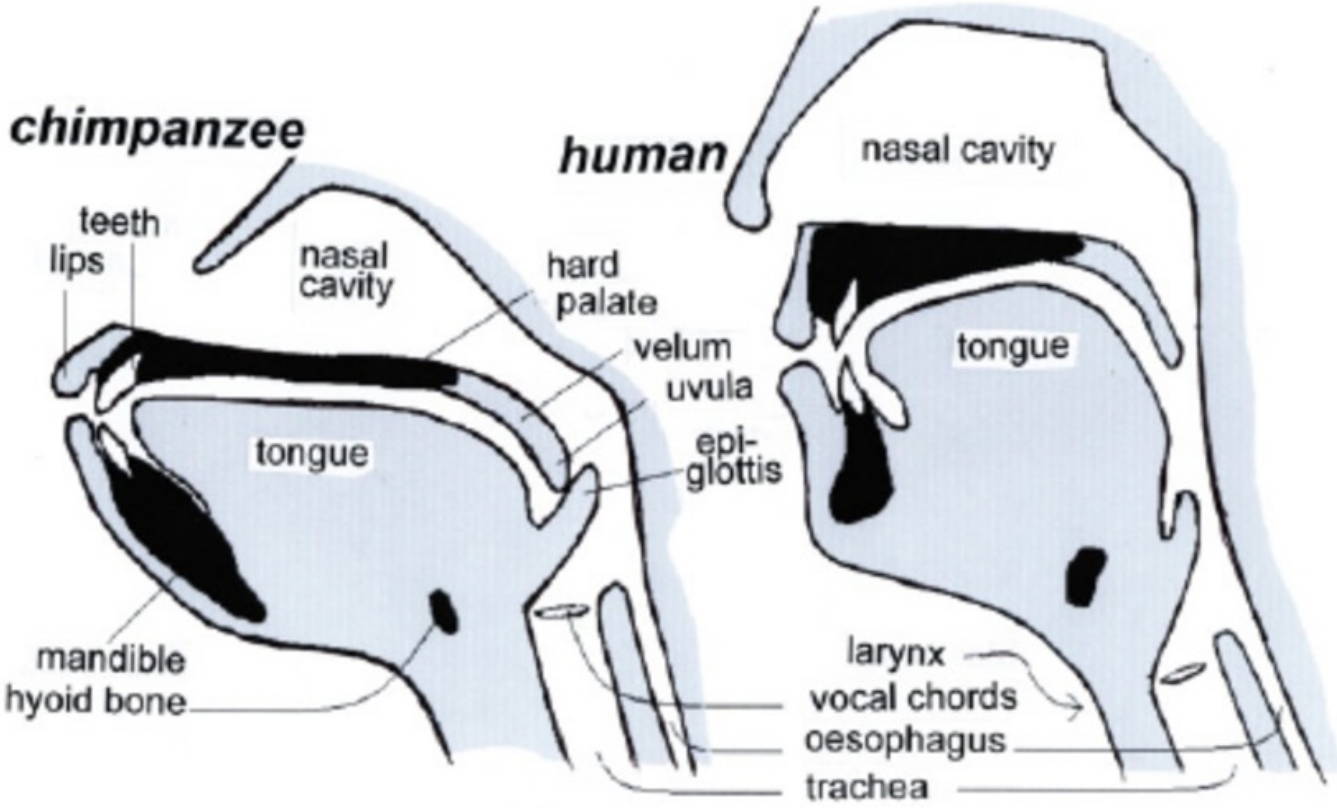
Speech Production and Perception ECE 448, Spring 2023

Mark Hasegawa-Johnson, 5/2023



Speech

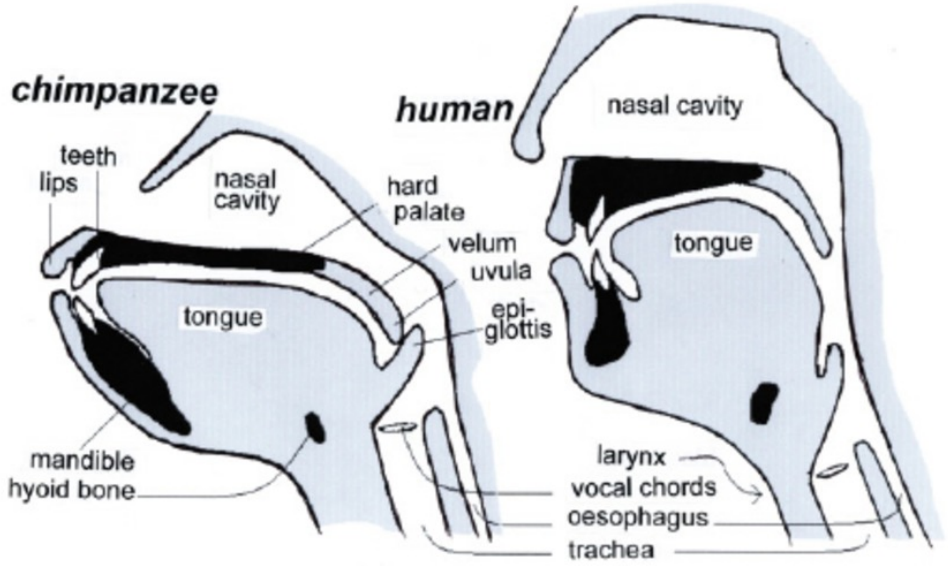
(Slide: Scharenborg, 2017)

