Are machines ready to listen?

Part-3 in Symposium on Dimension-based Attention in Learning and Understanding Spoken Language

Neeraj Sharma

Postdoctoral Researcher
Carnegie Mellon University, USA
Indian Institute of Science, INDIA

LEAP Labs

Carnegie Mellon University

A SHORT HISTORY OF THE VOICE REVOLUTION

1961 — IBM introduces the IBM Shoebox, the first digital speech recogntion tool. It recognized 16 words and digits.

1972 — Carnegie Mellon completes Harpy Program. It could understand about 1000 words.

1990 — Dragon launches Dragon Dictate, the first speech recognition product for consumers (for only $6,000).

1996 — Microsoft introduces Clippy

2001 — Microsoft introduces speech recognition feature for Office XP

2011 — **FEBRUARY 18** IBM's Watson wins Jeopardy!

2012 — **APRIL 14** Apple introduces Siri

16 words

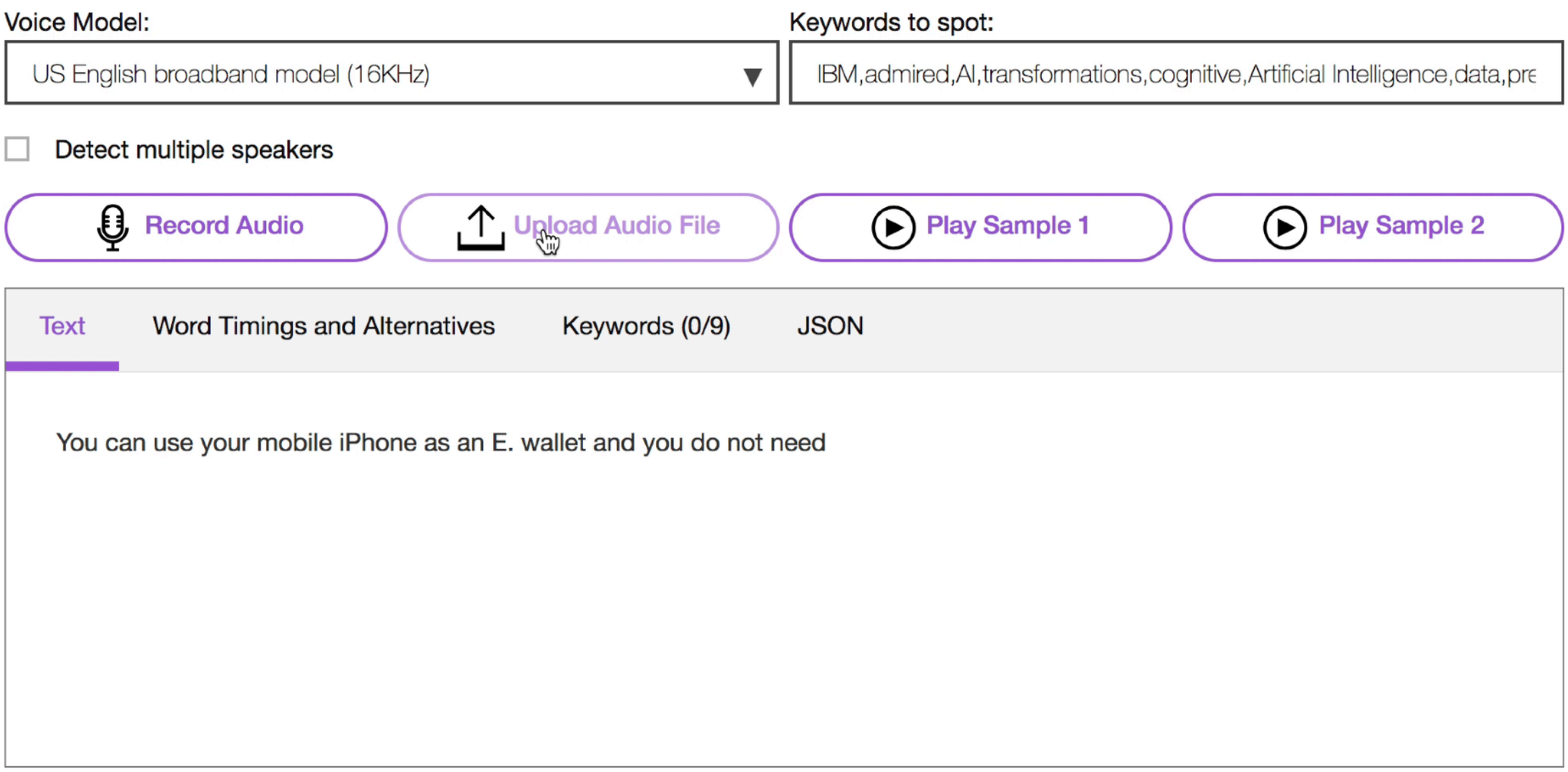
1000 words

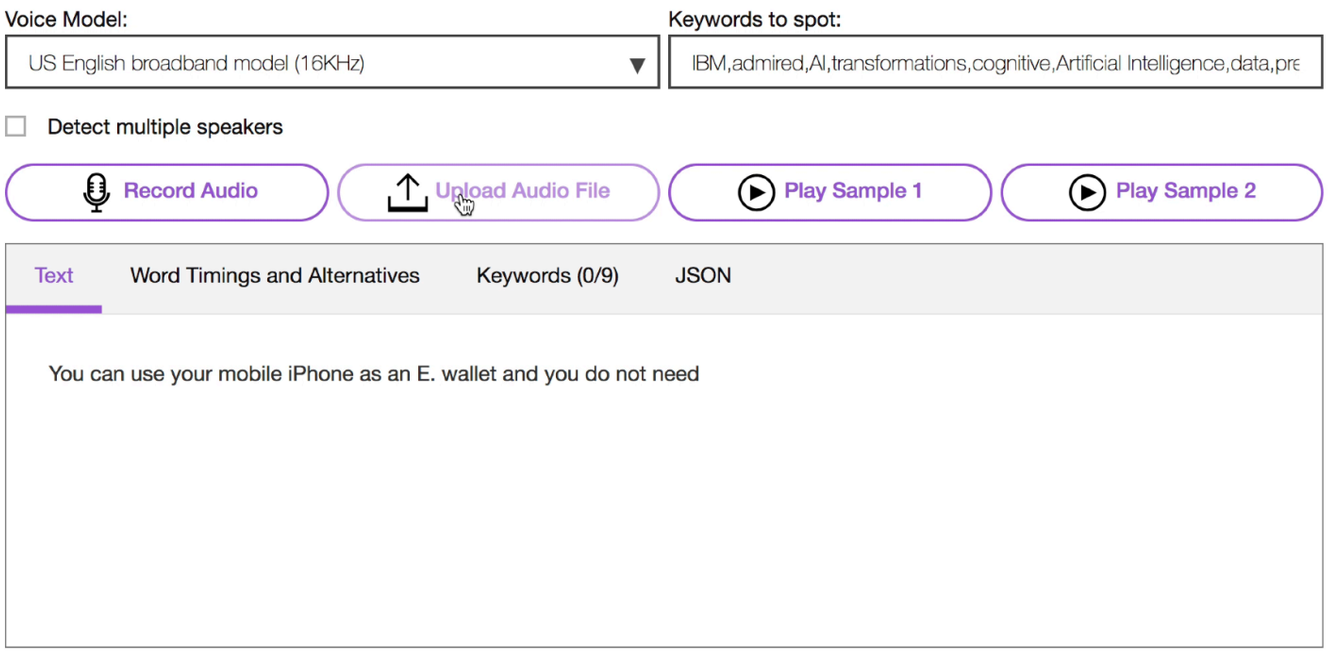
10,000+ words

1M+ words

Smartly.ai

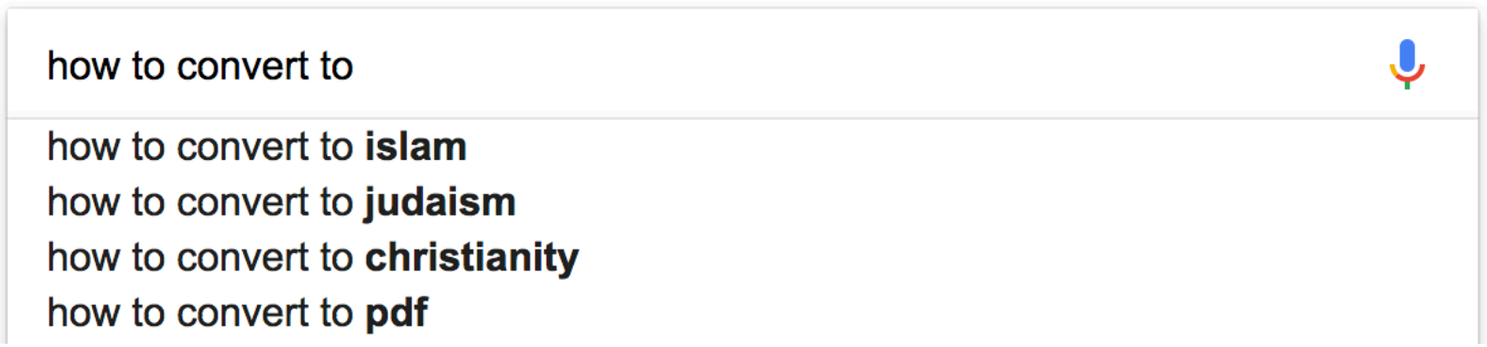
Automatic speech recognition (ASR)…





ASR Output:

You can use your mobile **iPhone** as an E. wallet and you do not need any hi fi smartphone for this because even with the help of your ordinary mobile **iPhone** itself you can make purchases from the neighborhood shops and make payments as well that is why I specially urge **I will work of brothers** and sisters to participate in the scheme because after all I took such a momentous decision for the benefit of the poor people.



When you blindly learn from data!

Performance is impacted by …

