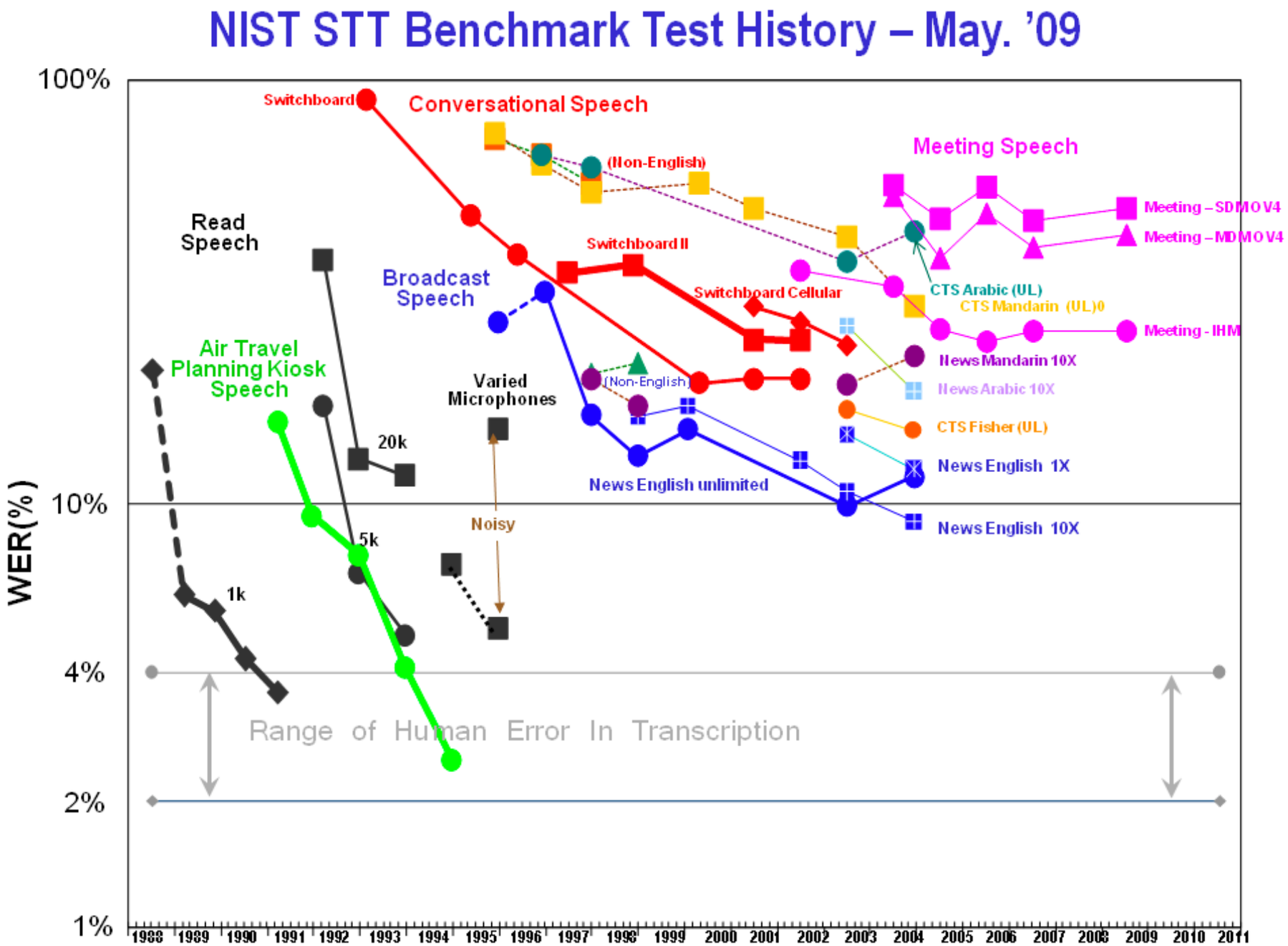
My 5 minute talk about 2 of my ongoing projects

