatic Speech Recognition (ASR)

- Speaker verification, identification and recognition

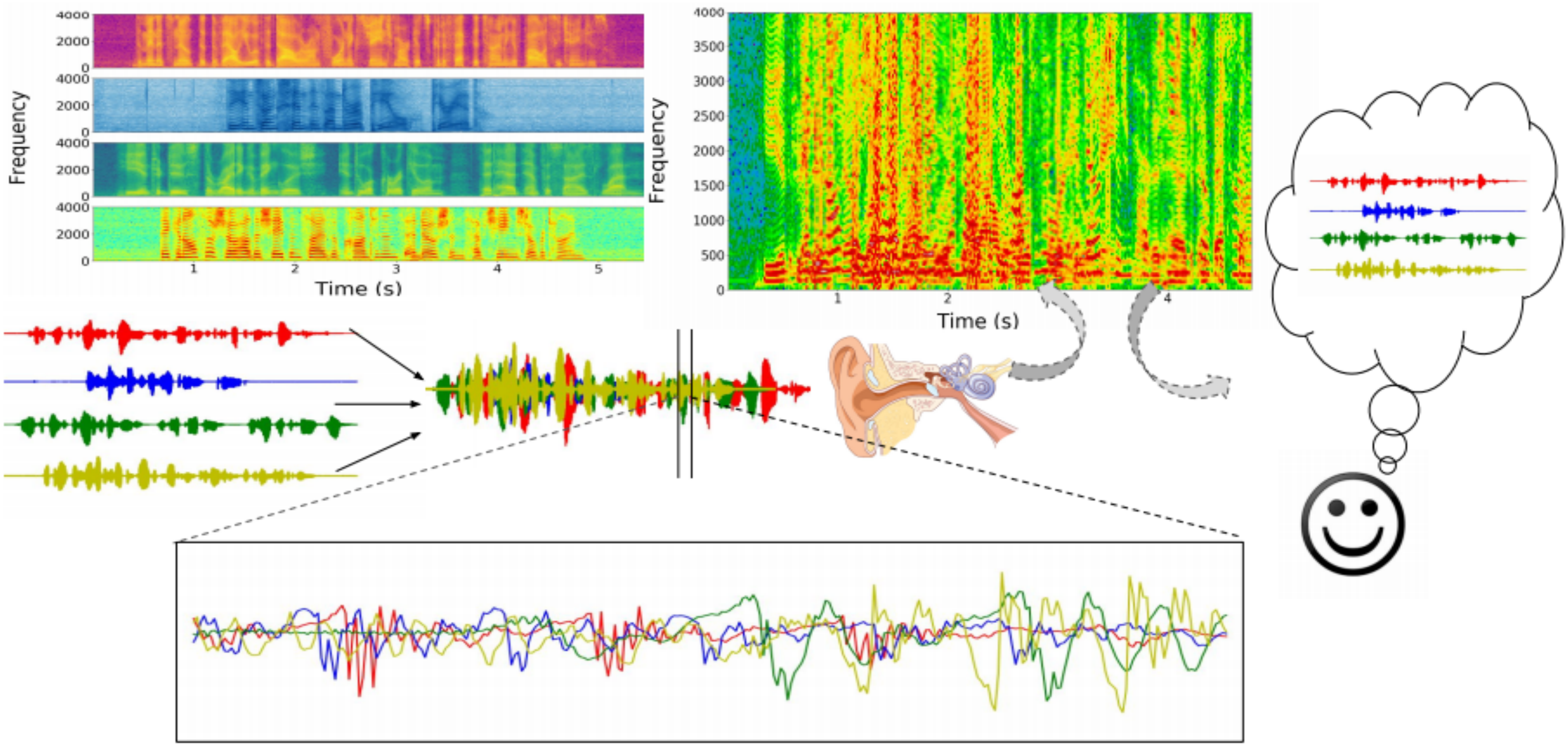
- Mobile voice telecommunication





Cocktail party problem

- Cocktail party problem is a psychoacoustic phenomena; refers to ability of human auditory system to selectively **attend**, **recognize** and **extract meaningful information** from complex auditory signals in noisy environments, where interference is from competing talkers.



