A Realism Objective for Speech Denoising with Deep Learning

Peter Plantinga

Outline

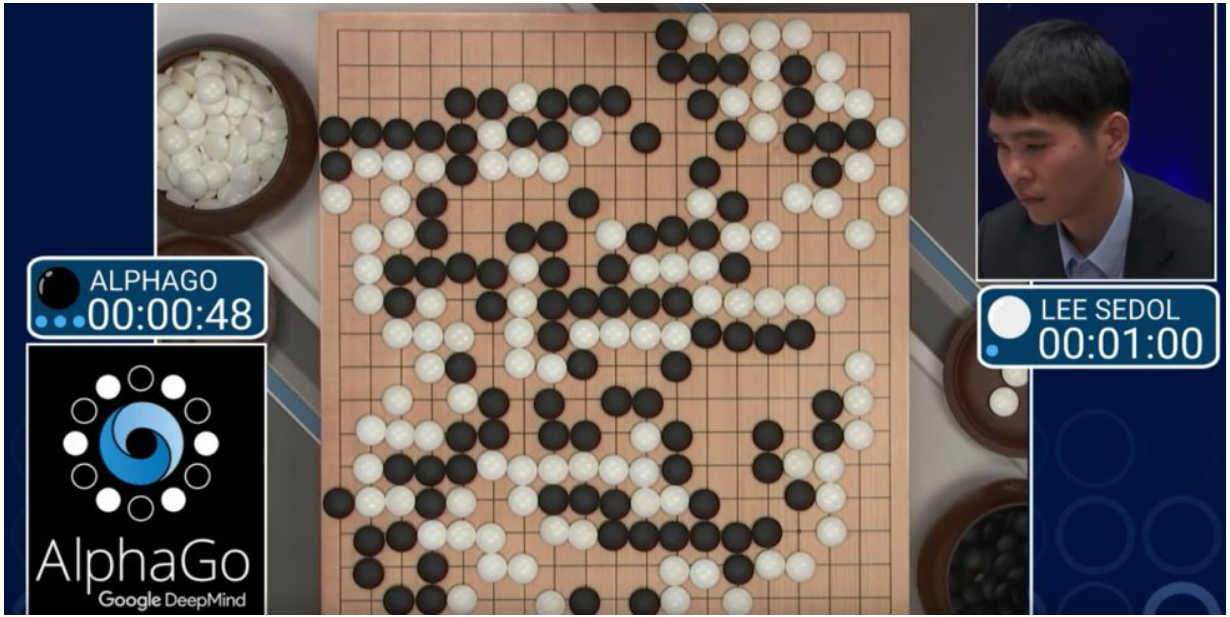
1. An introduction to deep learning
2. Adversarial networks
3. Speech denoising

Outline

1. An introduction to deep learning
2. Adversarial networks
3. Speech denoising

An Intro to Deep Learning

Lots of hype about DL. Example: AlphaGo

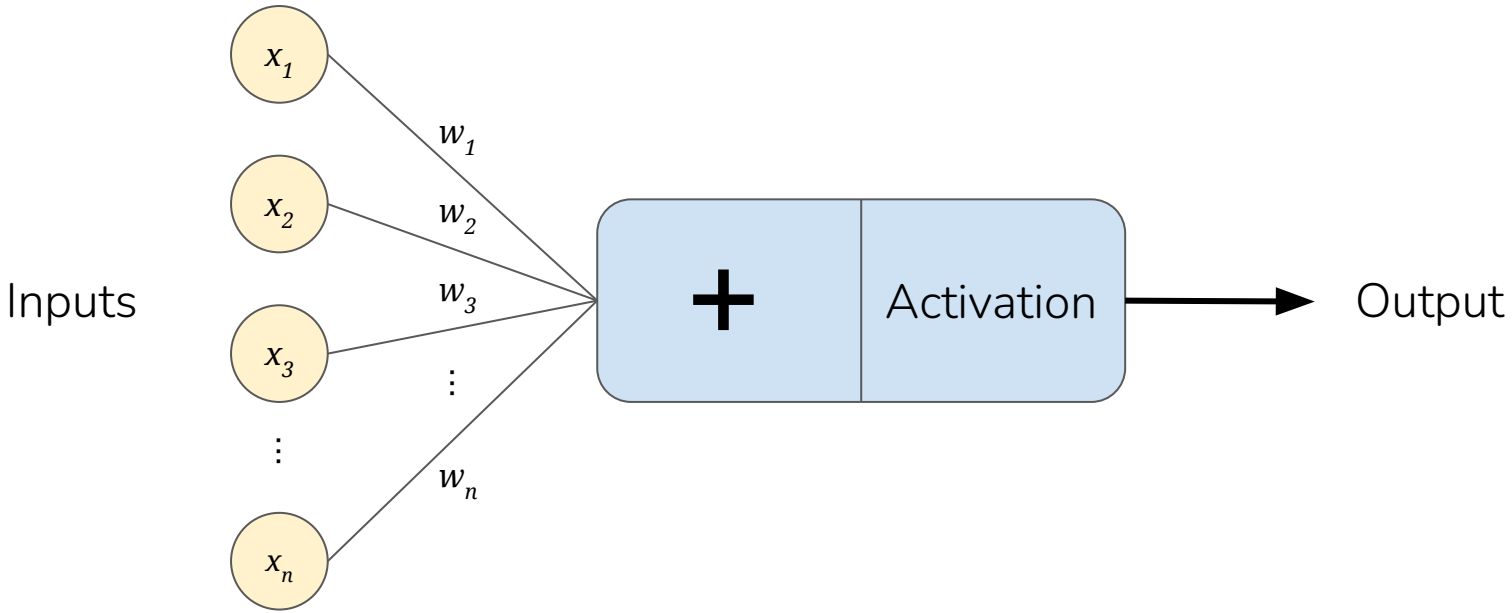


An Intro to Deep Learning

What is deep learning good for?

