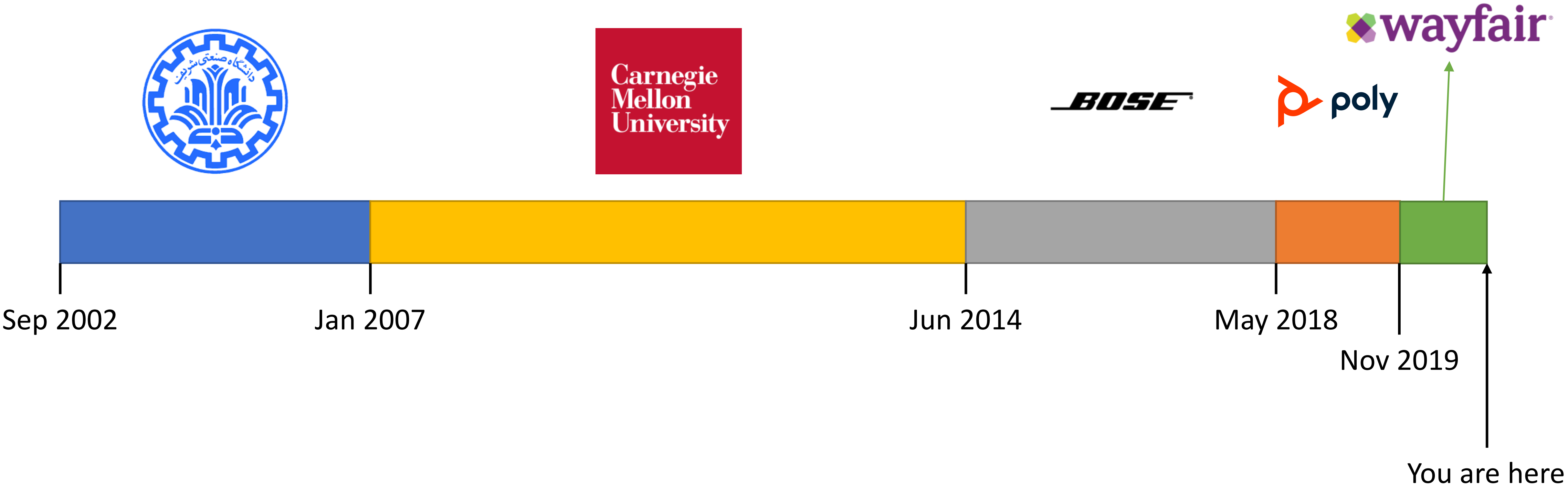
Applications of machine learning in speech and audio processing

Amir R. Moghimi

July 2020

My background



Topics for today

Automatic Speech Recognition (ASR)

What someone is saying (audio → text)

Voice Activity Detection (VAD)

Whether someone is speaking

Speech Enhancement

Making speech sound better

But first, a look at speech

Audio signals

Sound: longitudinal waves propagating through a medium (e.g., air)

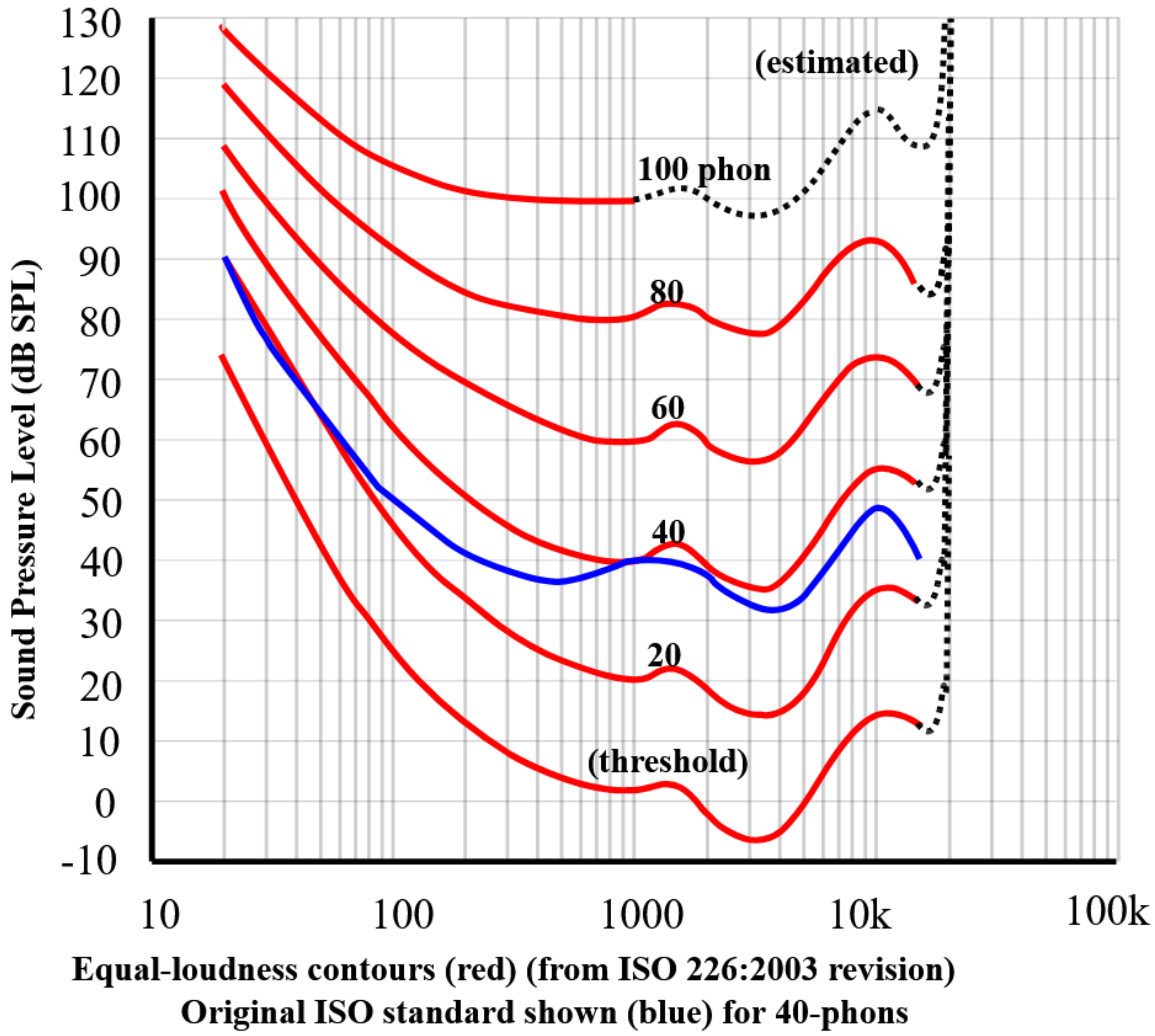
Audio signal: vibrations as received at a microphone or eardrum

Human hearing: $20\ \mathrm{Hz}$ to $20\ \mathrm{KHz}$. Typically sampled at

• $48\ \mathrm{KHz}$	“full-band audio”
• $44.1\ \mathrm{KHz}$	“CD quality”
• $16\ \mathrm{KHz}$	“wideband speech”
• $8\ \mathrm{KHz}$	“narrowband speech” or “telephone quality”
• …	

Human hearing

20 Hz to 20 KHz is very generous



https://en.wikipedia.org/wiki/Equal-loudness_contour

Speech: from language to sound

Language

consists of

Words

consist of

Phonemes

realized as

Phones

manifest as

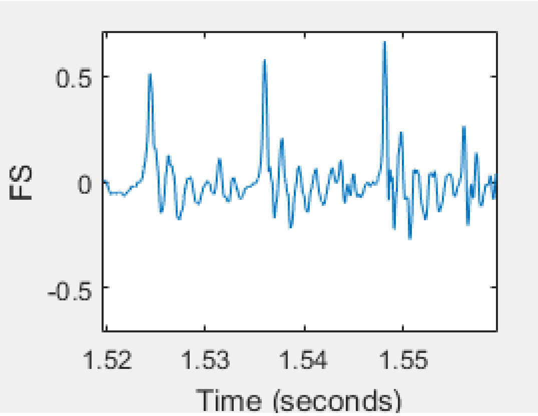
Sounds

Glass will clink when struck by metal.

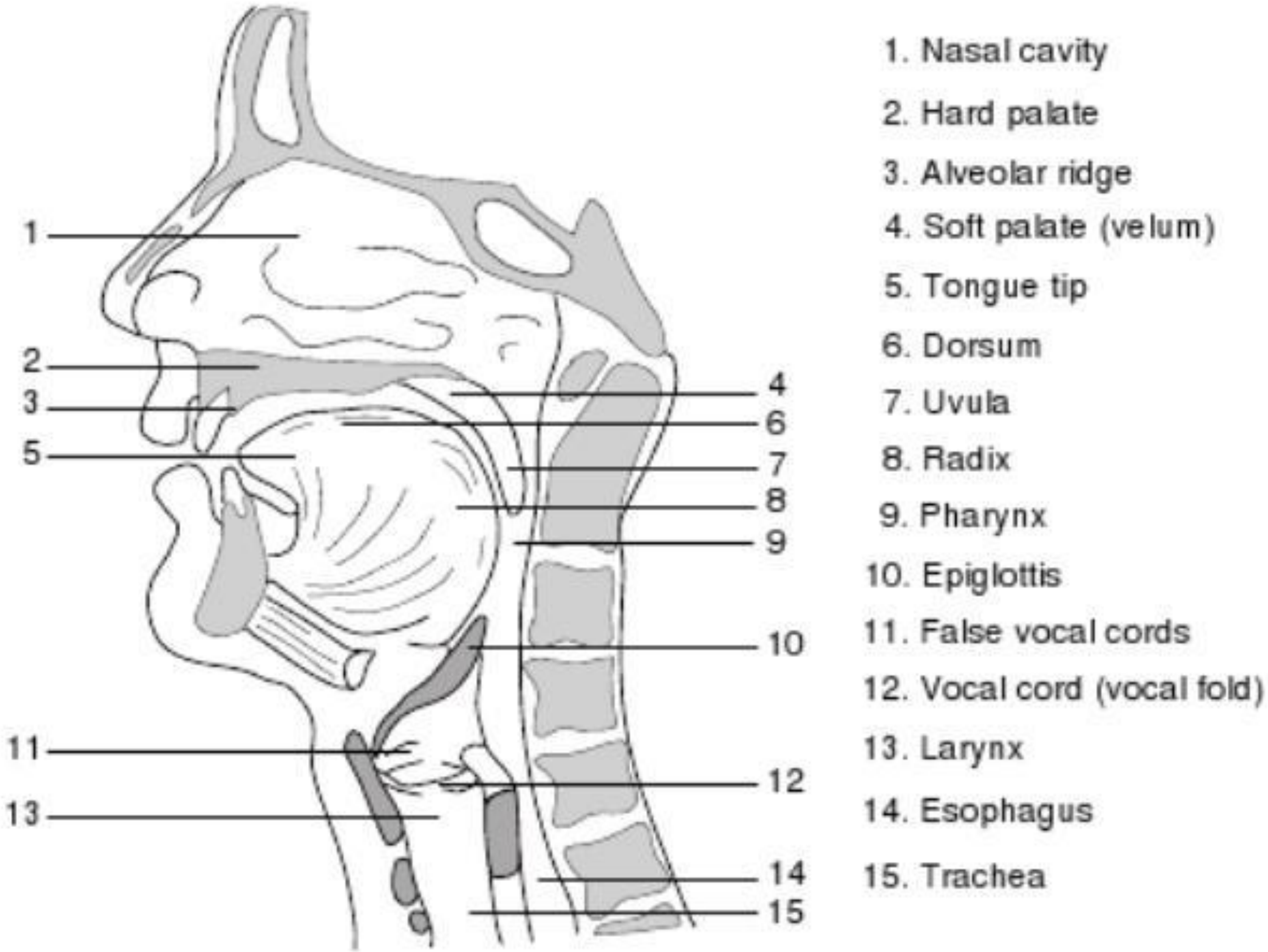
struck

[s t ɹ ʌ k]

This [ʌ]