Type of Utterance: isolated, continuous speech, conversational speech

Environment: clean versus background noise

Speaker Characteristics: age, accent, rate of speaking, …

Task Specific: constrained versus unconstrained

Quick Fixes for Alexa Voice Profiles

If you're having trouble with your voice profile, here are some solutions that may help.

- If there are any other Echo devices nearby, we recommend temporarily turning off their Microphones (press the **Microphone Off** button) while you complete voice training on your selected device.
- Sit or stand where you typically speak to your Alexa device.
- Make sure there isn't a lot of background noise where and when you're speaking.
- Try to speak like you do normally.
- Make sure the Alexa device you're interacting with is at least eight inches away from walls or other objects.

Improving performance …

`when sunlight strikes raindrops in the air`

$$\text{Word Error Rate (WER)} = \frac{\text{Insertion + Deletion + Substitution}}{\text{Total Number of words}}$$

Increase the training dataset size

Dataset	Specifics
TIMIT	630 speakers x 10 utterances
Wall Street Journal (WSJ)	30 k Vocabulary
Broadcast News	104 hrs
Switchboard	2000 hrs
DeepSpeech	5000 hrs read (Lombard) speech
YouTube	125,000 hrs aligned captions

Training does help in ASR

% of Child	# Utts	Adult	Child
0	2.6M	13.4	11.9
10	2.8M	13.6	11.7
20	3.0M	13.7	11.5
40	3.4M	13.7	11.1
80	4.1M	13.7	10.7
100	4.5M	13.4	10.2

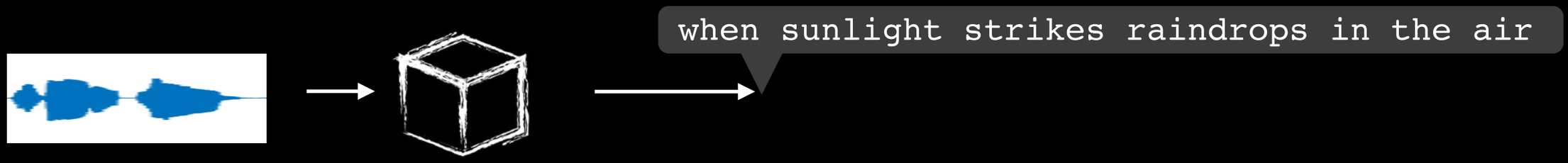
Large Vocabulary Automatic Speech Recognition for Children, Liao et al., Interspeech 2015

Are machines ready to listen?

