

2. How humans speak

The Monster Text to Speech & Voice Cloning Course

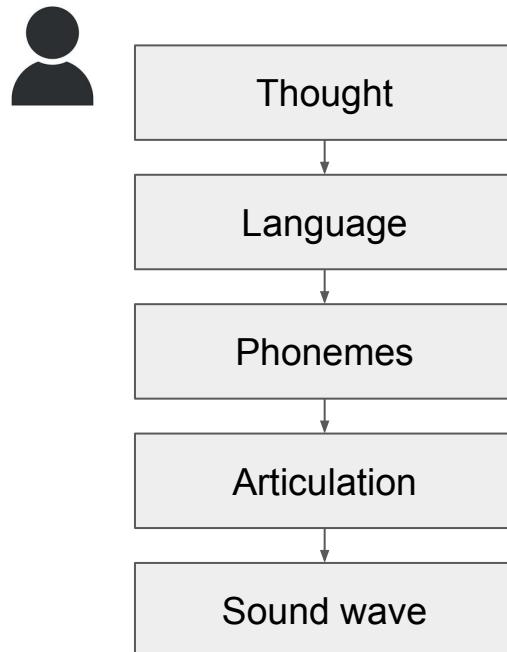
THE  SOUND OF AI

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

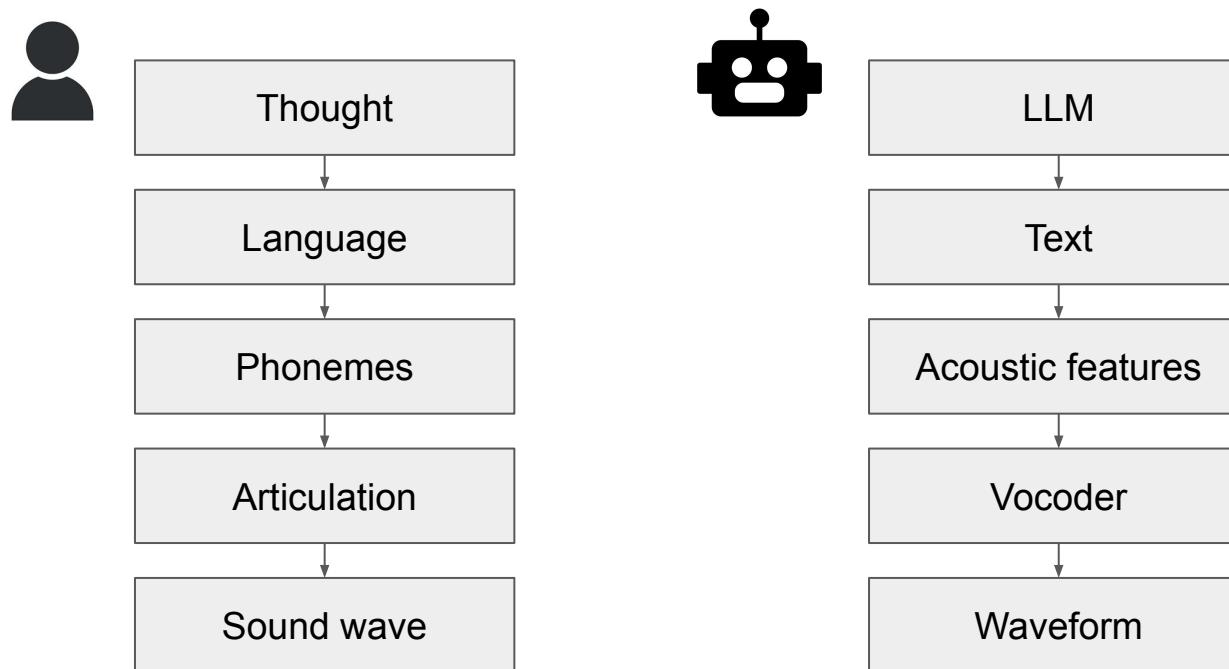
What makes speech special

- 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

Phrase 1 Bar 1 Bar 2 Bar 3 Bar 4

A musical staff in G clef and common time. It consists of four bars of music. The lyrics are: "Twin-kle Twin-kle lit-tle star, how I won-der what you are." The notes correspond to the syllables of the words, with a note for each phoneme.

