ECE 448
Speech

Mark Hasegawa-Johnson, 5/2022

Outline

- Speech Production: Source-Filter Model
- Speech Perception: Spectrogram and Filterbank Coefficients
- Phonemes: Vowels and Consonants
- Automatic Speech Recognition
- Text-to-Speech Synthesis

Speech

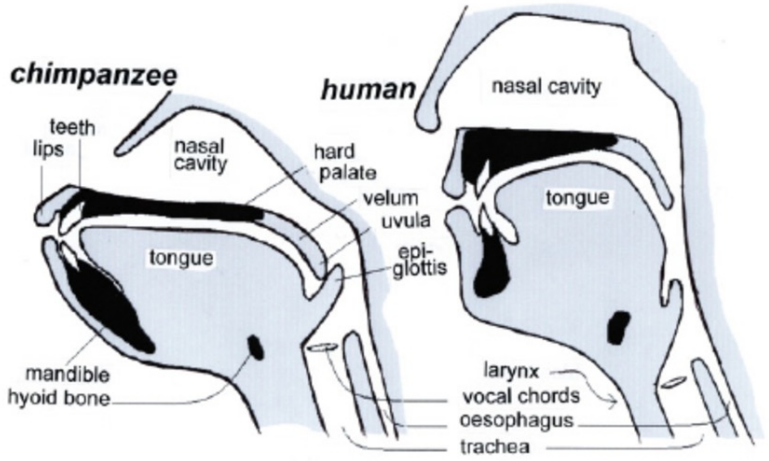
(Slide: Scharenborg, 2017)

- 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

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



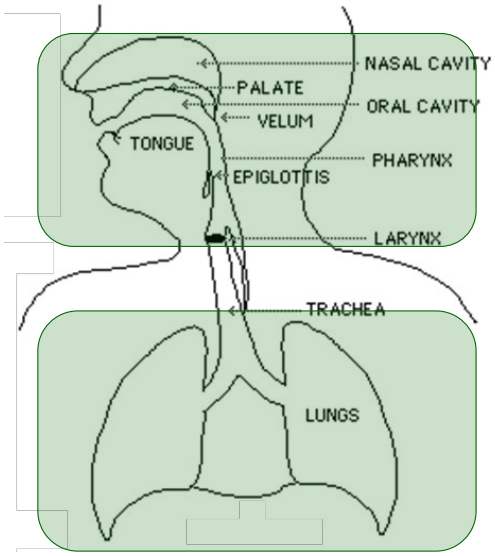
- Neanderthals were not capable of human speech
- Modern human vocal tract: since 50,000 years

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* =

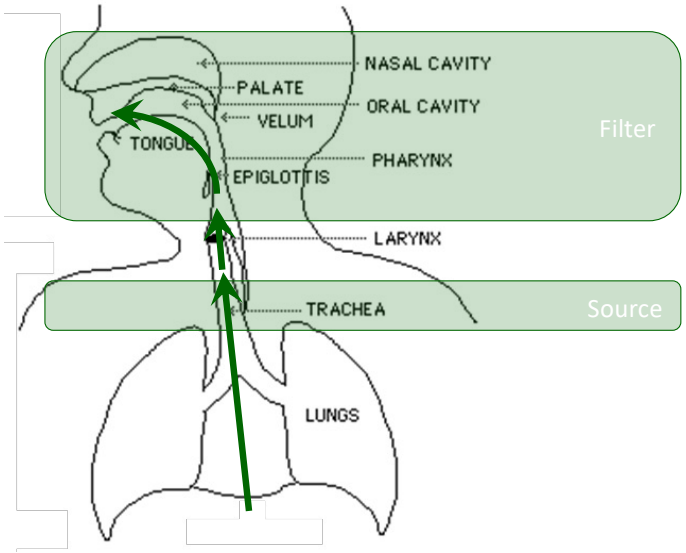
distortion of air

→ time-varying formant-frequency pattern

= speech

step 2: *phonation*

step 1: *initiation*



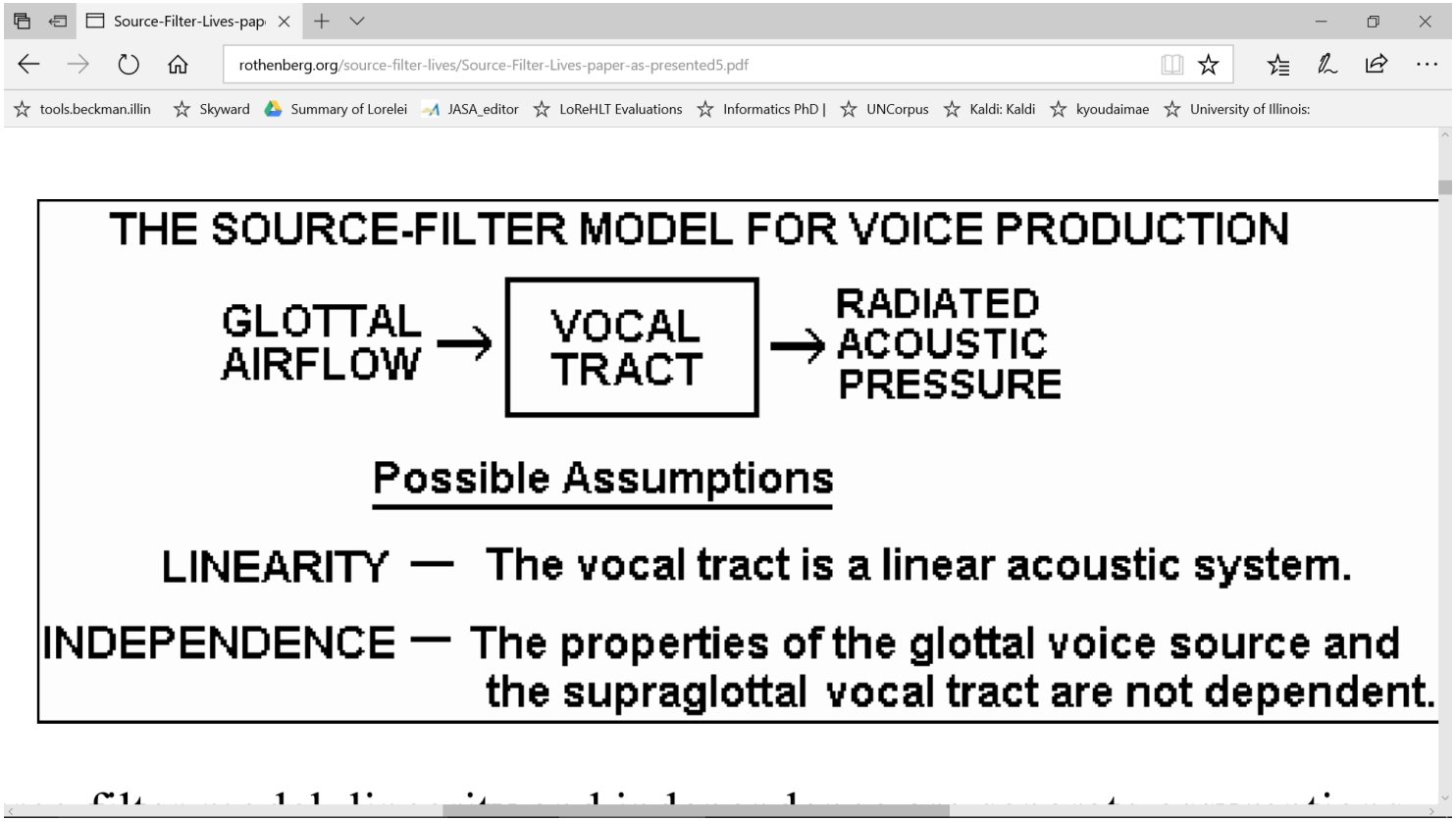
The Source-Filter Model of Speech Production
(Chiba & Kajiyama, 1940)

- 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

The Source-Filter Model of Speech Production

A picture from Martin Rothenberg's website



The Source-Filter Model

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

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

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

$$X(f) = H(f)E(f)$$

...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."

