2. How humans speak

The Monster Text to Speech & Voice Cloning Course

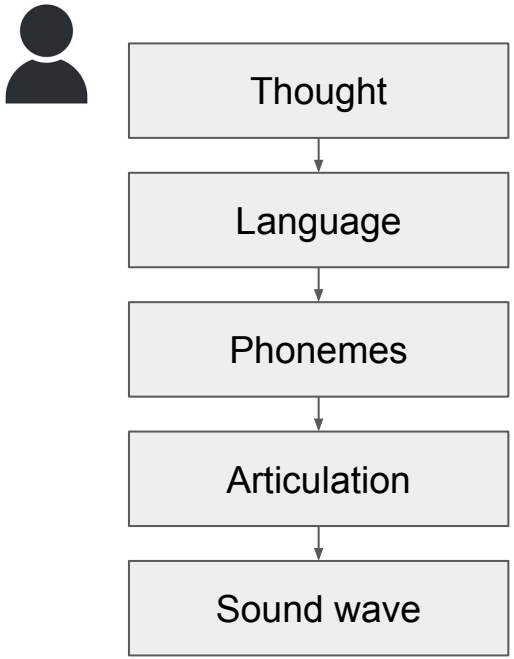
THE SOUND OF AI

Before we can make machines talk, we need to understand what talking is.

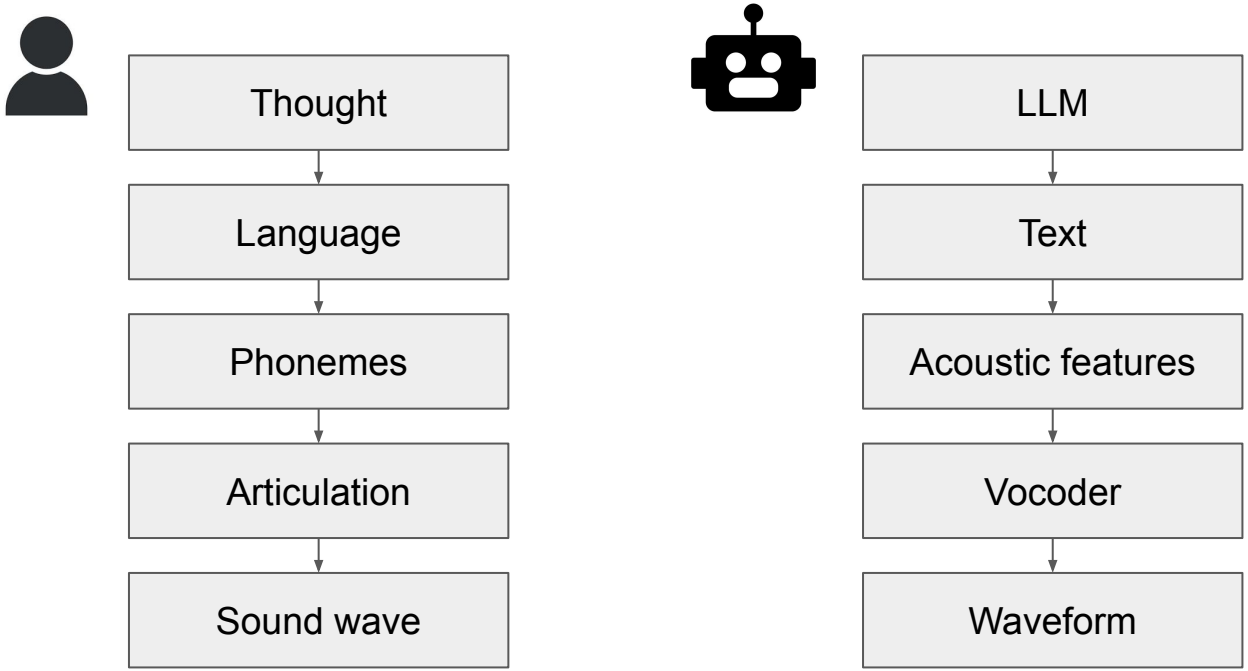
What makes speech special

- Information-dense
- Multidimensional: meaning, language, sound, emotion, and identity