- 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

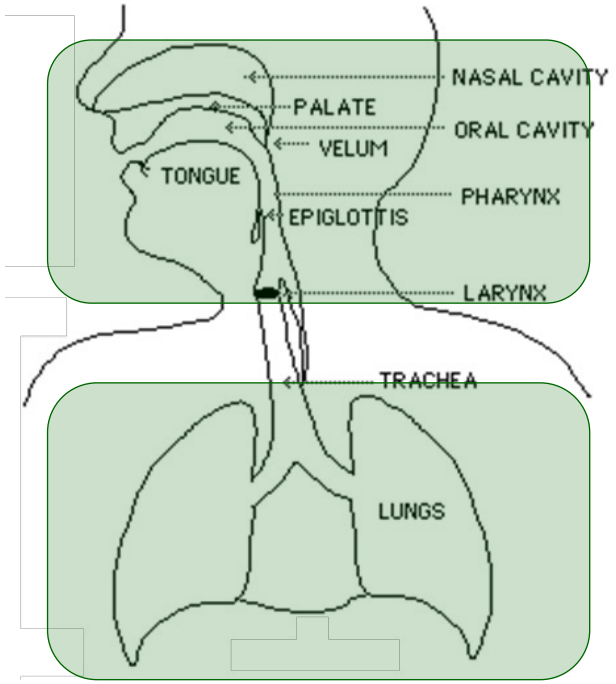
The anatomy and physiology of speech

(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* =

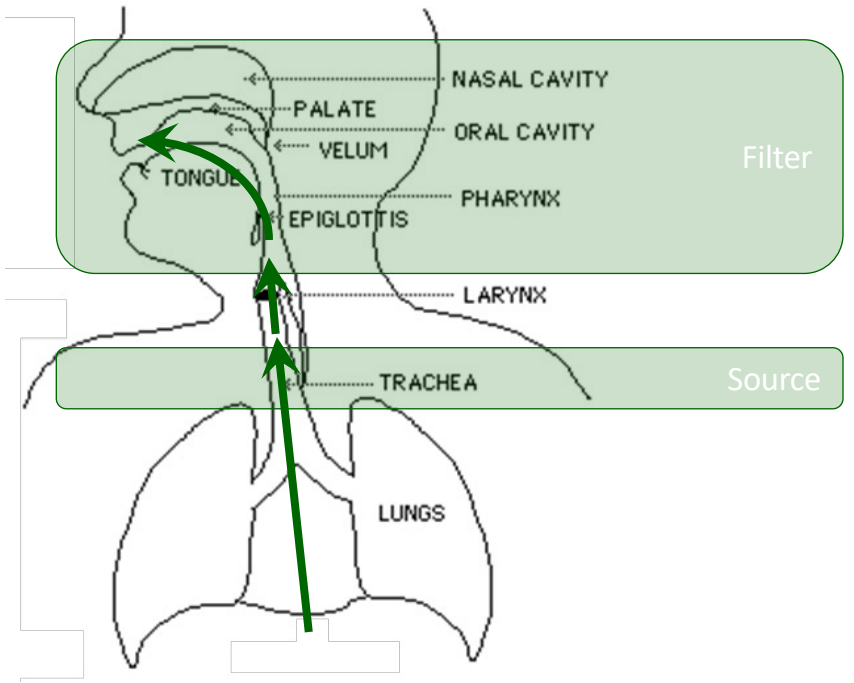
distortion of air

→ 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* =

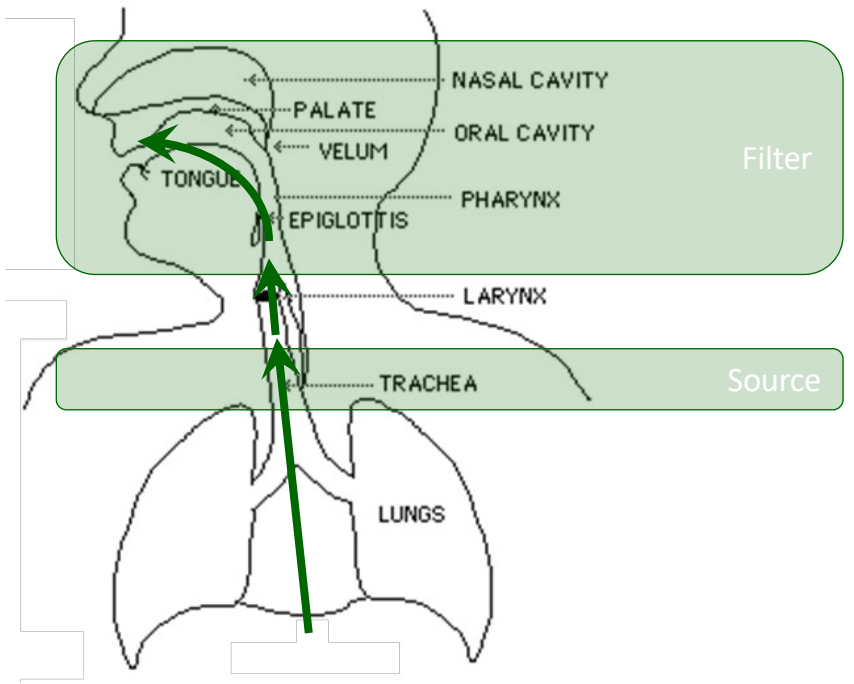
distortion of air

→ time-varying formant-frequency pattern

= speech

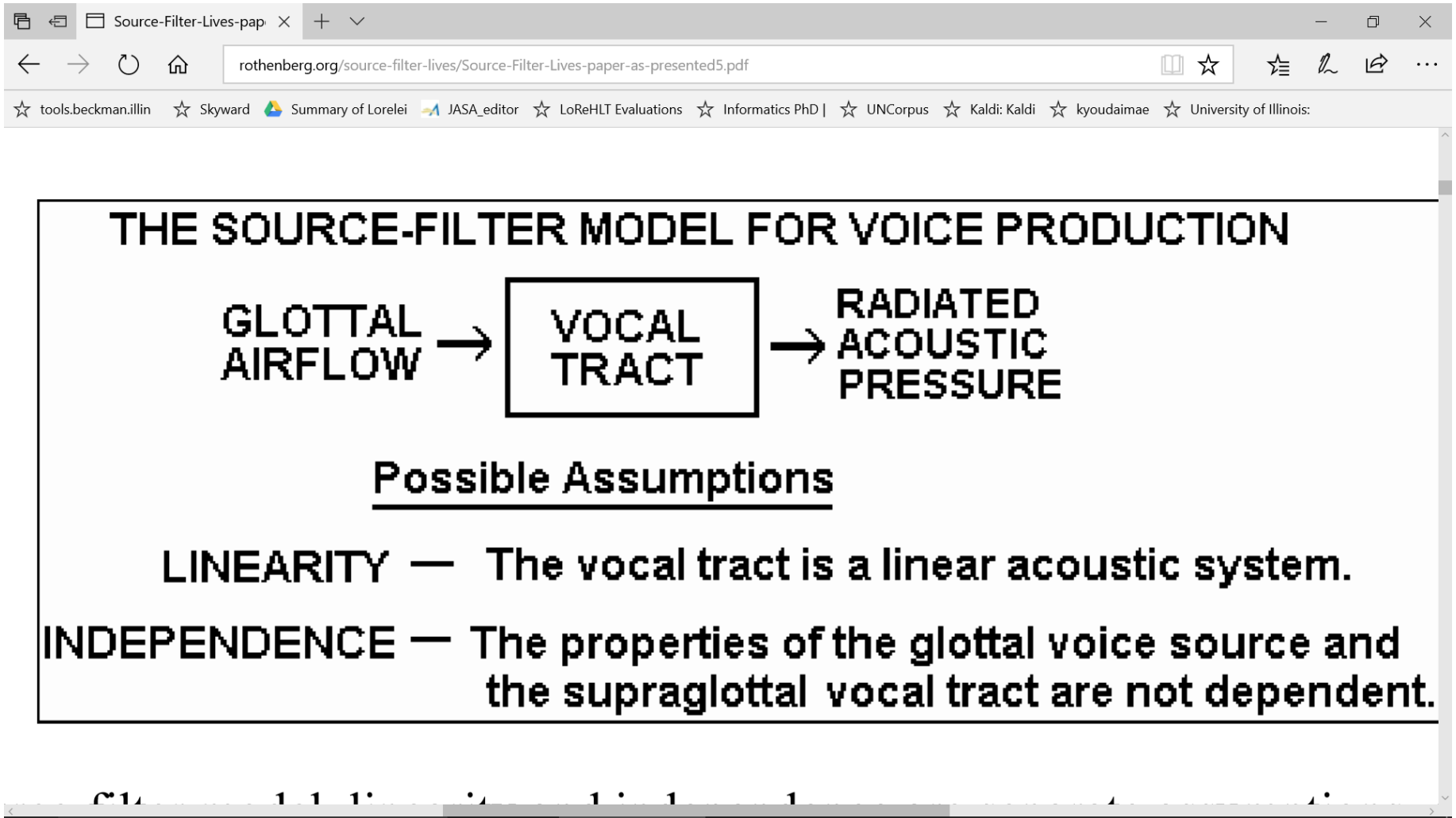
step 2: *phonation*

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 $h(t)$

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

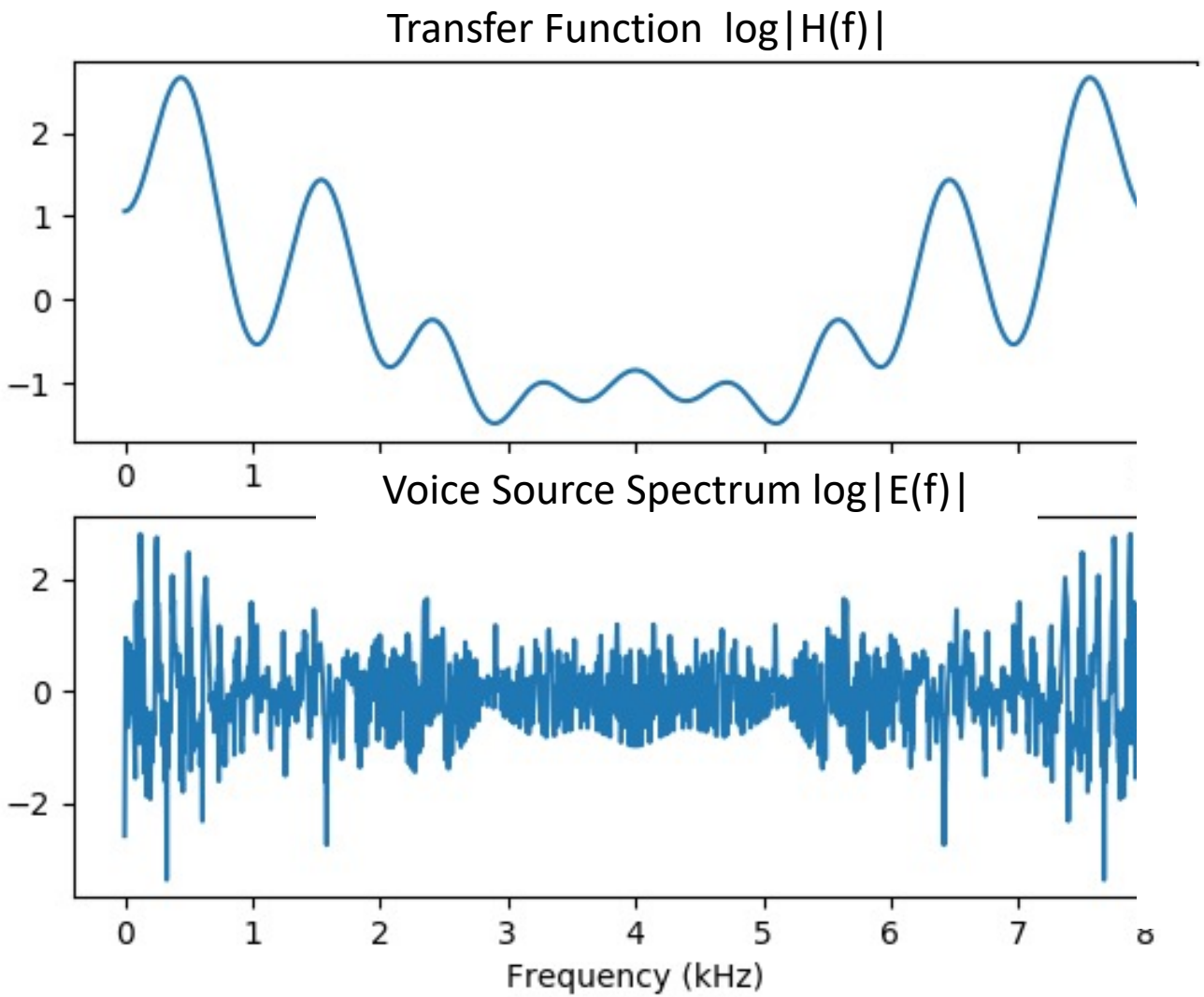
- The Fourier transform of speech is therefore the product of excitation times transfer function:

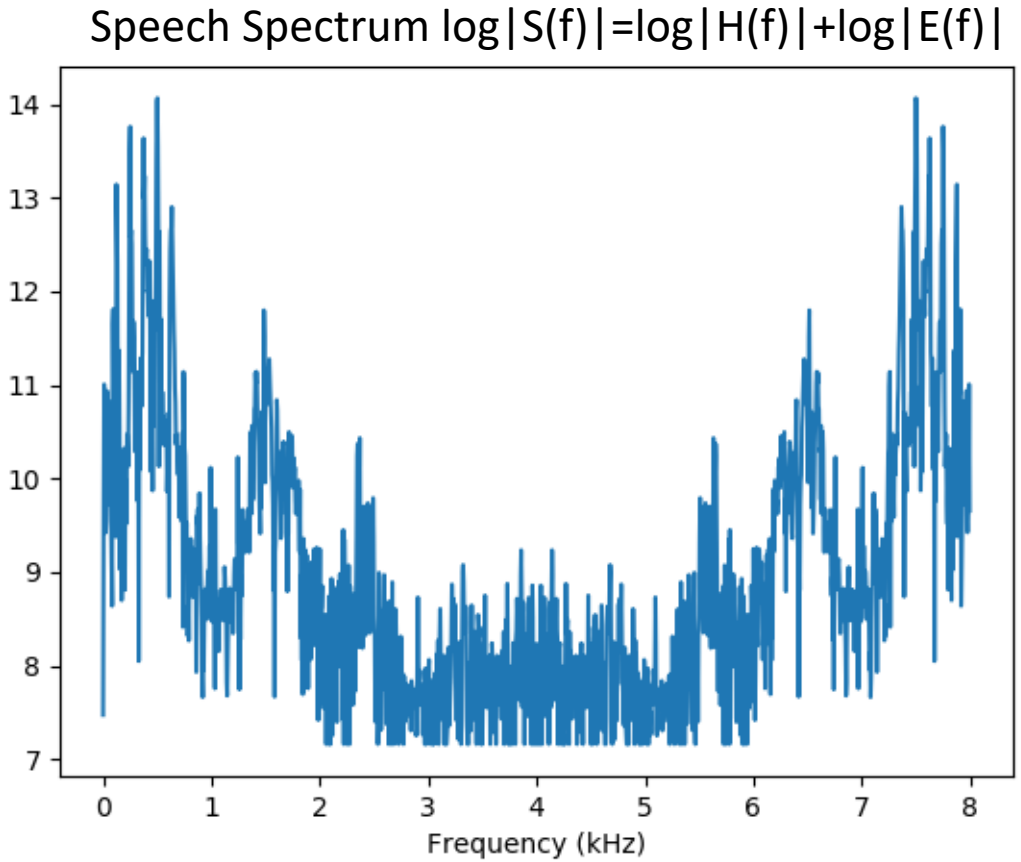
$$S(f) = H(f)E(f)$$

...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):

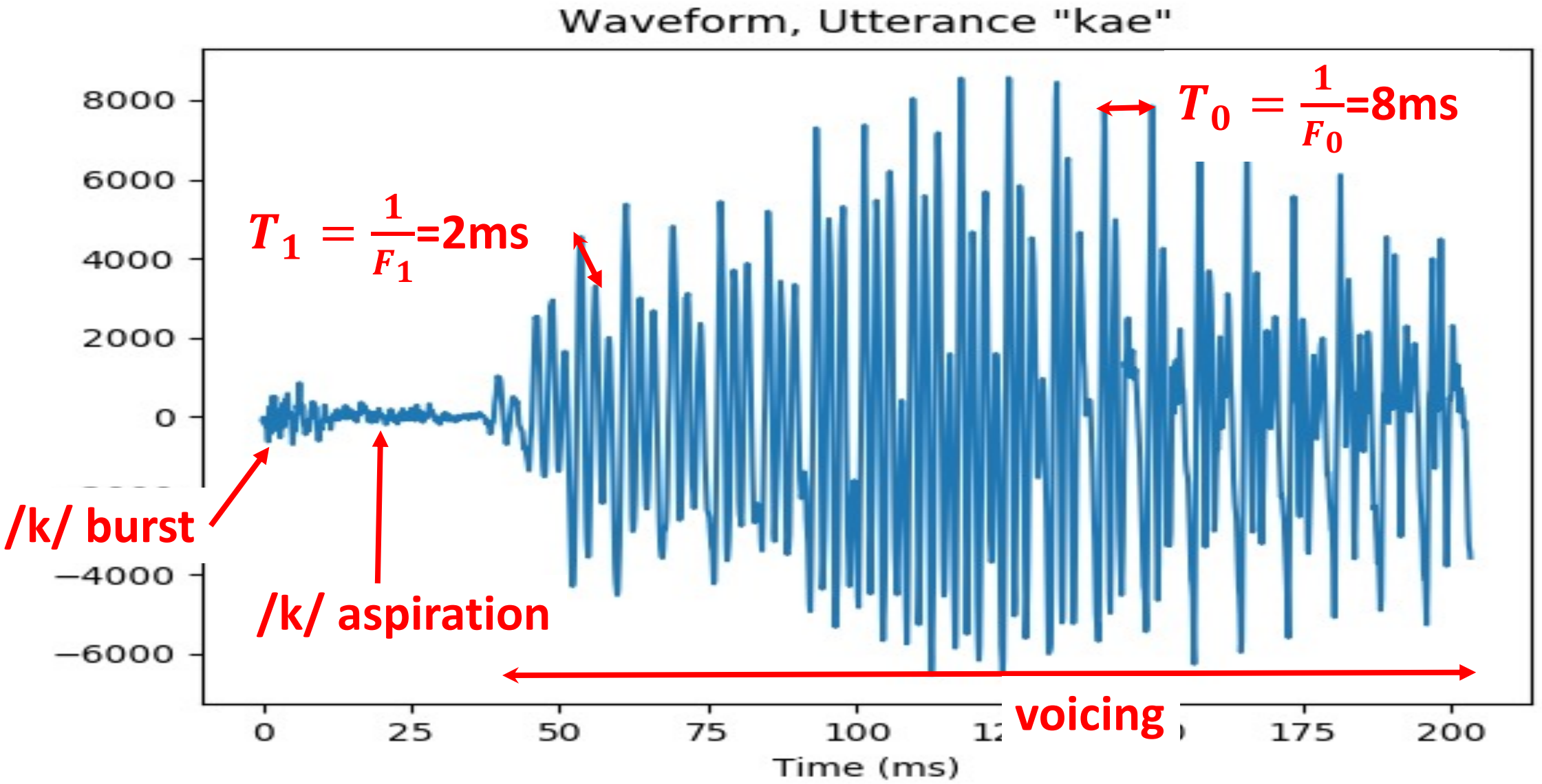
$$E(f) = \frac{1}{T_0} \sum_{k=-\infty}^{\infty} \delta(f - kF_0)$$

...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.