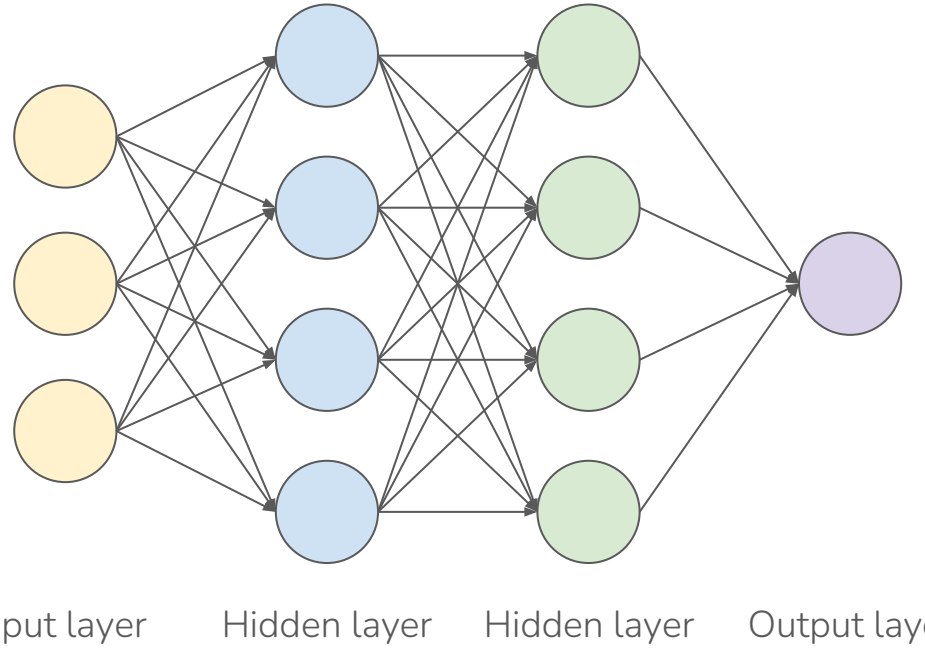
Cat

A Single Artificial Neuron



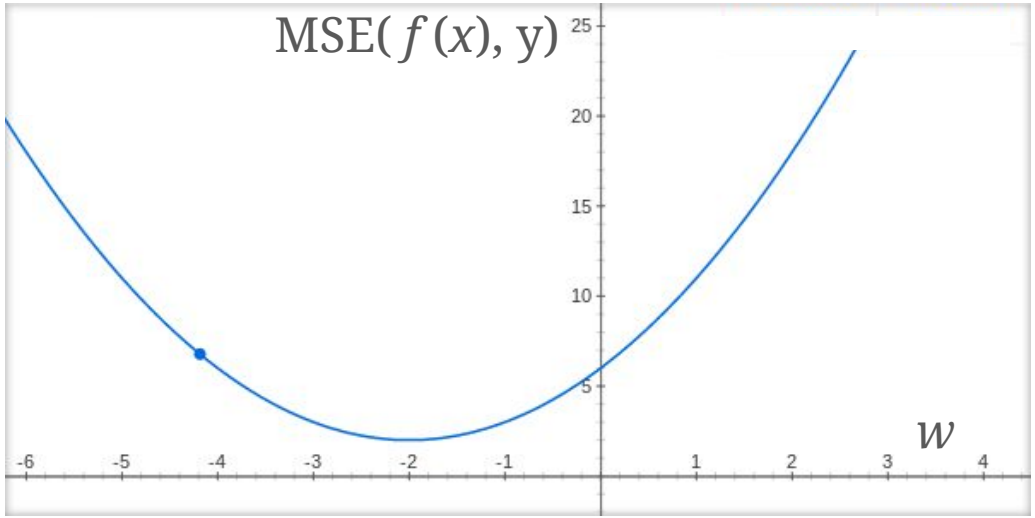
Training a DL Model

1. Design the network architecture
2. Pick an objective
3. Update parameters to improve objective



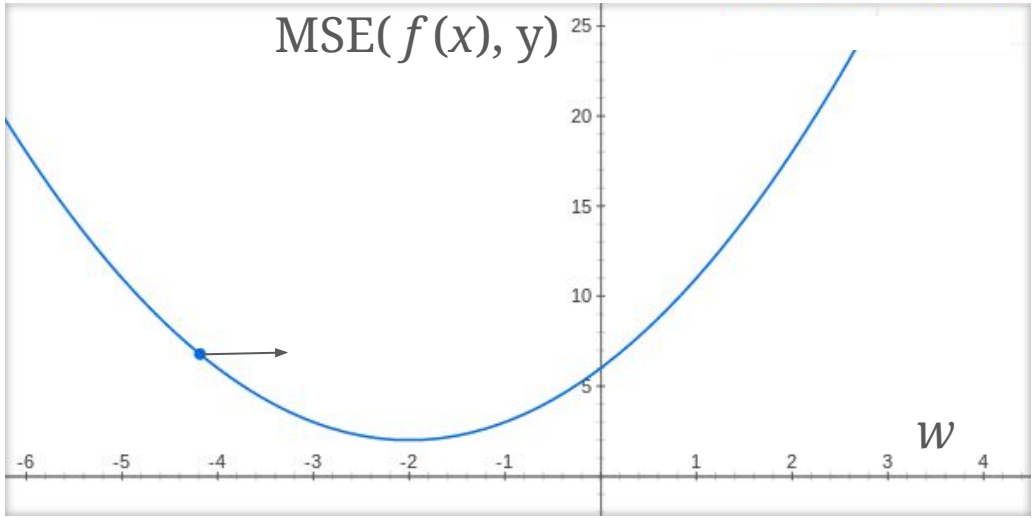
Training a DL Model

1. Design the network architecture
2. Pick an objective
3. Update parameters to improve objective