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):

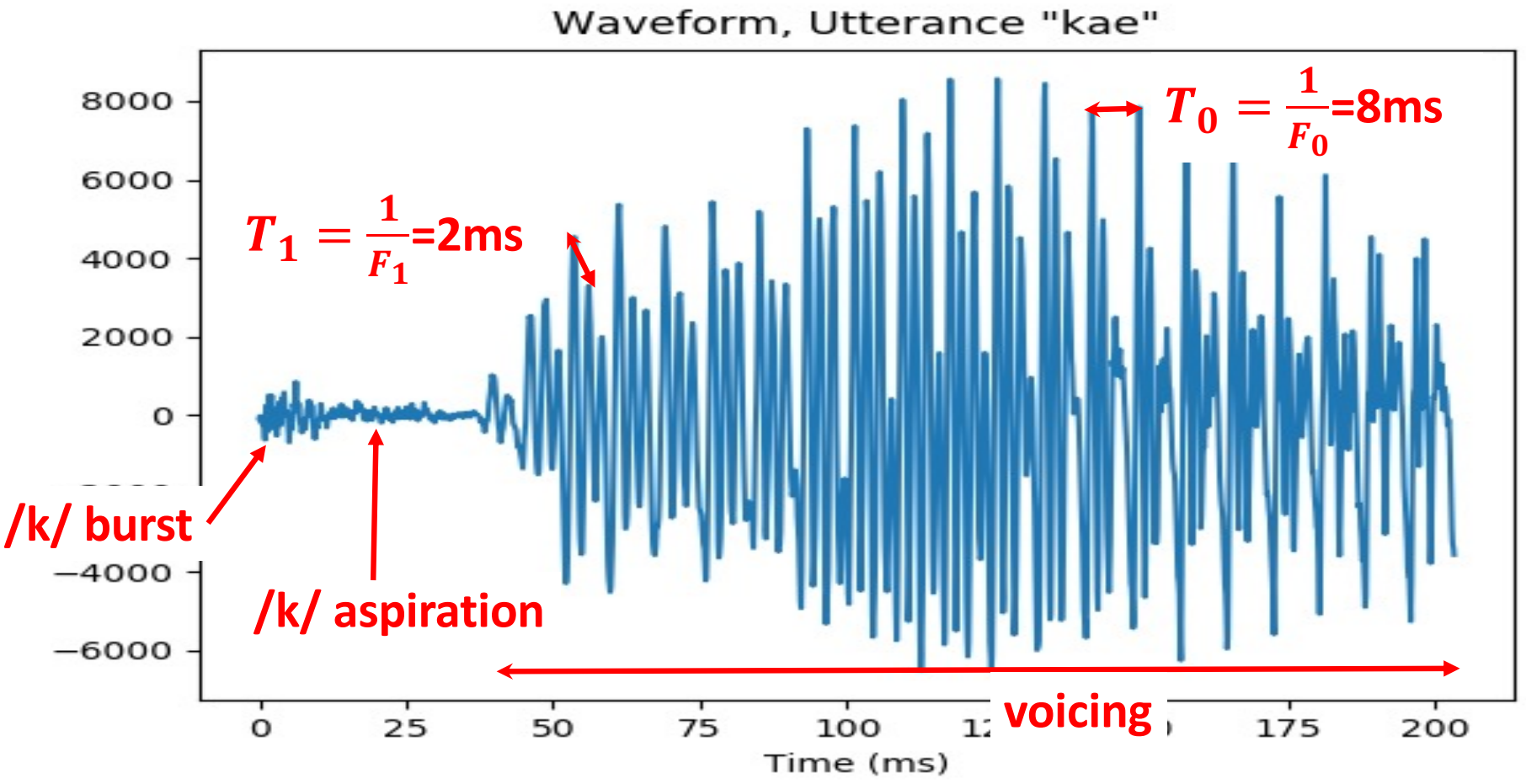
$$E(f) = \frac{1}{T_0} \sum_{k=-\infty}^{\infty} \delta(f - kF_0)$$

...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.