The short answer is - **Not yet.**

What is the concern?

It may be ready soon by learning from huge datasets.



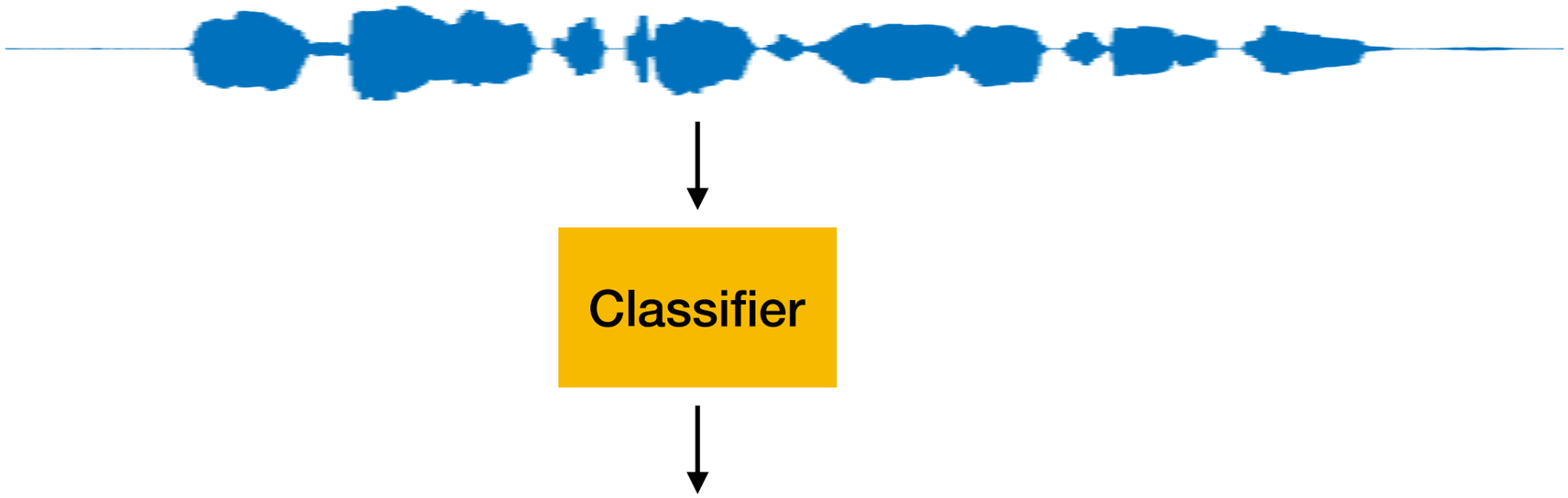
And then we will work on “machine perception” to uncover the “artificial auditory brain”!

What to do?

Analyze it while it is “in the making”

May clarify the unanswered questions in both human and machine perception.

So, lets understand ASR 1.0 ...



`when sunlight strikes raindrops in the air`

What is challenging here?

1. Word boundaries are not well defined.

2. Word duration is variable.

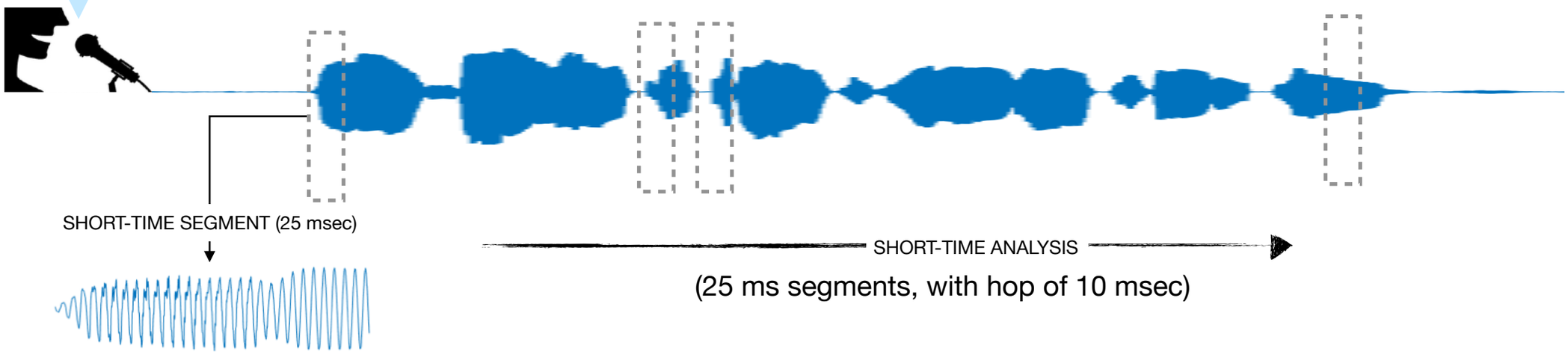
3. Need text transcripts.

