ECE 417, Lecture 11: MFCC

Mark Hasegawa-Johnson

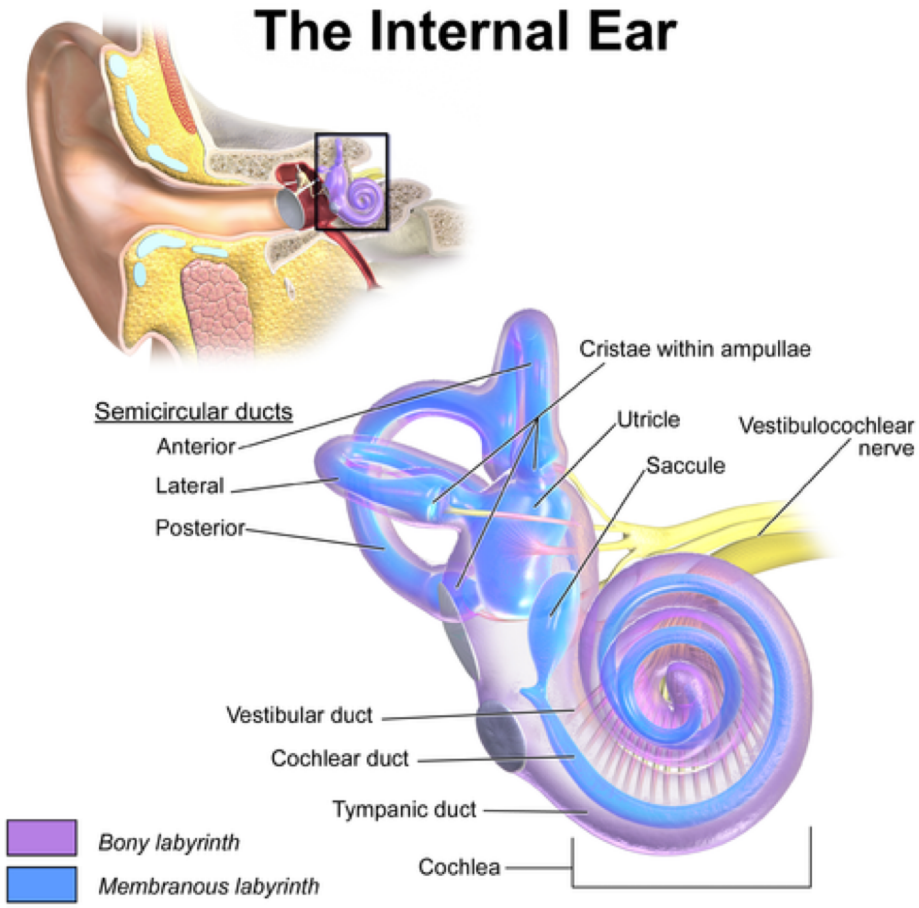
10/2/2018

Content

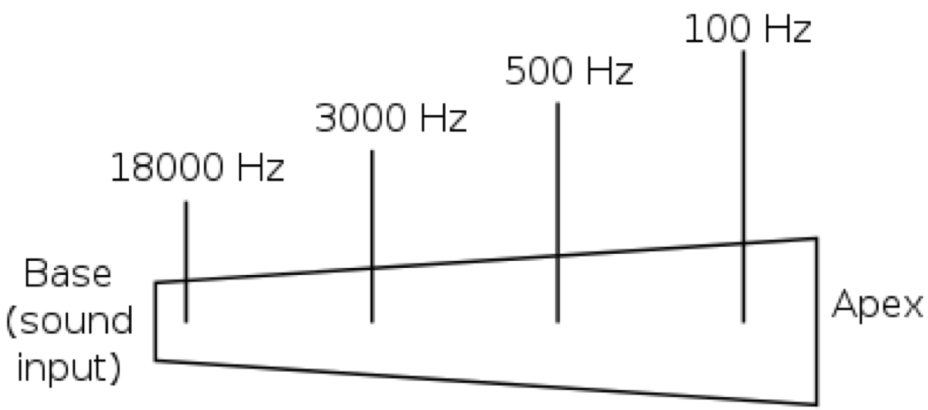
- What spectrum do people hear? The basilar membrane
- Frequency scales for hearing: mel scale
- Mel-filter spectral coefficients (also called “filterbank features”)
- Speech Production: Consonants and Vowels
- Parseval’s Theorem: Cepstral Distance = Spectral Distance

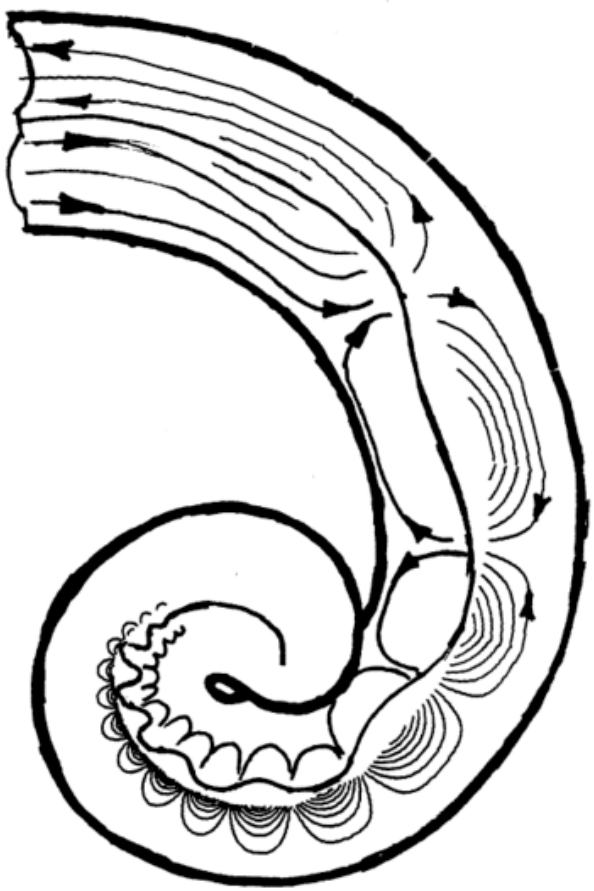
What spectrum do people hear? Basilar membrane

Inner ear


The Internal Ear
Cristae within ampullae
Semicircular ducts
Anterior
Lateral
Posterior
Utricle
Saccule
Vestibulocochlear nerve
Vestibular duct
Cochlear duct
Tympanic duct
Cochlea
Bony labyrinth
Membranous labyrinth

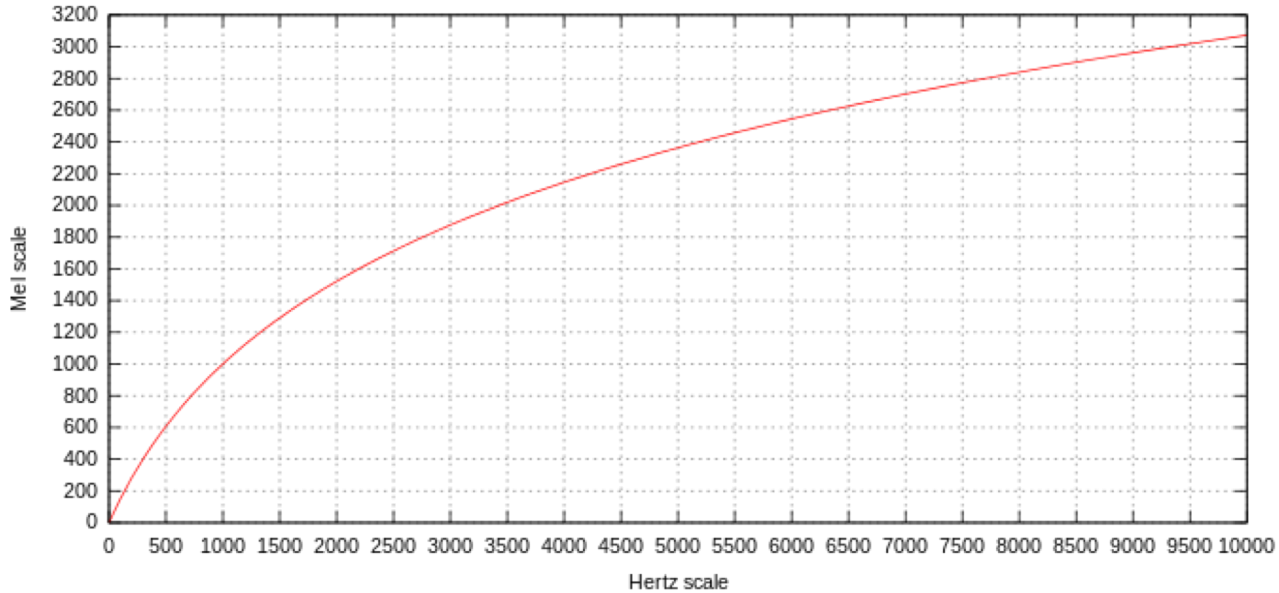
Basilar membrane of the cochlea = a bank of mechanical bandpass filters