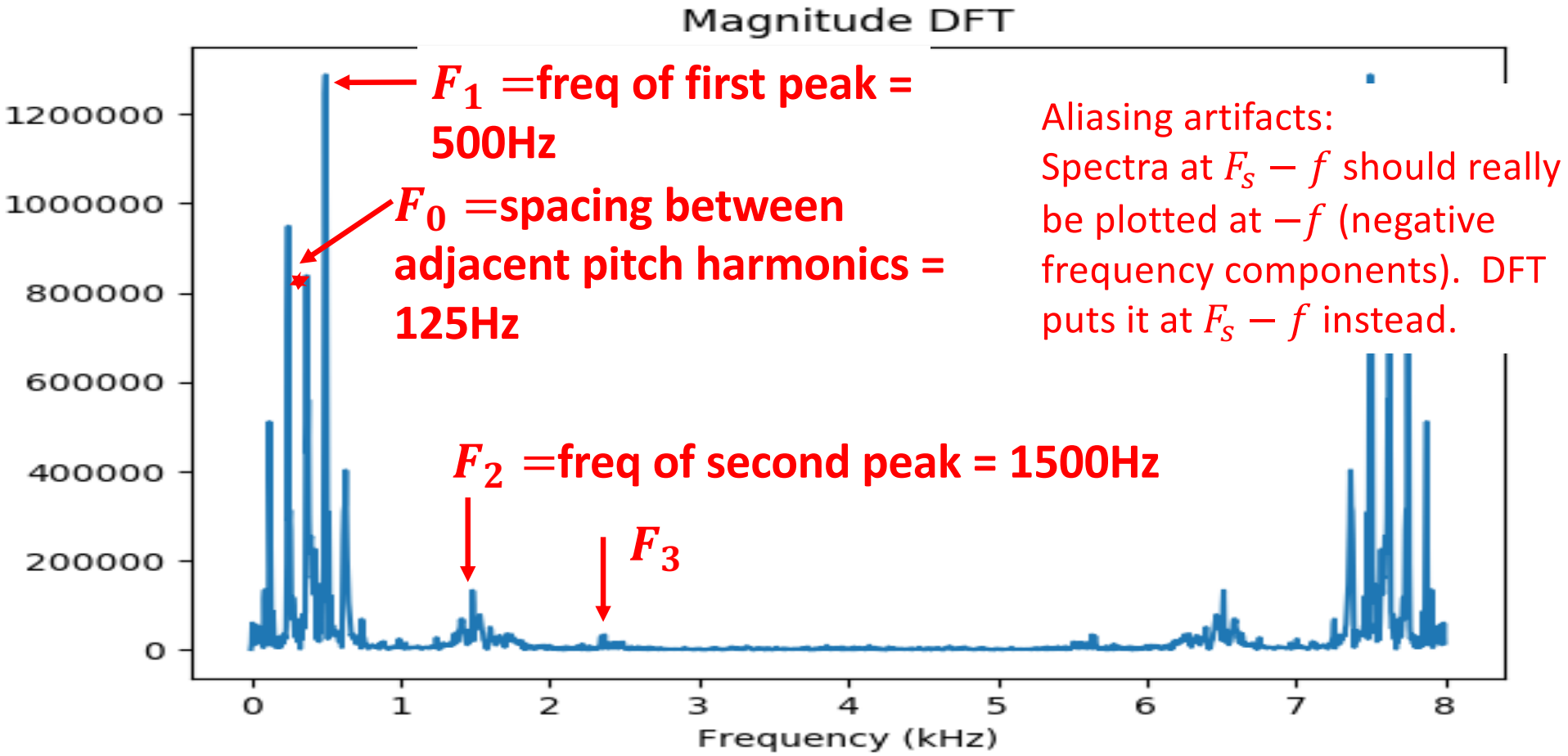
Source-Filter Model: Filter

- 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

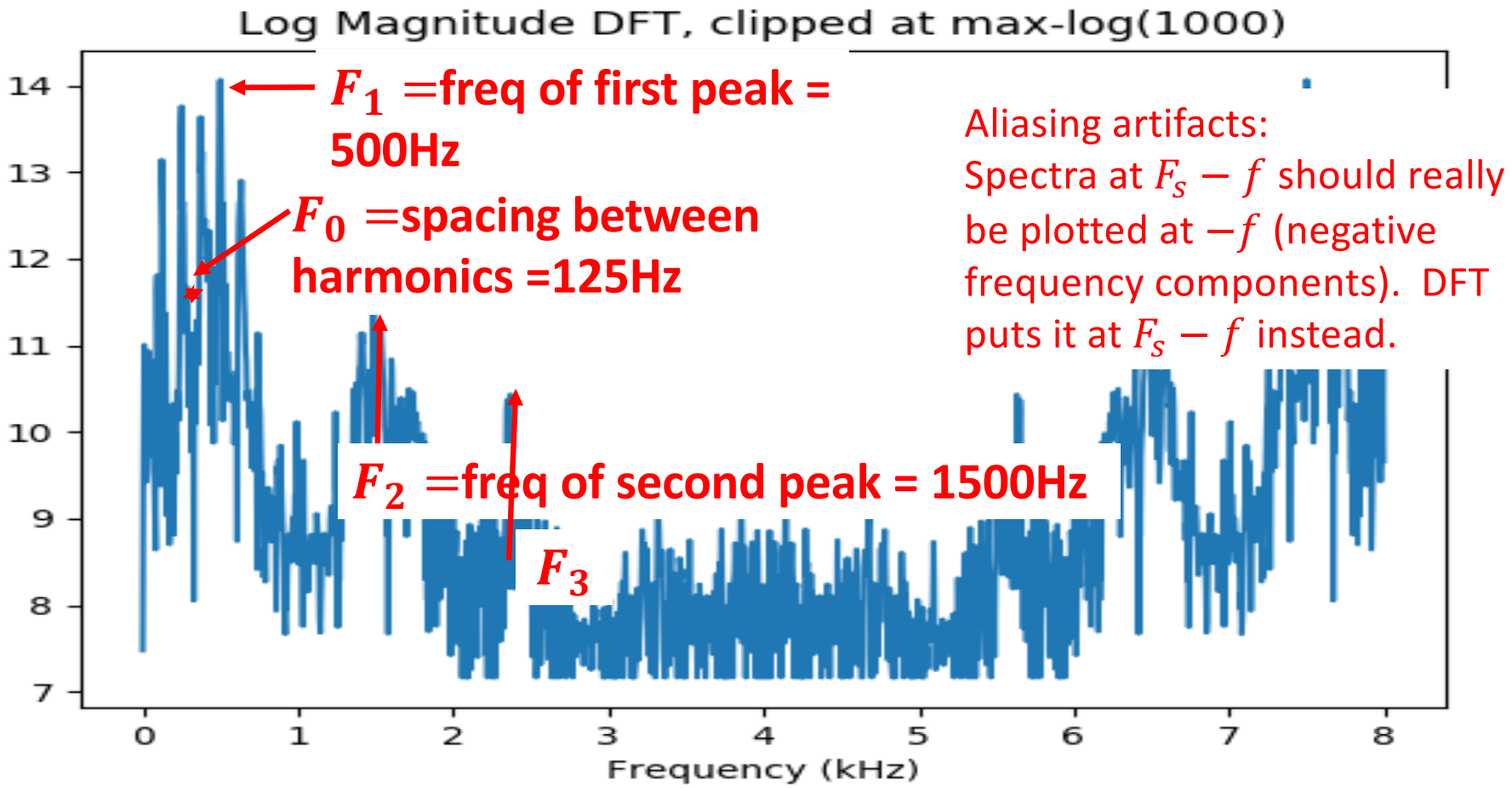


Speech signal: Magnitude Fourier Transform

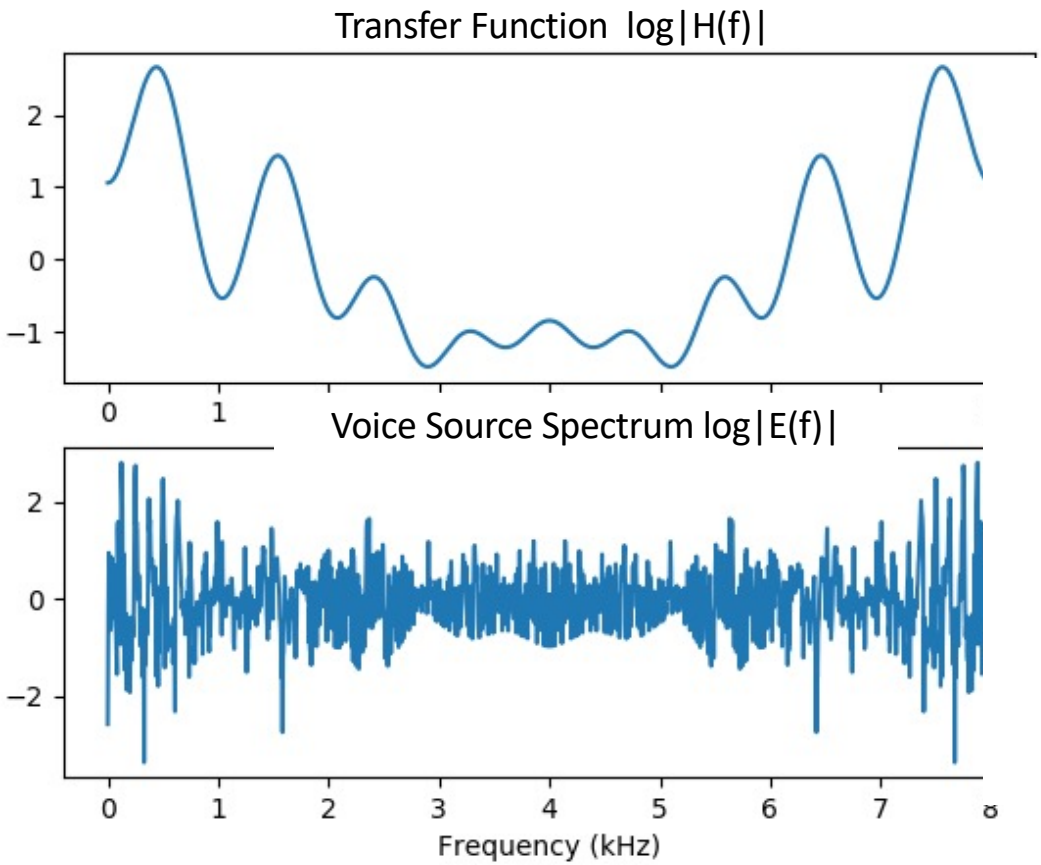


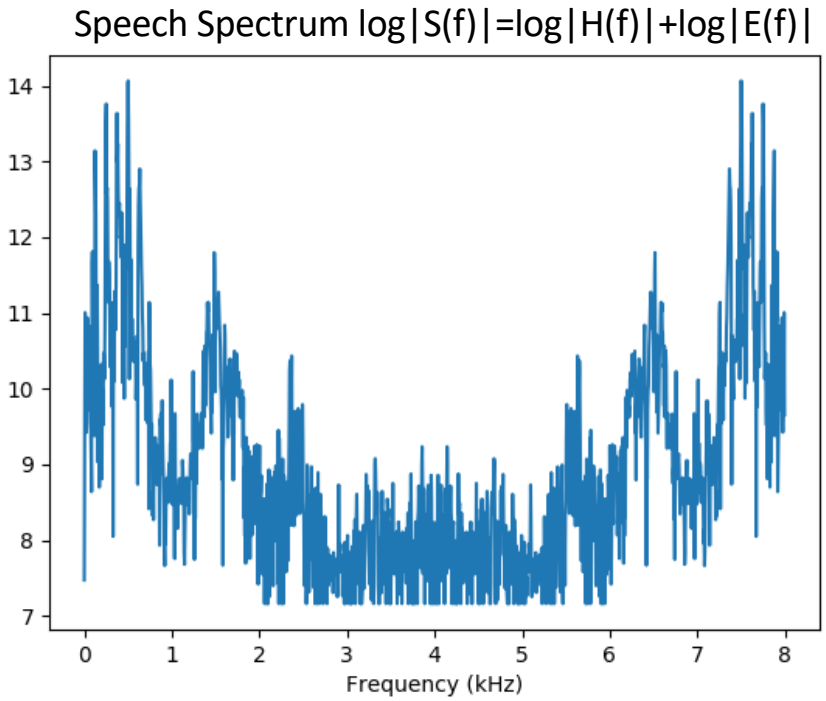
Aliasing artifacts:
Spectra at $F_s - f$ should really be plotted at $-f$ (negative frequency components). DFT puts it at $F_s - f$ instead.