Email: {Midia.Yousefi, John.Hansen}@utdallas.edu

Challenges

- Researchers have addressed co-channel speech challenge using two major approaches:

1) **Detecting** overlap speech segments & **removing** from the dataset

Results in building **better speaker-specific models** for speaker diarization/recognition

2) **Detecting** overlap speech segments & **separating** individual speech signals out of the mixture

Results in **better performance in identifying** active speakers and **recognizing** their associated speech content

Email: {Midia.Yousefi, John.Hansen}@utdallas.edu

Previous approaches

Unsupervised:

These approaches typically use signal processing methods to design suitable features for detecting overlapping segments.

1. Spectral Auto-correlation Peak Valley Ratio (SAPVR).
2. Measuring the Gaussianity of speech segment using Kurtosis
3. Zero crossing
4. Spectral flow
5. Harmonicity

Supervised:

Supervised approaches use model-based techniques to learn representations for both single speaker & overlapping speech segments

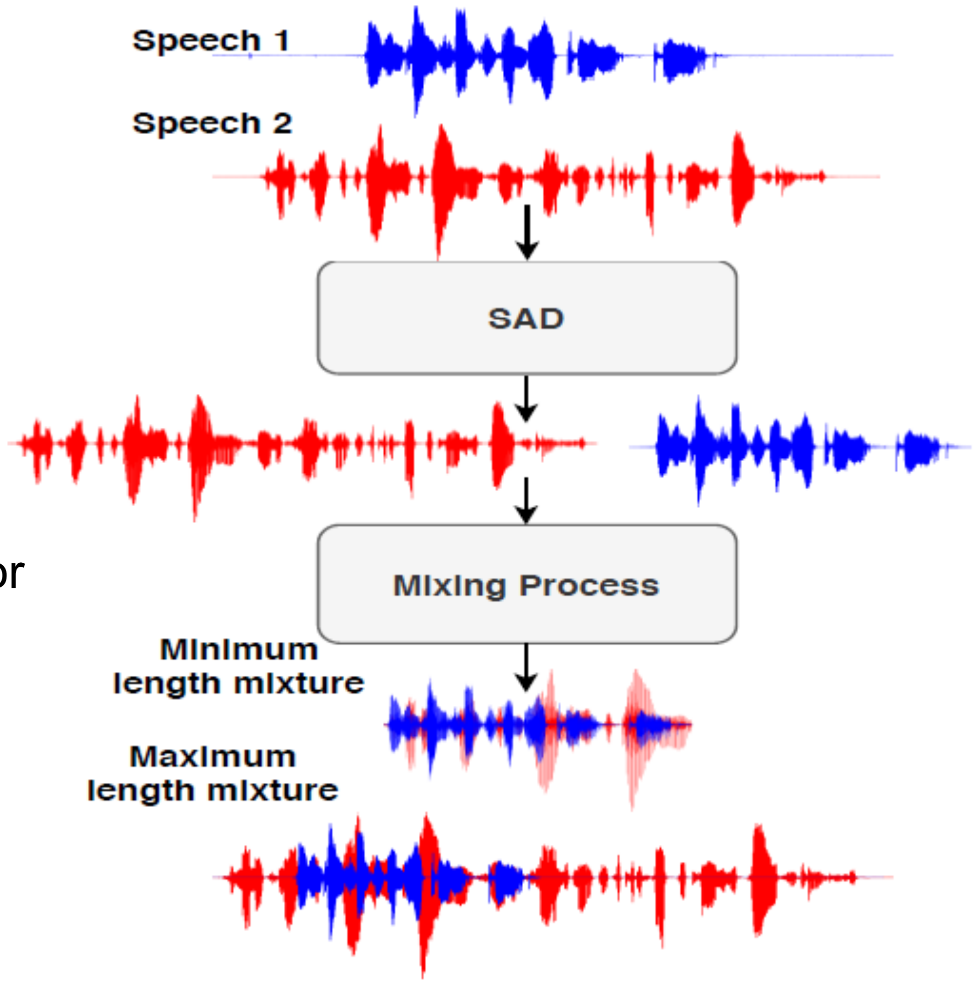
1. Non-negative Matrix Factorization (NMF)
2. Long Short Term Memory (LSTM) Networks
3. Convolutional Neural Networks (CNNs)