Frequency scales for hearing: mel scale

Mel-scale



- The experiment:
  - Play tones A, B, C
  - Let the user adjust tone D until pitch(D)-pitch(C) sounds the same as pitch(B)-pitch(A)
- Analysis: create a frequency scale m(f) such that m(D)-m(C) = m(B)-m(A)
- Result: $m(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$

Mel Frequency Spectral Coefficients (Filterbank Coefficients)

Mel filterbank coefficients: convert the spectrum from Hertz-frequency to mel-frequency

- Goal: instead of computing

$$C_k = \ln\left|S\left(\frac{(k+0.5)F_s}{N}\right)\right|$$

We want

$$C_k = \ln|S(f_k)|$$

Where the frequencies f_k are uniformly spaced on a mel-scale, i.e., $\mathrm{m}(f_{k+1}) - \mathrm{m}(f_k)$ is a constant across all k.

The problem with that idea: we don't want to just sample the spectrum. We want to summarize everything that's happening within a frequency band.

Mel filterbank coefficients: convert the spectrum from Hertz-frequency to mel-frequency

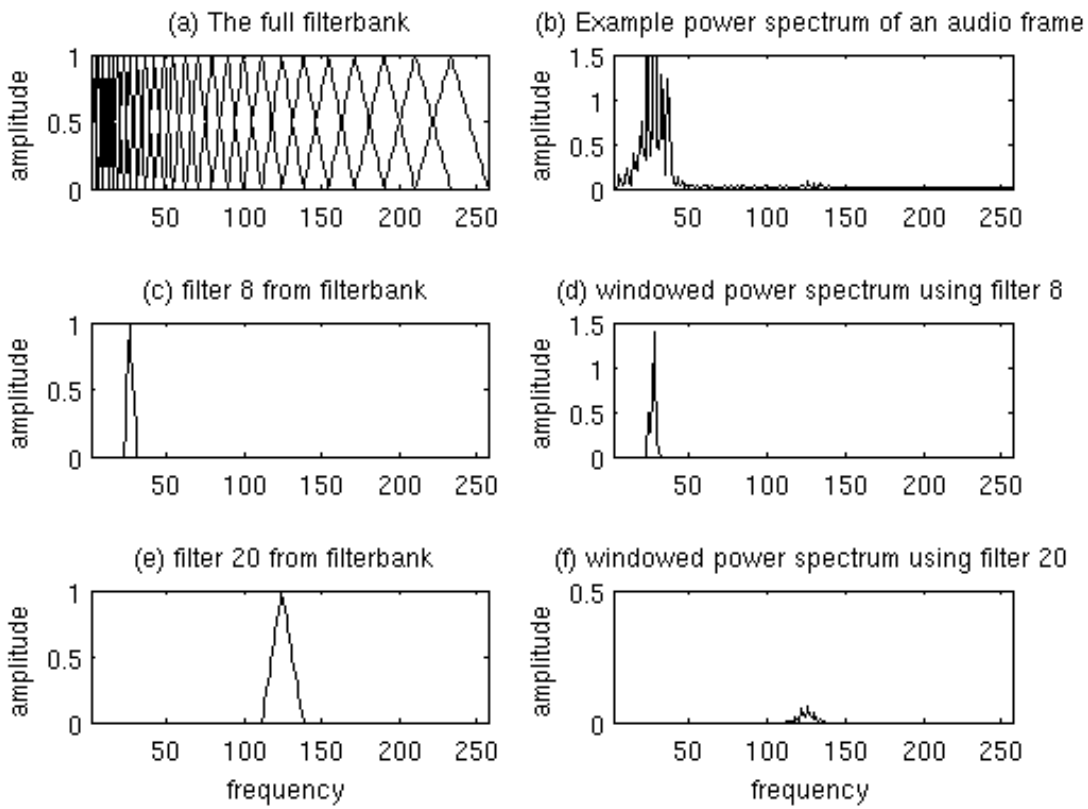
The solution:

$$C_m = \ln \sum_{k=0}^{\frac{N}{2}-1} W_m(k) \left| S\left(\frac{kF_s}{N}\right) \right|$$

Where

$$W_m(k) = \begin{cases} \dfrac{\dfrac{kF_s}{N} - f_{m-1}}{f_m - f_{m-1}} & f_m \geq \dfrac{kF_s}{N} \geq f_{m-1} \\ \dfrac{f_{m+1} - \dfrac{kF_s}{N}}{f_{m+1} - f_m} & f_{m+1} \geq \dfrac{kF_s}{N} \geq f_m \\ 0 & otherwise \end{cases}$$

Mel filterbank coefficients: convert the spectrum from Hertz-frequency to mel-frequency



Speech Production

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

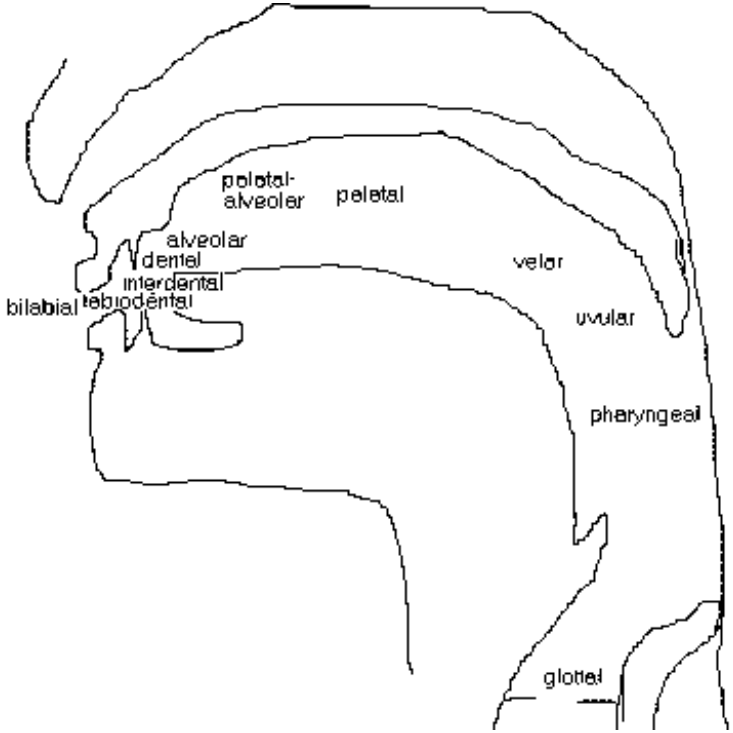
- Sources: there are only three, all of them have wideband spectrum
  - Voicing: vibration of the vocal folds, same type of aerodynamic mechanism as a flag flapping in the wind.
  - Frication or Aspiration: turbulence created when air passes through a narrow aperture
  - Burst: the “pop” that occurs 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