Speech pipeline



Speech pipeline



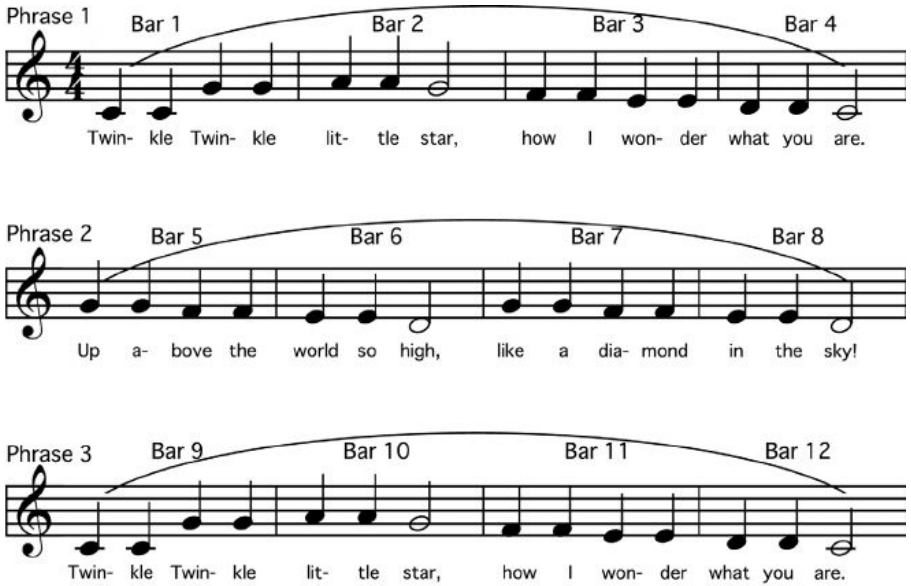
Thoughts to words (language)

- Speech is planned before it’s spoken
- Encode semantics + syntax
  - Semantic layer: what we want to say
  - Syntactic layer: how to say it
- Corresponds to text in TTS

Words to sounds (phonemes)

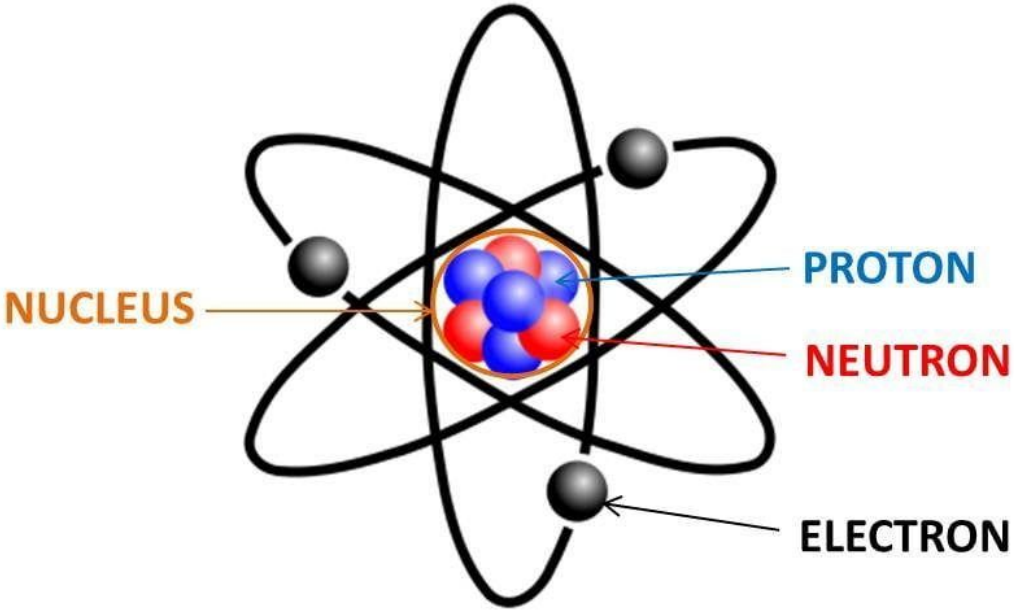
- Phoneme = atom of speech
- Vowels + consonants