4. Present has dependence on past.

when sunlight strikes raindrops in the air

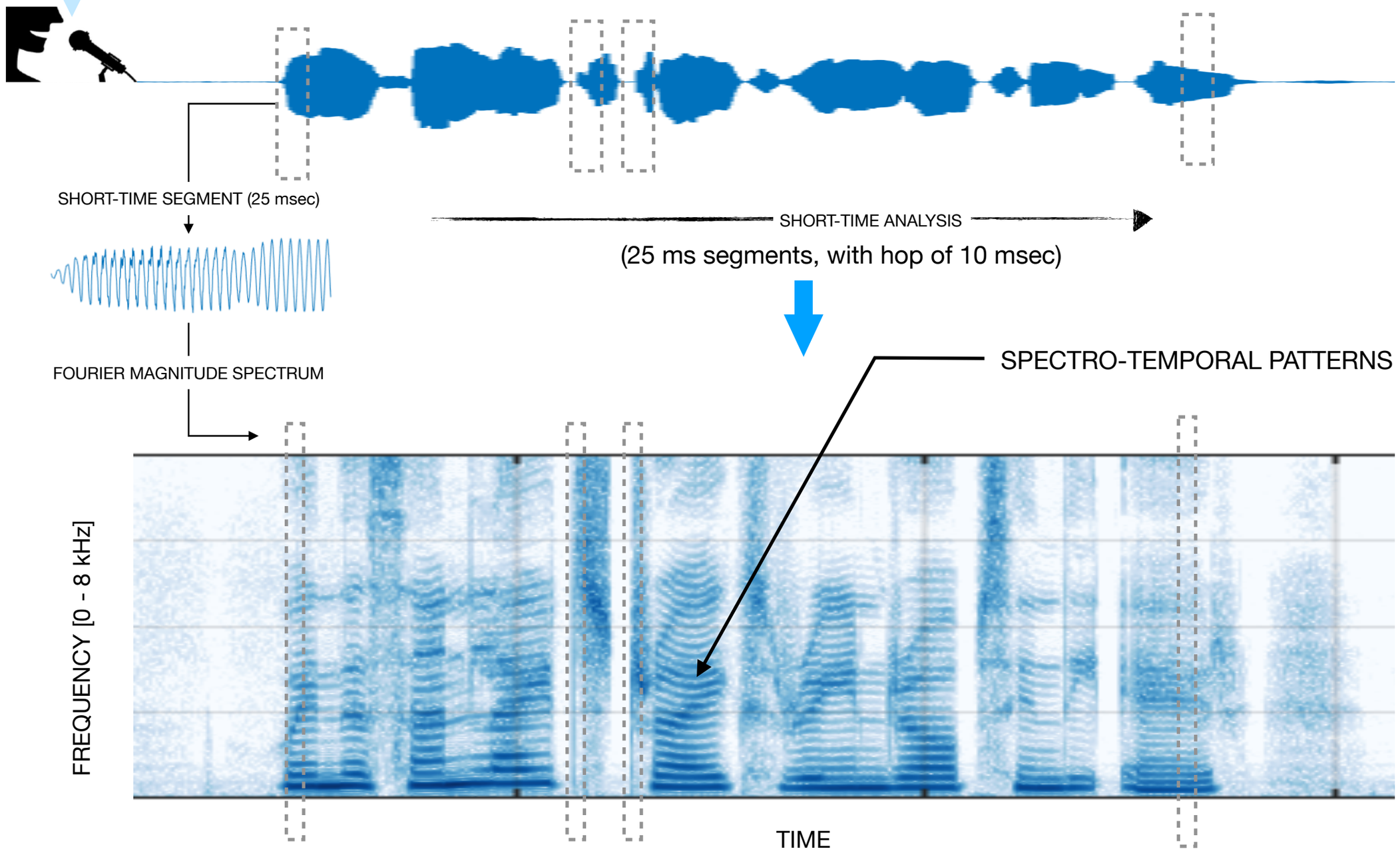
Speech Signal Basics …

when sunlight strikes raindrops in the air



Speech Signal Basics …

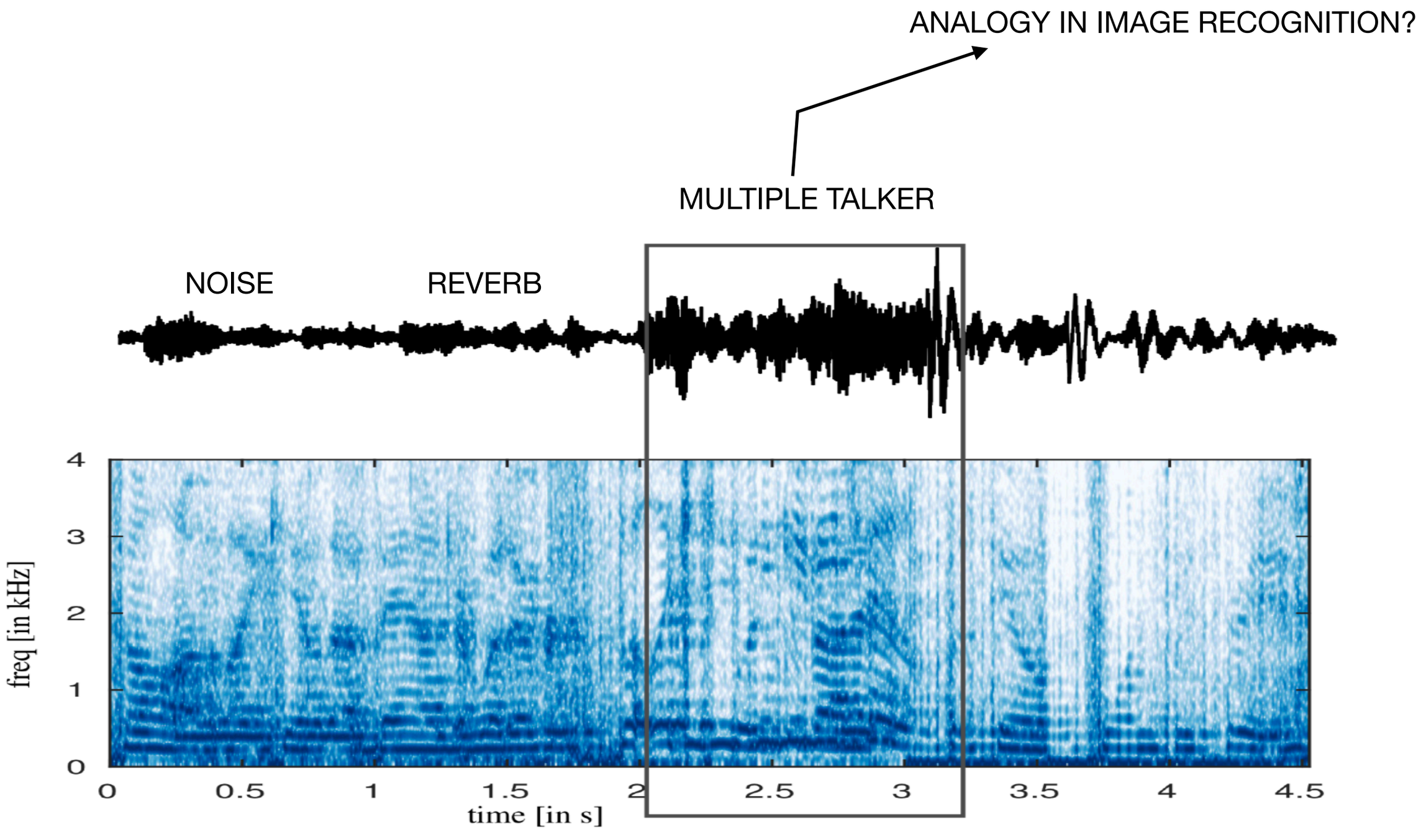
when sunlight strikes raindrops in the air



