University of Illinois at Urbana-Champaign
Dept. of Electrical and Computer Engineering

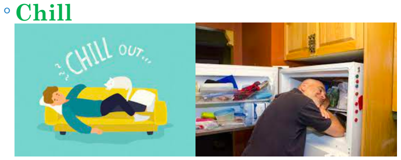
ECE 101: Exploring Digital Information Technologies for Non-Engineers

Speech and Natural Language

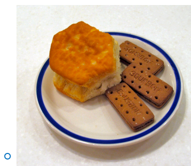
Today's Topic: Speech and Language

- **Torch vs. Flashlight**

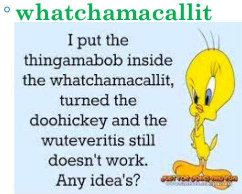
- **Chill**



- **Biscuit vs. Biscuit**



- **whatchamacallit**



What are the Steps in Our Cycle?

Let's start by thinking about the cycle:
sense, compute, actuate, communicate.

Consider a voice-controlled smart home (or an online assistant).

What is being sensed when interacting with the homeowner?

A human voice, captured by a mic.

Computing Requires Several Steps

Consider a voice-controlled smart home (or an online assistant).

Sense: human voice

What is computed?

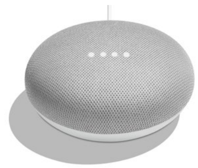
Here, we may **need several steps**.

Computing Steps in Voice Response

Consider a voice-controlled smart home (or an online assistant).

Computing:

1. **get rid of noise**: other voices, music, television, video games, pets, and so forth.
2. **perform "voice recognition"**: translate an audio signal into a sequence of words.
3. **understand** what the human is trying to communicate: process their **natural language** (English, for example).



Noise will Always Impair the Process

Step 1: get rid of noise.

This task is becoming feasible, but controlled environments are always easier, and results better.

In 2006, IBM transcribed news from Al Jazeera: formal tone, little/no background noise, intended for clarity, prosaic content—not poems (I asked! No such luck at that point).

THE IBM 2006 GALE ARABIC ASR SYSTEM

Hagen Soltau, George Saon,
Daniel Povey, Lidia Mangu, Brian Kingsbury, Jeff Kuo, Mohamed Omar and Geoffrey Zweig

IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598
e-mail: {hsoltau,gsaon}@us.ibm.com

using a super-computer

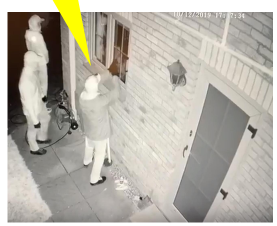
شبكة الجزيرة الإعلامية
ALJAZEERA MEDIA NETWORK

Unauthorized Voices Can be Treated as Noise

Noise today may also include unauthorized voices.

By parametrizing the range of frequencies, speeds, and accents for human speech in a given language,

- modern systems are able to record a voiceprint (a set of parameter values) and
- verify that the speaker is authorized to make use of the system.



Voice Recognition Success Depends Strongly on Context

Step 2: voice recognition

The context matters here, too.

Recognizing "zero" to "nine" in clear, crisp, and unaccented speech has been **possible for decades**.

Understanding a **non-native speaker** who mispronounces words and abuses grammar **on an unknown topic is** still **years away** on edge devices.

Variations of Speech Affect Success

Success depends on several aspects…
- **How many words** in the vocabulary?
- Do speakers **need to enunciate clearly** (Didja getdat?)?
- Are **euphemisms and idioms** allowed ("passed away" instead of "died")?
- How **precise**ly must speakers use **grammar**?
- Are **different accents** handled?

It's **easier to provide support for multiple languages** than to understand the vast number of pidgin languages that humans develop spontaneously as they learn new languages.

Hierarchical Models Share Information

Modern voice recognition uses a **hierarchy of interacting, probabilistic models**.

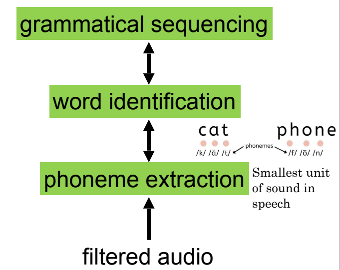
In the 90s and 00s, these systems made rapid progress.