Desh Raj

Noise-aware Training of Acoustic Models

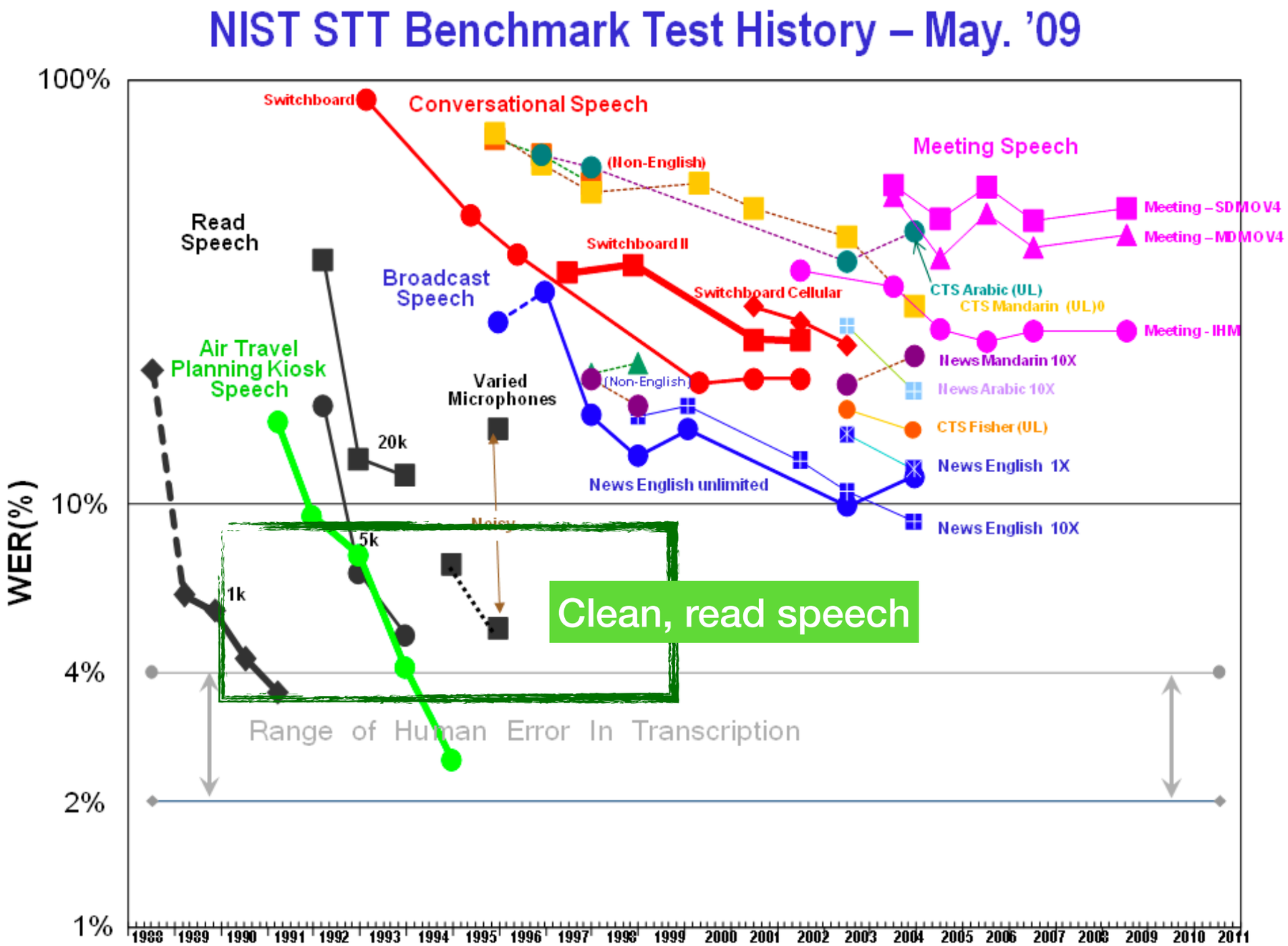
Noise-aware Training of Acoustic Models

History



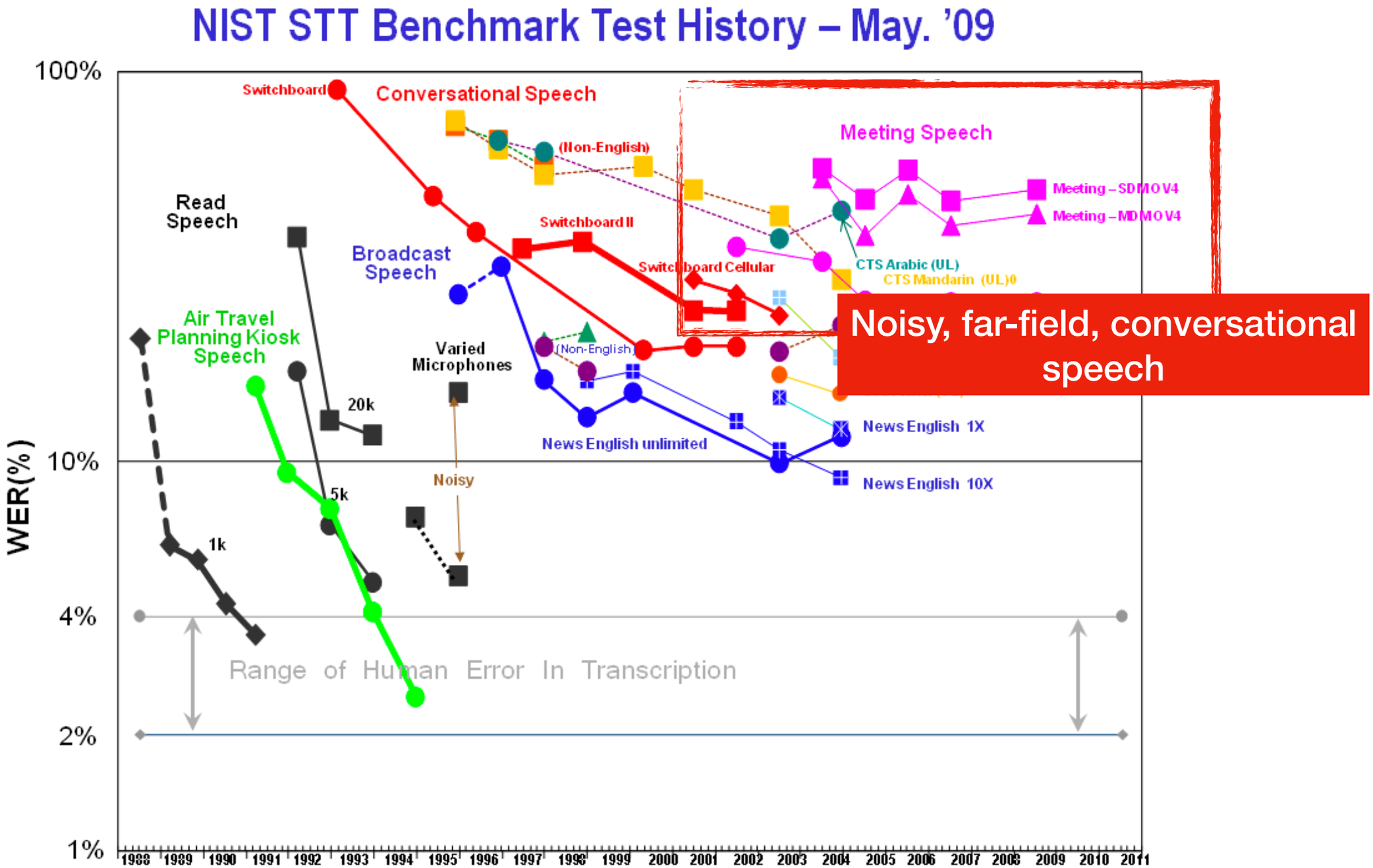
Noise-aware Training of Acoustic Models

History



Noise-aware Training of Acoustic Models

History



Noisy, far-field, conversational speech

Noise-aware Training of Acoustic Models

The Problem

- Speaker effects
- Background noise
- Reverberation

Noise-aware Training of Acoustic Models

The Problem

- Speaker effects
- Background noise
- Reverberation

Noise-aware Training of Acoustic Models

Well, actually…

VERY DEEP CONVOLUTIONAL NEURAL NETWORKS FOR ROBUST SPEECH RECOGNITION

Yanmin Qian[1,2], *Philip C Woodland*[2]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]Cambridge University Engineering Department, Cambridge CB2 1PZ, UK
yanminqian@sjtu.edu.cn, pcw@eng.cam.ac.uk

Noise-aware Training of Acoustic Models

Well, actually…

VERY DEEP CONVOLUTIONAL NEURAL NETWORKS FOR ROBUST SPEECH RECOGNITION

[1] Department of
[2]C

Very Deep Self-Attention Networks for End-to-End Speech Recognition

Ngoc-Quan Pham[1], *Thai-Son Nguyen*[1], *Jan Niehues*[1], *Markus Müller*[1], *Sebastian Stüker*[1], *Alex Waibel*[1,2]

[1]Interactive Systems Lab, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]Carnegie Mellon University, Pittsburgh PA, USA

ngoc.pham@kit.edu, thai.nguyen@kit.edu

Noise-aware Training of Acoustic Models

Well, actually…

VERY DEEP CONVOLUTIONAL NEURAL NETWORKS FOR ROBUST SPEECH RECOGNITION

1

Untangling in Invariant Speech Recognition

Cory Stephenson
Intel AI Lab
cory.stephenson@intel.com

Jenelle Feather
MIT
jfeather@mit.edu

Suchismita Padhy
Intel AI Lab
suchismita.padhy@intel.com

Oguz Elibol
Intel AI Lab
oguz.h.elibol@intel.com

Hanlin Tang
Intel AI Lab
hanlin.tang@intel.com

Josh McDermott
MIT/ Center for Brains, Minds, and Machines
jhm@mit.edu

SueYeon Chung
Columbia University/ MIT
sueyeon@mit.edu

peech Recognition

Sebastian Stüker[1], *Alex*

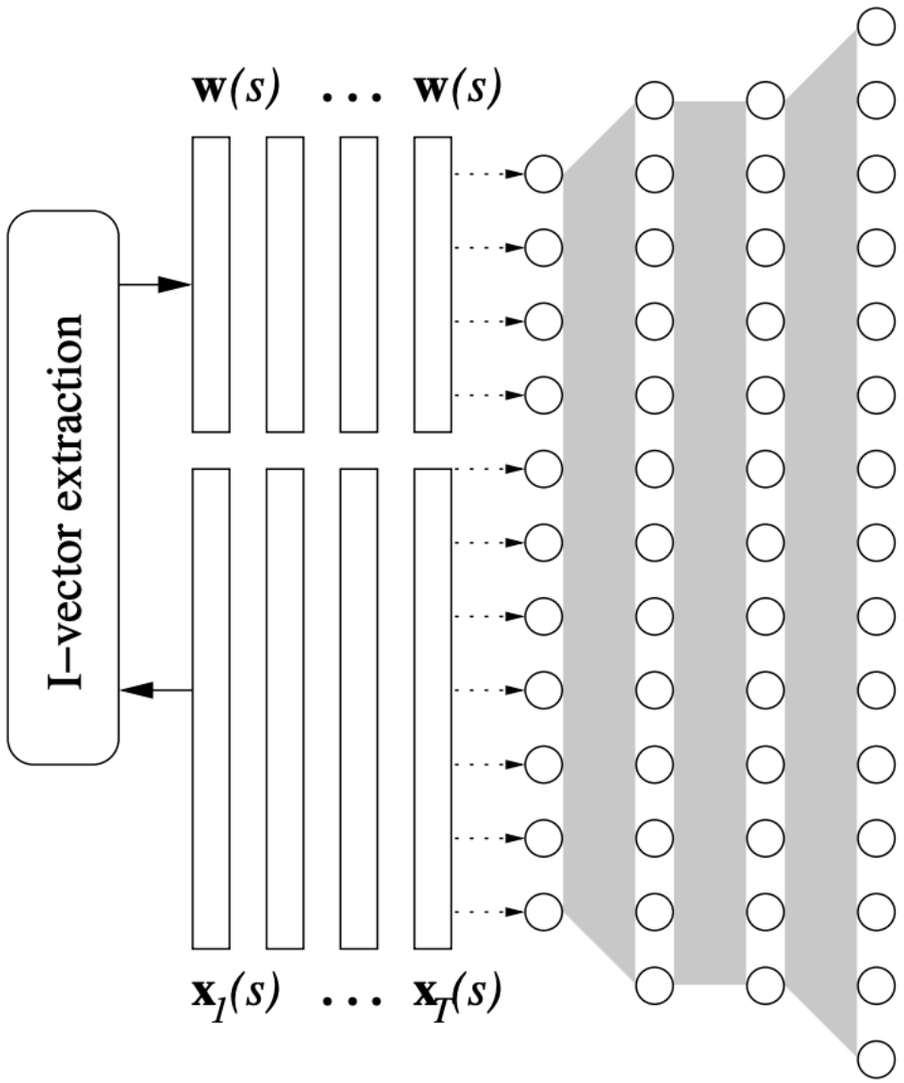
rlsruhe, Germany
A

“But we don’t have enough data or GPUs.”