SPECTROGRAM: A 2-D VISUALIZATION OF SPEECH

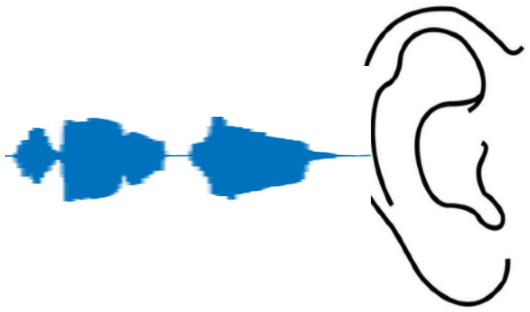
Each column is a FRAME

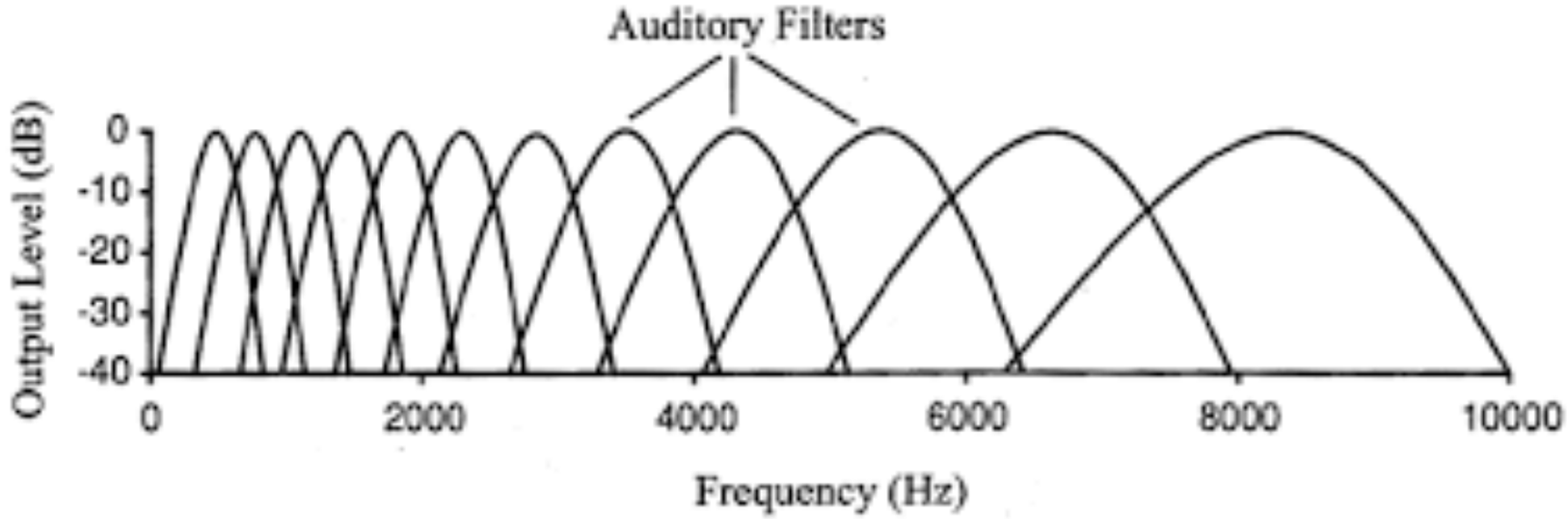
Impact of ambience



Step 1: Feature Extraction

1. Making a **crude approximation** of peripheral auditory system





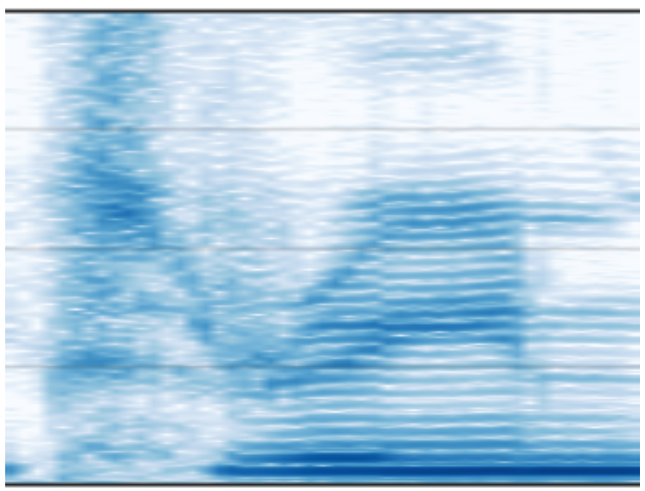
1. Ear is like a filter bank

2. Higher frequency resolution in lower frequencies (narrow filters).

2. Extract features robust to noise

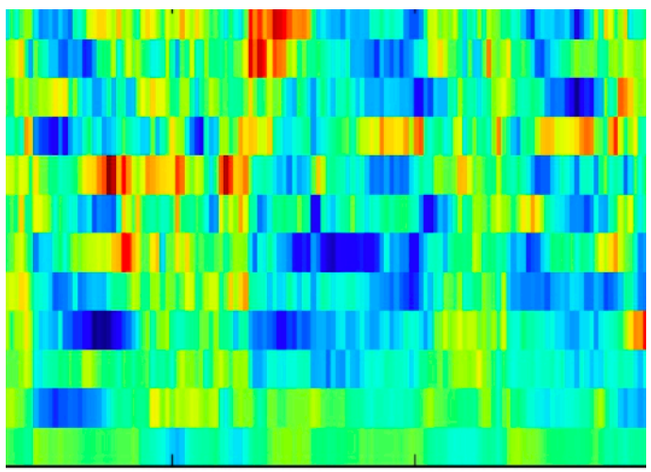
Step 1: Feature Extraction

3. MFCC: Mel-frequency cepstral coefficients (widely used)



[256 x N]