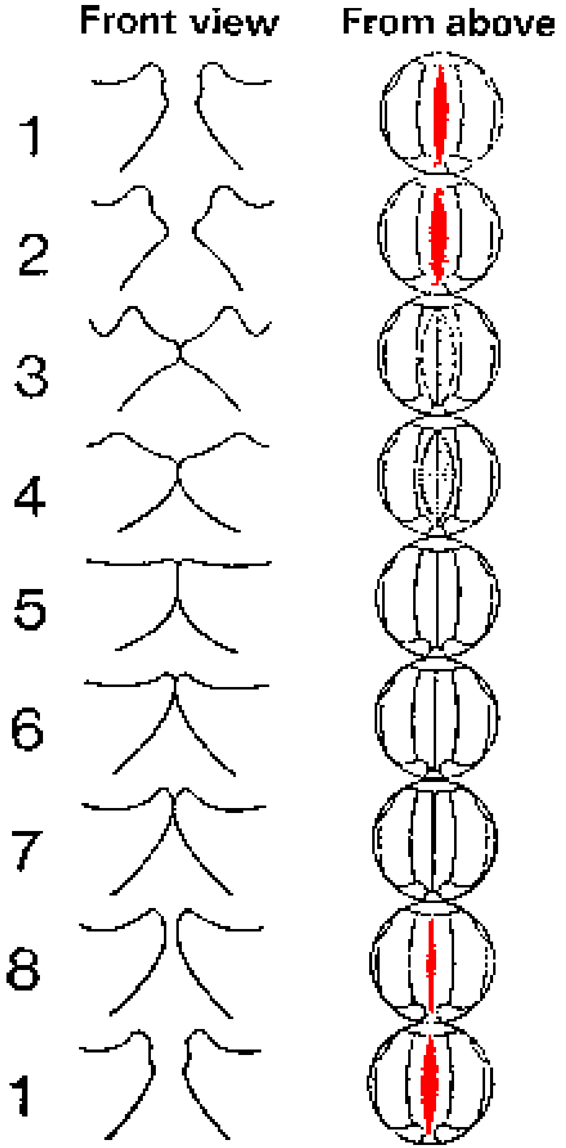


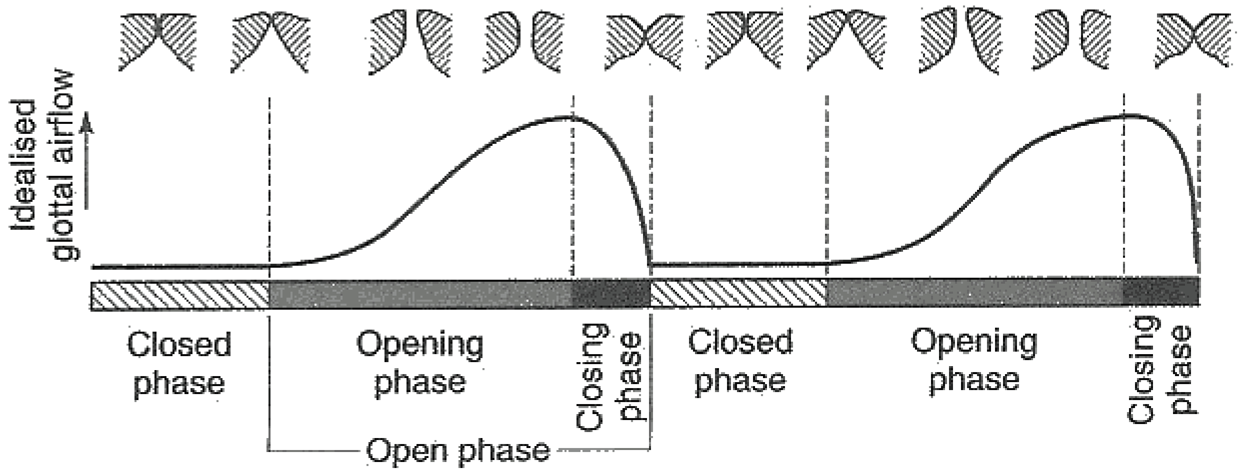
Anatomy of speech production



http://test.virtual-labs.ac.in/labs/ece01/exp3/theory/

Vocal cords for voiced speech

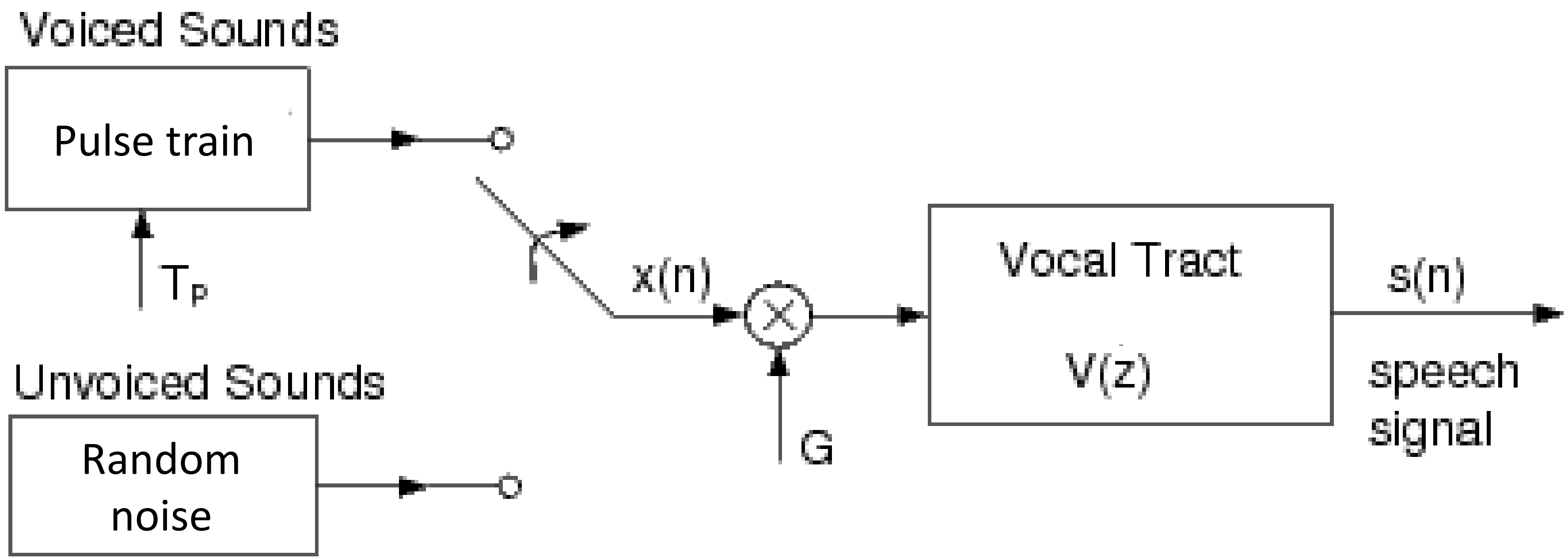




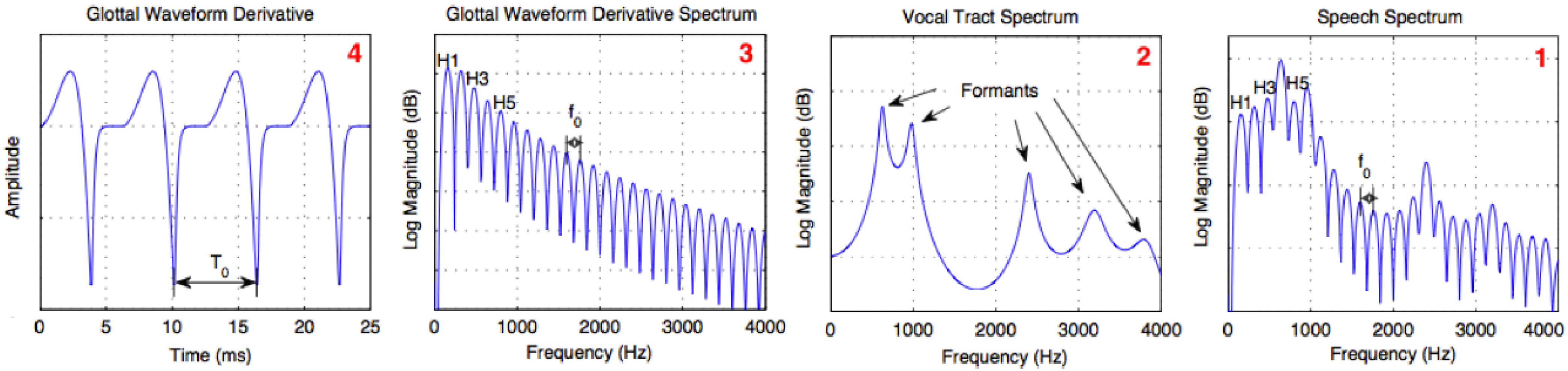
http://www.feilding.net/sfuad/musi3012-01/images/lectures/vocal_fold_cycles.gif

http://www.phy.duke.edu/~dtl/136126/restrict/Voice/foldwave.gif

Source-filter model of speech production

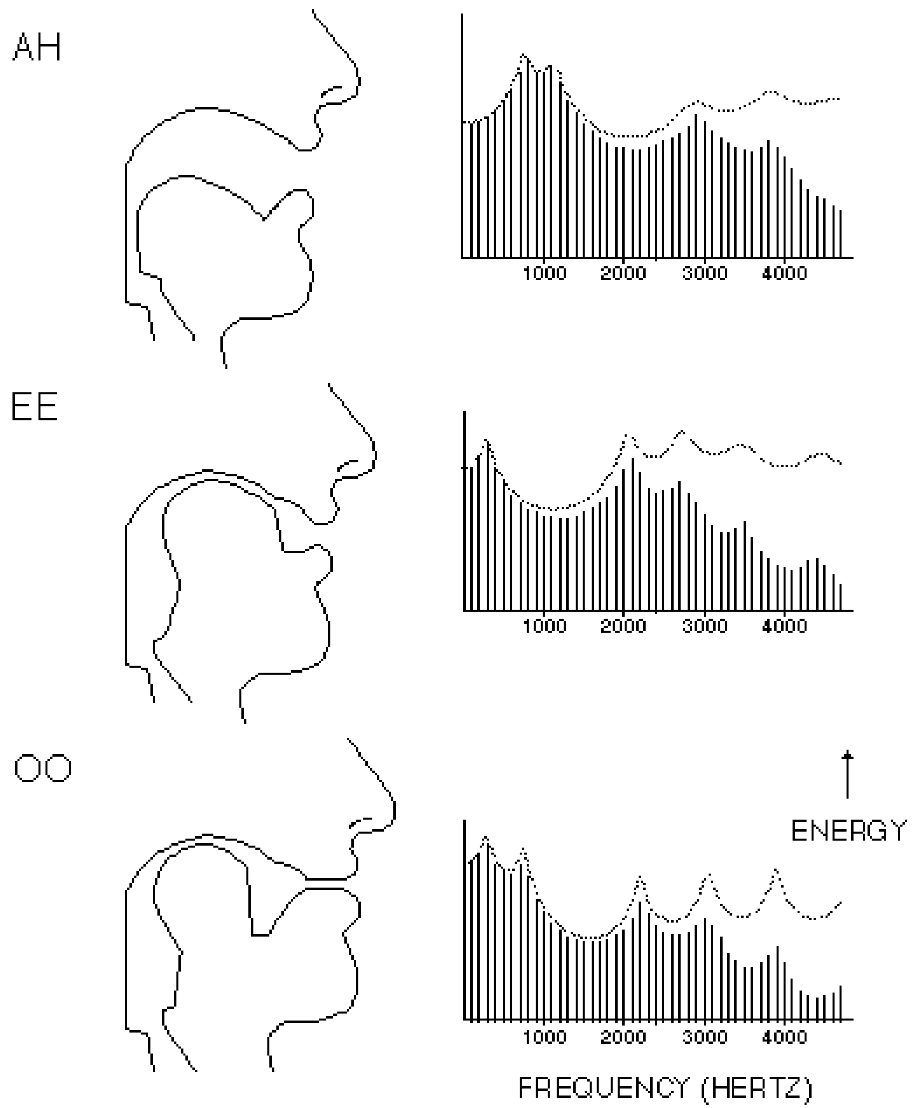


Cooking up a phone



Vandyke, D. J. (2014). *Glottal Waveforms for Speaker Inference & A Regression Score Post-Processing Method Applicable to General Classification Problems* (Doctoral dissertation, University of Canberra).