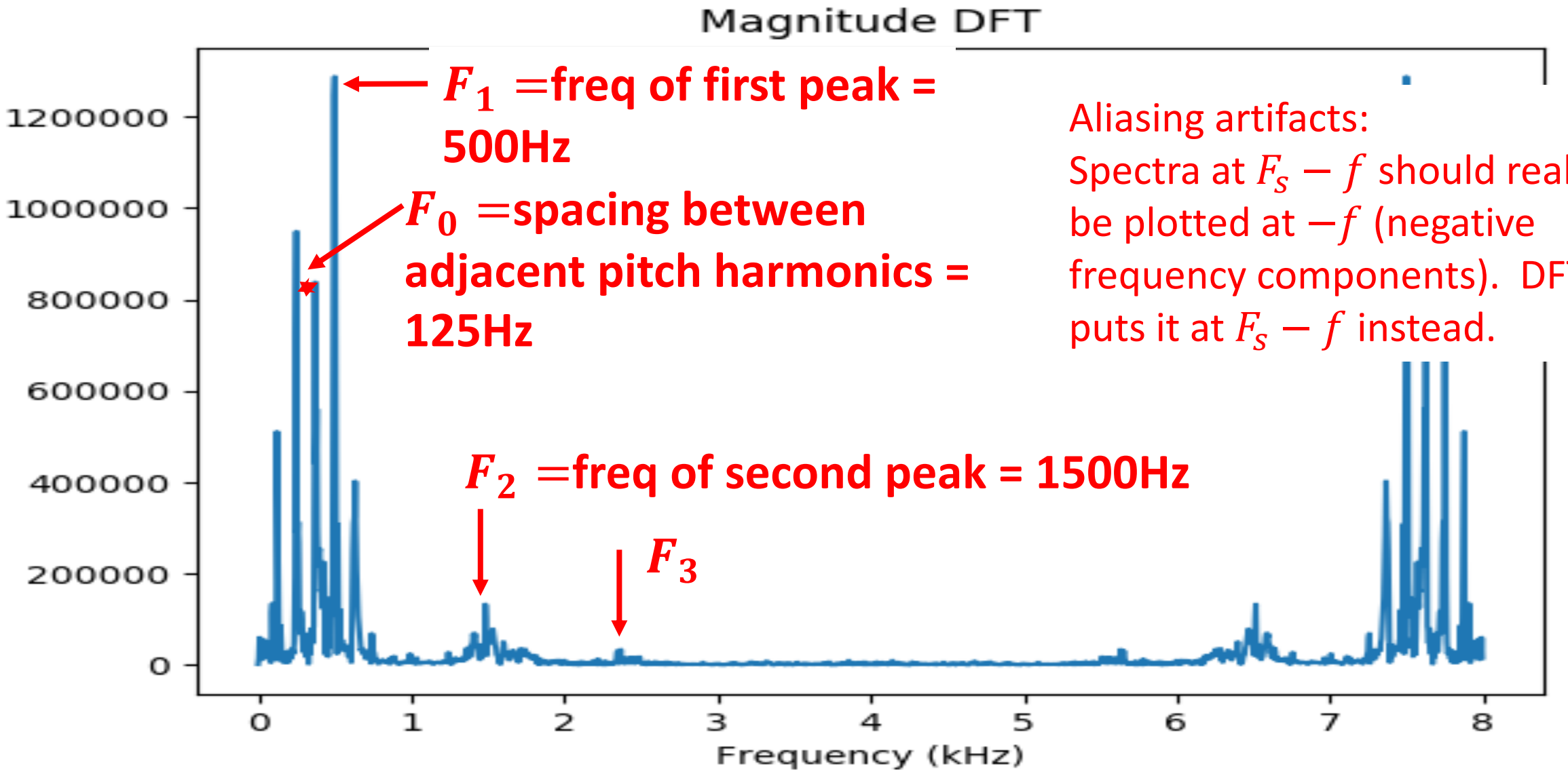
Source-Filter Model: Filter

- 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

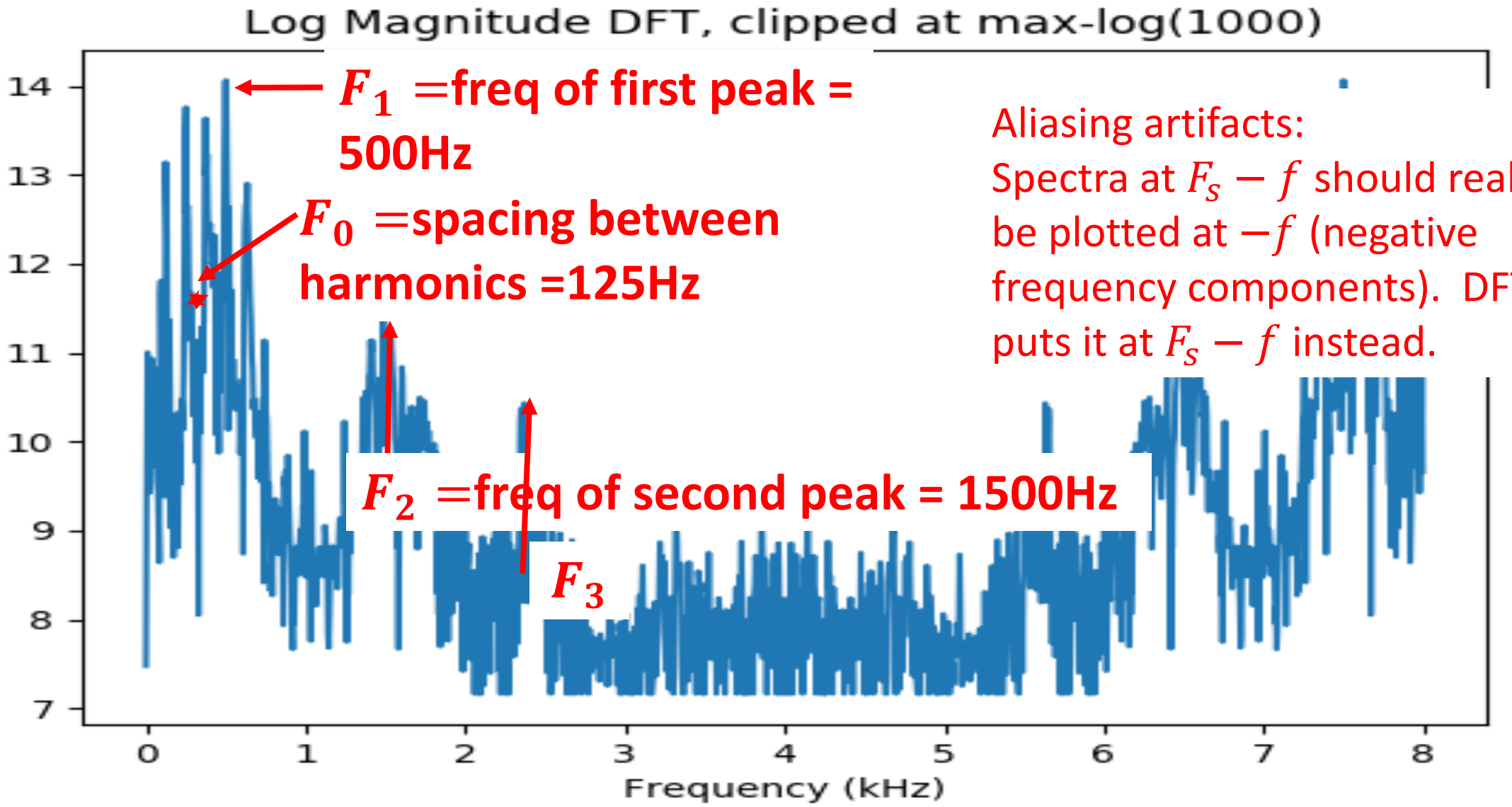


Speech signal: Magnitude Fourier Transform



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



Part 2: Linguistic units

Scharenborg, 2017

- Speech signal

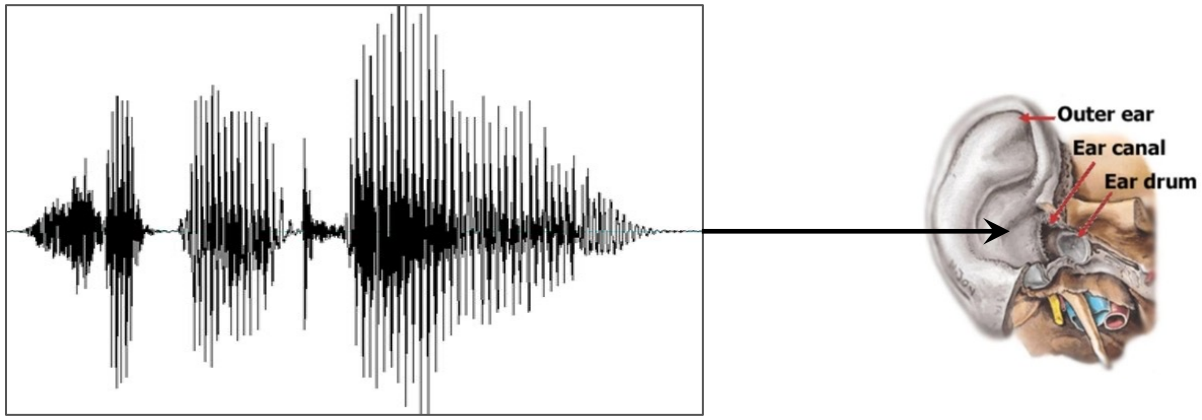
Linguistic units are:

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

Speech sounds

Scharenborg, 2017

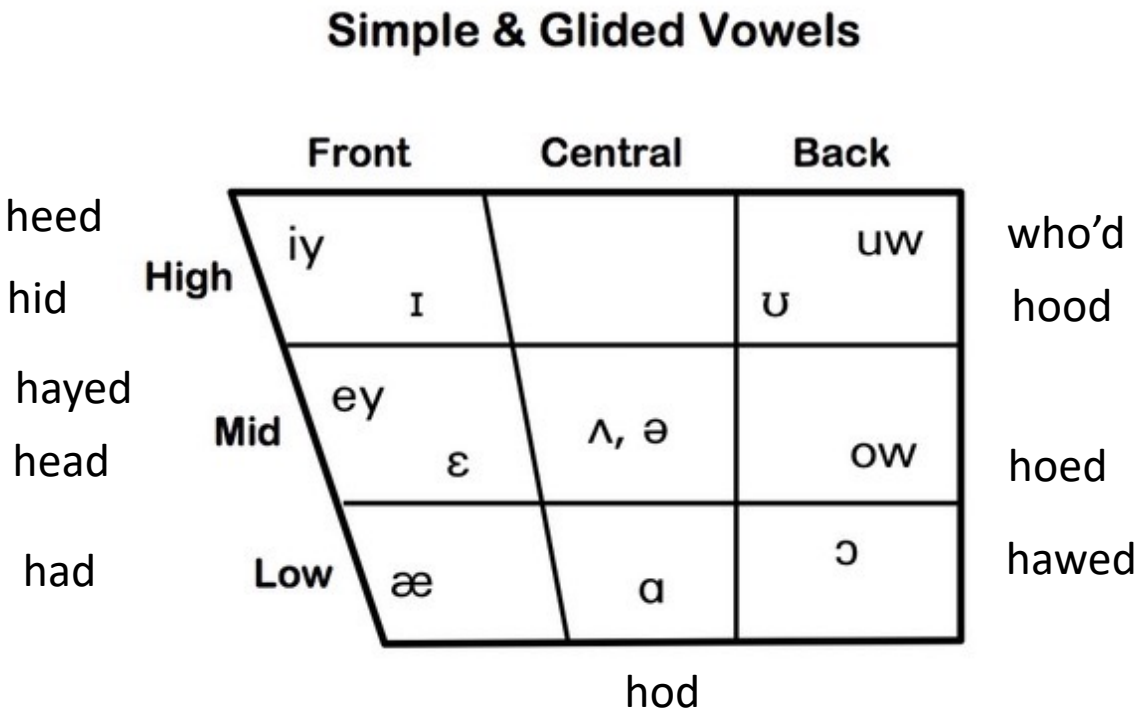
- Vowels: unblocked air stream

- Consonants: constricted or blocked air stream

Different sounds: Vowels

Scharenborg, 2017

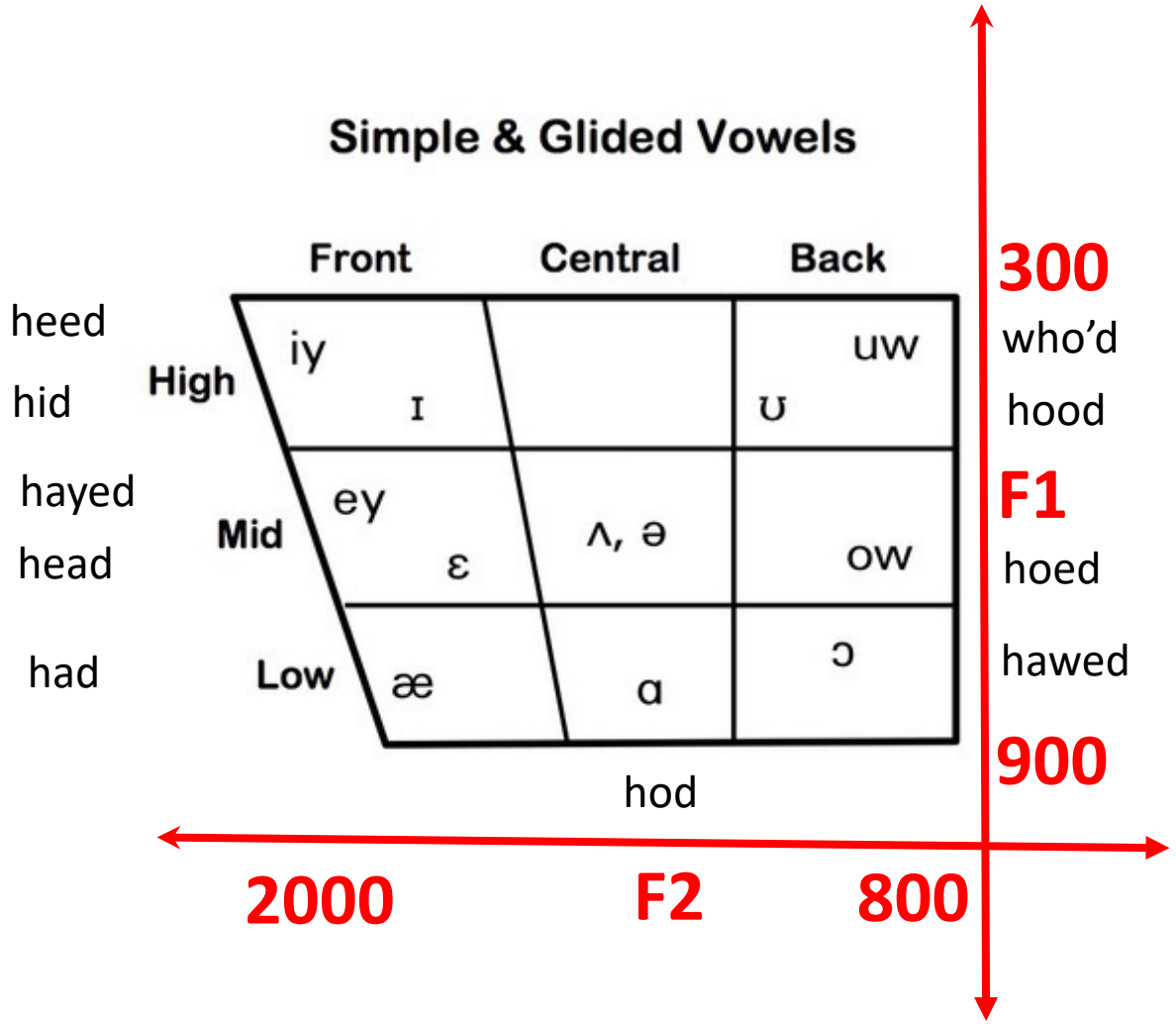
- Tongue height:
  - Low: e.g., /a/
  - Mid: e.g., /e/
  - High: e.g., /i/
- Tongue advancement:
  - Front : e.g., /i/
  - Central : e.g., /ə/
  - Back : e.g., /u/
- Lip rounding:
  - Unrounded: e.g., /ɪ, ɛ, e, ə/
  - Rounded: e.g., /u, o, ɔ/
- Tense/lax:
  - Tense: e.g., /i, e, u, o, ɔ, ɑ/
  - Lax: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



