DT2118
Speech and Speaker Recognition
Introduction

Giampiero Salvi

KTH/CSC/TMH giampi@kth.se

VT 2016

Outline

Outline

Contact Info

Giampiero Salvi (giampi@kth.se)

All communications handled through the course web:
https://www.kth.se/social/course/DT2118/

Course Objectives

after the course you should be able to:

- *implement* simple training and evaluation methods for Hidden Markov Models
- *train* and *evaluate* a speech recogniser using software packages
- *compare* different feature extraction and training methods
- *document* and *discuss* specific aspects related to speech and speaker recognition
- with the help of the literature, *review* and *criticise* other students' work in the subject

Schedule

- Part 1 Introduction, Speech Signal, Features, Statistics (ca 4 hours)
- Part 2 Hidden Markov Models, Training and Decoding, Acoustic Models (ca 4-6 hours)
- Part 3 Decoding and Search Algorithms (ca 2 hours)
- Part 4 Language Models (Grammars) (ca 2 hours)
- Part 5 Noise robustness and Speaker Recognition (ca 2-4 hours)

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 (course book)

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
	HTK tutorial (HTK book)		
Part 3	Basic Search Algorithms	591–643	53
	(Large-Vocabulary Search Algorithms)	(645–685)	(41)
	(Applications and User Interfaces)	(919–956)	(38)
Part 4	Language Modeling	545–590	46
Part 5	Speaker Recognition literature		

(Optional chapters in parentheses)

Requirements/Activities

Grades: **Pass/Fail**
In order to pass you have to:

1. carry out three **labs** and hand in the report
2. carry out **mini-project** in groups and write **report (paper)**
3. **review** other students' report
4. **present** your work at final seminar
5. **discuss** other student's work at final seminar
6. **submit** final version of the paper

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)
- Optional: implement DNNs using Theano, compare with GMM-HMMS

Project report

- Suggest a title or choose a topic from a list
- Project report in form of research paper
- Suggested topics:

Own work and experiments after discussion with the teacher
Limitations in standard HMM and a survey of alternatives
Pronunciation variation and its importance for speech recognition
Language models for speech recognition
New search methods
Techniques for robust recognition of speech
Confidence measures in speech recognition
The role of prosody for speech recognition
Speaker variability and methods for adaptation

Computational Resources at PDC

For Lab 3 and the Project

- apply for an account at https://www.pdc.kth.se/support/accounts/user
- use edu16.DT2118 when asked for time allocation

Important dates

- Thu 14 April: submit Lab 1 report
- Tue 19 April: decide topic for project
- Thu 28 April: submit Lab 2 report
- Thu 12 May: submit Lab 3 report
- Wed 25 May: submit project report (draft). Needed for the peer review.
- Sun 29 May: submit review of other report
- Tue 31 May: Final seminar: present own project results, and discuss others'
- Mon 6 Jun: Final report

KTH Social deadlines are set at 23:55

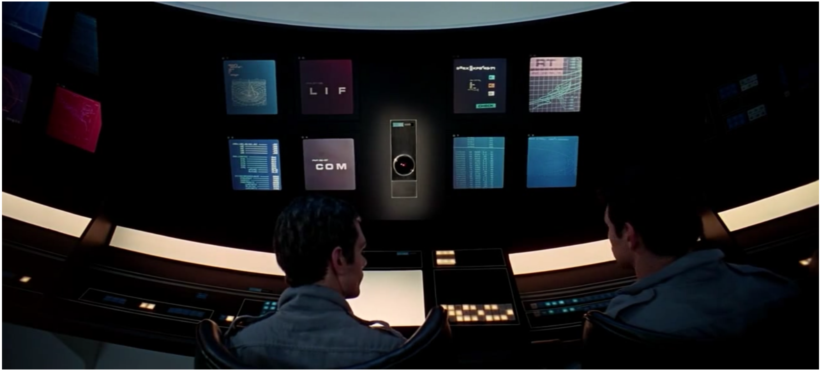
Part 1

Outline

Motivation

- 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)

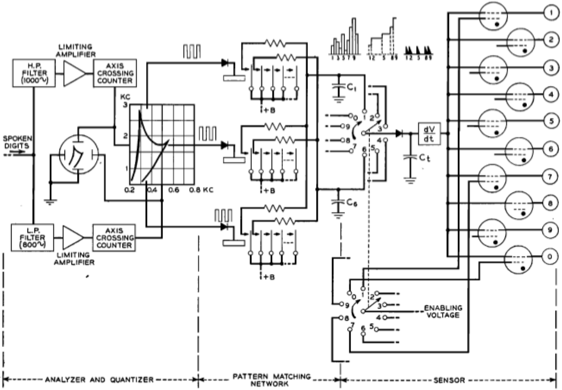
The dream of Artificial Intelligence



2001: A space odyssey (1968)