Speech signal: Log Magnitude Transform



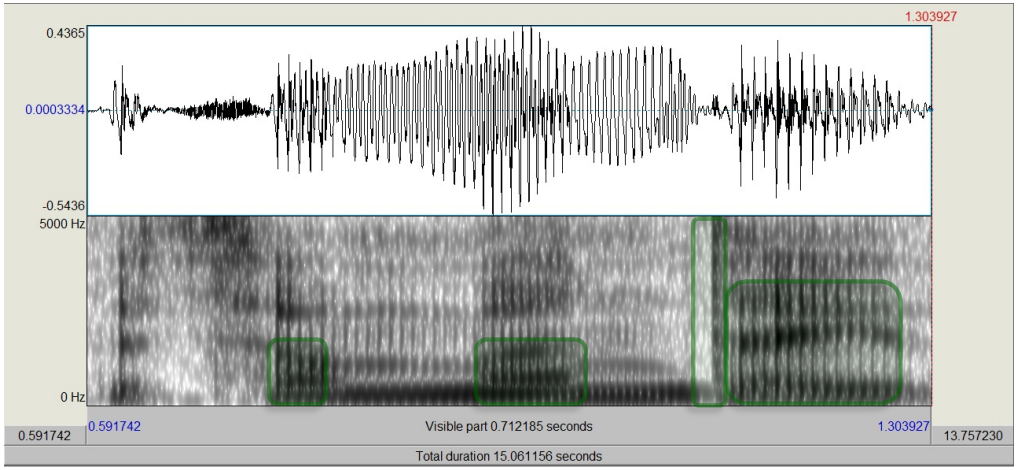
The Source-Filter Model





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.
Spectrogram = time on the horizontal axis, frequency on vertical axis.

Outline

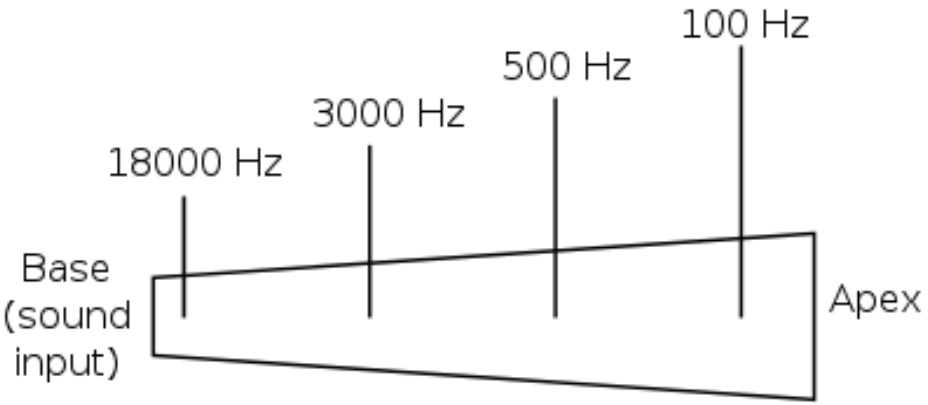
- Speech Production: Source-Filter Model
- Speech Perception: Spectrogram and Filterbank Coefficients
- Phonemes: Vowels and Consonants
- Automatic Speech Recognition
- Text-to-Speech Synthesis

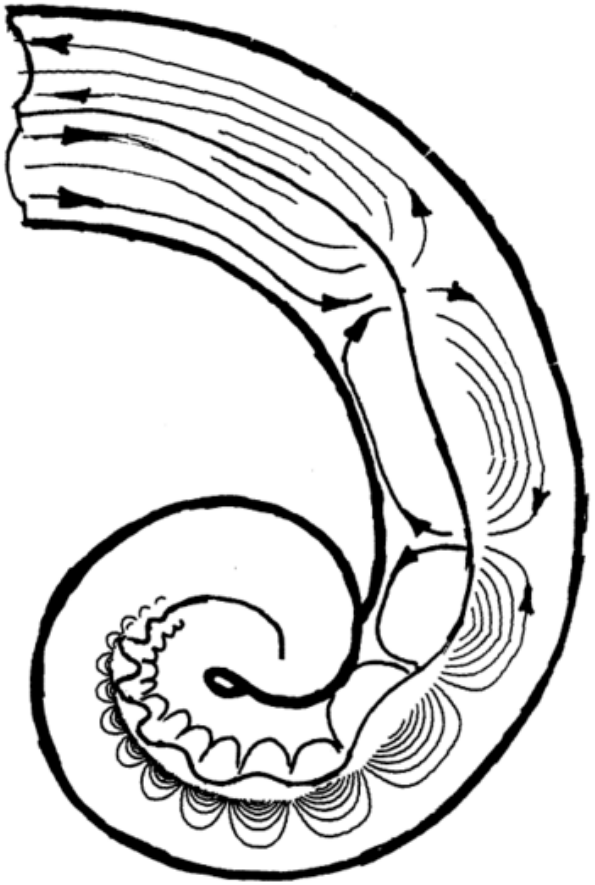
What spectrum do people hear? Basilar membrane

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





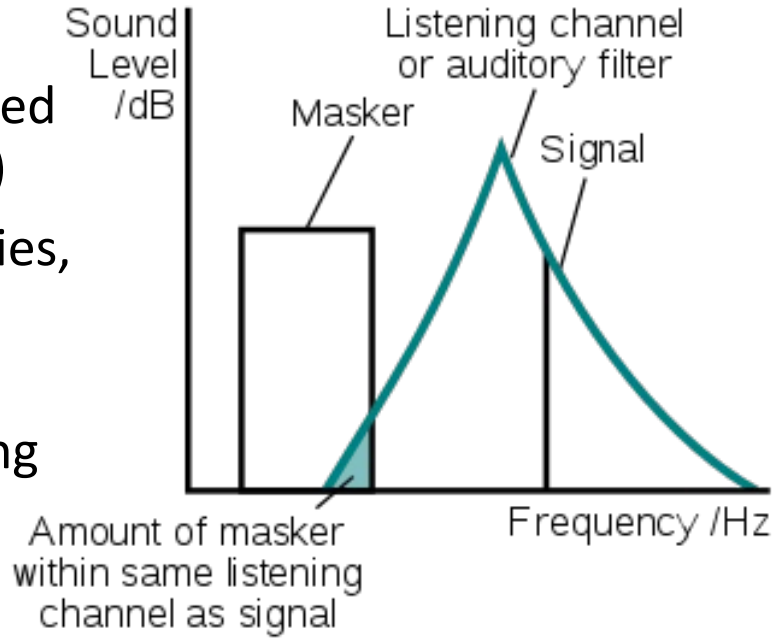
Frequency scales for hearing: mel scale, ERB scale

How is a spectrogram created? Short-Time Fourier Transform (STFT)