Phrase 2 Bar 5 Bar 6 Bar 7 Bar 8

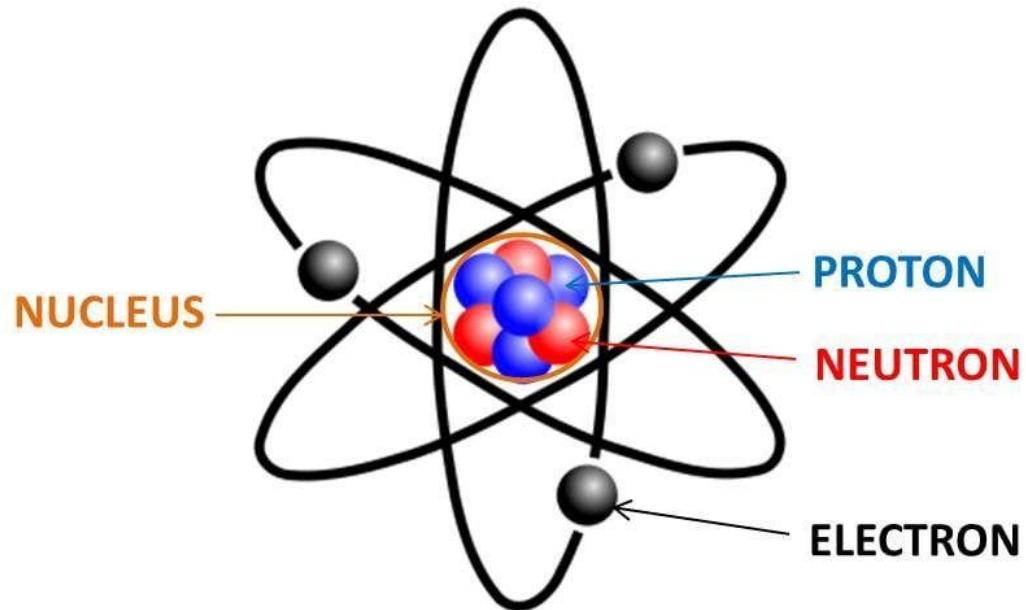
A musical staff in G clef and common time. It consists of four bars of music. The lyrics are: "Up a- bove the world so high, like a dia-mond in the sky!" The notes correspond to the syllables of the words, with a note for each phoneme.

Phrase 3 Bar 9 Bar 10 Bar 11 Bar 12

A musical staff in G clef and common time. It consists of four bars of music. The lyrics are: "Twin-kle Twin-kle lit-tle star, how I won-der what you are." The notes correspond to the syllables of the words, with a note for each phoneme.

- Word = musical phrase
- Phoneme = note

Understanding phonemes



International Phonetic Alphabet

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2020)												
© 2020 IPA												
CONSONANTS (PULMONIC)												
Plosive	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal	
p	b		t	d	t̪	d̪	c	j	k	g	q	g̪
Nasal	m	n̪		n	n̪	j̪	ŋ	ŋ̪		N		
Trill	B			r					R			
Tap or Flap		v̄		t̄								
Fricative	f̪	β	f	v	θ̪	s̪	z̪	ʃ̪	x̪	χ̪	h̪	h̪ f̪
Lateral fricative					ɬ̪							
Approximant		U		x̪		ɬ̪	j̪	w̪				
Lateral approximant	I			l̪		ɻ̪		L̪				

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)		
Clicks	Voiced implosives	Ejectives
O Bilabial	b	' Examples:
Dental	d̪	Bilabial
! (Postalveolar	f̪	p̪
# Palatoalveolar	g̪	t̪
Alveolar lateral	g̪ uvar	s̪ Alveolar fricative