A very long endeavour

1952, Bell laboratories, isolated digit recognition, single speaker, hardware based [1]

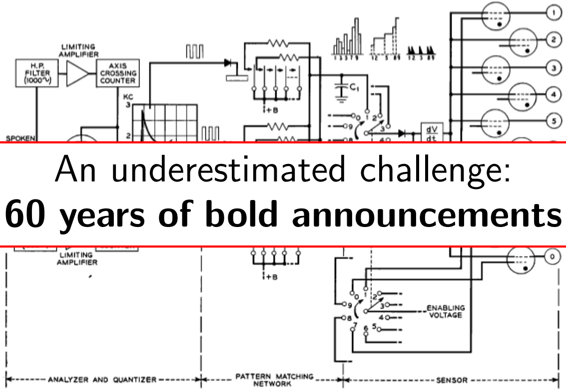


---

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

A very long endeavour

1952, Bell laboratories, isolated digit recognition, single speaker, hardware based [1]



An underestimated challenge:
60 years of bold announcements

[1] K. H. Davis, R. Biddulph, and S. Balashek. “Automatic Recognition of Spoken Digits”. In: *JASA* 24.6 (1952), pp. 637–642

Today's Reality

I Now Pronounce You Chuck & Larry (2007)

The ASR Goal (for this course)

Convert speech into text

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

The ASR Goal (for this course)

Convert speech into text

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

CC Please tell me your name

LV Larry Valentine

CC I'm sorry, I didn't quite get that

LV Larry Valentine

CC You said "Berry Schmallenpine". . . is that right?

LV Schmallenpine?!?!

CC You said "Schmallenpine". . . is that right?

The ASR Goal (for this course)

Convert speech into text

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

CC Please tell me your name

LV Larry Valentine

CC I'm sorry, I didn't quite get that

LV Larry Valentine

CC You said "Berry Schmallenpine"... is that right?

LV Schmallenpine?!?!

CC You said "Schmallenpine"... is that right?

The ASR Goal (for this course)

Convert speech into text

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

CC Please tell me your name

LV Larry Valentine

CC I'm sorry, I didn't quite get that

LV Larry Valentine

CC You said "Berry Schmallenpine"... is that right?

LV Schmallenpine?!?!

CC You said "Schmallenpine"... is that right?

The ASR Goal (for this course)

Convert speech into text

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

CC Please tell me your name

LV Larry Valentine

CC I'm sorry, I didn't quite get that

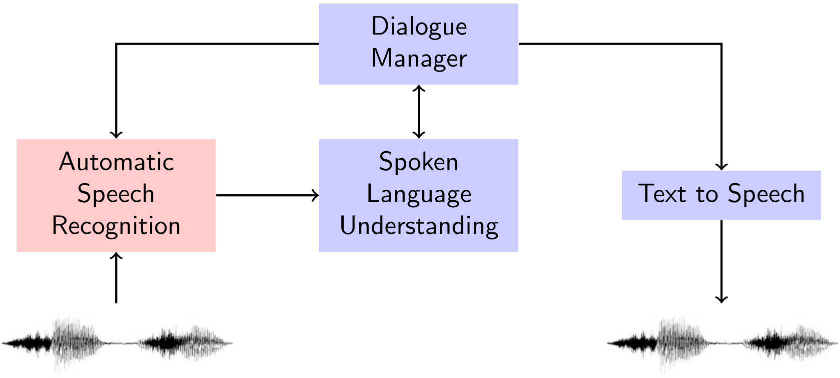
LV Larry Valentine

CC You said "Berry Schmallenpine"... is that right?

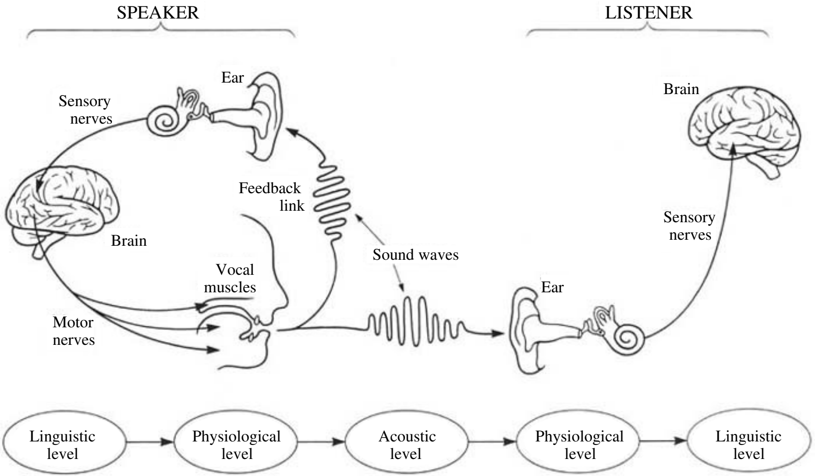
LV Schmallenpine?!?!

CC You said "Schmallenpine"... is that right?

