ECE 417 Lecture 8: Speech Production

Mark Hasegawa-Johnson, 9/2017

Speech
(Slide: Scharenborg, 2017)

- Specific to humans

- Allows us to convey information very fast

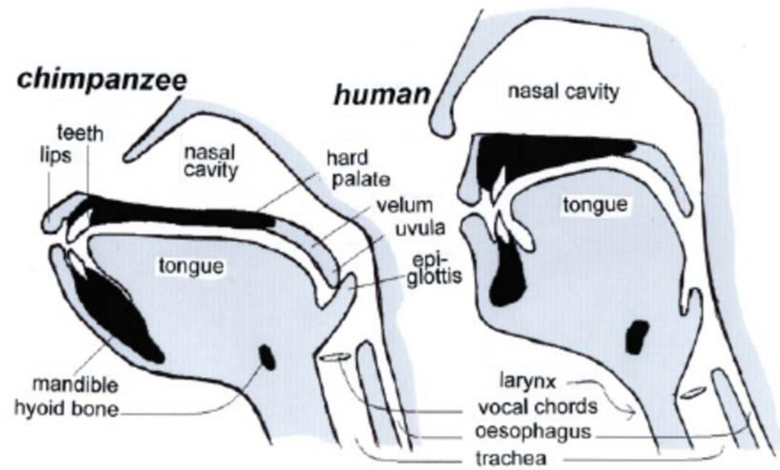
- Central role in many other language-related processes

- One of the most complex skills humans perform:
  - https://www.youtube.com/watch?v=DcNMCB-Gsn8
  - https://www.youtube.com/watch?v=KtN-FCOeWjI

Evolution of the vocal tract

(Slide: Scharenborg, 2017)

- Lowering of the tongue into the pharynx → lowering of the larynx
- Lengthening of the neck
- At the cost of an increase in the risk of choking on food

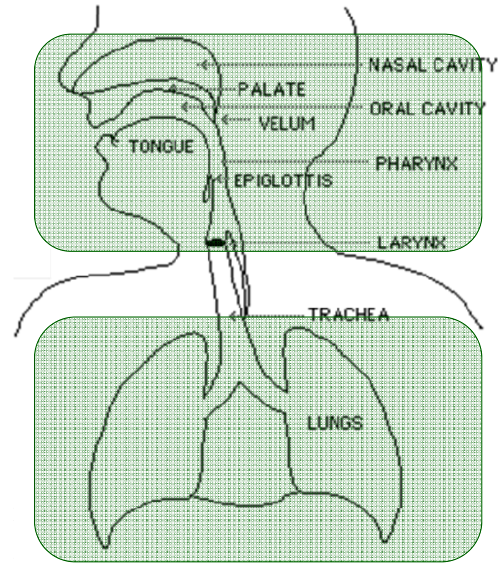


- Neanderthals were not capable of human speech
- Modern human vocal tract: since 50,000 years

The anatomy and physiology of speech
(Slide: Scharenborg, 2017)

Vocal tract

- Area between vocal cords and lips
- Pharynx + nasal cavity
  + oral cavity



and lungs

3 steps to produce sounds
(Slide: Scharenborg, 2017)

step 3: *articulation* =

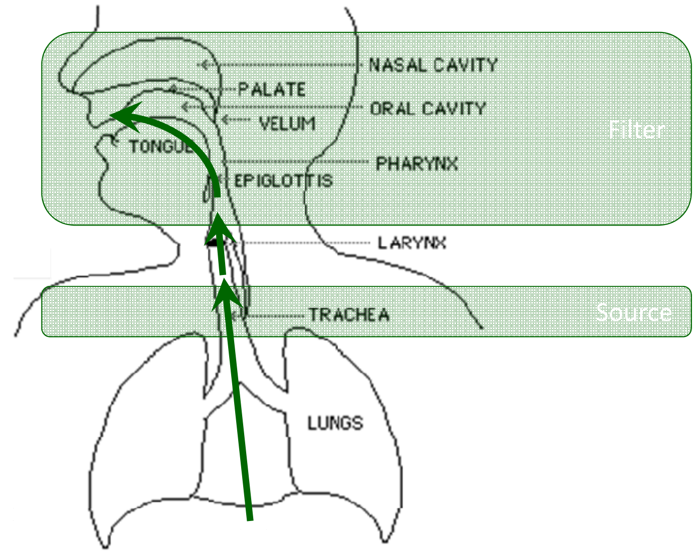
distortion of air

→ time-varying formant-frequency pattern

= speech

step 2: *phonation*

step 1: *initiation*



The Source-Filter Model of Speech Production
(Chiba & Kajiyama, 1940)

- Sources: there are only three, all of them have wideband spectrum
  - Voicing: vibration of the vocal folds, same type of aerodynamic mechanism as a flag flapping in the wind.
  - Frication or Aspiration: turbulence created when air passes through a narrow aperture
  - Burst: the “pop” that occurs when high air pressure is suddenly released
- Filter:
  - Vocal tract = the air cavity between glottis and lips
  - Just like a flute or a shower stall, it has resonances
  - The excitation has energy at all frequencies; excitation at the resonant frequencies is enhanced

3 steps to produce sounds

step 3: *articulation* =

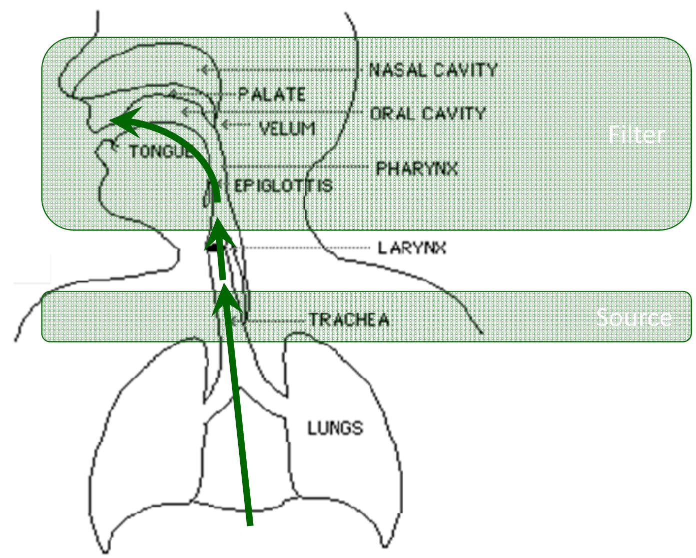
distortion of air

→ time-varying formant-frequency pattern

= speech

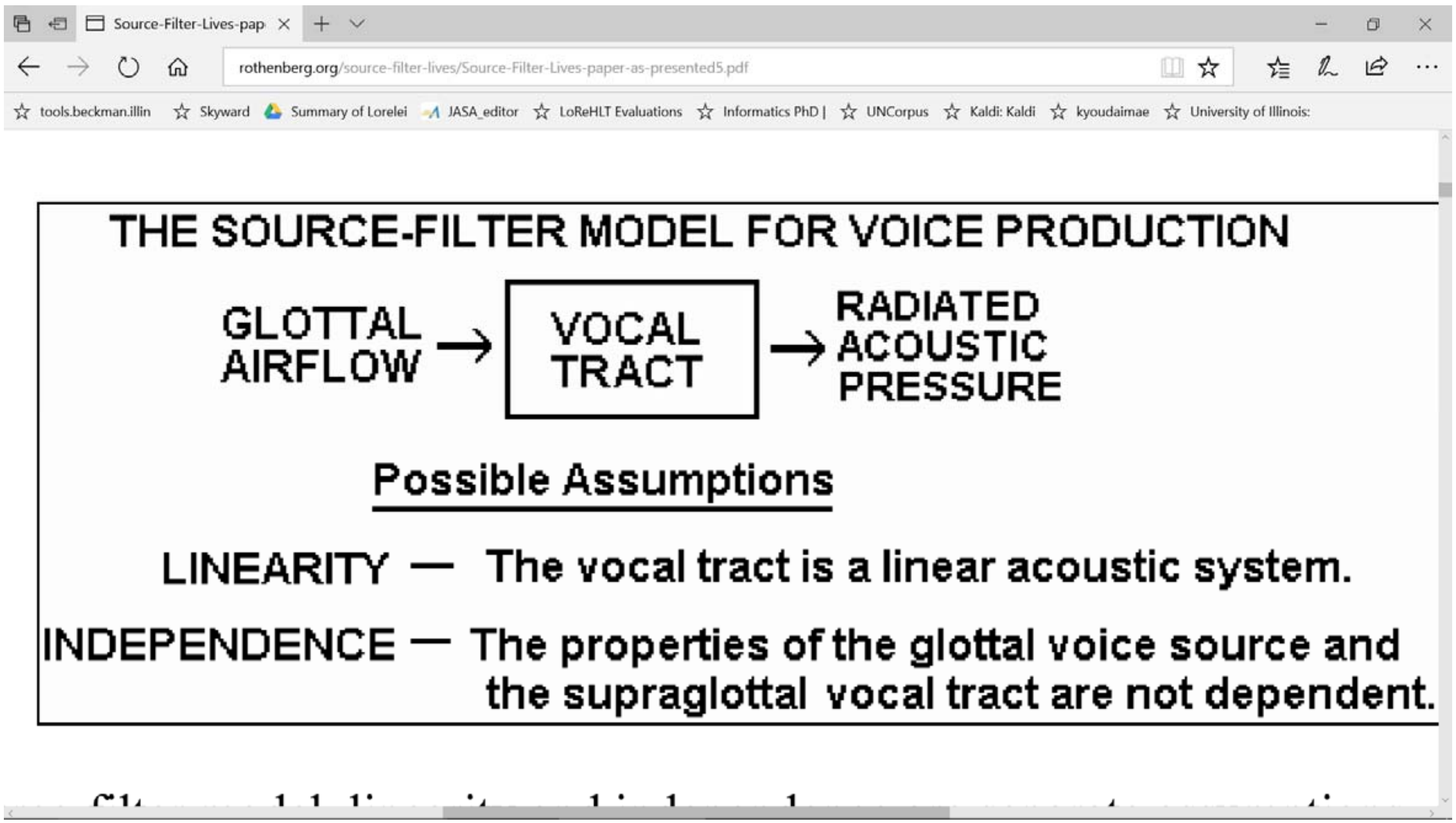
step 2: *phonation*

step 1: *initiation*



The Source-Filter Model of Speech Production

A picture from Martin Rothenberg's website



The Source-Filter Model

- The speech signal, $s(t)$, is created by convolving ($*$) an excitation signal $e(t)$ through a vocal tract transfer function $h(t)$