Different sounds: Vowels

Scharenborg, 2017

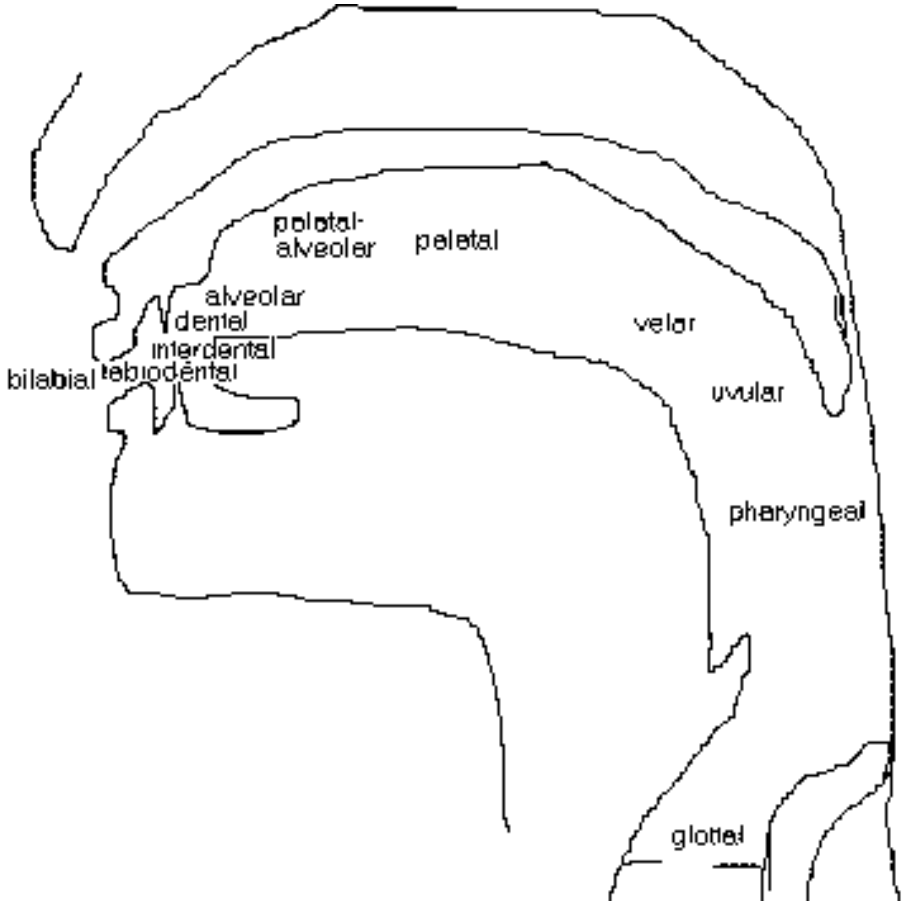
- Tongue height:
  - Low: e.g., /a/
  - Mid: e.g., /e/
  - High: e.g., /i/
- Tongue advancement:
  - Front : e.g., /i/
  - Central : e.g., /ə/
  - Back : e.g., /u/
- Lip rounding:
  - Unrounded: e.g., /ɪ, ɛ, e, ə/
  - Rounded: e.g., /u, o, ɔ/
- Tense/lax:
  - Tense: e.g., /i, e, u, o, ɔ, ɑ/
  - Lax: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

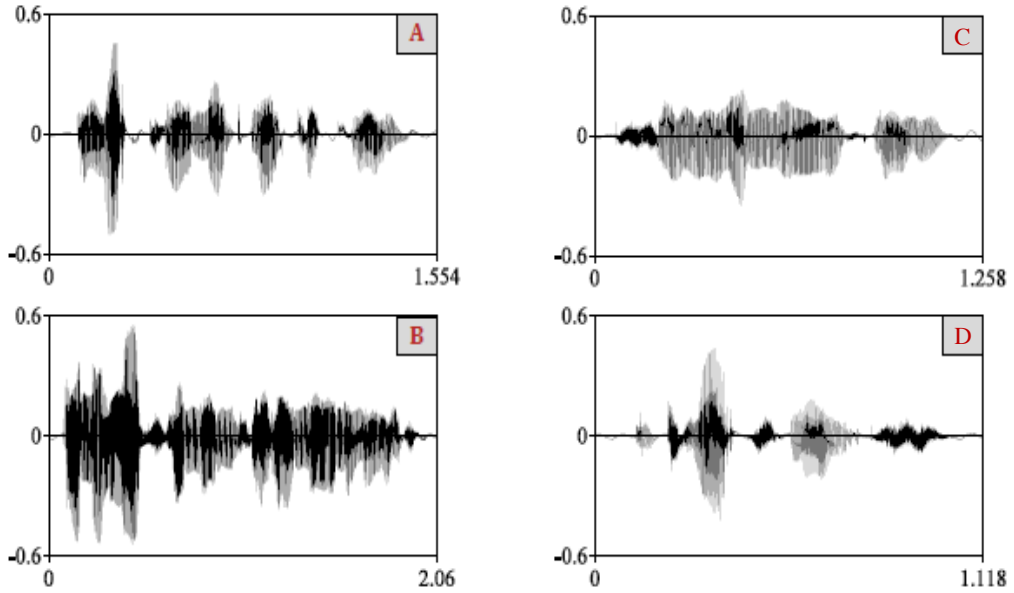
- https://www.youtube.com/watch?v=DcNMCB-Gsn8

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)

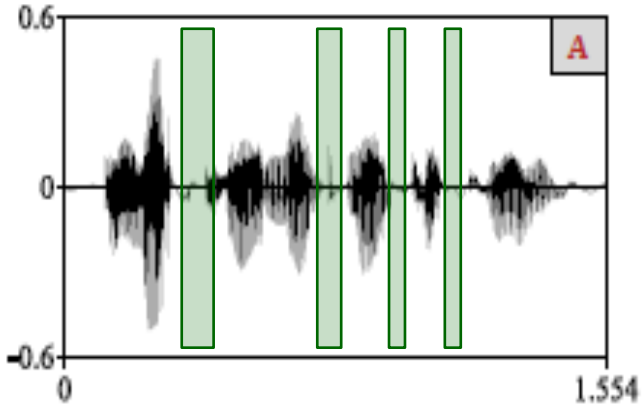
C: See you in the evening (5)

D: Attachment (1)

Electromagnetically

Scharenborg, 2017

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



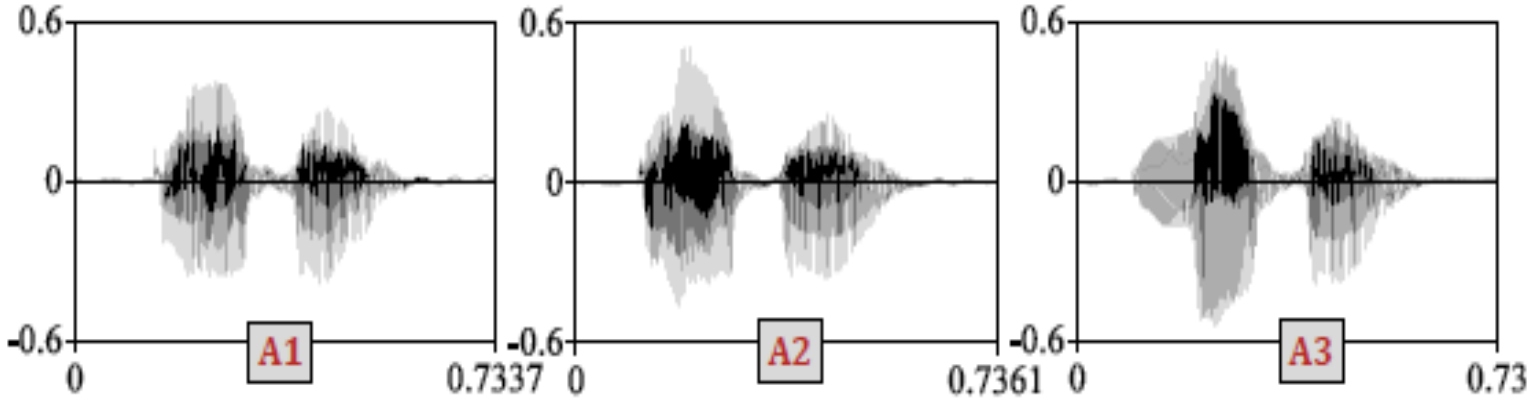
/i l ɛ kt romæ g nɛ t ɪ k ə l i/

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)

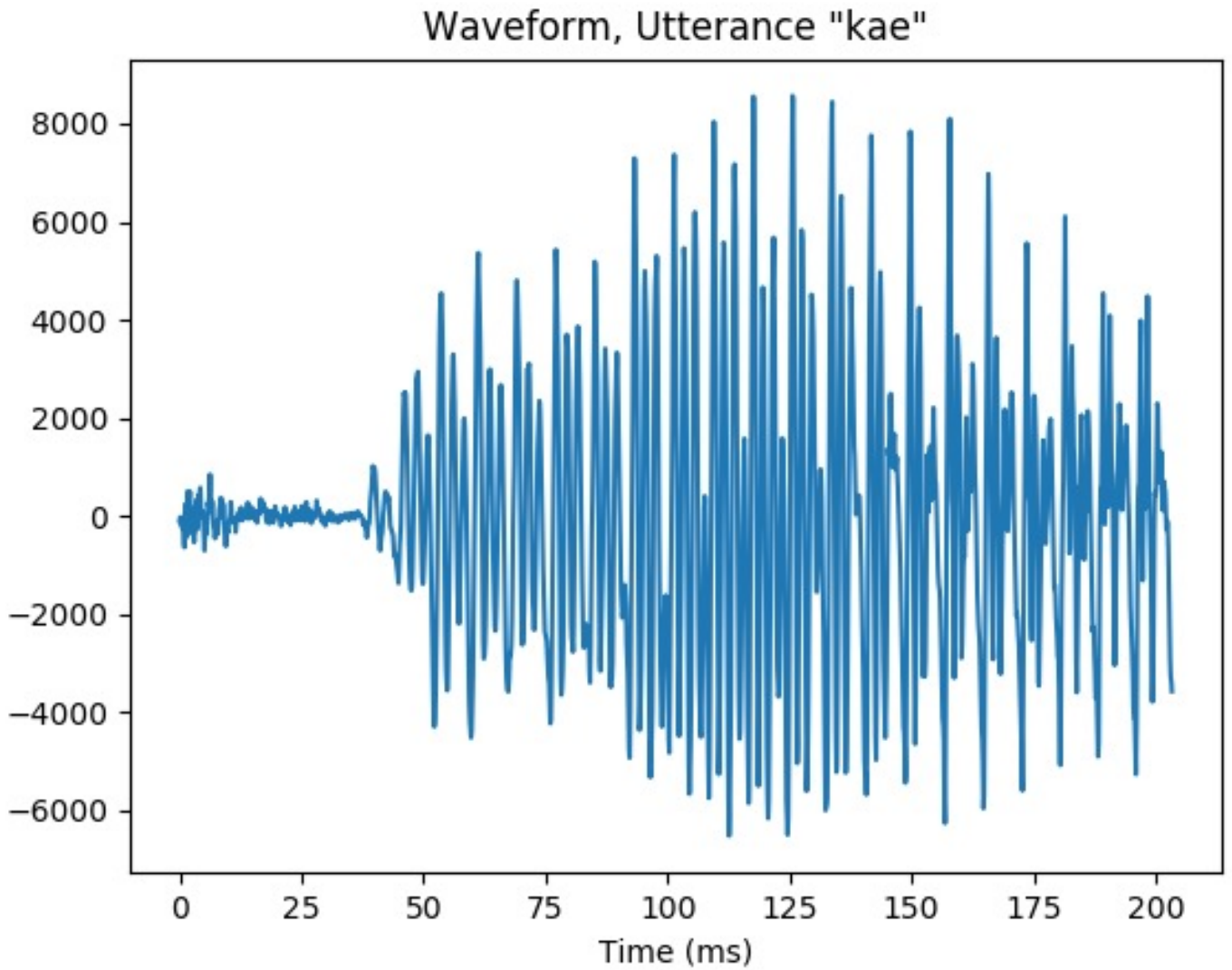
Enormous variability

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

Time domain signal: Hard to tell what he was saying

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



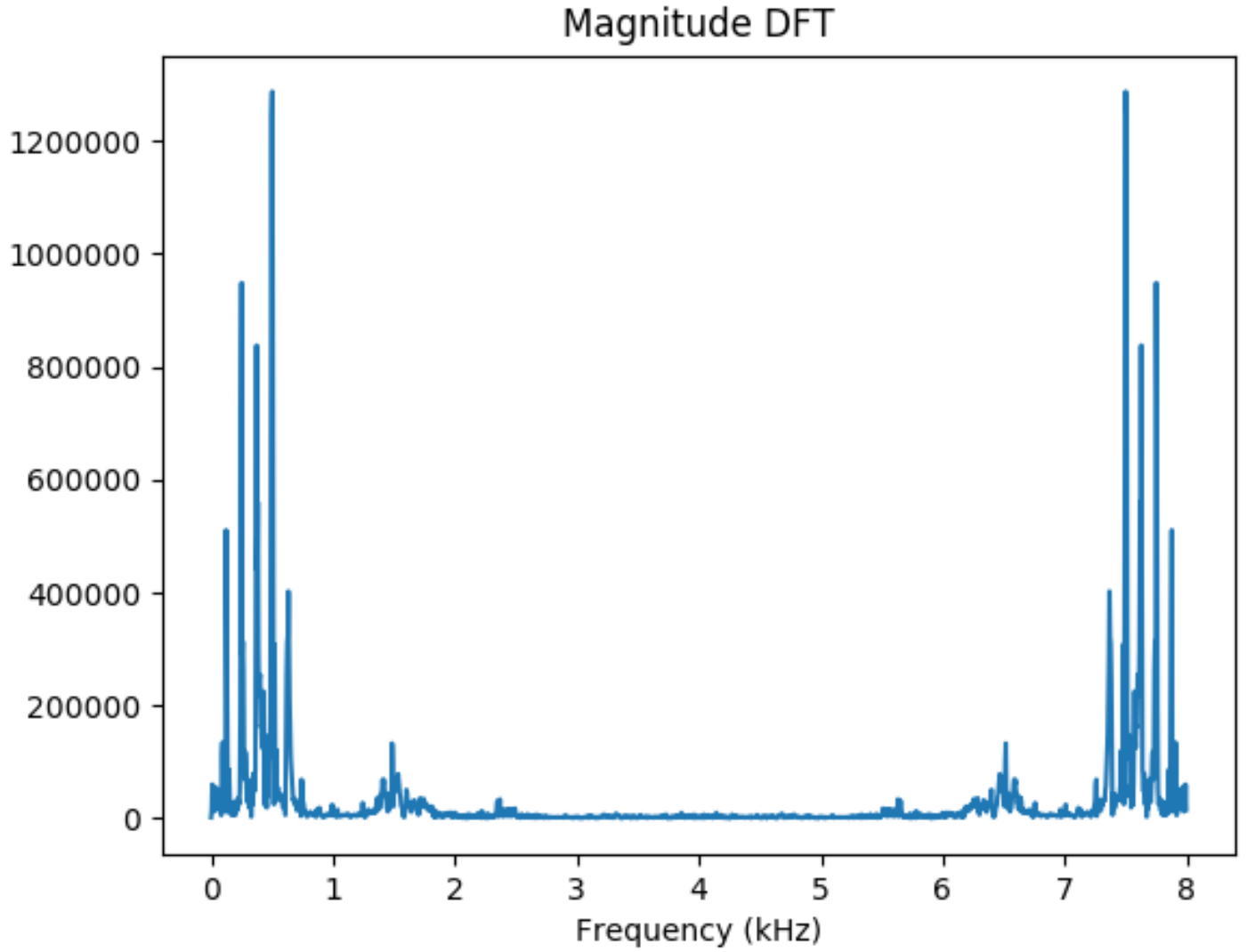
Magnitude spectrum: A little easier

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

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)