ASR in a Broader Context

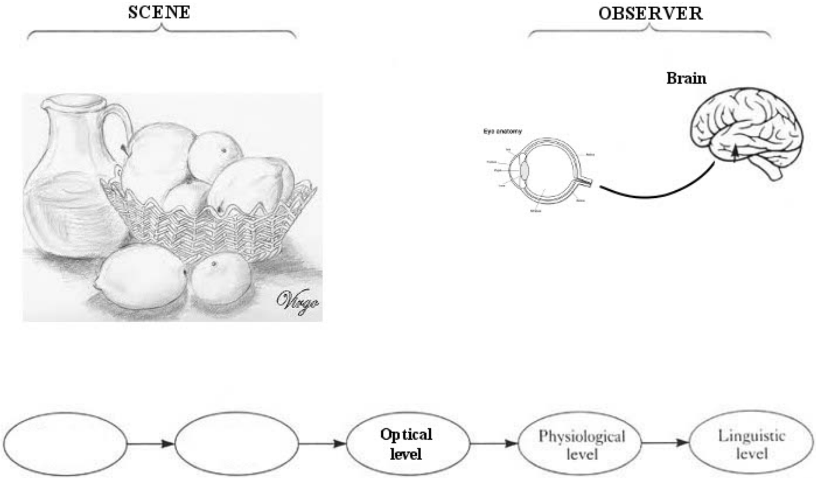


The Speech Chain



Peter Denes, Elliot Pinson, 1963

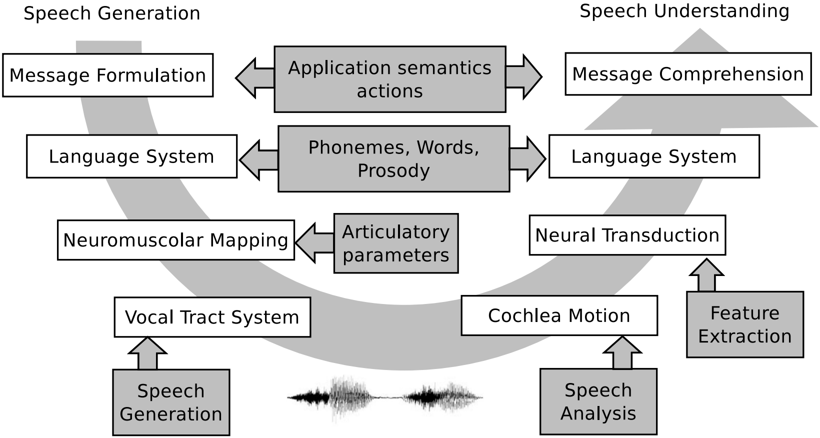
ASR versus Computer Vision



ASR versus Computer Vision

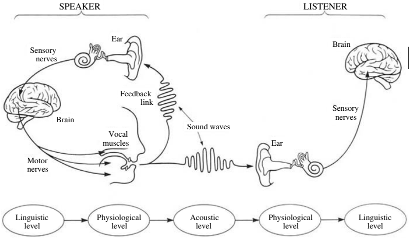
Property	ASR	Computer Vision
signal originates from:	cognition + physics	physics
persistence:	disappears as soon as heard	continually available (active perception)
across countries:	different languages	same objects
type of interaction:	two-way	one-way

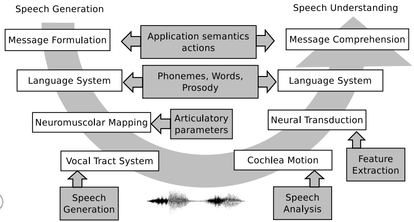
The Speech Chain (from the book)



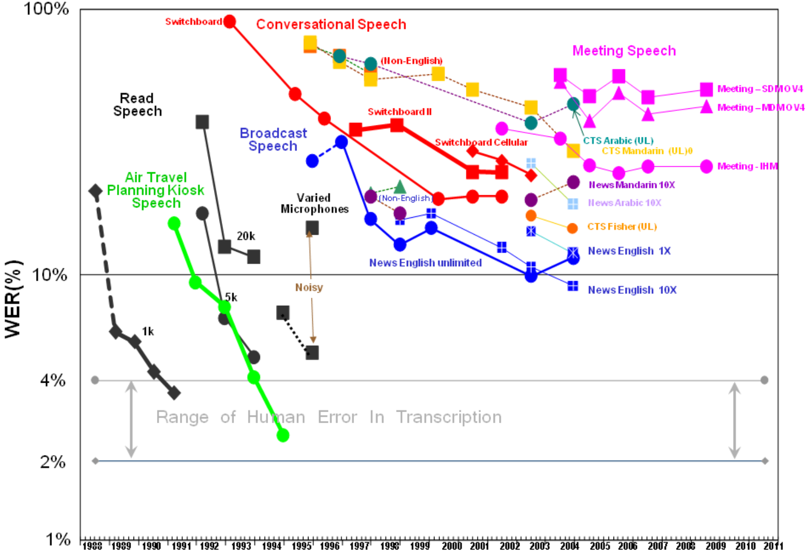
Not covered in this course:

- multimodality
- interaction (bi-directional)
- incrementality
- non-verbal communication





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

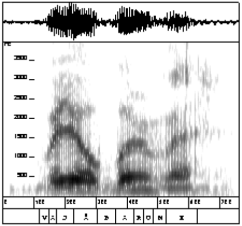
Environment

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

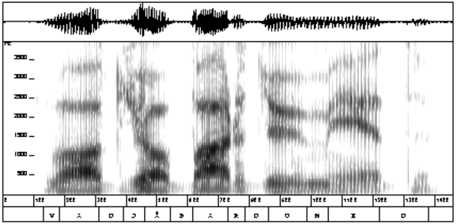
Listener

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

Example: spontaneous vs hyper-articulated



Va jobbaru me



Vad jobbar du med

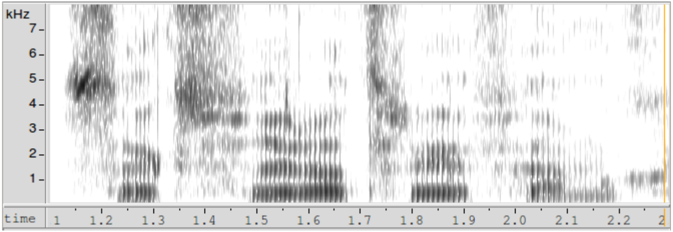
"What is your occupation"
("What work you with")

Examples of reduced pronunciation

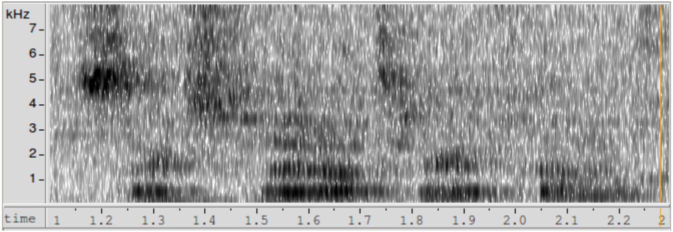
Spoken	Written	In English
Tesempel	Till exempel	for example
åhamba	och han bara	and he just
bafatt	bara för att	just because
javende	jag vet inte	I don't know

Microphone distance

Headset



2 m distance



Applications today

Call centers:

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

Accessibility

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

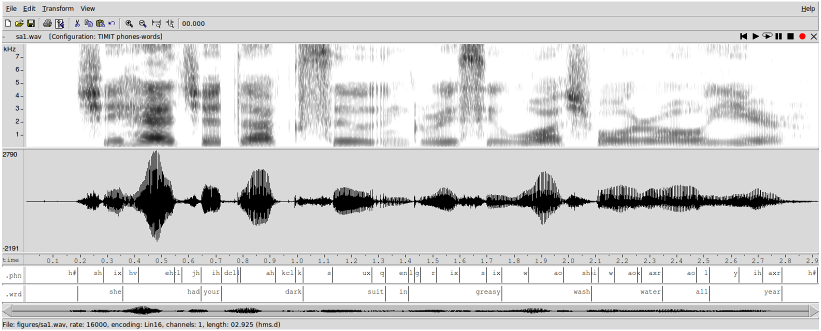
Smart phones

- Siri, Android...

Outline

Speech Examples

TIMIT database (American English)



example of “clean” speech

Elements of Signal Processing

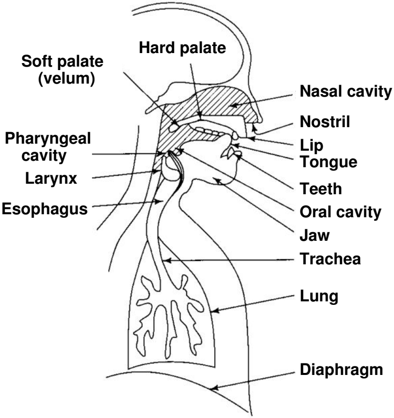
- continuous/digital signals
- Linear and Time Invariant (LTI) systems
- impulse response and convolution
- Fourier transform and transfer function
- sampling theorem
- short-time Fourier transform

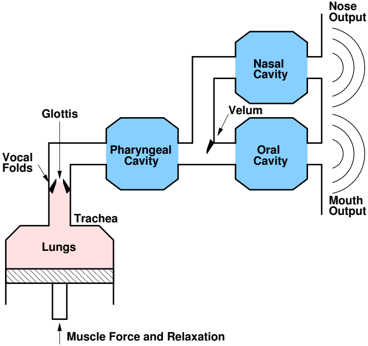
(Chapter 5 in the book)