Problem formulation

- **Overlapping speech dataset generated considering two scenarios:**

  1. Entire utterance contains overlapping speech.

  2. Utterance contains both overlapping & clean speech.



- **Drawbacks of manually designed features:**

  1. May not be best representation for modeling competing talker; could lead to sub-optimal results.

  2. Can be fragile in noisy conditions.

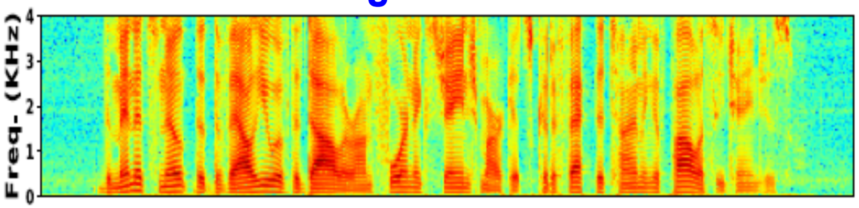
Email: {Midia.Yousefi, John.Hansen}@utdallas.edu

Spectral Features

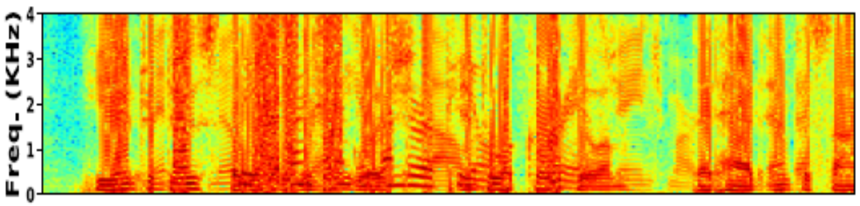
- **Feature extraction for classifier training:**
  - **256-dim spectral magnitude:**

    257-dim feature vector calculated using a 512-dim (STFT) computed over a 25 ms Hamming window with 10 ms of frame shift.

Single talker



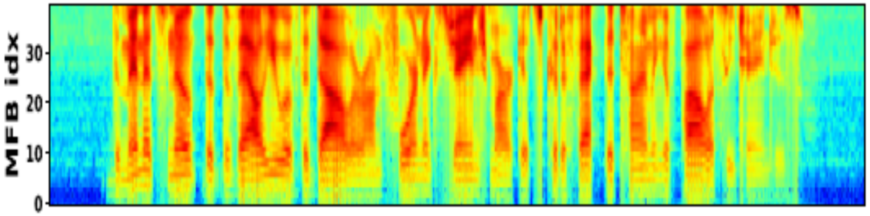
Overlapping speech



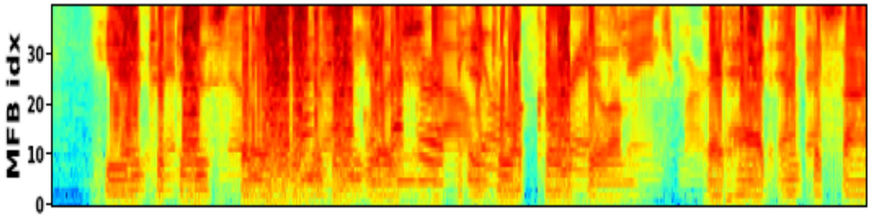
  - **40-dim Mel Filter Banks (MFB)**

    Calculated by applying the 40 Mel-scale filter banks to the power spectrum of the speech signal.