The secret?

Quantify success to enable competition!

For example, **sequences of three words** transcribed correctly.

Machine learning is now **used to solve** specific **sub-problems**.



Processing from the Nouns Up

Step 3: natural language processing (NLP)—understanding what the human meant

The most basic form is keyword search.

What does a human want to see when they type "Ukraine" into Google?

Interrogative Adverbs Add Clarity … Sometimes

What if we start to add grammatical elements?

what Ukraine

where Ukraine

when Ukraine

how Ukraine

why Ukraine

Even Full Sentences May be Ambiguous in Meaning

Or more complete questions…

Why is Putin in Ukraine?

[Do they mean the Russian army? A simple answer or a feasible explanation of the rationale?]

How long has Russia been in Ukraine?

[Again, the army? Or a history of the Soviet Union? Or an older history?]

Understanding human sentences is pretty hard.

IBM Watson: Jeopardy Champion through Web Crawling

In 2011, **IBM Watson**
- became the **world champion of Jeopardy**,
- a game in which a host gives an answer to a question of the form, "**What is X?**"

Example: "To marry Elizabeth, Prince Philip had to renounce claims to this southern European country's crown."

The question? "What is Greece?"

To compete, **Watson crawled the web and built a knowledge base** from which it could draw answers.

using a super-computer

Natural Language Models are Complex and Expensive

Natural language processing today uses a combination of probabilistic inference and machine learning.

One study* estimated that training a modern NLP model releases as much carbon as manufacturing and using five cars for their entire lifetimes.

*E. Strubell, A. Ganesh, A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," *57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Let's Learn Some Probability Tools

To better understand the ideas in voice recognition and NLP, let's talk about using probabilities to make good guesses.

Probabilities are always more fun as games…

Rules for 20 Questions

Have you played 20 Questions?

- One person picks a "thing."
- Others ask twenty yes/no questions
  (the first is allowed to have three answers:
  animal, vegetable, or mineral?).
- First person to guess the "thing" wins
  ("Is it a 'thing?' " Yes!).
- Picker wins if no one guesses
  within 20 questions.

Here's a Sample Game

Question 1: Animal, vegetable, or mineral?

Answer: Animal.

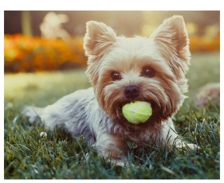
Question 2: Is it bigger than a dog?

????? Which dog ?????



Need Intuition about Dogs to Answer the Question!

To answer, we have to think …
- What's the typical size of a dog?
- What's the typical size of a thing?
- How likely is a thing to be bigger than a dog?



A Similar Question of Imagination

Similarly, say I tell you,
"The dog knocked over the child."

In your imagination, how big is the dog?

(Perhaps you want to know the child's age?)

Our Brains Use Probability … Minds … Maybe Not

How do we come up with probabilities based on observed facts?

Humans are generally pretty **bad**
- **at** reasoning consciously about **probability**
- AND at using probability subconsciously.

But **our brains are** reasonably **good at using probability** unconsciously **for language**.

Estimate Highly Biased by Experience

What if we ask two people:
how big is an average dog?
- Pat, who grew up in an apartment in downtown Chicago, and
- Jan, who grew up on a farm?

Pat will probably give a smaller size than Jan.

Why?

Their experience with dogs is likely to differ.

Using Probabilities in Reverse Makes No Sense

In many problems, however, we must **estimate values based on observations**.

Probabilities are **not invertable**:
- if I tell you that I flipped a coin,
- and the result was “tails,”*
- **what can you say about my coin?**

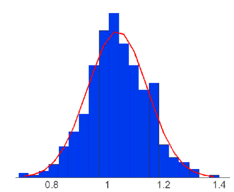
Only that one side is marked as “tails”—not two “heads”.

*For those who grew up without coins: “heads” means the side with a person’s head, and “tails” means the other side. Most coins in most countries allow this distinction.

MLE: Explanation Most Likely to Lead to Observation

To address this issue,
- people often use a technique called
- **maximum likelihood estimation** (MLE).