Training a DL Model

1. Design the network architecture
2. Pick an objective
3. Update parameters to improve objective

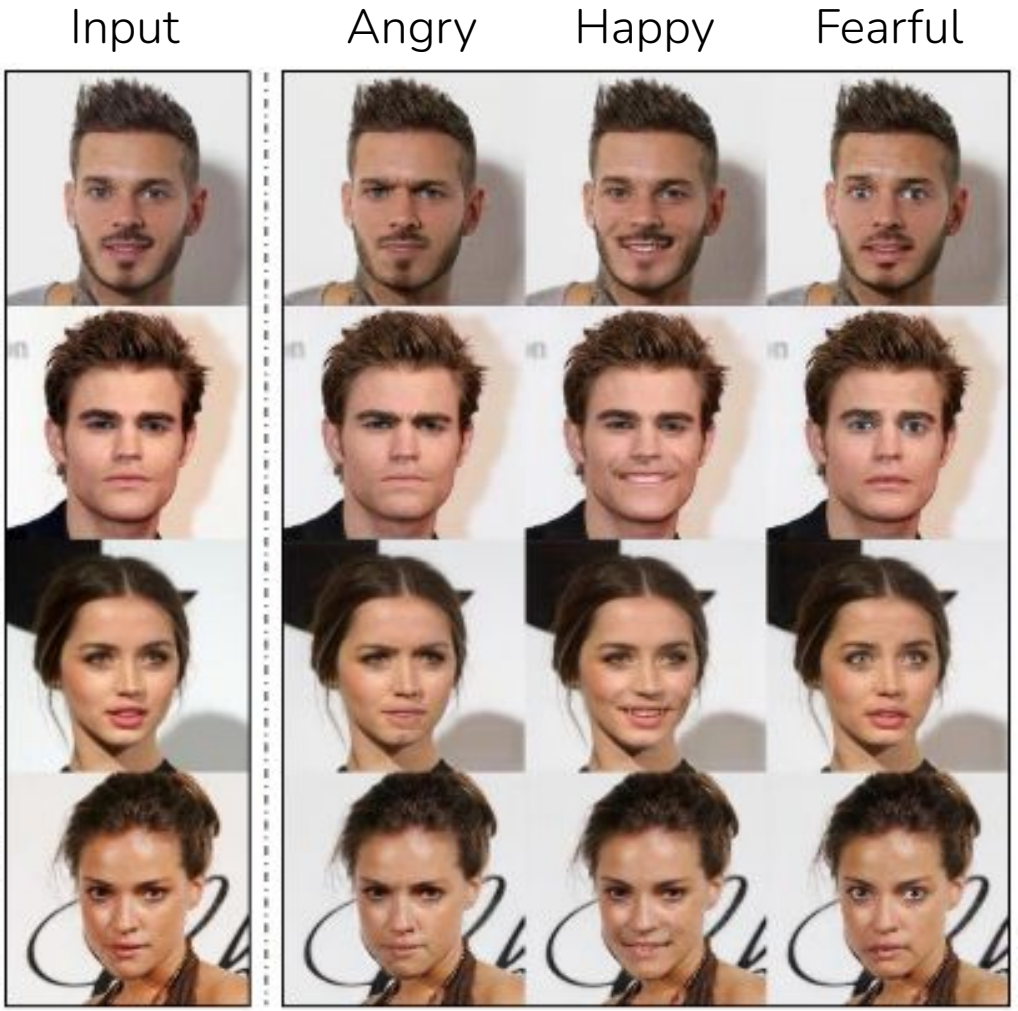


Outline

1. An introduction to deep learning
2. Adversarial networks
3. Speech denoising

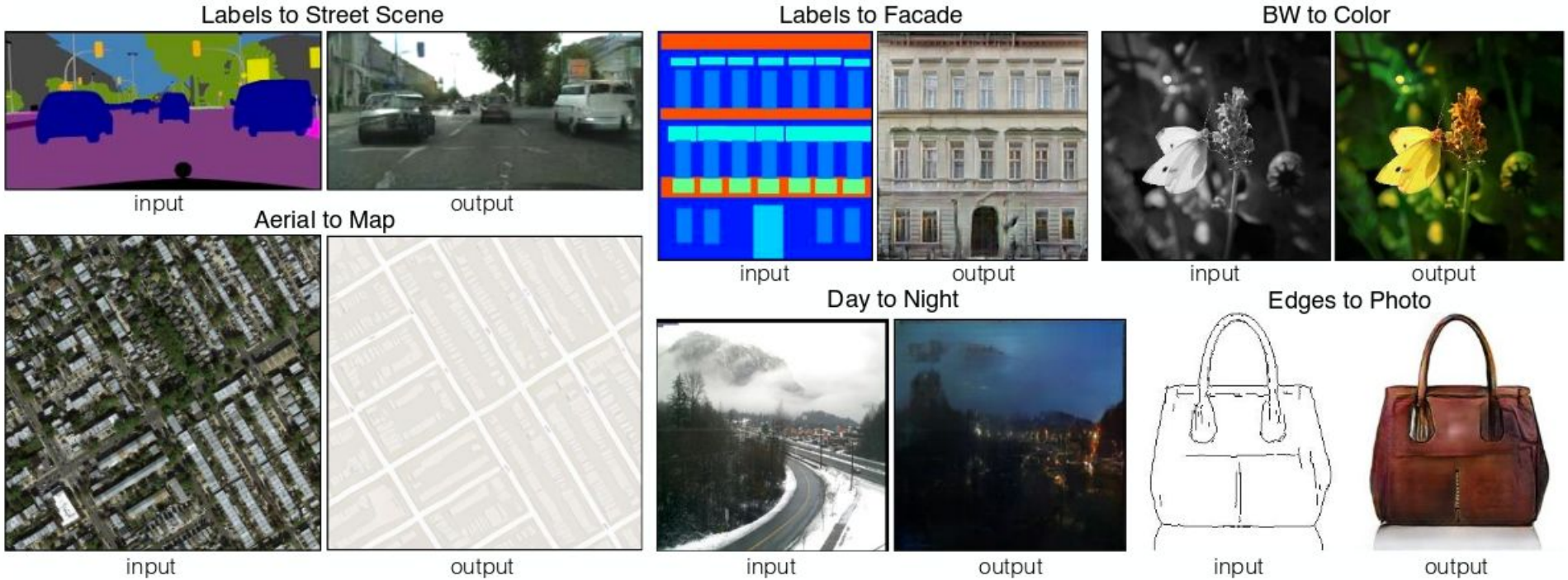
Expression Generation

Example: generate faces with a specific expression



