11-751
Speech Recognition and Understanding

Phonetics and Phonology

Florian Metze

August 28, 2013

Carnegie Mellon

Agenda

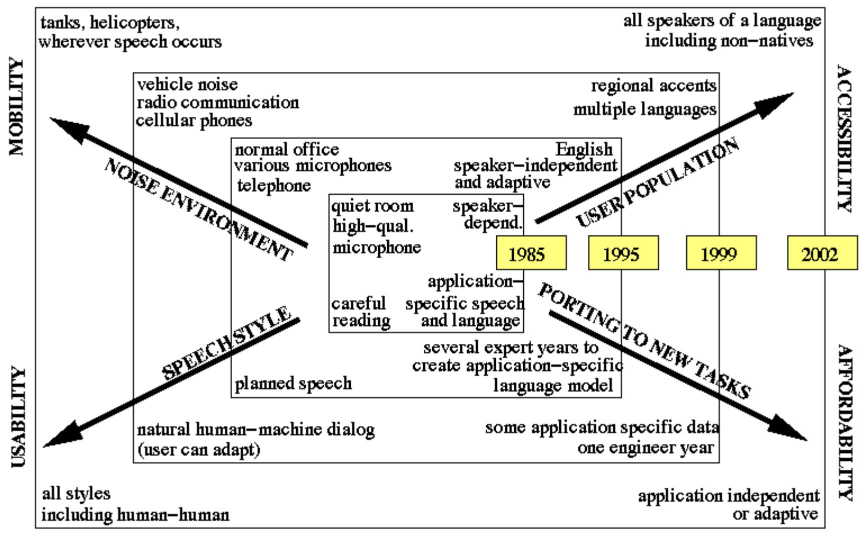
- What is Sound?
- Speech Production
- Signal Processing
- What is Speech – Speech Perception
- Phonetics, Phonology
- Words and beyond

Recap: Speech Research

Some Historic Milestones

1939	Voice Operated ReCorDER (Vocoder) by Dudley (Bell)
1946	"Visible Speech" by Bell (to teach deaf)
1965	Paper by Cooley & Tukey: "The Fast Fourier Transform"
1968	DTW for speech recognition by Vintsyuk
1971	DARPA starts ambitious speech understanding project (SUR)
1975	Statistical Models (HMMs) first proposed by J. Baker at CMU
1988	Speaker-independent continuous speech (>1000 words)
1992	Large vocabulary (isolated) vocabulary dictation
1995	Speaker-independent continuous speech (>60.000 words)
1997	Commercially available LVCSR >60.000 words (IBM, Dragon)
2000	Speech-to-speech translation for compact domains (Verbmobil)
2002	General English recognition in noisy environment (DARPA, RT)
2004	Speech Translation on a PDA (Transtac–DARPA)
2005	GALE (Global Autonomous Language Exploitation) that targets speech recognition, translation, and information extraction of unlimited domains in multiple languages
2009	Jibbigo – commercial mobile speech-to-speech translation

Resulting Research Directions



When Is a Recognizer Good?

Typical criteria for the evaluation of modern large vocabulary recognisers are

Word-Error-Rate: WER = #Errors / #Spoken_Words

Word Accuracy: WA = 1 – WER

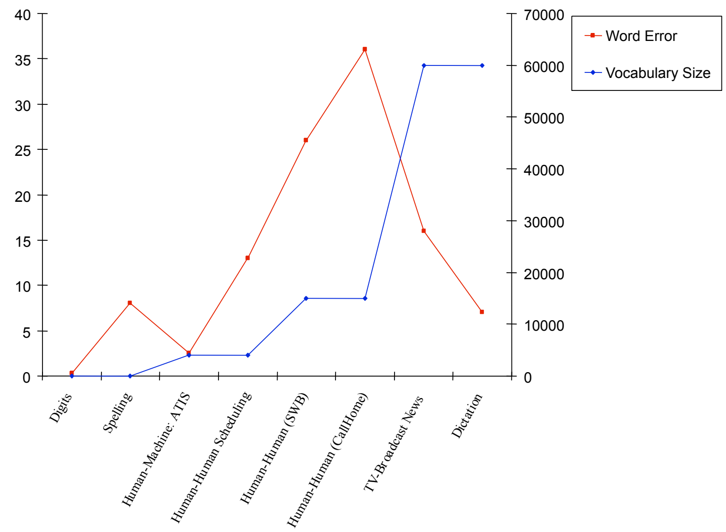
#Errors = #substitutions + #deletions + #insertions (alignment errors taken into account for WER)

Example: WER = ¾ = 75%

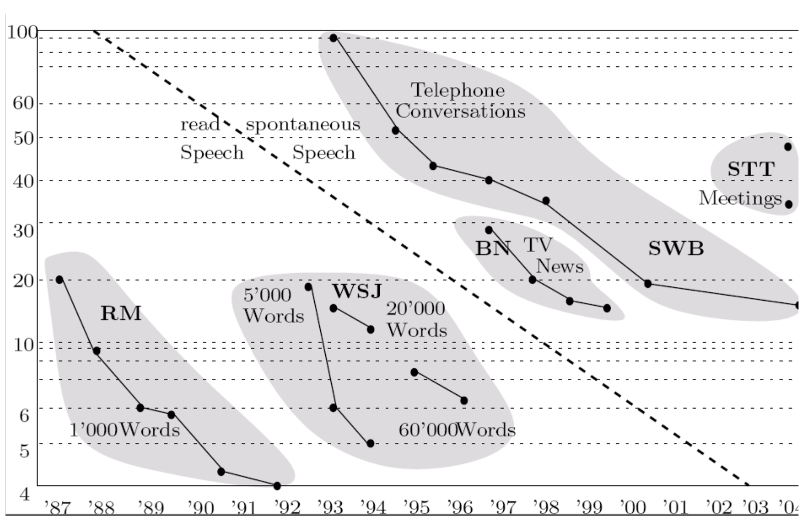
Reference:	`SHOW ME THE INTERFACE`
Hypothesis:	`I SHOW ME FACE`
Alignment:	`I D S`