OTHER SYMBOLS

ℳ Voiceless labial-velar fricative	ℳ Voiceless labio-palatal fricatives
W Voiced labial-lateral approximant	J Voiced alveolar lateral flap
ℳ Voiced labial-palatal approximant	J Simultaneous J and X
H Voiceless epiglottal fricative	Affricates and double articulations can be represented by two symbols joined by a bar if necessary.
ℳ Voiced epiglottal fricative	
? Epiglottal plosive	

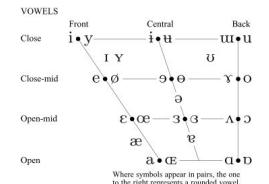
DIACRITICS

o Voiceless	n̪ d̪	.. Breathy voiced	b̄ ā	~ Dental	t̄ d̄
~ Voiced	s̪ t̪	~ Creaky voiced	b̄ ā	~ Apical	t̄ d̄
h Aspirated	t̪̄ d̪̄	~ Lingualized	t̄ d̄	~ Laminar	t̄ d̄
~ More rounded	χ̄	~ Labialized	t̄ w̄	~ Nasalized	ē̄
c Less rounded	χ̄	~ Palatalized	t̄j̄ d̄j̄	~ Nasal release	d̄n̄
~ Advanced	ū	~ Velarized	t̄v̄ d̄v̄	~ Lateral release	d̄l̄
~ Retracted	ē	~ Pharyngalized	t̄v̄ d̄v̄	~ No audible release	d̄v̄
~ Centralized	ē̄	~ Vocalized or pharyngalized	χ̄		
* Mid-centralized	ē̄	~ Raised	ē (ē = voiced alveolar fricative)		
Syllabic	n̄	~ Lowered	ē (β̄ = voiced bilabial approximant)		
Non-syllabic	ē	~ Advanced Tongue Root	ē		
~ Rhicity	ð̄ ð̄r̄	~ Retracted Tongue Root	ē		

Some diacritics may be placed above a symbol with a descender, e.g. ī̄

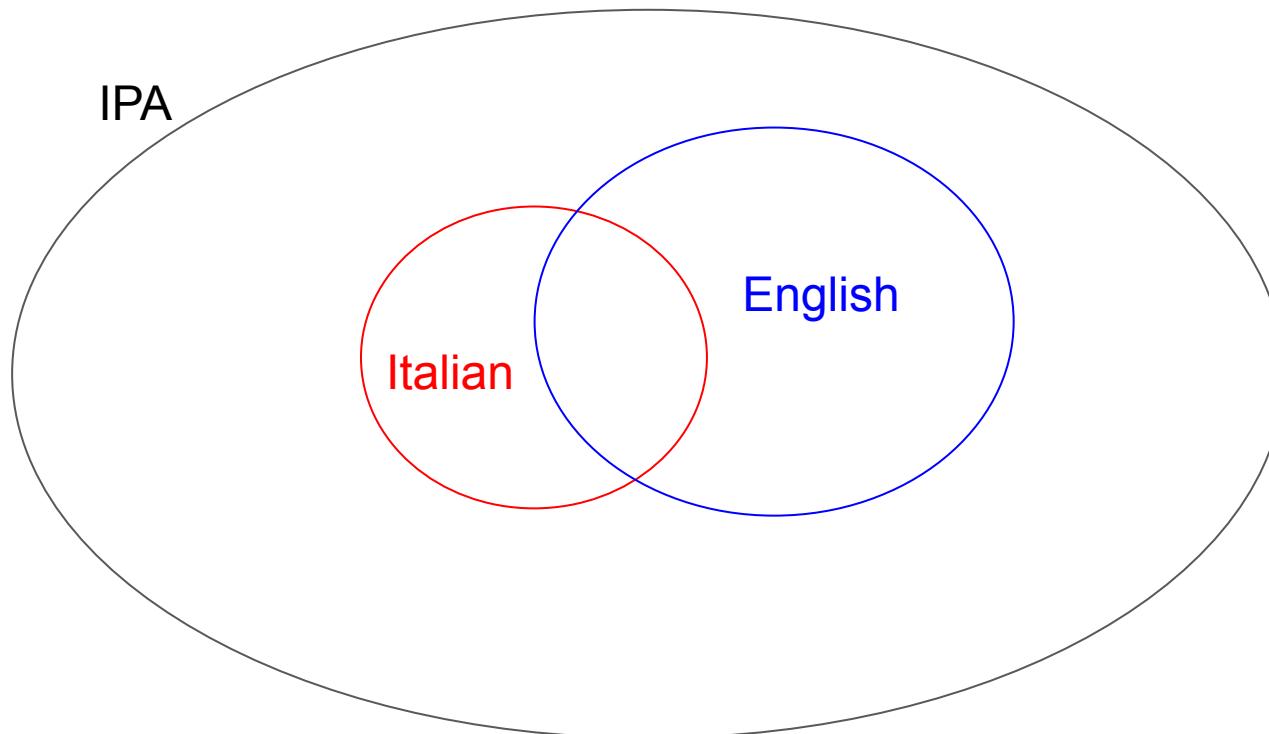
Typeface: Didot 14pt, italicized, sans-serif