Speech Examples

live examples

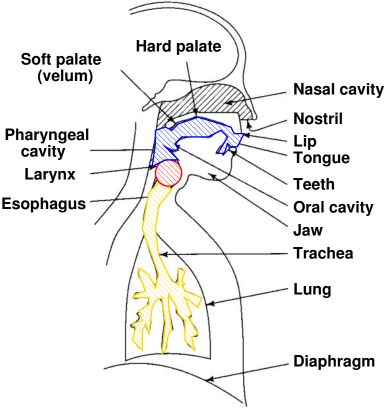
Physiology

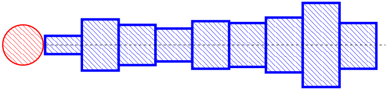

Hard palate
Soft palate
(velum)
Nasal cavity
Nostril
Pharyngeal
cavity
Lip
Tongue
Larynx
Teeth
Esophagus
Oral cavity
Jaw
Trachea
Lung
Diaphragm


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

Source/Filter Model, Vowel-like sounds

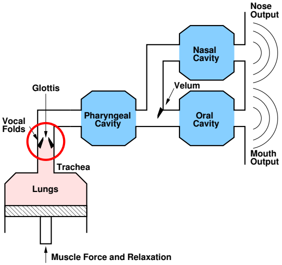
Vowels

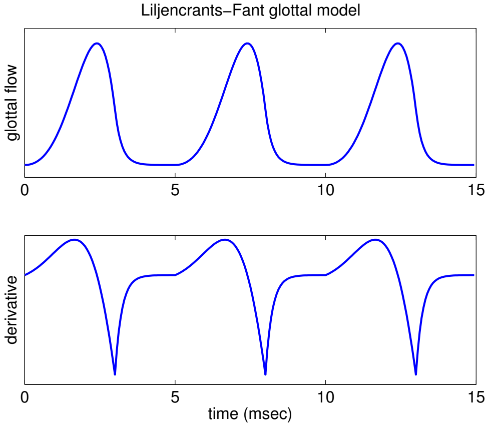




- ☐ Source (periodic)
- ☐ Front Cavity
- ☐ Back Cavity
- ☐ Back Cavity (2nd approx.)

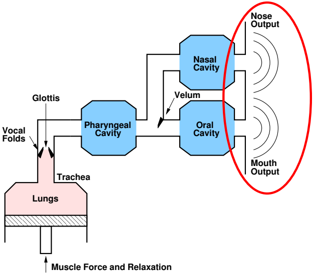
Glottal Flow





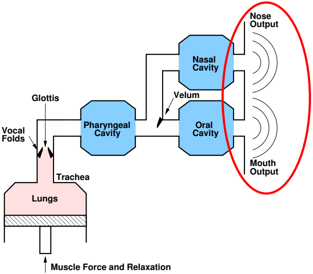
$$G(z) = \frac{1}{(1-\beta z)^2}, \quad \beta < 1$$

Radiation form the Lips/Nose

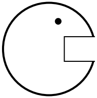


Problem of radiation at the lips plus diffraction about the head too complicated.

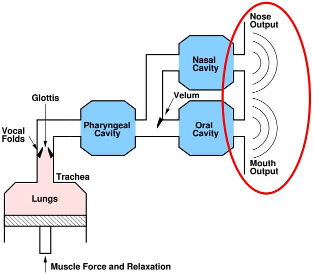
Radiation form the Lips/Nose



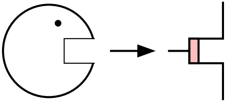
Approx. with a piston in a rigid sphere: solved but not in closed form



Radiation form the Lips/Nose

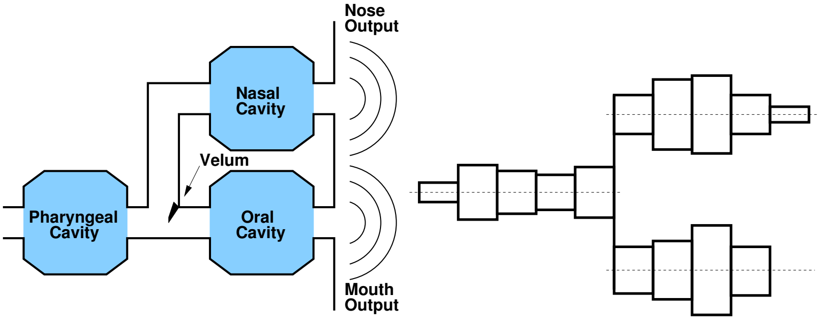


2nd approx: piston in an infinite wall

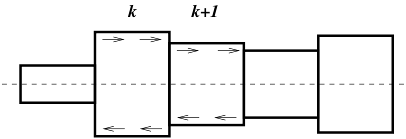


$$R(z) \approx 1 - \alpha z^{-1}$$

Tube Model of the Vocal Tract



Tube Model (cntd.)