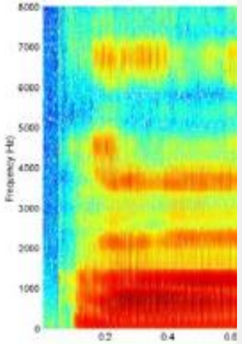
Anatomy and acoustics



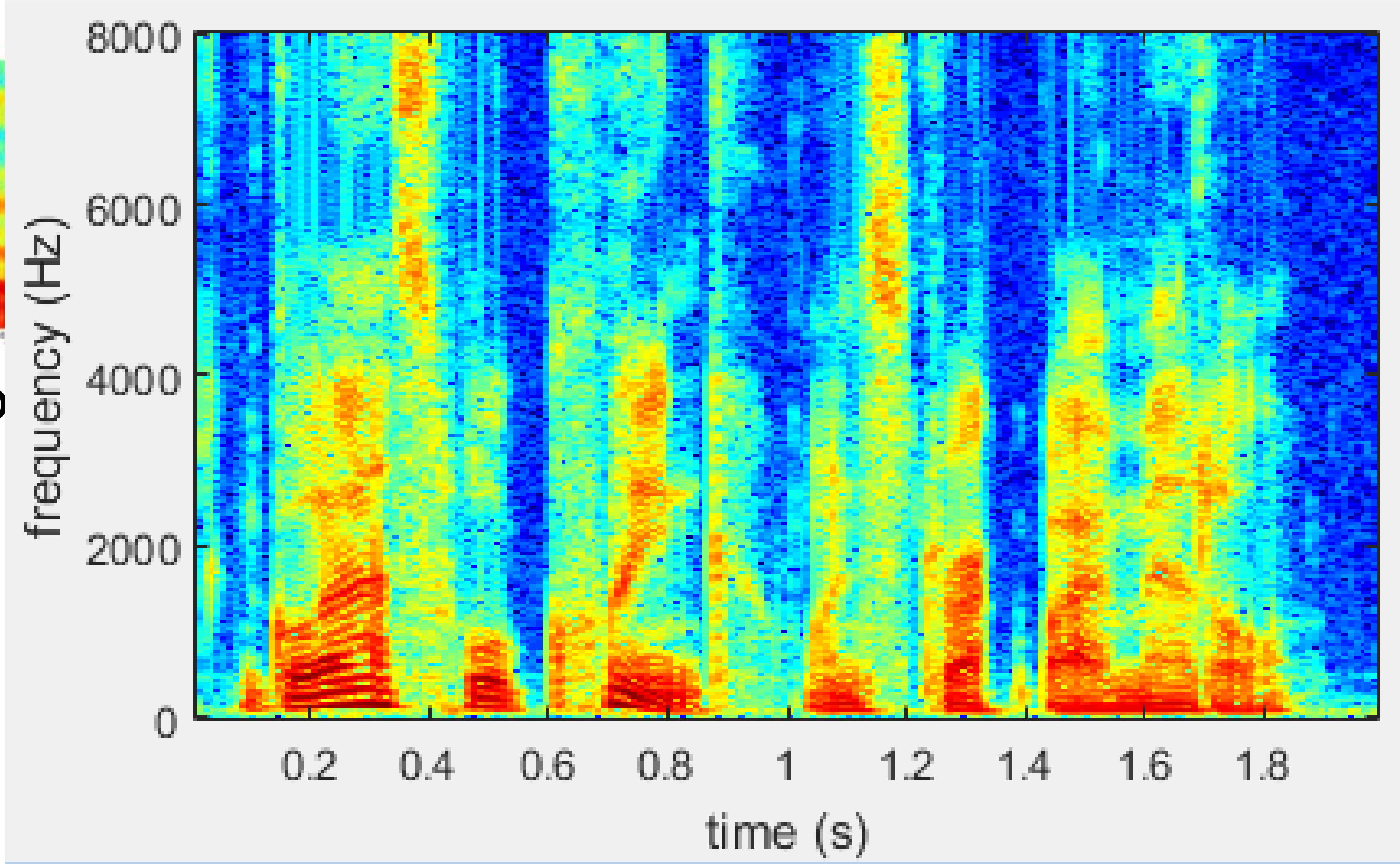
http://www.haskins.yale.edu/featured/heads/mmsp/intro.html

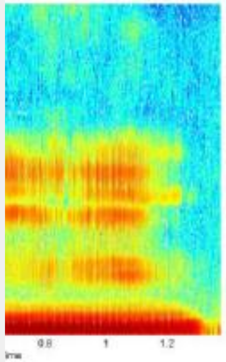
Spectral and temporal behavior

g l æ s wəl k l ɪ n k wɛn s t ɹ ʌ k b aɪ m ɛ r ə l



/b





n/

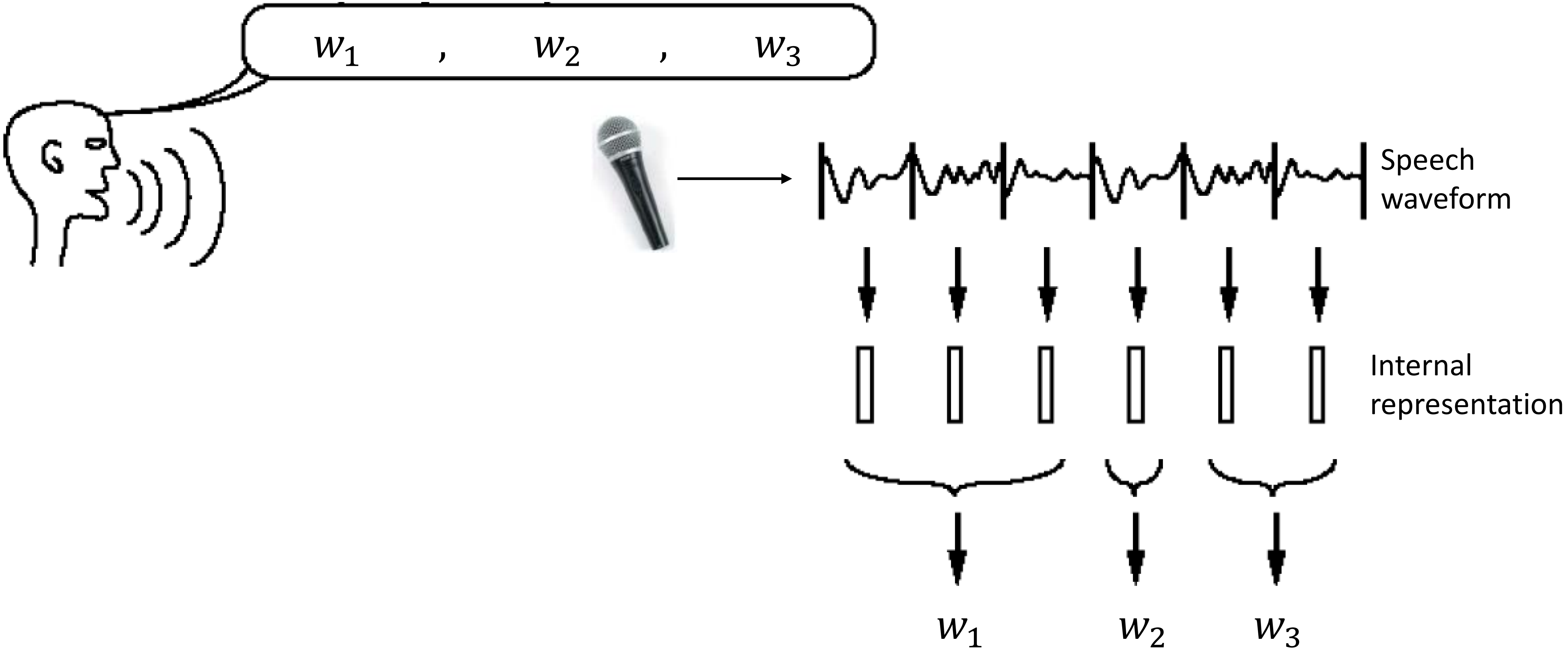
And now, some numbers

Spectral range:	50 Hz – 12 KHz	
Pitch (fundamental frequency):		
Male:	50 – 250 Hz	
Female:	120 – 500 Hz	
Vowel durations:	40 – 400 ms	(English)
Phone rate:	9.4 – 13.8 phone/s	(English) (***roughly***)

Speech processing problems

Automatic Speech Recognition (ASR)

ASR: Reverse-engineering biology & physics



Why is speech recognition hard?

Speech variances

- Environmental, natural, systemic

Continuous speech and audio

- e.g., “I scream” vs. “ice cream”

Vocabulary sizes

- English language: 100,000+ words to 1,000,000+ words
- Native English speaker (active): 20,000 words
- 95% of common text: 3000 words

Many other factors

Measuring ASR performance

Ground truth: it is great seeing you all here today

Hypothesis: let's great see you all here two day

$$\text{word error rate (WER)} = \frac{\text{substitutions} + \text{deletions} + \text{insertions}}{\text{words in correct text}}$$

Measuring ASR performance

Ground truth: it is great seeing you all here today

Hypothesis: let's great see you all here two day

$$\text{word error rate (WER)} = \frac{\text{substitutions} + \text{deletions} + \text{insertions}}{\text{words in correct text}}$$

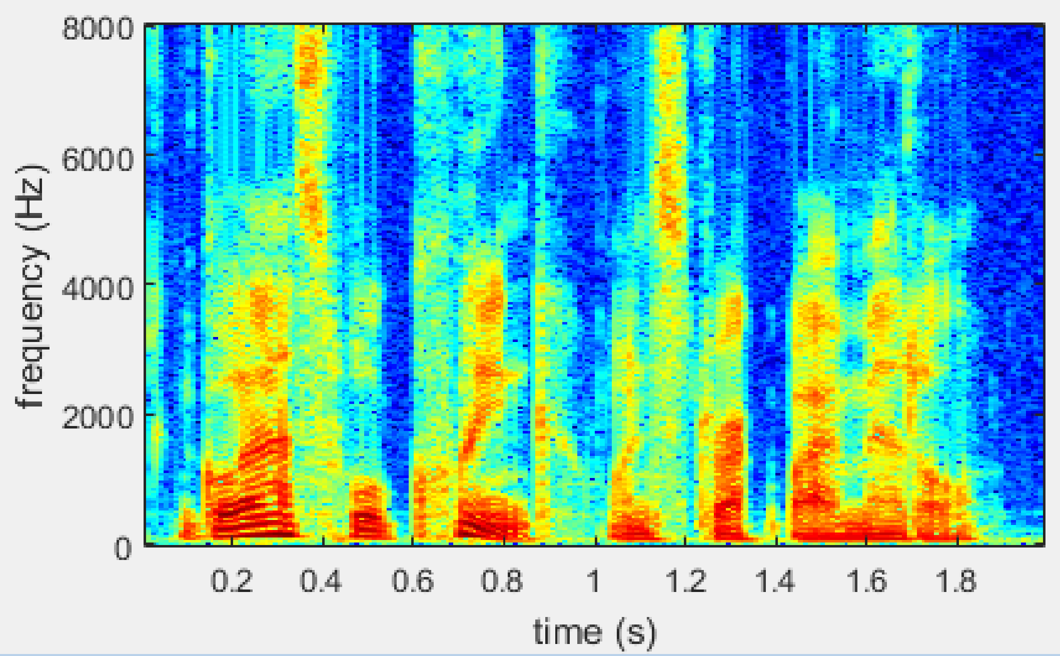
$$= \frac{3 + 1 + 1}{8} = 50\%$$

a.k.a. (relative) Levenshtein distance

ASR as a pattern matching problem

Feature extraction

Speech information: **spectral** and **temporal**



$$\mathbf{o}(t) = \begin{bmatrix} o_1(t) \\ \vdots \\ o_N(t) \end{bmatrix} = \begin{bmatrix} X(t, f_0) \\ \vdots \\ X(t, f_{N-1}) \end{bmatrix}$$

Commonly used speech features

Energy and pitch

Log power spectra

Linear Predictive Coding (LPC) coefficients

Mel-frequency Cepstral Coefficients (MFCC)

Perceptual Linear Prediction (PLP) coefficients

Power-normalized Cepstral Coefficients (PNCC)

…

Very early ASR

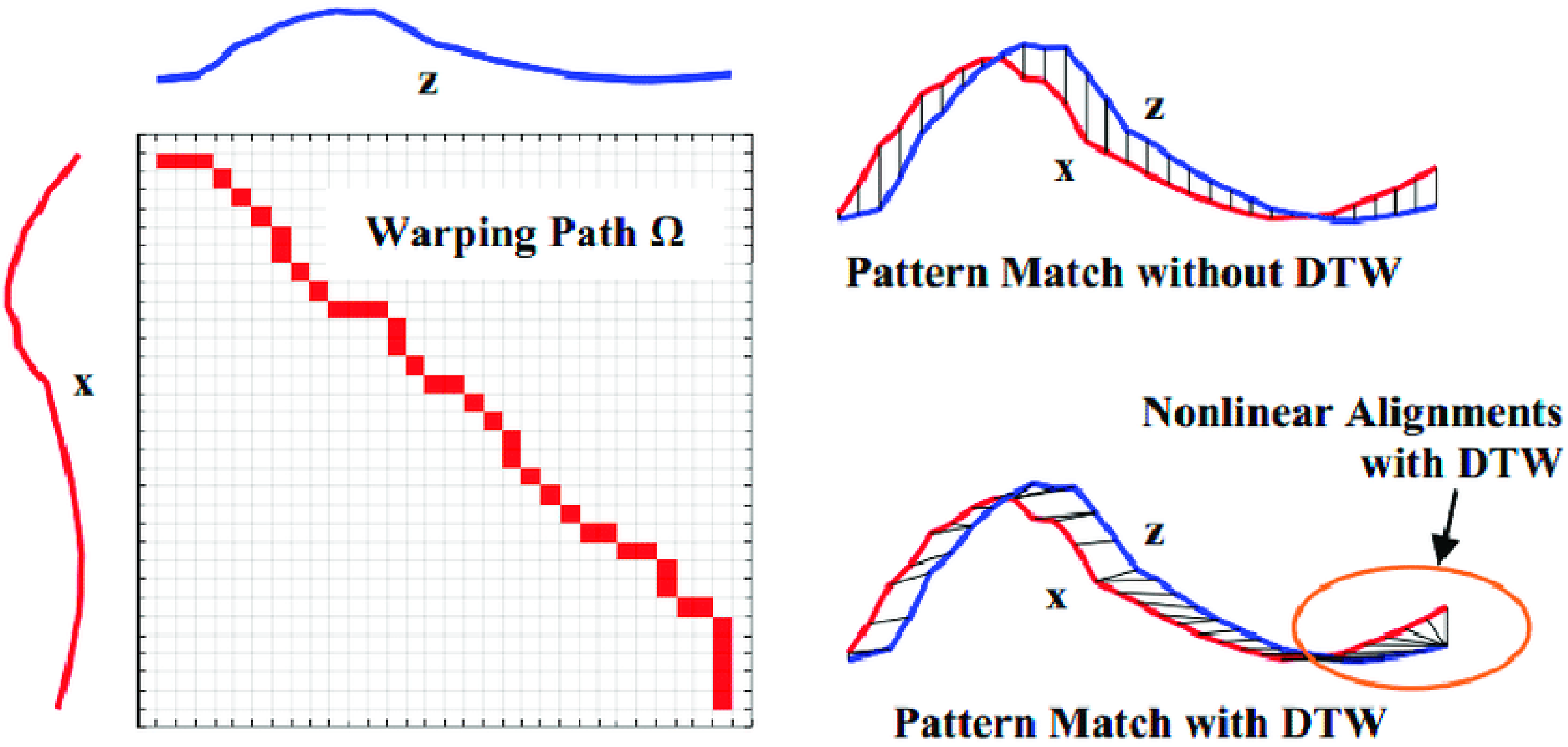
Template matching

- Compare new recording to pre-recorded samples

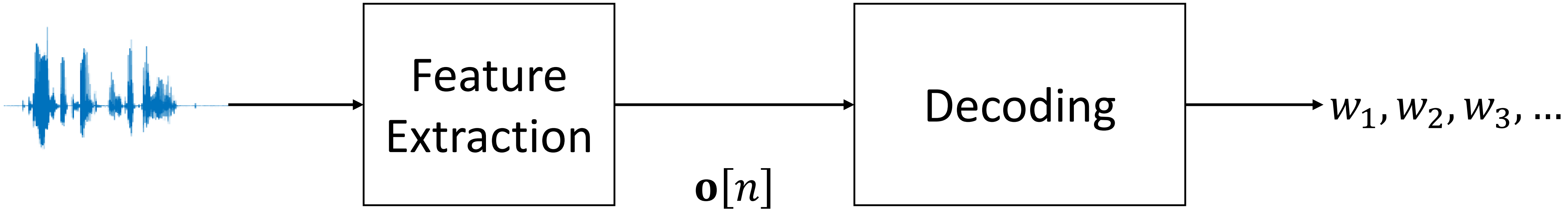
Dynamic Time Warping (DTW)