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



SUPRASEGMENTALS	
↑ Primary stress	,fōon̄'t̄fən̄
↓ Secondary stress	
— Long	ē:
— Half-long	ē'
— Extra-short	é̄
— Minor (foot) group	
— Major (intonation) group	
— Syllable break	.i..ækt̄
— Linking (absence of a break)	
TONES AND WORD ACCENTS	
— LEVEL	— CONTOUR
é̄ ↑ Extra high	é̄ ↗ Raising
é̄ ↑ High	é̄ ↘ Falling
é̄ ↓ Mid	é̄ ↙ High falling
é̄ ↓ Low	é̄ ↖ Low rising
é̄ ↓ low	é̄ ↘ Global rise
↑ Upstep	↓ Global fall

Language phoneme inventories



English phonetic chart

	monophthongs				diphthongs		
VOWELS	i: sheep	I ship	ʊ good	u: shoot	ɪə here	eɪ wait	
	e bed	θ teacher	ɜ: bird	ɔ: door	ʊə tourist	ɔɪ boy	əʊ show
	æ cat	ʌ up	a: far	ɒ on	eə hair	aɪ my	aʊ cow
CONSONANTS	p pea	b boat	t tea	d dog	tʃ cheese	dʒ June	k car
	f fly	v video	θ think	ð this	s see	z zoo	ʃ shall
	m man	n now	ŋ sing	h hat	l love	r red	w wet
				l	r	w	j yes

Phonemic Chart
voiced
unvoiced

Phonetic transcription

She bought two red apples

/ʃi əʊ:t tu: red 'æplz/

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?



