DT2119
Speech and Speaker Recognition
Introduction

Giampiero Salvi

KTH/CSC/TMH giampi@kth.se
NTNU/IE/IES giampiero.salvi@ntnu.no

VT 2020

Outline

Outline

Who are we? (Contact Info)

Teacher, course responsible, examiner:
Giampiero Salvi

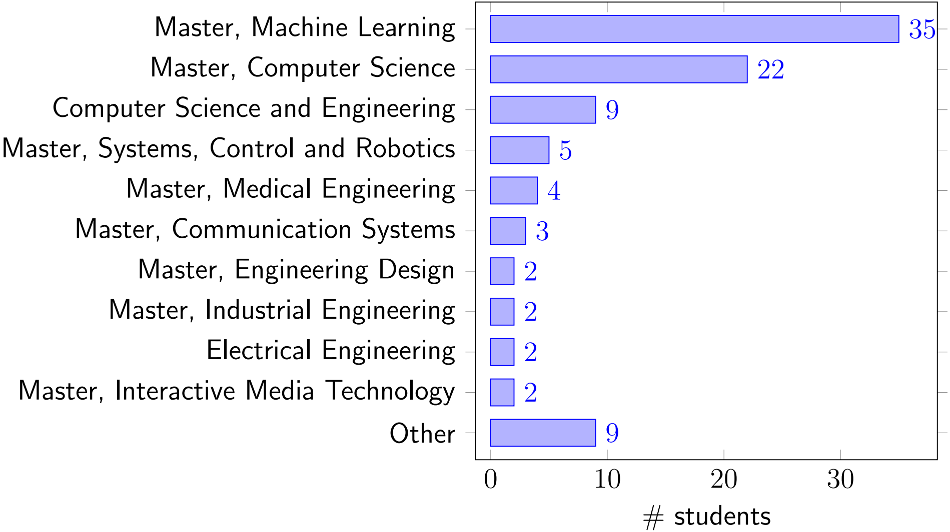
Teaching Assistants:

Gustav Henter Ronald Combal Polydefkis Gkagkos
Stavros Giorgis Alexandros Ferles

All communications handled through Canvas
`https://kth.instructure.com/courses/17109`

You are encouraged to use the "Discussions" forum

Who are you? (data from expected participants)



Course Objectives

after the course you should be able to:

- describe the state-of-the-art in a topic in speech and speaker recognition
- carry out experiments related to speech and speaker recognition
- present experimental results in written form and orally
- interpret and explain topics in speech production and recognition
- with the help of the literature, review and criticise other students' work in the subject

Course Objectives

after the course you should be able to:

- describe the state-of-the-art in a topic in speech and speaker recognition
- carry out experiments related to speech and speaker recognition
- present experimental results in written form and orally
- interpret and explain topics in speech production and recognition
- with the help of the literature, review and criticise other students' work in the subject

Detailed Grading Criteria in Canvas

Course Objectives (research perspective)

- explore literature
- carry out experiments
- produce documentation
- provide feedback (peer review)
- accept and use feedback (revision)
- present results

Topics

Part	Topic	time (hours)
1	Introduction, Speech Signal, Signal Processing, Features	~ 6
2	Hidden Markov Models, Training and Decoding, Acoustic Models	~ 6
3	Deep Learning for ASR	~ 4
4	Decoding and Search Algorithms	~ 2
5	Language Models (Grammars)	~ 2
6	Noise robustness and Speaker Recognition	~ 4
	Total	~ 24

Literature

- **Spoken Language Processing: A Guide to Theory, Algorithm, and System Development**

  Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, Prentice Hall

  - 3 at KTH library,
  - 6 at TMH library (against 300 SEK deposit)

- **Automatic Speech Recognition: A deep learning approach**

  Dong Yu and Li Deng, Springer 2015

  Available in PDF from SpringerLink (via KTH Biblioteket)

- **HTK manual** version 3.4

- selected research articles

Reading Instructions

These are indicative, check the schedule for more updated instructions

		pages	# pages
Part 1	(Spoken Language Structure)	(19–71)	(52)
	Digital Signal Processing	(201–273)	73
	Probability, Statistics and Inform. Theory	73–131	59
	Pattern Recognition	133–197	65
	Speech Signal Representations	275–336	62
Part 2	Hidden Markov Models	377–413	37
	Acoustic Modeling	415–475	61
	Environmental Robustness	477–544	68
Part 3	Deep Neural Nets and ASR	Ch4, Ch6[1]	35
Part 4	Basic Search Algorithms	591–643	53
	(Large-Vocabulary Search Algorithms)	(645–685)	(41)
	(Applications and User Interfaces)	(919–956)	(38)
Part 5	Language Modeling	545–590	46
Part 6	Speaker Recognition literature		

[1] Dong and Deng's book

Lab 1: Speech Feature Extraction

- implement extraction for typical speech features
- analyse the features on speech data
- compare utterances with Dynamic Time Warping

Lab 2: Gaussian Hidden Markov Models

- implement the decoding algorithms for HMMs
- implement the training algorithms for HMMs
- test the algorithms on isolated digits

Lab 3: Continuous Speech Recognition and Deep Learning

- Extend the training and testing algorithms to continuous speech
- test the algorithms on the TIDIGIT database (connected digits)
- implement DNNs using Keras and TensorFlow, compare with GMM-HMMs

Project

- A list of topics will be available shortly
- you are also welcome to suggest topics outside the list
- Project report in form of research paper
- Normally the project has an experimental part, but you can choose to do a literature study.

Grading Criteria: Prerequisite for Pass

In groups:

1. **present** the three labs
2. carry out **project work**
3. submit **report draft**
4. **present** at final seminar (the form will depend on COVID19)
5. submit **final report** including answers to reviews

Individually:

1. carry out **quizzes** on Canvas
2. **review** other students' report

Grading Criteria for E-A

Check the full grading criteria on Canvas.

Note that if you choose a literature study as project, you will be granted a maximum grade of **C**

GPU Resources (Lab 3 and Project)

Parallel Data Centre at KTH:

- PDC accounts will be created for all registered students
- alternatively, apply for an account at `https://pdc-web-01.csc.kth.se/accounts/`
- use `edu20.DT2119` when asked for time allocation

Google Cloud Platform (recommended):

- 50$ per student
- we are updating the procedure and instructions

Amazon Web Services through GitHub

Time Organisation (Preliminary)

Week 12 (March 16): Course start

Week 15 (April 8): Decide groups/project topics

Week 15 (before April 10): Present Lab 1

Week 18 (before April 30): Present Lab 2

Week 21 (before May 22): Present Lab 3

Week 21 (May 20): Submit first version of report

Week 22 (May 28): Submit review on report

Week 23 (June 1): Project poster presentations (depends on COVID19)

Week 23 (June 5): Submit final report.

always check Canvas for updated deadlines!