Single talker



Overlapping speech



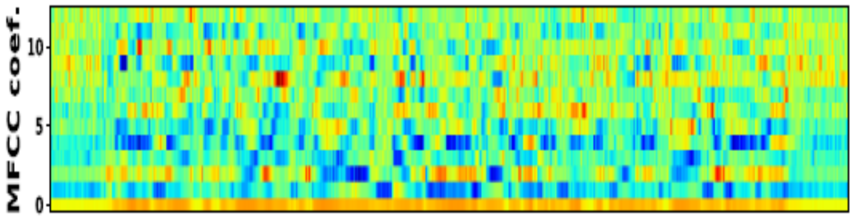
Email: {Midia.Yousefi, John.Hansen}@utdallas.edu

Spectral features

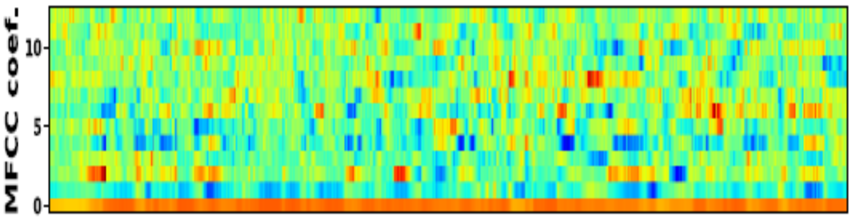
- **39-dim Mel Frequency Cepstral Coefficients (MFCCs)**

  Calculating Mel FilterBank, then logarithm of filter bank energies derived; Discrete Cosine Transform (DCT) is then applied.

Single talker



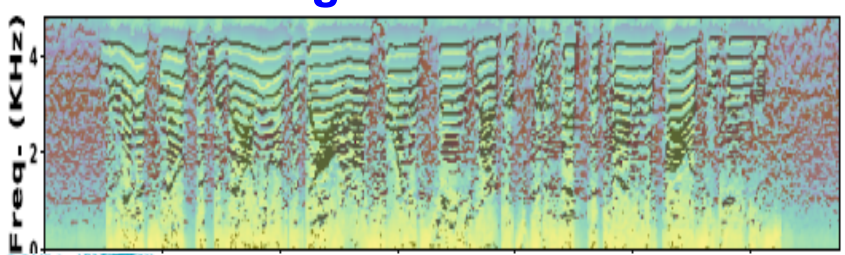
Overlapping speech



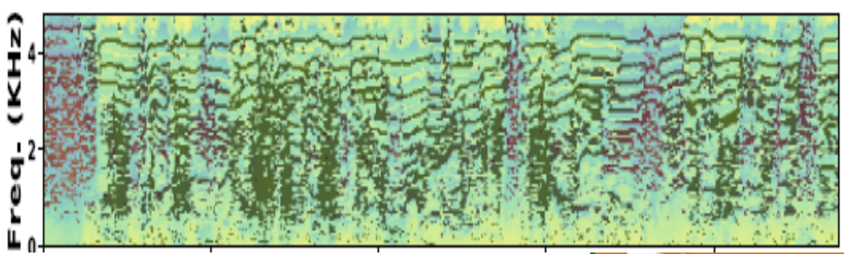
- **120-dim Pyknogram:**

  Pyknogram enhances speech spectrogram by performing AM-FM (Amplitude-Frequency Modulation) analysis; this decomposes speech spectral sub-bands into amplitude & frequency components.