Different sounds: Consonants

Scharenborg, 2017

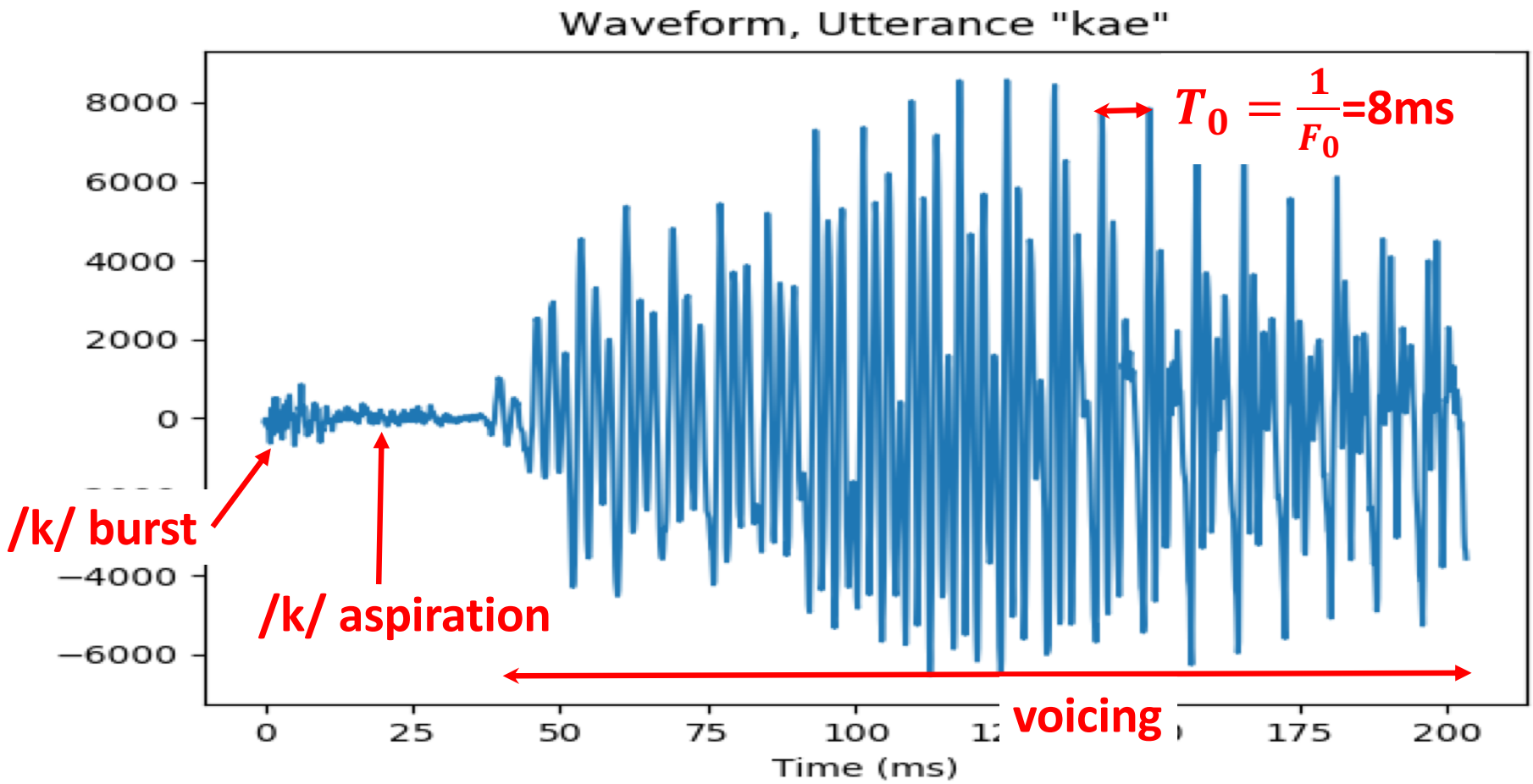
- Place of articulation
  - Where is the constriction/blocking of the air stream?



- 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 signal: Time domain



Speech sound production

Scharenborg, 2017

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

Recorded in 1962, Ken Stevens
Source: YouTube

The Source-Filter Model

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

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

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

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

- Excitation includes all of the information about voicing, frication, or burst.
- Filter includes all of the information about the vocal tract resonances, which are called “formants.”

Source: V/UV, F_0

- 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:

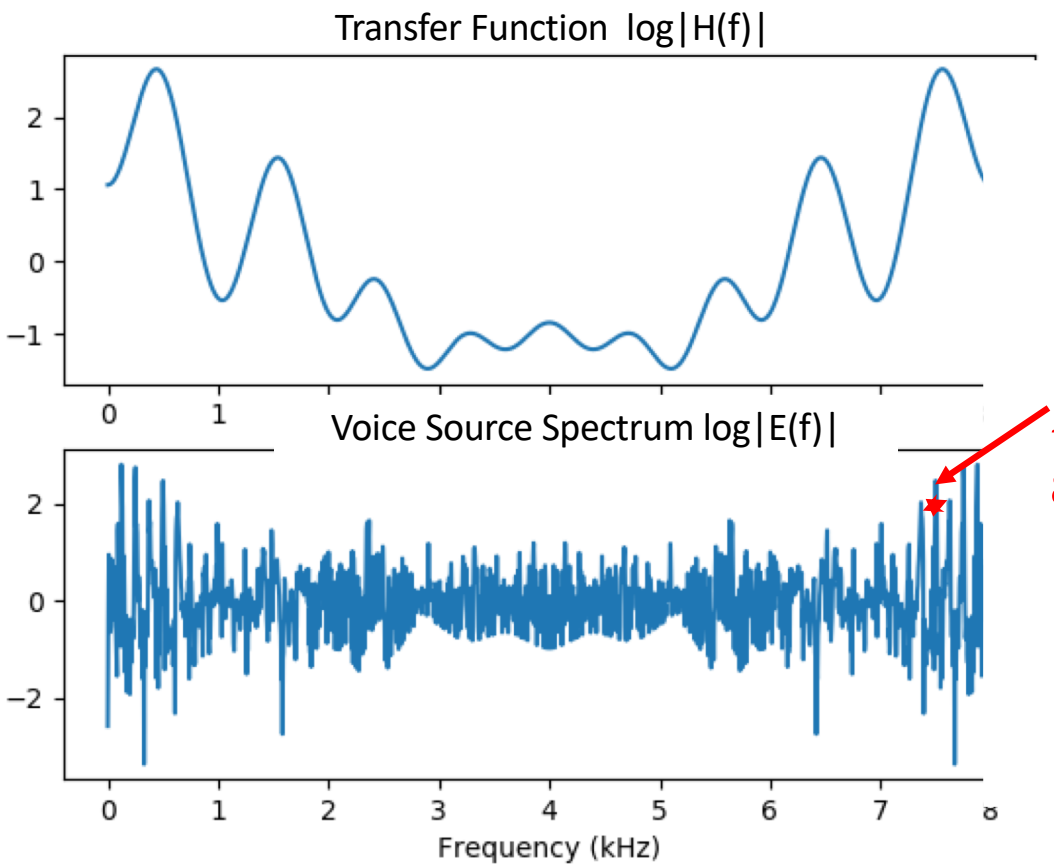
$$e(t) = \sum_{m=-\infty}^{\infty} \delta(t - mT_0)$$

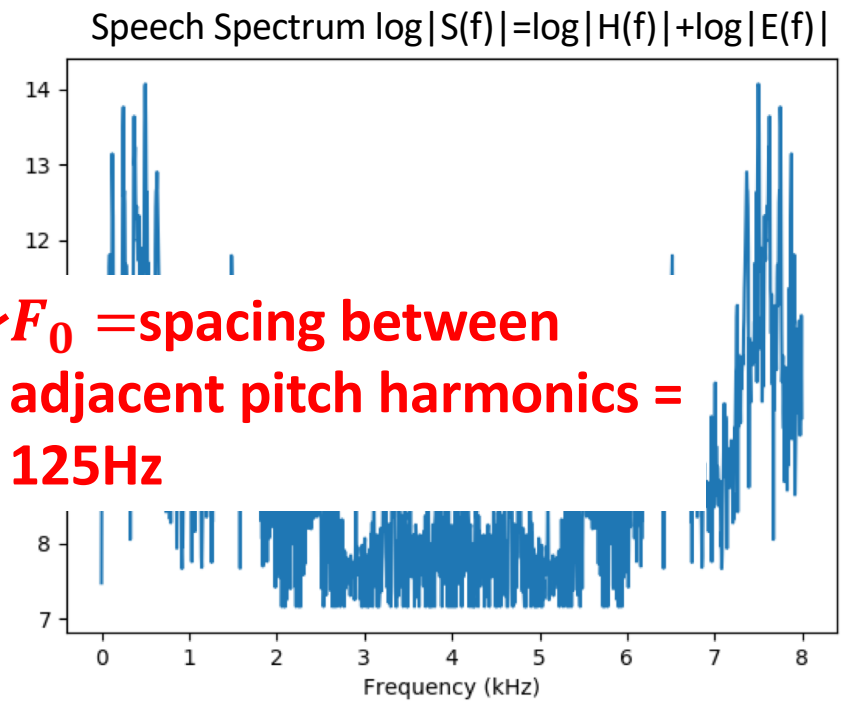
- The Fourier transform of an impulse train is an impulse train (to prove this: use Fourier series):

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

...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.