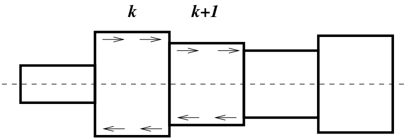


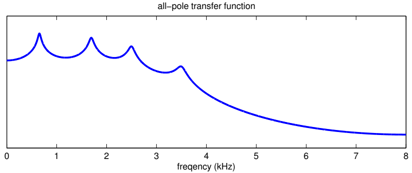
- assume planar wave propagation and lossless tubes
- solve pressure $p(x,t)$ and velocity $u(x,t)$ in each tube according to wave equation
- impose continuity of pressure and velocity at the junctions

$\Rightarrow$ 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.)



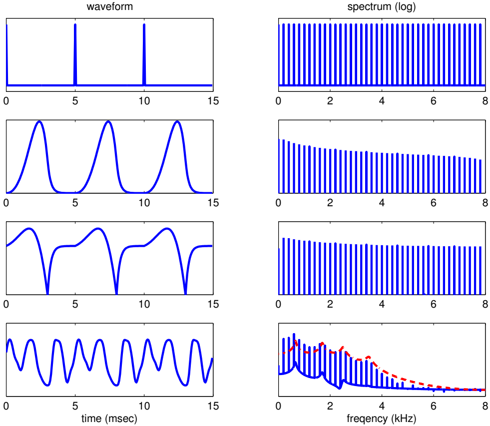


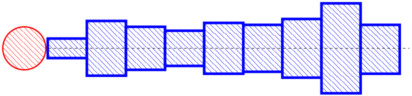
- assume planar wave propagation and lossless tubes
- solve pressure $p(x,t)$ and velocity $u(x,t)$ in each tube according to wave equation
- impose continuity of pressure and velocity at the junctions

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

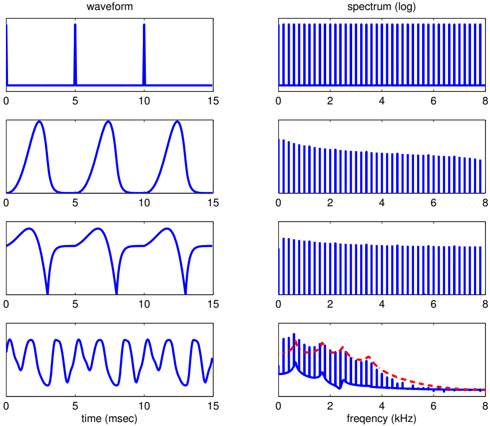
$$V(z) = \frac{Az^{-N/2}}{1 - \sum_{k=1}^{N} a_k z^{-k}}$$