$$s(t) = h(t) * e(t)$$

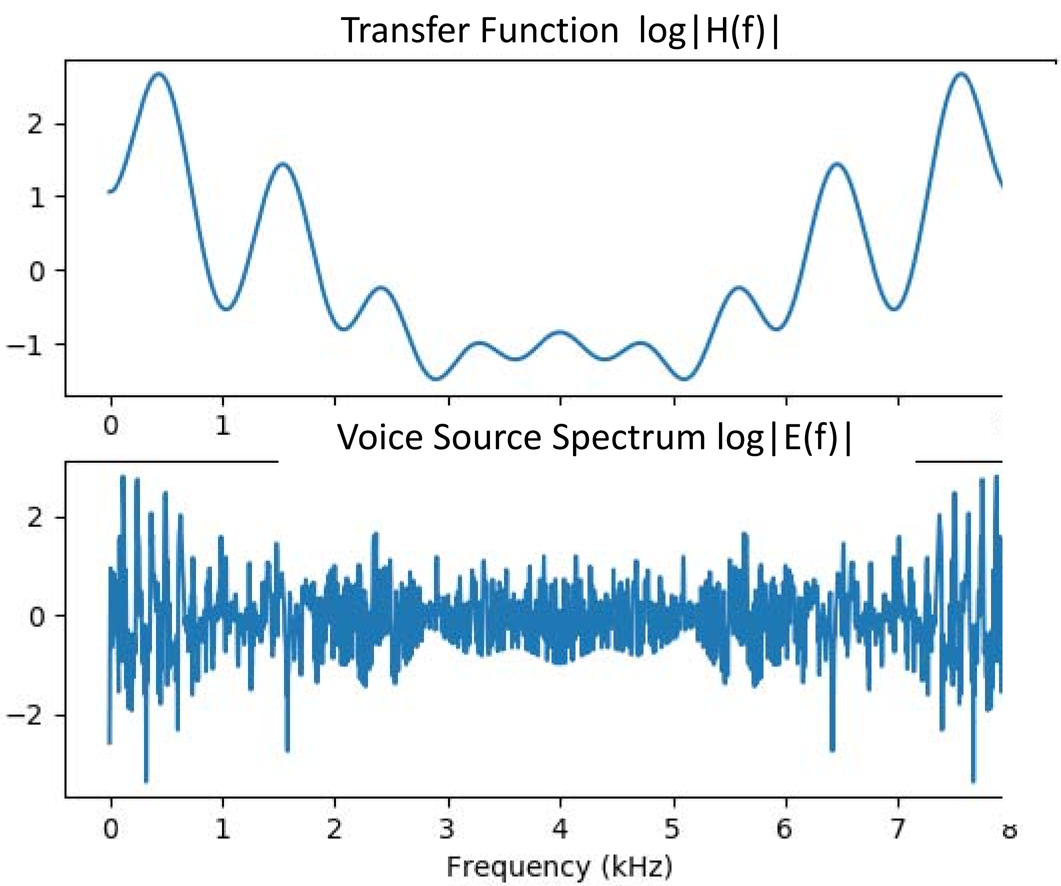
- The Fourier transform of speech is therefore the product of excitation times transfer function:

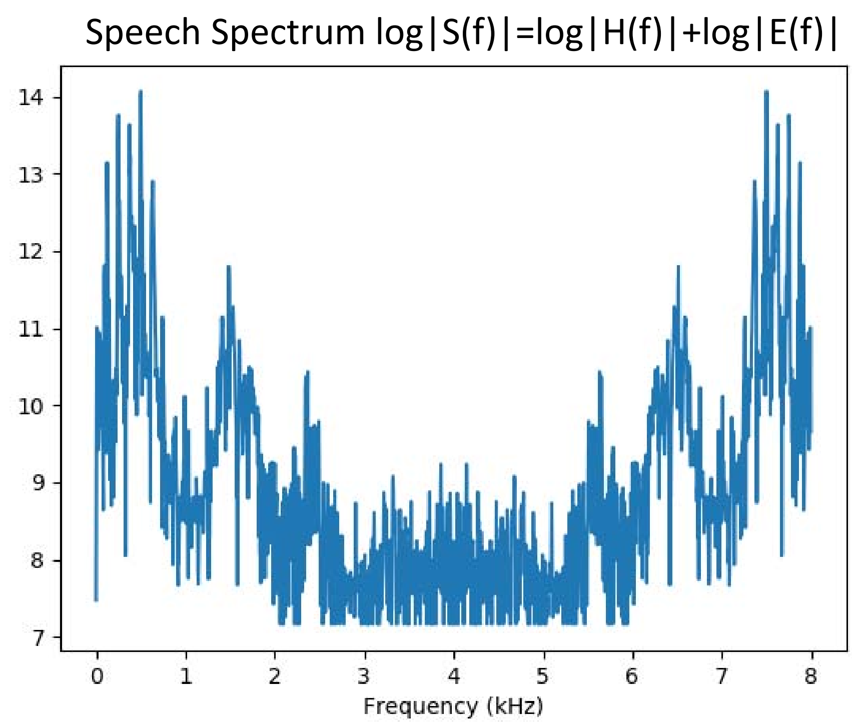
$$S(f) = H(f)E(f)$$

...engineers usually compute Fourier transform using $\Omega = 2\pi f$ rather than f. You can get one from the other if you remember that $\mathrm{d}\Omega = 2\pi\, df$.

- Excitation includes all of the information about voicing, frication, or burst. Transfer function includes all of the information about the vocal tract resonances, which are called "formants."

The Source-Filter Model





Source-Filter Model: Voice Source

- The most important thing about voiced excitation is that it is periodic, with a period called the "pitch period," T_0
- It's reasonable to model voiced excitation as a simple sequence of impulses, one impulse every T_0 seconds:

$$e(t) = \sum_{m=-\infty}^{\infty} \delta(t - mT_0)$$

- The Fourier transform of an impulse train is an impulse train (to prove this: use Fourier series):

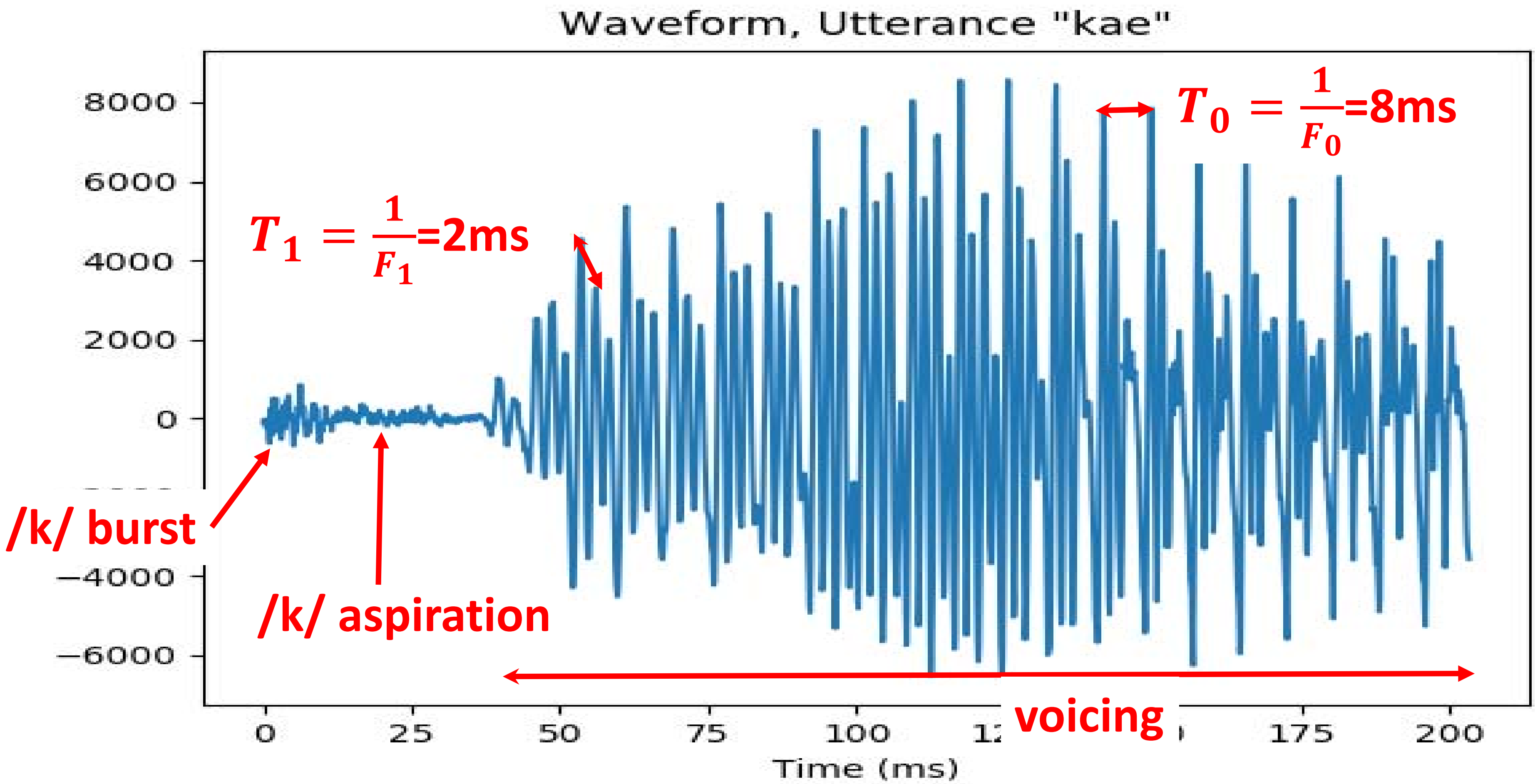
$$E(f) = \frac{1}{T_0} \sum_{k=-\infty}^{\infty} \delta(f - kF_0)$$

...where $F_0 = \frac{1}{T_0}$ is the pitch frequency. It's the number of times per second that the vocal folds slap together.