Understanding phonemes

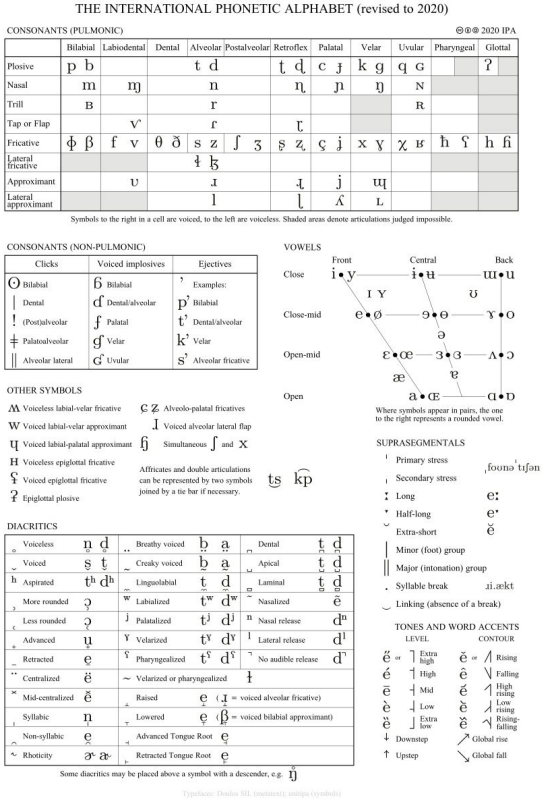


- Word = musical phrase
- Phoneme = note

Understanding phonemes



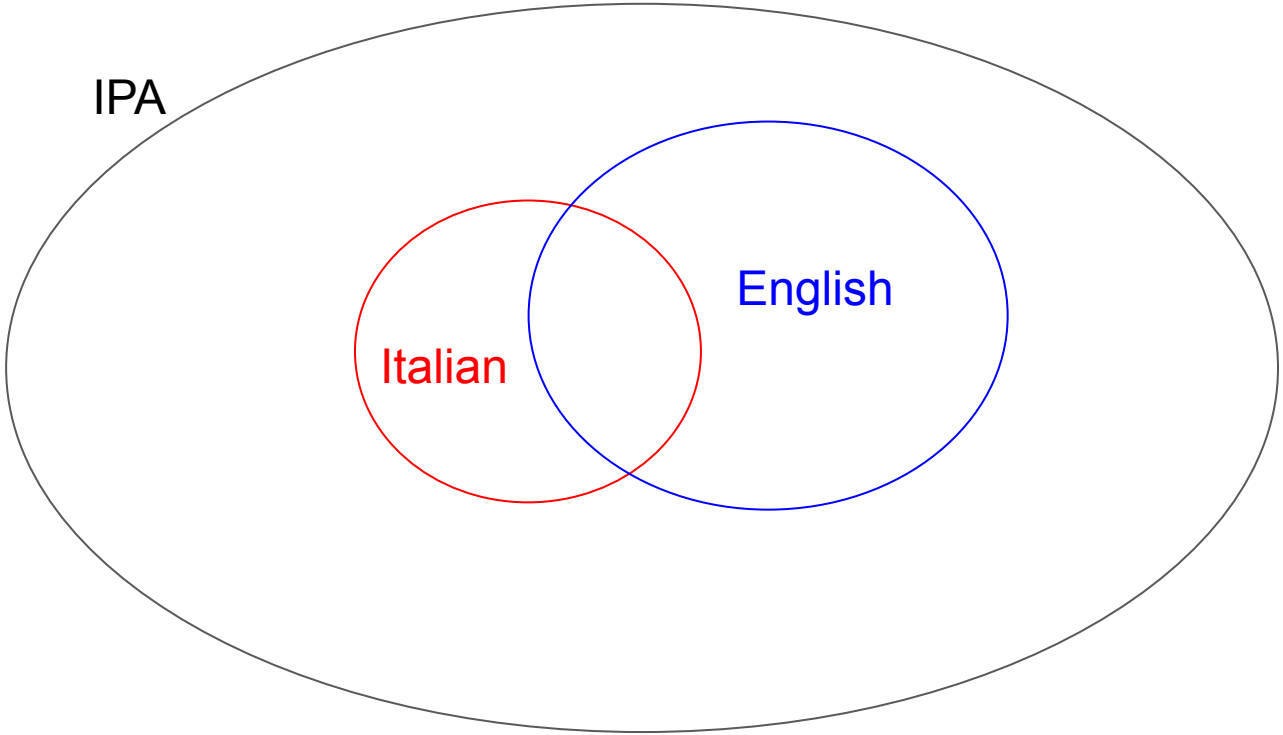
International Phonetic Alphabet



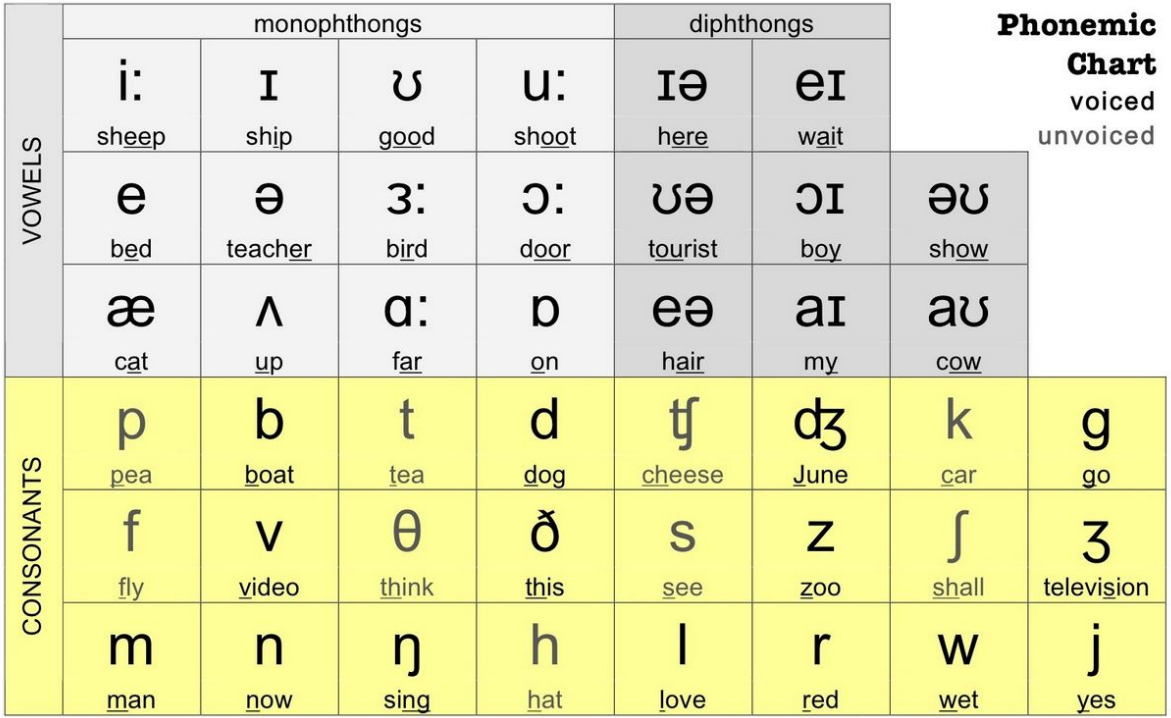
- Collection of all possible sounds
- 100-150 distinct base symbols
- Modifiers (diacritics)
- Hundreds of distinct sounds