Source/Filter Model: vowel-like sounds

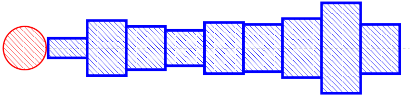




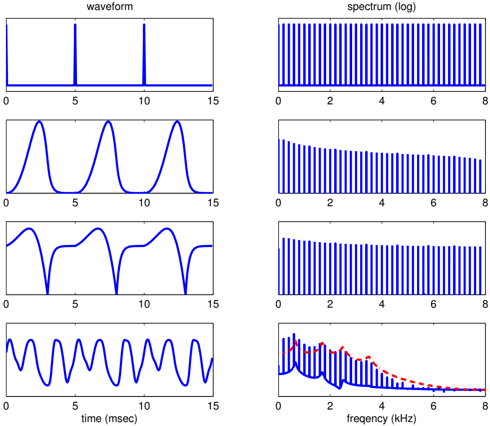
Source/Filter Model: vowel-like sounds



$\leftarrow p[n]$

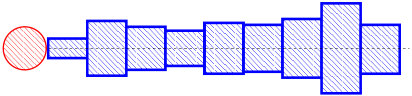


Source/Filter Model: vowel-like sounds

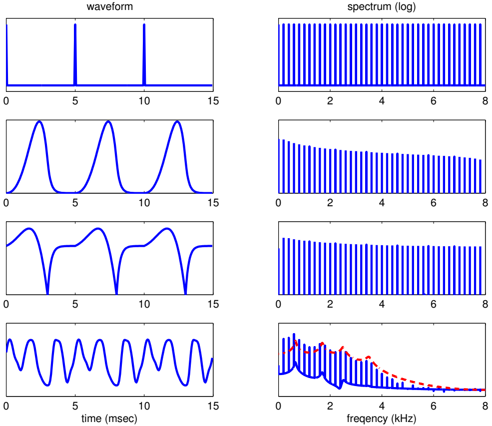


$\leftarrow p[n]$

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



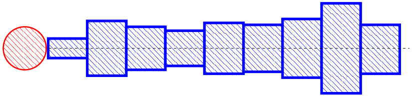
Source/Filter Model: vowel-like sounds



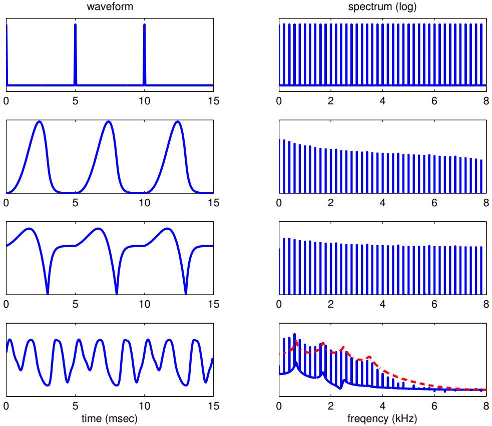
$\leftarrow p[n]$

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

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



Source/Filter Model: vowel-like sounds

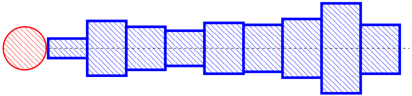


$\leftarrow p[n]$

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

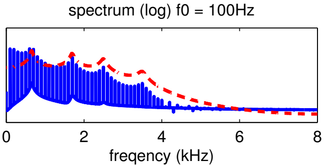
$\leftarrow p[n] * g[n] * r[n]$

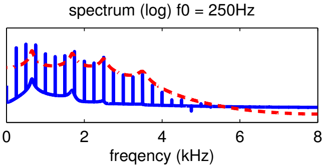
$\leftarrow p[n] * g[n] * r[n] * v[n]$



F_0 and Formants

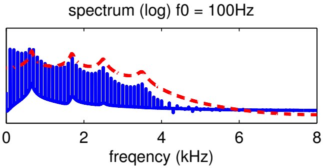
- Varying F_0 (vocal fold oscillation rate)

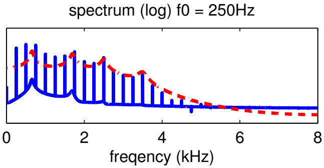




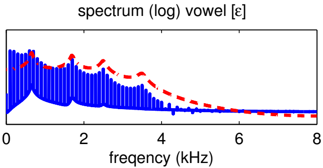
F_0 and Formants

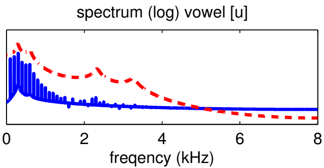
- Varying F_0 (vocal fold oscillation rate)





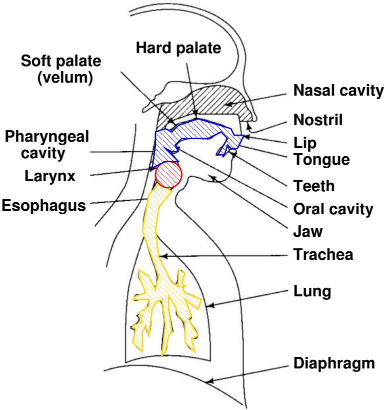
- Varying Formants (vocal tract shape)

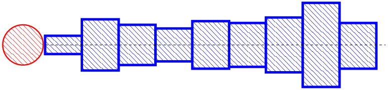




Source/Filter Model, General Case

Vowels

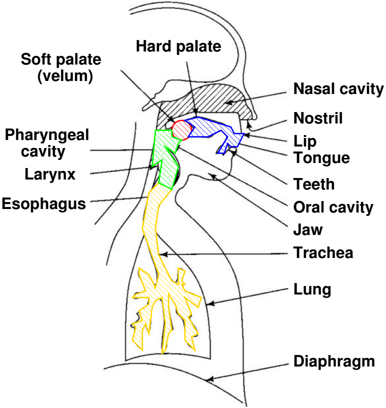


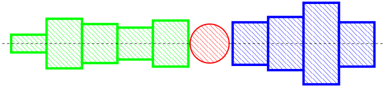


- Source (periodic)
- Front Cavity
- Back Cavity
- Back Cavity (2nd approx.)

Source/Filter Model, General Case

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

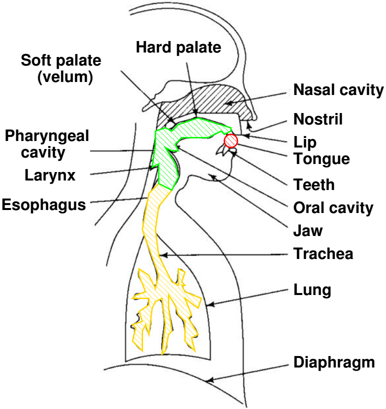


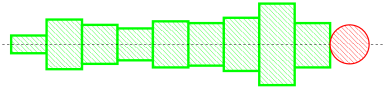


- ☐ Source (noise or impulsive)
- ☐ Front Cavity
- ☐ Back Cavity
- ☐ Back Cavity (2nd approx.)