Given an observation,
- **choose** the **explanation** that is
- **most likely to produce** the **observation.**



Applying MLE in Casino: Watch and Learn

In casinos, for example…

- **people think** that slot machines pay at different rates.
- **One machine may pay more** money more **often** than another.
- So they **stand and watch** other people play.

If one machine pays **10 times out of 100** plays, and a second machine pays **5 times out of 100** plays, the person then sits down at the first machine.

Likelihood Used to Estimate Win Probabilities

Why?

They are applying primitive MLE!

Assume each play is random, but

- first machine pays with probability $\mathbf{P_1}$, and
- second machine pays with probability $\mathbf{P_2}$.

If one sees 100 plays on a machine,

- and the machine pays **N** times,
- probability **N/100** is most likely for that machine.

Compare the Frequency of Payouts to Pick a Machine

If first machine pays **X** times, $\mathbf{P_1 = X/100}$.

If second machine pays **Y** times, $\mathbf{P_2 = Y/100}$.

So **X > Y implies $\mathbf{P_1}$ is probably > than $\mathbf{P_2}$**!

Most gamblers

- couldn't explain why at this level of detail
- let alone prove the MLE claims.

Here's a Easy Game to Play

Let's think about another game.

Pat will roll either

- one (six-sided) die or
- two dice and add up the numbers.

Then Pat tells us the amount rolled.

Can we guess whether Pat rolled one or two dice?

Some Cases are Easy, but Others are Hard

Some cases are easy.

For example, Pat rolled an 11. **One or two dice?**

Pat rolled a 1. **One or two dice?**

Other cases are harder…

Pat rolled a 4. **One or two dice?**

Calculate the Chance of a 4 for Each Choice

Let's imagine that Pat rolled one die.

What is the chance that Pat rolled a 4?

1 in 6

Now imagine that Pat rolled two dice.

What is the chance that Pat rolled a 4 (total)?

1+3, 2+2, or 3+1

3 in 36 (same as 1 in 12)

Choice Most Likely to Report 4 is the Best!

With maximum likelihood estimation,
- **we choose "one die"** because
- **probability (if Pat rolls one die, Pat gets a 4)**

  >

  probability (if Pat rolls two dice, Pat gets a 4).

But there's a tricky point.

What does "if Pat rolls one die" mean?

Conditional Probabilities: Chances in Specific Conditions

"If Pat rolls one die" is a **condition**.

In math and engineering,
- we call such probabilities
- **conditional probabilities**
- and we write them this way:

probability (get a 4 | Pat rolls one die)

The meaning is the same:

if Pat rolls one die, Pat gets a 4.

Did We Compute the Wrong Values?

But that's NOT what we wanted to know!

We **wanted to compare**

probability (Pat rolled one die AND got a 4)

with

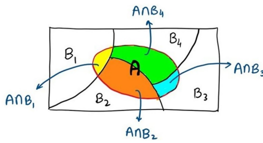
probability (Pat rolled two dice AND got a 4)

What can we do?

Bayes' Theorem to the Rescue

Fortunately, we can make use of a famous fact about probability called **Bayes' Theorem**:

probability (A AND B) =
probability (A) · probability (B | A)



The chance of A and B both happening is equal to the product of the chance of A happening and the chance of B happening if A has happened.

Apply Bayes' Theorem to Find Our Answer

So to find

probability (Pat rolled one die AND got a 4),

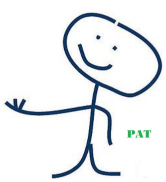
we compute

probability (Pat rolled one die) ·
probability (got a 4 | Pat rolled one die)

We know the second number: 1/6

But how can we know how Pat makes decisions?

We can't. Pat is a fictional character!



Assume Equal Chance of Both Options

In such cases, we often **assume** that **all** such **events** are **equally likely**.

It's a dumb assumption.

But what else can we do?