THE SOUND OF AI

Language phoneme inventories



English phonetic chart



The 44 phonemes of Received Pronunciation based on the popular Adrian Underhill layout

adapted by EnglishClub.com

THE SOUND OF AI

Phonetic transcription

She bought two red apples

/ʃi bɔːt tuː red ˈæplz/

Coarticulation

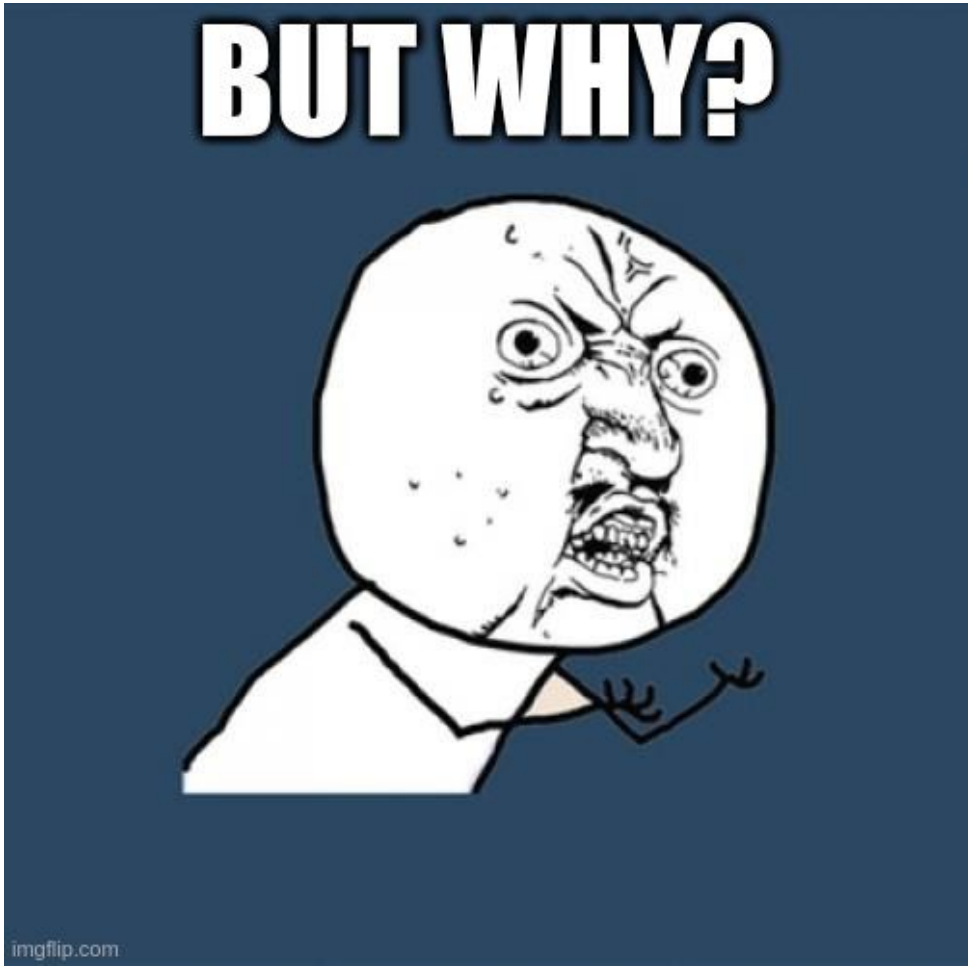
- Phonemes appear together

Coarticulation

- Phonemes appear together
- Idea: concatenate pre-recorded phonemes to generate speech

Coarticulation

- Phonemes appear together
- Idea: concatenate pre-recorded phonemes to generate speech
- Problem: sounds robotic


BUT WHY?
imgflip.com

Speech sound changes based on context

Prosody

- Rhythm + pitch + stress + intonation
- Encodes both linguistic structure and emotion
- Prosody = music of speech

Spot the difference in prosody

“We’re going.”

“We’re going?”

Spot the difference in prosody

“We’re going.”

“We’re going?”