The Source-Filter Model



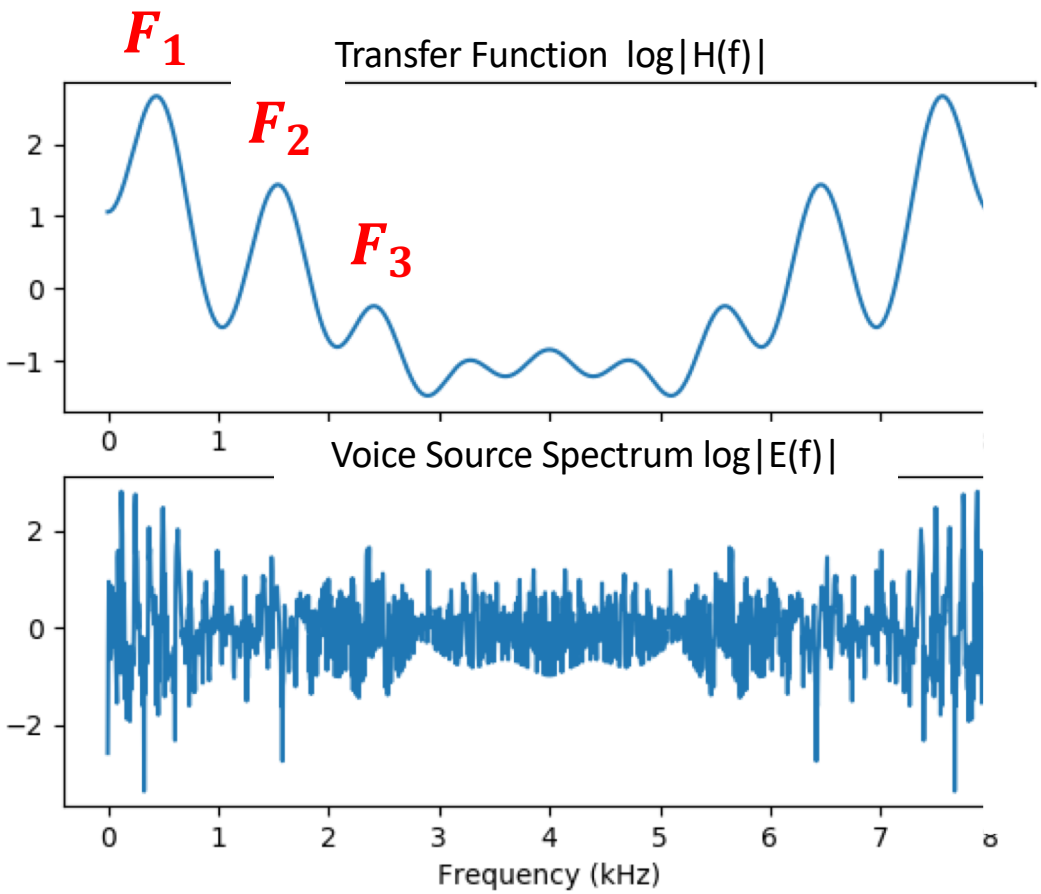


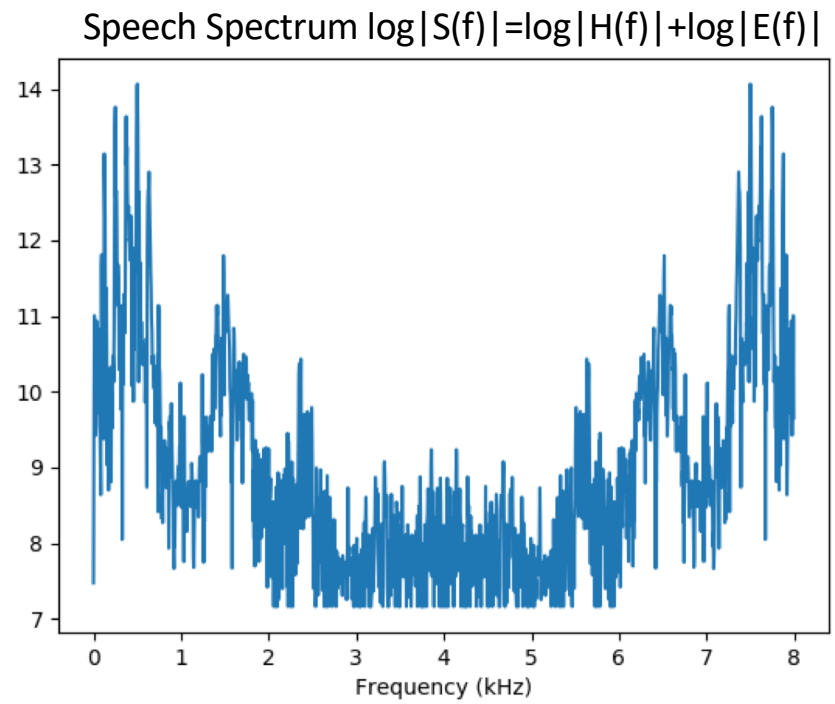
F_0 =spacing between adjacent pitch harmonics = 125Hz

Filter: $F_1, F_2, F_3, \ldots$

- 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.