Alignment is not unique, but error count is (`FACE` could also be aligned with `THE`)

Difficulty of Speech Tasks



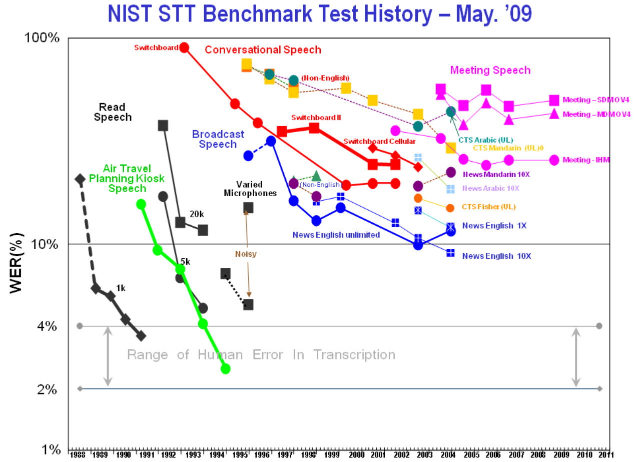
NIST Evaluations: State-of-the-Art



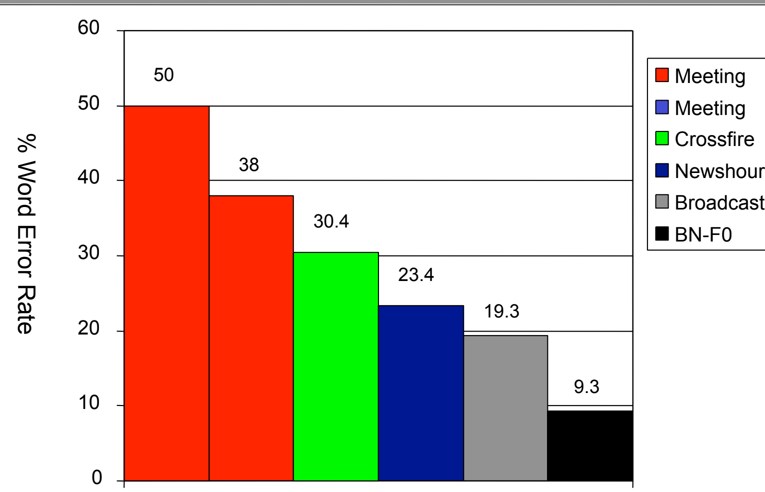
The performance improves on every task – new tasks keep on coming.

NIST Evaluations (2)

NIST STT Benchmark Test History – May. '09



Recognition of Speech in Mixed Meetings



The same recognizer delivers different performance on different data!

How Do Computers Fare Against Humans?

interACT

Tasks (English)	Vocabulary	Performance Humans	Machines
Connected Digits	10	0.009%	0.72%
Alphabet Letters	26	1%	5%
Read speech (WSJ)	5.000	0.9%	3%
WSJ noise (10db)	5.000	1.1%	8.6%
Conversational Telephone Task	25.000	3.8%	21%
Broadcast News (04)	100.000	(3% transcribers)	4%
Broadcast Conversations (noise, crosstalk)	100.000	4%	25-30%
Clean speech based on 3-gram	20.000	7.6%	4.4%

1) Humans at least 5 times better than machines, far more robust in noise and conv. env. (2005)!
2) Same syntactic and semantic model > the difference disappears (Microsoft, 2001)

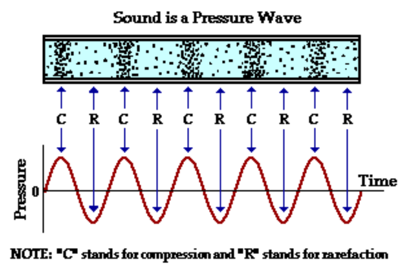
interACT

What is Sound?

What Is Sound?

interACT

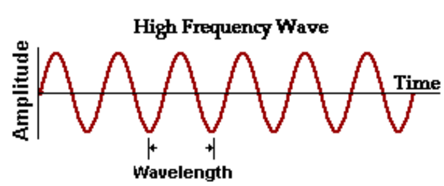
- Sound is a pressure wave which is created by a vibrating object.
- This vibrations set particles in the surrounding medium (typical air) in vibrational motion, thus **transporting energy** through the medium.
- Since the particles are moving in parallel direction to the wave movement, the sound wave is referred to as a **longitudinal** wave.
- The result of longitudinal waves is the creation of **compressions** and **rarefactions** within the air.
- The alternating configuration of C and R of particles is described by a **sine wave** (C~crests, R~troughs)
- The speed of a sound pressure wave in air is $331.5+0.6T_c$ m/s , T_c temperature in Celsius
- The particles do not move down the way with the wave, but oscillate back and forth about their individual equilibrium position.

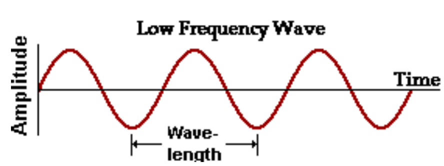


Wave Length, Amplitude, Frequency

interACT

- The amount of work done to generate the energy that sets the particles in motion is reflected in the <u>degree of displacement</u> which is measured as the **amplitude** of a sound.
- The **frequency** f **of a wave** is measured as the <u>number of complete back-and-forth vibrations</u> of a particle of the medium per unit of time.
  1 Hertz = 1 vibration/second
  f **= 1/time**
- Depending on the medium, sound travels at some speed *c* which defines the **wavelength** l**:** $l = c/f$





Picking up the Sound Pressure Wave

Pressure:

- air pressure is measured in Pascal (Pa) (1Pa = $1N/m^2 = 1kg/m \cdot s^2$)
- the standard air pressure on earth's surface is around 100000 Pa
- the softest audible sound modulates the pressure by 0.000001 Pa

$\Longrightarrow$ Standard pressure is 100 billion times higher than the softest audible sound

Challenge:

- Sound pressure decreases linearly with distance from the sound source
- Speed of excited air molecules decreases with the square of distance

$\Longrightarrow$ Problems for recording speech using a microphone (or hearing)