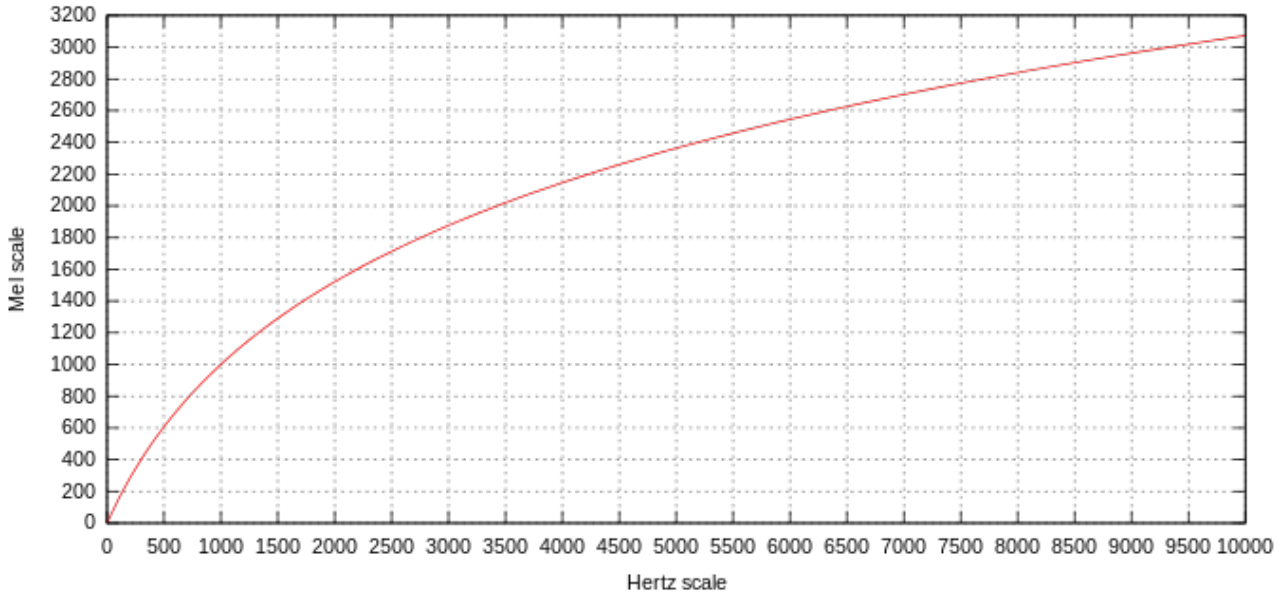
1. Chop up the speech signal, $x[n]$, into frames, $x_t[n]$
   - For example, 25ms duration, starting once every 10ms
   - t =frame index, n =sample index
2. Compute the Fourier transform of every frame
   - Calculate the energy in each frequency band, like human ear
   - $X_t[k] = \sum_{n=0}^{N-1} x_t[n] e^{-j2\pi kn/N}$
3. Optional: combine into mel-scale filterbank coefficients
4. Calculate the log magnitude in each frequency band
   - $S_t[k] = \log|X_t[k]|$
   - Simulate the ear's relative insensitivity to phase
   - Allows the neural net to have real-valued inputs, instead of complex

Critical bands

- When two tones play at exactly the same frequency, users can't tell the difference between x(t) versus x(t)+y(t) if y(t) is about 14dB below x(t) (in other words, the summed power is 1.03 times the power of x(t) alone)
- When x(t) and y(t) are at different frequencies, the masking power of x(t) is reduced
- Model: assume that the reduced masking power of x(t) is caused because x(t) is coming in through the tails of the bandpass filter centered at y(t).



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: $m(f) = \frac{1}{2595}\log_{10}\left(1 + \frac{f}{700}\right)$

ERB scale

- The experiment: find out the widths, B(f), of the critical-band filters centered at every frequency f.
- Analysis: create a scale $e(f)$ such that $e(f + 0.5B(f)) - e(f - 0.5B(f)) = 1$, for all frequencies
- Result: $\mathrm{e}(f) = 21.4 \log_{10}(1 + 0.00437f)$

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

- Goal: instead of computing

$$C_{t,k} = \ln|X_t[k]|$$

We want

$$C_{t,m} = \ln|S(f_m)|$$

Where the frequencies f_m 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.

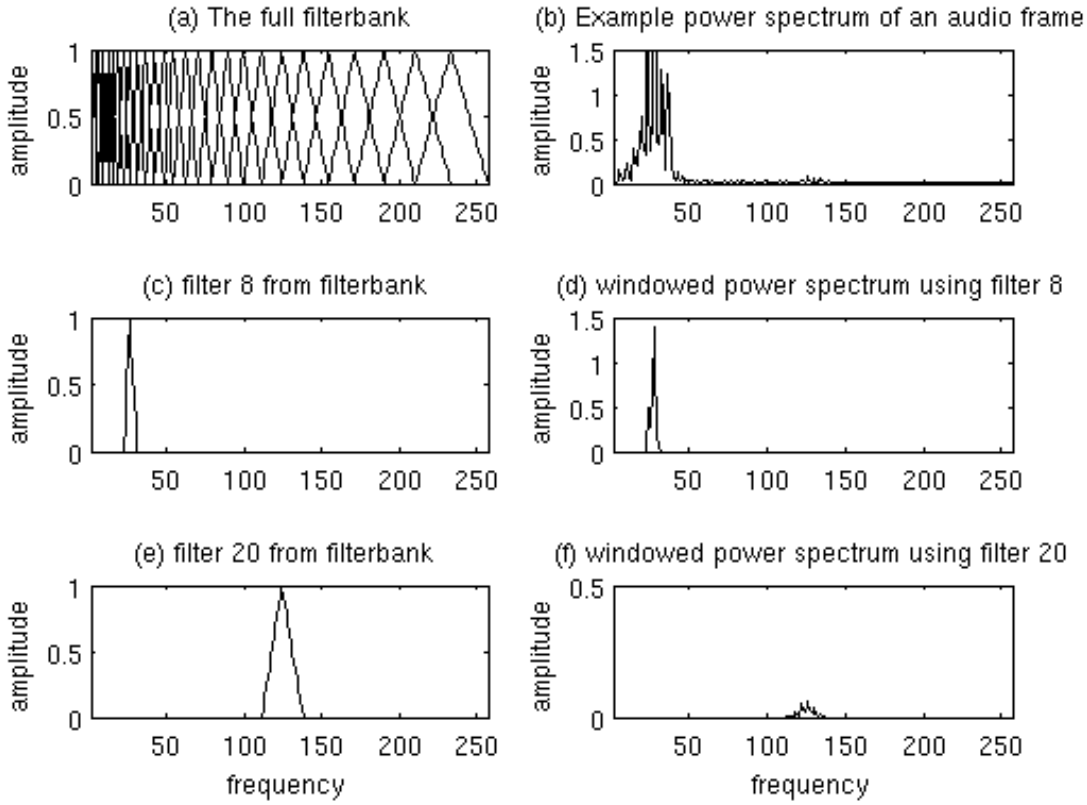
Mel filterbank coefficients: convert the spectrum from Hertz-frequency to mel-frequency

The solution:

$$C_{t,m} = \ln \sum_{k=0}^{\frac{N}{2}-1} W_m(k)|X_t[k]|$$

…where the $W_m(k)$ are filters approximately one ERB wide. They could be shaped to exactly match the shapes of human auditory filters, but usually we approximate that shape using triangular filters.

