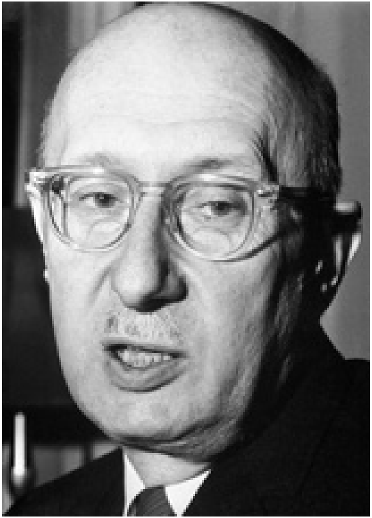
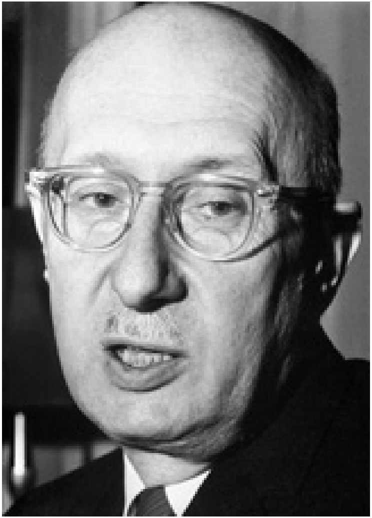
# Perception

# Georg

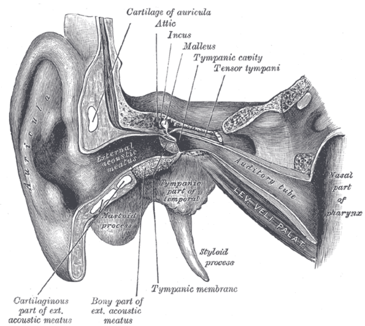


# Georg



Georg von Békésy
1899–1972

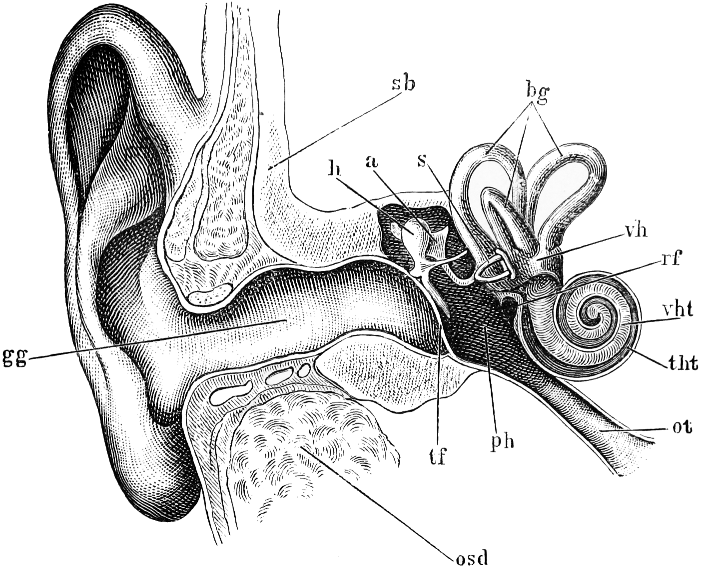
# Gray's anatomy




http://en.wikipedia.org/wiki/Auditory_system

- Find the bit where the noise goes in.
- Take it to bits.
- See how it works.

# Tidens Naturlære




http://en.wikipedia.org/wiki/Hearing_(sense)

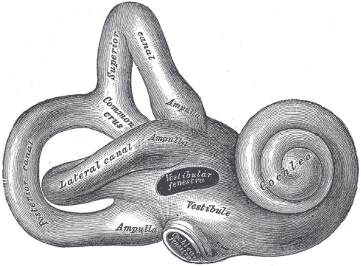
# Ear basics

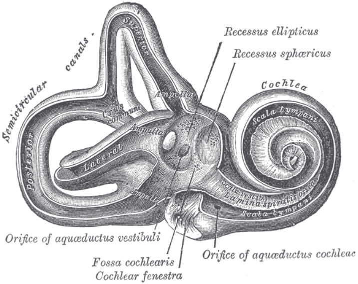
They're old pictures, but you can see:

- The ear is connected to the pharynx - it can't hear DC[1].
- It's mechanical up to a point, then something else happens.