- Every DL practitioner not working at Froogle

Noise-aware Training of Acoustic Models

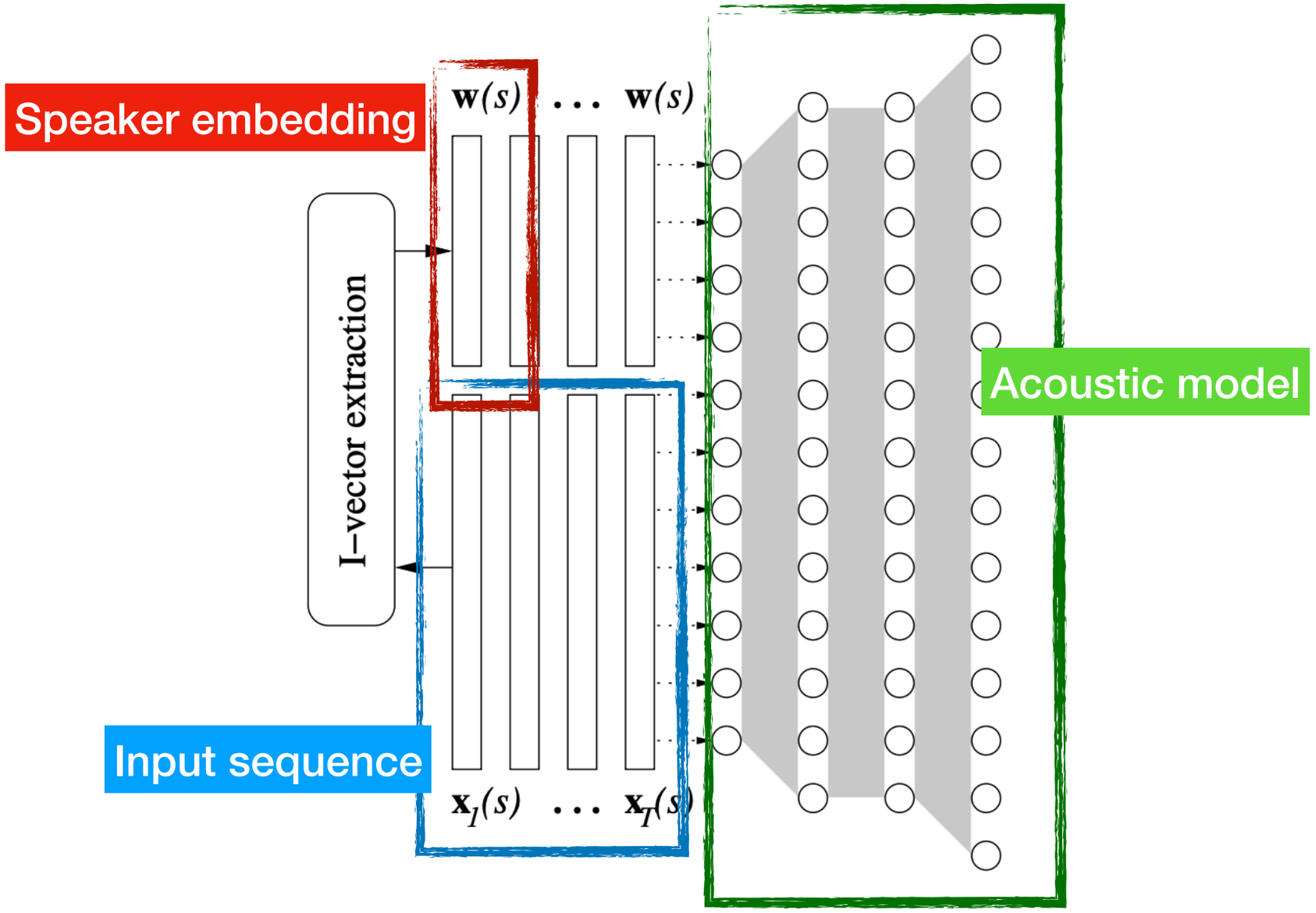
An idea from speaker adaptation



Saon, George, et al. "Speaker adaptation of neural network acoustic models using i-vectors." *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013.

Noise-aware Training of Acoustic Models

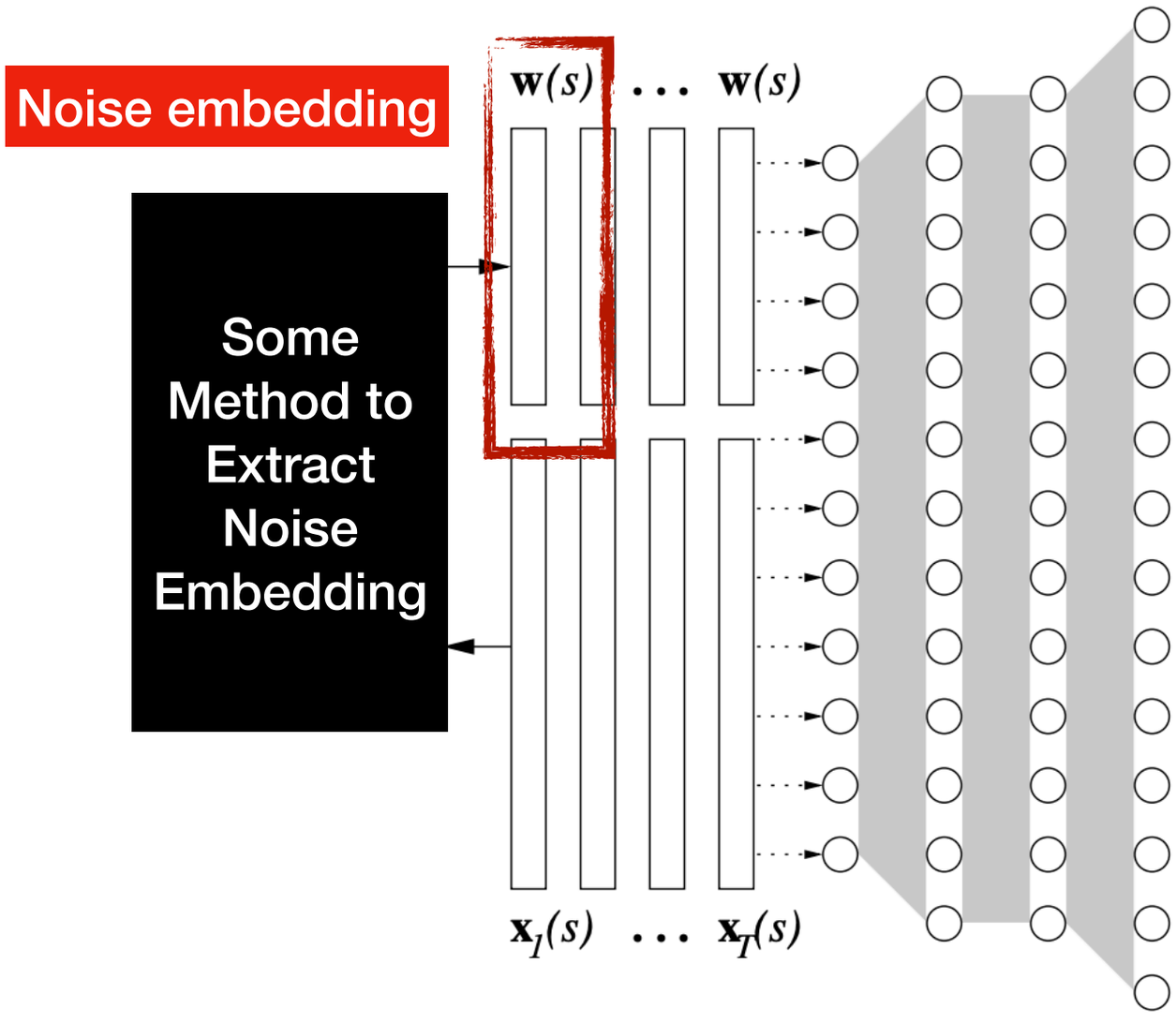
An idea from speaker adaptation



Saon, George, et al. "Speaker adaptation of neural network acoustic models using i-vectors." *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013.

Noise-aware Training of Acoustic Models

An idea from speaker adaptation



Noise-aware Training of Acoustic Models

Where to get noise embeddings?

Noise-aware Training of Acoustic Models

Where to get noise embeddings?

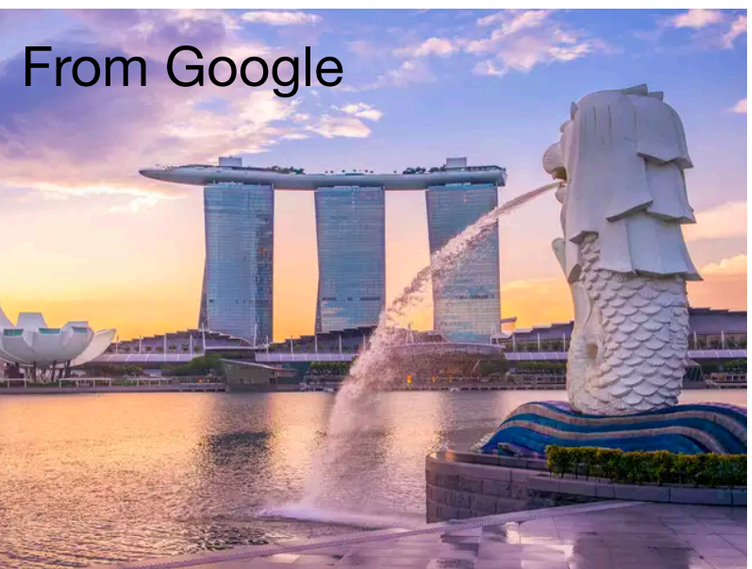
ASRU 2019

PROBING THE INFORMATION ENCODED IN X-VECTORS

Desh Raj, David Snyder, Daniel Povey, Sanjeev Khudanpur

Center for Language and Speech Processing & Human Language Technology Center of Excellence
The Johns Hopkins University, Baltimore, MD 21218, USA.

draj@cs.jhu.edu, {david.ryan.snyder, dpovey}@gmail.com, khudanpur@jhu.edu



Noise-aware Training of Acoustic Models

Where to get noise embeddings?

Discriminatively trained speaker embeddings

ASRU 2019

PROBING THE INFORMATION ENCODED IN X-VECTORS

Desh Raj, David Snyder, Daniel Povey, Sanjeev Khudanpur

Center for Language and Speech Processing & Human Language Technology Center of Excellence
The Johns Hopkins University, Baltimore, MD 21218, USA.

draj@cs.jhu.edu, {david.ryan.snyder, dpovey}@gmail.com, khudanpur@jhu.edu

Noise-aware Training of Acoustic Models

Where to get noise embeddings?