Single talker



Overlapping speech



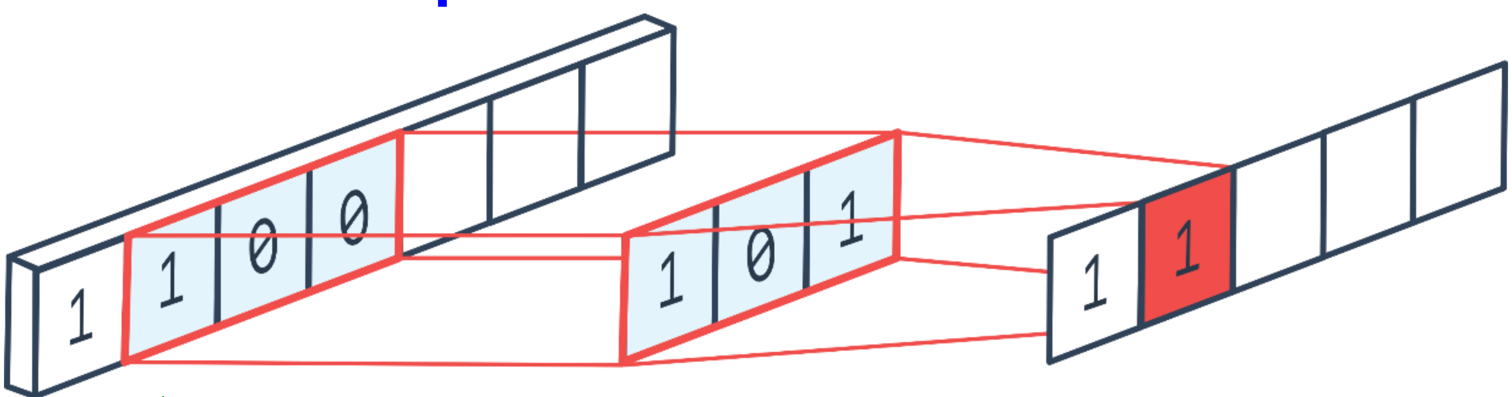
Email: {Midia.Yousefi, John.Hansen}@utdallas.edu

Classifier

Convolutional Neural Network:

- Classical approaches to problem involve hand crafting features from time series data; difficulty is that this uses fixed-sized windows.
- This feature engineering requires deep expertise in the field.
- Convolutional Neural Network (CNN) is the foundation of many supervised solutions for problems such as computer vision

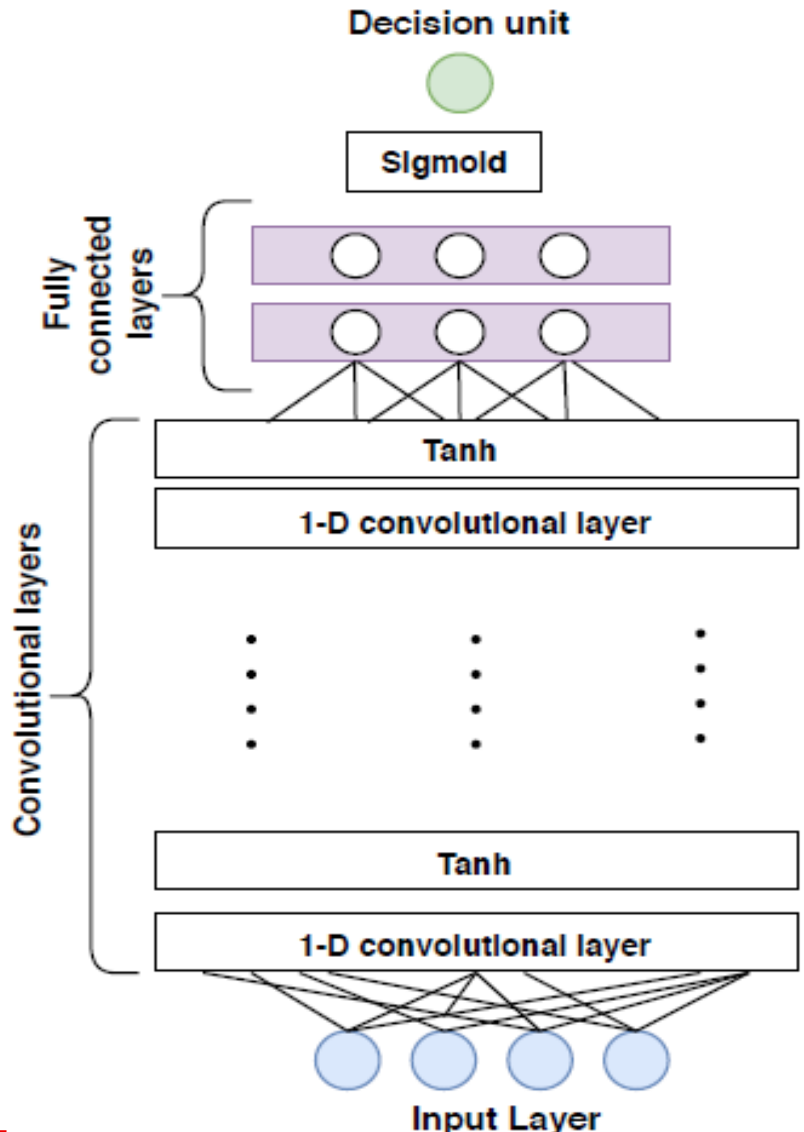
Convolutional operation:



- A “filter”, sometimes called a “kernel”, is passed over the feature vector, viewing a few samples at a time.

Proposed Architecture

- **Proposed design:**
  - **1-D Convolutional layer:**
    Convolution using a 'kernel' to extract certain 'features' from the input.
  - **Tanh activation function:**
    Activation layer introduces non-linearity to allow network to train itself
  - **Fully connected layer:**
    Used to reduce the dimensions of the extracted features by the CNN
  - **Sigmoid activation function:**
    Generating the probability of each class for the data samples



Experimental setup

- **Dataset:**
  - Naturalistic data, like AMI corpus has been used to evaluate systems for overlapping speech detection.
  - AMI dataset contains only 5-10% overlapping speech; not sufficient for training DNNs.

We generate overlapping speech based on the GRID corpus.

- **The GRID corpus:**
  - A multi-speaker, sentence-based corpus.
  - Contains 34 speakers, 18 males and 16 females each narrating 1000 sentences.

The generated overlapping speech dataset

Train	Development	Test
20hrs	3hrs	2hrs

Email: {Midia.Yousefi, John.Hansen}@utdallas.edu

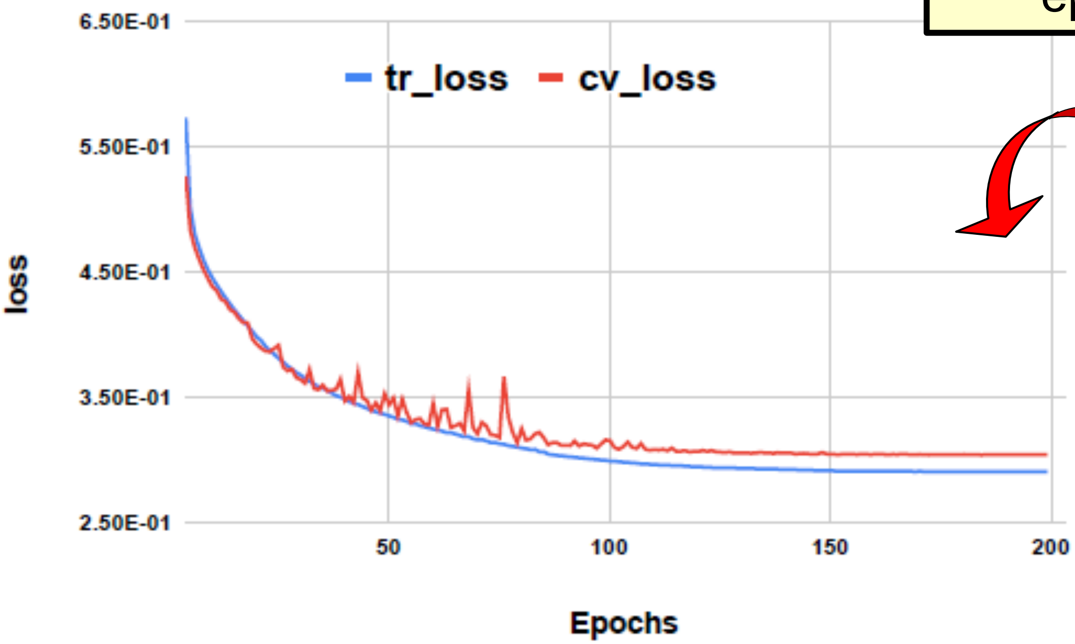
Model training