Part 1

Outline

Motivation

Human-Computer (or -Robot) Interaction

- Natural way of communication (No training needed)
- Leaves hands and eyes free (Good for functionally disabled)
- Effective (Higher data rate than typing)
- Can be transmitted/received inexpensively (phones)

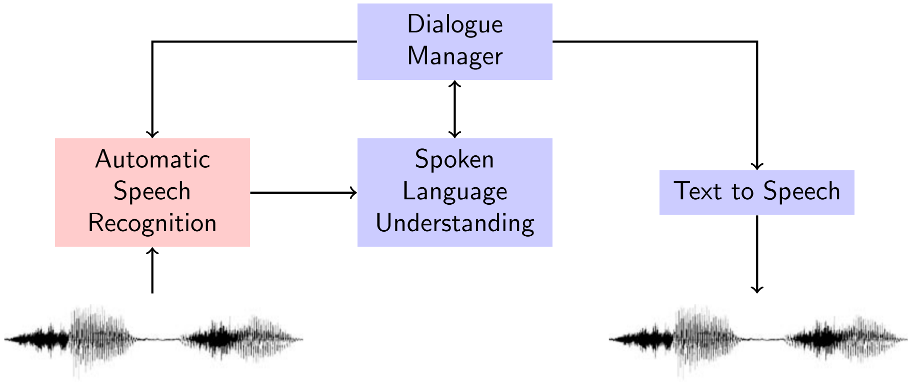
Surveillance/Search

- Transcribe human-human conversations
- produce indexing for broadcast material
- produce subtitles for movies/news

Dream and Reality in Artificial Intelligence

`https://youtu.be/JepKVUym9Fg`,
based on "2001: A space odyssey" (1968)

ASR in a Broader Context



The ASR Scope

Convert speech into text

→ Automatic Speech Recognition → "My name is . . ."
[confidence score]

The ASR Scope

Convert speech into text

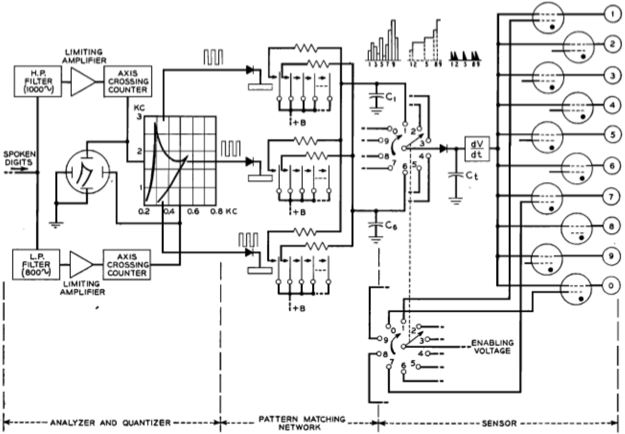
$\longrightarrow$ Automatic Speech Recognition $\longrightarrow$ "My name is . . . " [confidence score]

Not considered here:

- non-verbal signals
- prosody
- multi-modal interaction

A very long endeavour

1952, Bell laboratories, isolated digit recognition, single speaker, hardware based[2]

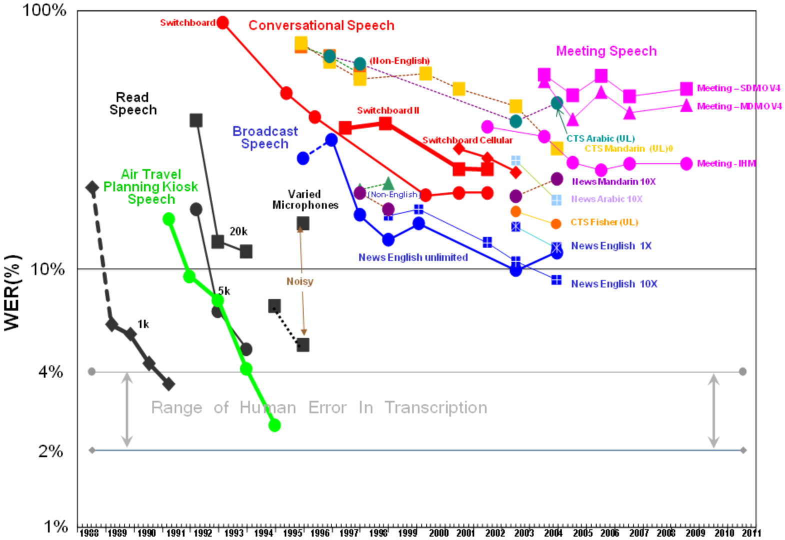


[2] K. H. Davis, R. Biddulph, and S. Balashek. "Automatic Recognition of Spoken Digits". In: 24.6 (1952), pp. 637–642.

Historical Perspective

- 1950's Bel lab: 10 digits, 1 speaker
- 1960's IBM: 16 words
- 1970's Large investments from DARPA
  - CMU Harpy: 1011 words (beam search)
  - Threshold Technology: first ASR company
  - Bell labs: multiple voices
- 1980's
  - from template matching to probabilistic models (Hidden Markov Models)
  - from hundreds to thousands of words
- 1990's Dragon Dictate and later Dragon NaturallySpeaking
- 2000's not big improvements
- 2010's Google, Apple, Microsoft, Amazon get heavily involved

NIST STT Benchmark Test History – May. '09



http://www.itl.nist.gov/iad/mig/publications/ASRhistory/

Main variables in ASR

Speaking mode isolated words vs continuous speech

Speaking style read speech vs spontaneous speech

Speakers speaker dependent vs speaker independent

Vocabulary small ($<$20 words) vs large ($>$50 000 words)

Robustness against background noise

Challenges — Variability

Between speakers

- Age
- Gender
- Anatomy
- Dialect

Within speaker

- Stress
- Emotion
- Health condition
- Read vs Spontaneous
- Adaptation to environment (Lombard effect)
- Adaptation to listener

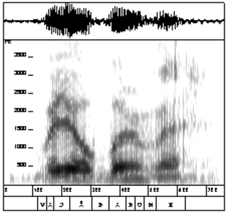
Environment

- Noise
- Room acoustics
- Microphone distance
- Microphone, telephone
- Bandwidth