ASRU 2019

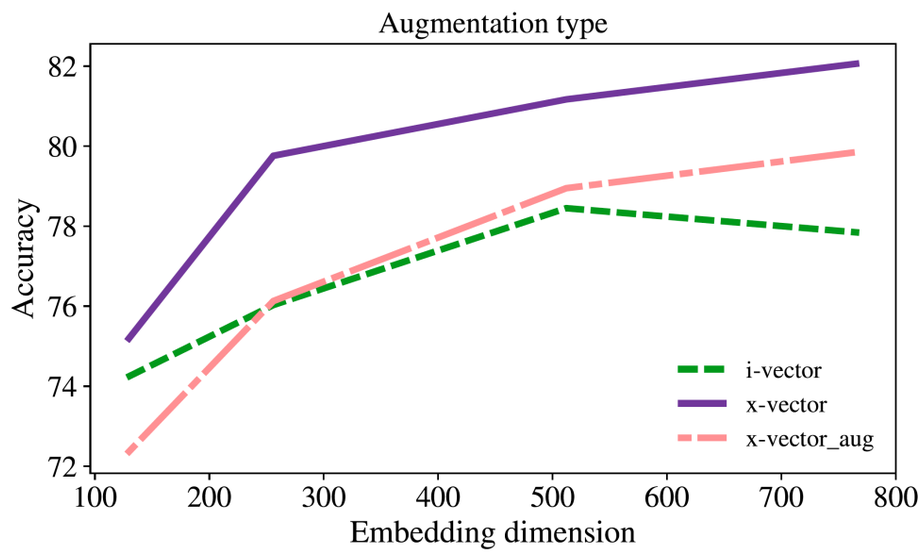
PROBING THE INFORMATION ENCODED IN X-VECTORS

Desh Raj, David Snyder, Daniel Povey, Sanjeev Khudanpur

Center for Language and Speech Processing & Human Language Technology Center of Excellence
The Johns Hopkins University, Baltimore, MD 21218, USA.

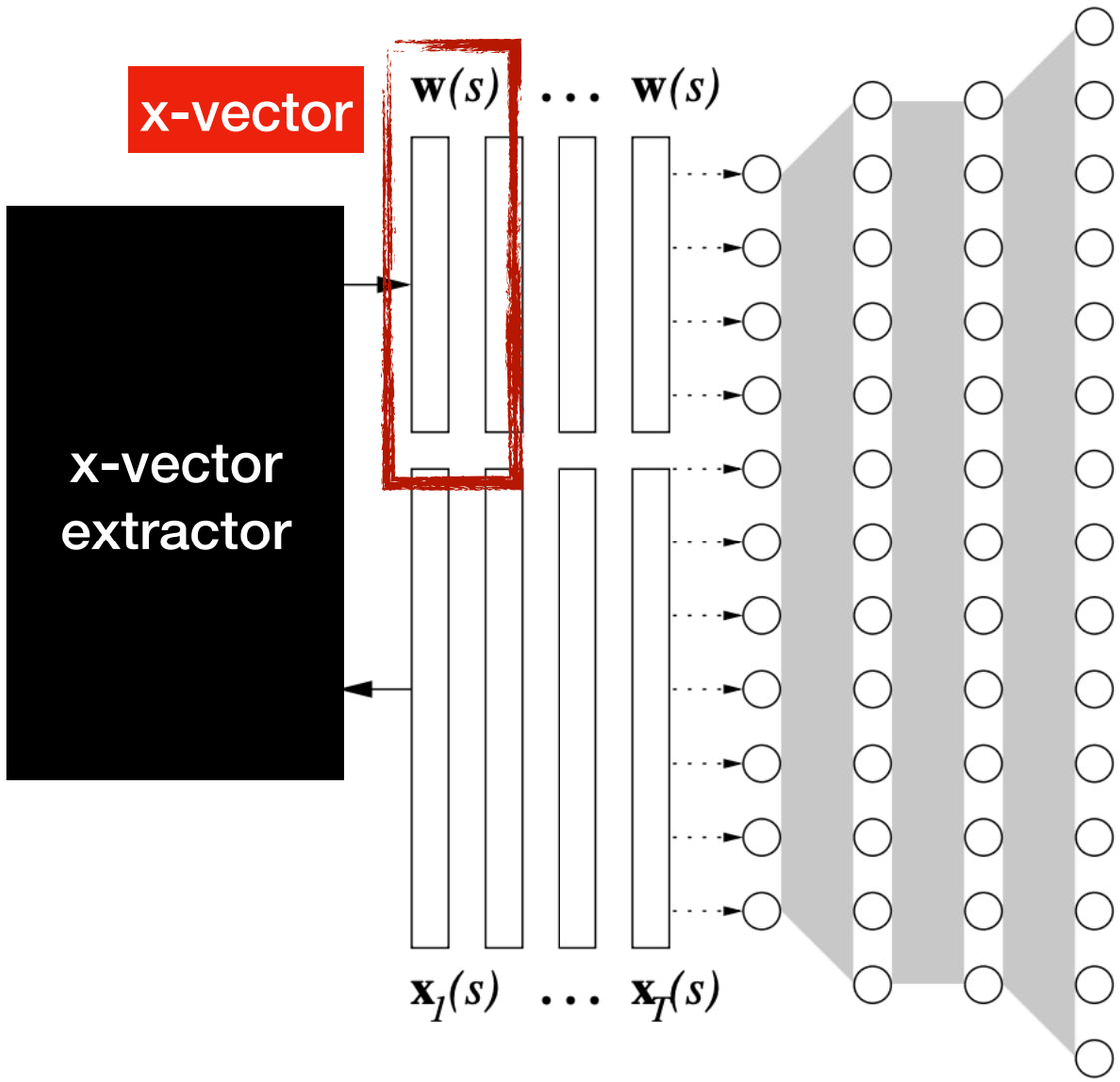
draj@cs.jhu.edu, {david.ryan.snyder, dpovey}@gmail.com, khudanpur@jhu.edu