Model training and experiments:

Hyper-parameters tuning:

1. Number of layers
2. Kernel size
3. Channel size
4. Learning rate

1. 6 1-D convolutional layer
2. Kernel size of 3
3. 128 output channel
4. 200 epoch, batch size 32
5. Learning rate 0.001, reduced by 0.5 if no improvement in validation loss for 3 successive epochs.



Training & validation loss depicts ability of the network to generalize to unseen speech segments in the development phase.

Email: {Midia.Yousefi, John.Hansen}@utdallas.edu

Experimental Results

- **Precision:** correctly detected overlapping segments vs. the total number of overlapping segments:
- **Recall:** ability of model to find all overlapping segments in dataset; measured as ratio of correctly detected overlap segments to total number of actual overlapping.
- **Fscore:** defined as harmonic mean of recall & precision.
- **Time:** processing time per epoch for each experiment is also captured.

Considering classification measures & time efficiency, **MFCC** outperforms other features

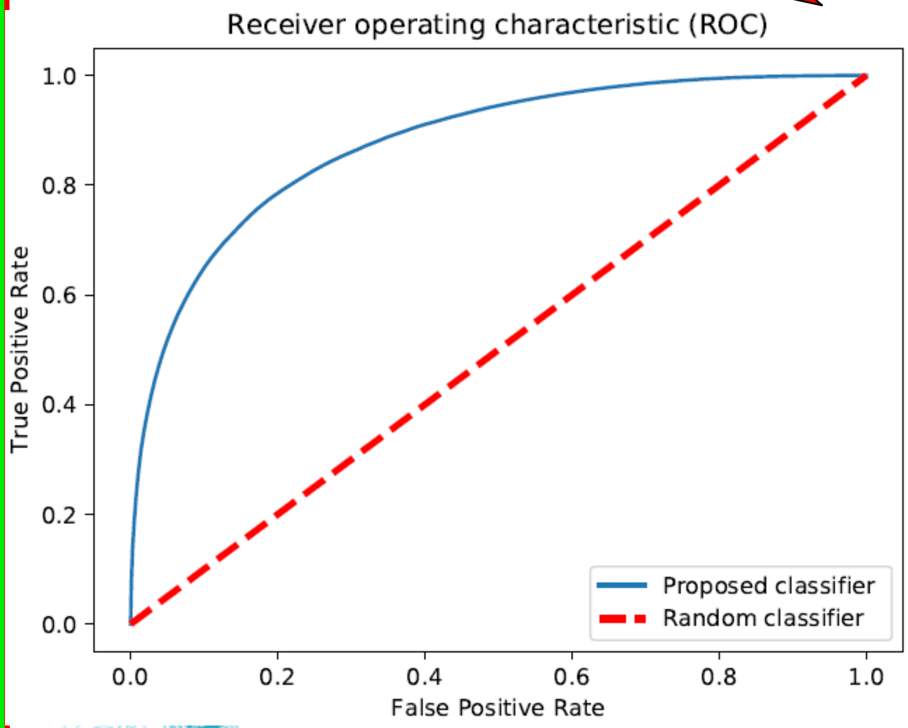
Male-Male	MagSpec	Pykno	MFB	MFCC
Accuracy	79%	82%	78%	81%
Precision	80%	84%	81%	82%
Recall	90%	91%	91%	90%
Fscore	85%	87%	86%	86%
Time	898s	530s	247s	220s