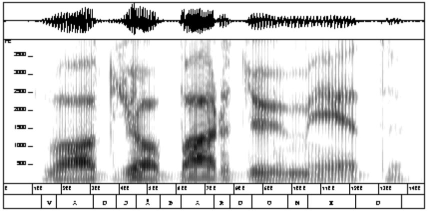
Listener

- Age
- Mother tongue
- Hearing loss
- Known / unknown
- Human / Machine

Example: spontaneous vs hyper-articulated



Va jobbaru me



Vad jobbar du med

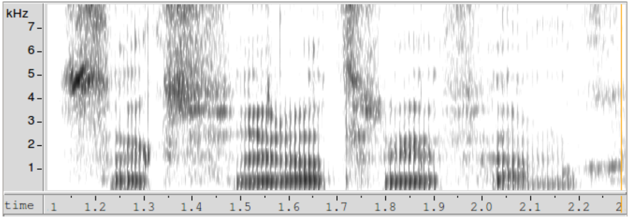
"What is your occupation"
("What work you with")

Examples of reduced pronunciation

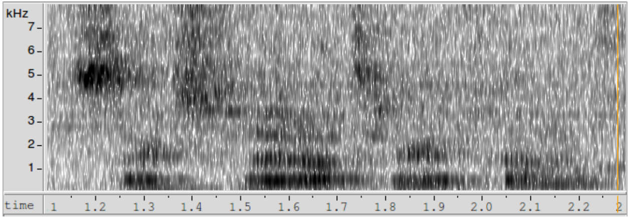
Spoken	Written	In English
Tesempel	Till exempel	for example
åhamba	och han bara	and he just
bafatt	bara för att	just because
javende	jag vet inte	I don't know

Microphone distance

Headset



2 m distance



Applications today

Call centers:

- traffic information
- time-tables
- booking...

Accessibility

- Dictation
- hand-free control (TV, video, telephone)

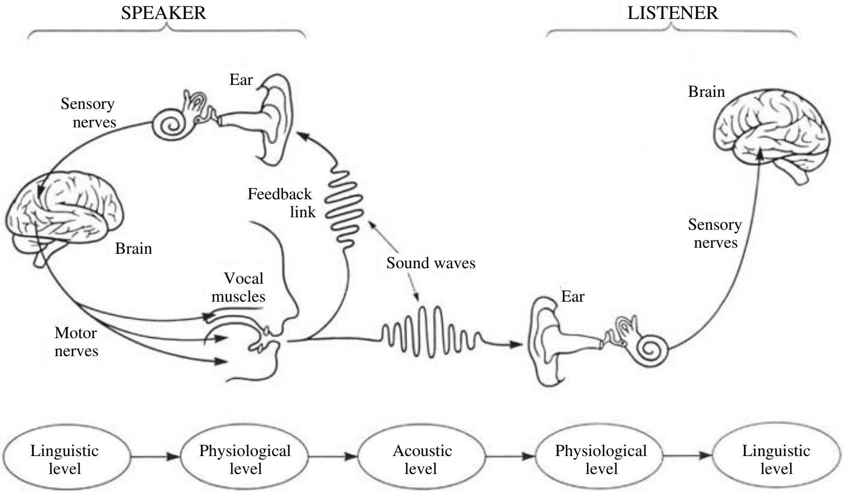
Smart phones

- Siri, Android, ...

Smart speakers

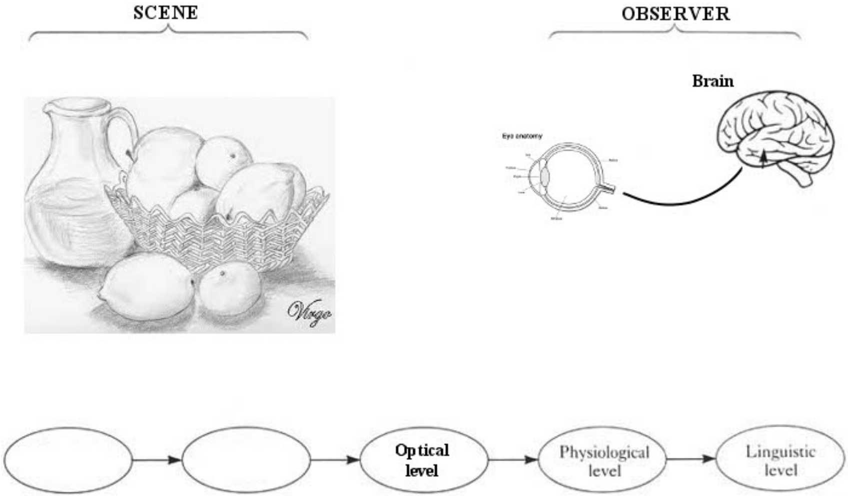
- Amazon Echo, Google Home, ...

The Speech Chain



Peter Denes, Elliot Pinson, 1963

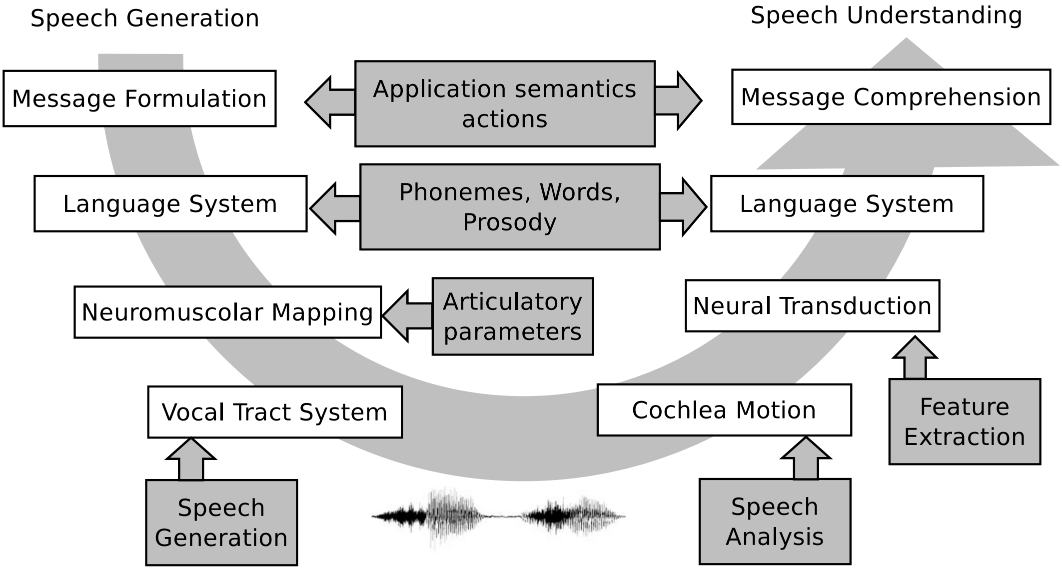
ASR versus Computer Vision



ASR versus Computer Vision

Property	ASR	Computer Vision
signal originates from:	cognition + physics	physics
persistence:	disappears as soon as heard	continually available (active perception)
across countries:	different languages	same objects
type of interaction:	two-way	one-way