- Allows for timing fluidity

Dynamic Time Warping (DTW)



ASR as MAP detection



$$\boldsymbol{O} = [\mathbf{o}[0] \quad \mathbf{o}[1] \quad \ldots] \qquad \boldsymbol{W} = [w_1 \quad w_2 \quad \ldots]$$

$$\widehat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\operatorname{argmax}} \Pr\{\boldsymbol{W} | \boldsymbol{O}\}$$

The models of ASR

$$\widehat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\operatorname{argmax}} \Pr\{\boldsymbol{W}|\boldsymbol{O}\}$$

$$\boldsymbol{O} = [\mathbf{o}[0] \quad \mathbf{o}[1] \quad \ldots]$$

$$\boldsymbol{W} = [w_1 \quad w_2 \quad \ldots]$$

$$= \underset{\boldsymbol{W}}{\operatorname{argmax}} \frac{\Pr\{\boldsymbol{O}|\boldsymbol{W}\}\Pr\{\boldsymbol{W}\}}{\Pr\{\boldsymbol{O}\}}$$

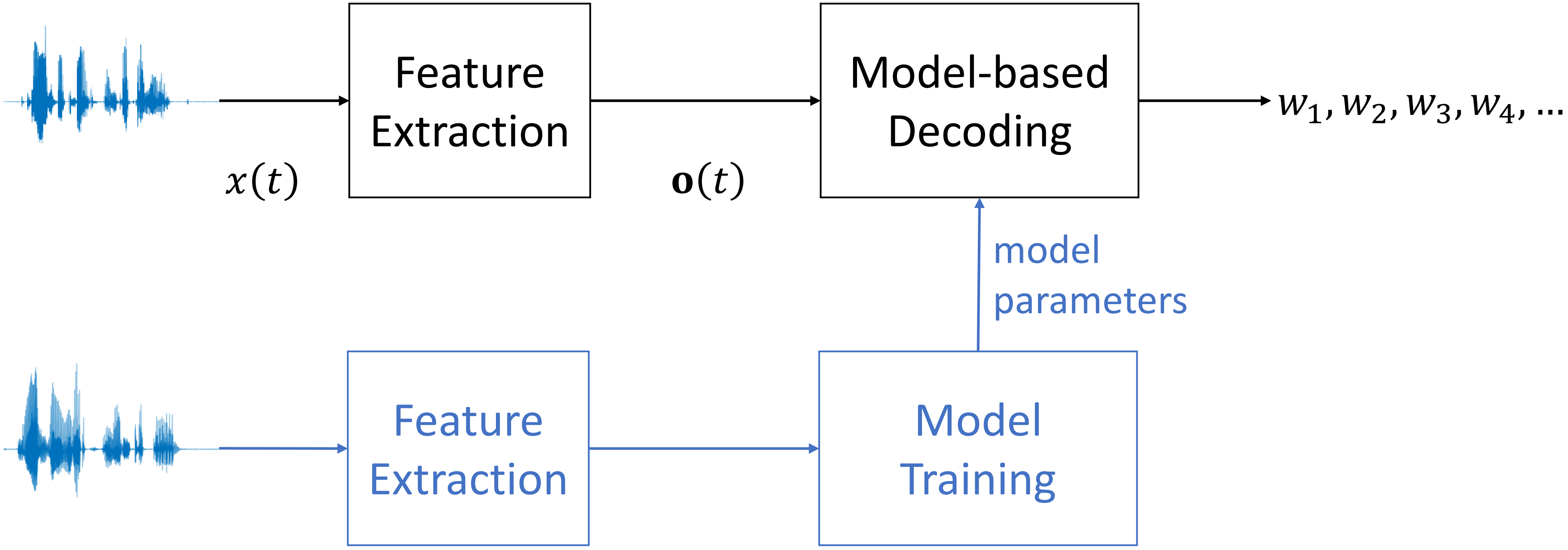
$$= \underset{\boldsymbol{W}}{\operatorname{argmax}} \Pr\{\boldsymbol{O}|\boldsymbol{W}\}\Pr\{\boldsymbol{W}\} \quad = \underset{\boldsymbol{W}}{\operatorname{argmax}} \Pr\{\boldsymbol{O}|\boldsymbol{P}\}\Pr\{\boldsymbol{P}|\boldsymbol{W}\}\Pr\{\boldsymbol{W}\}$$

Acoustic Model (AM)

Dictionary (a.k.a. Lexicon)

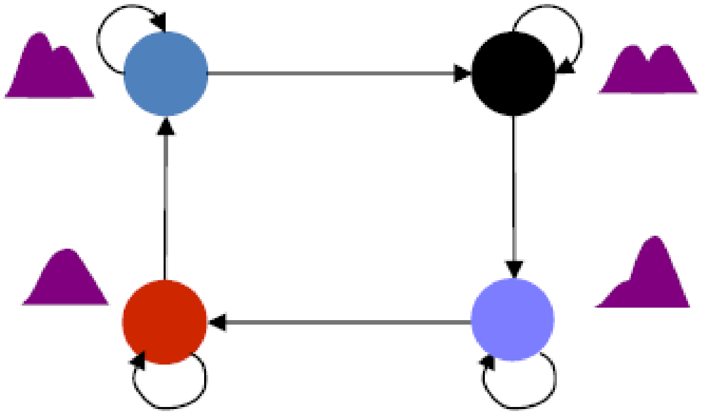
Language Model (LM)

ASR as a machine learning problem



Hidden Markov Model (HMM)

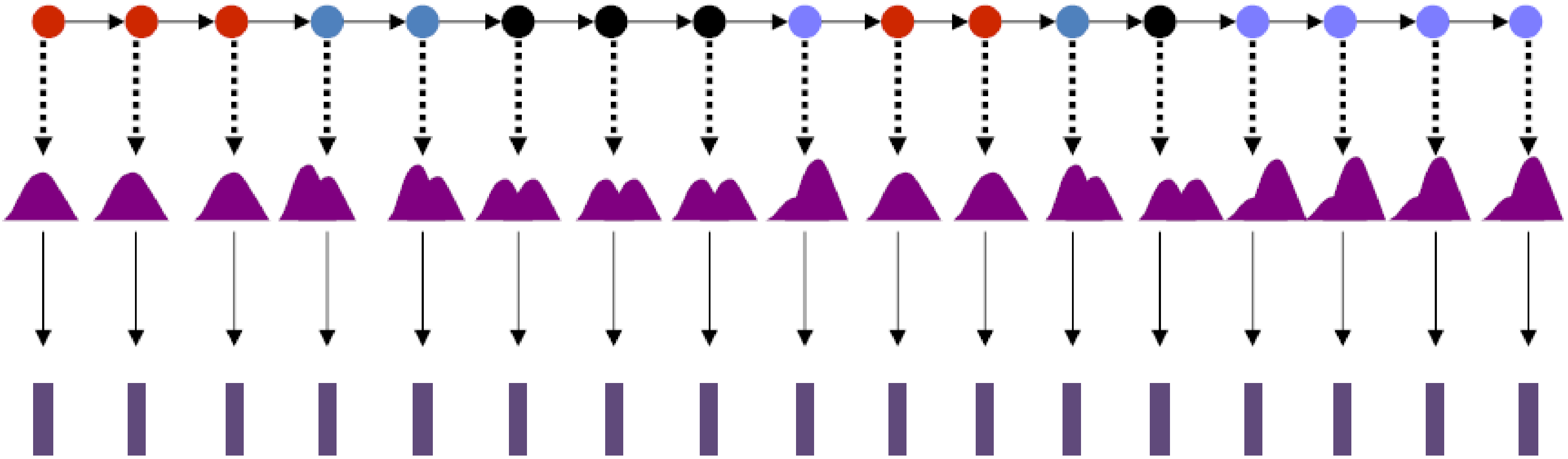
HMM assumed to be generating data



state sequence

state distributions

observation sequence

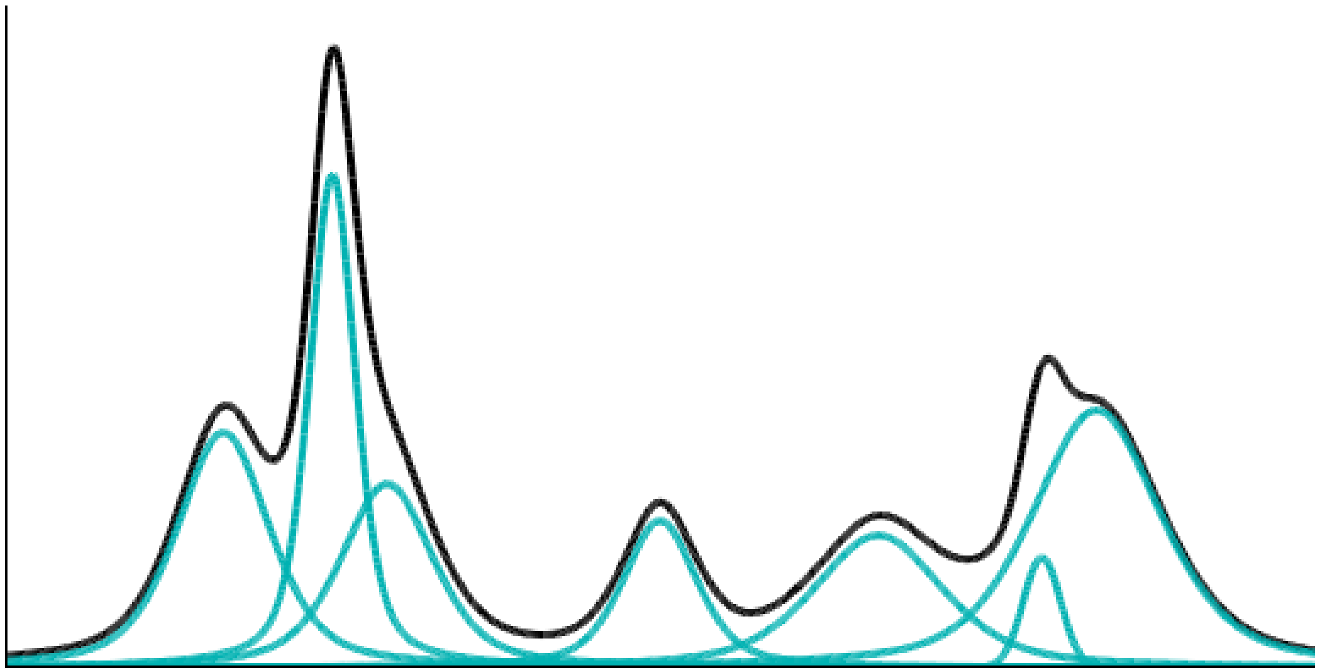


Gaussian mixture model (GMM)