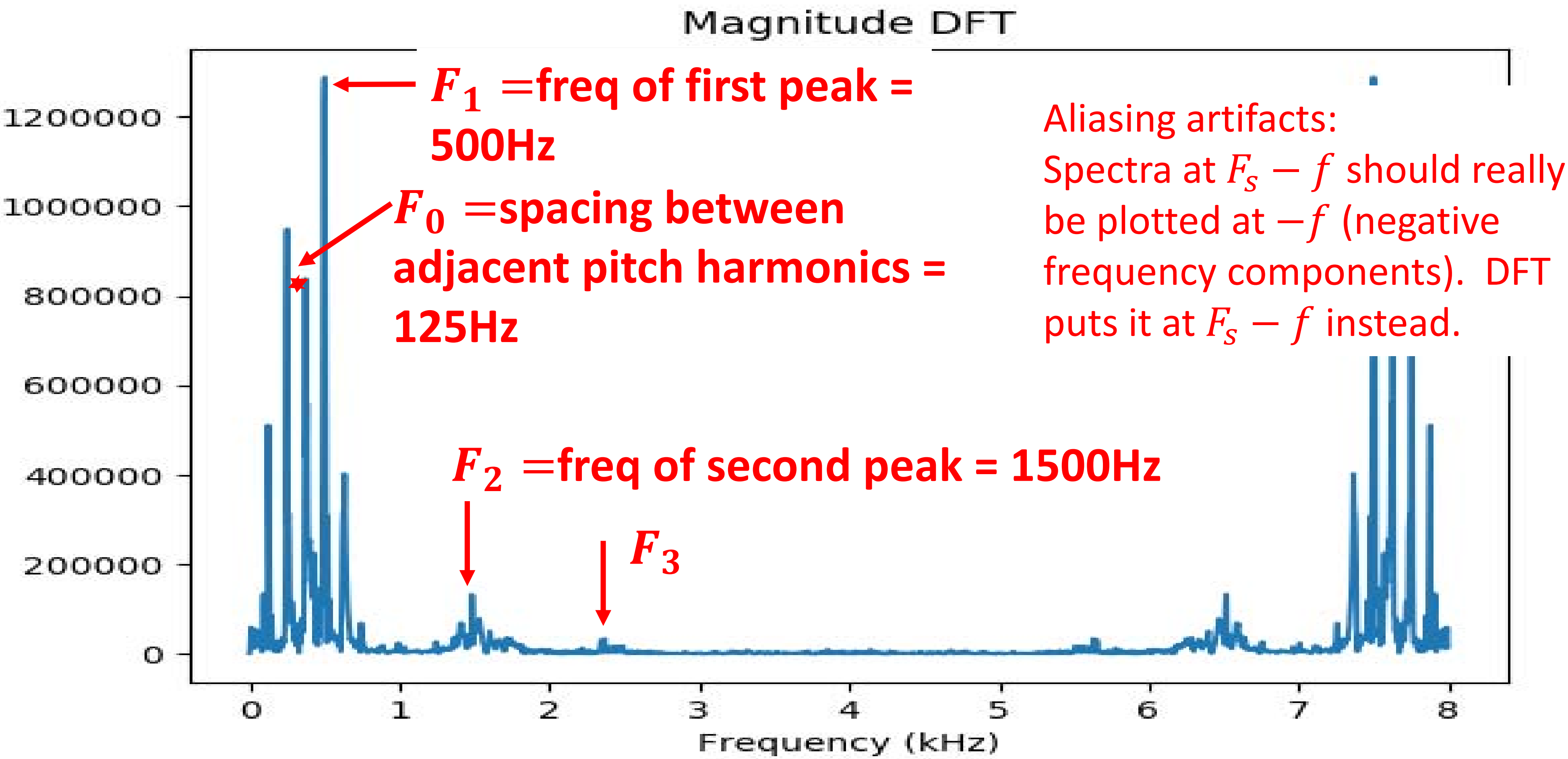
Source-Filter Model: Filter

- The vocal tract is just a tube. At most frequencies, it just passes the excitation signal with no modification at all ($H(f) = 1$).
- The important exception: the vocal tract has resonances, like a clarinet or a shower stall. These resonances are called "formant frequencies," numbered in order: $F_1 < F_2 < F_3 < \cdots$. Typically $0 < F_1 < 1000 < F_2 < 2000 < F_3 < 3000$Hz and so on, but there are some exceptions.
- At the resonant frequencies, the resonance enhances the energy of the excitation, so the transfer function $H(f)$ is large at those frequencies, and small at other frequencies.

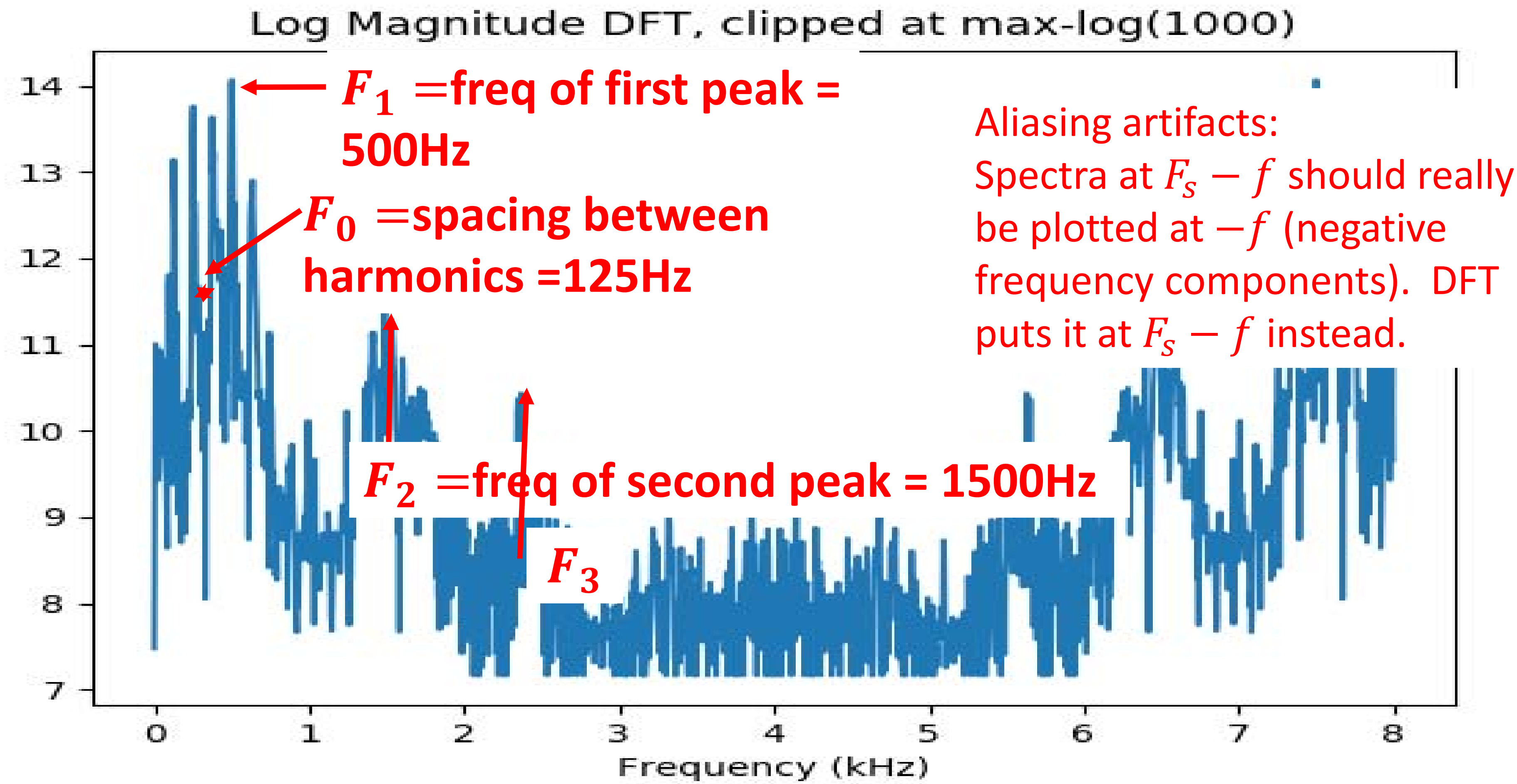
Speech signal: Time domain



Speech signal: Magnitude Fourier Transform



Speech signal: Log Magnitude Transform



Part 2: Linguistic units

Scharenborg, 2017

- Speech signal

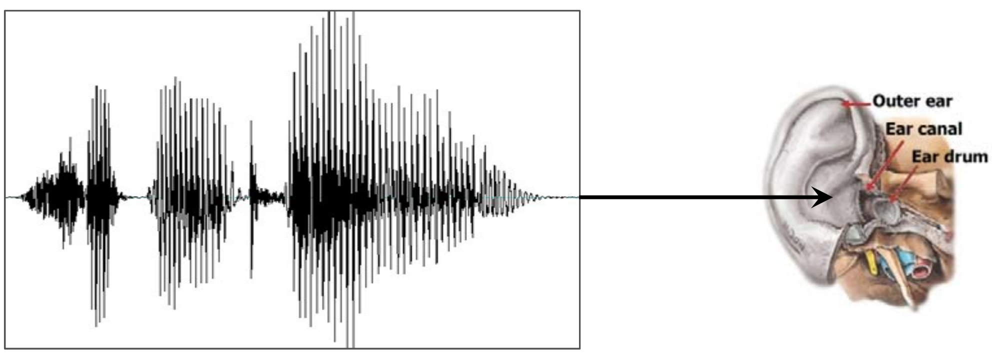
Linguistic units are:

- Phone(me)s
- Words

Linguistic units

Scharenborg, 2017

- Speech = sound
- Sound = differences in air pressure
- Air pressure waves perceived as different phone(me)s, phone(me) sequences, and (partial or multi) words
- Via eardrum, cochlea, and auditory nerve to brain



speech signal

Some terminology

Scharenborg, 2017

- Phoneme: the smallest contrastive linguistic unit that distinguishes meaning, e.g., *tip* vs. *dip*
- Allophone: a variation of a phoneme, eg., *p^hot* vs. *spot*
- Phone: a distinct speech sound
- Word: the smallest distinct unit that can be uttered in isolation which has meaning

Speech sounds

Scharenborg, 2017

- Vowels: unblocked air stream
- Consonants: constricted or blocked air stream

Different sounds: Vowels

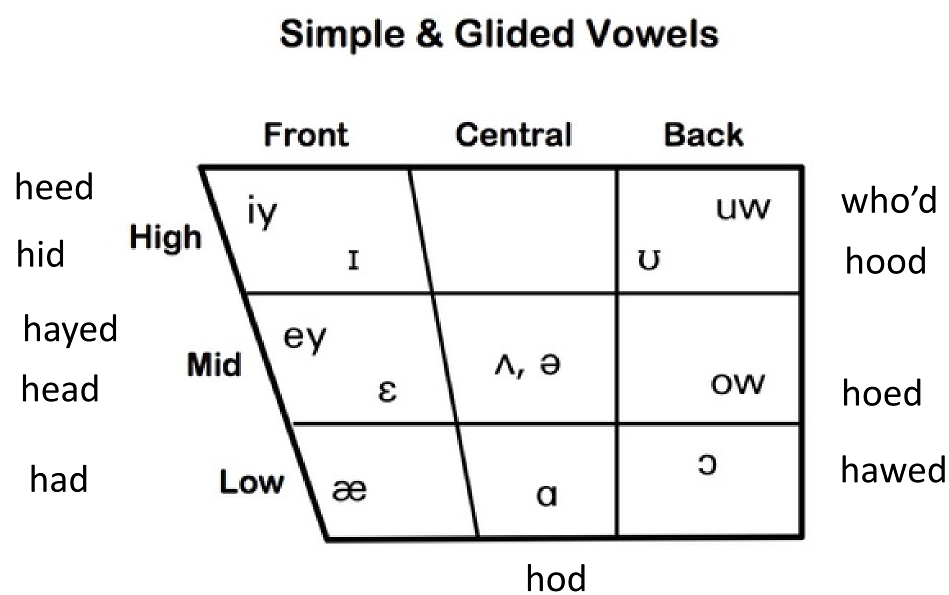
Scharenborg, 2017

- Tongue height:
  - Low: e.g., /a/
  - Mid: e.g., /e/
  - High: e.g., /i/

- Tongue advancement:
  - Front : e.g., /i/
  - Central : e.g., /ə/
  - Back : e.g., /u/