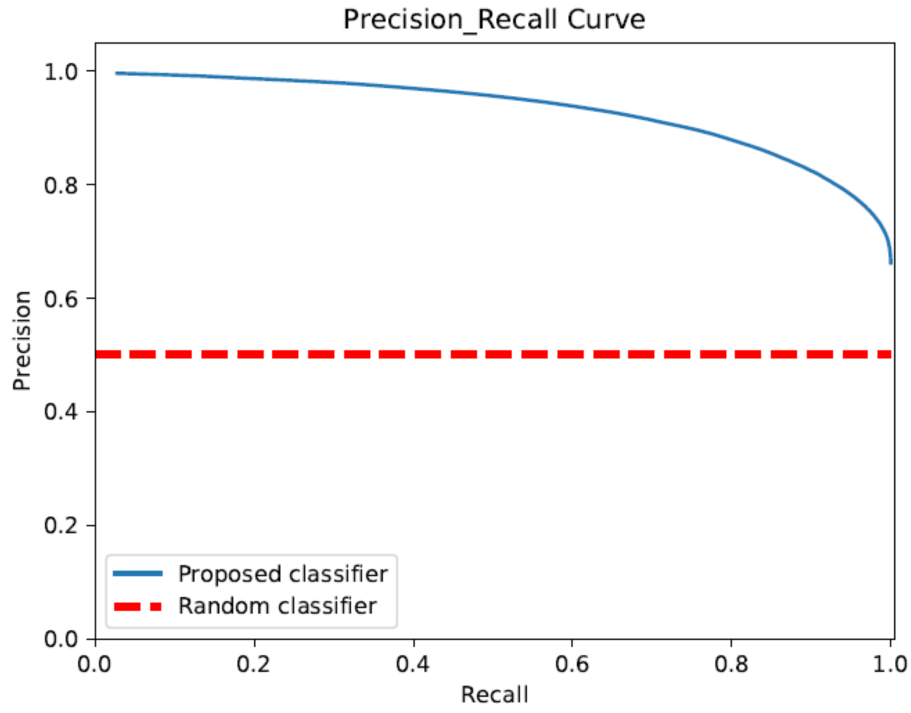
Female-Female	MagSpec	Pykno	MFB	MFCC
Accuracy	82%	84%	82%	83%
Precision	83%	86%	84%	85%
Recall	91%	91%	91%	91%
Fscore	87%	88%	86%	88%
Time	998s	536s	250s	216s

Experimental results

ROC & **Precision-Recall** curves based on MFCC feature derived from male-male data test-set.

Male-Female	MagSpec	Pykno	MFB	MFCC
Accuracy	88%	89%	89%	89%
Precision	91%	92%	92%	92%
Recall	91%	91%	92%	91%
Fscore	91%	92%	92%	92%
Time	933s	510s	230s	217s





Discussion and conclusions

Conclusions on proposed overlapping speech detection system:

- A CNN architecture was introduced to classify overlapping speech on frames as short 25 ms.
- Proposed CNN architecture was trained using 4 spectral features:

  Spectral magnitude, **MFB**, **MFCC**, **Pyknogram**.

- Accuracy for spectral magnitude is **79%** for male-male dataset.
- Fscore of male-male dataset is **85%;** generally a good performance for classification; however precision is **80%,** which is 10% lower than recall (90%).
- Magnitude spectra is a dense feature; processing time is high in each epoch.
- Second largest feature is Pyknogram; outperforms spectrogram in both classification metrics & processing time; not computationally efficient compared to 39-dim MFCC and 40-dim MFB.
- **Pyknograms & MFCC have competitive classification performance, while MFCC is a lower dimensional feature & reduces processing time by 60%.**

Email: {Midia.Yousefi, John.Hansen}@utdallas.edu

Questions

?

Email: {Midia.Yousefi, John.Hansen}@utdallas.edu