Audibility: Speech needs to be perceivable by microphone

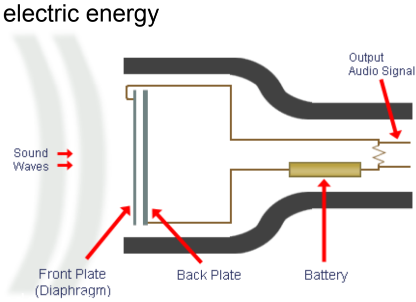
- Interference: Speech disturbs others (no speaking in libraries, theaters, meetings)
- Privacy: Speech signal can be captured by others (no confidentiality in public places)

Robustness: Signal is corrupted by noisy environment

Condensor Microphones

Microphones convert the acoustic energy (sound wave) into electrical energy (energy wave in voltage)

- They measure the molecular speed or the sound pressure
  - Close-speaking vs far-field,
  - Unidirectional vs omnidirectional
- Most popular type is the condenser microphones, i.e. they use a capacitor to convert acoustic into electric energy
- Capacitor has two plates with a voltage between them (requires battery or external phantom power)
- One plate is made of light material and acts as the diaphragm. It vibrates when struck by sound waves, changing the distance and thus changing the capacitance.



Measuring the Intensity of Sound

interACT

- The softest audible sound modulates the air pressure by ~ 10^{-5} Pascal (Pa – Pressure Unit: $1Pa = 1N/m^2 = 1\ kg / m \cdot s^2$)
- The loudest (pain inflict) audible sound does it at ~ 10^2 Pa
- Because of this wide range it is convenient to measure sound amplitude on a logarithmic scale in ***Decibel [dB]***.
- Decibel is no physical unit, it expresses a *relation* for **comparing the intensity** of two sounds I and I_o: $10 \log_{10} (I/I_o)$ (e.g. a channel amplifies the sound by 3 dB = output is 3 dB louder than the input)
- The sound pressure level (SPL) is a measure of absolute sound pressure P in dB: $20 \log_{10} (P/P_0)$, where $P_0 = 2 \cdot 10^{-5}$ Pa, which corresponds to the threshold of hearing = 0dB
- Thus +20 dB denotes a pressure increase by a factor of 10, and an intensity increase by a factor of 100
- Face-to-face speech conversation (1feet away) is ~ 60dB SPL,
- Close-talking microphone ~ 1Pa = 94dB

Examples for Sound Levels in Decibel

interACT

threshold of hearing (TOH)	0 dB	softest audible 1000 Hz sound	6 dB
quiet living room	20 dB	soft whispering	25 dB
refrigerator	40 dB	soft talking	50 dB
normal conversation	60 dB	busy city street noise	70 dB
passing motorcycle	90 dB	somebody shouting	100 dB
pneumatic drill	100 dB	helicopter	110 dB
loud rock concert	110 dB	air raid siren	130 dB
pain threshold	120 dB	gunshot	140 dB
rocket launch	180 dB	instant perforation of eardrum	160 dB

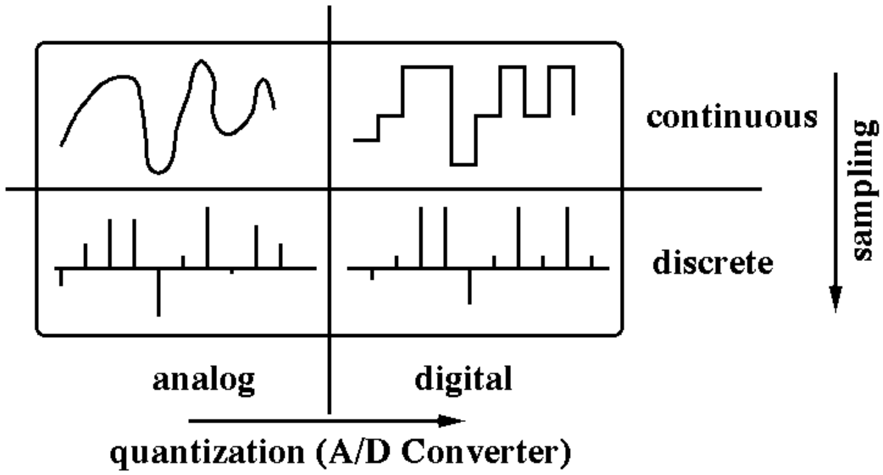
1) TOH: One-billionth of a centimeter of molecular motion
2) The most intense sound (without physical damage) is one trillion times more intense

What is Speech?

Representation of Speech

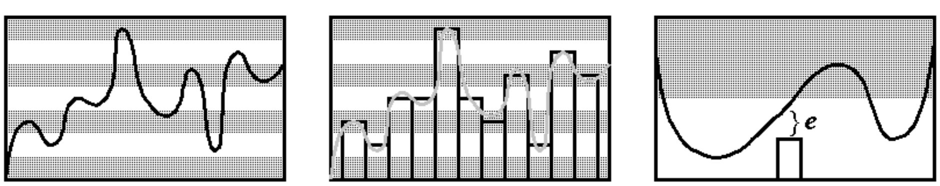
- By definition: Digital representation of speech
  - Represent speech as a sequences of numbers
  - As a prerequisite for automatic processing using computers
- Direct representation of speech waveform:
  - Represent speech waveform as accurately as possible, so that an acoustic signal can be reconstructed
- Parametric representation
  - Represent a set of properties/parameters wrt. a certain model
- Decide the targeted application first
  - Speech coding
  - Speech synthesis
  - Speech recognition

Different Kinds of (Speech) Signals



Quantization of Signals

- Given a discrete signal $f[i]$ to be quantized into $q[i]$
- Assume that f is between f_{min} and f_{max}
- Partition y-axis into a fixed number n of (equally sized) intervals
- Usually $n=2^b$, in ASR typically b=16 > n=65536, then
  - $q[i]$ can only have values that are centers of the intervals
  - **Quantization:** assign $q[i]$ the center of the interval in which lies $f[i]$