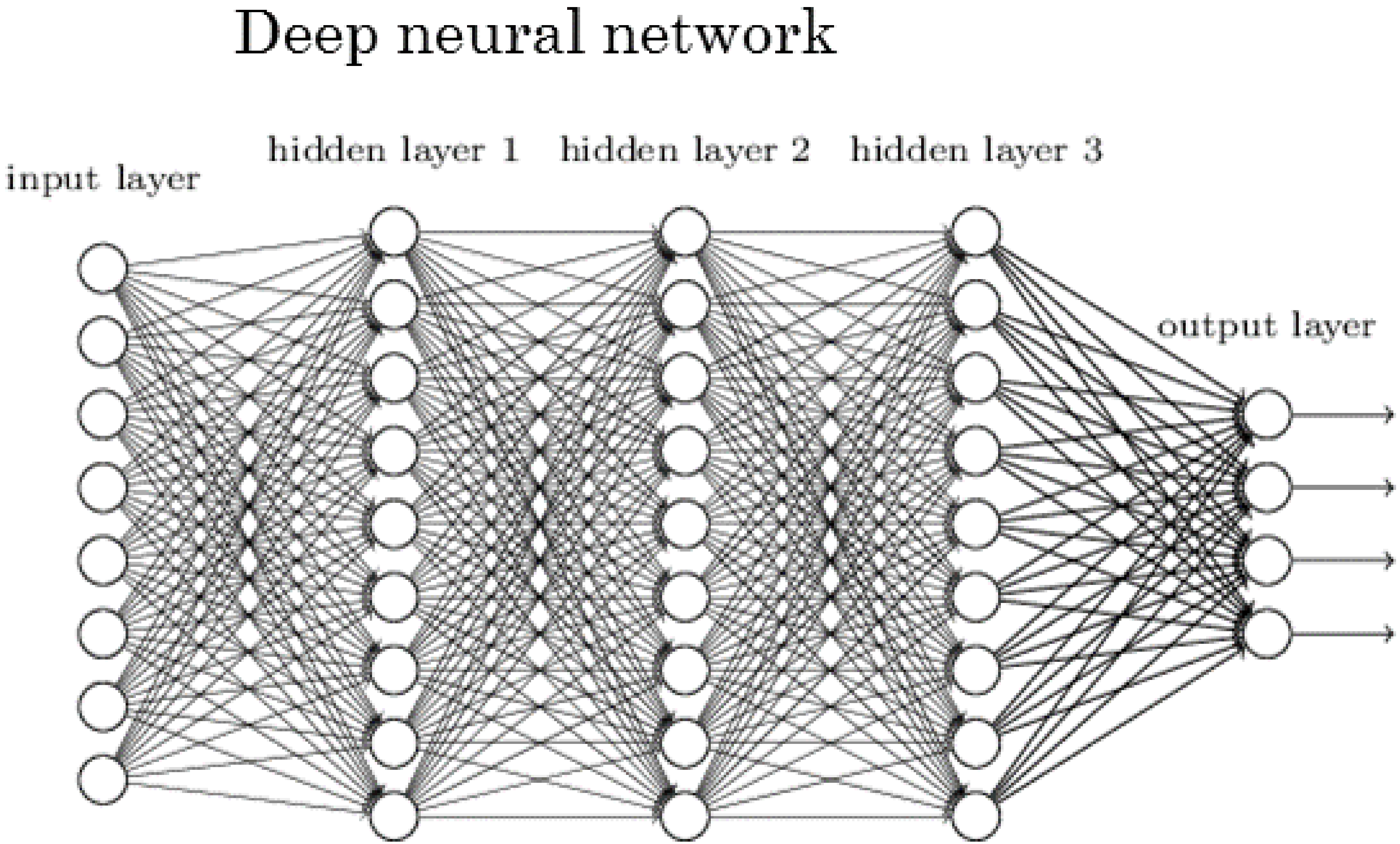
Family of probability distributions (parameterized)

$$f_{\mathbf{o}|s}(\mathbf{o}) = \sum_{k=1}^{K} \alpha_k g(\mathbf{o}; \boldsymbol{\mu}_k, C_k)$$

$$g(\mathbf{o}; \boldsymbol{\mu}, C) = \frac{1}{\sqrt{(2\pi)^D |C|}} e^{-\frac{1}{2}}$$



Deep learning comes to ASR



The new acoustic model: HMM-DNN

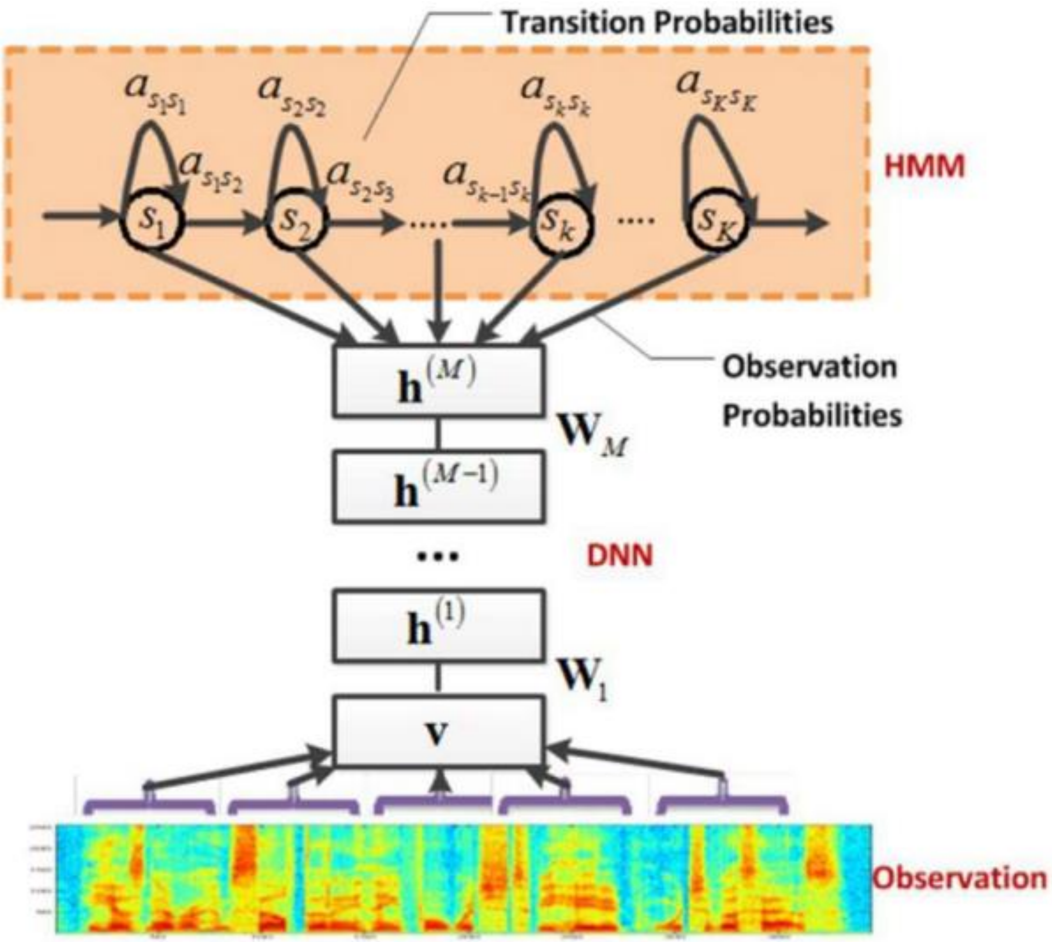
HMM output distributions:

~~GMM~~ DNN

Typical DNNs for speech have **millions** of parameters **per state**

→ huge training data

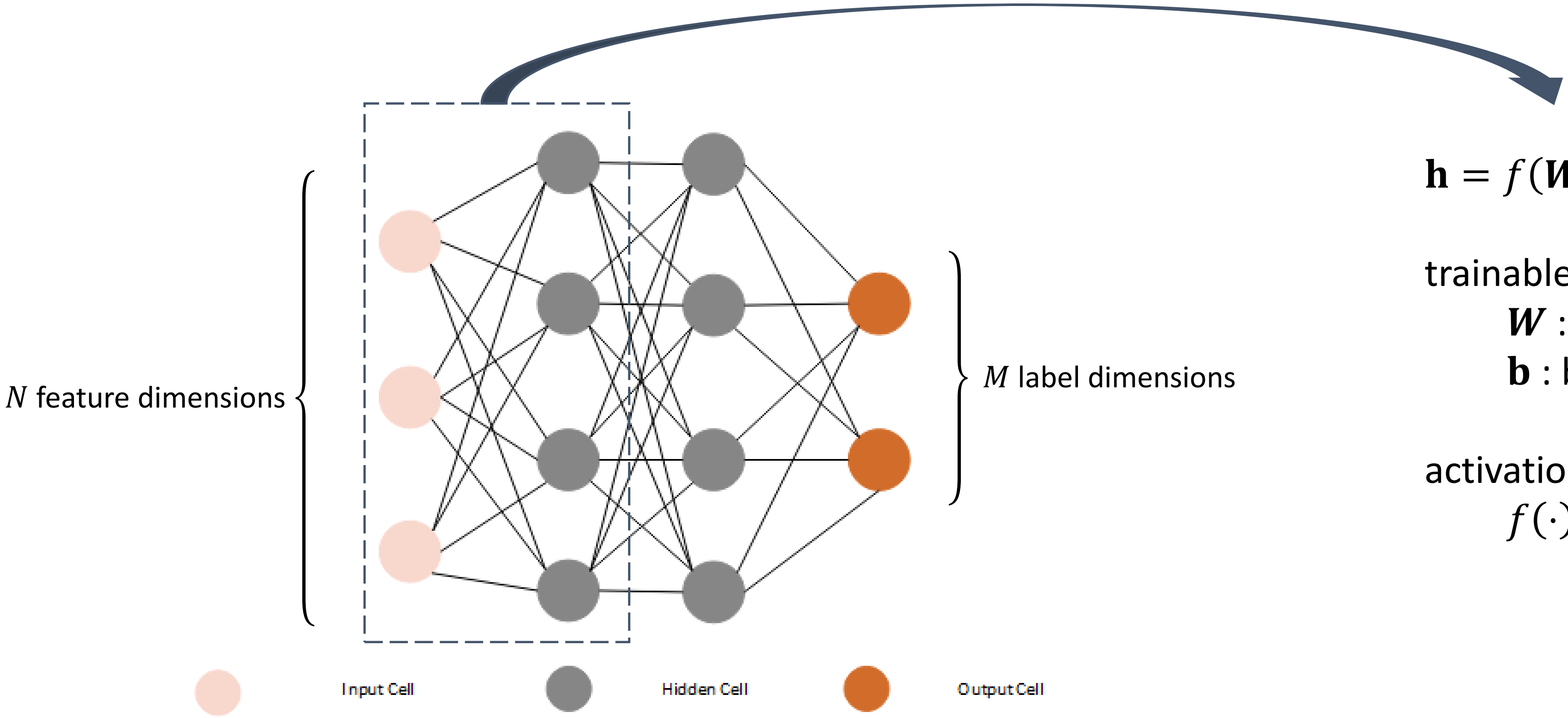
(thousands of hours)



Dahl, George E., et al. "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition." *Audio, Speech, and Language Processing, IEEE Transactions on* 20.1 (2012): 30-42.

Detour:
Recurrent Neural Networks (RNNs)

Feed-forward (densely connected) neural network



$$\mathbf{h} = f(\boldsymbol{W}\mathbf{x} + \mathbf{b})$$

trainable parameters:

$\boldsymbol{W}$: weight matrix

$\mathbf{b}$: bias vector

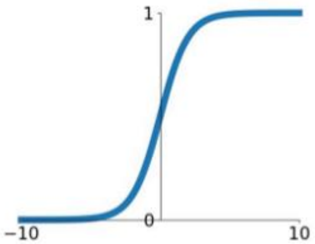
activation function:

$f(\cdot)$

Common activation functions

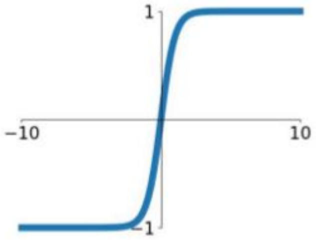
Sigmoid

$\sigma(x) = \frac{1}{1+e^{-x}}$



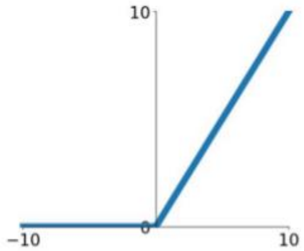
tanh

$\tanh(x)$



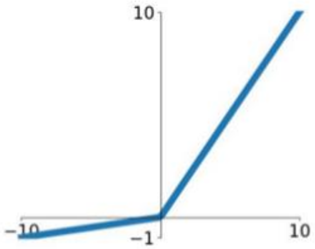
ReLU

$\max(0, x)$



Leaky ReLU

$\max(0.1x, x)$

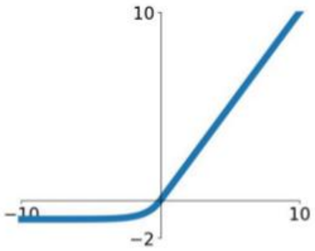


Maxout

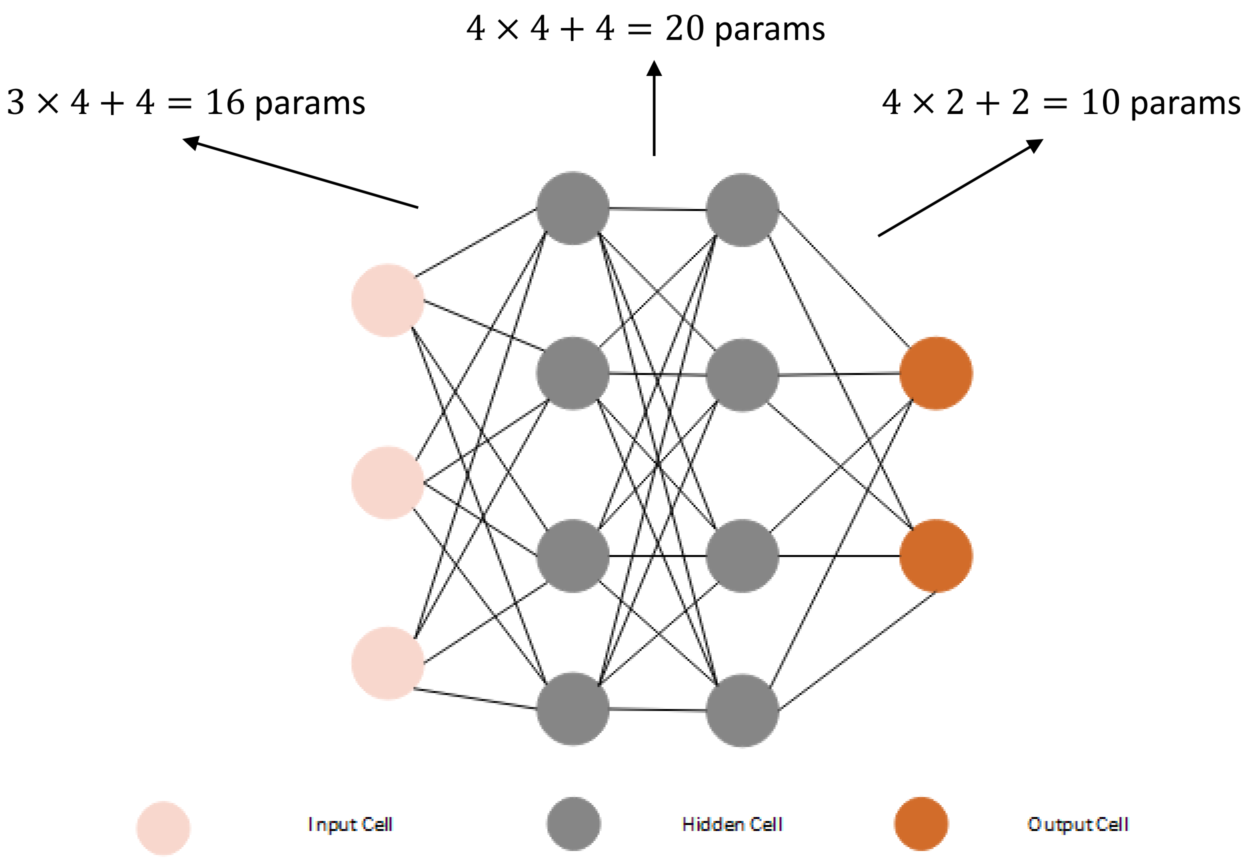
$\max(w_1^T x + b_1, w_2^T x + b_2)$

ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



Dense DNNs can get big fast



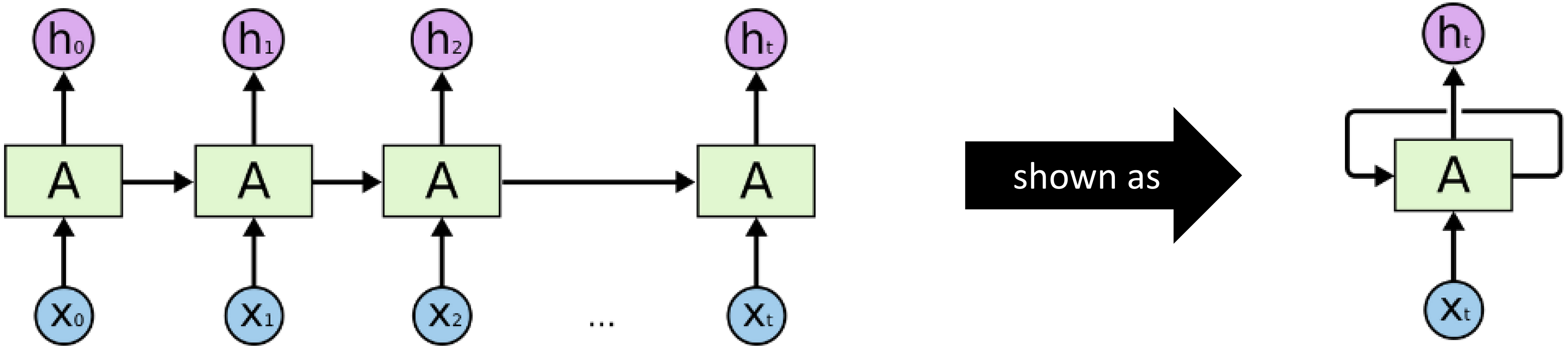
total: 46 trainable parameters

More parameters means more …

- memory & computational resources
- “capacity” for learning
- data required for training

Recurrence: weight-sharing across time

Recurrent layer has an output and a “state” that is fed back into the next copy as input:

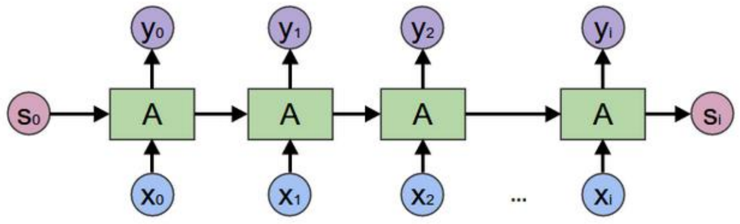


The “state” gathers information across time → arbitrary length sequences (on paper)

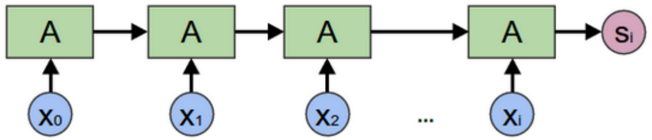
https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Sequence modeling with RNNs: variations

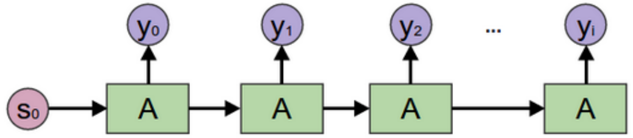
Many-to-many (sequence-to-sequence) RNN



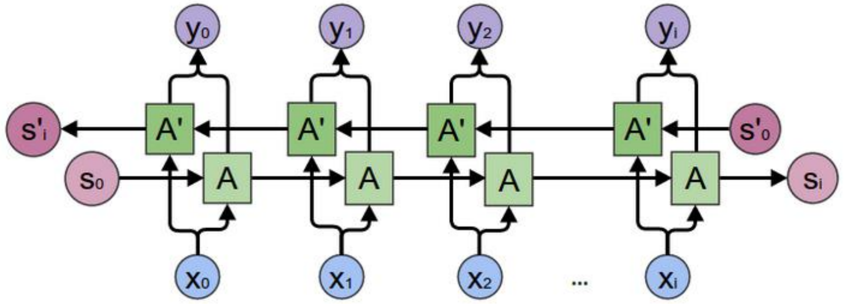
Many-to-one (encoding) RNN



One-to-many (generating) RNN

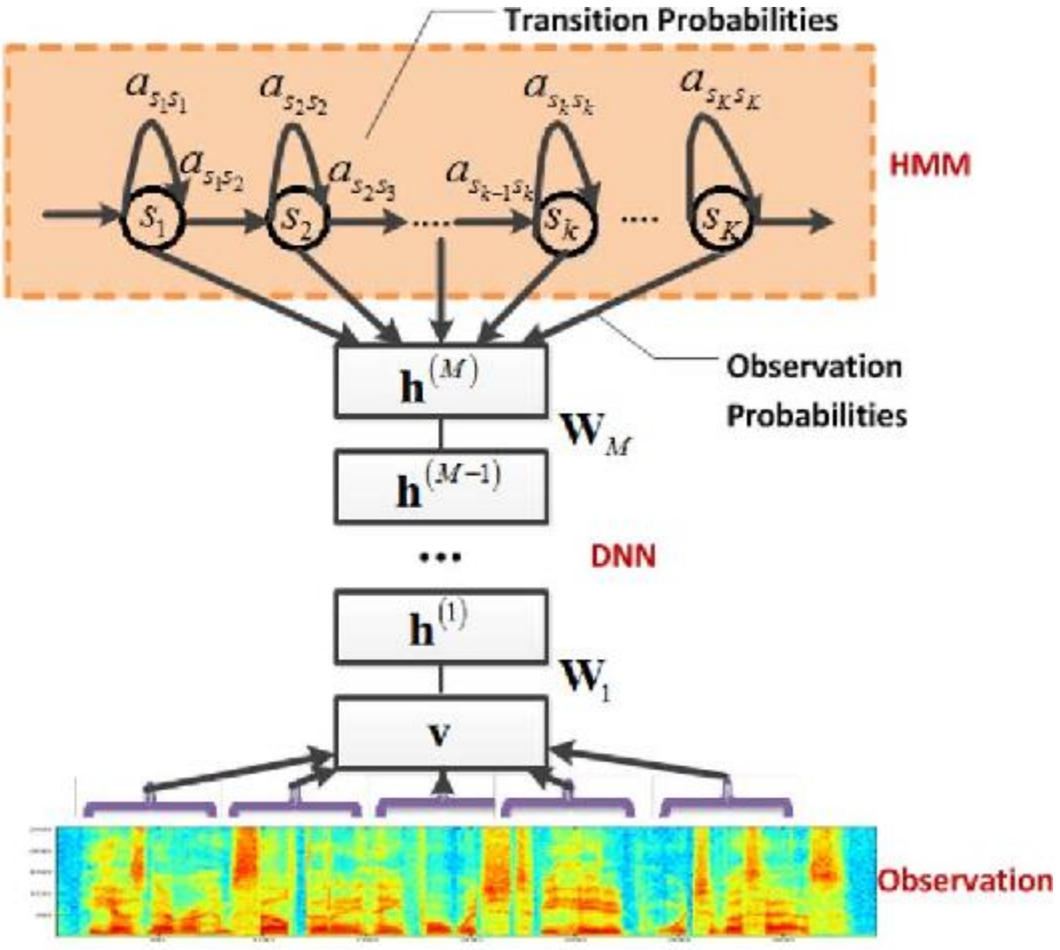


Bidirectional RNN (e.g., BLSTM)