*F*ilter bank + DCT

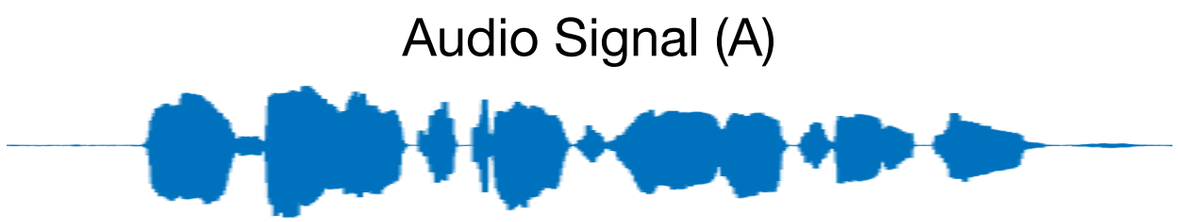


[13 x N]

4. Many other variants exist based on similar concepts

5. Imparts some robustness against noise

Step 2: Modeling the data



`when sunlight strikes raindrops in the air`

Word String (W)

Prior

Likelihood

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{w}} P(\mathbf{W}|\mathbf{A}) = \arg\max_{\mathbf{w}} \frac{P(\mathbf{W})P(\mathbf{A}|\mathbf{W})}{P(\mathbf{A})}$$

Posterior Probability

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{w}} P(\mathbf{W})P(\mathbf{X}|\mathbf{W})$$

Note: A = [X_1 X_2 ….. X_n] = sequence observations (MFCC vectors)

Spoken language modeling … building statistical models

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{w}} P(\mathbf{W})P(\mathbf{X}|\mathbf{W})$$

$P(\mathbf{X}|\mathbf{W})$: Acoustic model (trained with audio + transcription to give words)

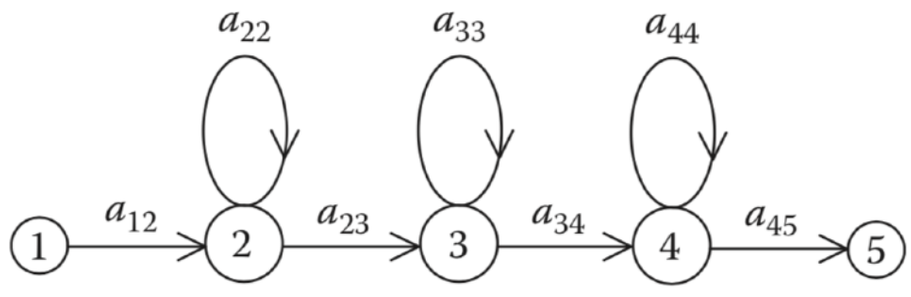
1. Word space is huge in large vocabulary speech.

2. Break word into pronunciation sequence of phones

   `COGNITIVE` ⟶ `K AA G N IH T IH V`

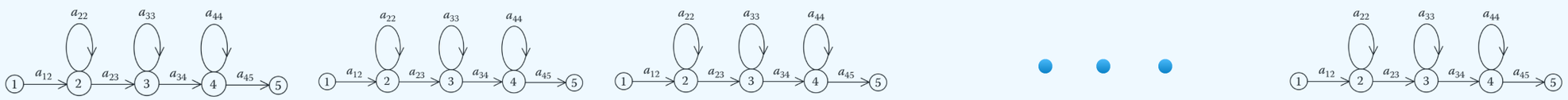
3. Make mono-phone or tri-phone models

4. Modeled using HMMs



/IH/

WORD MODEL P(X|W) “COGNITIVE”



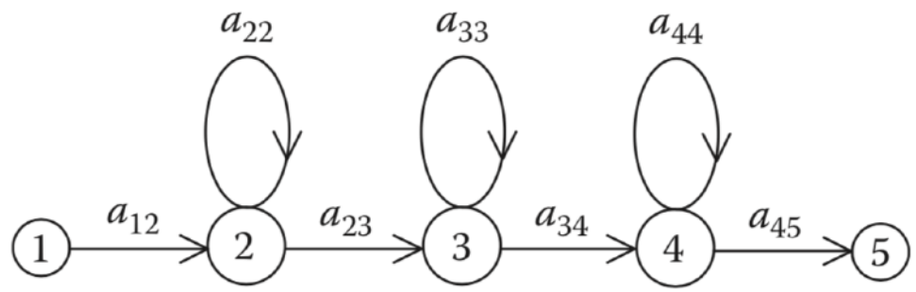
1. For each HMM:

$$a_{ij} = \Pr(s(t) = j | s(t-1) = i)$$

$$\sum_{N} a_{ij} = 1$$

$$b_j(x) = \sum_{m=1}^{M} c_{j,m} N(x; \mu_{jm}, \Sigma_{jm})$$

$$N(x; \mu_{jm}, \Sigma_{jm}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_{jm}|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_{jm})^T \Sigma_{jm}^{-1} (x-\mu_{jm})}$$



1. This is a huge parameter space to learn from data.
2. Requires fast and accurate algorithms (developments from 1972 - 2008)

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{w}} P(\mathbf{W})P(\mathbf{X}|\mathbf{W})$$