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

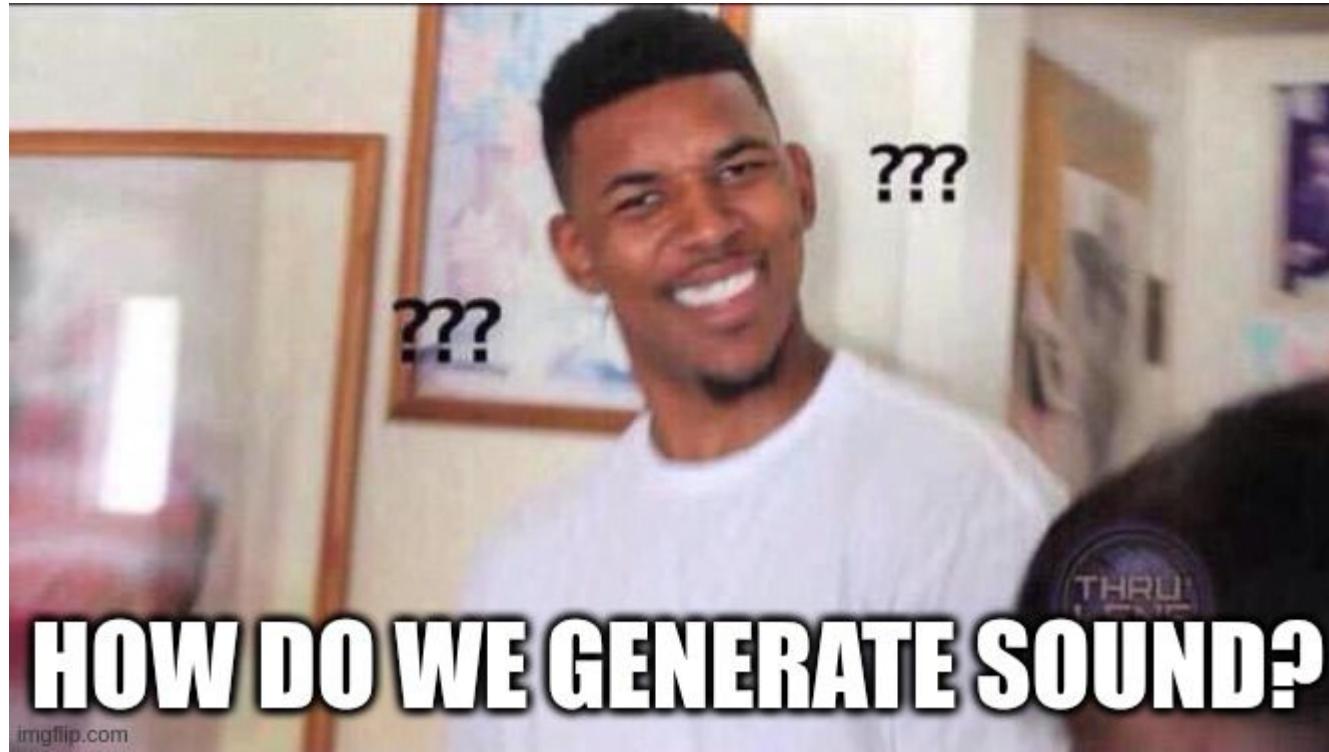
- Timbre is the “color” or quality of a voice
- Why every person sounds unique:
 - Vocal tract shape
 - Resonance of the nasal and oral cavities
 - Habitual pitch, age, health