- Lip rounding:
  - Unrounded: e.g., /ɪ, ε, e, ə/
  - Rounded: e.g., /u, o, ɔ/

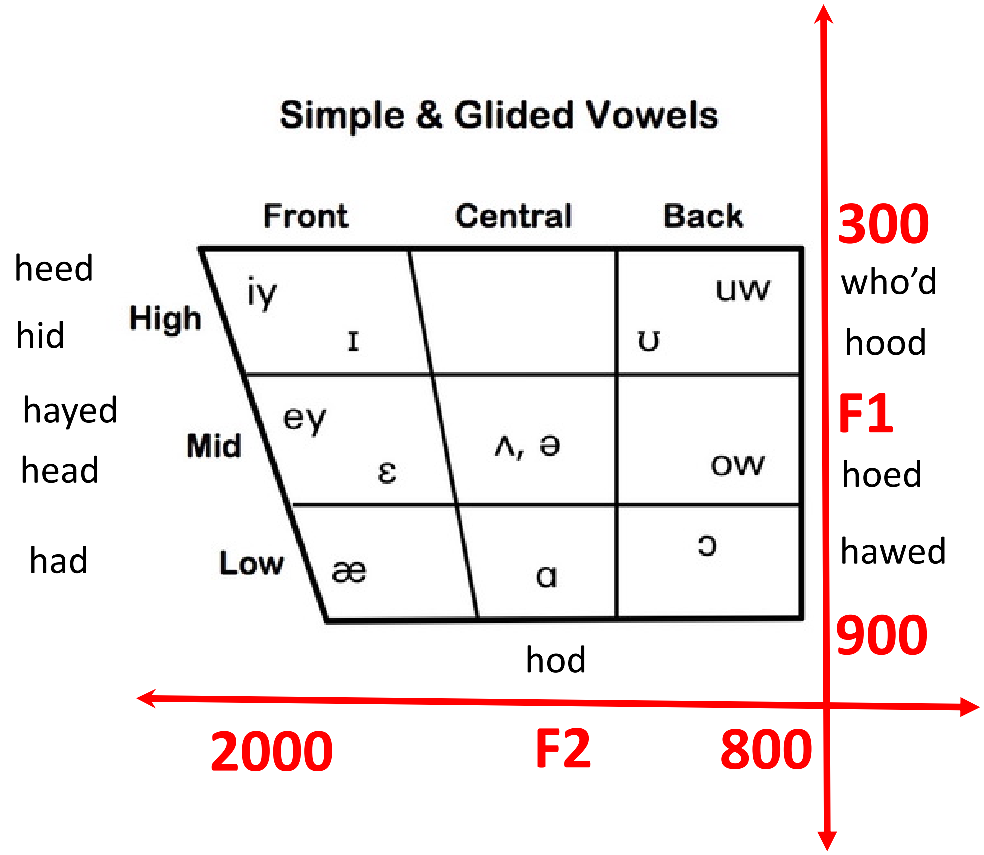
- Tense/lax:
  - Tense: e.g., /i, e, u, o, ɔ, ɑ/
  - Lax: e.g., /ɪ, ε, æ, ə/



Different sounds: Vowels

Scharenborg, 2017

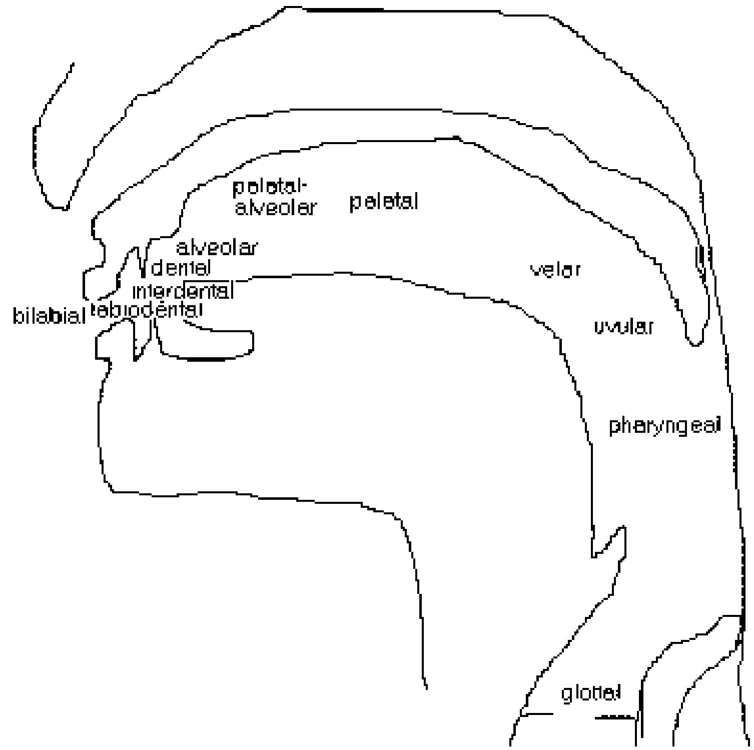
- Tongue height:
  - Low: e.g., /a/
  - Mid: e.g., /e/
  - High: e.g., /i/
- Tongue advancement:
  - Front : e.g., /i/
  - Central : e.g., /ə/
  - Back : e.g., /u/
- Lip rounding:
  - Unrounded: e.g., /ɪ, ɛ, e, ə/
  - Rounded: e.g., /u, o, ɔ/
- Tense/lax:
  - Tense: e.g., /i, e, u, o, ɔ, ɑ/
  - Lax: e.g., /ɪ, ɛ, æ, ə/



Different sounds: Consonants

Scharenborg, 2017

- Place of articulation
  - Where is the constriction/blocking of the air stream?



- Manner of articulation
  - Stops: /p, t, k, b, d, g/
  - Fricatives: /f, s, S, v, z, Z/
  - Affricates: /tS, dZ/
  - Approximants/Liquids: /l, r, w, j/
  - Nasals: /m, n, ng/
- Voicing

Speech sound production

Scharenborg, 2017

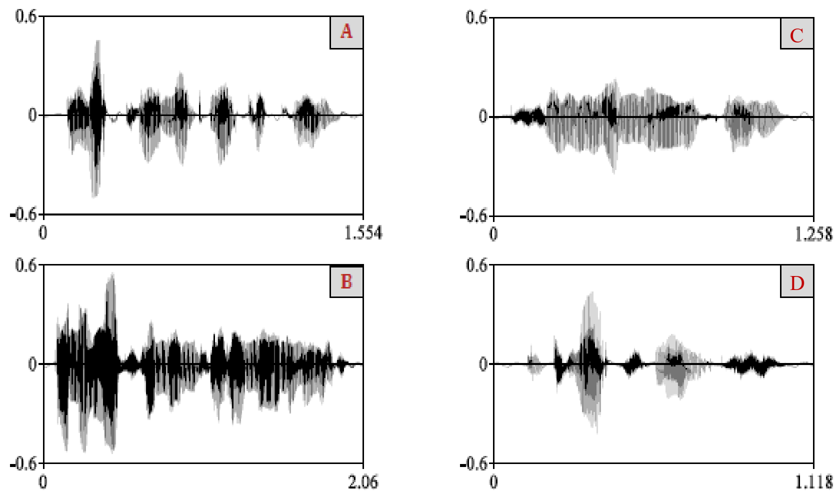
- https://www.youtube.com/watch?v=DcNMCB-Gsn8

Recorded in 1962, Ken Stevens
Source: YouTube

Quiz 1: How many words are there?

Scharenborg, 2017

Each picture shows a waveform of a short stretch of speech:



A: Electromagnetically (1)

B: Emma loves her mum's yellow marmelade (6)

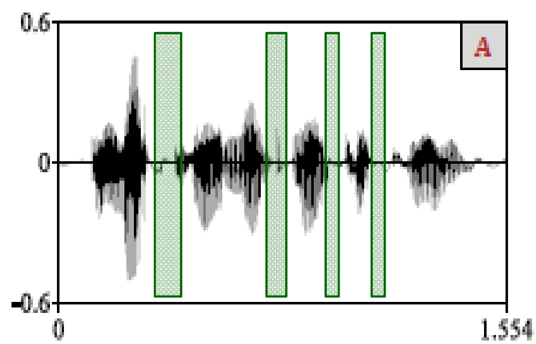
C: See you in the evening (5)

D: Attachment (1)

Electromagnetically

Scharenborg, 2017

Why is it so hard to determine the number of words?



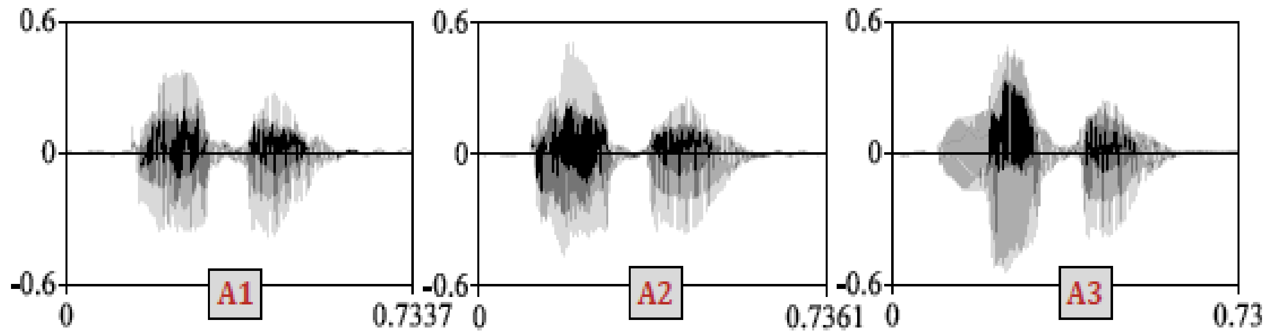
/i l ɛ kt romæ g nɛ t ɪ k ə l i/

silence ≠ word boundary

Quiz 2: Can you spot the odd one out?