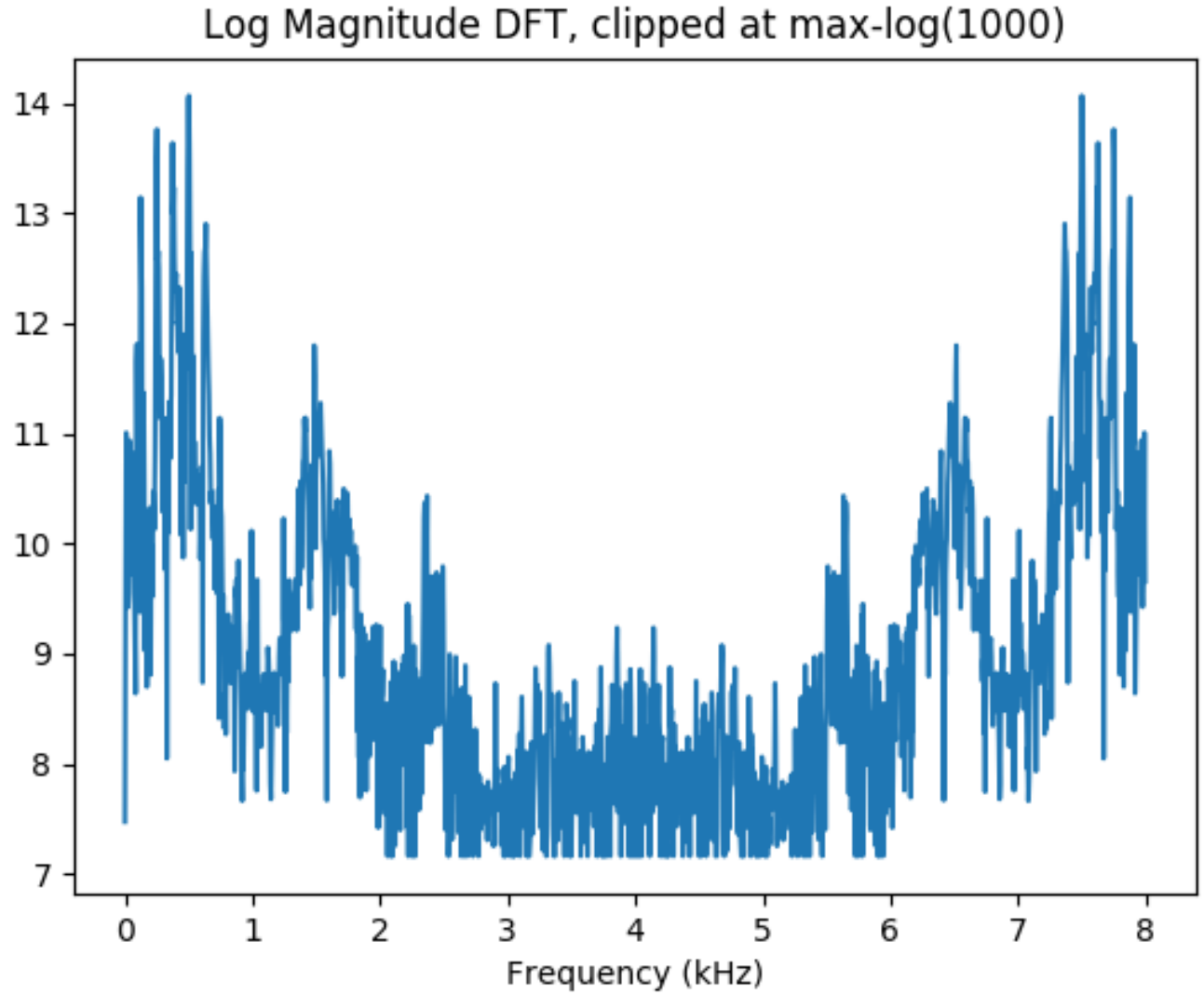
Log magnitude spectrum: A lot easier

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

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

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

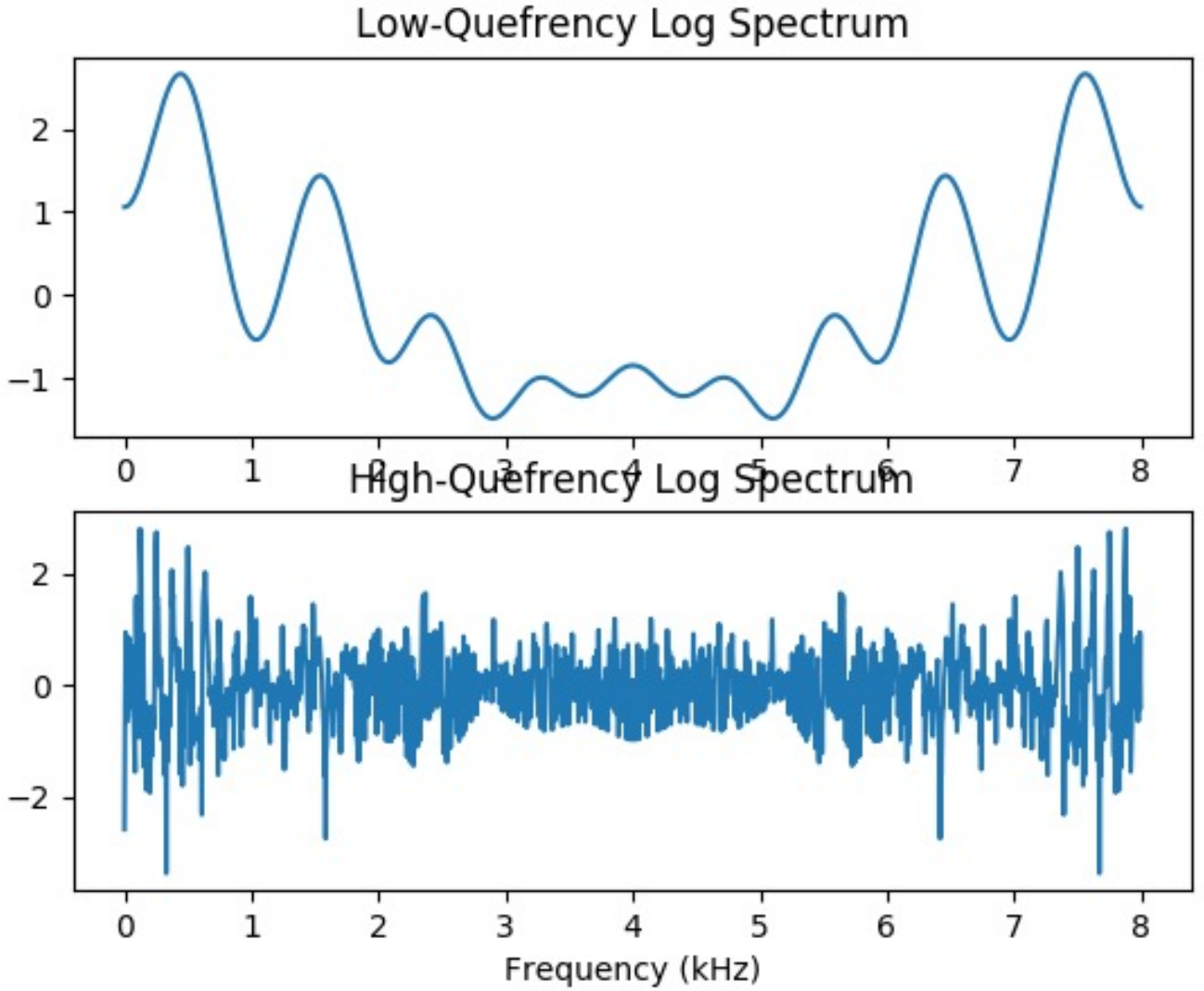
(Formants≈phone-dependent, F0≈person-dependent, though there's a lot of cross-talk)



Log spectrum = log filter + log excitation

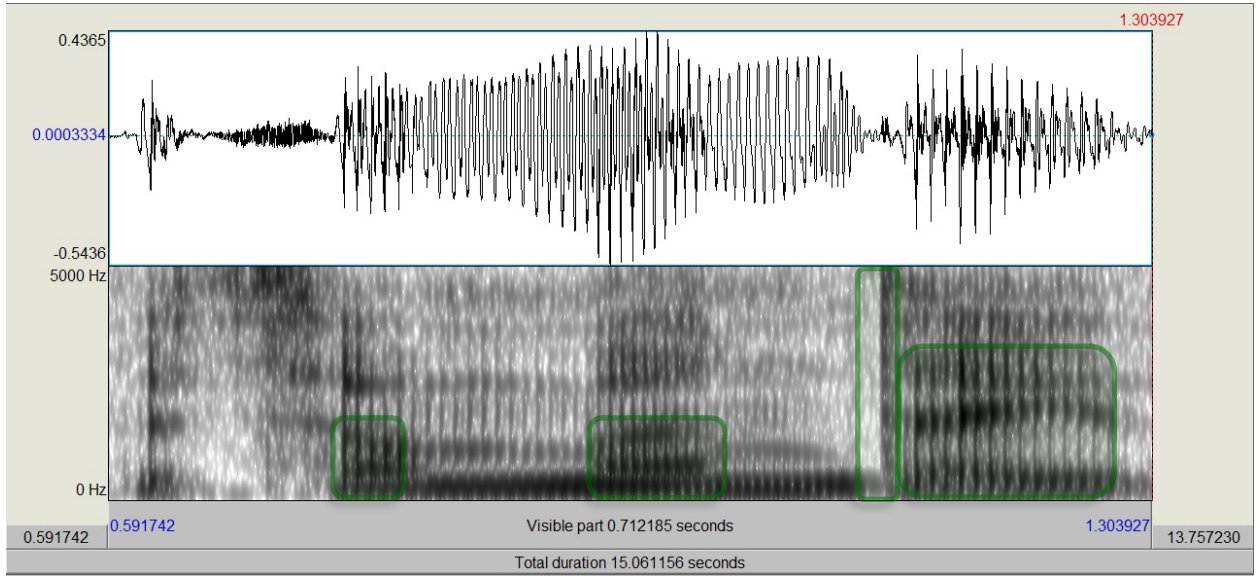
$$\ln|S(f)| = \ln|H(f)| + \ln|E(f)|$$

- 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)



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.

Summary: Speech Production

- 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

Speech Perception: Outline

- Parseval's Theorem: Cepstral Distance = Spectral Distance
- What spectrum do people hear? The basilar membrane
- Frequency scales for hearing: mel, ERB
- Filterbank coefficients and MFCC

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:

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

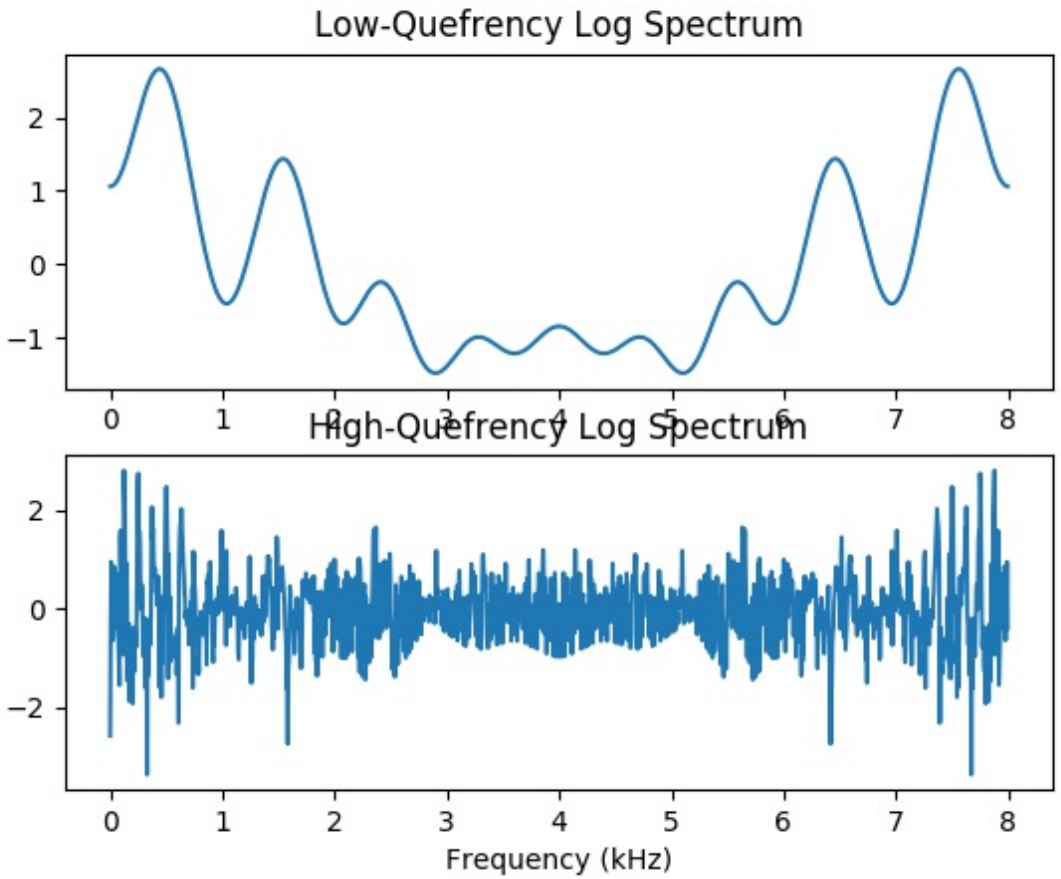
What it means for deep learning

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 L2 difference between their log spectra:

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

$$= \sum_{k=0}^{M-1} (X_k - Y_k)^2 = \sum_{n=0}^{M-1} (x_n - y_n)^2 = \|\vec{x} - \vec{y}\|^2 = \left\|\vec{X} - \vec{Y}\right\|^2$$

Low-pass liftering smooths the spectrum

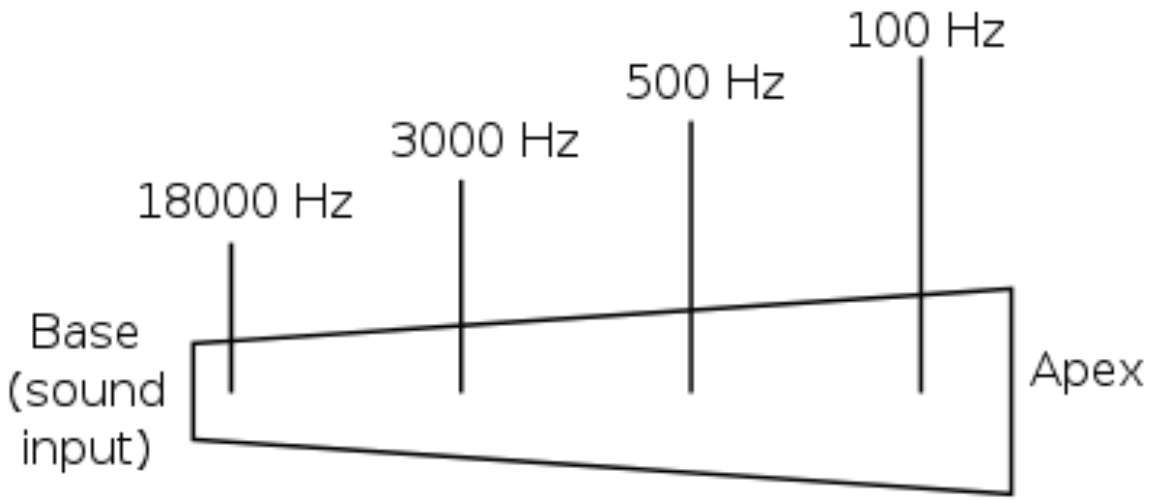


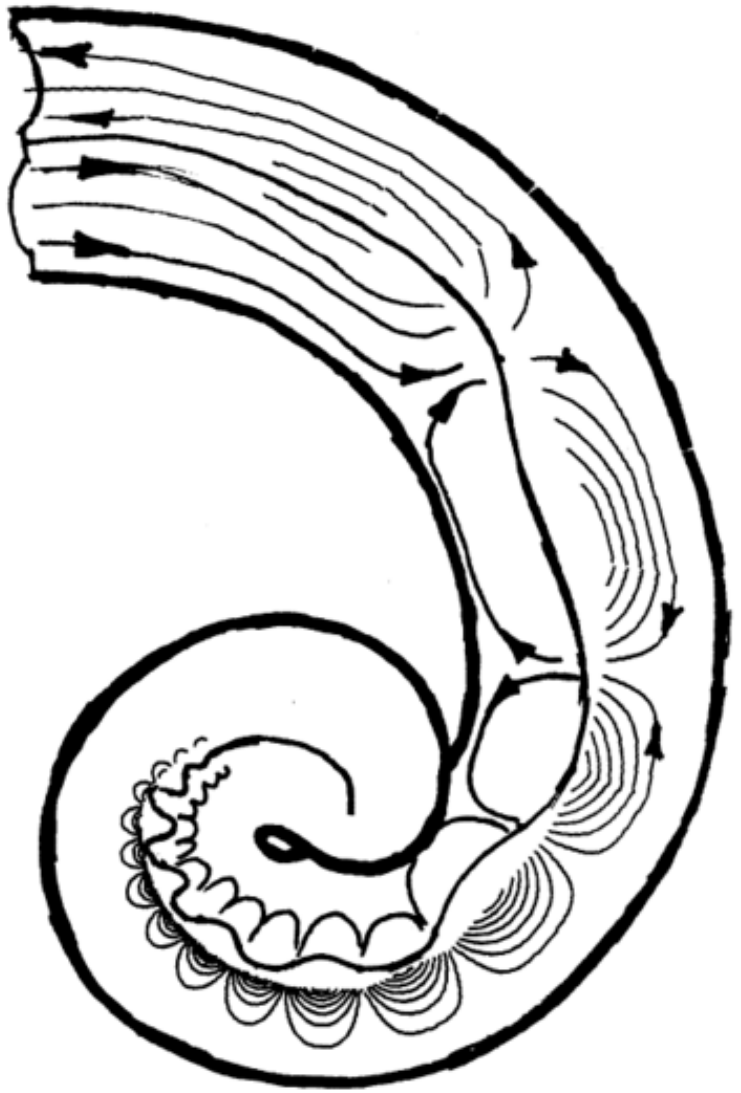
What spectrum do people hear? Basilar membrane

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

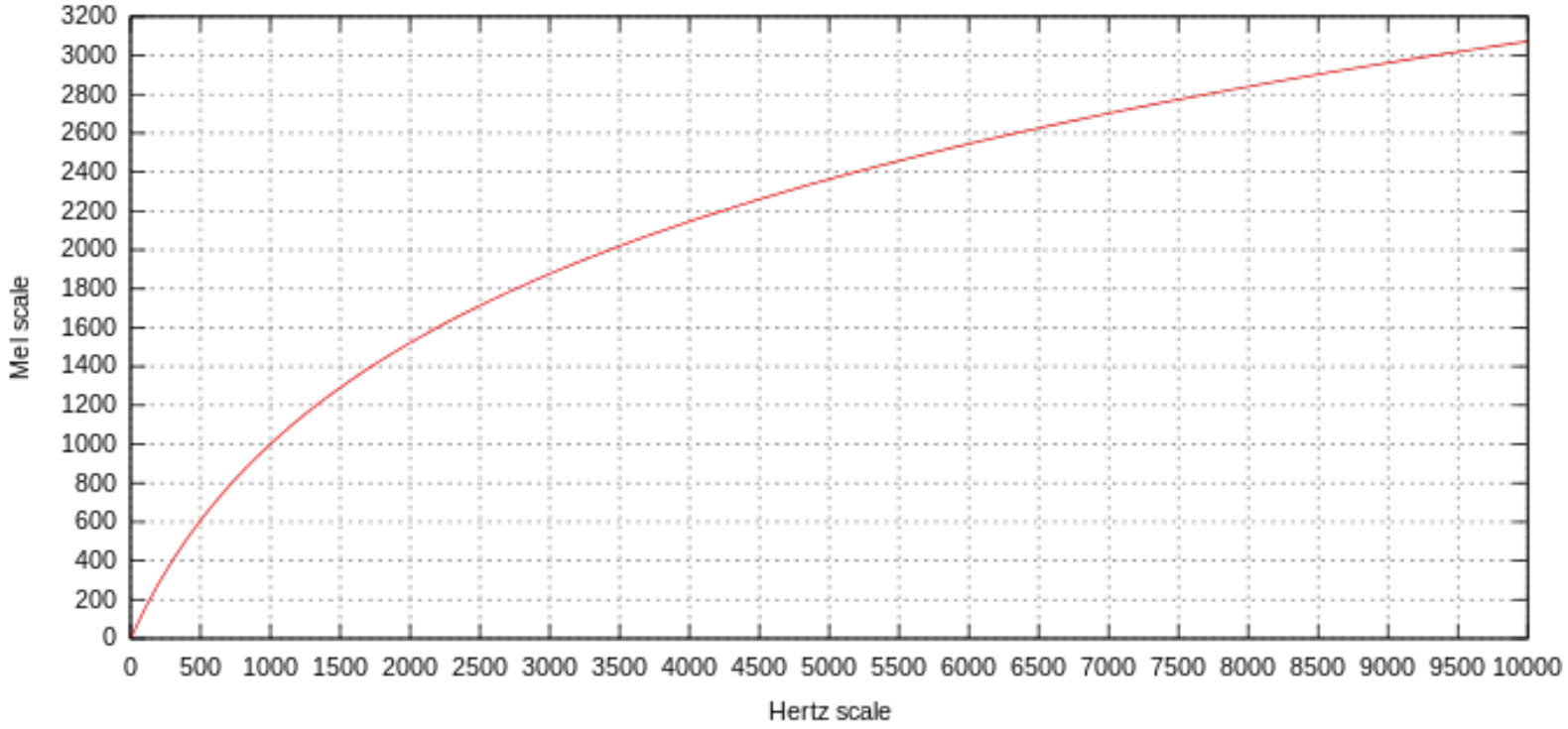
Basilar membrane of the cochlea = a bank of mechanical bandpass filters





Frequency scales for hearing: mel scale, ERB scale

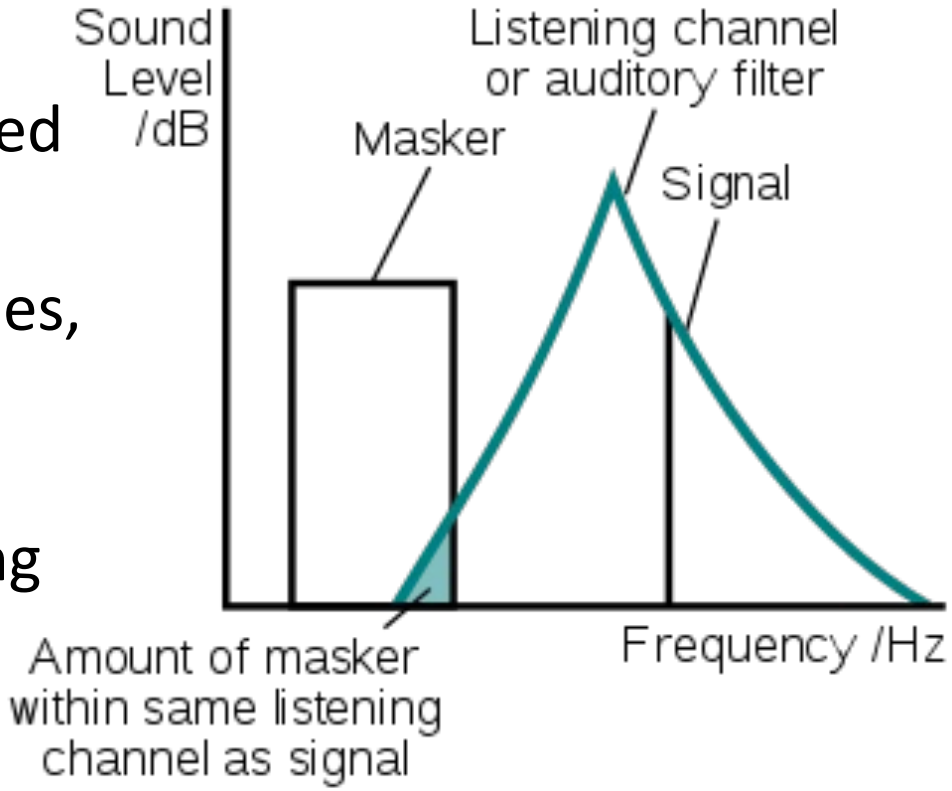
Mel-scale



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

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).



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-frequency 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

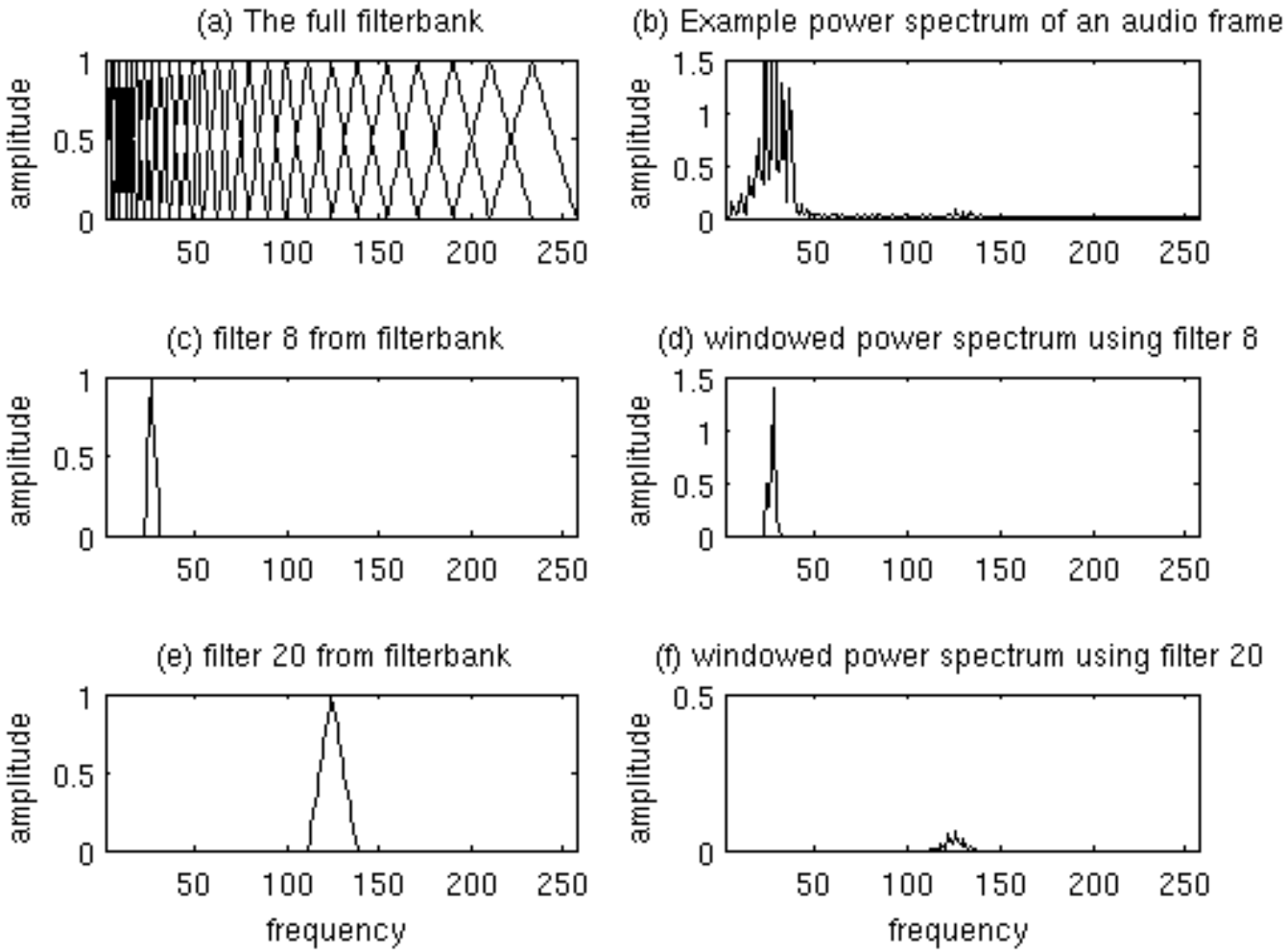
The solution:

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

Where

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

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



Inclusive Speech Technology

Outline

- The many dimensions of ability and disability
- Speech technology: the Transformer
- Speech technology: Pre-training and Fine-tuning
- Speech technology: Audiobooks
- Inclusive speech recognition
- The Speech Accessibility Project
- Disability in the public sphere

Abilities that are financially useful

- Bench-press 225 pounds (weight of a washing machine)
- Rap 200 words/minute
- Differentiate a softmax w.r.t. the Fourier transform of its input
- Debug an operating system
- Notice when somebody is lying or unsure
- Conceptualize deliveries for a 200-component just-in-time assembly

What is disability?