Yunjei Choi et al. *StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation*. 2017

Image-to-Image Translation

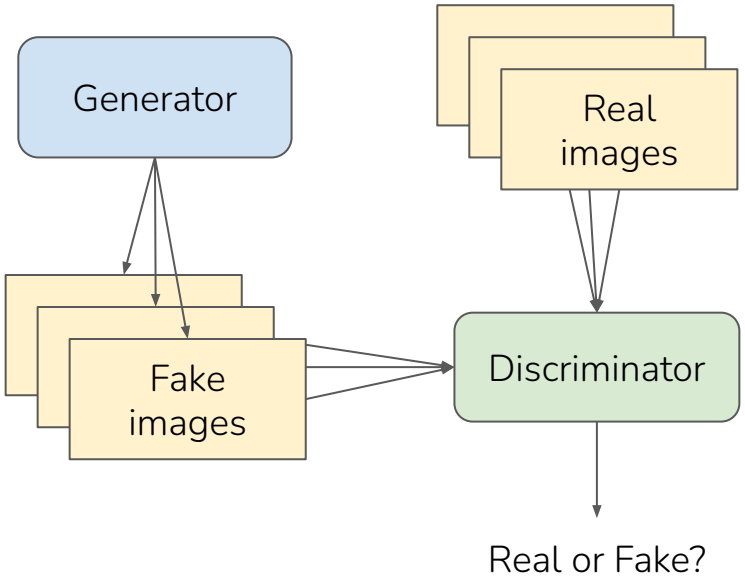


Phillip Isola et al. *Image-to-image translation with conditional adversarial networks*. CVPR 2017

Generative Adversarial Networks

GANs work by providing a *realism* objective

Is this image real or fake?