In that case, our earlier comparison makes sense

½ · probability (got a 4 | Pat rolled one die)
= ½ · 1/6 = 1/12

>

½ · probability (got a 4 | Pat rolled two dice)
= ½ · 1/12 = 1/24

Initial Probabilities are Important to Correct Choices

What **if Pat tells us** that

probability (Pat rolls one die) = ¼ and

probability (Pat rolls two dice) = ¾ ?

In that case, **our guess changes**, as

¼ · probability (got a 4 | Pat rolled one die)
= ¼ · 1/6 = 1/24

<

¾ · probability (got a 4 | Pat rolled two dice)
= ¾ · 1/12 = 1/16

Recognizing Digits Also Uses MLE

One can also interpret systems that we've already seen as examples of MLE:

Given a picture of a digit, which digit most likely produced the picture?

And context (initial probabilities) DOES matter.

What is this number? 

And when it's in context?

The student collapsed, so we called 411.

MLE Solves the Voice Recognition Problem

How is MLE useful in speech recognition?

Voice recognition answers the question, **"Given an audio input, what sequence of words was spoken?"**

A solution is generated by finding the sequence of words that is most likely to have generated the audio input.

(Our brains are good with this question.)

MLE Solves the Natural Language Problem

How is MLE useful in NLP?

Natural language processing answers the question, "**Given a sequence of words, what did the speaker want to communicate?**"

A solution is generated by finding the meaning that is most likely to have generated the sequence of words.

(Our brains are also good with this question.)

Guessing Words Easier with Words on Both Sides

Understanding how words fit together is an important element. For example, you hear,

"I took my … [more words]."

What's the next word?

In 1953, journalists* realized that **the words AFTER the missing would help** in guessing.

Everyone else already knew.

*W. L. Taylor, "Cloze procedure: A new tool for measuring readability," *Journalism Bulletin*, 30(4):415–433, 1953.

Google Applied Bidirectional Idea to Create BERT

"I took my … for a walk."

What's the next word?

Dog, perhaps?

Could be lots of words, but dog may be a good MLE choice.

"Cat," "snake," "Ferrari," maybe not so good.

In 2019, Google* realized the same thing, and natural language processing changed forever.

*Everyone except engineers, I meant: J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL, 2019. To their credit, the authors DID mention the journalists.

Making Use of Big Language Models

Models like BERT*
- can be connected to task-specific networks
- then either used directly
- or fine-tuned to the specific task.

*"Bidirectional Encoder Representations from Transformers" … yeah, sure. See the picture.

Examples of NLP Applications

Classification: How much did a reviewer like a movie? What sentiment did they express? Did they mention or suggest any movie categories?

Interpretation: Is anything in a patient's electronic medical records relevant to the patient's current symptoms? Which sentence in a page of text answers a particular question?

Organization: How are documents and terms related? Which documents are relevant to a given question of interest?

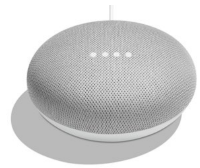
The Last Phase: Communication

Let's close the loop by returning to the cycle:
sense, compute, actuate, communicate.

Once a smart home unit has understood a human and performed any necessary actions, it **needs to respond verbally**.

The process is similar and uses similar models:

1. **Convert** the response **into an intelligible sequence of words** in the speaker's language.
2. **Convert** the words **into** an audio output, **a synthetic voice**.



Voice Synthesis Allows a "Human" Response

The last step
- is called **voice synthesis**, or text to speech,
- **generation of human voice from text**.

The "voice" can be parametrized
- and thus **tuned to the listener's preferences**
- or to match their verbal style and accent.

Synthesis is also **useful**
- **for entertainment and accessibility**,
- such as reading aloud for the vision-impaired or while humans are busy with other tasks.

Terminology You Should Know from These Slides

- voice/speech recognition
- Natural Language Processing (NLP)
- Maximum Likelihood Estimation (MLE)
- conditional probability
- Bayes' Theorem
- BERT (the bidirectional language model)

Concepts You Should Know from These Slides

- steps computation: audio → noise removal → word sequence → meaning
- sources of noise
- challenging aspects of speech recognition
- hierarchy of models for speech: phonemes, words, and grammar
- impact of human experience on probabilistic "reasoning"
- how MLE can be used to solve problems
- bidirectionality of natural languages