Source/Filter Model, General Case

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

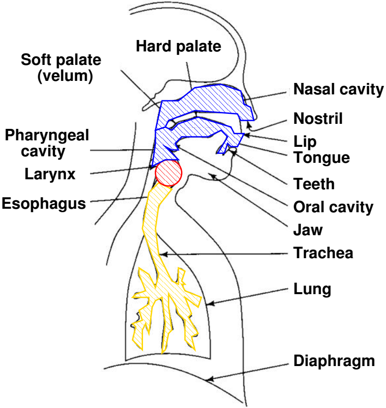


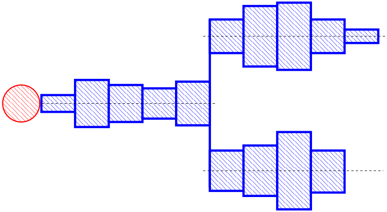


- ☐ Source (noise or impulsive)
- ☐ Front Cavity
- ☐ Back Cavity
- ☐ Back Cavity (2nd approx.)

Source/Filter Model, General Case

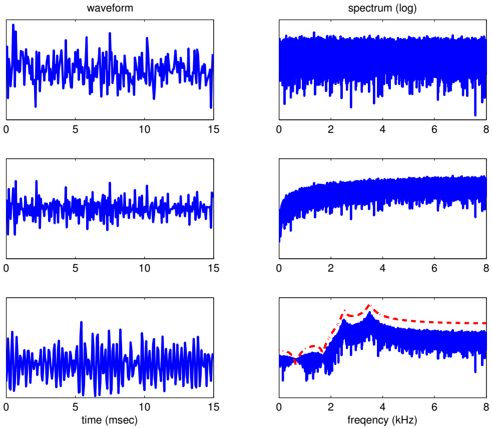
Nasalised Vowels





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

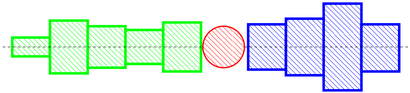
Source/Filter Model: fricative sounds



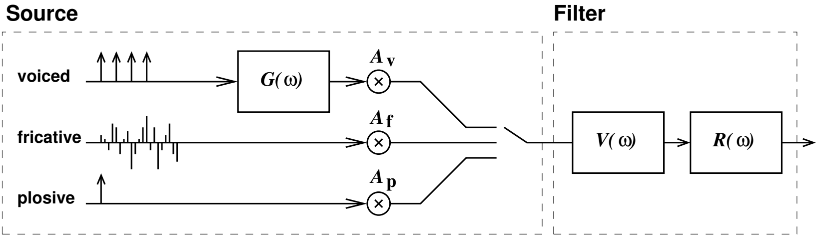
$\leftarrow p[n]$

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

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



Complete Source/Filter Model



IPA Chart: Consonants

THE INTERNATIONAL PHONETIC ALPHABET (2005)

CONSONANTS (PULMONIC)

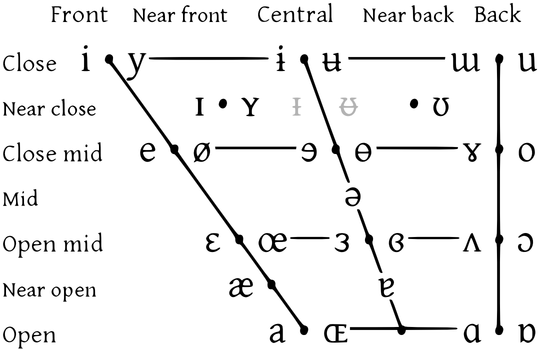
	LABIAL		CORONAL				DORSAL			RADICAL		LARYNGEAL
	Bilabial	Labio-dental	Dental	Alveolar	Palato-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ			
Plosive	p b	ȹ ȸ		t d		ʈ ɖ	c ɟ	k g	q ɢ		ʡ	ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	ʜ ʢ	h ɦ
Approximant		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

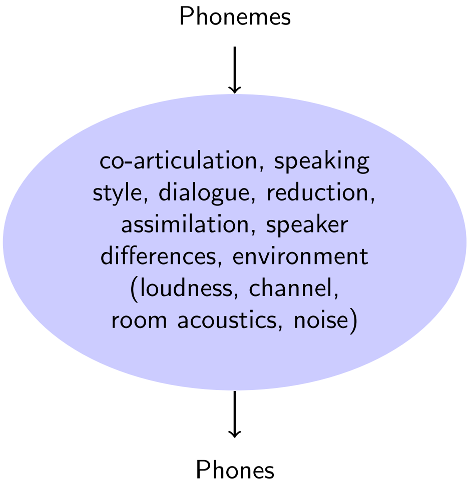
THE INTERNATIONAL PHONETIC ALPHABET (2005)

VOWELS



Vowels at right & left of bullets are rounded & unrounded.

Phonology vs Phonetics



Phonology vs Phonetics

Phonemes Words

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

Phones Sounds

Components of ASR System