Scharenborg, 2017

- Below are three waveforms each containing a single word:



Every time you produce a word it sounds differently

A3 (brother, brother, mother)

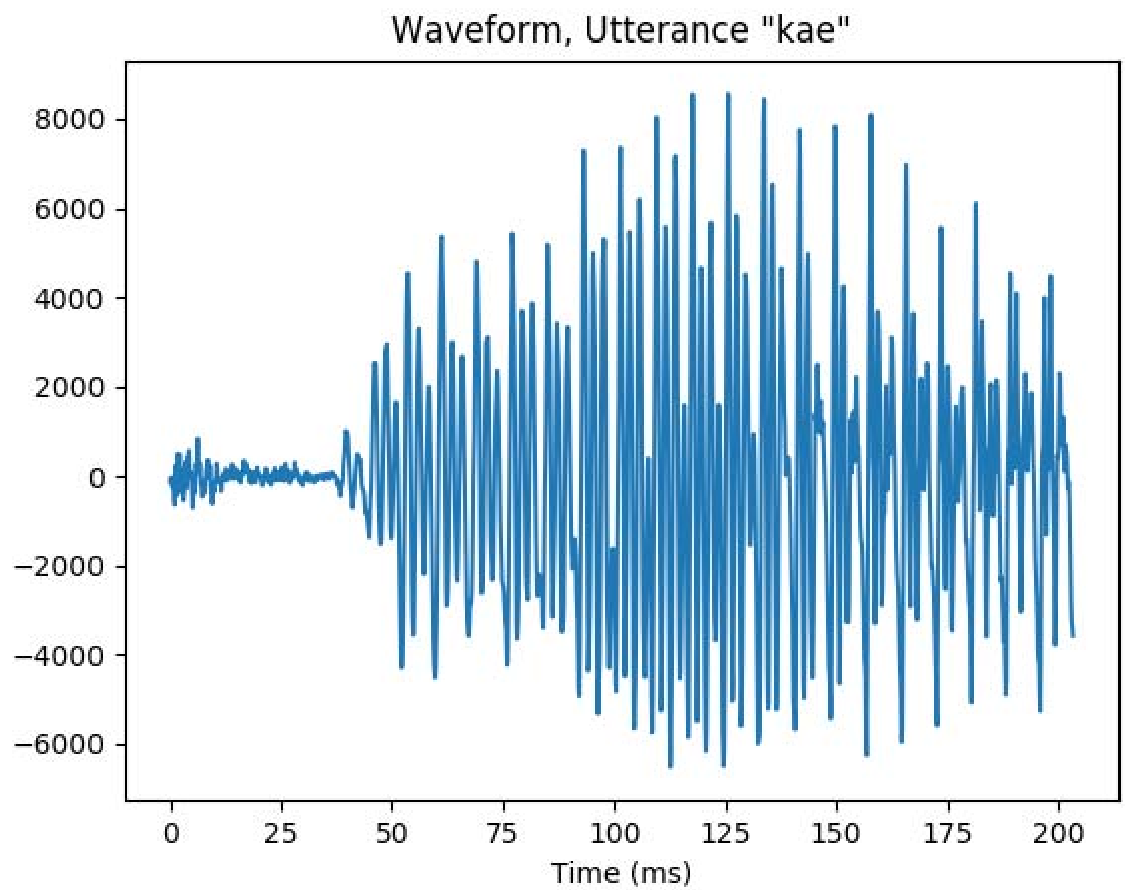
Enormous variability

Scharenborg, 2017

- Speaker differences, e.g., gender, vocal tract length, age
- Speaker idiosyncracies , e.g., lisp, creaky voice
- Accent: dialects, non-nativeness
- Coarticulation: production of a speech sound becomes more like that of a preceding/following speech sound
- Speaking style → reductions

Time domain signal: Hard to tell what he was saying

$$s(t) = h(t) * e(t)$$



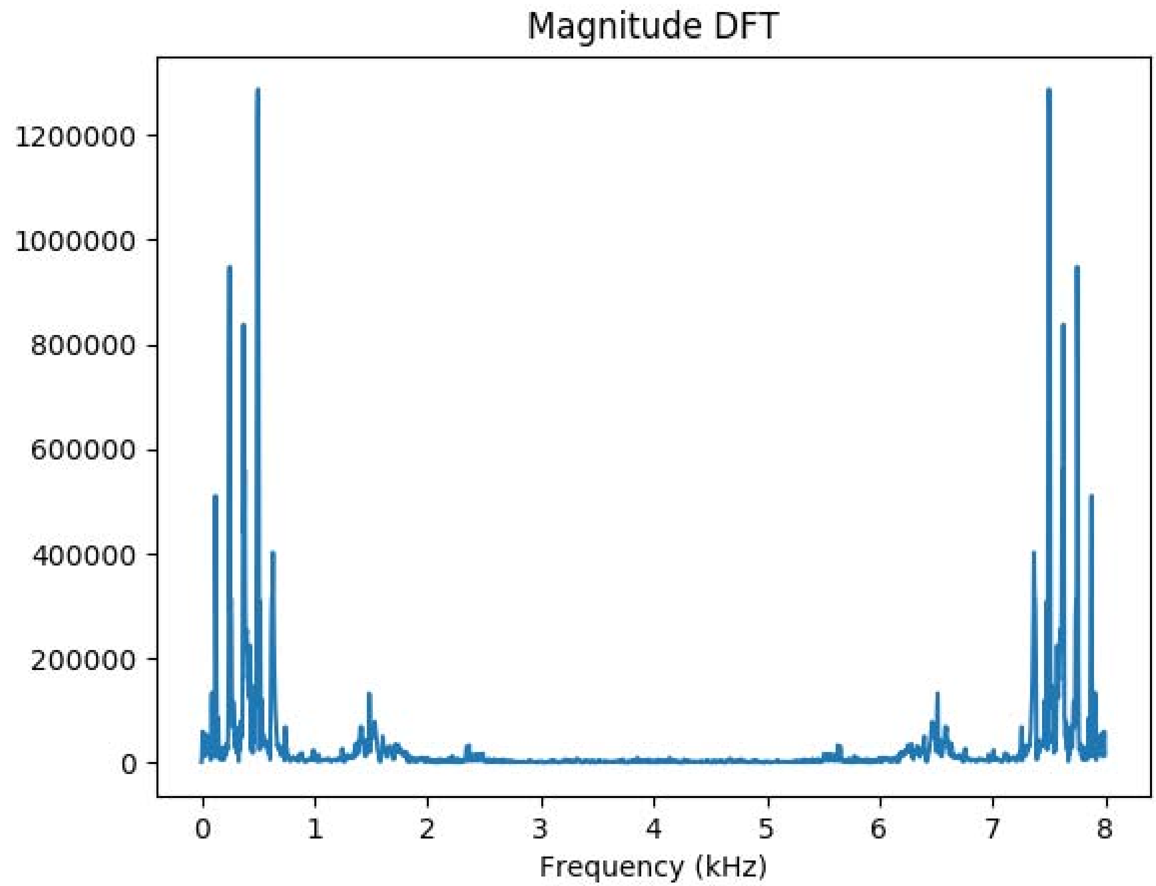
Magnitude spectrum: A little easier

$$S(f) = H(f)E(f)$$

Easier to measure formants→easier to guess what he's saying.

Still easy to measure F0→can still guess who he is.

(Formants≈phone-dependent, F0≈person-dependent, though there's a lot of cross-talk)



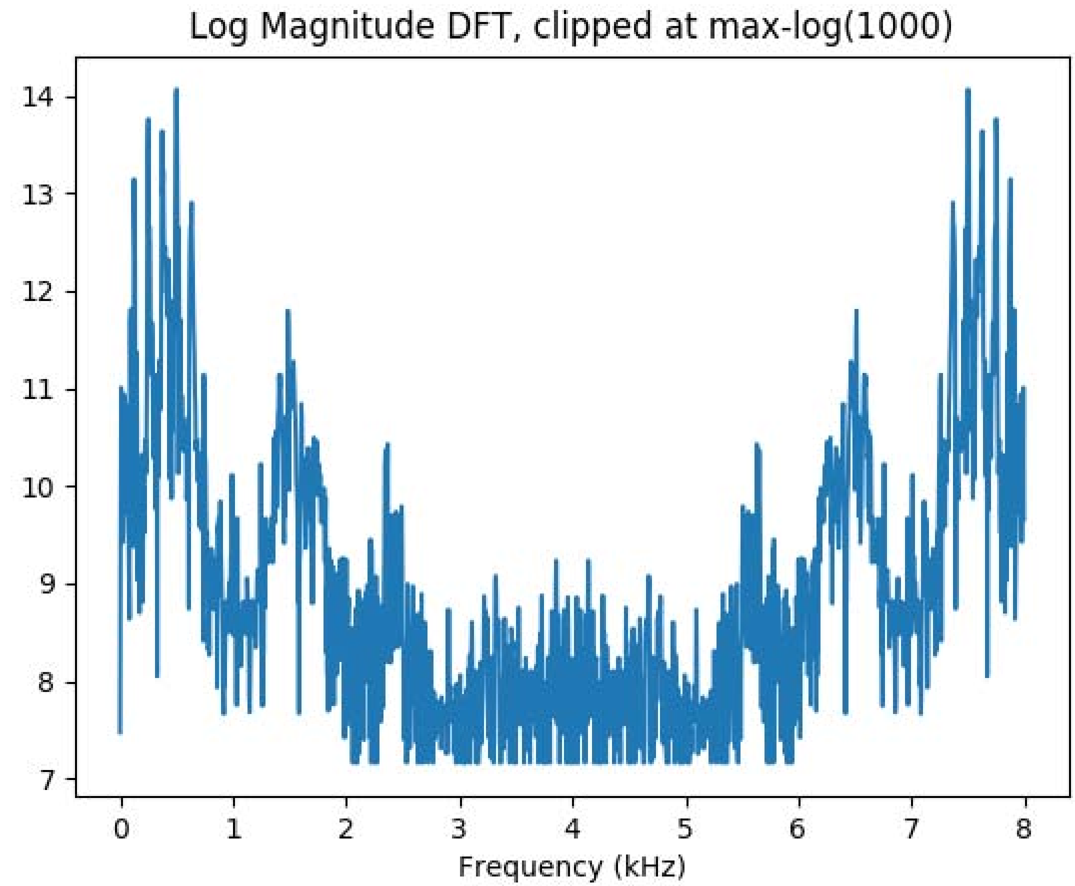
Log magnitude spectrum: A lot easier

$$\ln|S(f)| \\ = \ln|H(f)| + \ln|E(f)|$$

Easier to measure formants→easier to guess wha he's saying.