A disability is any condition of the body or mind (impairment) that makes it more difficult for the person with the condition to do certain activities (activity limitation) and interact with the world around them (participation restrictions). - CDC

Why is disability a social construct?

The following things are socially constructed:

- How to do an activity
- How to interact with the world

…because sometimes comedy is the best way to describe the world…

https://www.tinafriml.com

…also…

https://www.instagram.com/p/Ck6X6NCgc1c/?hl=en

Why aren’t devices, buildings, and organizations accessible to everyone?

- Availability bias: Engineers, architects, and entrepreneurs overestimate the frequency of ability profiles that are most familiar to them personally

Some neurological conditions that affect speech

- Progressive neurological conditions
  - Parkinson's disease (PD) and related cell-death conditions (~1m in USA)
  - Multiple sclerosis (MS) and related demyelination conditions (~800k in USA)
  - Amyotrophic lateral sclerosis (ALS)
  - Muscular dystrophies (MD: ~250k in USA)
- Trauma
  - Stroke/CVA (~4m in USA)
  - Traumatic brain injury/Concussion (~one quarter of people surveyed)
- Lifelong, non-progressive neurological conditions
  - Down Syndrome (~400k in USA)
  - Cerebral Palsy (~764k in USA)

The engineering perspective

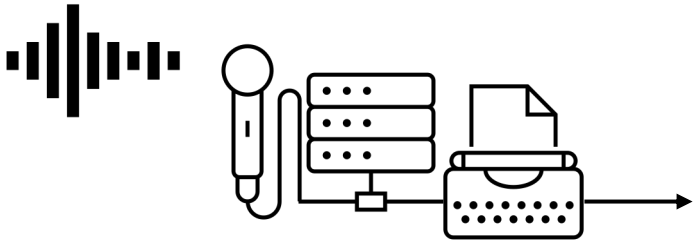
- If a device is usable by 20% of people, that’s good.
- If it is usable by 80% of people, that’s better.
- If it is usable by 95% of people, that’s even better.
- If it is usable by 99% of people, that’s even better.
- If everyone who wants to use the device can use it successfully, that’s best of all.

Outline

- The many dimensions of ability and disability
- Speech technology: the Transformer
- Speech technology: Pre-training and Fine-tuning
- Speech technology: Audiobooks
- Inclusive speech recognition
- The Speech Accessibility Project
- Disability in the public sphere

Automatic speech recognition (ASR)

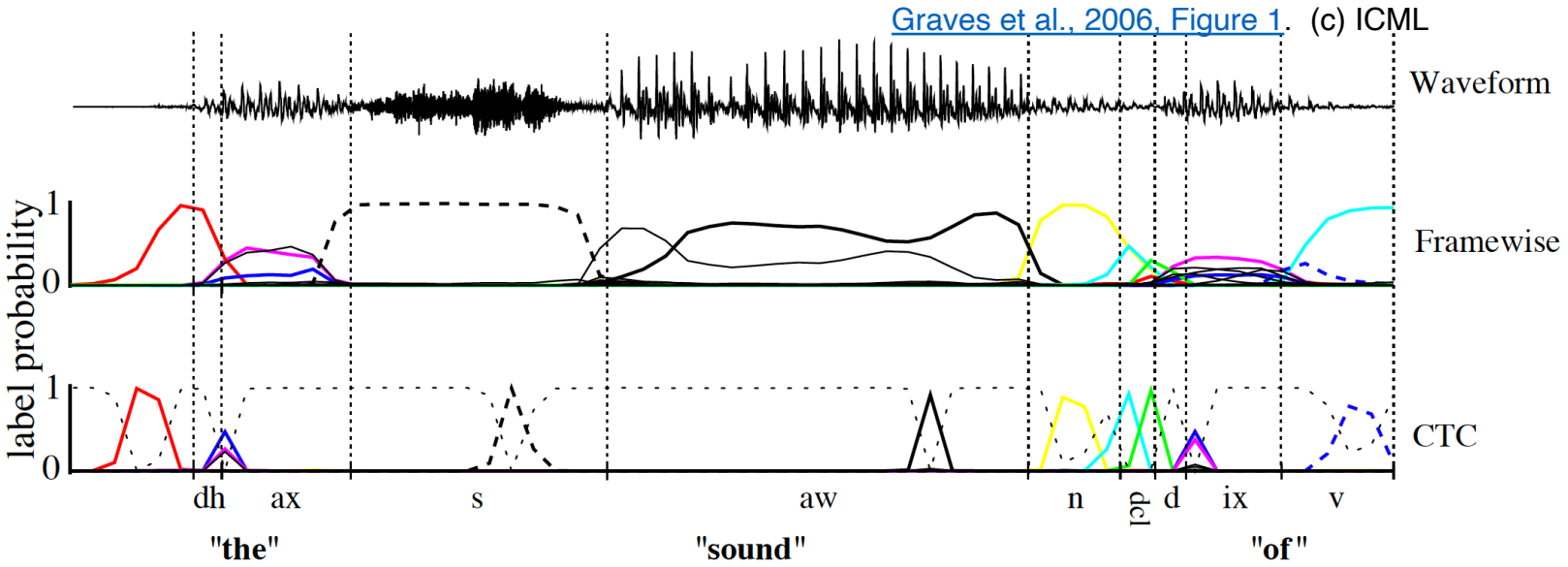
CC-BY 2.0, JCS, 2021

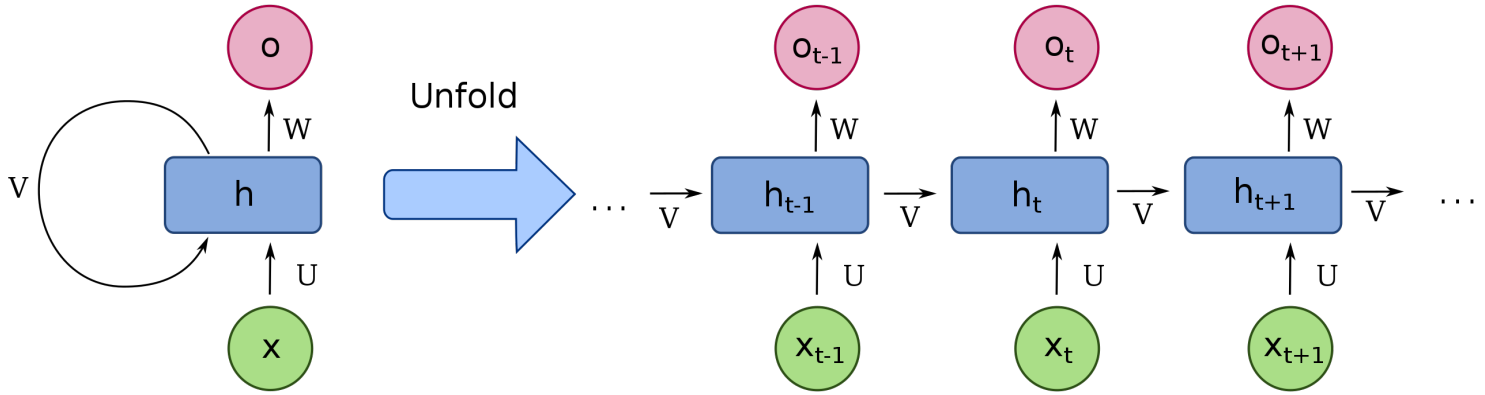


With every breath from my bronze-pounded chest, we will raise this wounded world into a wondrous one.
Amanda Gorman, 2021

- Input: sampled audio waveform, $x = [x[1], \ldots, x[T]]$
- Output: characters, $y = \{y_1, \ldots, y_N\}$

Recurrent neural network





By fdeloche - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=60109157

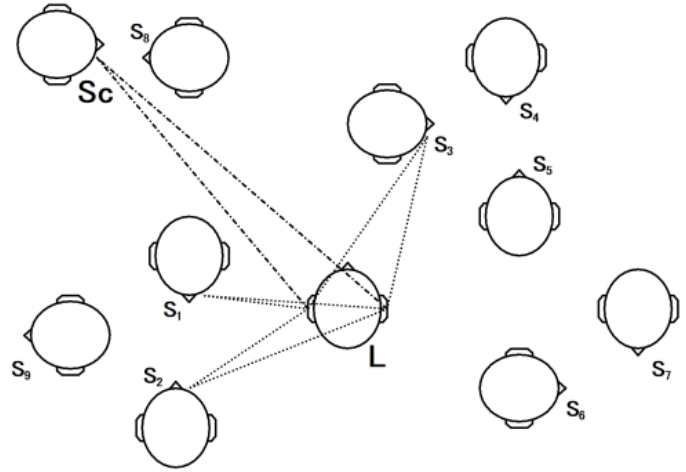
- In a recurrent neural network (RNN), the hidden node activation vector, h_t, depends on the value of the same vector at time $t-1$.
- From 2014-2017, the best speech recognizers used RNNs.
- The input is x_t=speech (waveform or spectrogram).
- The output is o_t=text characters.

The Cocktail-Party Effect

- If you are focusing on one person's voice, but hear your name spoken by another person, your attention immediately shifts to the second voice.
- This "cocktail-party effect" suggests a model of hearing in which all sounds are processed preconsciously. Trigger sounds in an unattended source will cause attention to re-orient to that source.

https://commons.wikimedia.org/wiki/File:Cocktail_party_attendees_at_Fuller_Lodge,_1946.jpg

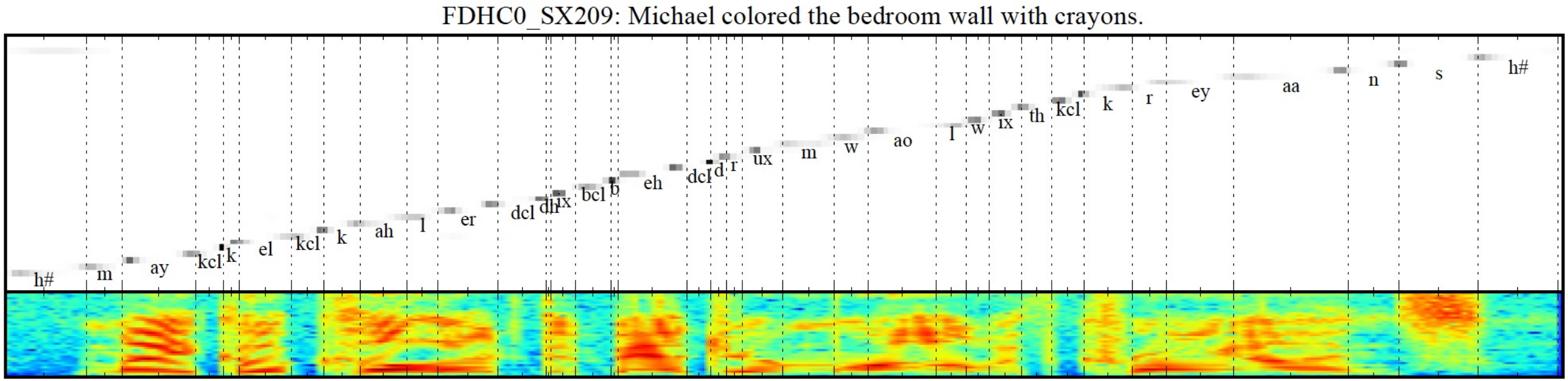
https://commons.wikimedia.org/wiki/File:Cocktail_Party_At_The_Imperial_Hotel_March_13,_1961_(Tokyo,_Japan)_(4966610682).jpg



https://commons.wikimedia.org/wiki/File:Cocktail-party_effect.svg

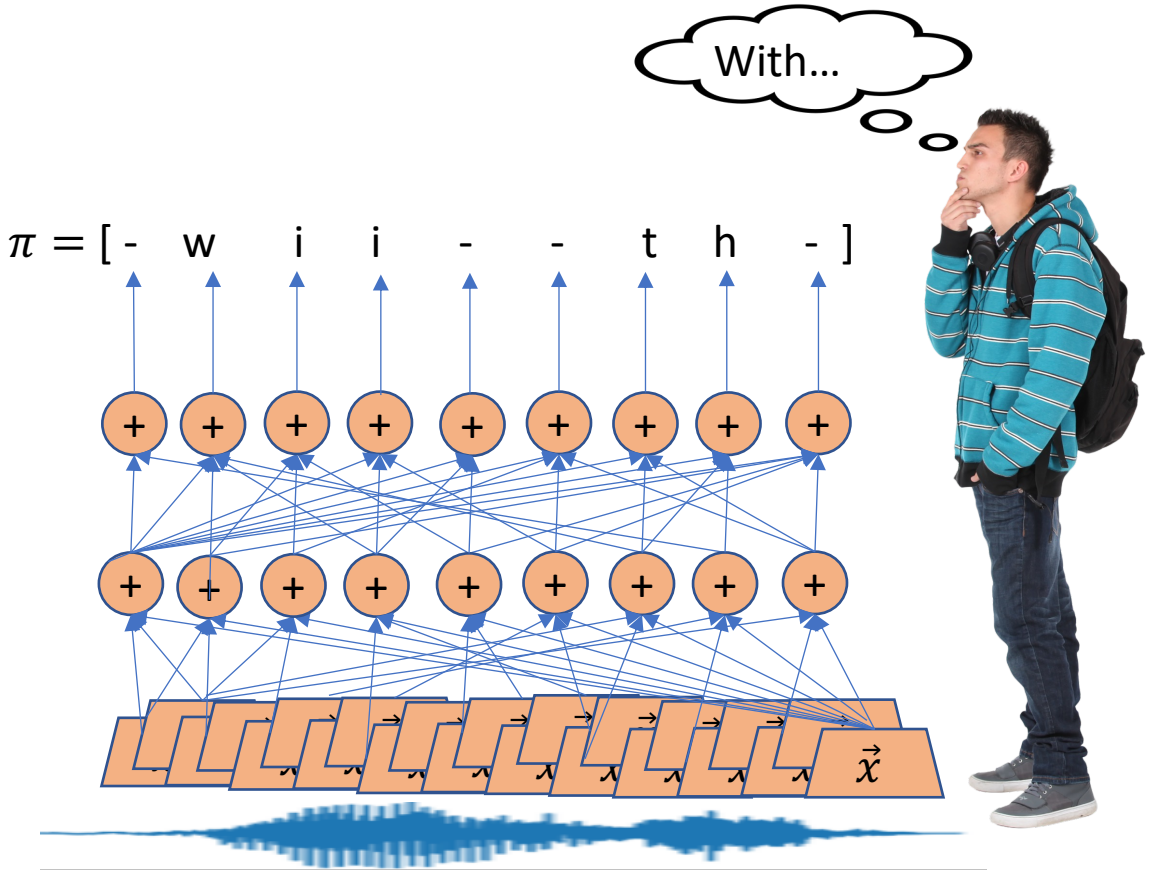
Bottom-up attention as a strategy for machine listening