- Quantization makes errors, i.e. adds noise to the signal f[i]=q[i]+e[i]
- The average **quantization error** $e[i]$ is $(f_{max}-f_{min})/(2n)$
- Define **signal to noise ratio** SNR[dB] = $E\{f^2[i]\} / E\{e^2[i]\}$

Quantization of Speech Signals

interACT

Choice of sampling depth:

- Dynamics of speech signals are usually in the range of 50 to 60 dB
- The lower the SNR (signal-to-noise ratio), the lower ASR performance
- To get a reasonable SNR, *b* should be at least 10 to 12
- Typically in ASR the samples are quantized with 16 bits

Speech signals' amplitudes are not equally distributed

- Speech amplitudes are exponentially distributed around their mean
- So the information-theoretic optimum is not to partition the *y*-axis into equal intervals
  - use ***μ*-law** encoding:
  - $f^{(\mu)}[n] = f_{max} \cdot \mathrm{sgn}(f[n]) \cdot \log(1+\mu|f[n]|/f_{max}) / \log(1+\mu)$
    $\mu=100,\ldots,500$
- μ-law encoding makes speech amplitudes equally distributed before quantization → 8 bit sufficient

Speech Coding

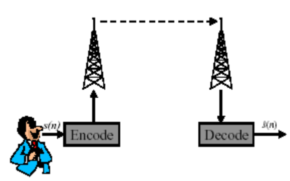
interACT

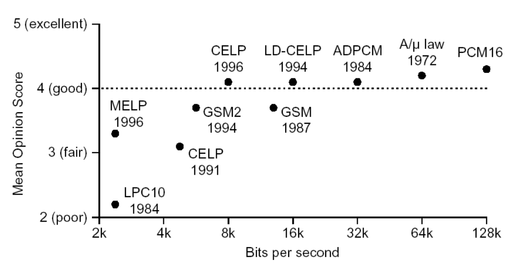
Objectives of Speech Coding:

- Quality versus bit rate
- Quantization Noise
- High measured intelligibility
- Low bit rate (b/s of speech)
- Low computational requirement
- Robustness to transmission errors
- Robustness to successive encode/decode cycles

Objectives for real-time:

- Low coding/ decoding delay
- Work with non-speech signals (e.g. touch tone)

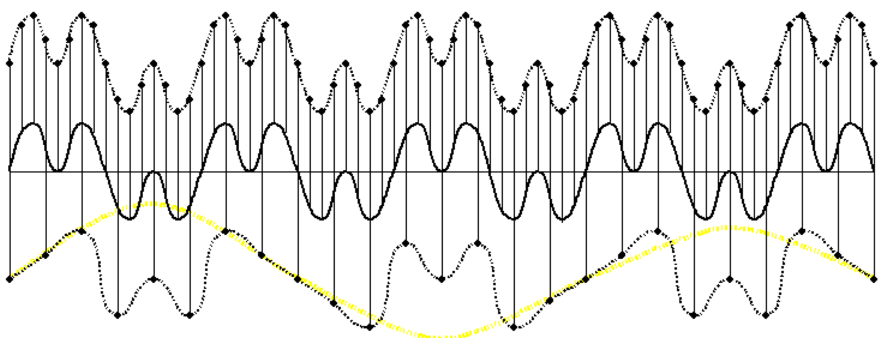




The Aliasing Effect

Nyquist theorem: When a f_l-band-limited signal is sampled with a sampling rate of at least $2f_l$ then the signal can be exactly reproduced from the samples

- When the sampling rate is too low, the samples can contain "incorrect" frequencies:

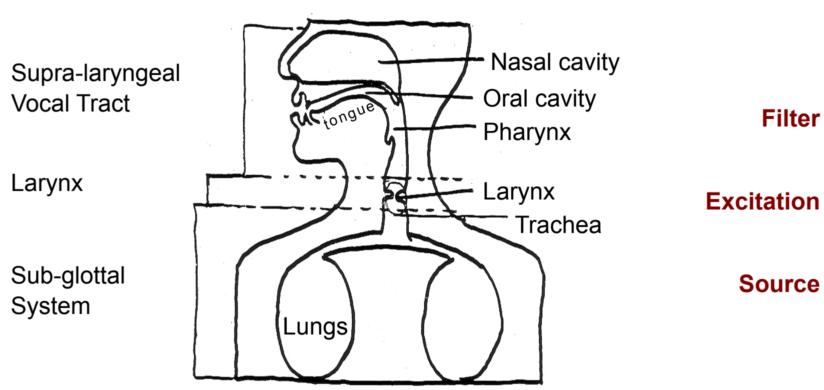


Acoustic Features in the Sampled Signal

- The **amplitude** and the **envelope** of the amplitude
- The envelope of the amplitude is correlated to the **power** (= integral of the squared signal)
- Power is useful for detecting speech vs. silence, also syllables, phrase boundaries, prosody
- **Peak to peak**
  - (correlated to amplitude)
- **Zero-crossing** rate
  - (can help distinguish some weak sounds from silence)
- **Voiced/ Unvoiced**
  - (homogeneous periodic signal vs. white noise in fricatives)