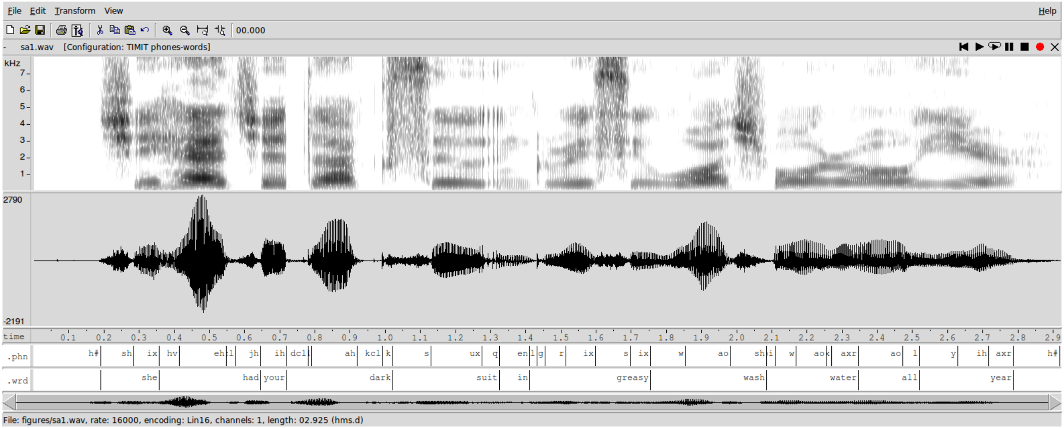
The Speech Chain (from the book)



Outline

Speech Examples

TIMIT database (American English)



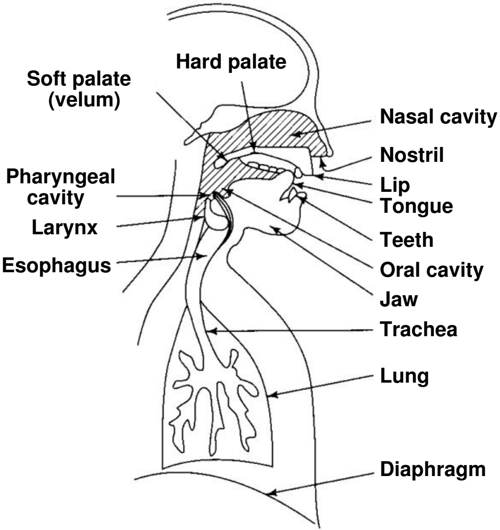
example of “clean” speech

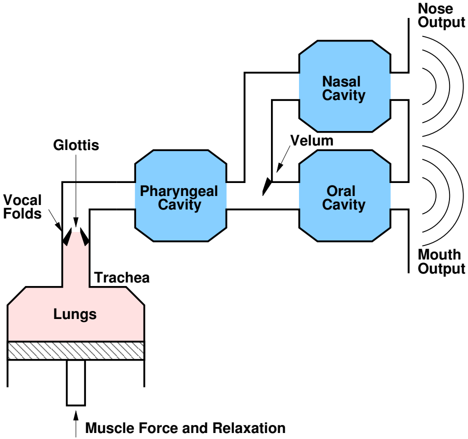
Speech Examples

live examples

https://sourceforge.net/projects/wavesurfer/

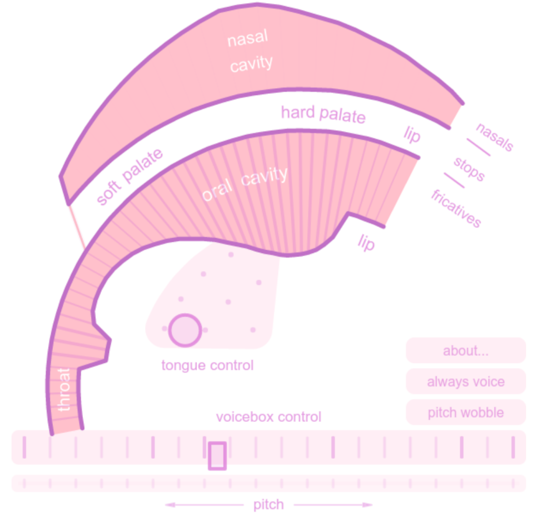
Physiology


Hard palate
Soft palate
(velum)
Nasal cavity
Nostril
Lip
Tongue
Pharyngeal
cavity
Larynx
Esophagus
Teeth
Oral cavity
Jaw
Trachea
Lung
Diaphragm


Nose
Output
Nasal
Cavity
Velum
Glottis
Pharyngeal
Cavity
Oral
Cavity
Vocal
Folds
Mouth
Output
Trachea
Lungs
Muscle Force and Relaxation

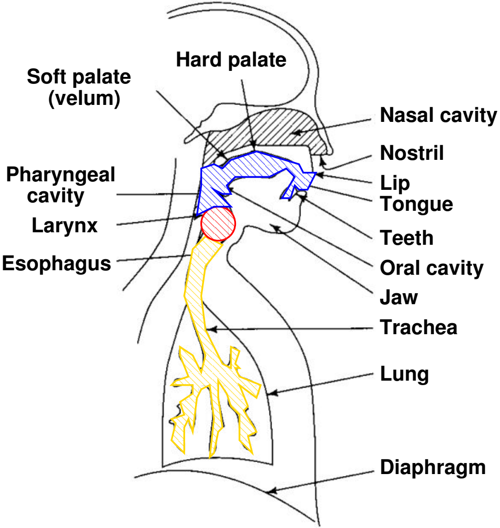
Pink Trombone!