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



THE SOUND OF AI

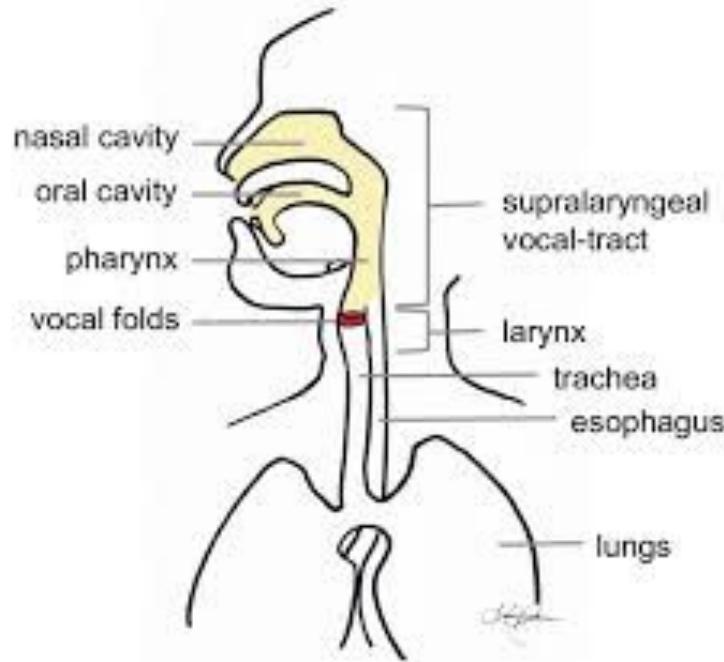
Source-filter model

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

Source-filter: Subtractive synth

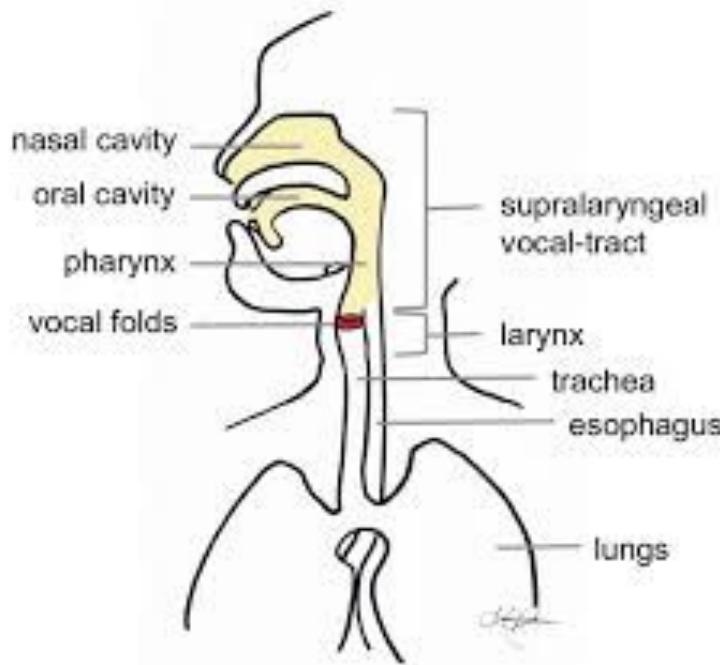


Source-filter model: Revisited



1. Lungs = energy source

Source-filter model: Revisited



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

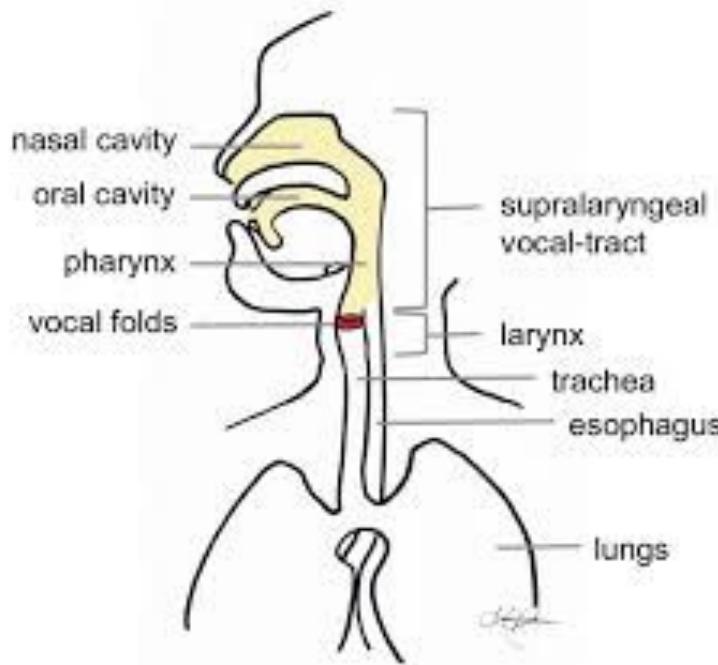