But also contain noise information if trained without augmentation



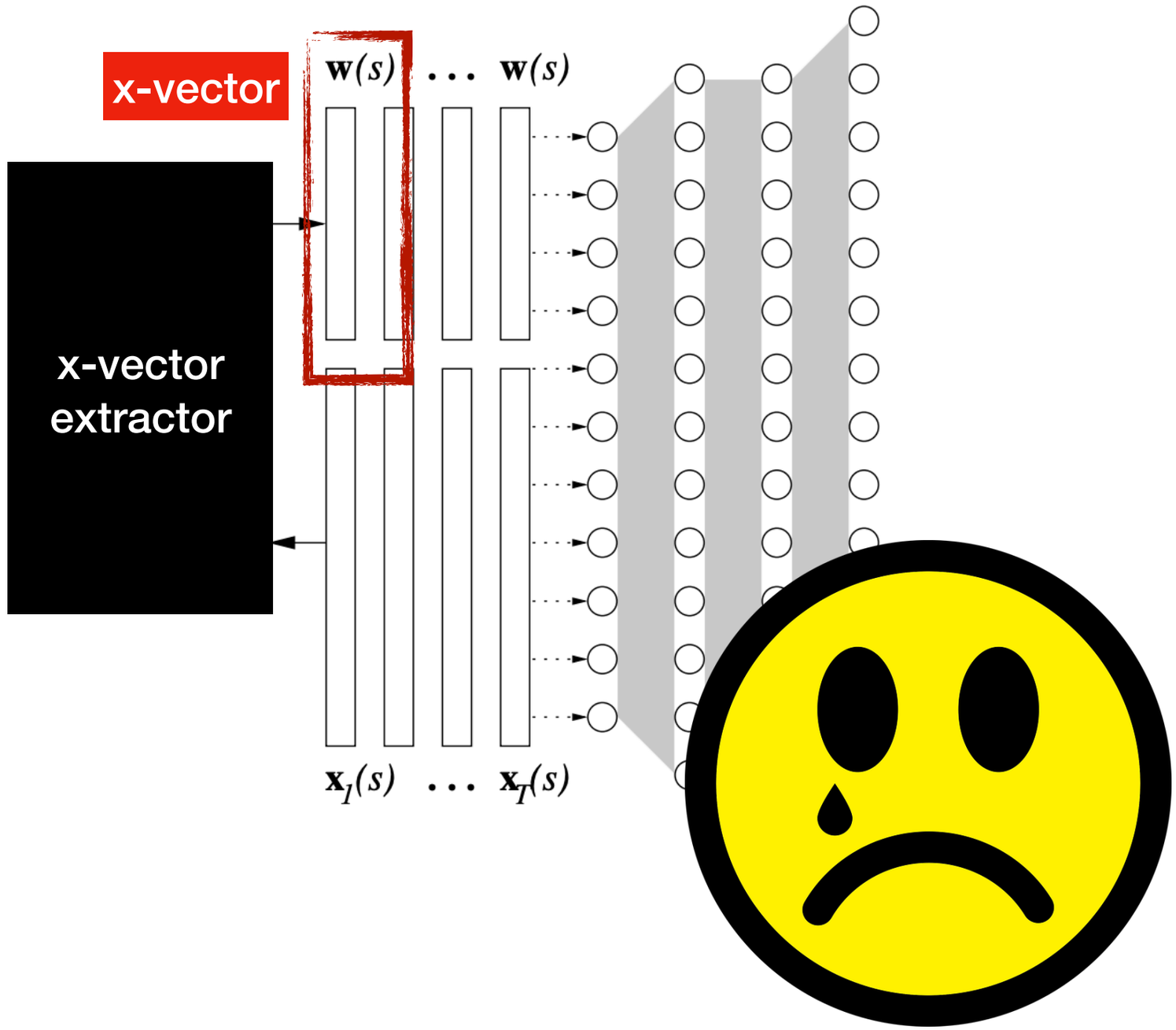
Noise-aware Training of Acoustic Models

Using x-vectors



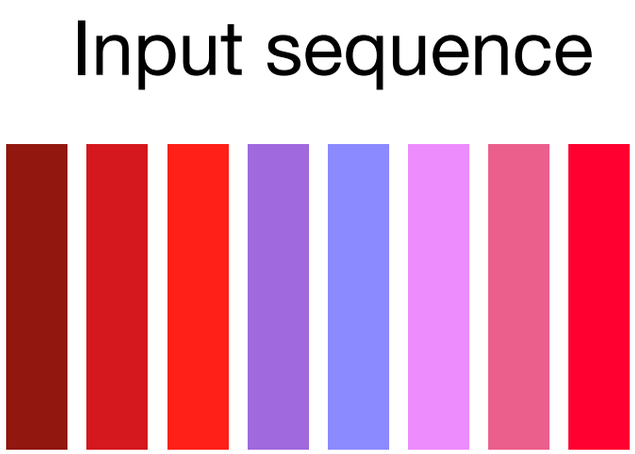
Noise-aware Training of Acoustic Models

Using x-vectors



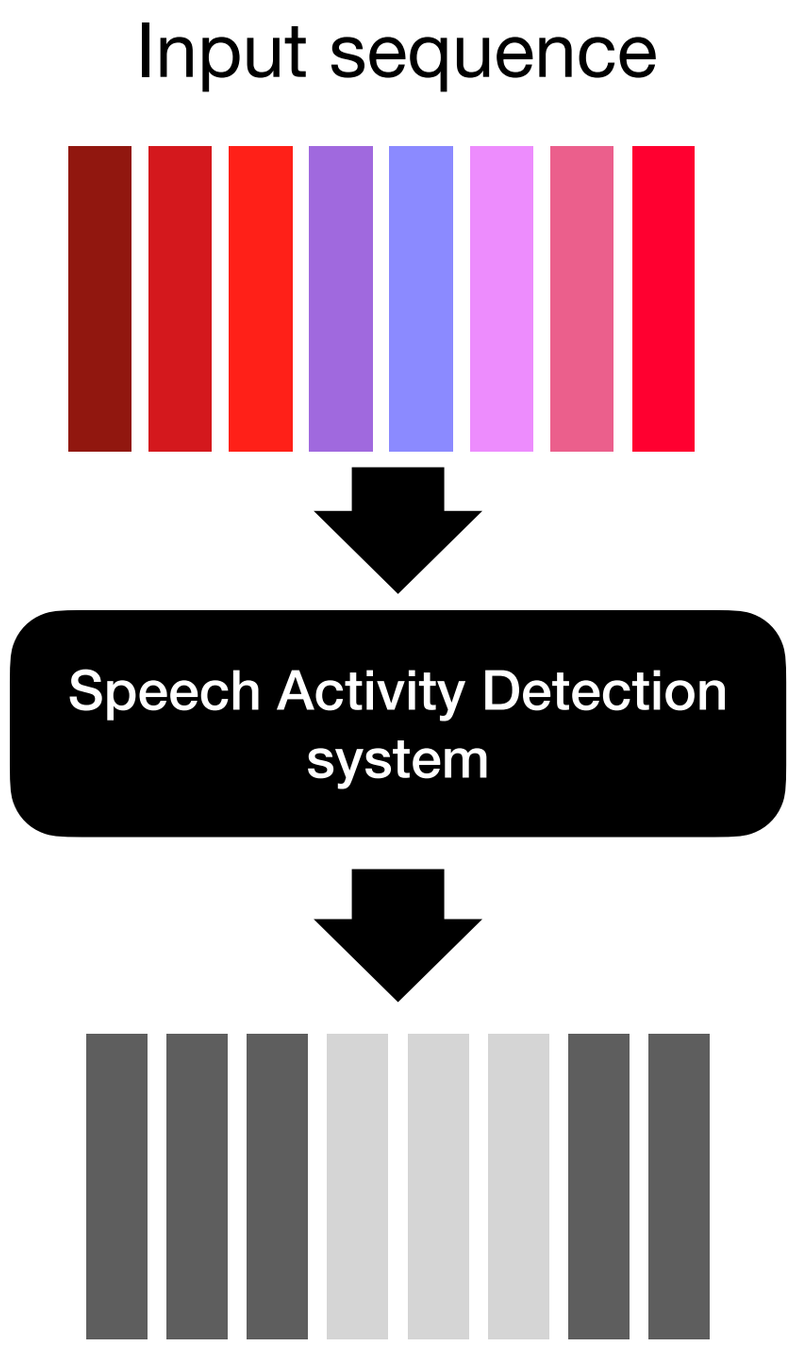
Noise-aware Training of Acoustic Models

Speech/silence classification



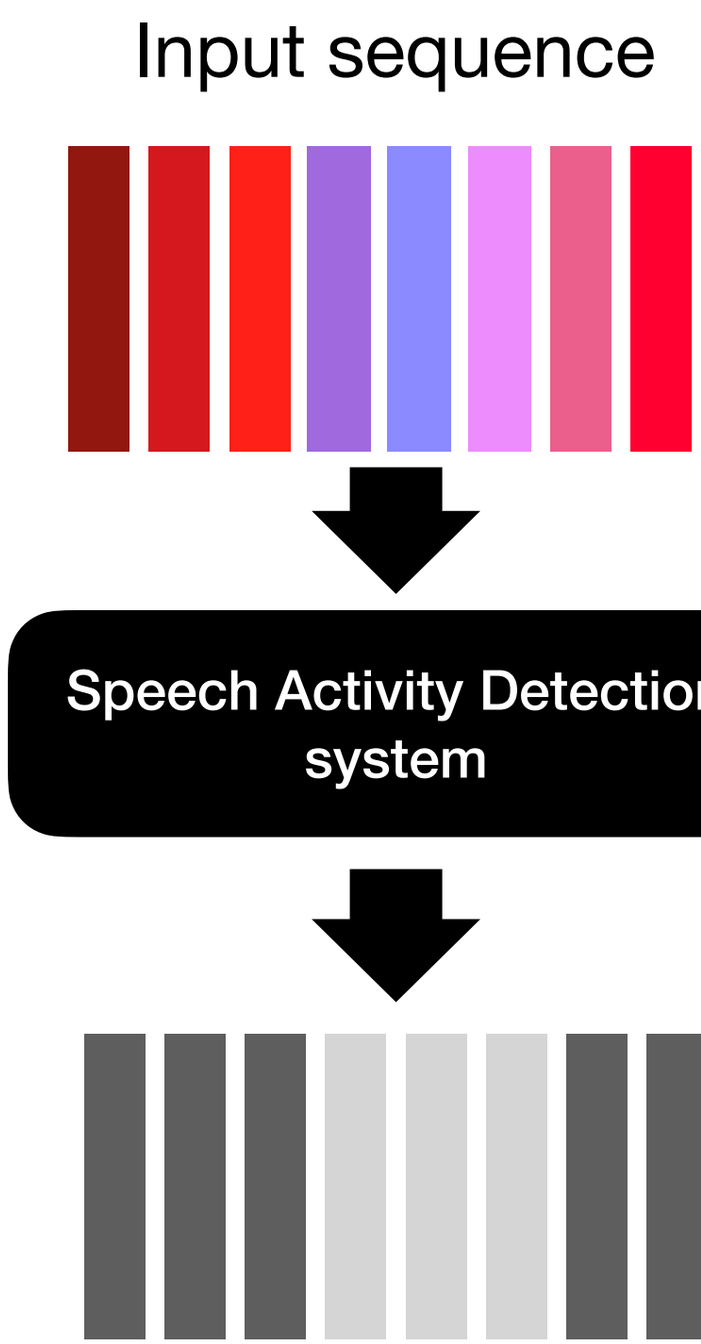
Noise-aware Training of Acoustic Models

Speech/silence classification



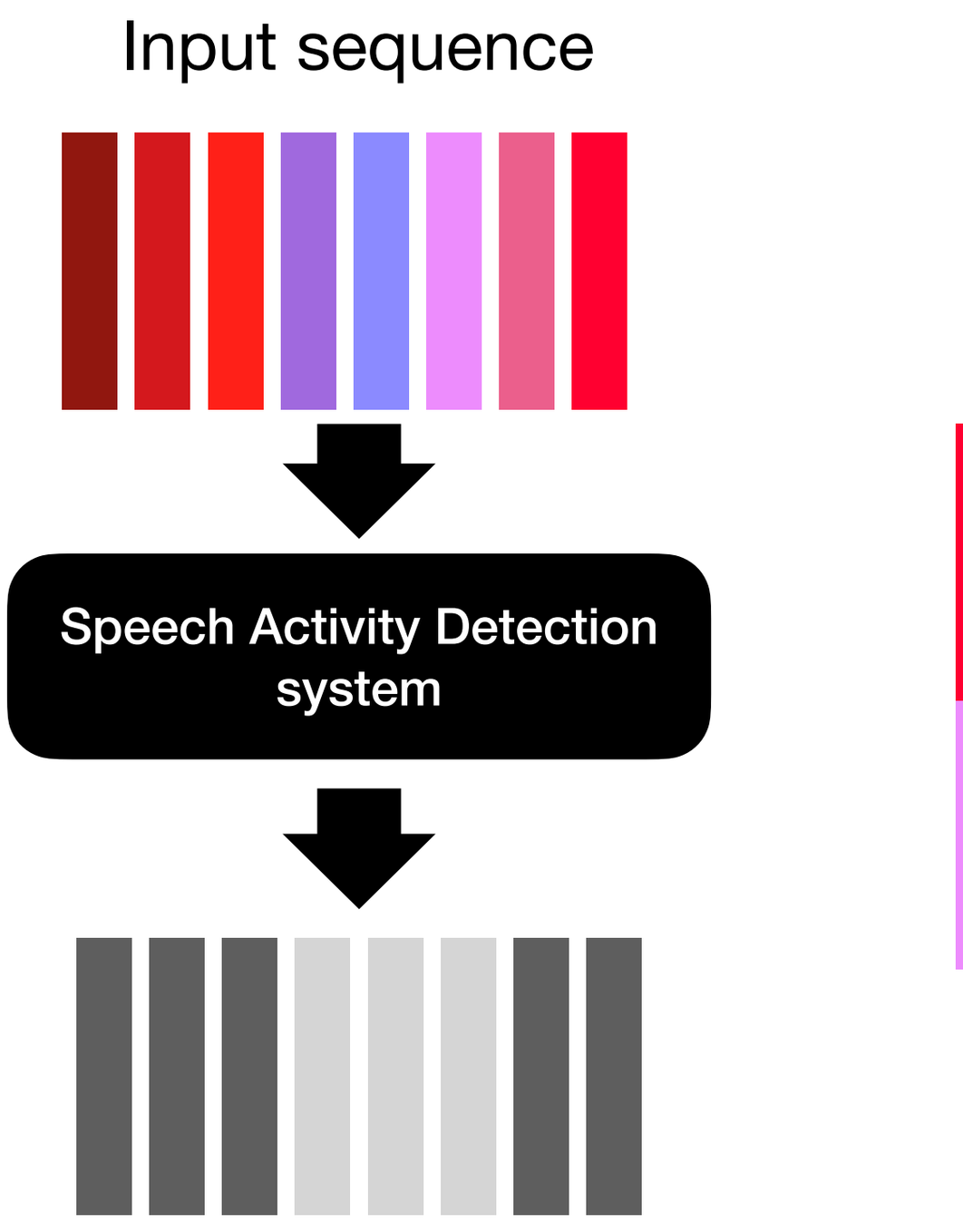
Noise-aware Training of Acoustic Models

Speech/silence classification



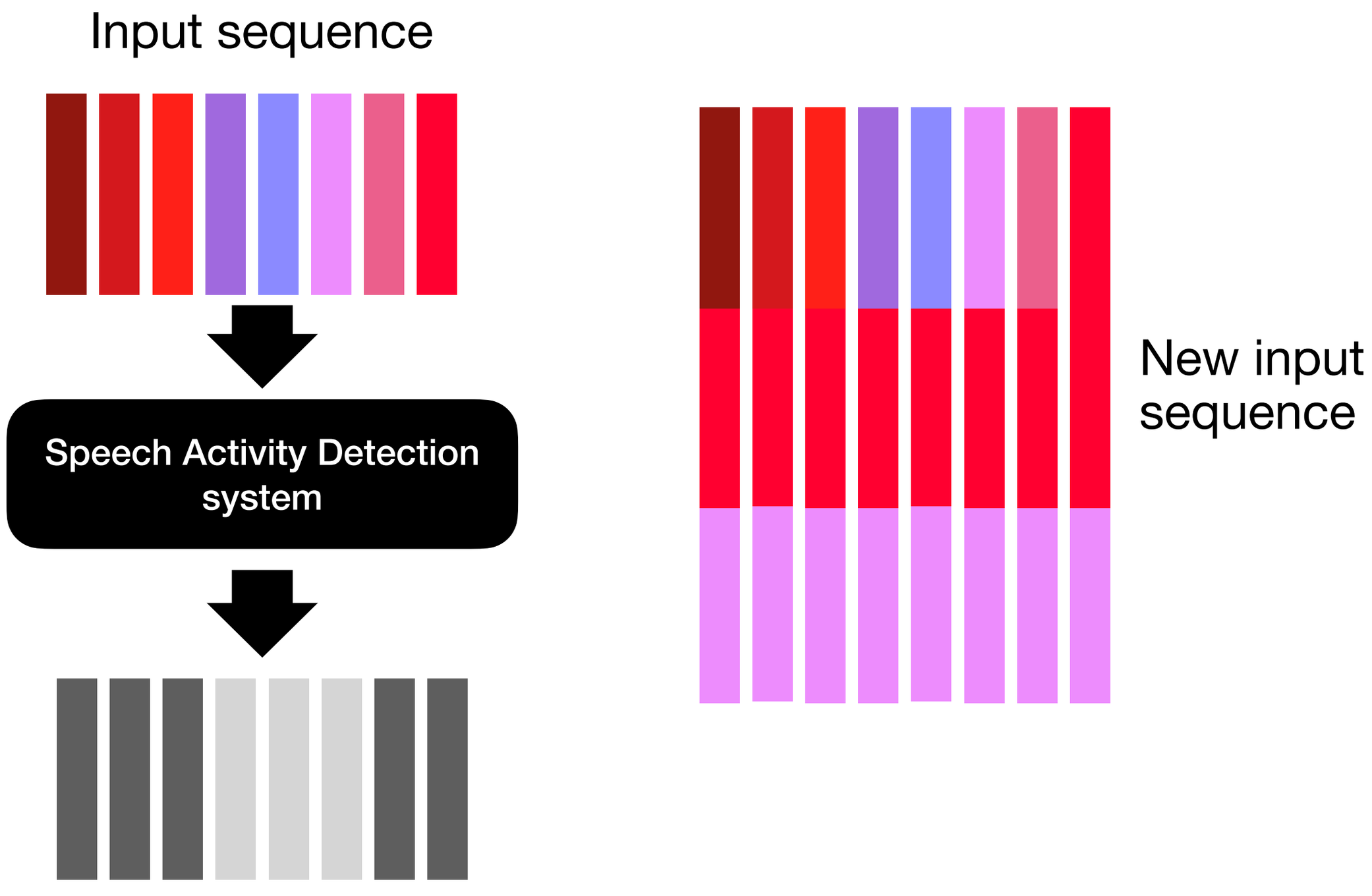
Noise-aware Training of Acoustic Models

Speech/silence classification



Noise-aware Training of Acoustic Models

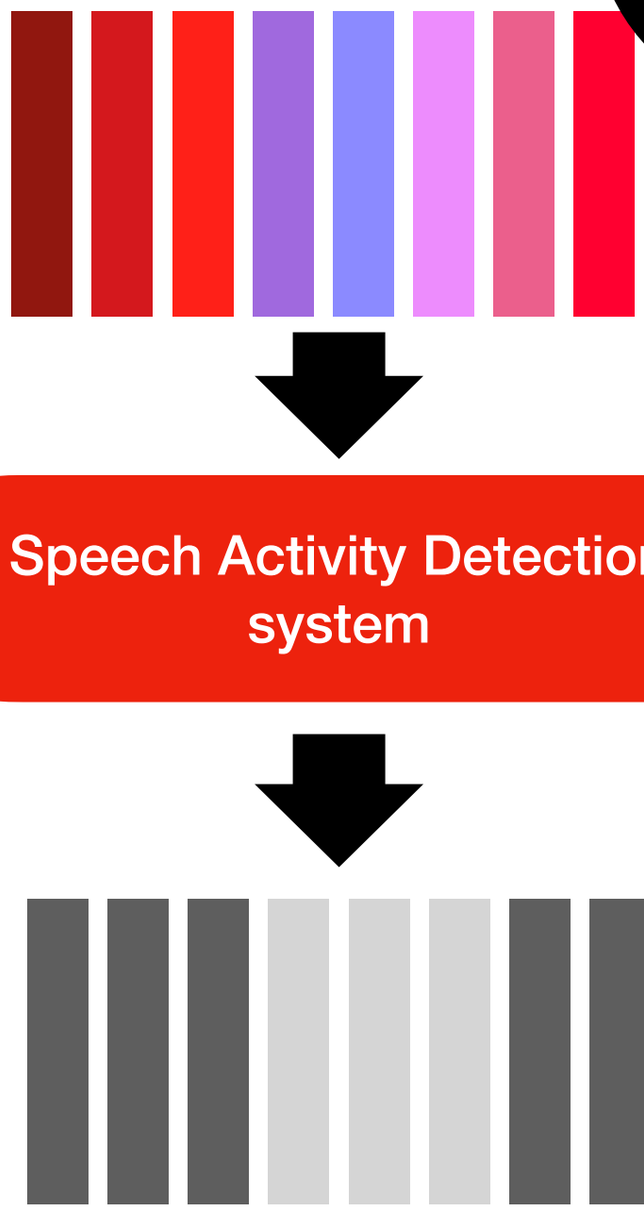
Speech/silence classification



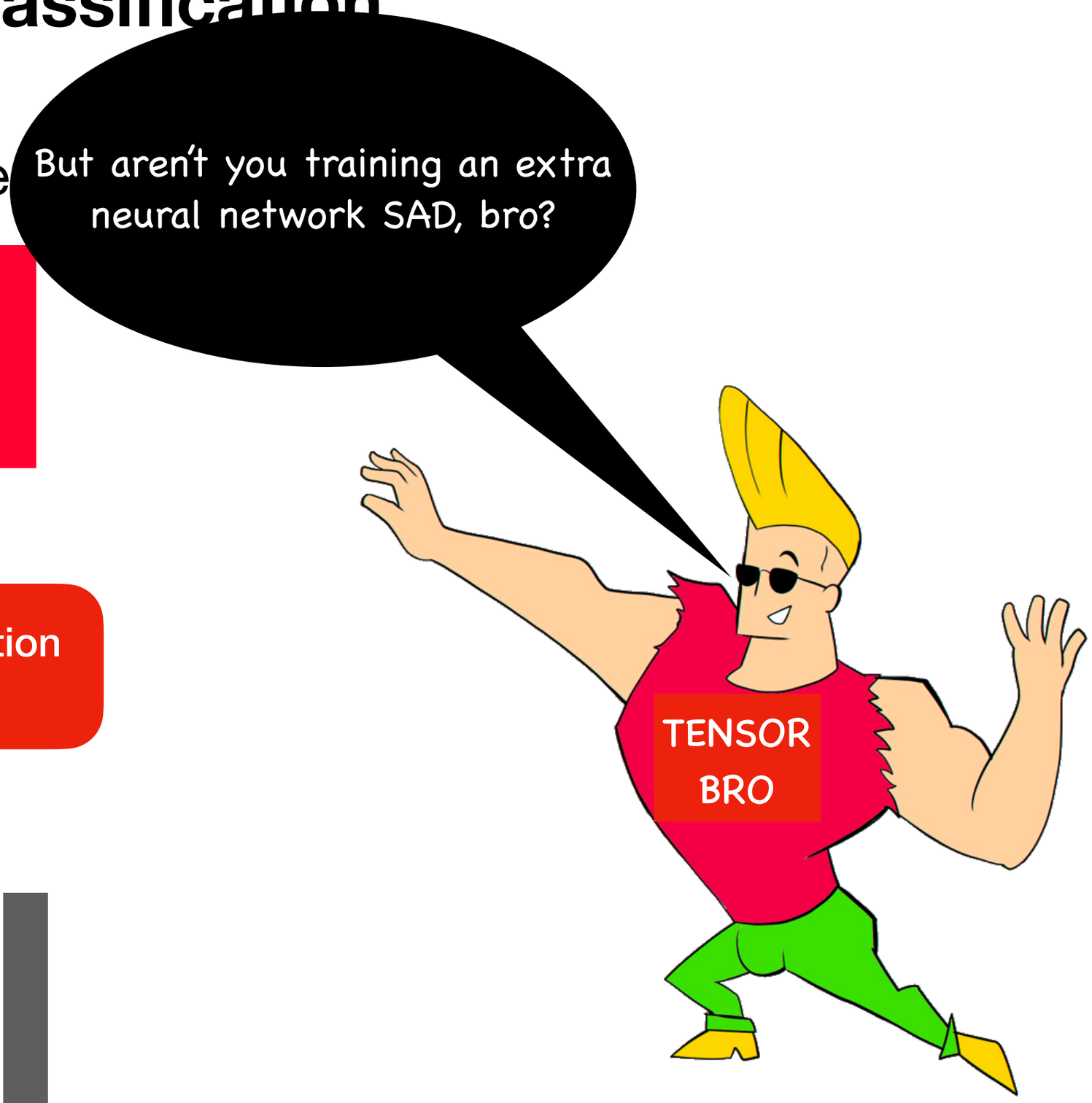