The Source-Filter Model

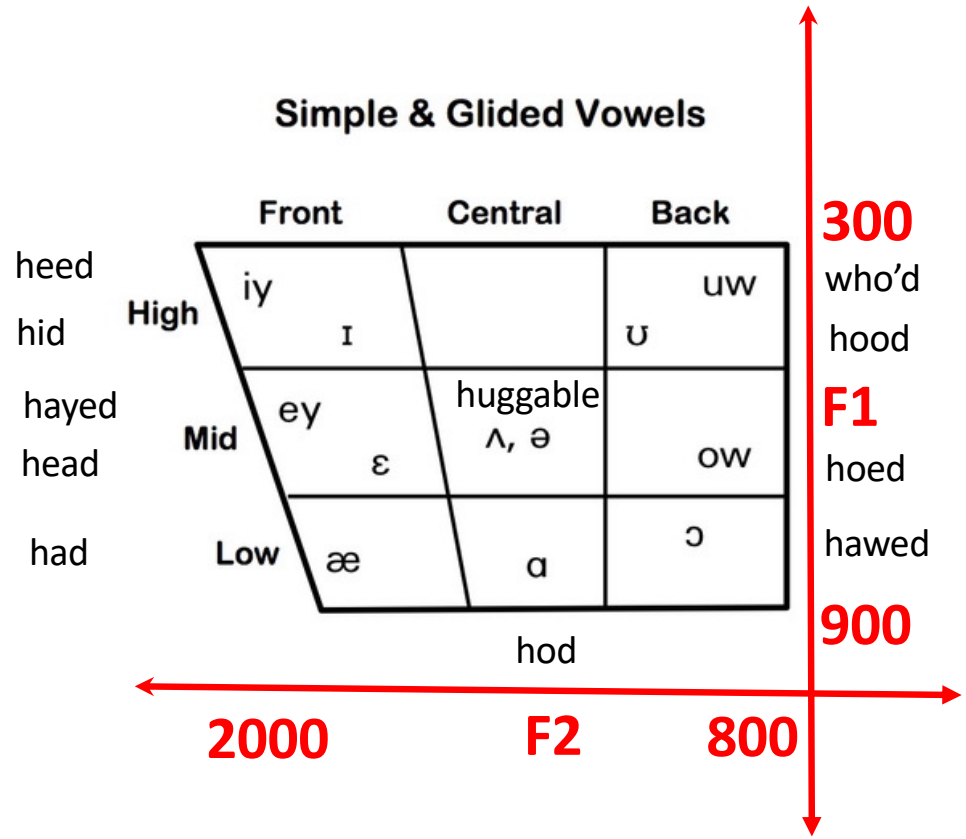




Different sounds: Vowels

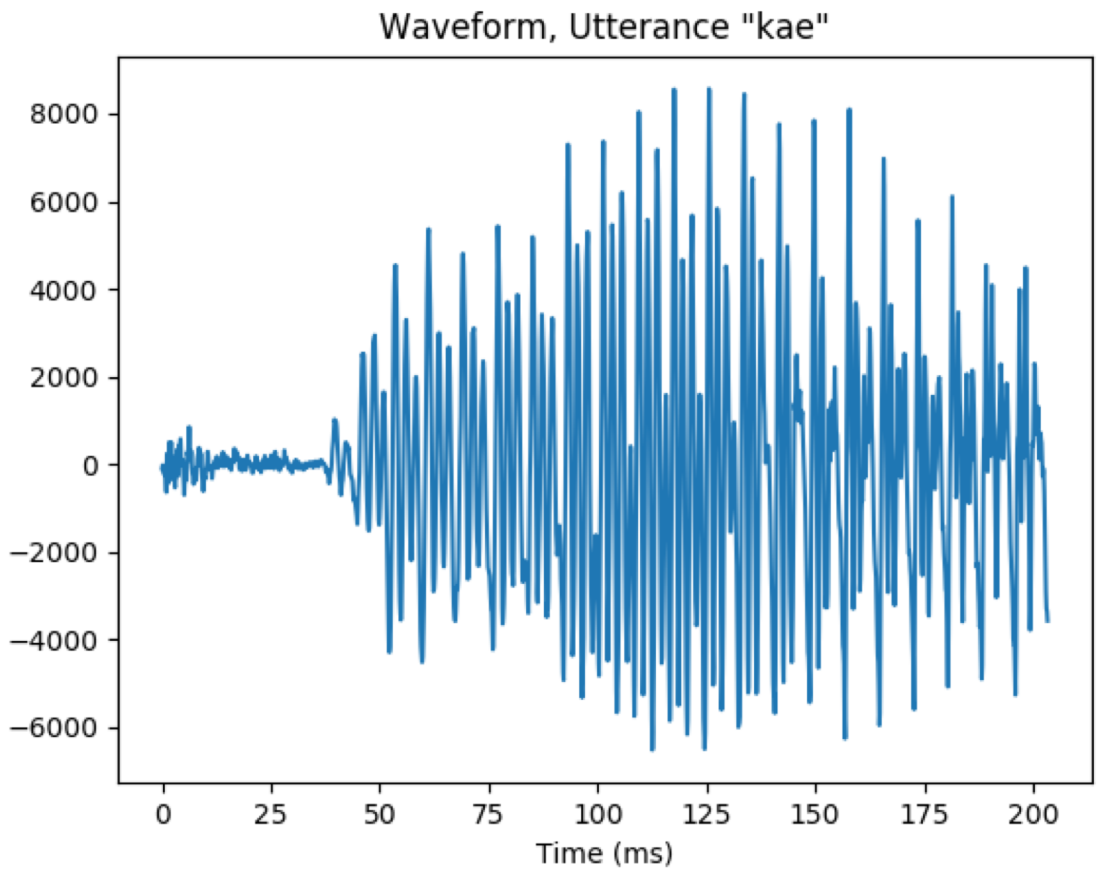
Scharenborg, 2017

- 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/



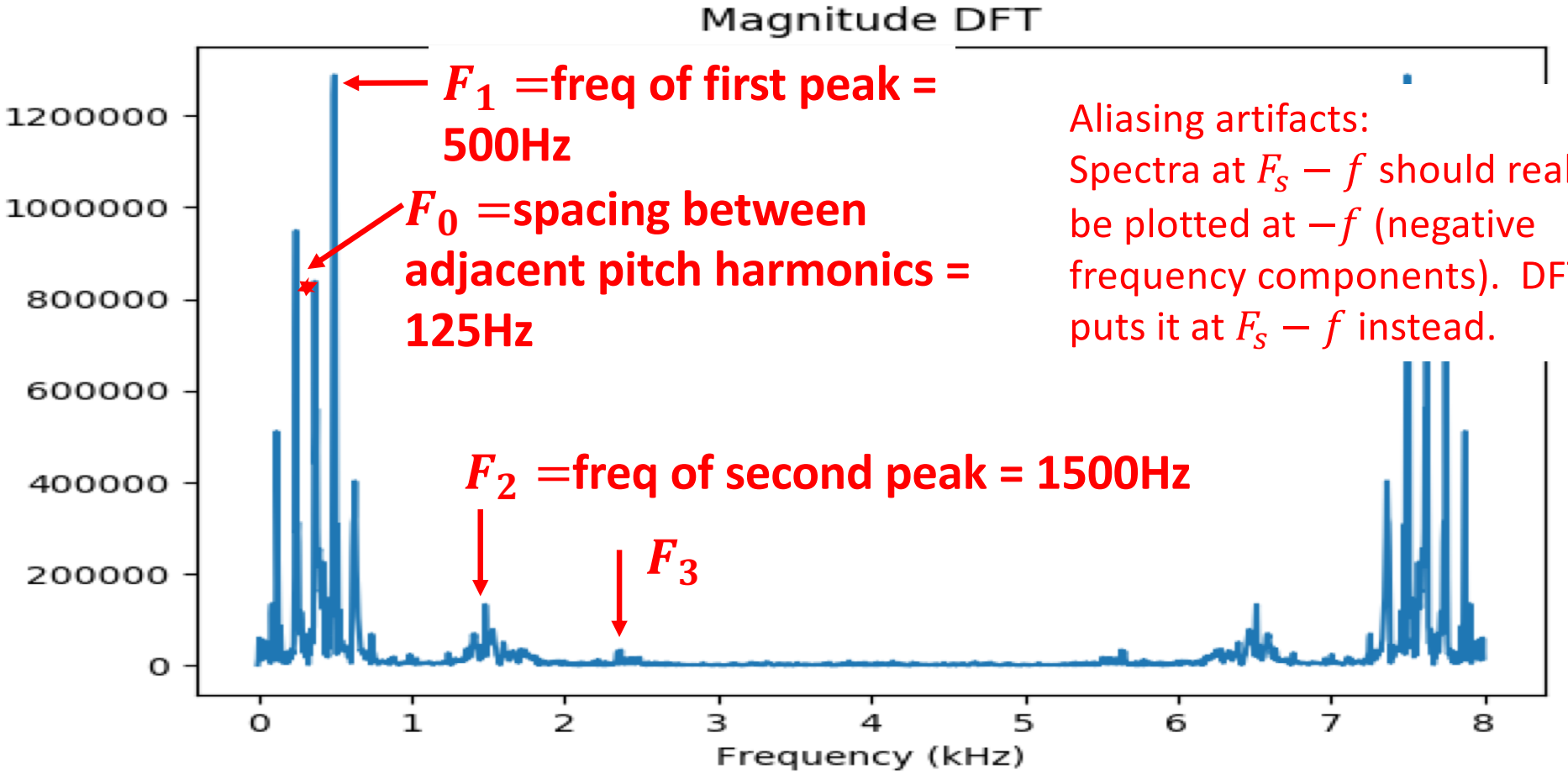
Time domain signal: Hard to tell what he was saying