Glottal sound

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

Types of glottal sounds

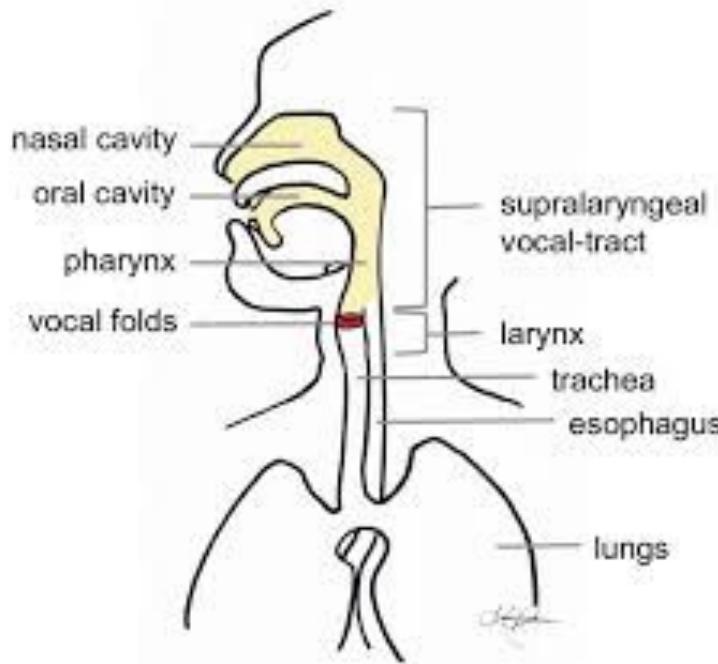
- Voiced sounds: vocal folds vibrate
→ /a/, /b/, /d/
- Voiceless sounds: folds open →
noise-like → /s/, /f/, /t/

Source-filter model: Revisited



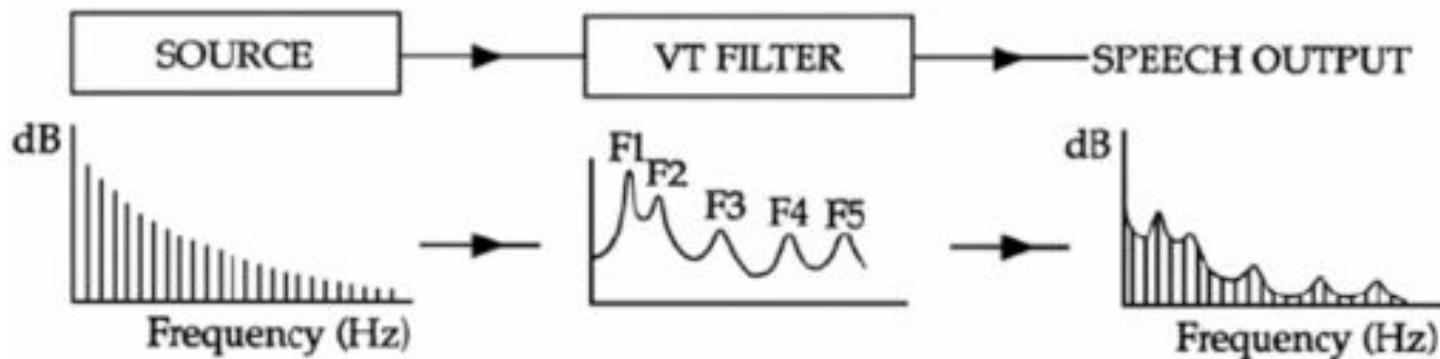
1. Lungs = energy source
2. Vocal folds = oscillator (i.e., glottal sound)
3. Vocal tract = resonant filter