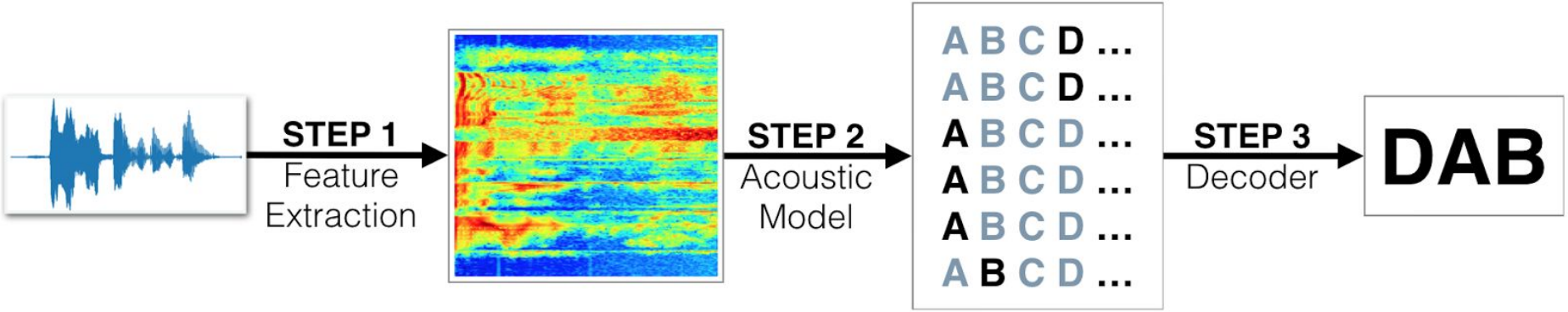
Outline

1. An introduction to deep learning
2. Adversarial networks
3. Speech denoising

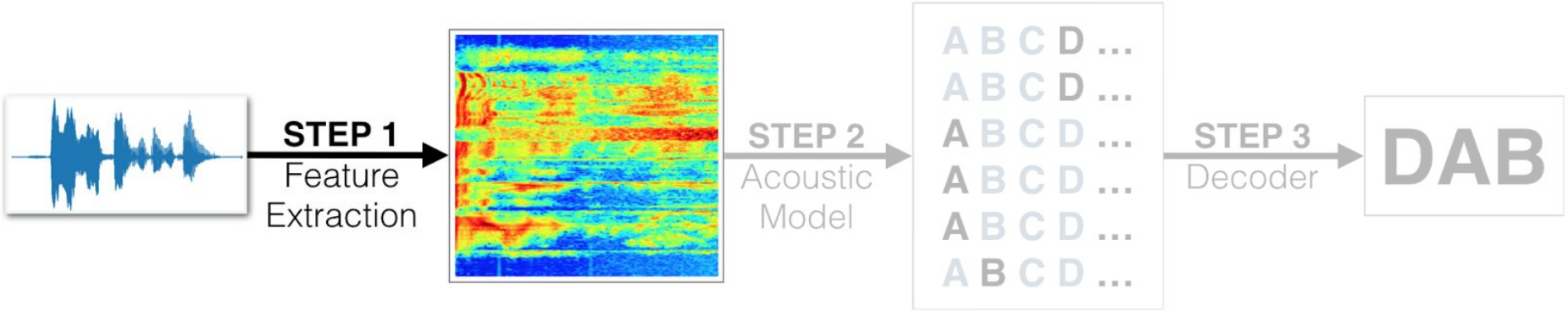
Speech Denoising Applications

- Speech recognition in noisy environments
- Hearing aids
- Teleconferencing
- Automatic captioning
- etc.