Still easy to measure F0→can still guess who he is.

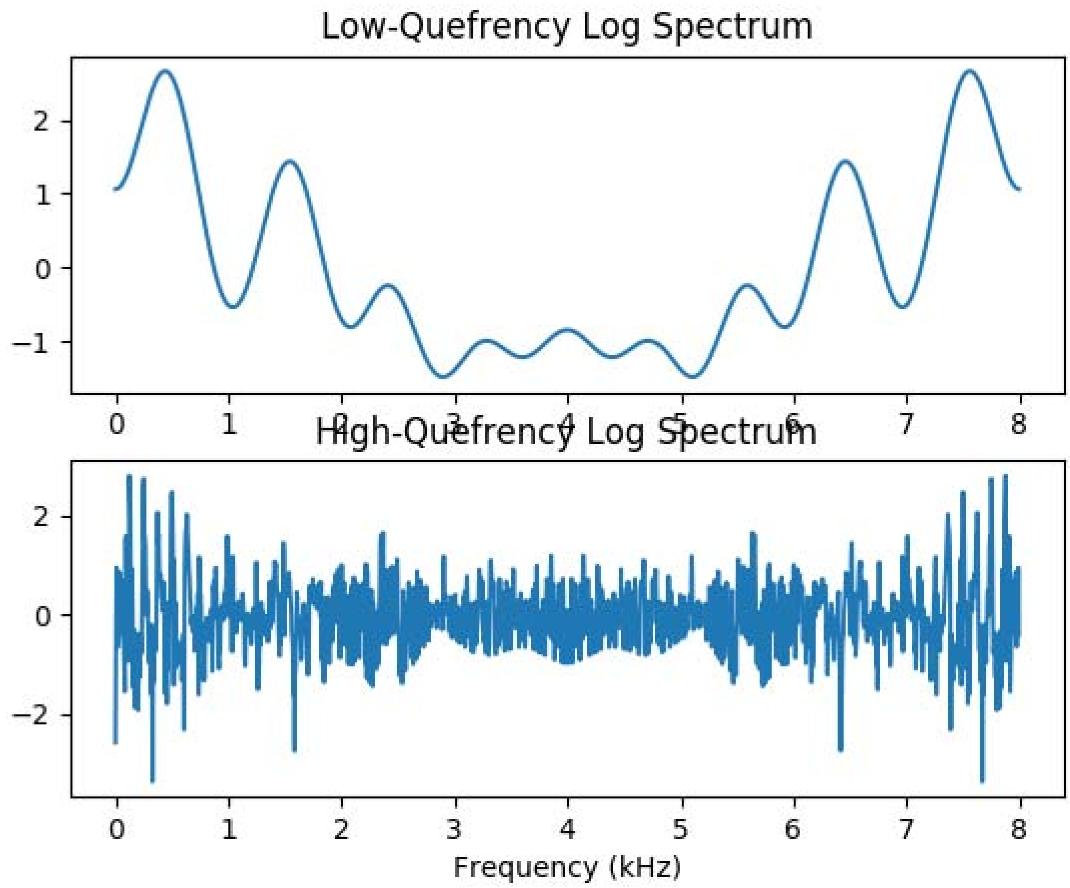
(Formants≈phone-dependent, F0≈person-dependent, though there's a lot of cross-talk)



Log spectrum = log filter + log excitation

$$\ln|S(f)| \\ = \ln|H(f)| + \ln|E(f)|$$

- But how can we separate the speech spectrum into the transfer function part, and the excitation part?
- Bogert, Healy & Tukey:
  - Excitation is high "quefrency" (varies rapidly as a function of frequency)
  - Transfer function is low "quefrency" (varies slowly as a function of frequency)



Cepstrum = inverse FFT of the log spectrum

(Bogert, Healy & Tukey, 1962)

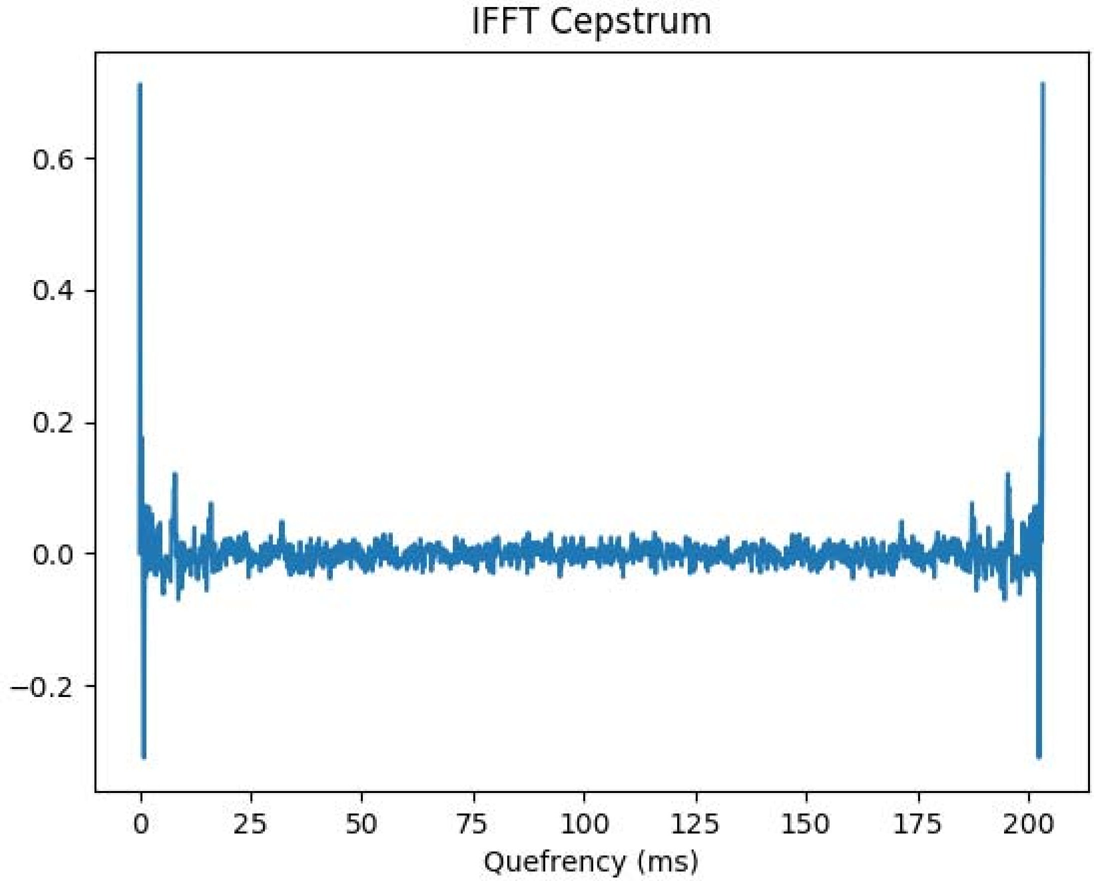
$$\hat{s}[q] = IFFT(\ln|S(f)|)$$

- q =quefrency. It has units of time.
- IFFT is linear, so since

$$\hat{s}[q]=\hat{h}[q]+\hat{e}[q]$$

…the transfer function and excitation are added together. All we need to do is separate two added signals.

- Transfer function and Excitation are separated into low-quefrency ($0 < q < 2$ms) and high-quefrency ($q > 2$ms) parts.



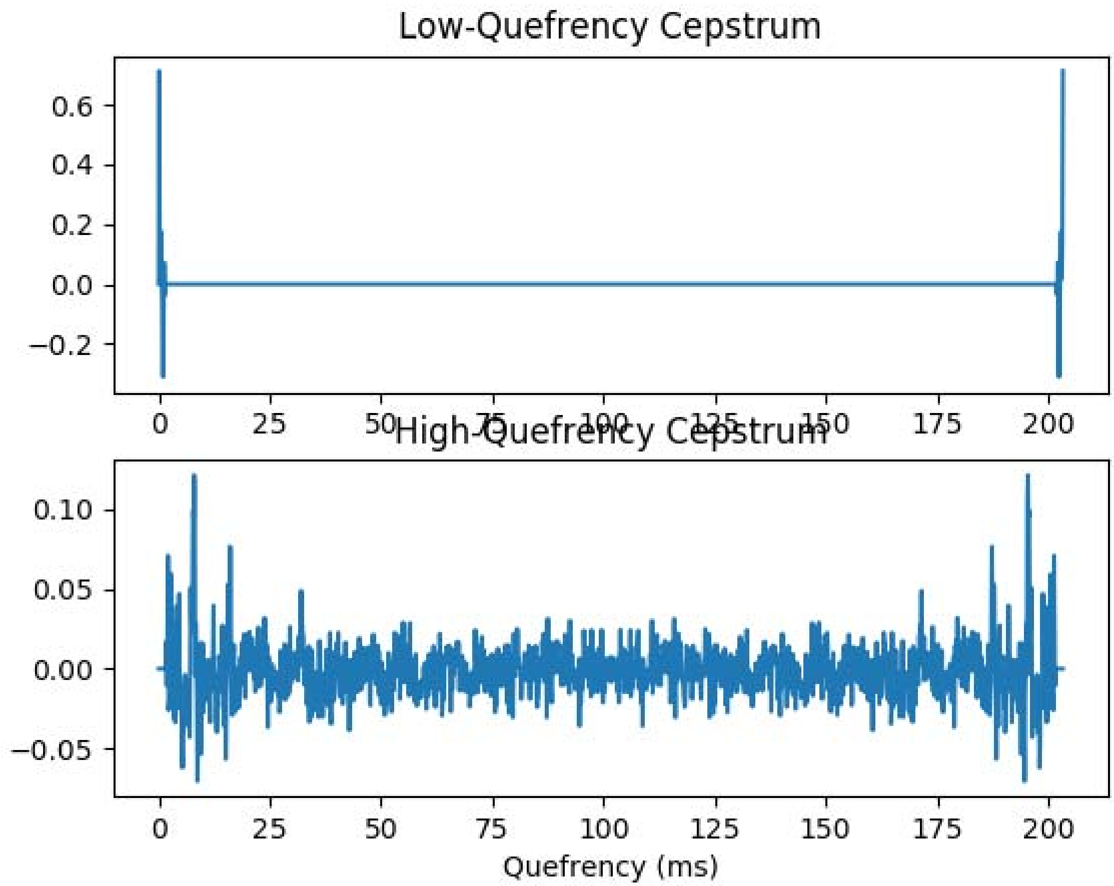
Liftering = filter(spectrum) = window(cepstrum)
(Bogert, Healy & Tukey, 1962)