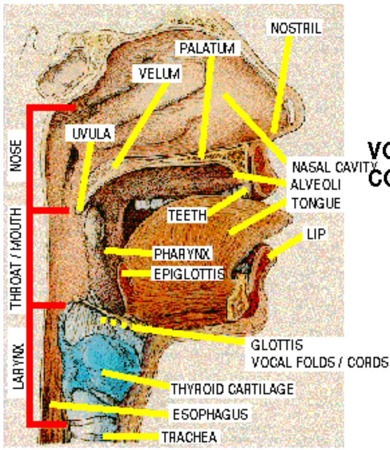
Speech Production

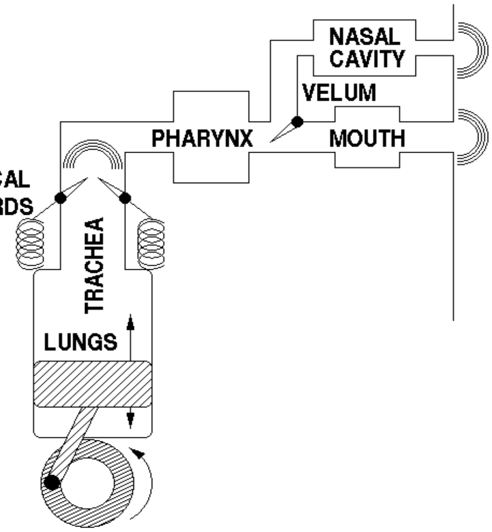
Speech Production



The three physiologic components of human speech production

Models of the Vocal Tract



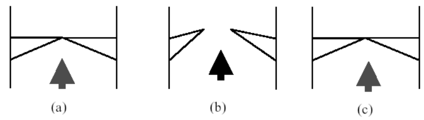


Models of Speech Production

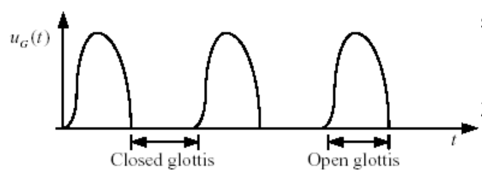
- The acoustic theory analyzing the physics of the propagation of sound waves through the vocal tract should consider:
  - three dimensional wave propagation
  - variation of the vocal tract shape with time
  - viscous fiction at the walls,
  - softness of the tract walls,
  - radiation of sound at the lips,
  - nasal coupling,
  - excitation of sound
- Model that considers all of the above is not yet available, but some models provide good approximation and good understanding of physics involved
  - Glottal Excitation Model
  - Lossless Tube Concatenation
  - Source Filter Models

Glottal Excitation Model

Remember the oscillation for voiced sounds:



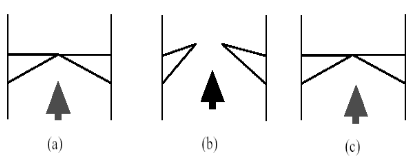
- When the vocal cords (glottis) are together airflow from lung to vocal tract is blocked
- The *pressure* builds up behind the vocal cords
- At a certain pressure the vocal cords are forced to open and allow air to flow through the glottis
- When pressure in the glottis falls, it allows the cords to come together and the cycle is repeated
- In the *closed-phase (*glottis is closed) the *volume velocity* is zero
- In the *open-phase (*non-zero volume velocity) the lungs and the vocal tract are coupled



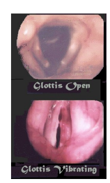
Rosenberg's glottal model defines the shape of the volume velocity u_G with the:

1) open quotient ratio of pulse duration to pitch period
2) speed quotient: ratio of rising to falling pulse durations

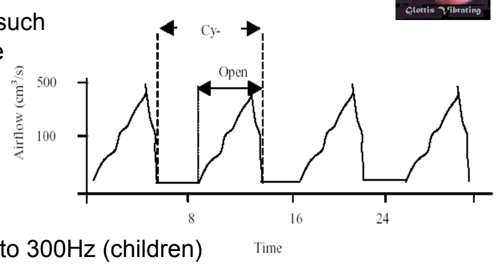
Fundamental Frequency



Vocal fold cycling: (a) closed vocal folds build a barrier for the air stream from lungs; (b) air pressure under the barrier overcomes the resistance of the vocal fold closure and blows them apart; (c) elasticity in tissue and muscles make them fall back into place

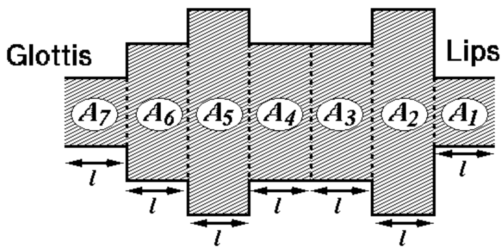


- The time for a single open-close cycle depends on the stiffness and size of the vocal folds and the amount of air pressure. This can be controlled (in some range) by a speaker to raise and lower the pitch of a voiced sound
- We can measure the number of such cycles per second. It is called the **fundamental frequency F_o**
- It sets the periodic baseline for all higher-frequency harmonics by the resonance cavities above



- F_o varies from 60Hz (large men) to 300Hz (children)

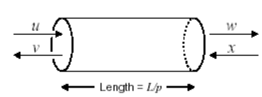
Lossless Tube Concatenation Model



- Idea: The vocal tract can be represented a s a concatenation of *p* lossless tubes
- It consists of a series of cylinders (Helmholtz-Resonator) of equal length *l*
- The cross-sections A_i approximate the area function *A(x)* of the vocal tract
- If *p* is large, and *l* is short, the frequency response is expected to be close to those of tubes with continuously varying area functions
- For waves with wavelength >> dimensions of the vocal tract, waves propagate along the axis of the tubes
- Further assume: No loss due to viscosity or thermal conduction, and area *A* remains constant over time

Lossless Tube Concatenation Model (2)

- Acoustic Signal is a superposition of two waves *u* in for *v* in reverse direction

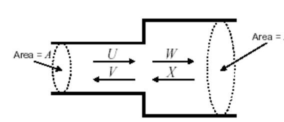


- If we pick l=L/p as tube length, then the time to travel along the tube is = L/cp and hence:

$$v(t)=x\left(t-\frac{L}{cp}\right) \quad \text{and} \quad u(t)=w\left(t+\frac{L}{cp}\right)$$

- If we take z-transforms, then the time delay corresponds to multiplying by $z^{-1/2}$, so:

$$V(z)=z^{-1/2}X(z) \quad \text{and} \quad U(z)=z^{+1/2}W(z)$$



- Reflexion Coefficients r: $r=\frac{B-A}{B+A}$
- Vocal Tract transfer function for a 2-segmental tube (glottal U_g > lips U_l)