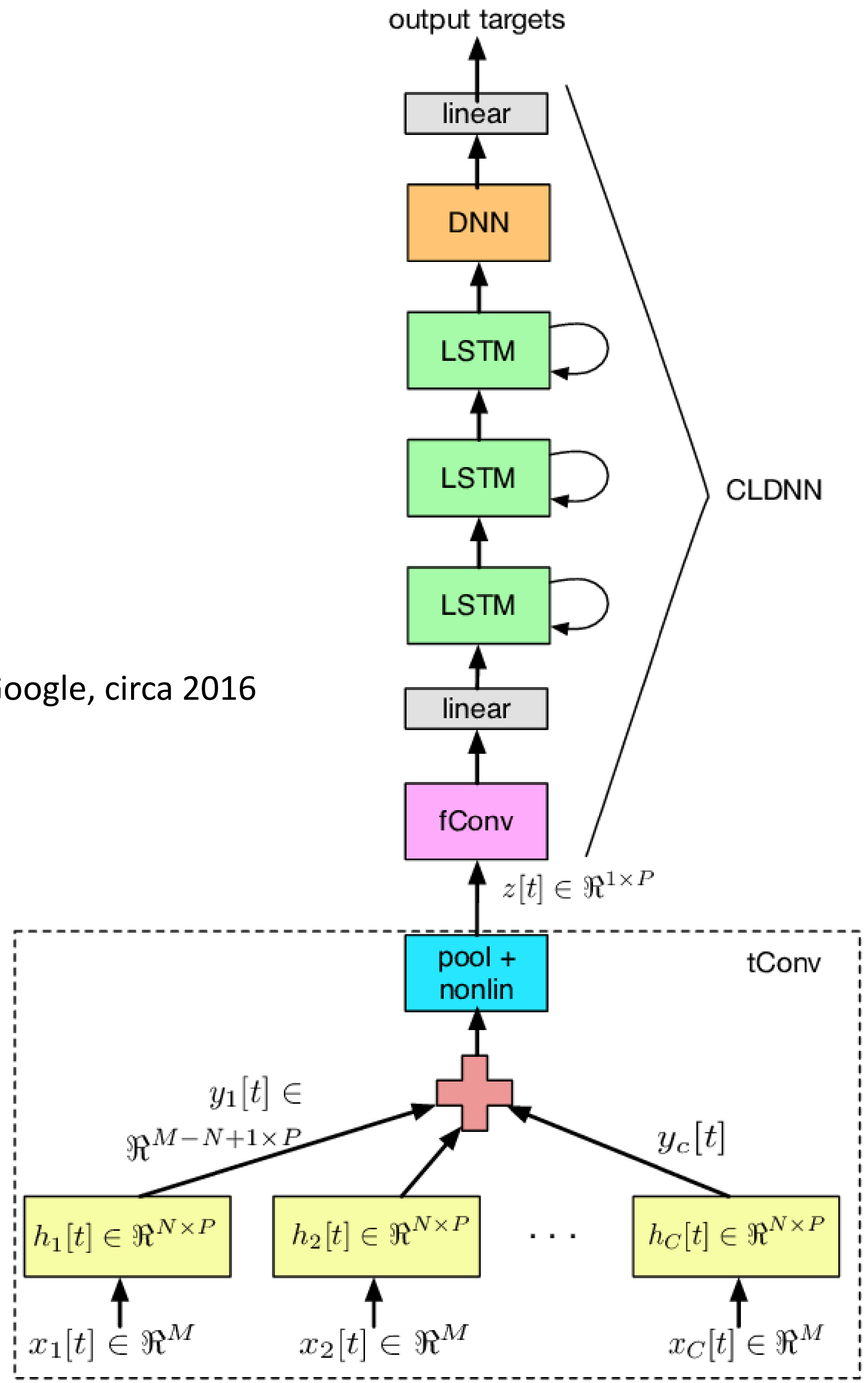
And now, back to our show

ASR with RNNs



Microsoft, circa 2010

Google, circa 2016



The impact of deep learning on speech

Model variations in large, diverse datasets

- Pronunciation
- Speaker
- Environment

Learn feature extraction

Signal → Pre-emph → STFT → Energy → Mel filtering → Log → DCT → Δ & ΔΔ → MFCC

The impact of deep learning on speech

The demise of HMMs

- Model temporal dynamics with RNNs

Merging acoustic and language models

Grapheme-based ASR

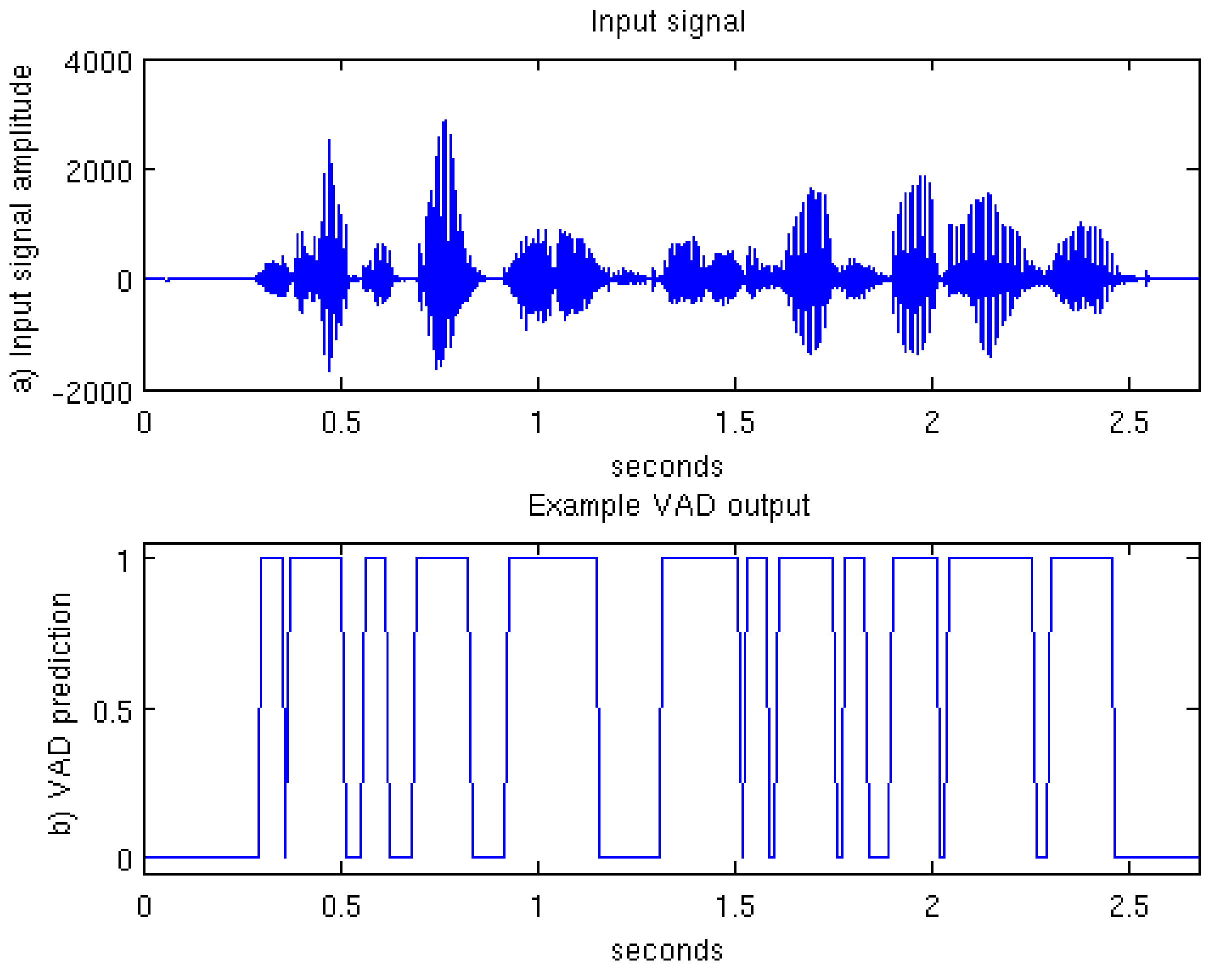
- No acoustics/phonetics knowledge at all!

Voice Activity Detection (VAD)

At what times in an audio recording or stream is there someone speaking?

Used in:

- Speech coding
- Speech recognition
- Speech communication (e.g., telephony)
- …



Evaluating a VAD

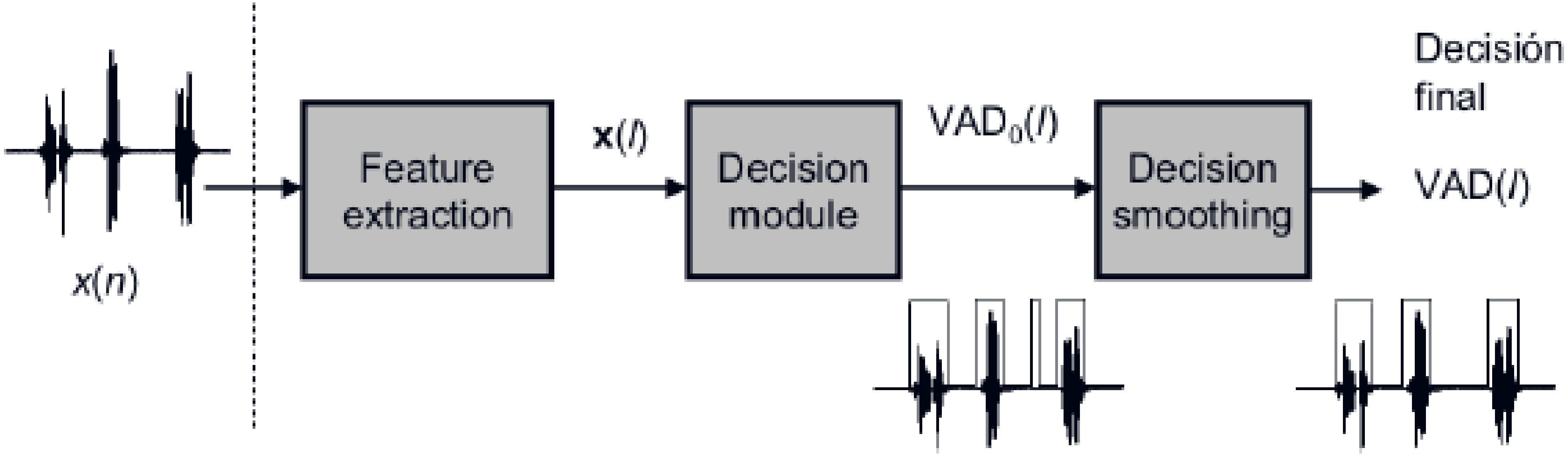
Signal-level metrics:

- FEC (Front End Clipping): clipping introduced in passing from noise to speech
- MSC (Mid Speech Clipping): clipping due to speech misclassified as noise
- OVER: noise interpreted as speech in passing from speech activity to noise
- NDS (Noise Detected as Speech): noise interpreted as speech

Application-based metrics:

- Speech quality (intelligibility, MOS, …)
- ASR performance (word error rate, …)
- …

VAD system components



Conventional VAD systems

Energy thresholding

Pitch detection & tracking

- Phase-lock loops, etc.

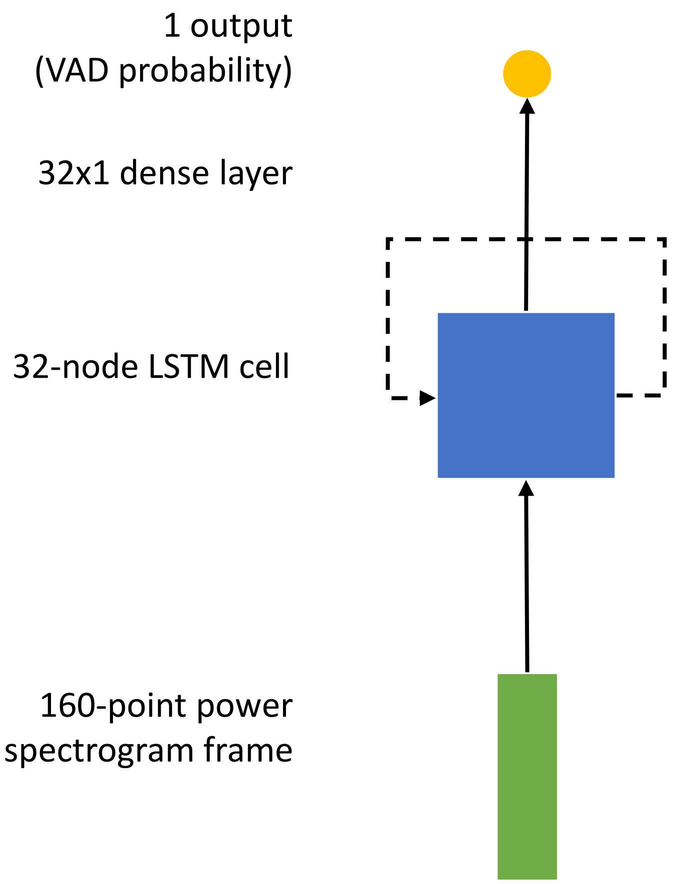
Frame-by-frame classification of:

- Auto-correlation function
- Power Spectral Density (PSD)
- Other features

Combinations of the above

…

Simple VAD with an RNN (LSTM)




```
from keras.layers import Input, Dense, concatenate
from keras.models import Model

# Create an input. This time, the input has a leading dimension of
# unspecific length to allow for arbitrary length sequences:
inputs = Input(shape=(None, 160))

# Define the LSTM cell:
hidden = LSTM(32, return_sequences=True)(inputs)

# Apply a Dense layer to the ouput to map it down to a single
# VAD probability per frame:
outputs = Dense(1, activation='sigmoid')(hidden)

# Define and compile the model
model = Model(inputs=inputs, outputs=outputs)
model.compile(...)
```

Speech enhancement

Speech quality can be degraded by

- Additive noise
- Reverberation
- Filtering
- Distortion
- Spectral processing
- Audio coding
- Network effects (e.g., packet loss)

Problem: Reconstruct original, “clean” speech from degraded speech

Speech quality evaluation

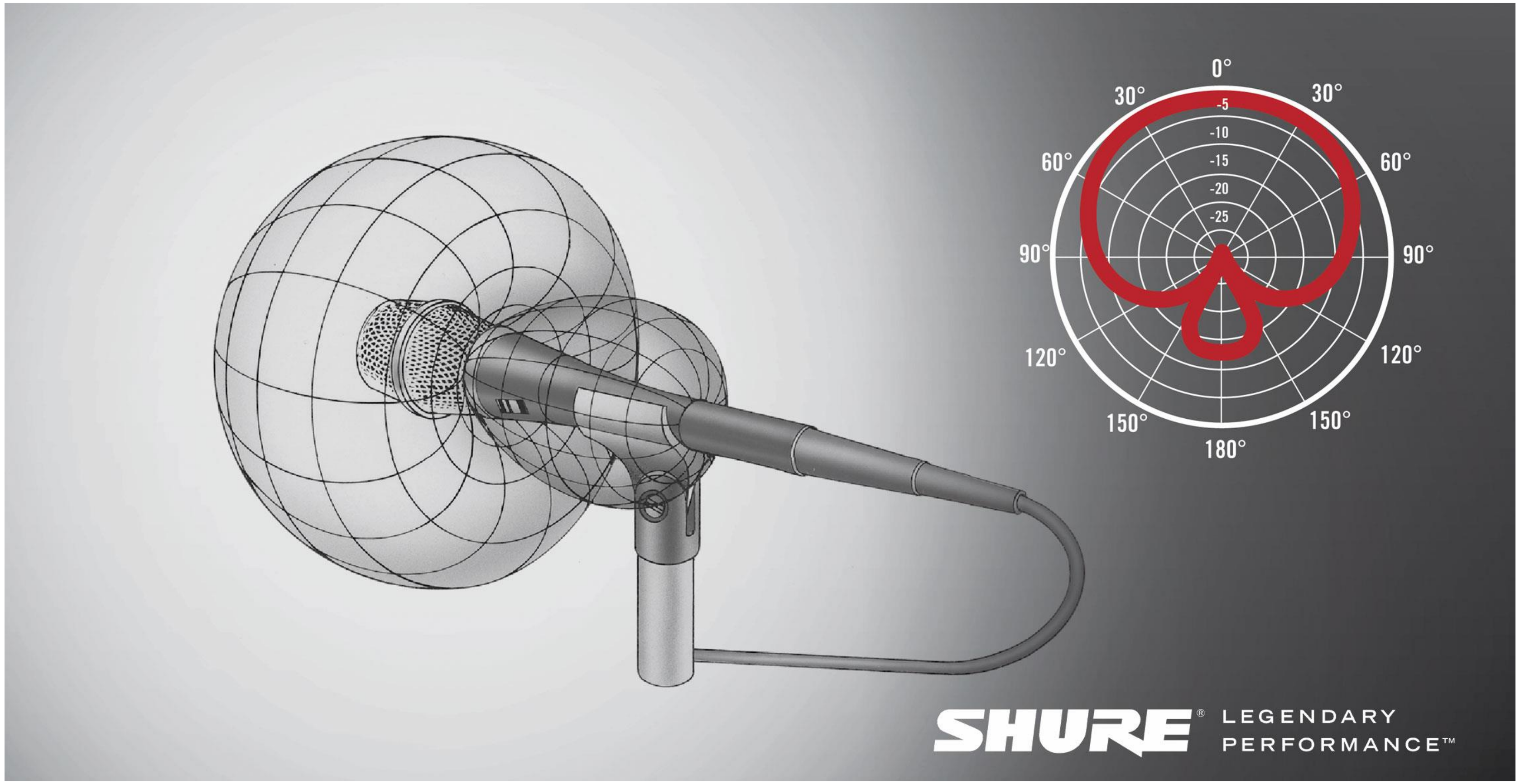
Intelligibility

- Human tests → measure WER
- Objective (calculable) metrics (e.g., STI or STOI)