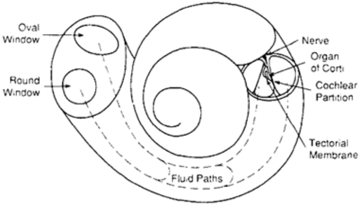
[1] Although the eustachian tube is mostly closed.

# Cochlea


Superior
canal
Common
crus
Ampulla
Ampulla
Lateral canal
Posterior canal
Vestibular
fenestra
Cochlea
Vestibule
Ampulla

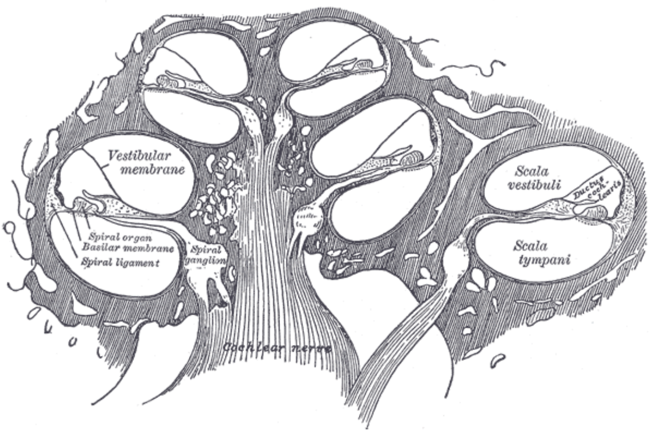

Recessus ellipticus
Recessus sphæricus
Cochlea
Superior
canals
Semicircular
Ampulla
Crus commune
Ampulla
Scala tympani
Posterior
Lateral
Scala vestibuli
Lamina spiralis ossea
Scala tympani
Orifice of aquæductus vestibuli
Fossa cochlearis
Cochlear fenestra
Orifice of aquæductus cochleae

# Mechanics of the cochlea

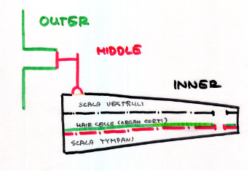



- Sound is introduced via the stapes and the oval window.
- The round window is there for pressure relief.
- The coil contains two separate, fluid filled chambers.

# Cochlear section



# Hynek's slide 1




## OUTER EAR

- FOCUSES THE SOUND
- PROTECTS MIDDLE EAR
- AMPLIFIES 3–5 kHz RANGE ( $\lambda/4$ RESONATOR)

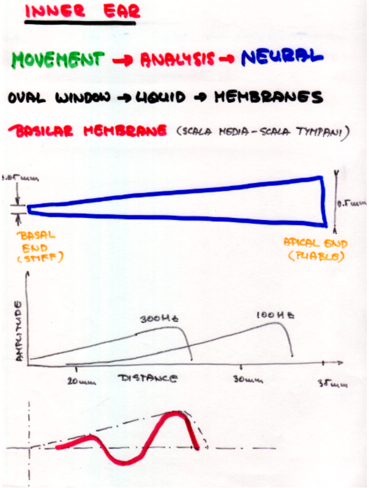
## MIDDLE EAR

- CONVERTS IMPEDANCE OF AIR TO IMPEDANCE OF COCHLEAR LIQUID ( $Z_{AIR} : Z_{LIQ} = 1:4000$ => 99.9% LOSS OF ENERGY)
- PROTECT INNER EAR (REACTIONS TO INTENSE SOUNDS – ABOUT 60-120 ms REACTION TIME => NO GOOD FOR IMPULSES !)
- LOW-PASS 15 dB/oct FROM 1 kHz

## INNER EAR

- ANALYSIS AND MECHANICAL → ELECTRIC

# Hynek's slide 2




- The basilar membrane resonates.
- It's stiffer at the base.