Timbre and voice identity

- Timbre is the “color” or quality of a voice

Timbre and voice identity

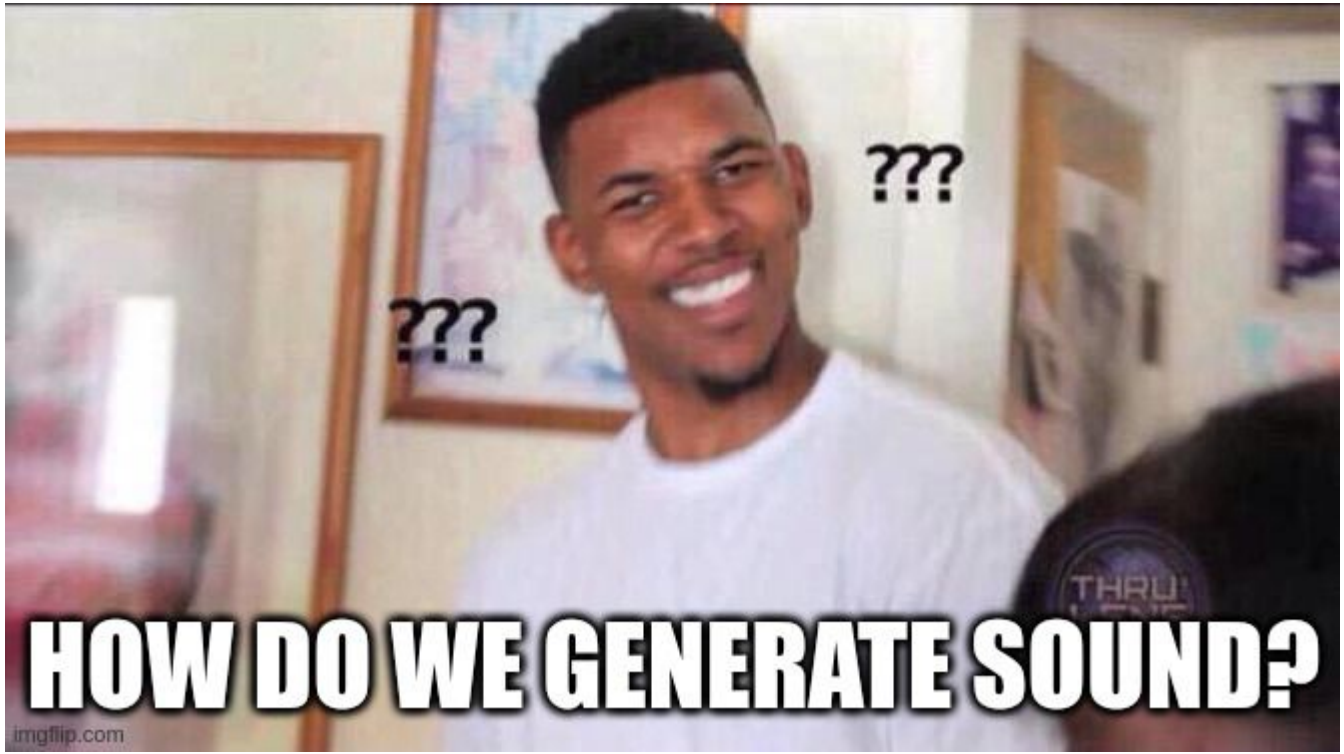
- Timbre is the “color” or quality of a voice
- Why every person sounds unique:
  - Vocal tract shape
  - Resonance of the nasal and oral cavities
  - Habitual pitch, age, health

Timbre and voice identity

- Timbre is the “color” or quality of a voice
- Why every person sounds unique:
  - Vocal tract shape
  - Resonance of the nasal and oral cavities
  - Habitual pitch, age, health
- Timbre = fingerprint of a speaker

Timbre and voice identity

- Timbre is the “color” or quality of a voice
- Why every person sounds unique:
  - Vocal tract shape
  - Resonance of the nasal and oral cavities
  - Habitual pitch, age, health
- Timbre = fingerprint of a speaker
- Voice cloning models try to capture timbre