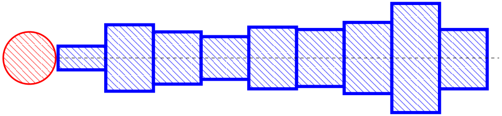
http://dood.al/pinktrombone/



Source/Filter Model, Vowel-like sounds

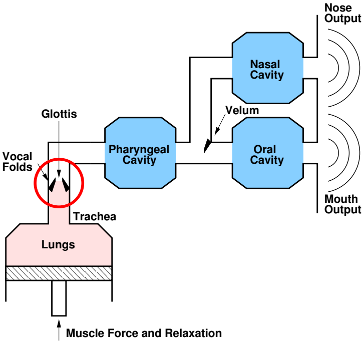
Vowels

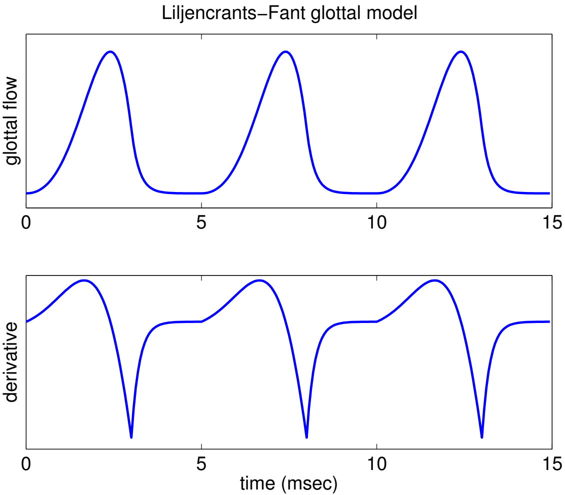




- Source (periodic)
- Front Cavity
- Back Cavity
- Back Cavity (2nd approx.)

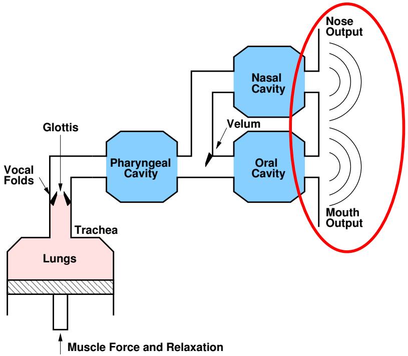
Glottal Flow





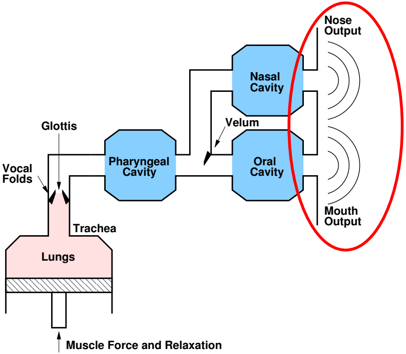
$$G(z) = \frac{1}{(1 - \beta z)^2}, \quad \beta < 1$$

Radiation form the Lips/Nose

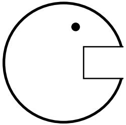


Problem of radiation at the lips plus diffraction about the head too complicated.

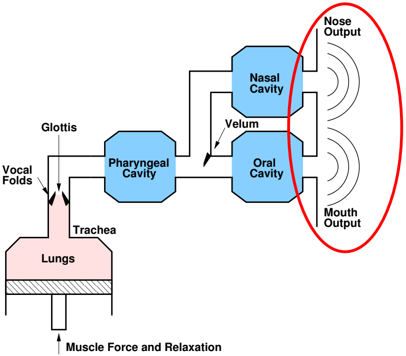
Radiation form the Lips/Nose



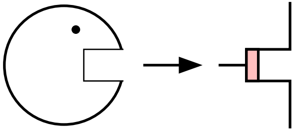
Approx. with a piston in a rigid sphere: solved but not in closed form



Radiation form the Lips/Nose

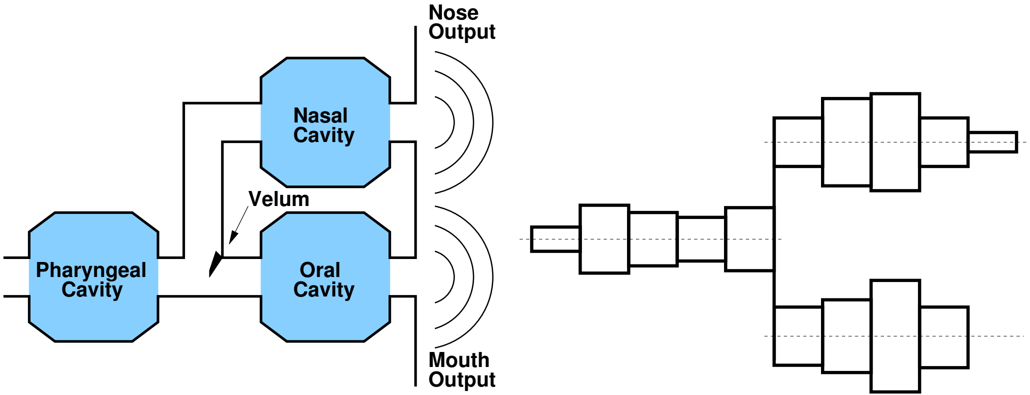


2nd approx: piston in an infinite wall

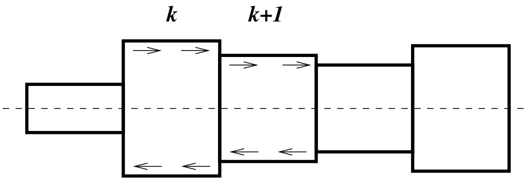


$$R(z) \approx 1 - \alpha z^{-1}$$

Tube Model of the Vocal Tract



Tube Model (cntd.)

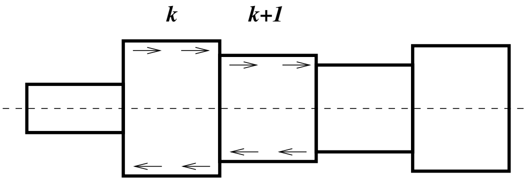


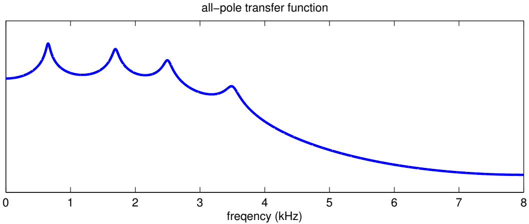
- assume planar wave propagation and lossless tubes
- solve pressure $p(x,t)$ and velocity $u(x,t)$ in each tube according to wave equation
- impose continuity of pressure and velocity at the junctions

$\Rightarrow$ all-pole transfer function ($N =$ number of tubes)

$$V(z) = \frac{Az^{-N/2}}{1 - \sum_{k=1}^{N} a_k z^{-k}}$$

Tube Model (cntd.)