# Basilar membrane

The Basilar membrane is really central to human hearing.

- It supports the organ of Corti.
  - You can think of this as a kind of microphone.
- It disperses frequency
  - Stiffer at the base than the apex.
  - High frequencies concentrate at the base, low frequencies at the apex (this is what Békésy showed).
  - All the above is non-linear. The non-linearity is important when designing speech signal processing systems.

# Summary

- The ear is a kind of amplifier.
- Sound is perceived as vibrations of the basilar membrane.
- The frequencies overlap.

More when we talk about PLP...

# Sampling speech

# Basics: Sample rate

Some common sample rates:

| | |
|---|---|
| 44.1 kHz | CD players. |
| 48 kHz | Pro. audio. |
| 96 kHz | Really pro. audio. |
| 20 kHz | Speech researchers (also 16 kHz is common). |
| 11.025 kHz | Good compromise for ASR. |
| 8 kHz | Telephony. |

All derive from two clocks:

- An 8 kHz multiple for pro. audio.
- A 44.1 kHz for consumer audio.

# Basics: Sample resolution

Some common sample resolutions:

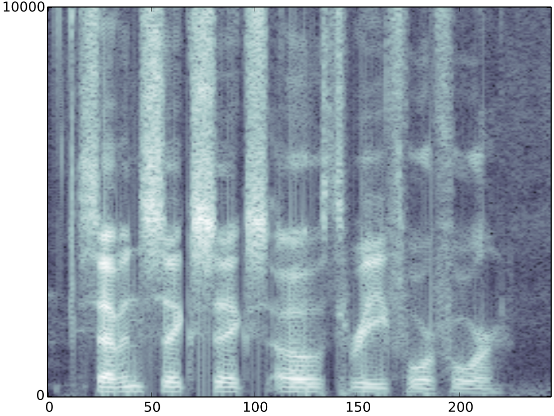
- **16 bit** Useful, ubiquitous.
- **24 bit** Pro audio
- **8 bit** Useless!
- **8 bit, companded** Telephony.
- **12 bit** The actual resolution of a crappy 16 bit ADC.

You get about 6 dB per bit.

- Speech can be up to 50 dB.

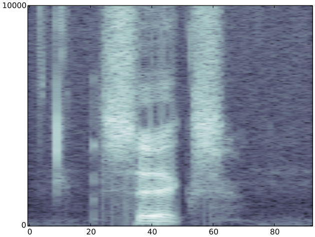
# 20 kHz spectrogram

An American male from TI-Digits.




“Seven six six oh three four four”

# 20k sampling



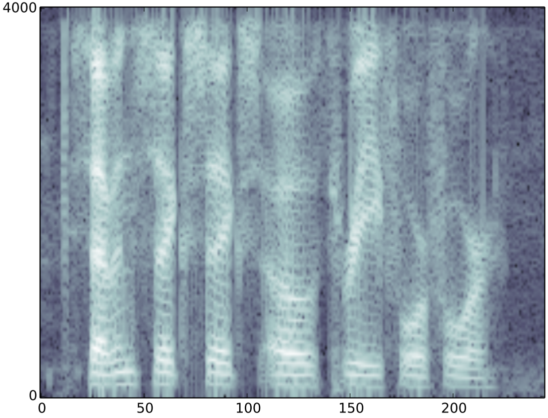

“Six” [sɪks]

This uses a 512 point DFT (257 bins visible), 10ms frame period

- Range is 0–10kHz.
- Already half the rate of CDs.
- The [s] takes the whole spectrum, the [ɪ] doesn’t.