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



Outline

- Speech Production: Source-Filter Model
- Speech Perception: Filterbank Coefficients
- Phonemes: Vowels and Consonants
- Automatic Speech Recognition
- Text-to-Speech Synthesis

Linguistic units

Scharenborg, 2017

- Speech signal

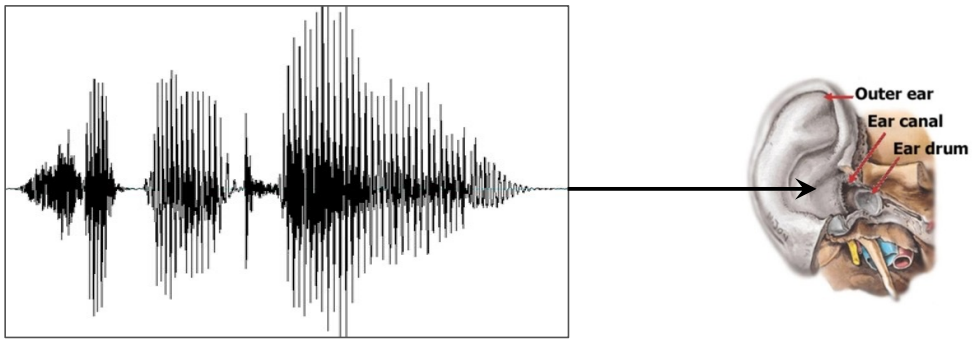
Linguistic units are:

- 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., *t*ip vs. *d*ip
- Allophone: a variation of a phoneme, eg., p^hot vs. *s*pot
- Phone: a distinct speech sound
- Word: the smallest distinct unit that can be uttered in isolation which has meaning

Speech sounds

Scharenborg, 2017

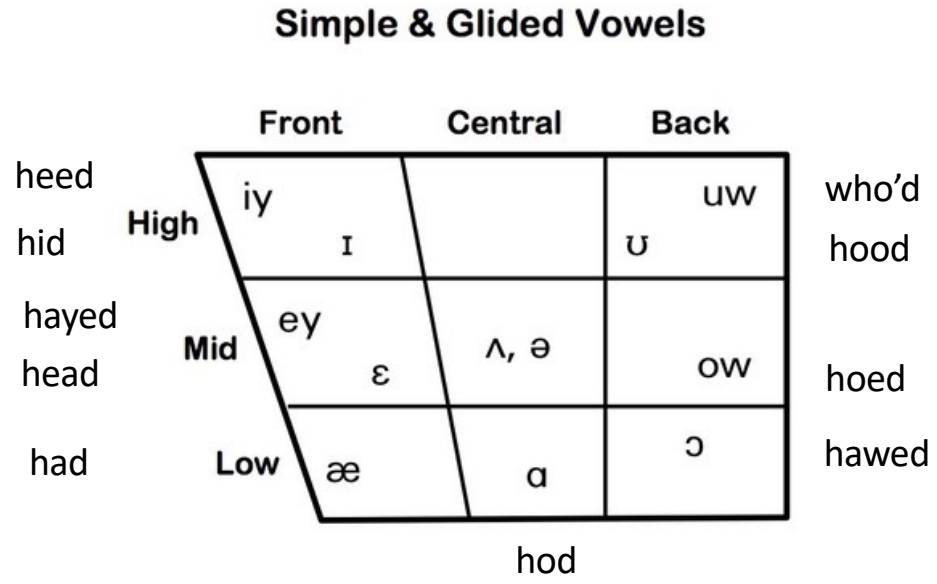
- Vowels: unblocked air stream

- Consonants: constricted or blocked air stream

Different sounds: Vowels

Scharenborg, 2017

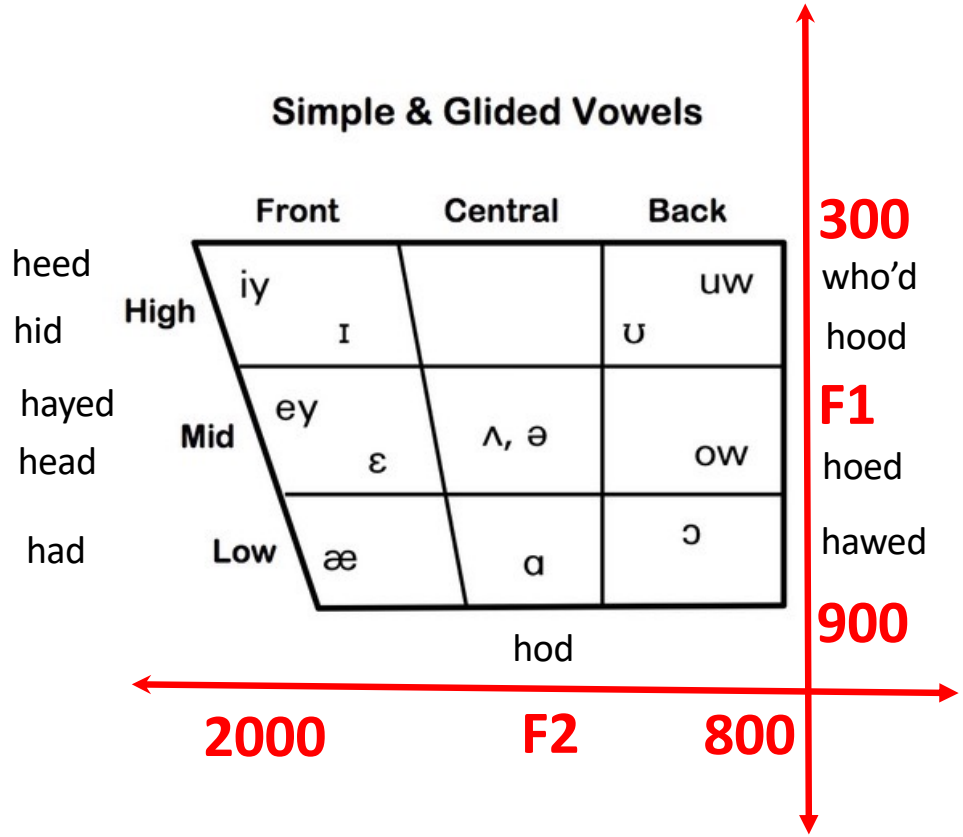
- Tongue height:
  - Low: e.g., /a/
  - Mid: e.g., /e/
  - High: e.g., /i/
- Tongue advancement:
  - Front : e.g., /i/
  - Central : e.g., /ə/
  - Back : e.g., /u/
- Lip rounding:
  - Unrounded: e.g., /ɪ, ɛ, e, ə/
  - Rounded: e.g., /u, o, ɔ/
- Tense/lax:
  - Tense: e.g., /i, e, u, o, ɔ, ɑ/
  - Lax: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



Different sounds: Vowels

Scharenborg, 2017

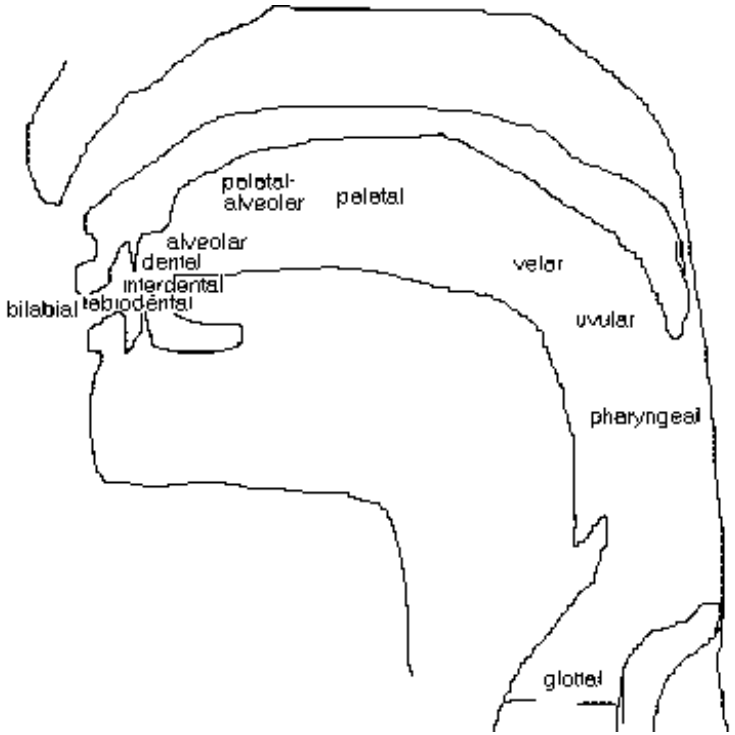
- Tongue height:
  - Low: e.g., /a/
  - Mid: e.g., /e/
  - High: e.g., /i/
- Tongue advancement:
  - Front : e.g., /i/
  - Central : e.g., /ə/
  - Back : e.g., /u/
- Lip rounding:
  - Unrounded: e.g., /ɪ, ɛ, e, ə/
  - Rounded: e.g., /u, o, ɔ/
- Tense/lax:
  - Tense: e.g., /i, e, u, o, ɔ, ɑ/
  - Lax: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