Perceptual quality

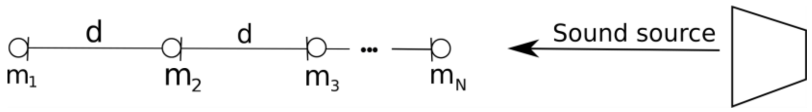
- Subjective listening tests → measure MOS, MUSHRA, etc.
- Objective (calculable) metrics (e.g., PESQ)

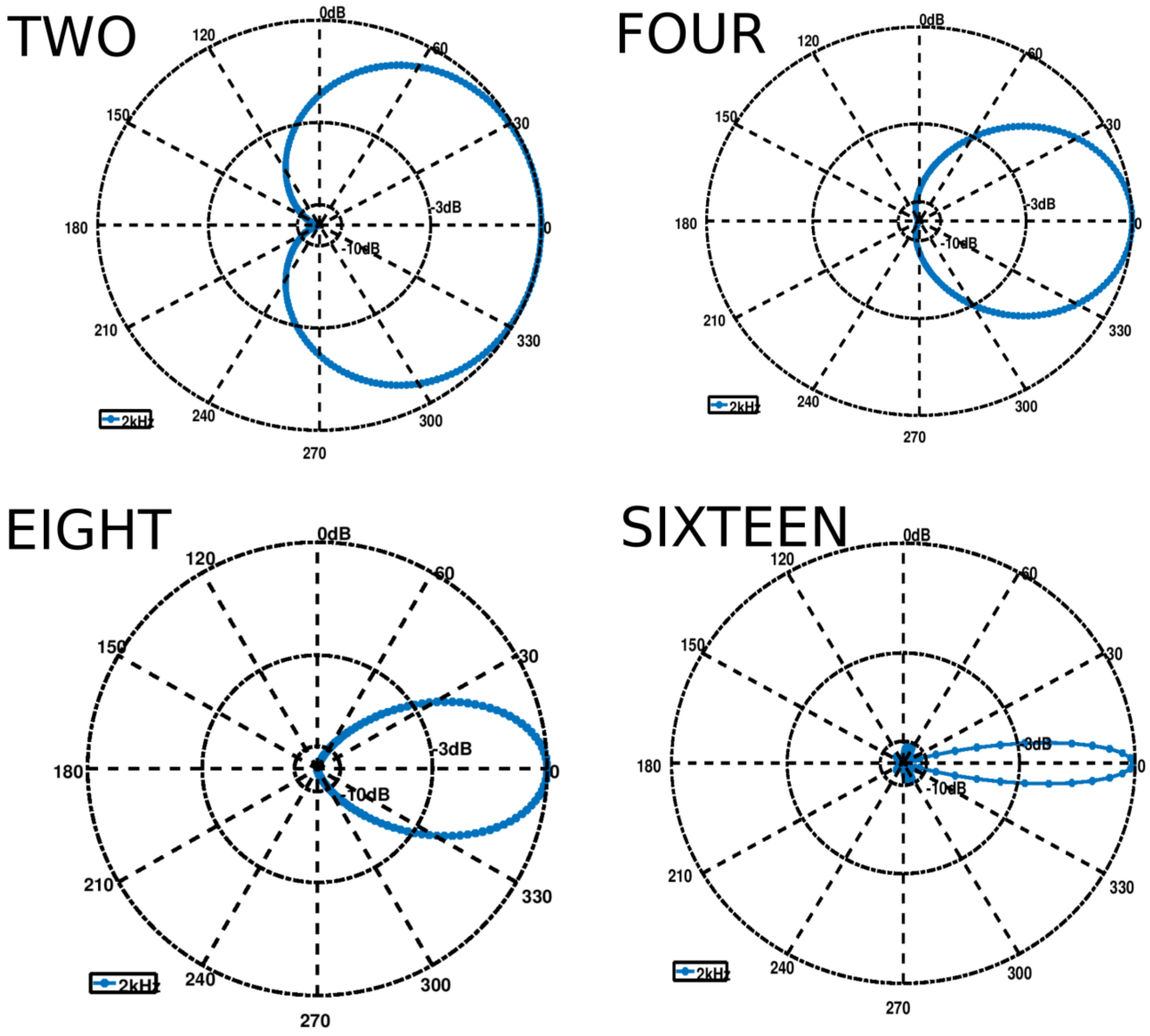
Directional microphones



Microphone array beamforming

E.g., endfire arrays

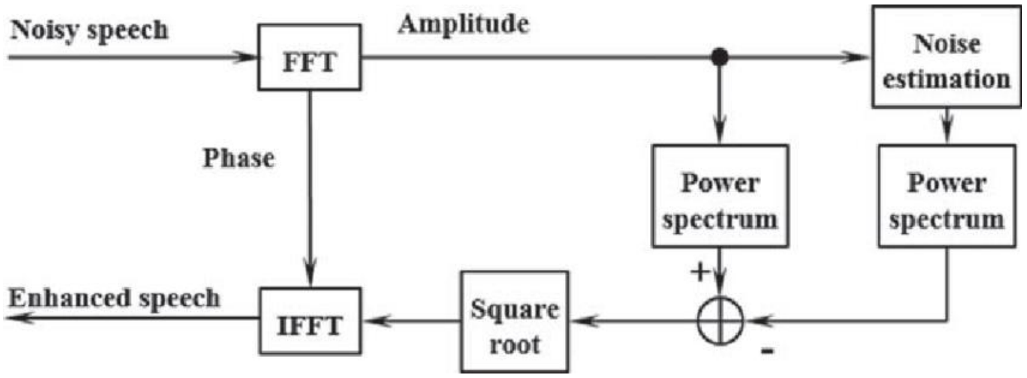




Wide variety of adaptive beamforming and dynamic Wiener filtering (DWF) techniques

De-noising with signal processing

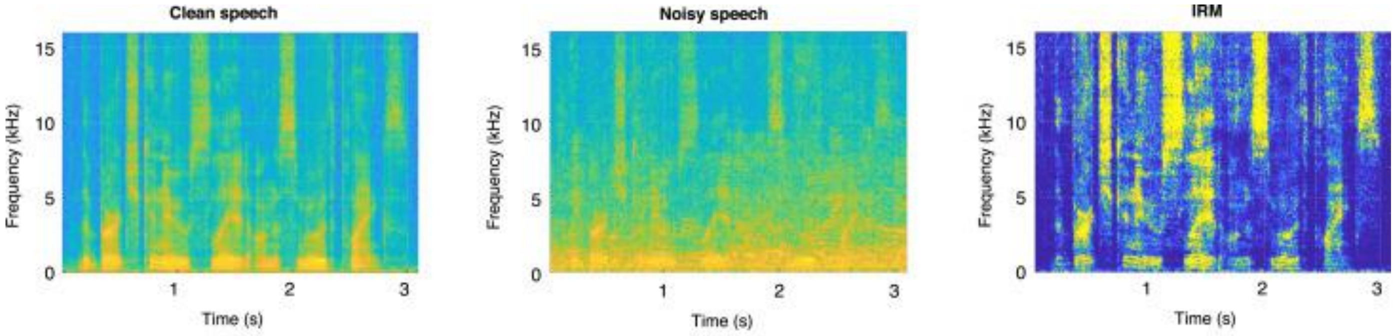
Spectral subtraction



Nonnegative matrix factorization

Adaptive noise reduction

Time-frequency masking



(a) (b) (c)

Speech enhancement with neural networks

Adaptive filtering/beamforming

Network estimates spectral or other characteristics of signal

Time-frequency mask (or gain) estimation

Network estimates gain to apply per time-frequency cell

Auto-encoder

Network models degraded-to-clean transformation

A real example: RNNoise

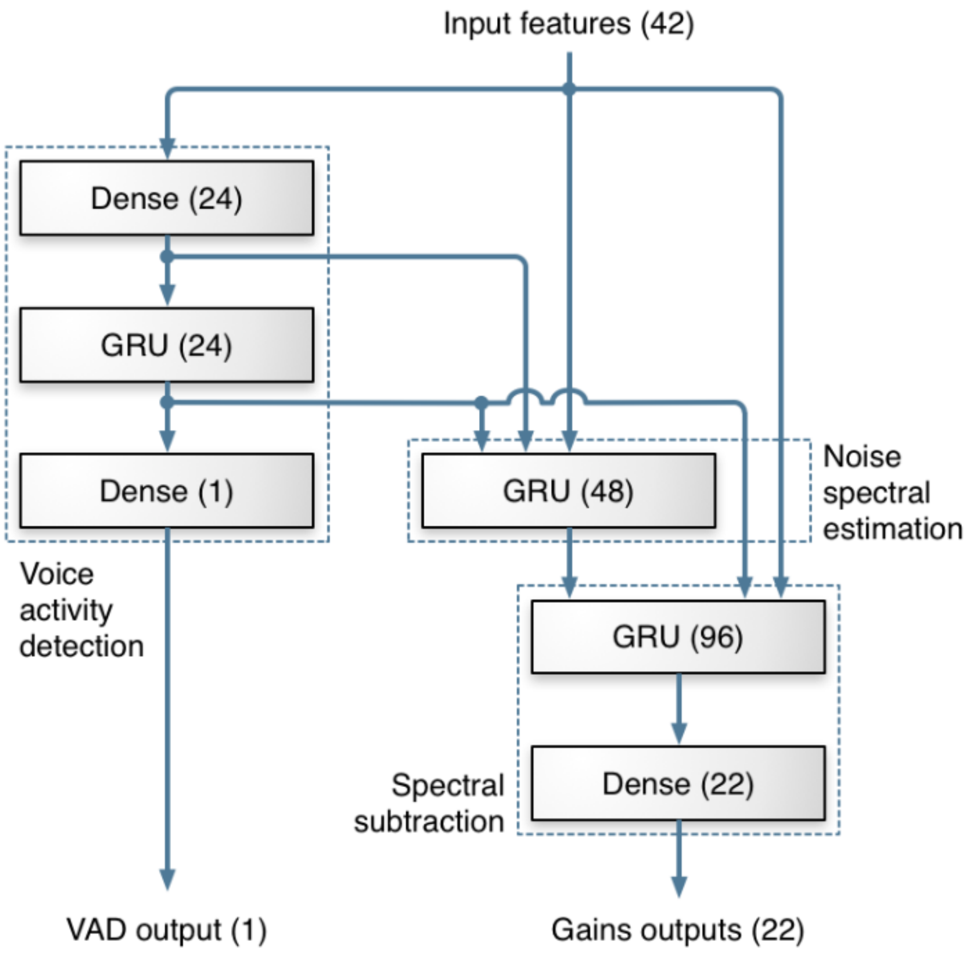
Real-time (causal) speech enhancement / denoising

Inputs:

- Frames of audio, 42 features per frame
- 22 band energy values
- 20 other features

Outputs:

- Ideal (Wiener) gains for each of the 22 bands per frame
- VAD estimate for the frame (auxiliary, just helps training)



J.-M. Valin, A Hybrid DSP/Deep Learning Approach to Real-Time Full-Band Speech Enhancement, International Workshop on Multimedia Signal Processing, 2018

Where we are now

Machine learning (especially deep learning) has completely overrun speech processing research

Promises of deep learning:

- Solves unsolvable problems
- Finds unintuitive solutions
- Removes the need for detailed expertise and handcrafting

Pitfalls of deep learning:

- Behavior difficult to explain/predict
- Too easy to apply (and misapply)
- Blind spots / false confidence / catastrophic failures

Thank you!