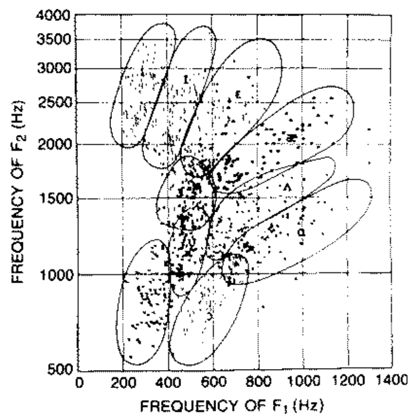
$$\frac{U_l}{U_g}=\frac{\prod_{k=0}^{2}(1+r_k)\times z^{-1}}{1+(r_0r_1+r_1r_2)z^{-1}+r_0r_2z^{-2}}=\frac{Gz^{-1}}{1+(r_0r_1+r_1r_2)z^{-1}+r_0r_2z^{-2}}=\frac{Gz^{-1}}{1-a_1z^{-1}-a_2z^{-2}}$$

- For a p-segmental vocal tract we get: denominator is a p-th order all pole filter

$$V(z)=\frac{U_l}{U_g}=\frac{Gz^{-1/2\,p}}{1-a_1z^{-1}-a_2z^{-2}-\ldots-a_pz^{-p}}$$

Formants

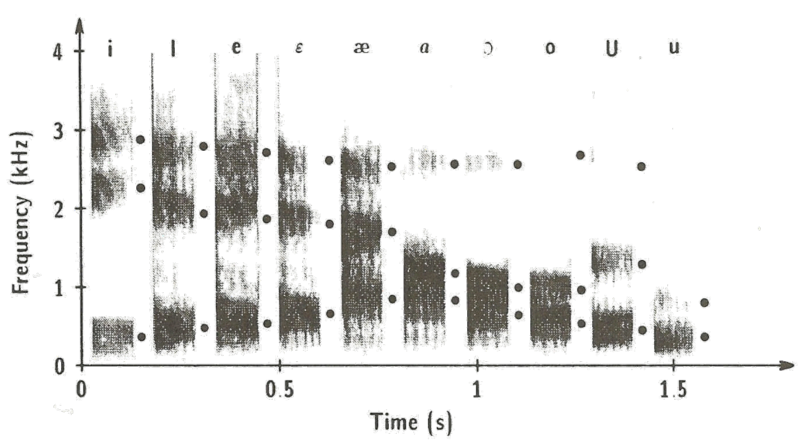
- In general the concatenation of p lossless tubes results in an p-pole system with p/2 conjugated poles at the most.
- These poles are called resonances or formants.
- They occur when a given frequency "gets trapped".
- Relationship between p and F:
  p = 2 L F / c
- Example: F=8000Hz, c=340m/s, L=17cm (adult vocal tract)
  ⇒ p=8 ⇒ 4 formants
- Experimentally shown, the vocal tract transfer function (of a male adult) has about 1 formant per kilohertz



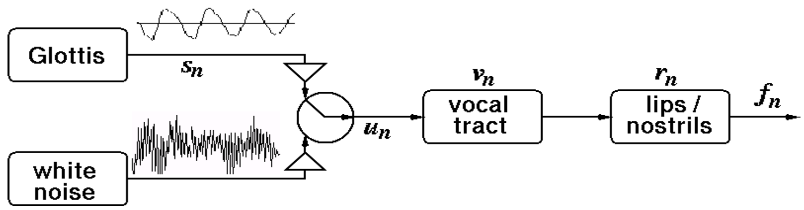
Formants measured for vowels from different speakers

Spectrograms

Short-time Fourier Analysis of Speech Signal



The Source-Filter Model of Speech Production



- Sounds are produced by either
  - vibrating the vocal cords (voiced sounds) or
  - random noise resulting from friction of the airflow (unvoiced sounds)
  - voiced fricatives need a mixed excitation model
- The signal u_n is modulated by the vocal tract, whose impulse response is v_n
- Resulting signal is modulated by the lips' and nostrils' radiation response r_n.
- Eventually, the resulting signal f_n is emitted.
- The modulation y_n of a signal x_n by a channel c can be computed as the **convolution** of the signal with the channel's impulse response $y_n = x_n * c_n$
  Thus: $\boldsymbol{f_n = u_n * v_n * r_n}$

Convolution

The convolution of two functions *f*(*x*) * *g*(*x*) is defined as:

$$(f \star g)(t) = \int_{-\infty}^{\infty} f(\tau - t) \cdot g(\tau) \mathrm{d}\tau$$

or in the discrete case:

$$(f \star g)[t] = \sum_{\tau=-\infty}^{\infty} f[\tau - t] \cdot g[\tau]$$

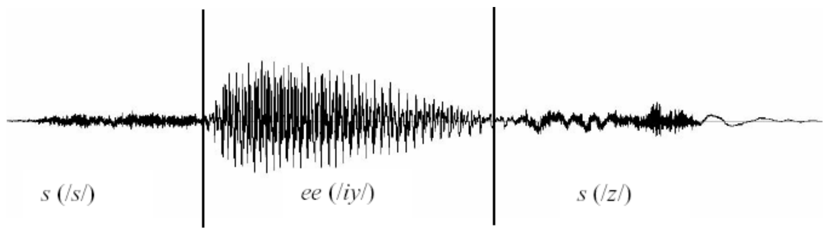
Let *F* be a *z*-transform of *f*, and let *G* be the corresponding *z*-transform of *g*, then the *z*-transform of (*f***g*) at *t* simply is $\boldsymbol{F(t) \cdot G(t)}$

Now the inverse transform of $F \cdot G$ is the convolution of *f* and *g*.