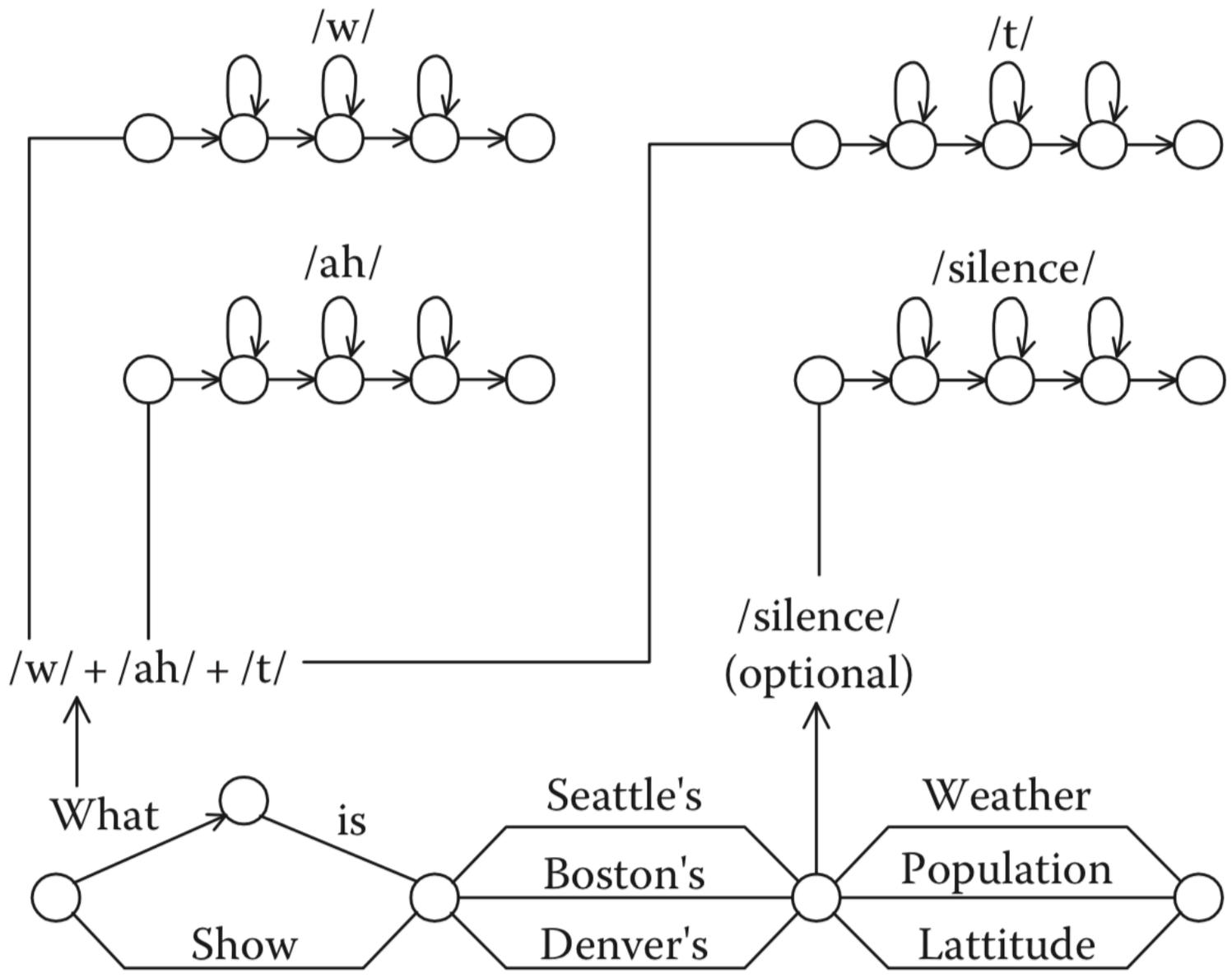
$P(\mathbf{W})$: Language model (trained from huge text corpus)

1. Bi-gram models:

P(word1 word2 word3) = P(sil|word3)P(word3|word2)P(word2|word1)P(word1|sil)