Noise-aware Training of Acoustic Models

Speech/silence classification

Input sequence



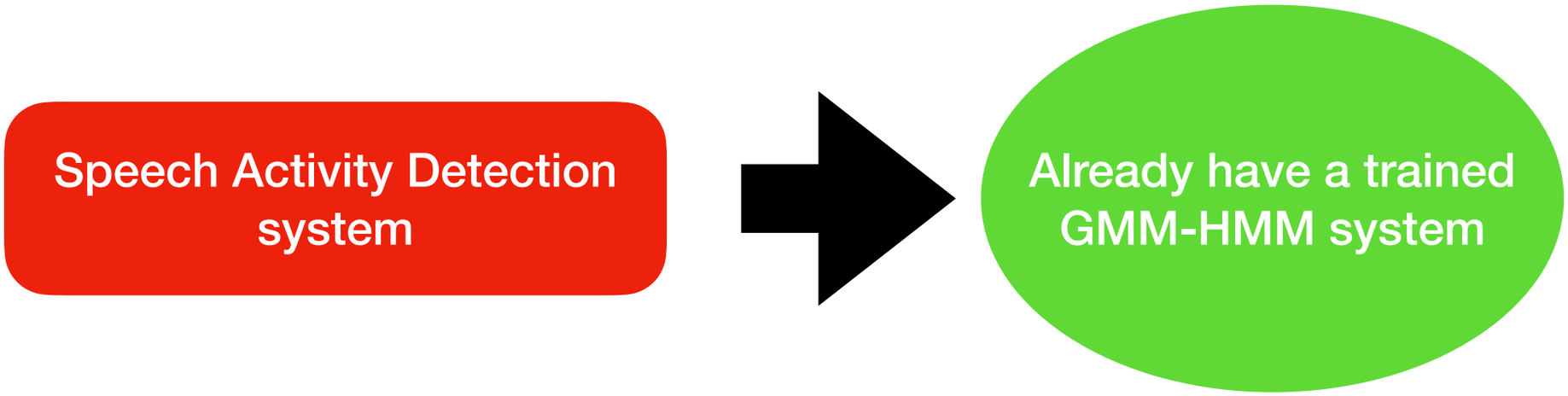
Speech Activity Detection system



Noise-aware Training of Acoustic Models

Speech/silence classification

No extra training needed!



Noise-aware Training of Acoustic Models

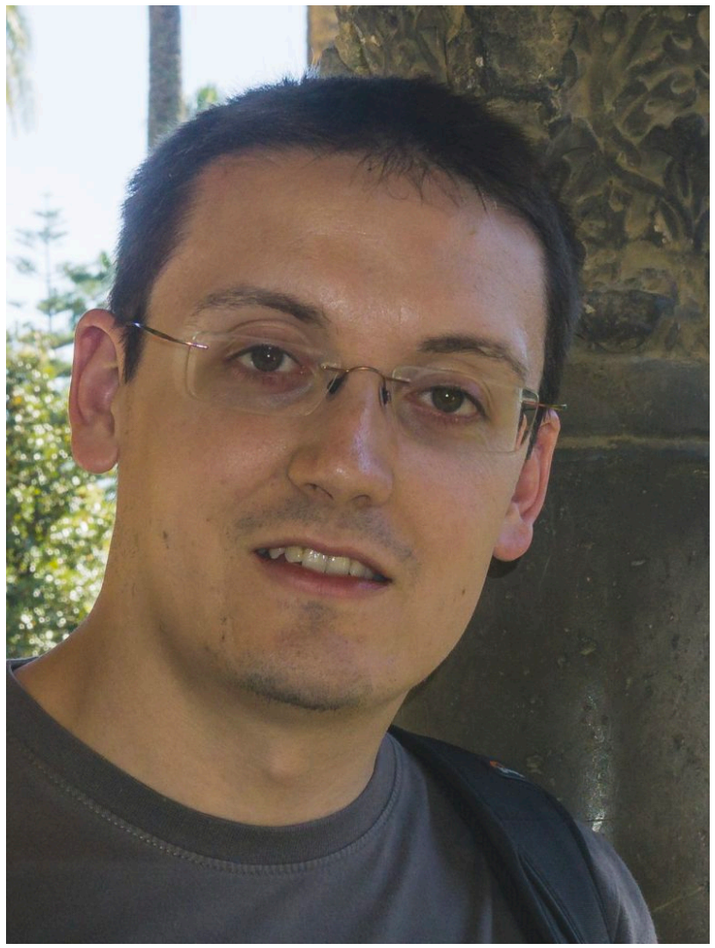
Extension to online ASR

How to estimate noise embedding in streaming ASR?

Noise-aware Training of Acoustic Models

Extension to online ASR

How to estimate noise embedding in streaming ASR?



Bayesian Model

$$\boldsymbol{\mu}_i = \begin{bmatrix} \boldsymbol{\mu}_{\mathrm{s}_i} \\ \boldsymbol{\mu}_{\mathrm{n}_i} \end{bmatrix}$$

$$\mathbf{K}_i = \begin{bmatrix} (1 + r_{\mathrm{s}_i} N_{\mathrm{s}_i}) \boldsymbol{\Lambda}_{\mathrm{s}} & -\boldsymbol{\Lambda}_{\mathrm{s}} \mathbf{B} \\ -\mathbf{B}^{\mathrm{T}} \boldsymbol{\Lambda}_{\mathrm{s}} & (1 + r_{\mathrm{n}_i} N_{\mathrm{n}_i}) \boldsymbol{\Lambda}_{\mathrm{n}} + \mathbf{B}^{\mathrm{T}} \boldsymbol{\Lambda}_{\mathrm{s}} \mathbf{B} \end{bmatrix}$$

$$\mathbf{Q}_i = \begin{bmatrix} \boldsymbol{\Lambda}_{\mathrm{s}} (\mathbf{a} + r_{\mathrm{s}_i} \mathbf{F}_{\mathrm{s}_i}) \\ \boldsymbol{\Lambda}_{\mathrm{n}} (\boldsymbol{\mu}_{\mathrm{n}} + r_{\mathrm{n}_i} \mathbf{F}_{\mathrm{n}_i}) + \mathbf{B}^{\mathrm{T}} \boldsymbol{\Lambda}_{\mathrm{s}} \mathbf{a} \end{bmatrix} .$$

The CHiME-6 Challenge

The CHiME-6 Challenge

What is it?

https://chimechallenge.github.io/chime6/overview.html

The CHiME-6 Challenge

What is it?

Far-field

Noisy

Overlapping

Real conversations

https://chimechallenge.github.io/chime6/overview.html

The CHiME-6 Challenge

How to solve it?

The CHiME-6 Challenge

How to solve it?