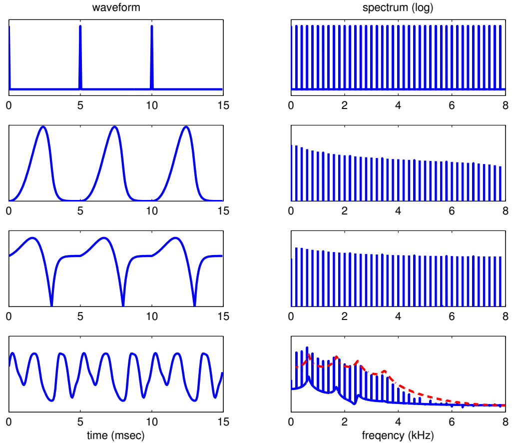


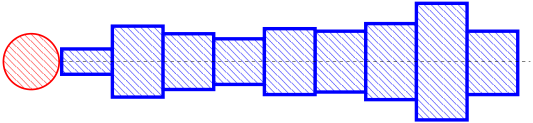
- assume planar wave propagation and lossless tubes
- solve pressure $p(x, t)$ and velocity $u(x, t)$ in each tube according to wave equation
- impose continuity of pressure and velocity at the junctions

$\Rightarrow$ all-pole transfer function ($N =$ number of tubes)

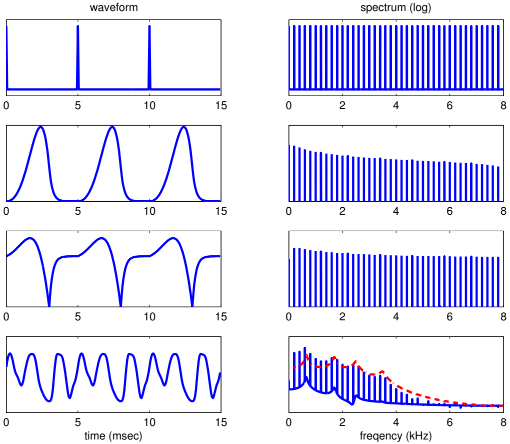
$$V(z) = \frac{Az^{-N/2}}{1 - \sum_{k=1}^{N} a_k z^{-k}}$$

Source/Filter Model: vowel-like sounds

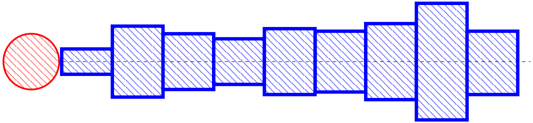




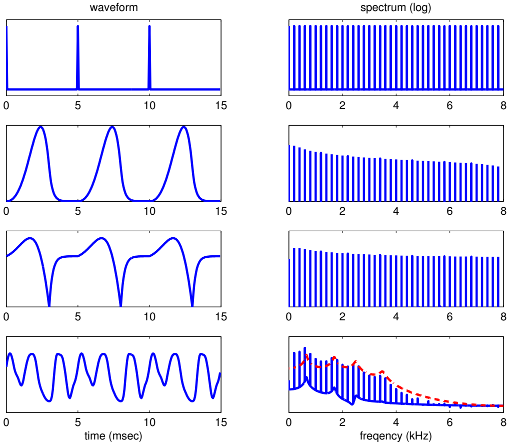
Source/Filter Model: vowel-like sounds



$\leftarrow p[n]$

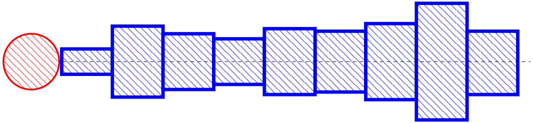


Source/Filter Model: vowel-like sounds

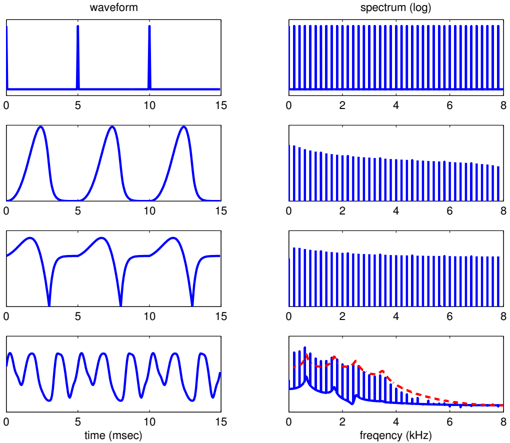


$\leftarrow p[n]$

$\leftarrow p[n] * g[n]$



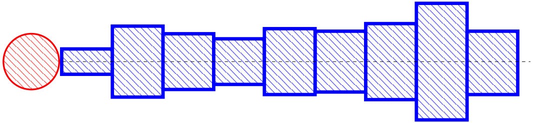
Source/Filter Model: vowel-like sounds



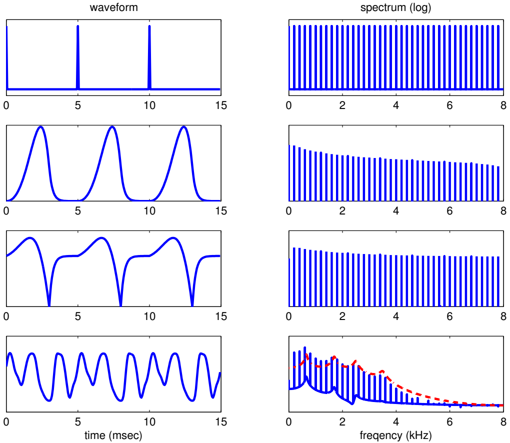
$\leftarrow p[n]$

$\leftarrow p[n] * g[n]$

$\leftarrow p[n] * g[n] * r[n]$



Source/Filter Model: vowel-like sounds

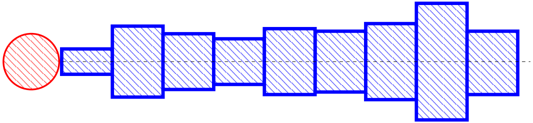


$\leftarrow p[n]$

$\leftarrow p[n] * g[n]$

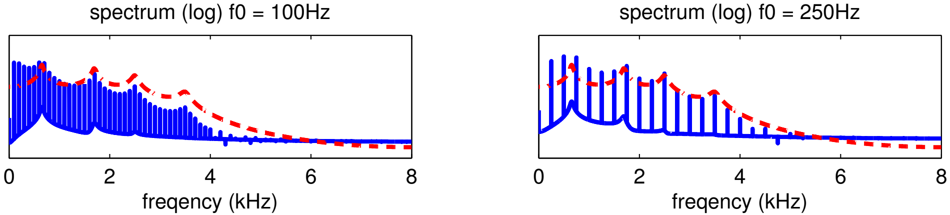
$\leftarrow p[n] * g[n] * r[n]$

$\leftarrow p[n] * g[n] * r[n] * v[n]$



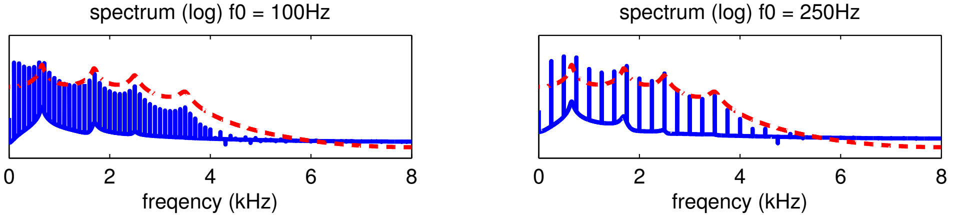
F_0 and Formants

- Varying F_0 (vocal fold oscillation rate)