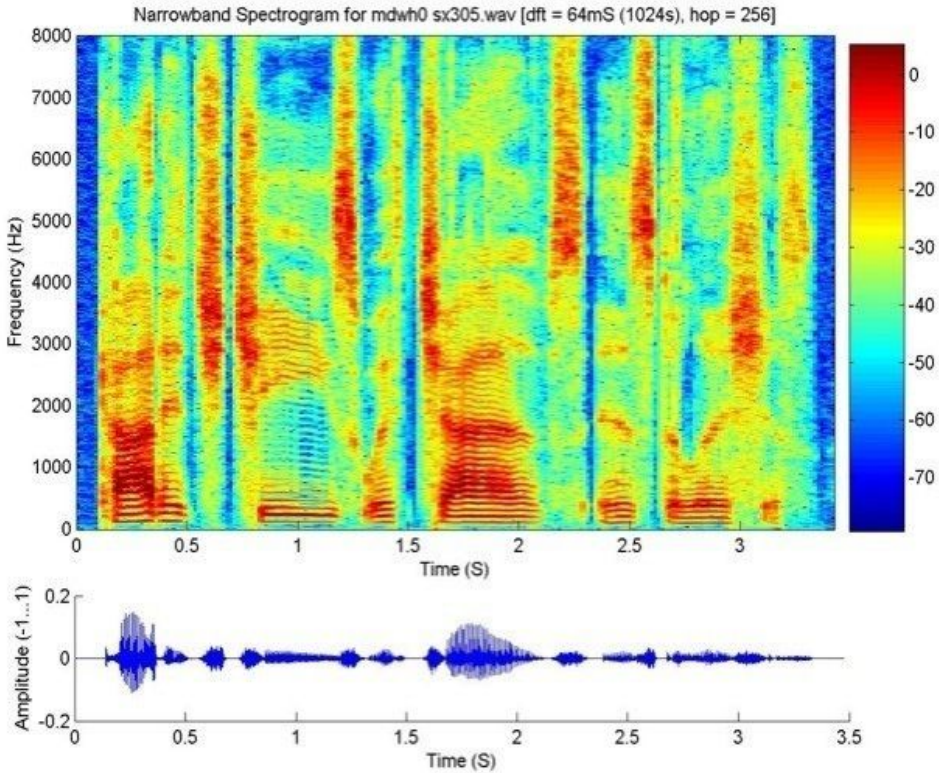
Traditional ASR Pipeline



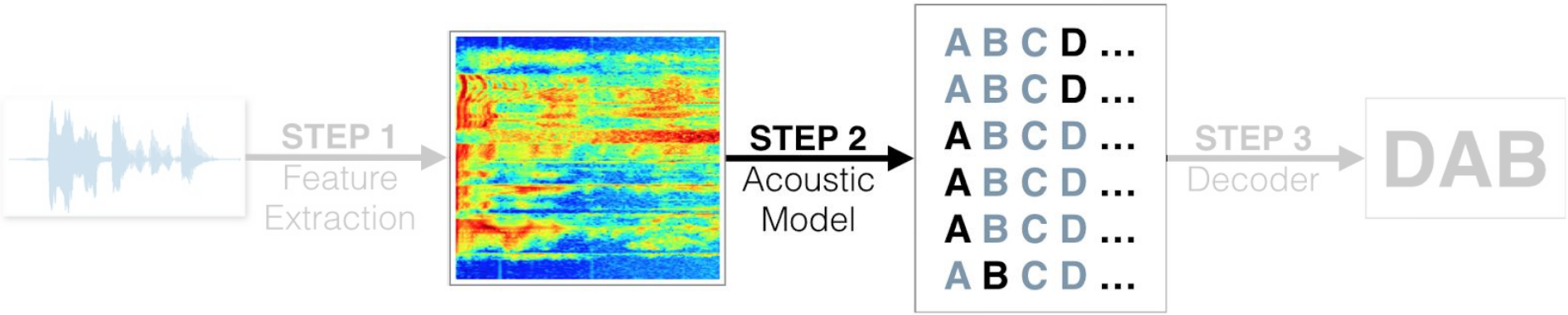
Traditional ASR Pipeline



Feature Extraction

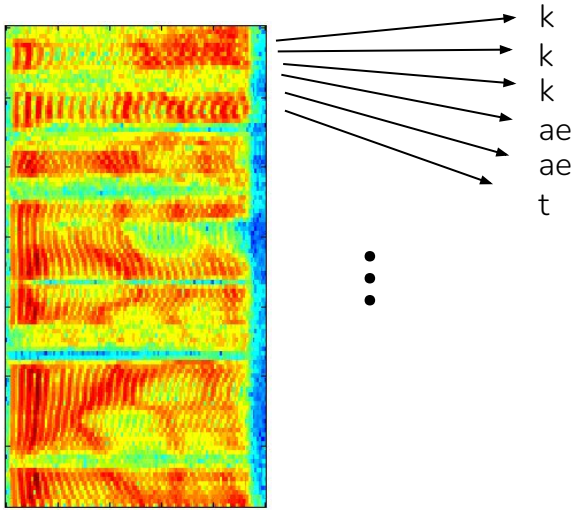


Traditional ASR Pipeline

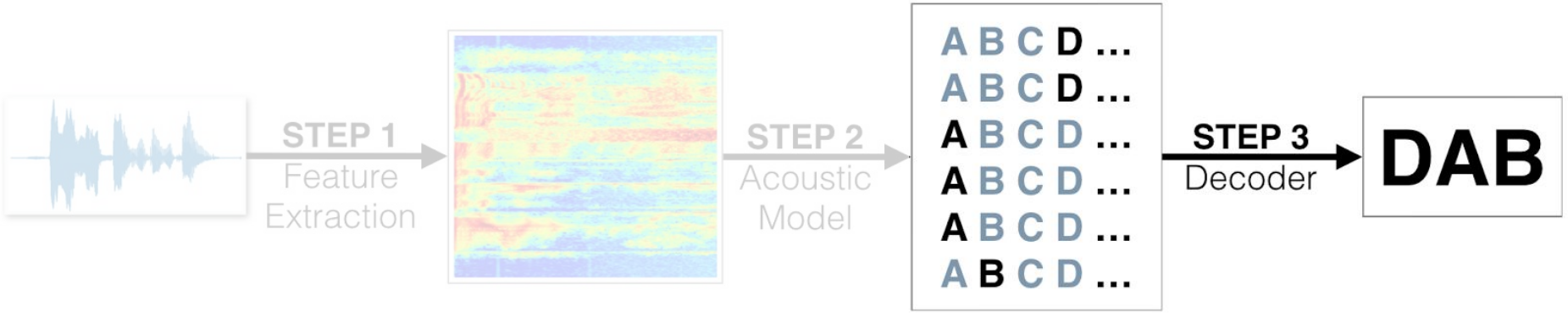


Acoustic Model

Map from spectrogram frames to phonemes



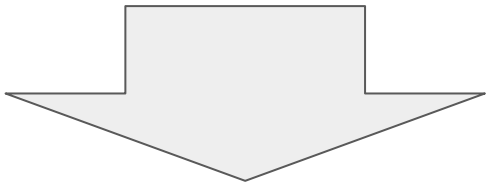
Traditional ASR Pipeline



Decoder

Map from phonemes to words

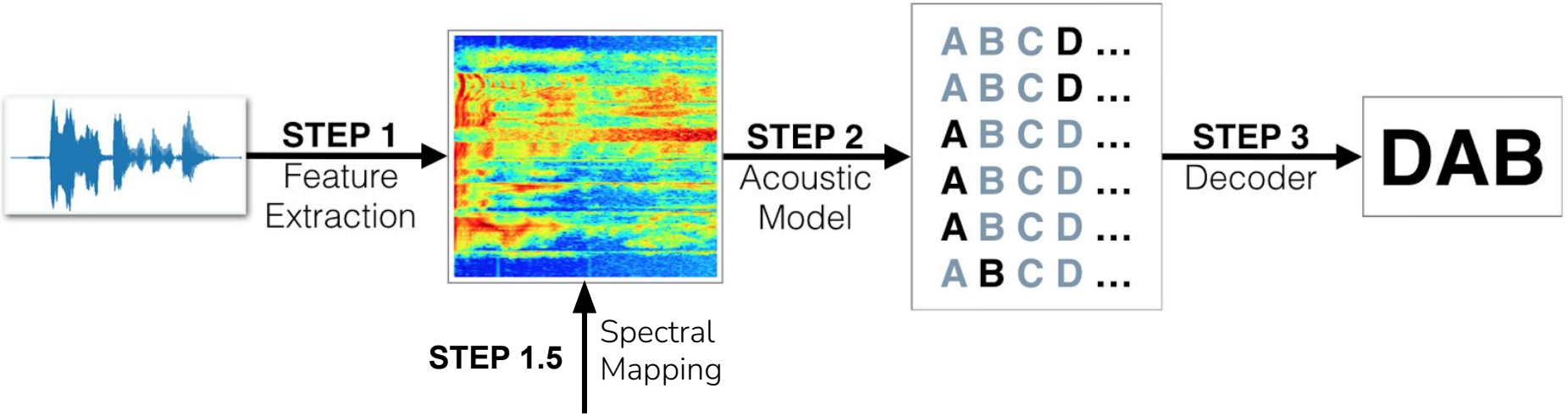