Transfer function and Excitation are separated into low-quefrency ($0 < q < 2$ms) and high-quefrency ($q > 2$ms) parts. So we can recover them by just windowing:

$$\hat{h}[q] \approx w[q]\hat{s}[q]$$

$$\hat{e}[q] \approx (1 - w[q])\hat{s}[q]$$

$$w[q] = \begin{cases} 1 & 0 < q < 2ms \\ 0 & q > 2ms \end{cases}$$



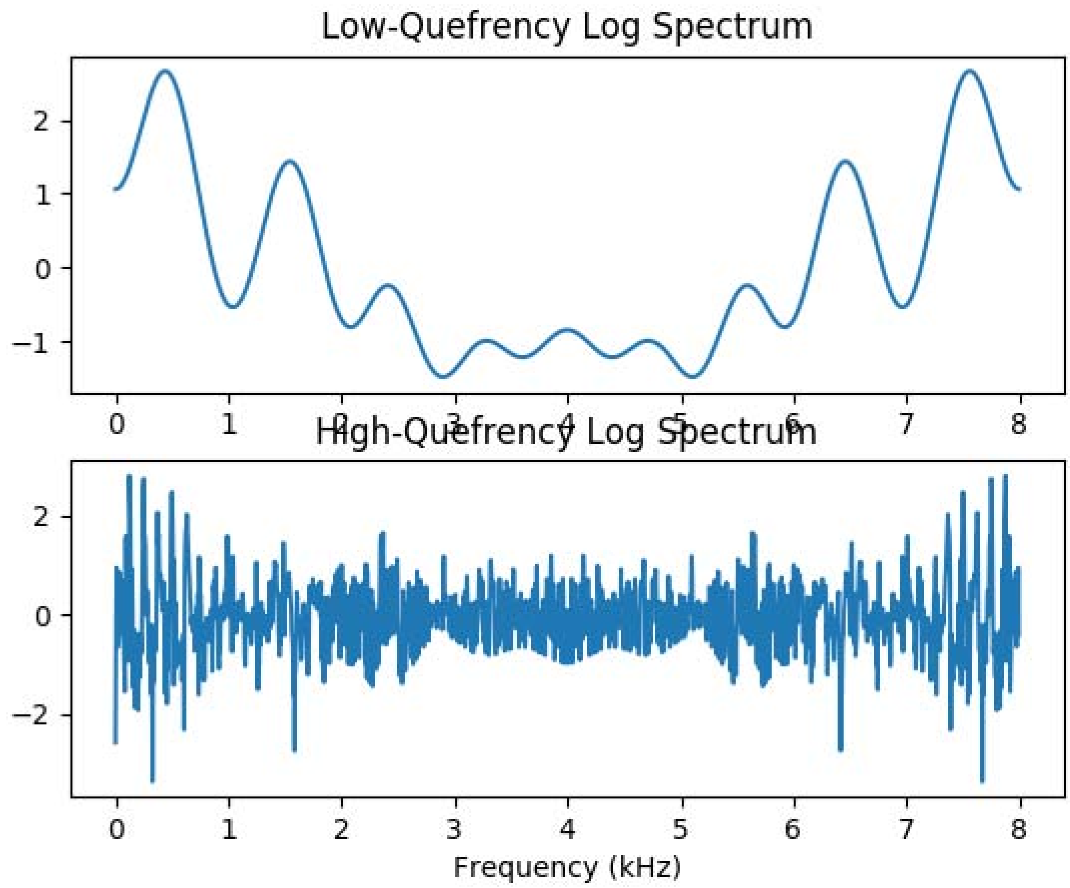
Liftering = filter(spectrum) = window(cepstrum)

(Bogert, Healy & Tukey, 1962)

Then we estimate the transfer function and excitation spectrum using the FFT:

$$\ln|H(f)| \approx FFT(\hat{h}[q])$$

$$\ln|E(f)| \approx FFT(\hat{e}[q])$$



Inverse Discrete Cosine Transform

- Log magnitude spectrum is symmetric: $\ln|S(f)| = \ln|S(-f)|$.
- In the IFFT definition, the real part is symmetric, and the imaginary part is antisymmetric. Suppose we define $S_k = \ln\left|S\left(\frac{kF_s}{N}\right)\right|$, then the definition of IFFT is

$$\hat{s}[q] = IFFT(\ln|S(f)|) = \frac{1}{N}\sum_{k=0}^{N-1} S_k e^{j\frac{2\pi kq}{N}}$$

…but since S_k is real, $S_{N-k} = S_k$ so

$$\hat{s}[q] = \frac{S_0 - (-1)^q S_M}{2M} + \frac{1}{M}\sum_{k=1}^{M-1} S_k \cos\left(\frac{\pi kq}{M}\right)$$

This is called the "inverse discrete cosine transform" or IDCT. It's half of the real symmetric IFFT of a real symmetric signal. (note M=N/2).

Type I DCT, IDCT, and Parseval's Theorem

$$S_k = \frac{\hat{s}[0] - (-1)^k \hat{s}[M]}{2} + \sum_{q=1}^{M-1} \hat{s}[q] \cos\left(\frac{\pi k q}{M}\right)$$

$$\hat{s}[q] = \frac{S_0 - (-1)^q S_M}{2M} + \frac{1}{M} \sum_{k=1}^{M-1} S_k \cos\left(\frac{\pi k q}{M}\right)$$

$$\hat{s}[0]^2 + \hat{s}[M]^2 + 2 \sum_{q=1}^{M-1} \hat{s}[q]^2 = \frac{1}{2M}\left({S_0}^2 + {S_M}^2 + 2 \sum_{k=1}^{M-1} {S_k}^2\right)$$

Type II Discrete Cosine Transform

- Suppose we define $C_k = \ln\left|S\left(\frac{(k+0.5)F_s}{N}\right)\right|$, and $c[n] = M\hat{s}[n]$. Then

$$c[n] = \frac{N}{2} IFFT(\ln|S(f)|) = \frac{1}{2}\sum_{k=0}^{N-1} C_k e^{j\frac{2\pi(k+0.5)n}{N}}$$

…but now $S_{N-1-k} = S_k$ so

$$c[n] = \sum_{k=0}^{M-1} C_k \cos\left(\frac{\pi(k+0.5)n}{M}\right)$$

This is called the “Type II DCT,” and it’s a lot more common than the Type I DCT because it eliminates the special handling of the k=0 and k=M terms.

Type II DCT, IDCT, and Parseval's Theorem

$$c[n] = \sum_{k=0}^{M-1} C_k \cos\left(\frac{\pi(k+0.5)n}{M}\right)$$

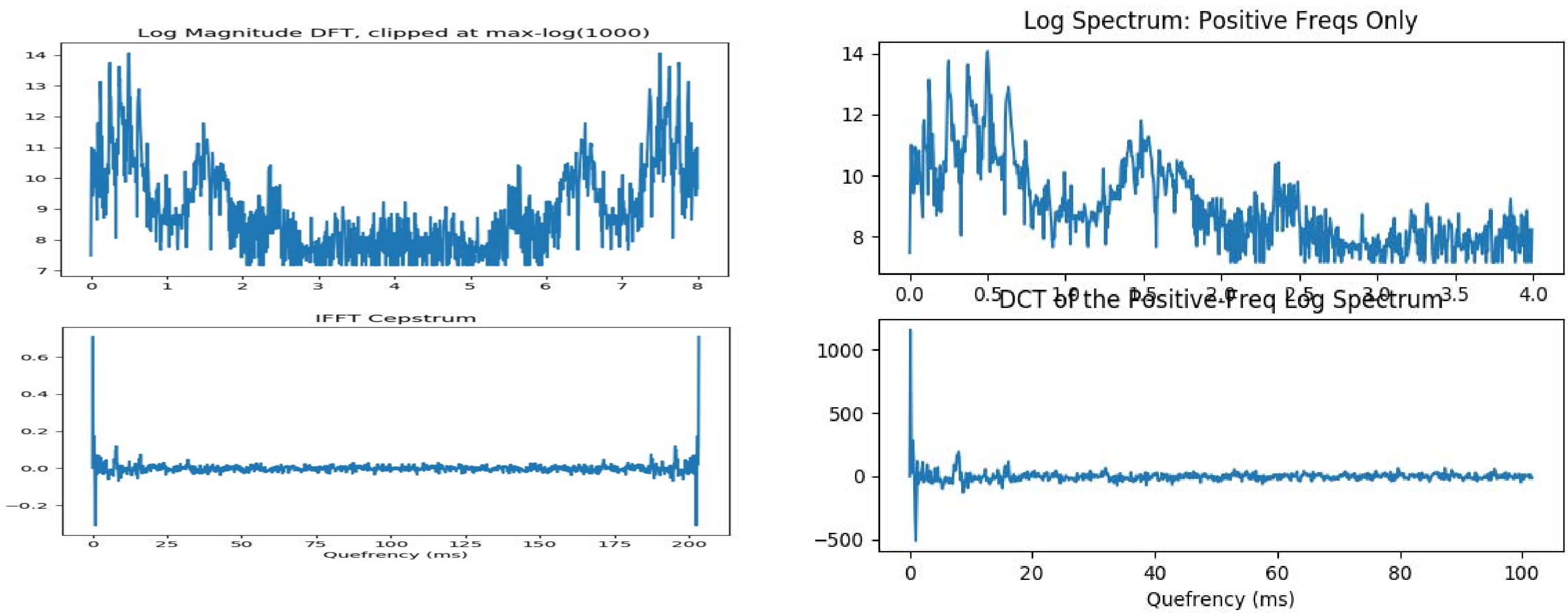
$$C_k = \frac{1}{M} \sum_{k=1}^{M-1} c[n] \cos\left(\frac{\pi(k+0.5)n}{M}\right)$$

$$\frac{1}{M}\left(c[0]^2 + 2\sum_{n=1}^{M-1} c[n]^2\right) = \sum_{k=0}^{M-1} {C_k}^2$$