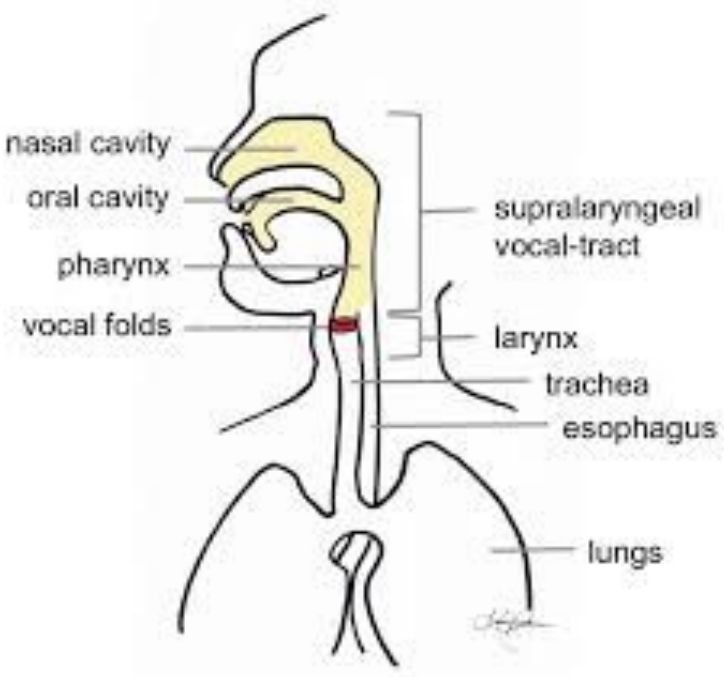

???
???
HOW DO WE GENERATE SOUND?
imgflip.com

Source-filter model

- Vocal folds = vibration source
- Vocal tract = resonant filter

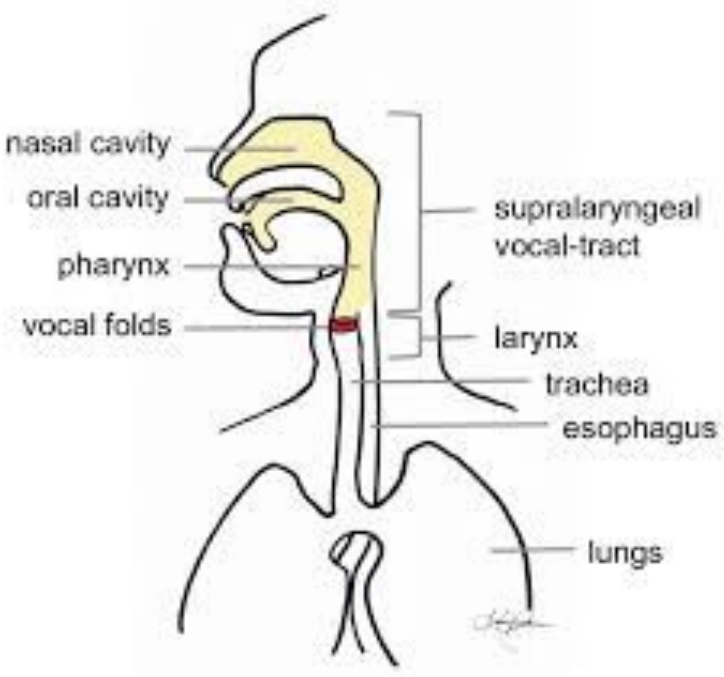
Source-filter: Subtractive synth

Source-filter model: Revisited



1. Lungs = energy source

Source-filter model: Revisited



1. Lungs = energy source
2. Vocal folds = oscillator (i.e., glottal sound)

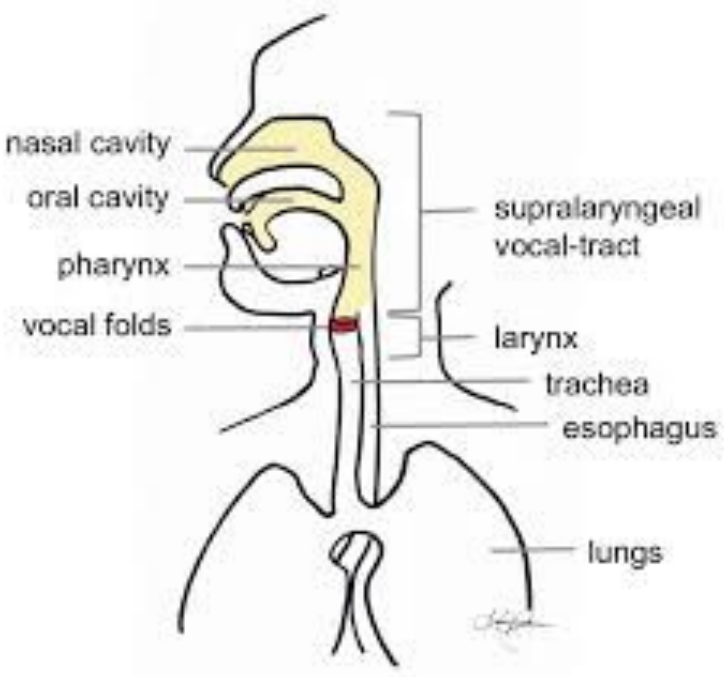
Glottal sound

- Vocal folds opening-closing cycle
- Quasi-periodic waveform
- The vibration frequency = fundamental frequency (f_0)
- The spectrum contains harmonics