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



Magnitude spectrum: A little easier

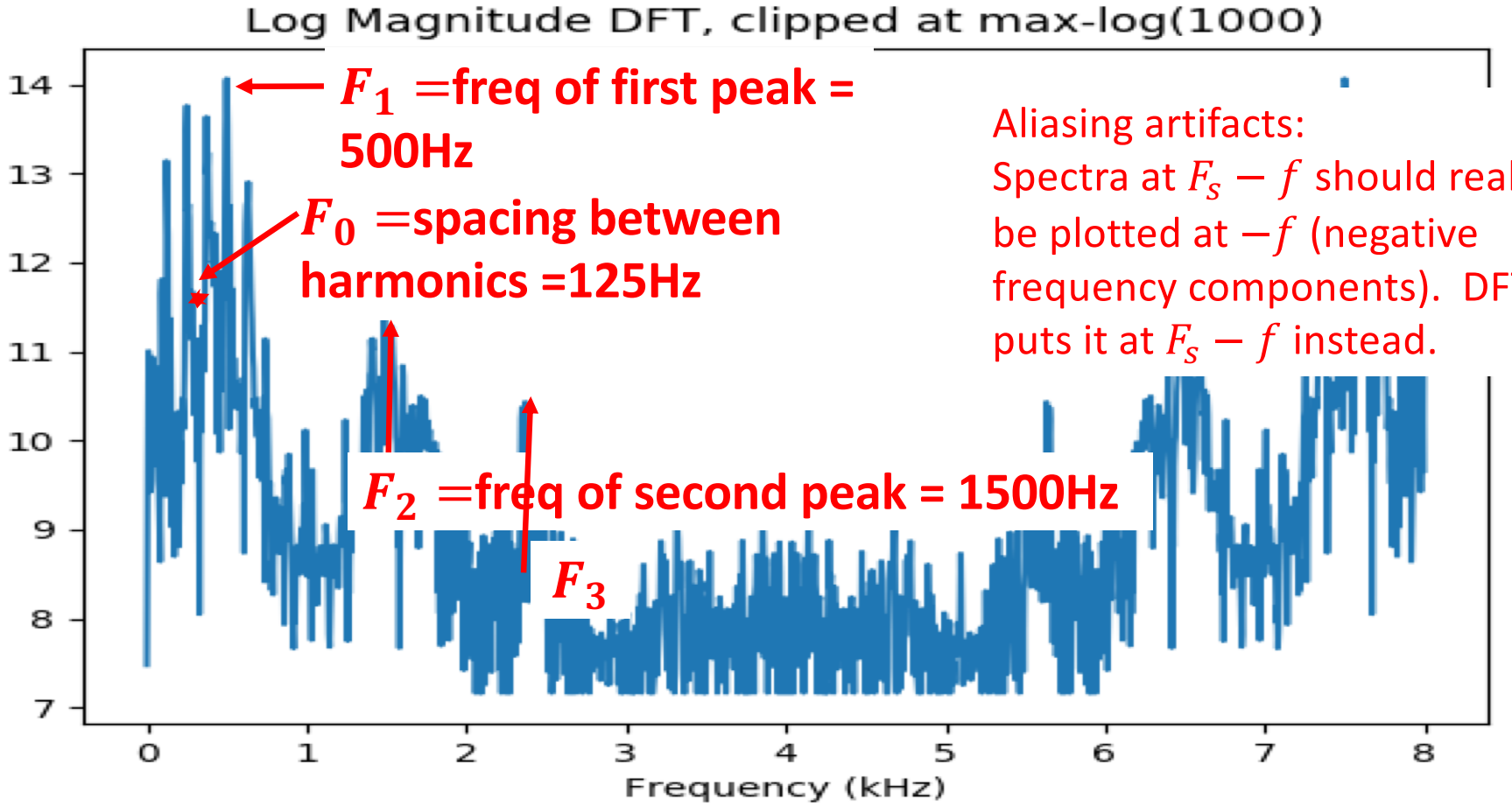
$$S(f) = H(f)E(f)$$



Aliasing artifacts:
Spectra at $F_s - f$ should really be plotted at $-f$ (negative frequency components). DFT puts it at $F_s - f$ instead.

Log magnitude spectrum: A lot easier

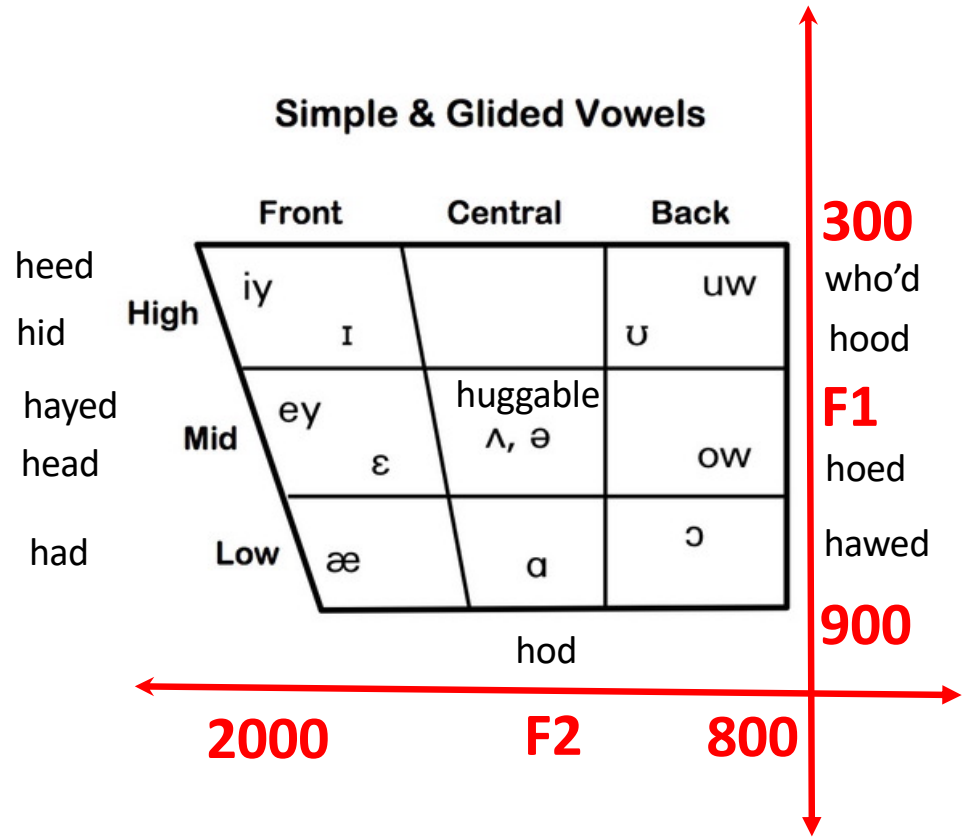
$$\ln|S(f)| = \ln|H(f)| + \ln|E(f)|$$



Different sounds: Vowels

Scharenborg, 2017

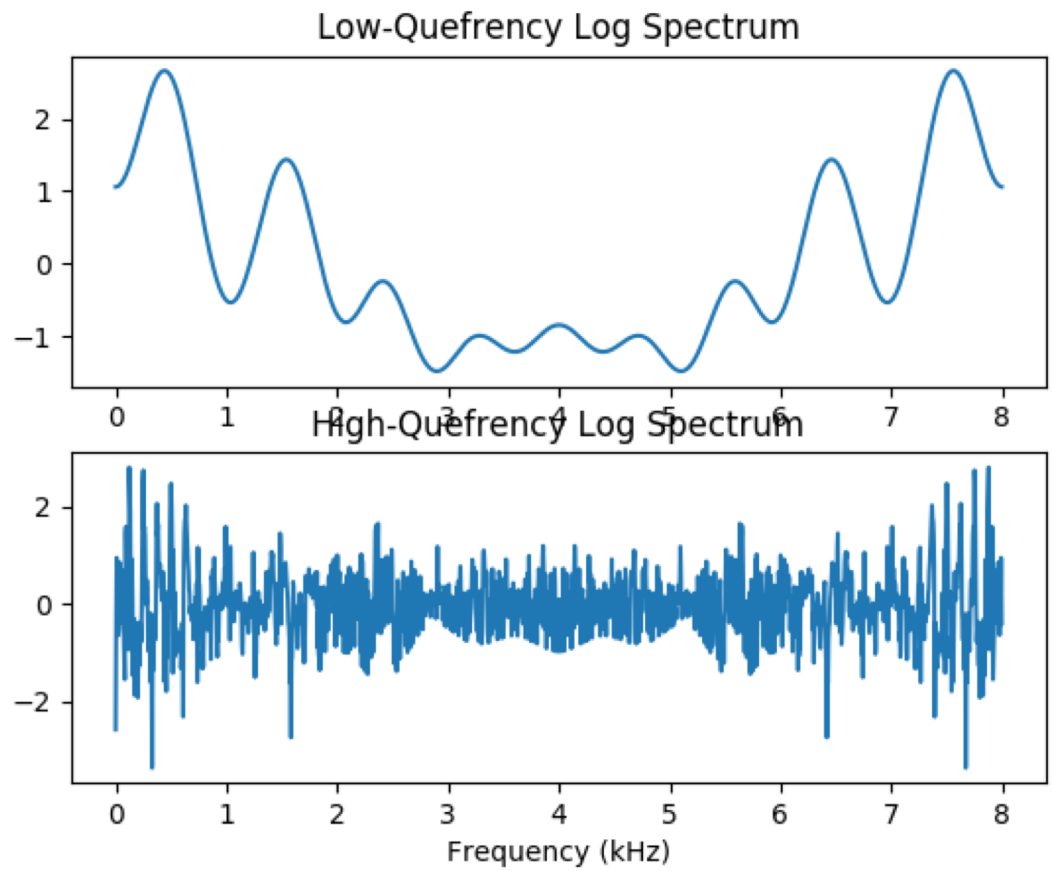
- Tongue height:
  - Low: e.g., /a/
  - Mid: e.g., /e/
  - High: e.g., /i/
- Tongue advancement:
  - Front : e.g., /i/
  - Central : e.g., /ə/
  - Back : e.g., /u/
- Lip rounding:
  - Unrounded: e.g., /ɪ, ɛ, e, ə/
  - Rounded: e.g., /u, o, ɔ/
- Tense/lax:
  - Tense: e.g., /i, e, u, o, ɔ, ɑ/
  - Lax: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