k,k,k,k,ae,ae,ae,ae,ae,ae,t,t,t,ih,ih,ih,ih,n,n,n,n,th,th,th,uh,uh,uh,h,h,ae,ae,t,t,t,t



Cat in the hat

Speech Denoising

Add a step to ASR pipeline, to clean up features

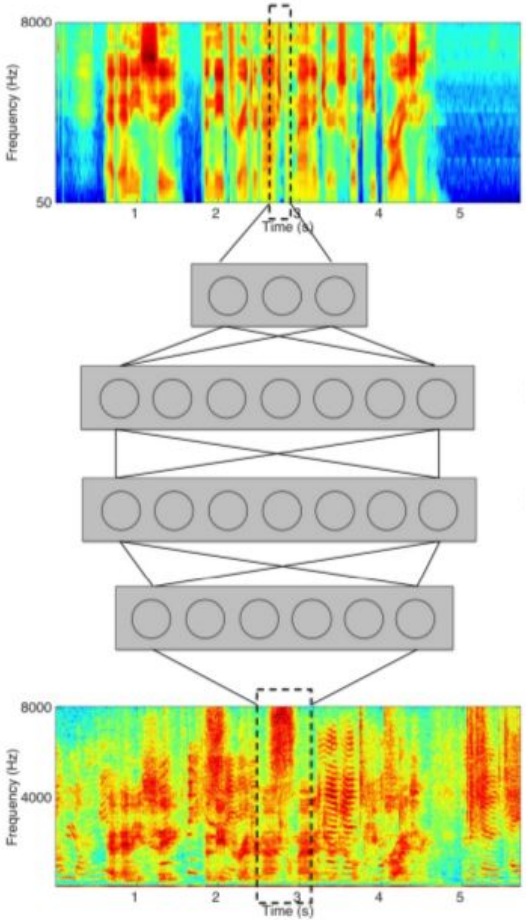


Spectral Mapping

Inputs: clean speech segment artificially mixed with noise

Labels: clean speech segment without added noise

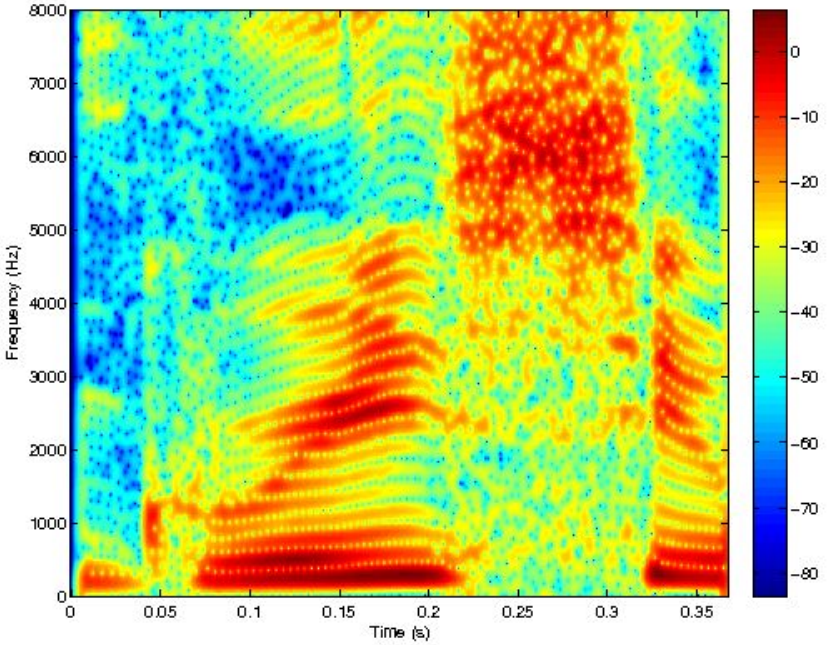
Objective: minimize MSE between denoised speech and clean speech (fidelity objective)