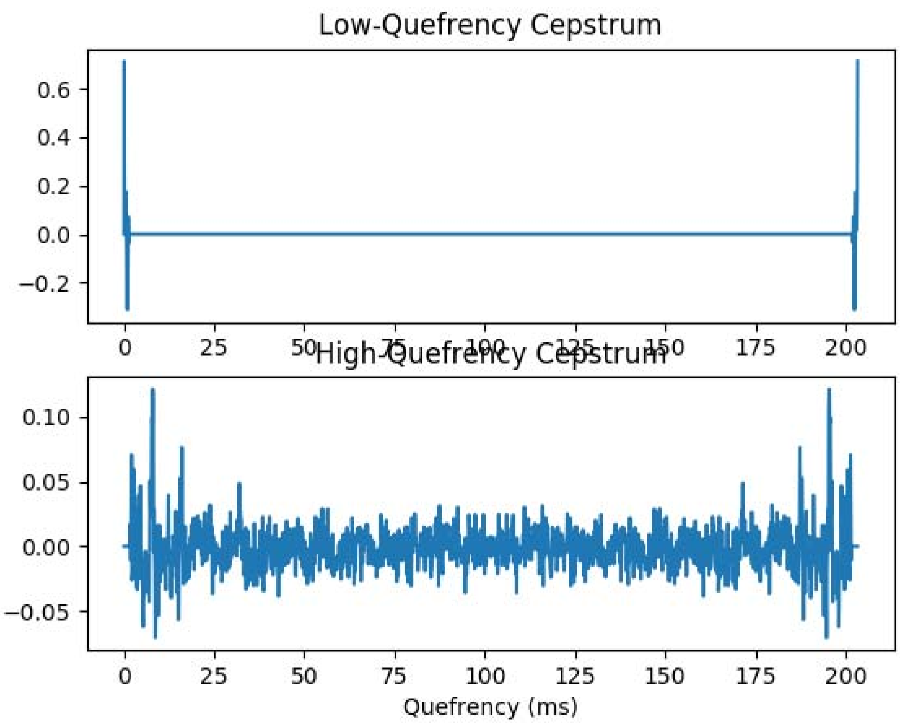
Details about type II DCT

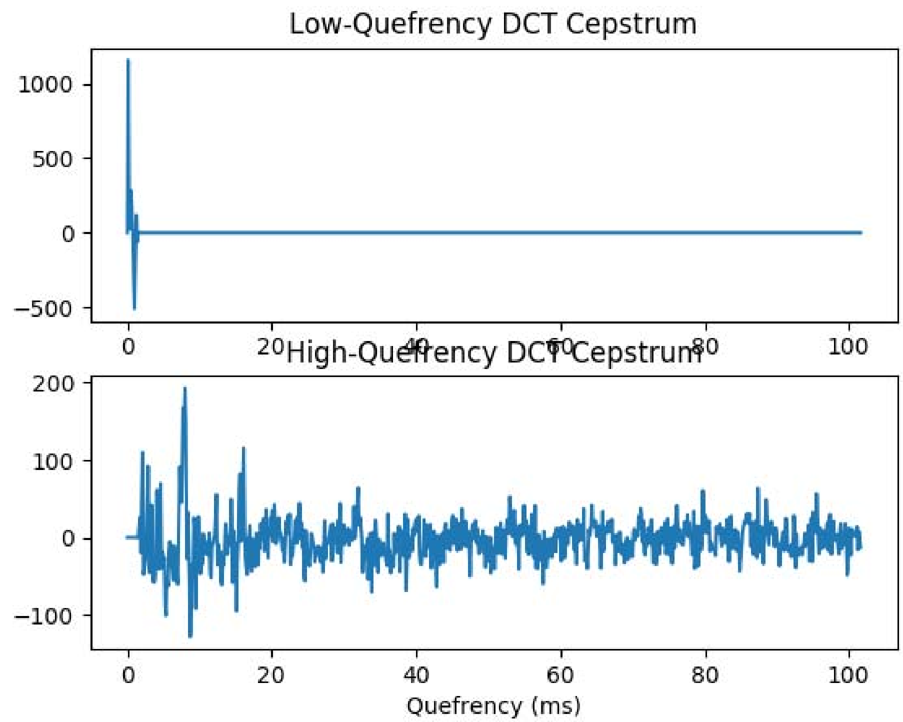
- It was defined as $C_k = \ln\left|S\left(\frac{(k+0.5)F_s}{N}\right)\right|$, but in practice we usually just use the FFT coefficients, $C_k \approx \ln\left|S\left(\frac{kF_s}{N}\right)\right|$. This approximation has no real impact on automatic speech recognition, but it might have some impact on pitch tracking – if you're trying to find out exactly what is the pitch frequency, then shifting by $\frac{F_s}{2N}$ might matter.
- The DCT and IDCT formulas are now easy, but Parseval's theorem still has a funny extra term for c[0]. But it doesn't matter because…
- Remember $c[0] = \sum_{k=0}^{M-1} C_k$ is the average log magnitude of the spectrum, i.e., a measure of the loudness. Loudness can be increased by just turning up the volume on the microphone, so we probably want to treat $c[0]$ differently from all of the other $c[n]$.

Discrete Cosine Transform = Half of the real symmetric IFFT of a real symmetric signal

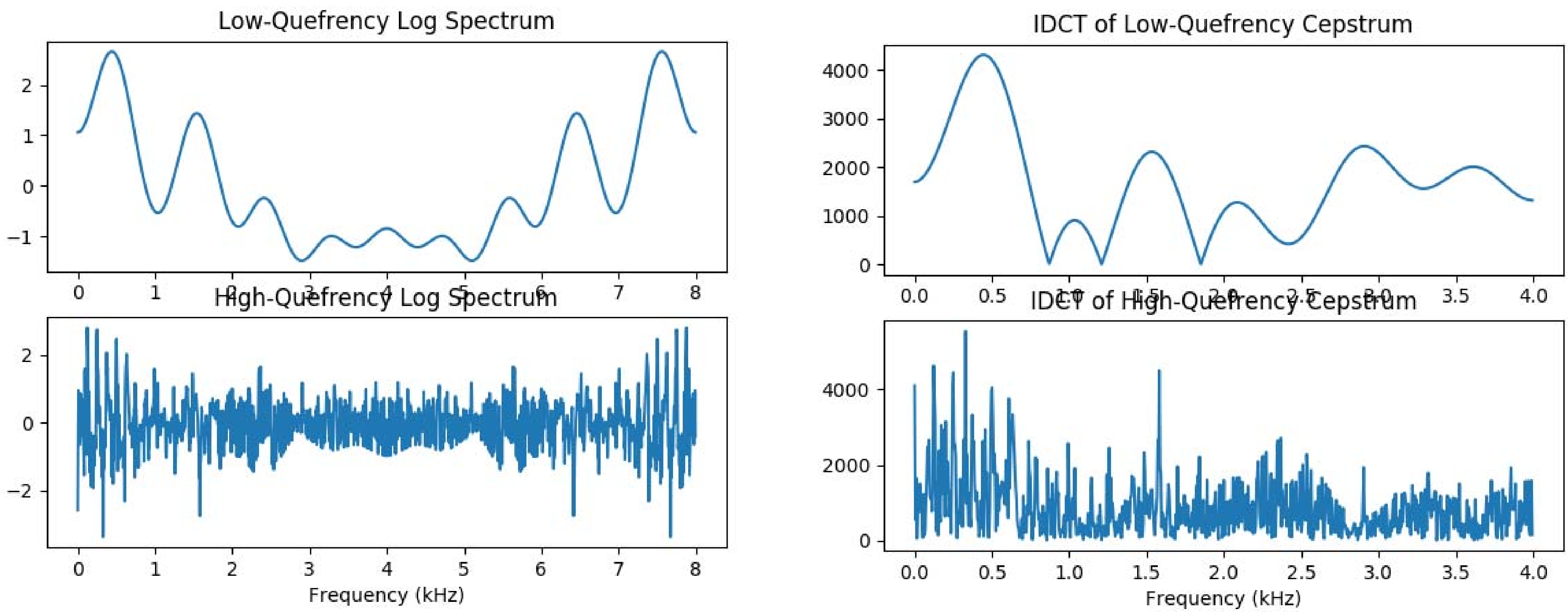


Lifter = window the IFFT (left) or DCT (right) cepstrum



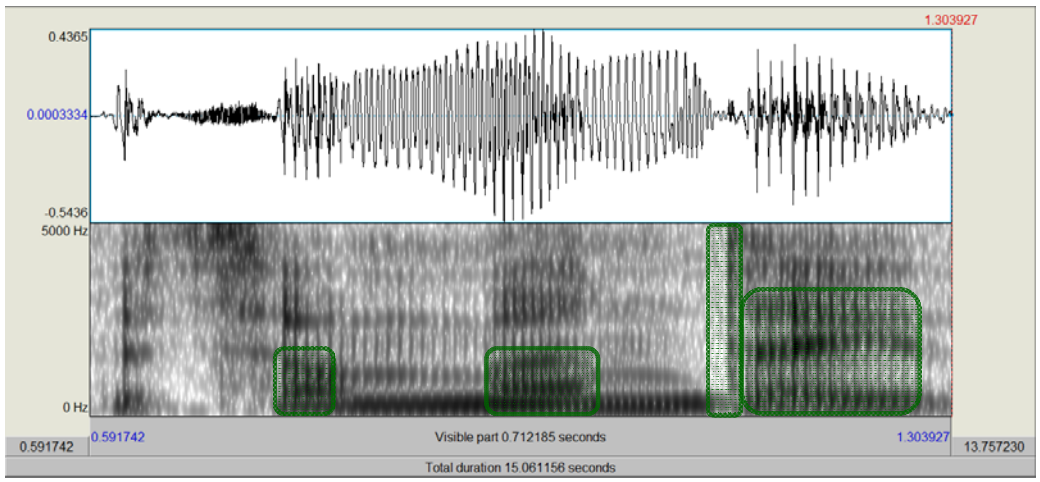


Both kinds of liftering give the same transfer function and excitation estimates



Spectrogram: ln(energy(frequency,time))

Scharenborg, 2017



bu t o nM o n d a y

Spectrum lets you measure formants, so it gives some information about vowels.

Timing is important to know about consonants.

<u>Spectrogram</u> = time on the horizontal axis, frequency on vertical axis.

Summary

- Source-filter model: $S(f) = H(f)E(f)$
  - Voiced excitation is an impulse train in time (with period = the pitch period T_0), whose Fourier transform is an impulse train in frequency (with inter-harmonic spacing equal to the pitch frequency F_0)
  - Transfer function is nearly $H(f) = 1$ at most frequencies, but with big peaks near the resonant frequencies, which are called formants
- Phones, phonemes, and allophones
- Estimating the transfer function and excitation
  - $\ln|S(f)| = \ln|H(f)| + \ln|E(f)|$
  - The transfer function is low-quefrency, excitation is high-quefrency
  - Cepstrum = $IFFT(\ln|S(f)|)$= $DCT(\ln|S(f)|)$
  - Liftering = windowing the cepstrum
  - DCT = half of the real symmetric IFFT of a real symmetric signal