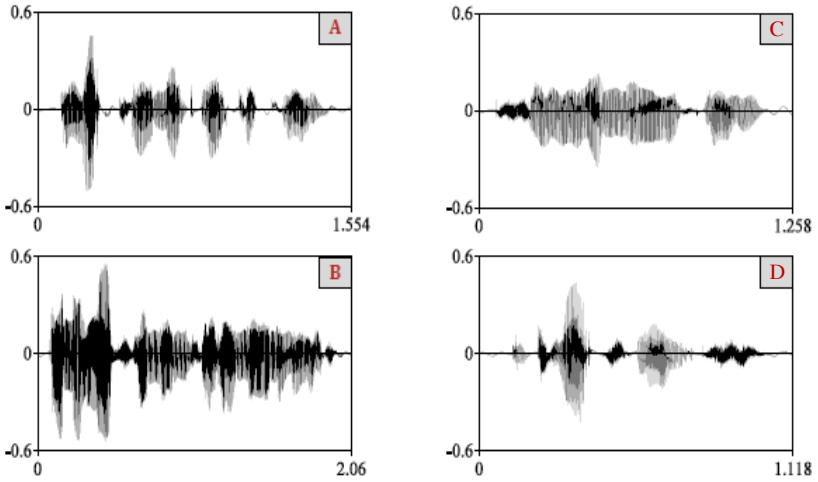
- https://www.youtube.com/watch?v=DcNMCB-Gsn8

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)

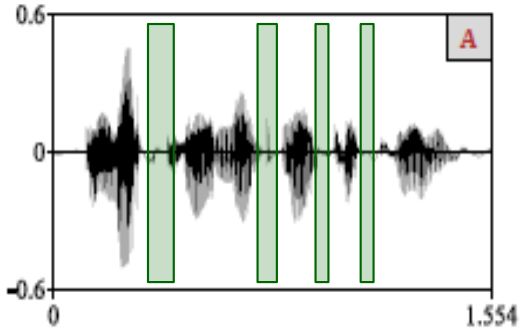
C: See you in the evening (5)

D: Attachment (1)

Electromagnetically

Scharenborg, 2017

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



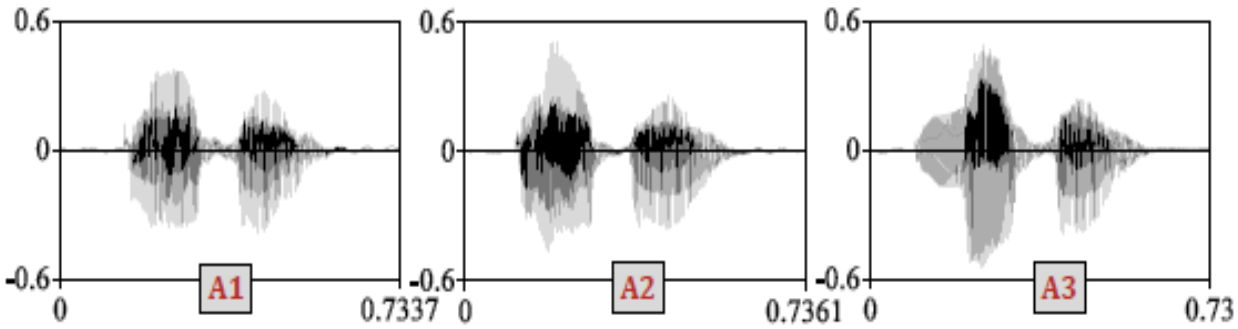
/i l ɛ kt romæ g nɛ t ɪ k ə l i/

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

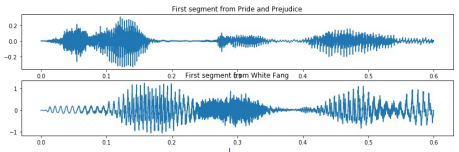
Outline

- Speech Production: Source-Filter Model
- Speech Perception: Filterbank Coefficients
- Phonemes: Vowels and Consonants
- Automatic Speech Recognition
- Text-to-Speech Synthesis

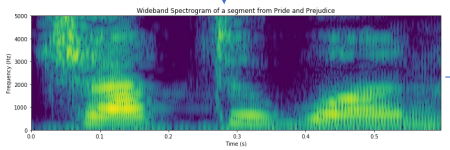
Speech Technologies

- speech → text or data
  - Automatic speech recognition (ASR: speech → text)
  - Speaker verification or identification
  - Emotion recognition
  - Speaker attribute recognition (drunk, sleepy, …)
  - Intent recognition (speech → meaning)
- text or data → speech
  - Text-to-speech (TTS) synthesis
  - Speech enhancement, source separation
  - Voice conversion
  - Image-to-speech (automatic spoken captioning of images)

ASR using spectrograms



STFT



LSTM, Transformer, or Conformer Neural Net

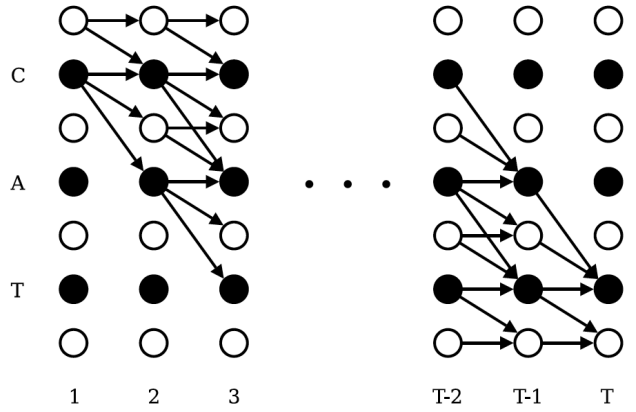
---cc--aaa-----t---

During training: NN loss calculated using an HMM (connectionist temporal classification)

How is the neural net trained? Connectionist Temporal Classification (CTC)

- Problem: the LSTM softmax layer outputs a vector of probabilities, once per 10ms:
$y_c[t] = P(\text{character} = c | x_t[n] \text{ and its LSTM context})$
- This is a problem because text has far fewer characters
- Solution: use an HMM to convert the LSTM output probabilities into the total probability of the correct transcript ℓ, then train the LSTM to maximize that probability

$$\mathcal{L} = -\ln P(\ell | x[n]) = -\ln \sum_{\ell = HMM(\pi)} \prod_{t=1}^{T} y_{\pi_t}[t]$$



HMM trellis used to discover that the neural network output:

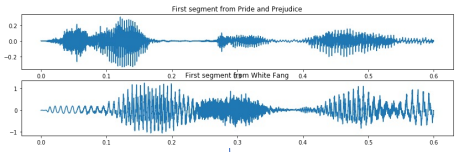
π = ---cc--aaa-----t---

…is a valid spelling of the word:

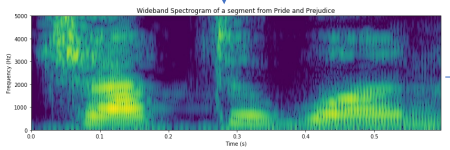
ℓ = cat

Graves et al., ICML 2006, "Connectionist Temporal Classification…"