Types of glottal sounds

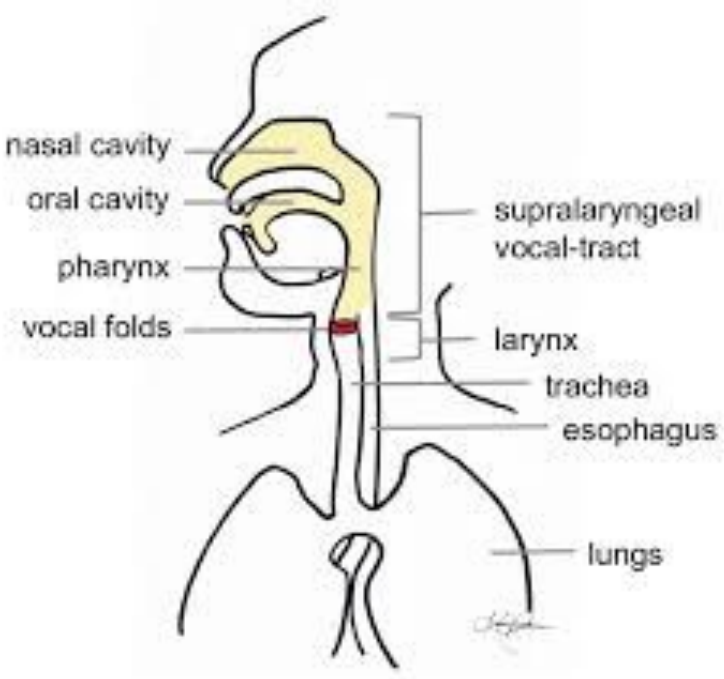
- Voiced sounds: vocal folds vibrate
  → /a/, /b/, /d/
- Voiceless sounds: folds open →
  noise-like → /s/, /f/, /t/

Source-filter model: Revisited



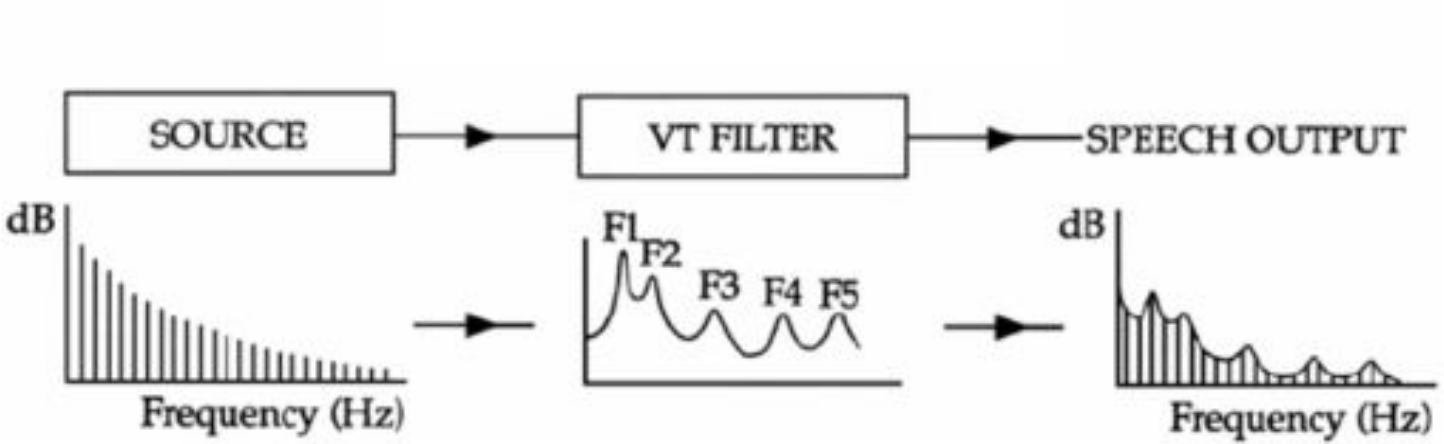
1. Lungs = energy source
2. Vocal folds = oscillator (i.e., glottal sound)
3. Vocal tract = resonant filter