Definition of the z-transform of a digital signal h[n]: $H(z) = \sum_{n=-\infty}^{\infty} h[n] z^{-n}$

Voiced vs. Unvoiced Sounds

- The most fundamental distinction between sound types in speech is the voiced/voiceless distinction
- Voiced sounds have a roughly regular pattern in time and frequency structure, voiceless sounds do NOT have this
- Mechanism: if the vocal cords vibrate during articulation the sound is voiced
- Waveform of the words *sees* consisting of an unvoiced consonant /s/, a vowel /iy/ and a voiced consonant /z/

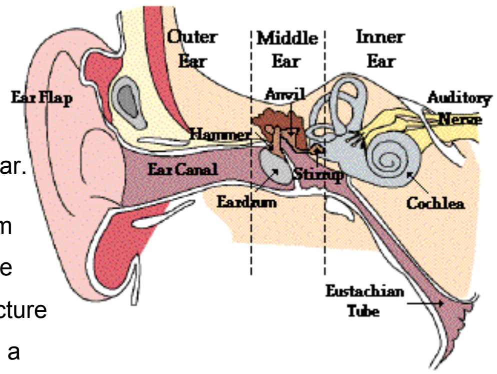


Speech Perception

Physiology of the Ear

interACT

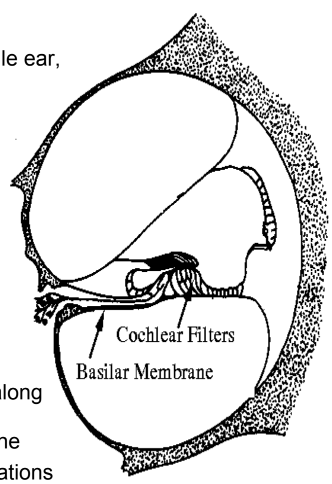
Three basic parts:



1. The **outer ear** serves to collect and channel sound to the middle ear.
2. The **middle ear** serves to transform the energy of a sound wave into the internal vibrations of the bone structure and transform these vibrations into a compressional wave in the inner ear.
3. The **inner ear** serves to transform the energy of a compressional wave within the inner ear fluid into nerve impulses which can be transmitted to the brain.

Biology of Perception

interACT



- Sound waves are guided from the outer ear to the middle ear, where they make the eardrum move
- A mechanical transducer (hammer, anvil, stirrup) adjacent to the eardrum's opposite side converts the sound waves to mechanical vibrations on the oval window, the entrance to the cochlea
- The cochlea is a spiral tube (3.5cm long that coils about 2.6 times) filled with fluid in which standing waves are excited
- The waves of the fluid make the cochlear filters swing along
- The cochlear filters are attached to the basilar membrane which responds to different frequencies at different locations
- The movement of the filters is transferred to the brain along the cochlear nerve

Physical versus Perceptual Attributes

interACT

Basic distinction between **perception of a sound** and its measurable **physical properties.** Below listed items have a strong correlation but the connection is complex because other physical properties may affect the perception.

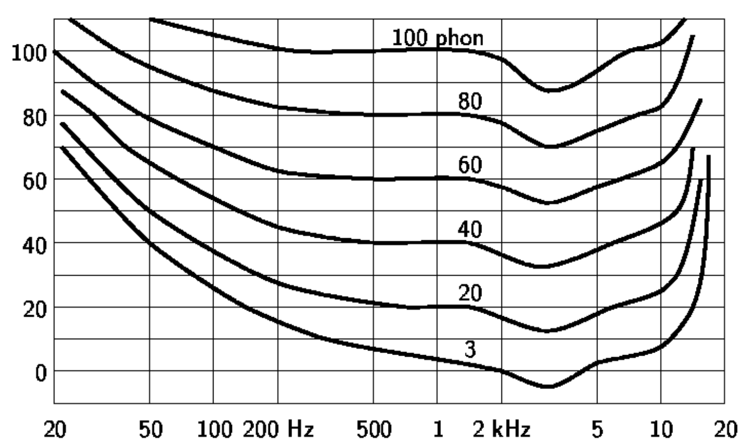
Physical Quantity	Perceptual Quality
Intensity	Loudness
Fundamental Frequency	Pitch
Spectral Shape	Timbre
Onset/offset time	Timing
Phase differences	Location

Non-uniform Equal Loudness

interACT

One divergence between perceptual and physical quality is the **non-uniform equal loudness**: the sensitivity of the ear varies with frequency (Loudness !~ Intensity)

Insensitive at low freq with low-moderate intensity, max at ~4kHz (resonance f at outer ear, another peak at ~13kHz 2nd resonance f)



equal loudness contours **(dB)**

What can you hear?

10Hz 100Hz 500Hz 1000Hz 2000Hz

4KHz 8KHz 10KHz 12KHz 14KHz

16KHz 18Khz 20KHz

Human frequency perception

Highest perception 20Khz

But it degrades with age.

The older you are the less high frequencies

Starts degrading as late teenager!

But is it important?

- Speech
  - F0 (intonation contour) 80-300Hz
  - F1/F2 250-3000Hz
  - Fricatives, higher maybe 4KHz-8KHz

Sampling Frequency

How many samples a second

- To capture an 8KHz signal?
- To capture a 16KHz signal?

At least 2 times the signal

- Nyquist frequency (half the sample rate)

So why is CD sampling rate 44.1KHz?