Source-filter model: Revisited

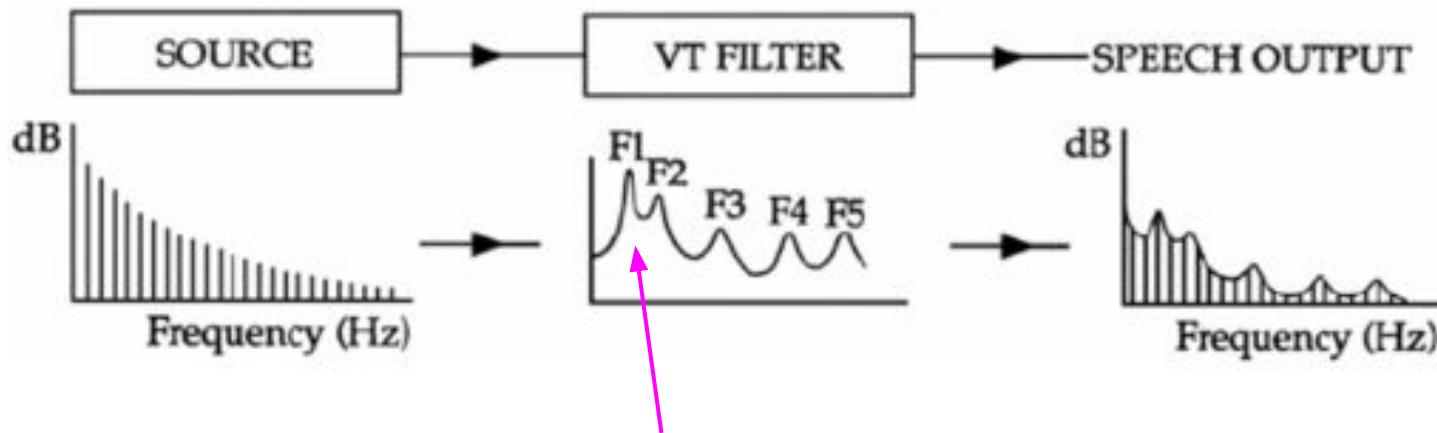


1. Lungs = energy source
2. Vocal folds = oscillator (i.e., glottal sound)
3. Vocal tract = resonant filter
4. Lips, tongue, jaw = articulators

Source-filter model: Revisited

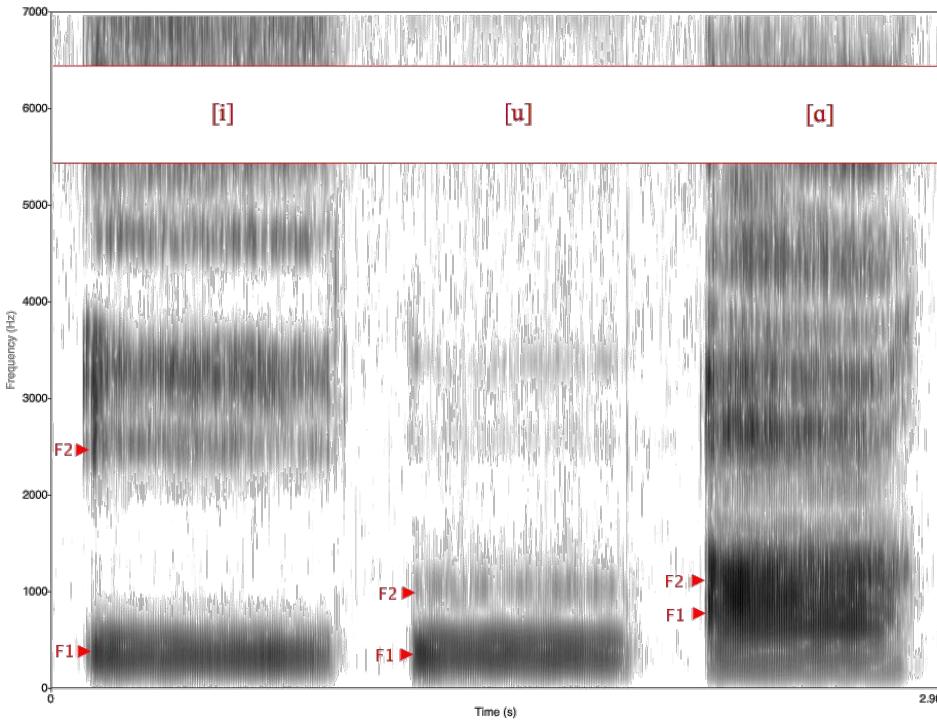


Source-filter model: Revisited



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?**

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