2. Needs a huge text corpus to learn the bi-gram/n-gram probabilities

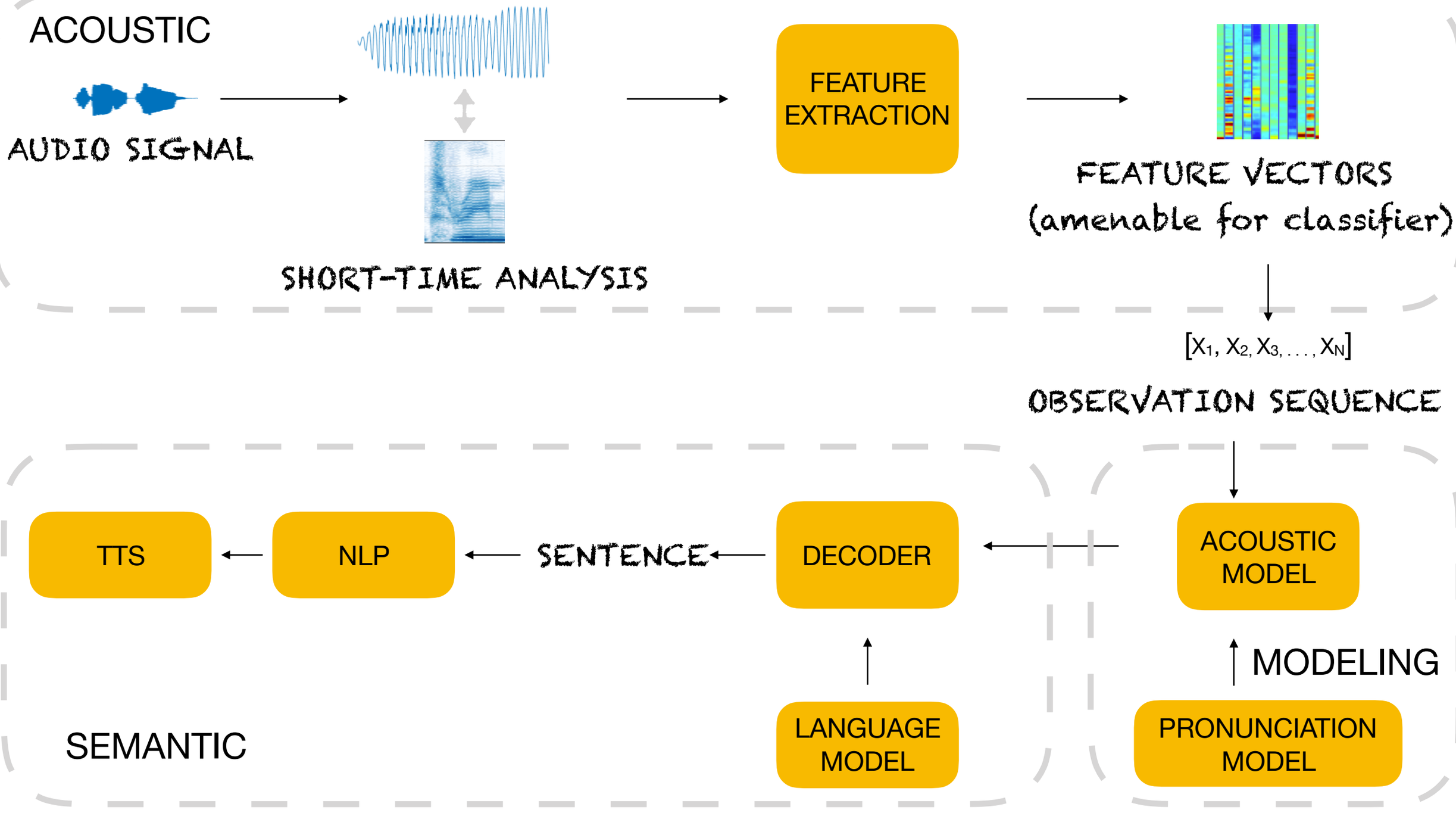
Step 3: Recognition - Decoding in the huge space



From: An overview of Modern Speech Recognition, Huang and Deng

Concepts from finite state machines, lattice search, etc..

ASR: Step 1 - Step 2 - Step 3



ASR: Step 1 - Step 2 - Step 3

1. How to make ASR adapt to new accent, language, ambience?
    a. Adapting the parameter to speaker data … a number of techniques exist …. But don't generalize
    b. Data augmentation via speech synthesis

2. Are we in the correct perceptual space to build speech models?
    a. Psychoacoustic studies generalizable to natural listening

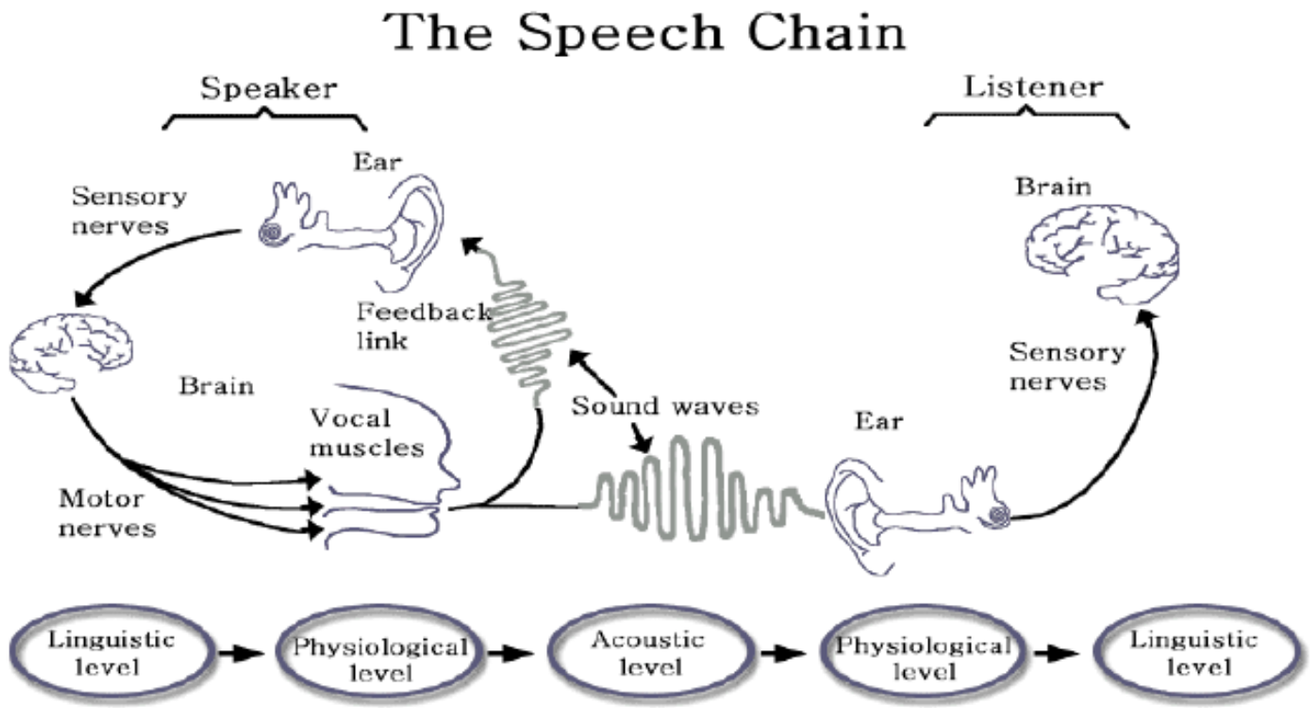
From humans perspective …

The Bell System Technical Journal

Vol. XIX *October, 1940* *No. 4*

The Carrier Nature of Speech

By HOMER DUDLEY



The Speech Chain: The Physics and Biology of Spoken Language, Denes and Pinson, 1963

We speak to be understood…hence, just fit the data. Not really, Speech has multiple cues.

Thank you!

Neeraj Sharma

web: www.neerajww.github.io