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)

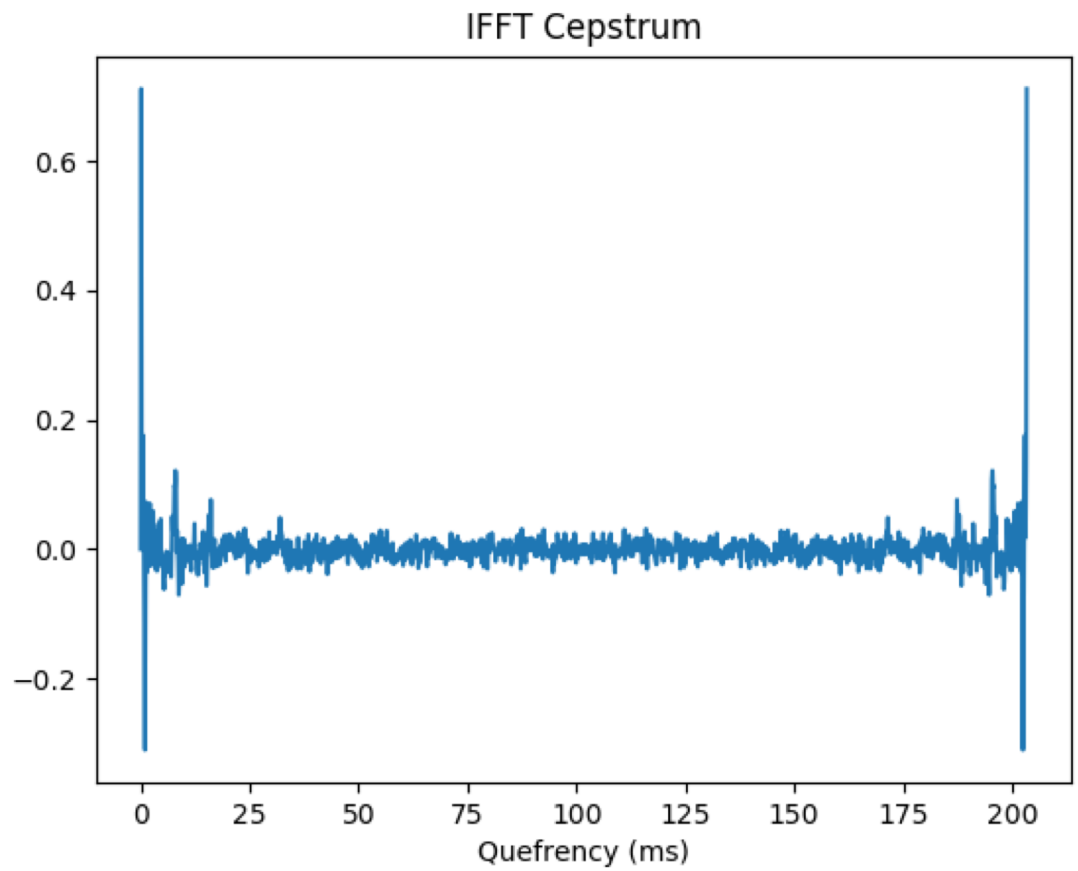
$$\hat{s}[q] = IFFT(\ln|S(f)|)$$

- 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.



Inverse Discrete Cosine Transform

Log magnitude spectrum is symmetric: $\ln|S(f)| = \ln|S(-f)|$. Suppose we assume that $\ln|S(0)| = 0$ and $\ln\left|S\left(\frac{F_s}{2}\right)\right| = 0$, then

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

$$= \frac{2}{N}\sum_{k=1}^{N/2-1} \ln\left|S\left(\frac{kF_s}{N}\right)\right| \cos\left(\frac{\pi kq}{M}\right)$$

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).

Liftering = filter(spectrum) = window(cepstrum)

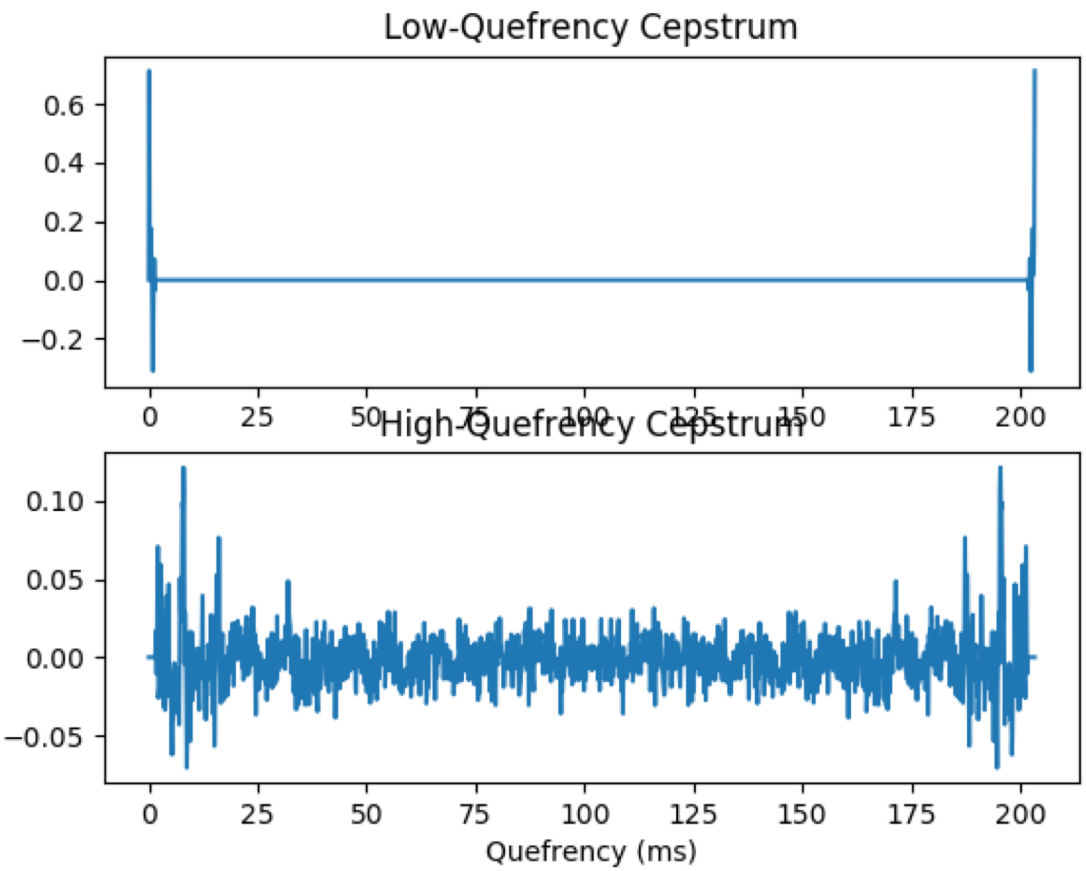
(Bogert, Healy & Tukey, 1962)

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:

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

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

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



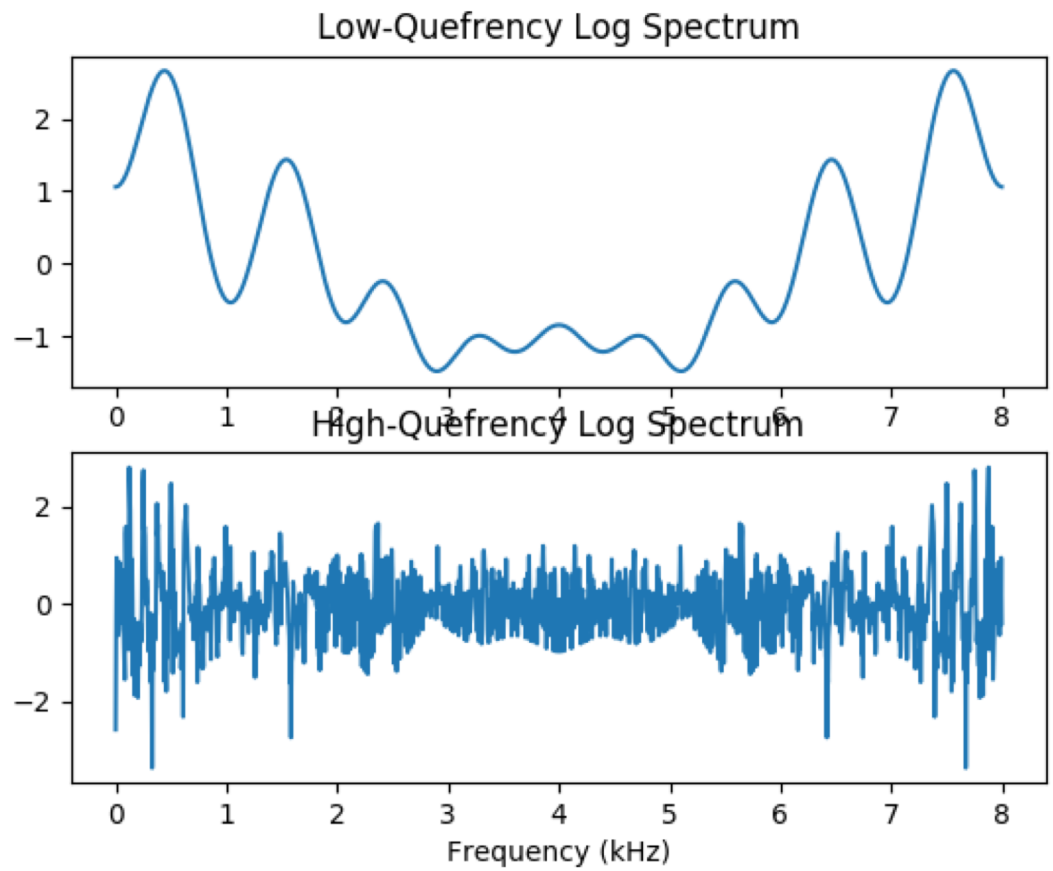
Liftering = filter(spectrum) = window(cepstrum)

(Bogert, Healy & Tukey, 1962)

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

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

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



Parseval's Theorem

L2 norm of a signal equals the L2 norm of its Fourier transform.

Parseval's Theorem: Examples

- Fourier Series:

$$\frac{1}{T}\int_0^T |x(t)|^2 dt = \sum_{k=-\infty}^{\infty} |X_k|^2$$

- DTFT:

$$\sum_{n=-\infty}^{\infty} |x[n]|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |X(\omega)|^2 d\omega$$

- DFT:

$$\sum_{n=0}^{N-1} |x[n]|^2 = \frac{1}{N}\sum_{k=0}^{N-1} |X[k]|^2$$

Parseval's Theorem: Vector Formulation

Suppose we define the vectors $\vec{c}$ and $\vec{C}$ as the cepstrum and the log spectrum, thus

$$\vec{c} = \begin{bmatrix} c_1 \\ \cdots \\ c_{N/2-1} \end{bmatrix}, \vec{C} = \begin{bmatrix} C_1 \\ \cdots \\ C_{N/2-1} \end{bmatrix}$$

Parseval's Theorem: Vector Formulation

That way Parseval's theorem can be written very simply as

$$\sum_{n=1}^{N/2-1} {c_n}^2 = \sum_{k=1}^{N/2-1} {C_k}^2$$

…or even more simply as…

$$\|\vec{c}\|^2 = \left\|\vec{C}\right\|^2$$

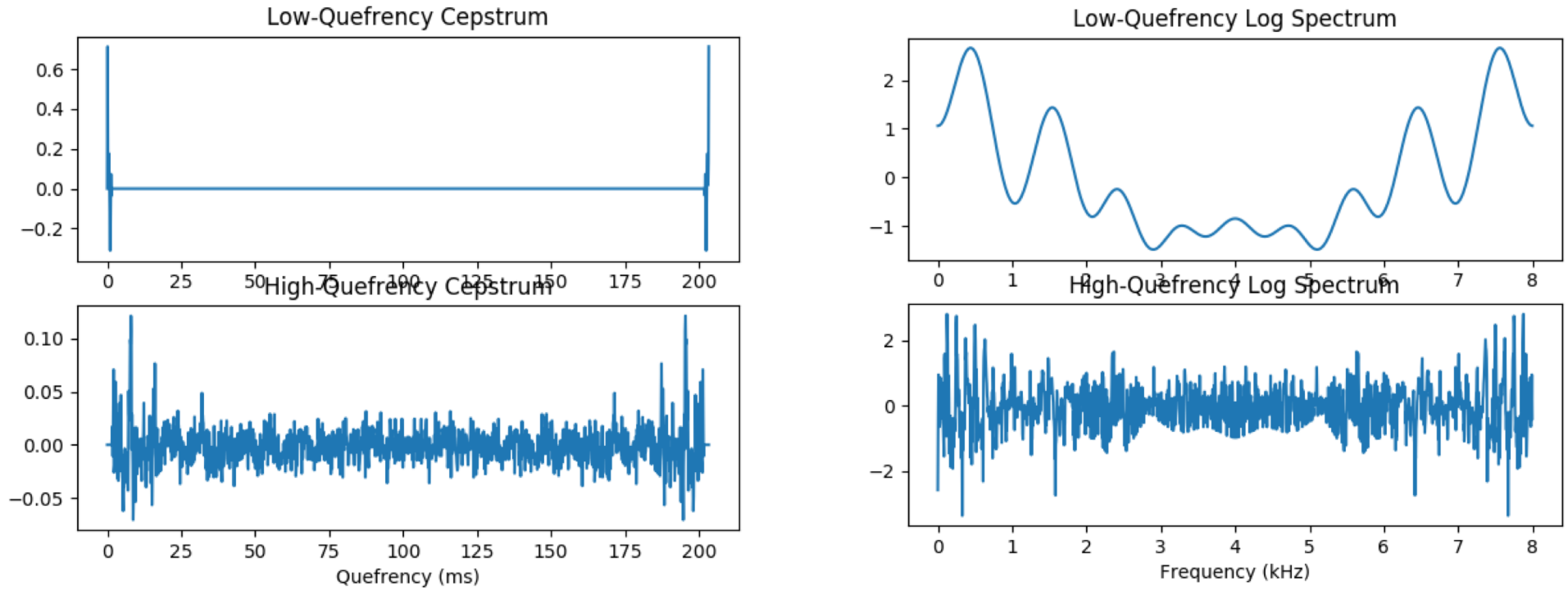
i.e., the L2 norm of the cepstrum equals the L2 norm of the log spectrum.

What it means for Gaussian classifiers

Suppose we have two acoustic signals $x(t)$ and $y(t)$, and we want to find out how different they sound. If they have static spectra, then a good measure of their difference is the distance between their log spectra:

$$D = \sum_{k=1}^{N/2-1} (X_k - Y_k)^2 = \sum_{n=1}^{N/2-1} (x_n - y_n)^2 = \|\vec{x} - \vec{y}\|^2 = \left\|\vec{X} - \vec{Y}\right\|^2$$

Low-pass liftering smooths the spectrum



Low-pass liftered L2 norm

If you want to know whether two signals are the same vowel, then you want to know how different their smoothed spectra are. Let H(k) be your liftering function. You lifter the log spectrum = windowing the cepstrum, then find the distance:

$$D = \sum_{k=0}^{M-1} (H(k) * X_k - H(k) * Y_k)^2 = \sum_{n=0}^{M-1} h^2[n](x_n - y_n)^2$$

Low-pass liftered L2 norm

In particular, suppose

$$h[n] = \begin{cases} 1 & 0 < n \leq 15 \\ 0 & n > 15 \end{cases}$$

Then

$$\sum_{k=0}^{M} \left(H(k) * \ln \left| X\left(\frac{(k+0.5)F_s}{N} \right) \right| - H(k) * \ln \left| Y\left(\frac{(k+0.5)F_s}{N} \right) \right| \right)^2$$

$$= \sum_{n=1}^{15} (x_n - y_n)^2$$

MFCC: mel-frequency cepstral coefficients

- Divide the acoustic signal into frames
- Compute the magnitude FFT of each frame
- Filterbank coefficients: $C_m = \ln \sum_{k=0}^{\frac{N}{2}-1} W_m(k) \left| S\left(\frac{kF_s}{N}\right) \right|$
- MFCC: $c[n] = \sum_{m=0}^{M-1} C_m \cos\left(\frac{\pi(m+0.5)n}{M}\right)$
- Liftering: keep only the first 12-15 MFCC coefficients, set the rest to zero.