ASR using 1-D Convolutional Neural Net



1D CNN with trainable filters instead of the STFT



LSTM, Transformer, or Conformer Neural Net

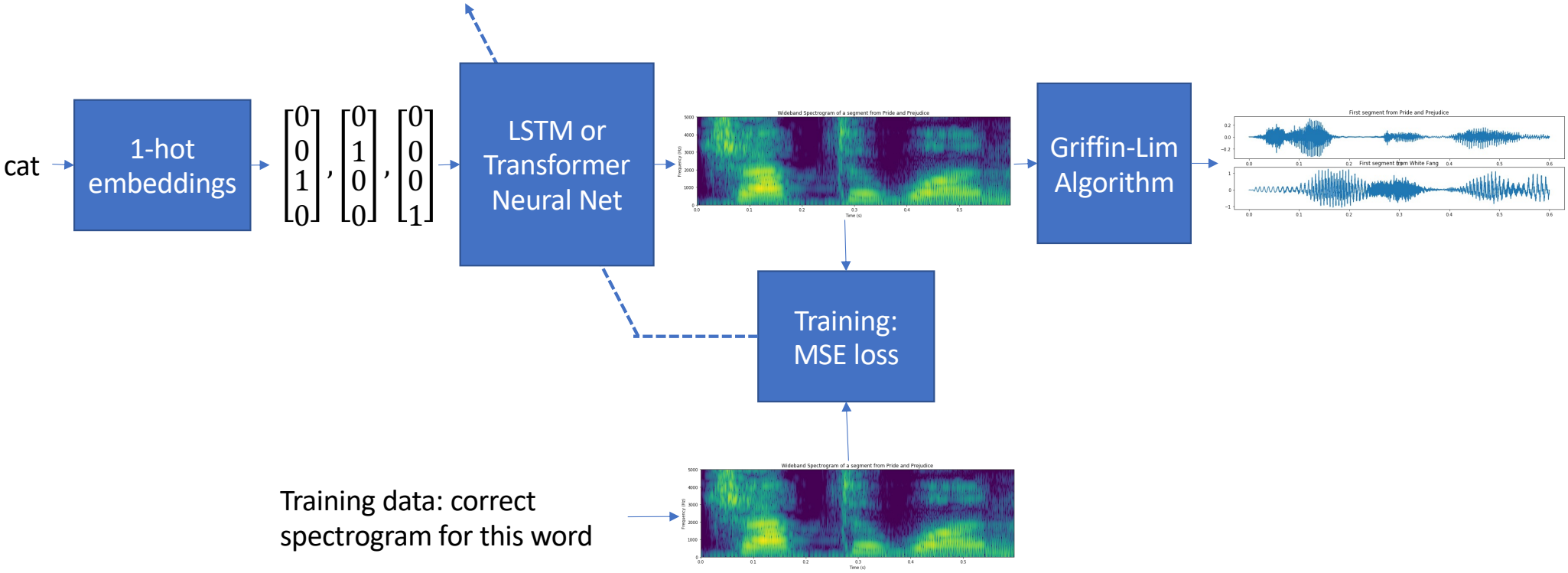
---cc--aaa-----t---

During training: NN loss calculated using an HMM (connectionist temporal classification)

Outline

- Speech Production: Source-Filter Model
- Speech Perception: Filterbank Coefficients
- Phonemes: Vowels and Consonants
- Automatic Speech Recognition
- Text-to-Speech Synthesis

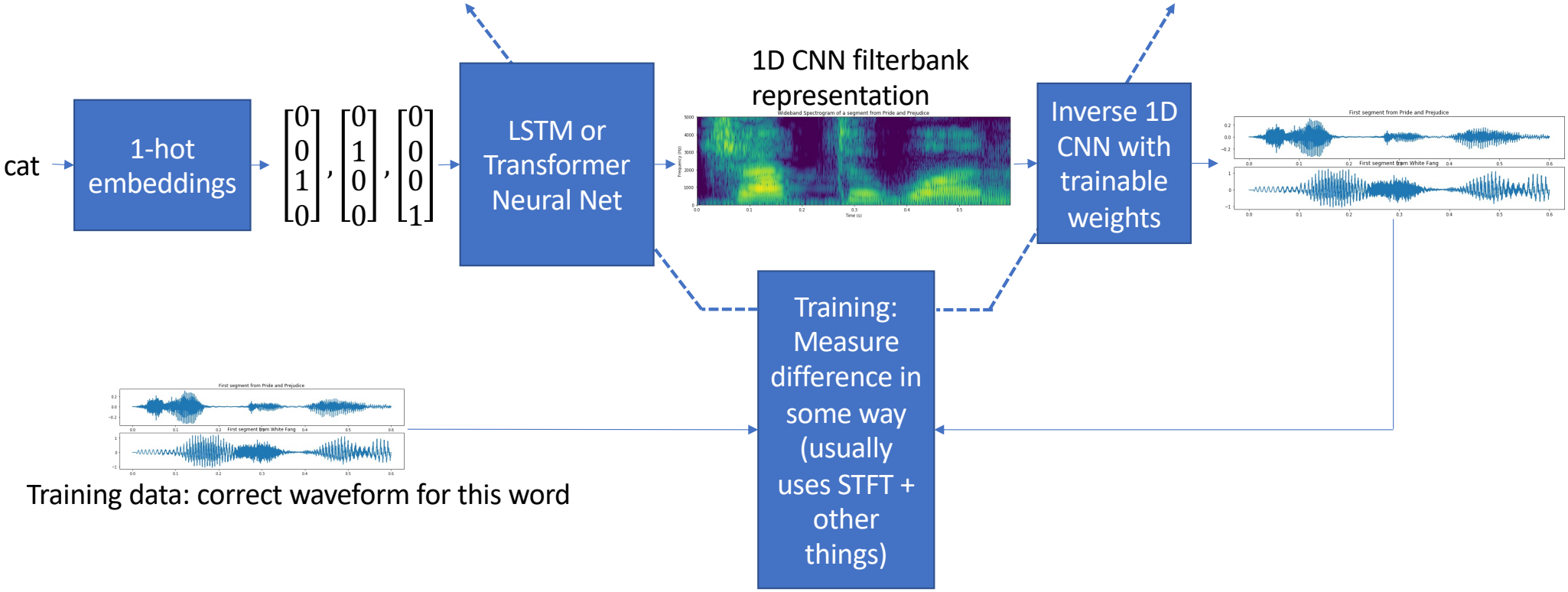
TTS using spectrograms



Griffin-Lim Algorithm

- Problem: synthesized spectrogram might not have a valid inverse STFT
- Solution: Among all complex $X_t[k]$ that have the desired $A_t[k] = |X_t[k]|$, choose the one that is closest to a valid STFT
- Resulting algorithm: choose an initial $X_t^{[0]} = A_t[k]e^{j\varphi_{t,k}}$ with random phases $\varphi_{t,k}$, then iterate the following two steps:
  1. Modify it so it has a valid inverse STFT: $\tilde{X}_t^{[m]}[k] = GG^{\dagger}X_t^{[m]}[k]$, where G is a matrix that computes the vectorized STFT of a waveform, and $G^{\dagger}$ is its pseudo-inverse
  2. Modify it so it has the right amplitudes: $X_t^{[m+1]}[k] = A_t[k]\frac{\tilde{X}_t^{[m]}[k]}{\left|\tilde{X}_t^{[m]}[k]\right|}$

TTS using 1-D Convolutions



Outline

- Speech Production: Source-Filter Model

$$X(f) = H(f)E(f)$$

- Speech Perception: STFT

$$X_t[k] = \sum_{n=0}^{N-1} x_t[n] e^{-j2\pi kn/N}$$

- Phonemes: Vowels and Consonants
- Automatic Speech Recognition: CTC

$$\mathcal{L} = -\ln P(\ell|x[n]) = -\ln \sum_{\ell=HMM(\pi)} \prod_{t=1}^{T} y_{\pi_t}[t]$$

- Text-to-Speech Synthesis: Griffin-Lim

$$\tilde{X}_t^{[m]}[k] = GG^{\dagger} X_t^{[m]}[k],\ X_t^{[m+1]}[k] = A_t[k] \frac{\tilde{X}_t^{[m]}[k]}{\left|\tilde{X}_t^{[m]}[k]\right|}$$