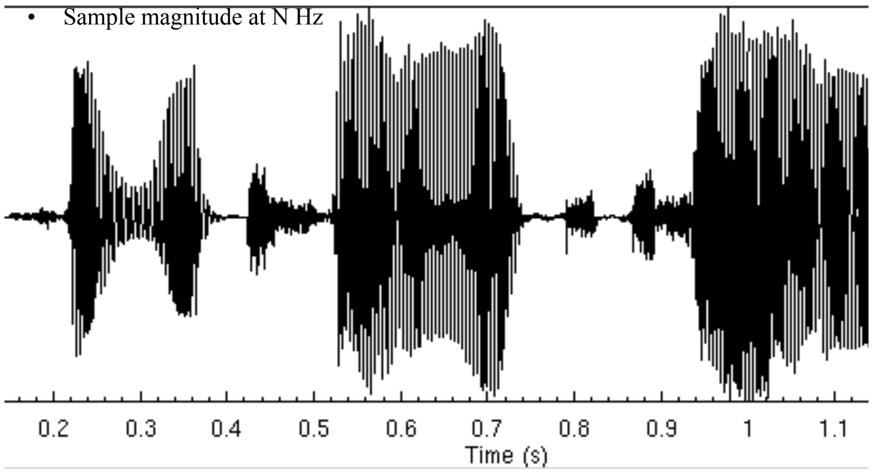
Human Speech

Human speech and sampling frequencies

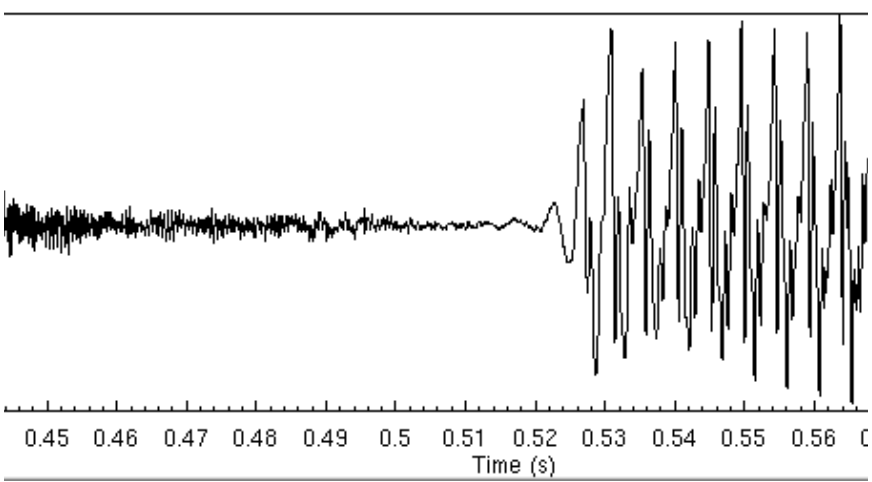
32000Hz	22500Hz	16000Hz
11250Hz	8000Hz	6000Hz
4000Hz	2000Hz	1000Hz

Waveform Representation

- Sample magnitude at N Hz



Waveform Representation



Analog to Digital

Speech (sound) is analog

- Computers are digital
  - We need to convert

Sample from A-D converter

- N times a second

How many times a second?

Waveform Encoding

PCM (Pulse code modulation)

- Simple +/-32768

But human hearing is logarithmic

- Changes are smaller amplitudes more important than changes at higher amplitudes
- mulaw (alaw) encodings

Human speech conventions

- Wide band speech 16KHz
- Narrow band speech 8KHz (telephone speech)

Speech Compression

interACT

- Bandwidth is money (or time)
- Telephone Speech
  - 64KBs (8KHz/8bit ulaw/alaw)
- Wide band:
  - 256KBs (16KHz/16bit)
- CDs
  - 1.4MBs (44.1KHz 16bit stereo)
- Mp3s (music)
  - 128KBs (expands to 44.1KHz stereo)
- Cell phone
  - 9.8KBs (or even 4.8KBs)

Time vs Frequency Domain

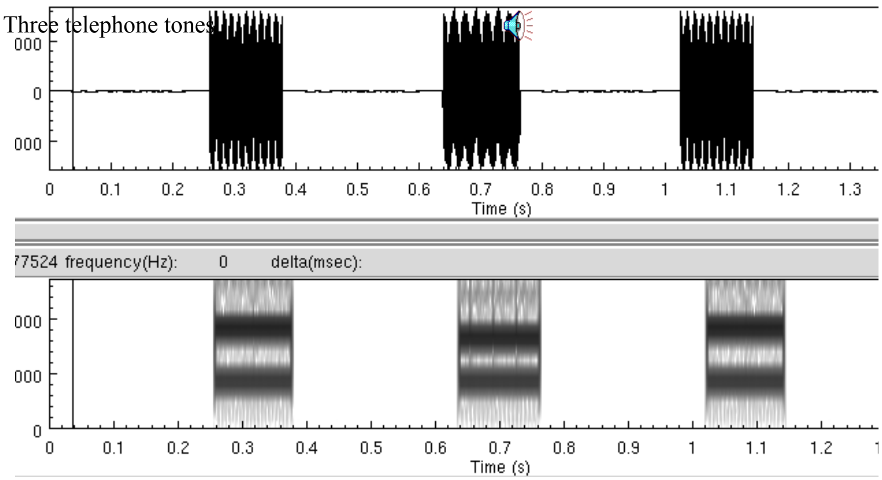
interACT

- All signals can be constructed
  - From sum of sine waves
- We can convert any signal into a set of sine waves
- Fourier Transform
  - Conversion of time signal to frequency spectrum
- Fast Fourier Transform
  - An efficient computer algorithm to do it

Spectrogram vs Time domain

interACT

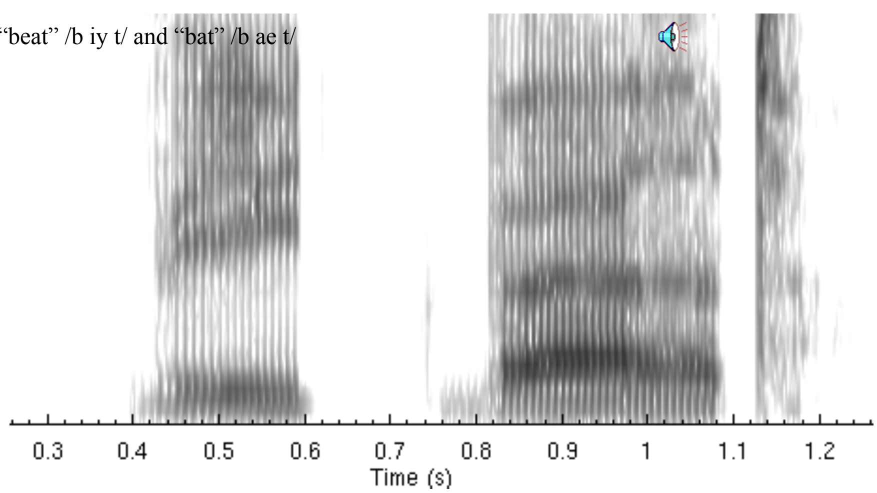
- Three telephone tones



/iy/ vs /ae/

interACT

- "beat" /b iy t/ and "bat" /b ae t/



Suggested Reading

- In: Dan Jurafsky & James Martin, „Speech and Language Processing", 2nd ed., Prentice Hall.
  - Chap. 9, „Automatic Speech Recognition", in particular:
    - Sec. 9.1, „Speech Recognition Architecture"
    - Sec 9.8, „Evaluation: Word Error Rate"
    - Sec 9.9, „Summary"
  - Chap. 7, „Phonetics"