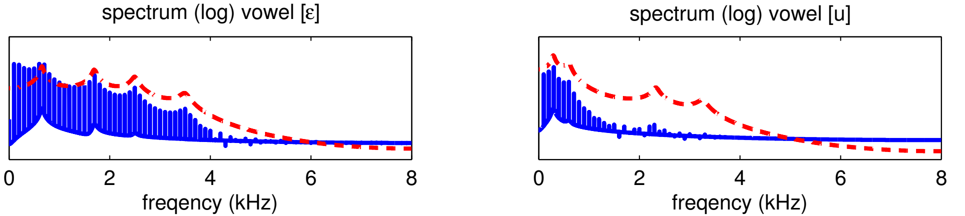


F_0 and Formants

- Varying F_0 (vocal fold oscillation rate)

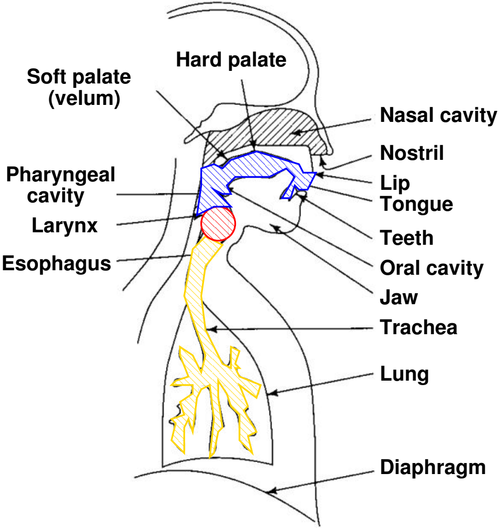


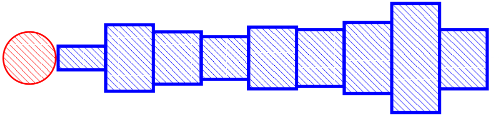
- Varying Formants (vocal tract shape)

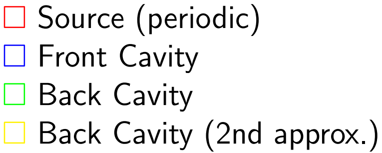


Source/Filter Model, General Case

Vowels

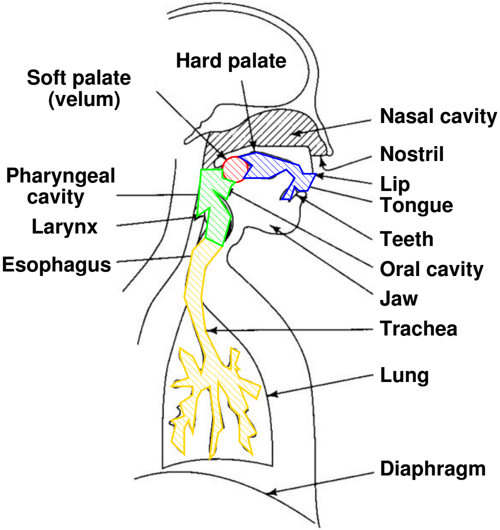


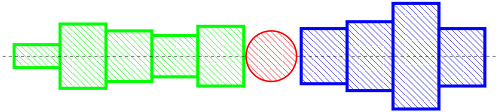




Source/Filter Model, General Case

Fricatives (e.g. sh) or Plosive (e.g. k)

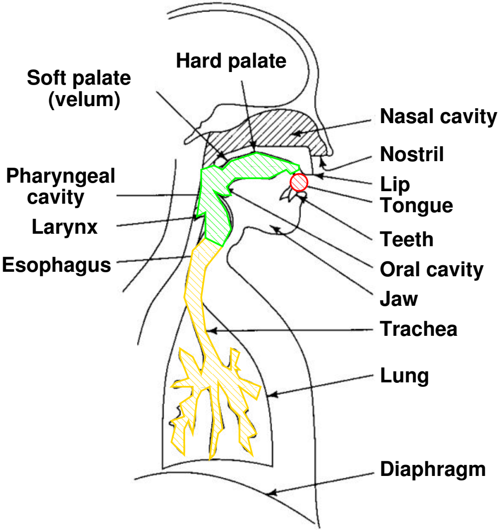


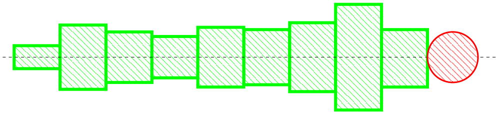


- Source (noise or impulsive)
- Front Cavity
- Back Cavity
- Back Cavity (2nd approx.)

Source/Filter Model, General Case

Fricatives (e.g. s) or Plosive (e.g. t)

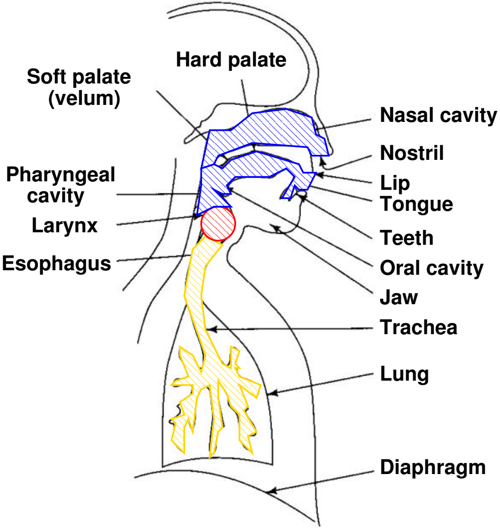


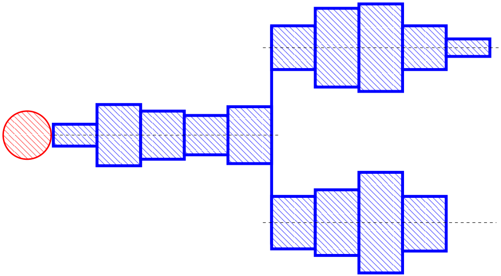


- Source (noise or impulsive)
- Front Cavity
- Back Cavity
- Back Cavity (2nd approx.)