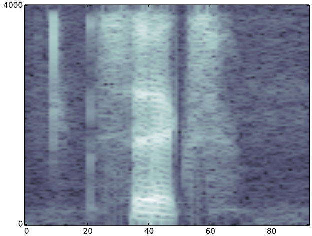
# 8 kHz spectrogram

The same American male from TI-Digits, same utterance




"Seven six six oh three four four"

# 8 kHz sampling



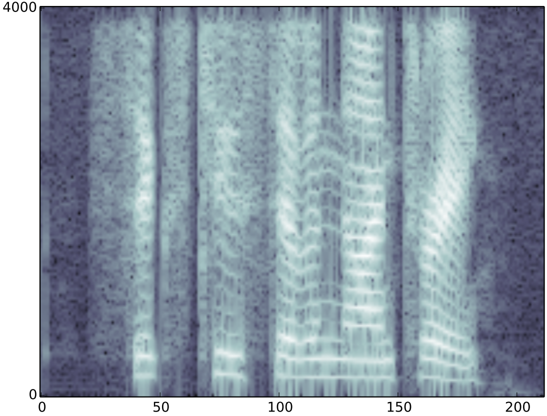

“Six” [sɪks]

Also uses a 256 point DFT (129 bins visible), 10ms frame period

- Range is 0–4kHz.
- Formants are more distributed across the spectrum.
- Horizontal striations!
    - Why?

# 8 kHz spectrogram, female

An American female from TI-Digits

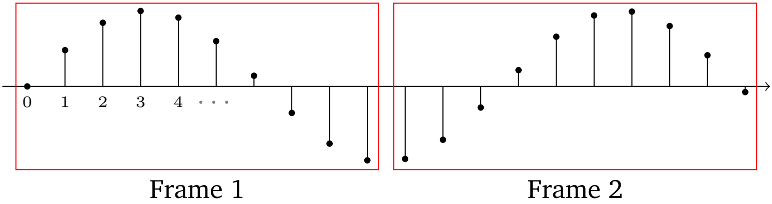



"Six two seven nine three"

# Framing

# Basic framing

- We have a one dimensional signal
- Framing gives a multi-dimensional signal
  ...possibly at a different rate



# Parameters

There are basically two parameters to control:

1. Frame rate
   How often does a new frame start? Two obvious choices:
   1.1 Every time a frame ends (no duplication)
   1.2 Every time a sample appears (frames overlap a lot)

   Perception suggests 10–15 ms
2. Frame size
   Boils down to bandwidth
   Normally means "big enough to capture the features you want"

# Narrow- vs. wide-band analysis

Thus far, we have used narrow-band spectra

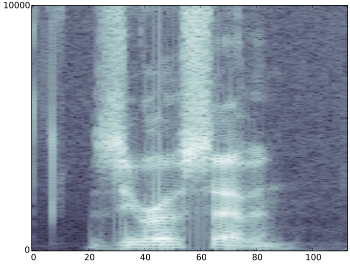
- The window is at least two pitch periods.
- Wide window mean narrow features in the frequency domain.

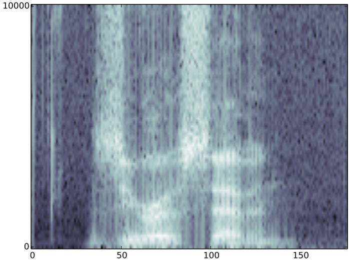
There is a different type: wide-band spectra

- The window is less than one pitch period.
- Narrow window leads to wide features in the spectrogram.

# Bad framing

If the frame period is too long, and the frame too short, the framing beats with the voicing.



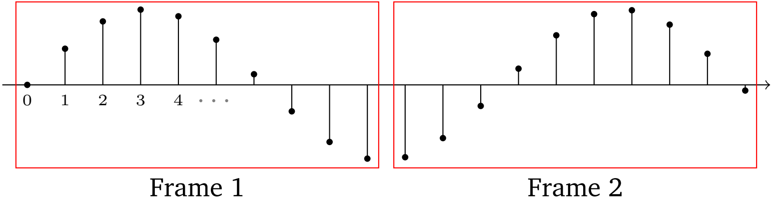




“Zero seven”, at 20 kHz.
Left: 512 point DFT, overlapped
Right: 128 point DFT and 128 sample period

# Overlap Add

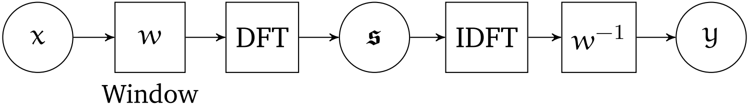
# Re-synthesis from DFT

Analysis with DFT is frame based:



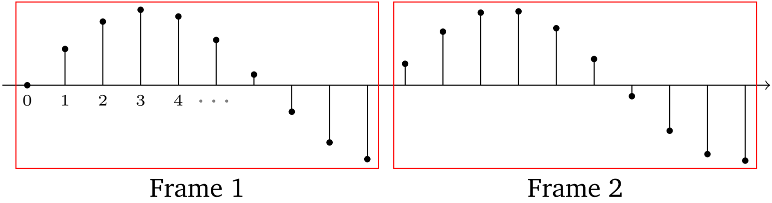

How do you reconstruct a frame based signal?

# Naive approach




- Synthesis is the inverse of analysis!
- Invert everything that was done in the analysis.

# The naive approach is dangerous




You tend to get discontinuities at the ends,

- Especially if you process the DFT data (which is most of the time).
- The $w^{-1}$ exagerrates it.

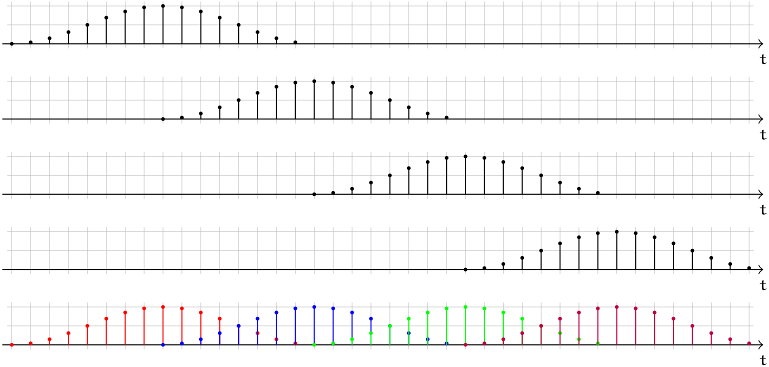
# Overlap add

Over-Lap Add (OLA) is basically a heuristic.
However, it works well in practice.

- Instead of inverting the window, design such that the window is cancelled out.
- Restricts the type of window that can be used. Typically Hann.
- Allows use of a synthesis window too.

Basically, though, you do exactly what it says on the tin.

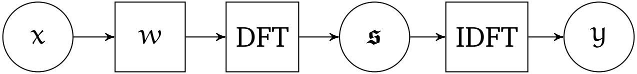
# Overlapped block sampling




Hann windows shifted by N/2

If using an even frame size, make sure both ends are not zero

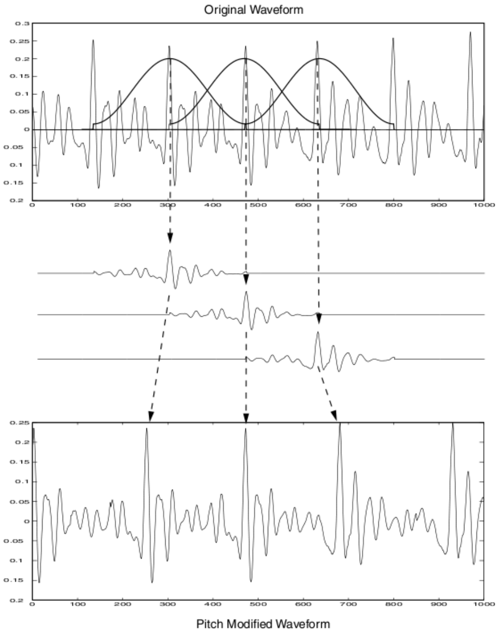
# OLA in practice



# PS-OLA

If you start messing with the pitch, the OLA has to be pitch-synchronous.

- Align the windows with the pitch periods.
- The windows are asymmetric


Original Waveform
Pitch Modified Waveform

# Summary

- There is some maths to go with it, but it's just not worth it.
- You can't use, e.g., Hamming windows.
  - Not in the usual case anyway
  - In general, check the window that you're using for COLA: Constant OverLap Add