Weakness of Fidelity Objective

What parts of the denoised spectrogram are more important to get right?

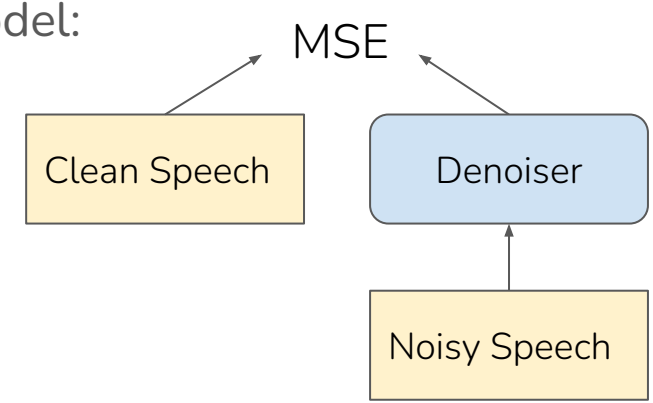
How do we know what looks like *real speech*?



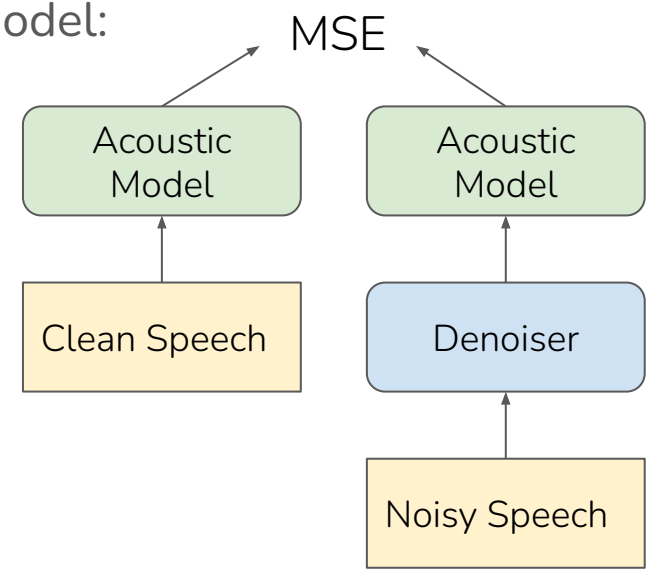
Realism Objective

A realism objective can teach our model how to extract features helpful for speech recognition, rather than just trying to imitate clean speech.

Old Model:



New Model:

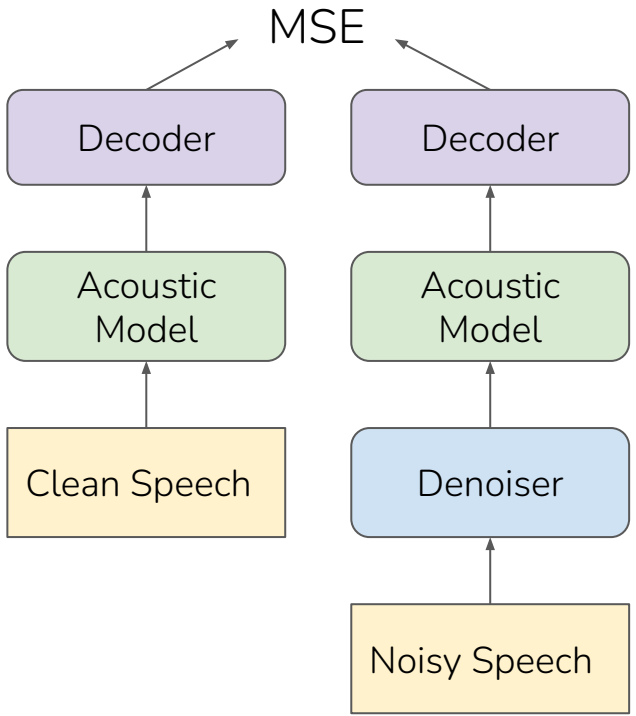


Results

Inputs to ASR pipeline	Word Error Rate
Noisy input	17.3
Fidelity objective	16.2
Realism objective	—
Joint objective	14.8

Future Work

- Add Decoder module on top of acoustic model
- Use more sophisticated neural network model
- Remove dependence on parallel clean/noisy speech



Questions?