Source/Filter Model, General Case

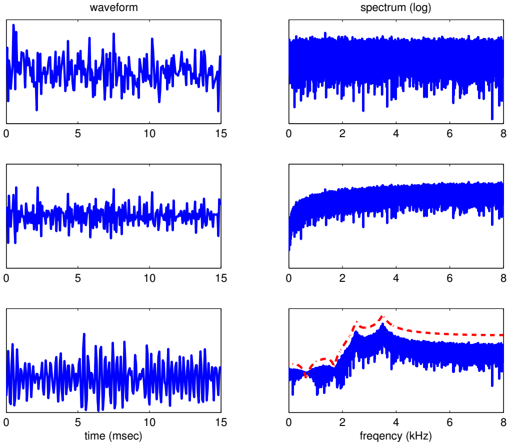
Nasalised Vowels





- Source (periodic)
- Front Cavity
- Back Cavity
- Back Cavity (2nd approx.)

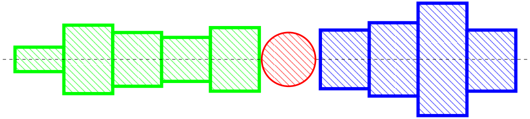
Source/Filter Model: fricative sounds



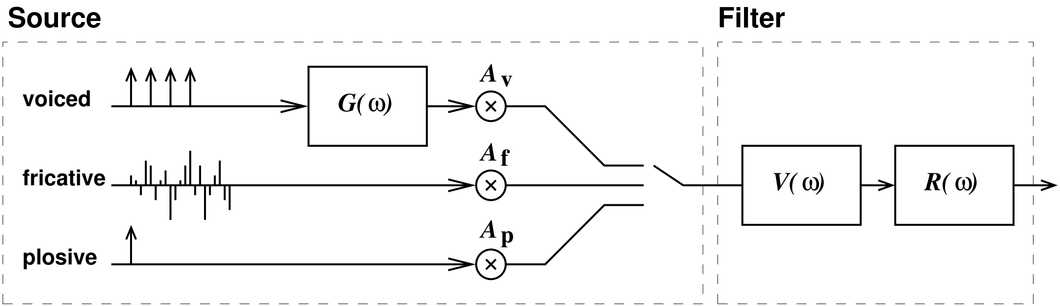
$\leftarrow p[n]$

$\leftarrow p[n] * r[n]$

$\leftarrow p[n] * r[n] * v[n]$



Complete Source/Filter Model



IPA Chart: Consonants

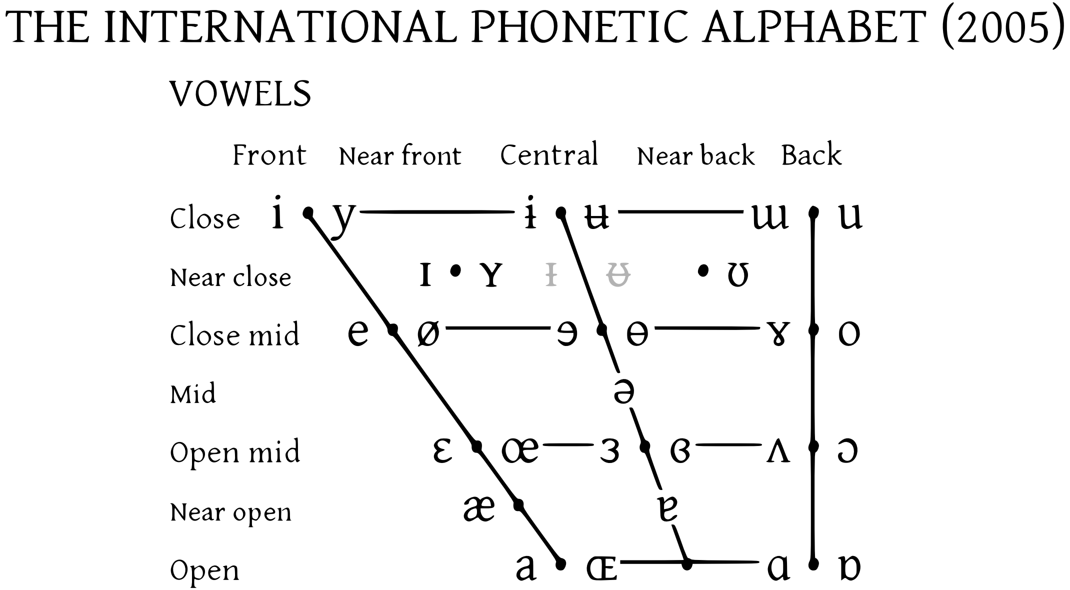
THE INTERNATIONAL PHONETIC ALPHABET (2005)

CONSONANTS (PULMONIC)

	LABIAL		CORONAL				DORSAL			RADICAL		LARYNGEAL
	Bilabial	Labio-dental	Dental	Alveolar	Palato-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ			
Plosive	p b	ȹ ȸ		t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʡ	ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	ʜ ʢ	h ɦ
Approximant		ʋ		ɹ		ɻ	j	ɰ				
Trill	ʙ			r					ʀ		ᴙ	
Tap, Flap		ⱱ		ɾ		ɽ						
Lateral fricative				ɬ ɮ		ꞎ	𝼆	𝼄				
Lateral approximant				l		ɭ	ʎ	ʟ				
Lateral flap				ɺ		𝼈						

Where symbols appear in pairs, the one to the right represents a modally voiced consonant, except for murmured *ɦ*.
Shaded areas denote articulations judged to be impossible. Light grey letters are unofficial extensions of the IPA.

IPA Chart: Vowels



Phonology vs Phonetics

Phonemes

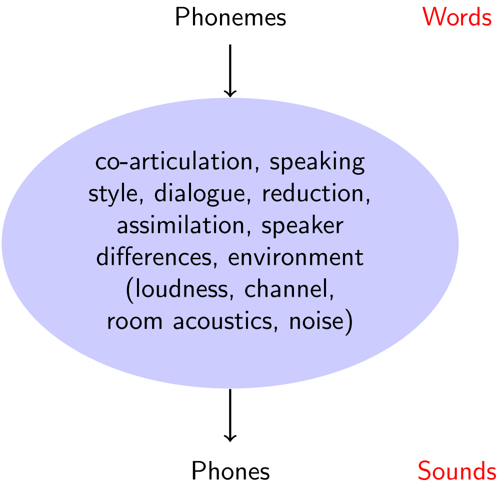
$\downarrow$

co-articulation, speaking style, dialogue, reduction, assimilation, speaker differences, environment (loudness, channel, room acoustics, noise)

$\downarrow$

Phones

Phonology vs Phonetics



Components of ASR System