Source-filter model: Revisited

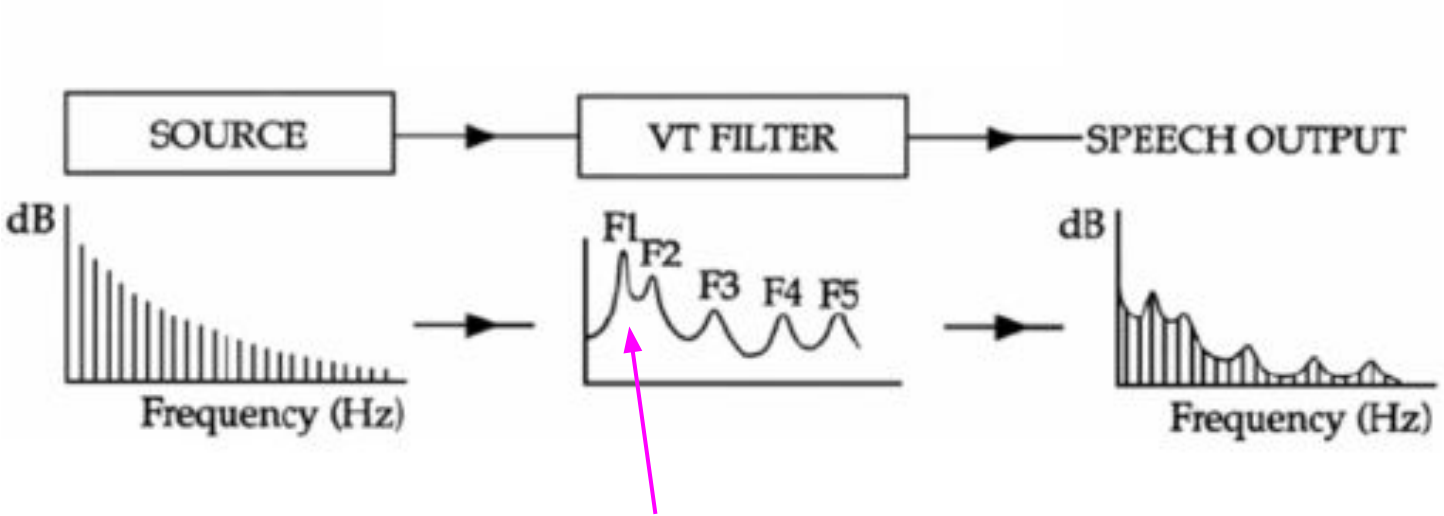


1. Lungs = energy source
2. Vocal folds = oscillator (i.e., glottal sound)
3. Vocal tract = resonant filter
4. Lips, tongue, jaw = articulators

Source-filter model: Revisited

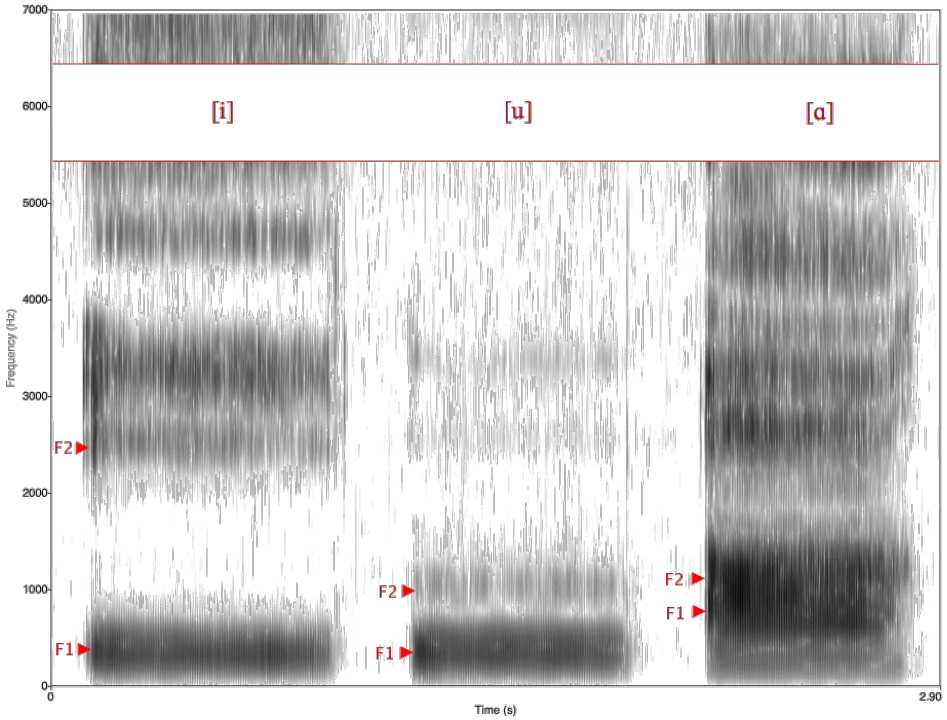


Source-filter model: Revisited



Amplified frequency regions = Formants

Formants



- Resonances
- Different vowels, different formants
- Changing tongue and lip position shifts formant frequencies → changes vowel
- Formants make you sound like you


WHY SHOULD WE
CARE ABOUT
SOURCE-FILTER AND FORMANTS?
imgflip.com

Many TTS systems use source-filter and formants

Emotion and expressivity

- Speech is deeply emotional and context-driven
- Emotion changes prosody, pitch, energy, tempo
- Example: “I’m fine”

Speech is multilayered

Layer	What It Encodes	AI Analogy
Thought	Meaning	Text input
Language	Words, grammar	Text-to-phoneme
Prosody	Rhythm, melody	Acoustic model
Timbre	Identity	Speaker embedding
Waveform	Air vibration	Vocoder output

In this course, we'll learn how AI replicates each of these layers

Why is TTS/voice cloning hard?

Why is TTS/voice cloning hard?

- Speech combines structure (phonemes) + fluidity (prosody) + individuality (timbre) + dynamics (emotion)

Why is TTS/voice cloning hard?

- Speech combines structure (phonemes) + fluidity (prosody) + individuality (timbre) + dynamics (emotion)
- Humans learn it effortlessly; machines have to model each layer