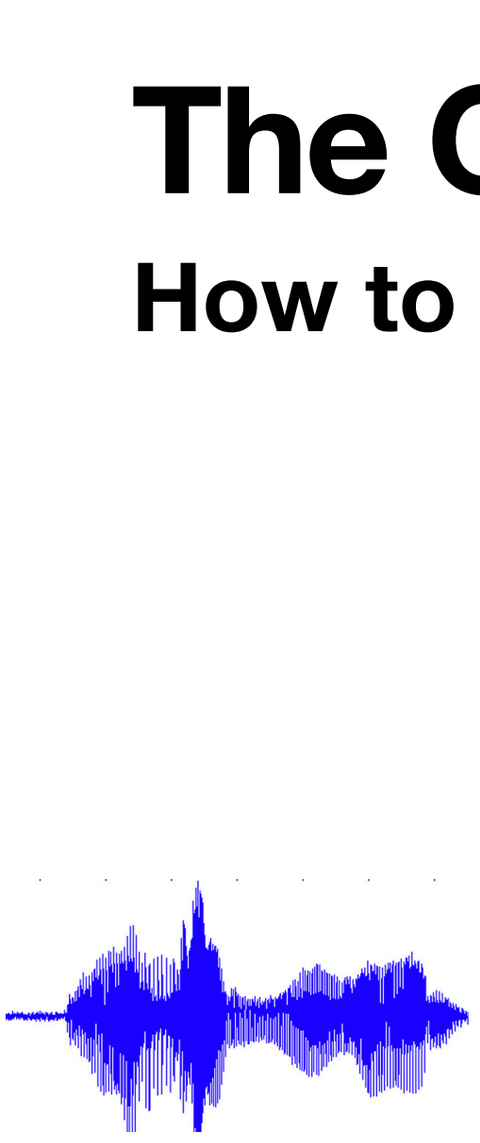
END-TO-END NEURAL NETWORK

Num. Params = Number of atoms in the universe + National debt

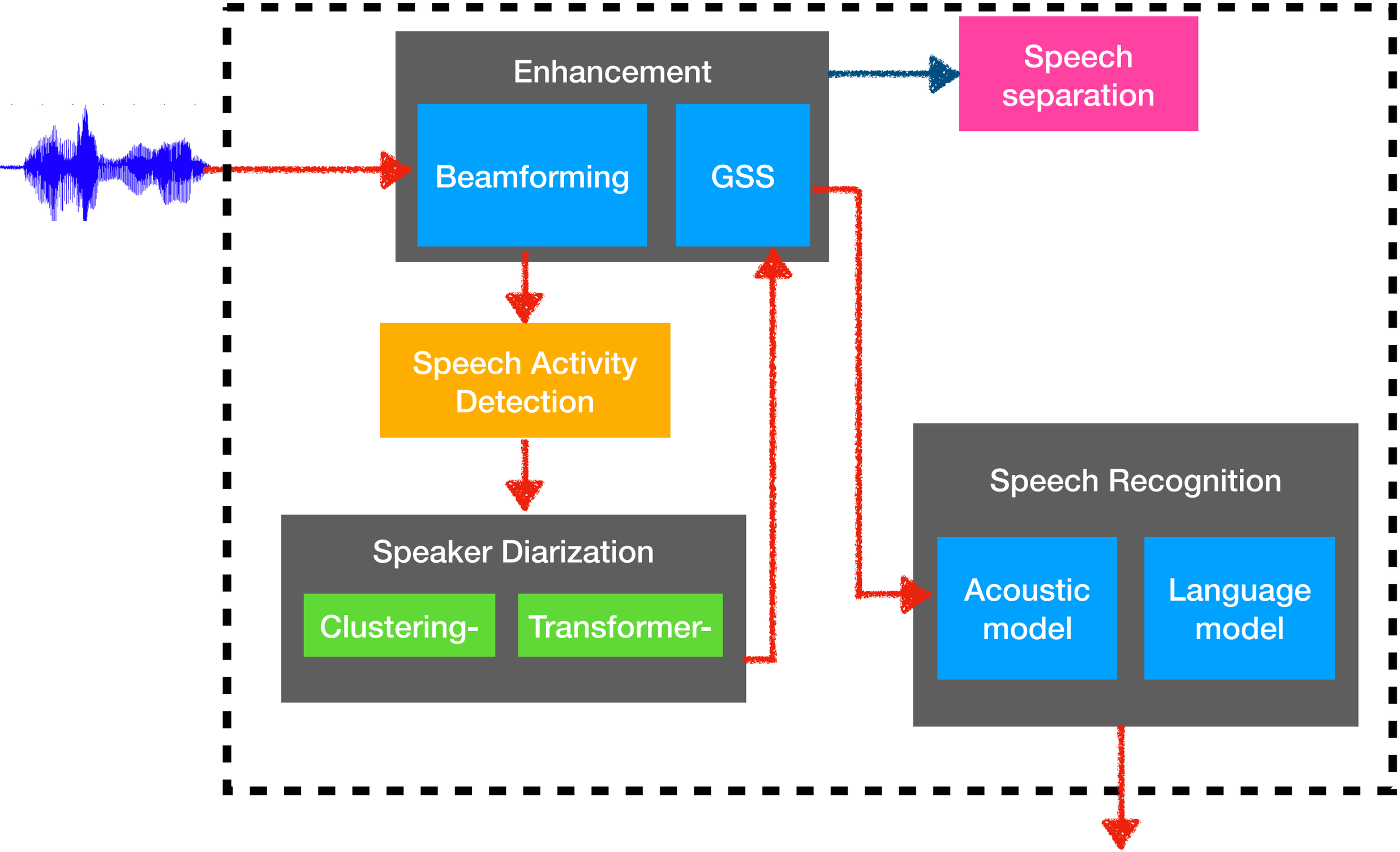
```
S01-0001845-0002138 1 0.820 0.060 we
S01-0001845-0002138 1 0.880 0.120 can
S01-0001845-0002138 1 1.000 0.110 do
S01-0001845-0002138 1 1.110 0.270 that
S01-0001845-0002138 1 1.410 0.400 after
S01-0002175-0003003 1 0.120 0.160 so
S01-0002175-0003003 1 0.280 0.340 maybe
S01-0002175-0003003 1 0.870 0.320 someone
S01-0002175-0003003 1 1.190 0.150 can
S01-0002175-0003003 1 1.340 0.180 help
S01-0002175-0003003 1 1.660 0.190 cook
```

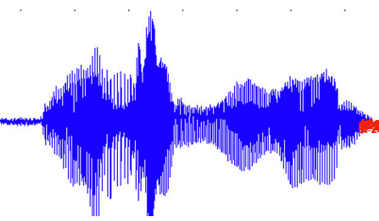
The CHiME-6 Challenge

How to solve it?




```
S01-0001845-0002138 1 0.820 0.060 we
S01-0001845-0002138 1 0.880 0.120 can
S01-0001845-0002138 1 1.000 0.110 do
S01-0001845-0002138 1 1.110 0.270 that
S01-0001845-0002138 1 1.410 0.400 after
S01-0002175-0003003 1 0.120 0.160 so
S01-0002175-0003003 1 0.280 0.340 maybe
S01-0002175-0003003 1 0.870 0.320 someone
S01-0002175-0003003 1 1.190 0.150 can
S01-0002175-0003003 1 1.340 0.180 help
S01-0002175-0003003 1 1.660 0.190 cook
```


Enhancement
Beamforming
GSS
Speech separation
Speech Activity Detection
Speaker Diarization
Clustering-
Transformer-
Speech Recognition
Acoustic model
Language model
S01-0001845-0002138 1 1.410 0.400 after
S01-0002175-0003003 1 0.120 0.160 so
S01-0002175-0003003 1 0.280 0.340 maybe
S01-0002175-0003003 1 0.870 0.320 someone
S01-0002175-0003003 1 1.190 0.150 can
S01-0002175-0003003 1 1.340 0.180 help
S01-0002175-0003003 1 1.660 0.190 cook



Enhancement

Beamforming

GSS

Speech separation

Speech Activity Detection

Speaker Diarization

Clustering-

Transformer-

Speech Recognition

Acoustic model

Language model

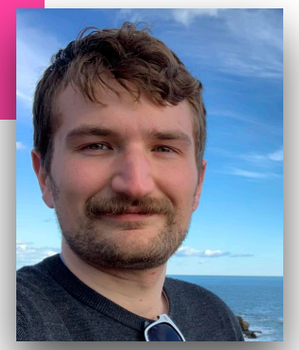
```
S01-0001845-0002138 1 1.410 0.400 after
S01-0002175-0003003 1 0.120 0.160 so
S01-0002175-0003003 1 0.280 0.340 maybe
S01-0002175-0003003 1 0.870 0.320 someone
S01-0002175-0003003 1 1.190 0.150 can
S01-0002175-0003003 1 1.340 0.180 help
S01-0002175-0003003 1 1.660 0.190 cook
```

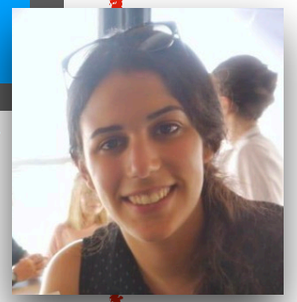
Enhancement

Beamforming

GSS

Speech separation





Speech Activity Detection

Speaker Diarization

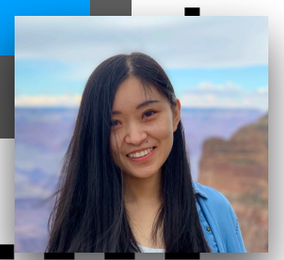
Clustering-

Transformer-

Speech Recognition

Acoustic model

Language model



```
S01-0001845-0002138 1 1.410 0.400 after
S01-0002175-0003003 1 0.120 0.160 so
S01-0002175-0003003 1 0.280 0.340 maybe
S01-0002175-0003003 1 0.870 0.320 someone
S01-0002175-0003003 1 1.190 0.150 can
S01-0002175-0003003 1 1.340 0.180 help
S01-0002175-0003003 1 1.660 0.190 cook
```

Desh, Paola, Aswin, Bar, Matt, Piotr, Ashish, Ke, Sanjeev, Shinji

Talk to me about…

(Inspired from Suzanna's talk)