- In 2014, researchers proposed that the past 200ms of RNN state vectors should be stored in a “short-term memory buffer”
- A speech recognizer can attend to several centiseconds, all at one time, to decide what words it thinks it is hearing



Chorowski, Bahdanau, Serdyk, Cho & Bengio, Attention-Based Models for Speech Recognition, Fig. 1

The Transformer: “Attention is all you need”



- In 2017, researchers proposed that the short-term memory buffer should contain raw signals, not processed signals.
- All processing is done using a model of bottom-up attention.
- The “context” for each part of the signal is computed by a model of perceptual similarity:
  - Perceptual similarity in the first layer
  - Semantic similarity in the top layer

Word Error Rates using Transformers

By 9/2020, transformers had error rates of:

- 2%: English, quiet recording conditions
- 4%: Chinese or Japanese, quiet recording conditions
- 5-7%: if the reference transcript has errors
- 14%: 2-talker mixtures, synthetic reverberation
- 38%: actual in-home recordings in noisy households

Table 1. CER/WER results on various open source ASR corpora. Both Transformer and Conformer models are implemented based on ESPnet toolkit. ∗ marks ESPnet2 results. † and ‡ indicate only w/ speed or only w/ SpecAugment, respectively. § denotes w/o any data augmentation.

Dataset	Vocab	Metric	Evaluation Sets	Transformer	Conformer
AIDATATANG	Char	CER	dev / test	(†) 5.9 / 6.7	**4.3 / 5.0**
AISHELL-1	Char	CER	dev / test	(†) 6.0 / 6.7	(*) **4.4 / 4.7**
AISHELL-2	Char	CER	android / ios / mic	(†) 8.9 / 7.5 / 8.6	**7.6 / 6.8 / 7.4**
AURORA4	Char	WER	dev_0330 (A / B / C / D)	**3.3 / 6.0 / 4.5** / 10.6	4.3 / 6.0 / 5.4 / **9.3**
CSJ	Char	CER	eval{1, 2, 3}	(∗) 4.7 / 3.7 / 3.9	(∗) **4.5 / 3.3 / 3.6**
CHiME4	Char	WER	{dt05, et05}_{simu, real}	(†) 9.6 / 8.2 / 15.7 / 14.5	**9.1 / 7.9 / 14.2 / 13.4**
Fisher-CallHome	BPE	WER	dev / dev2 / test / devtest / evltest	22.1 / 21.5 / 19.9 / 38.1 / 38.2	**21.5 / 21.1 / 19.4 / 37.4 / 37.5**
HKUST	Char	CER	dev	(†) 23.5	(†) **22.2**
JSUT	Char	CER	our split	(†) 18.7	**14.5**
LibriSpeech	BPE	WER	{dev, test}_{clean, other}	2.1 / 5.3 / 2.5 / 5.5	**1.9 / 4.9 / 2.1 / 4.9**
REVERB	Char	WER	et_{near, far}	(†) 13.1 / 15.4	(†) **10.5 / 13.9**
Switchboard	BPE	WER	eval2000 (callhm / swbd)	17.2 / 8.2	**14.0 / 6.8**
TEDLIUM2	BPE	WER	dev / test	9.3 / 8.1	**8.6 / 7.2**
TEDLIUM3	BPE	WER	dev / test	10.8 / 8.4	**9.6 / 7.6**
VoxForge	Char	CER	our split	(§) 9.4 / 9.1	(§) **8.7 / 8.2**
WSJ	BPE	WER	dev93/ eval92	(‡) **7.4 / 4.9**	(‡) 7.7 / 5.3
WSJ-2mix	Char	WER	tt	(§) 12.6	(§) **11.7**

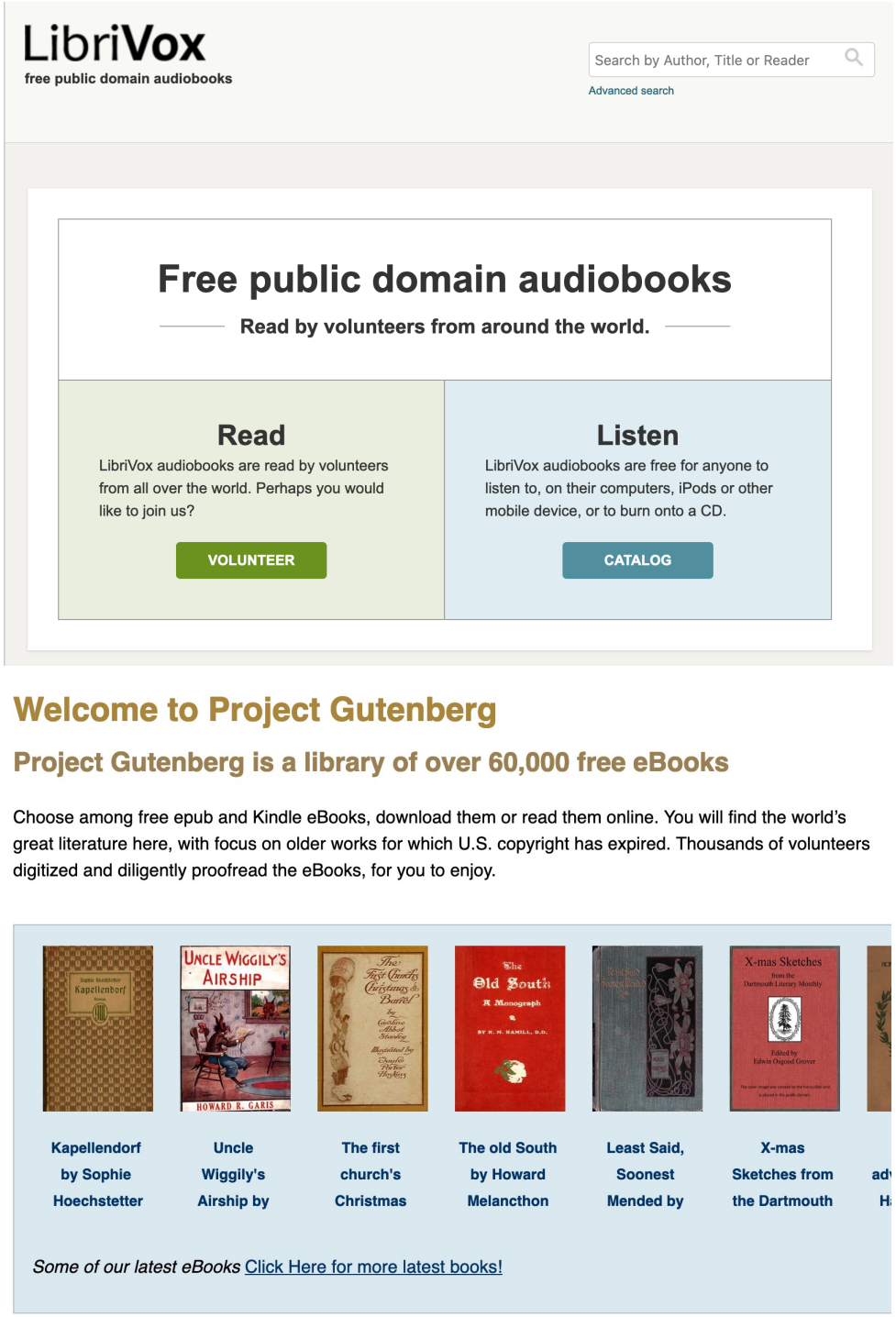
Guo, Boyer, Chang, Hayashi, Higuchi et al., ICASSP 2021, © IEEE

Outline

- The many dimensions of ability and disability
- Speech technology: the Transformer
- Speech technology: Pre-training and Fine-tuning
- Speech technology: Audiobooks
- Inclusive speech recognition
- The Speech Accessibility Project
- Disability in the public sphere

Label quality

- Audiobooks on librivox.org are readings of texts from Gutenberg.org.
- Often, readers make mistakes, or read different versions, or change the title enough to make it hard to know which Gutenberg text they read.
- Until 2020, speech technology was trained using “labeled data:”
  - 400 hours of librivox books with verified transcripts (“Librispeech Clean”)
  - 600 hours of librivox books with acoustic noise or transcript errors (“Librispeech Other”)



What do babies hear?

How much unlabeled speech does a baby hear?

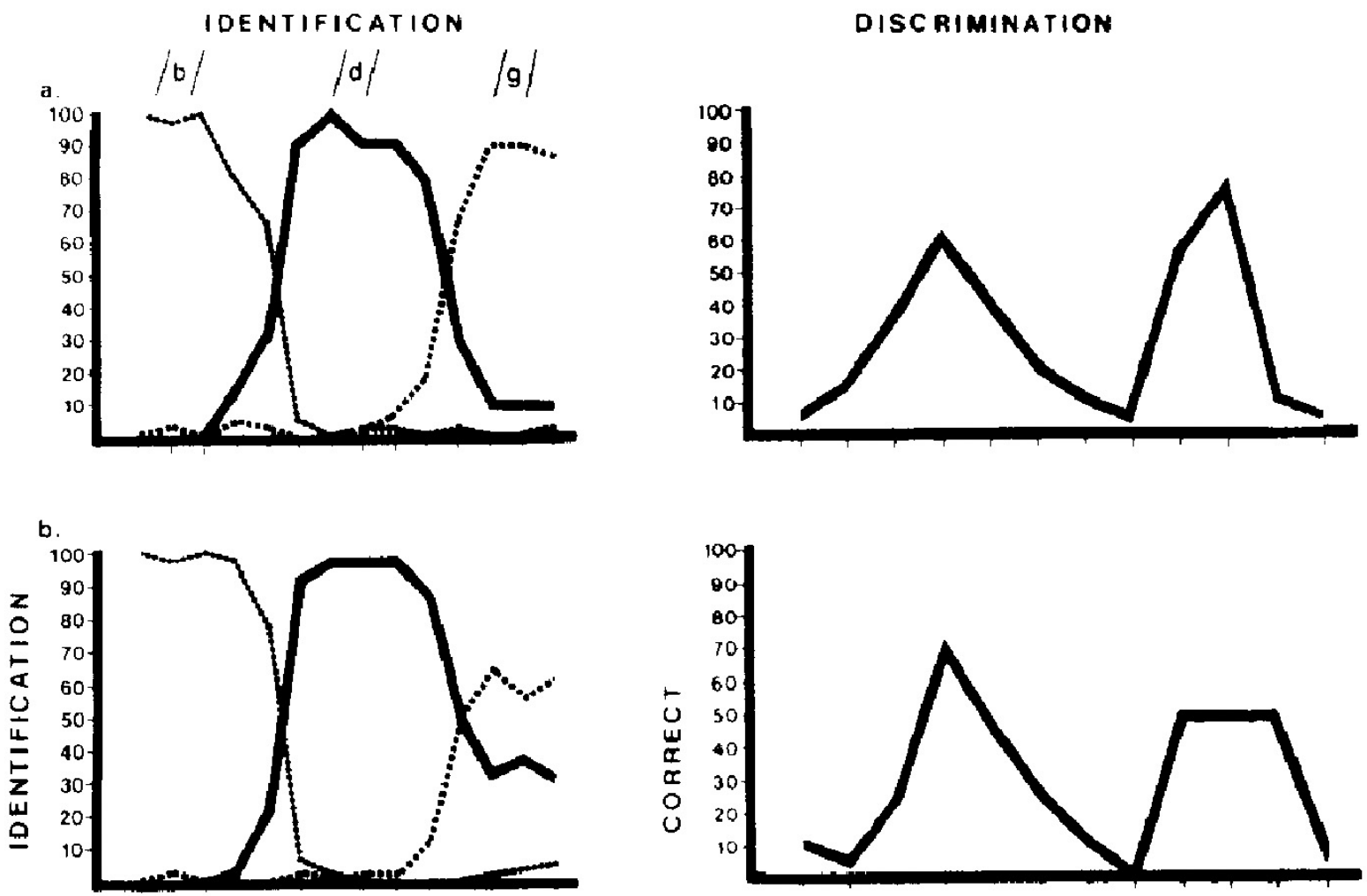
- 2000-15000 words/day = 600-4500 hours of speech by age 6 (Weisleder & Fernald, 2013)

How much labeled speech does a baby hear?

- 30 (?) words/day accompanied by referential gestures = 9.1 hours of speech by age 6

By Steve Jurvetson from Los Altos, USA - A Proper Space Book for Babies, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=105132804

Is speech learned, or innate? (Hint: it's a trick question)

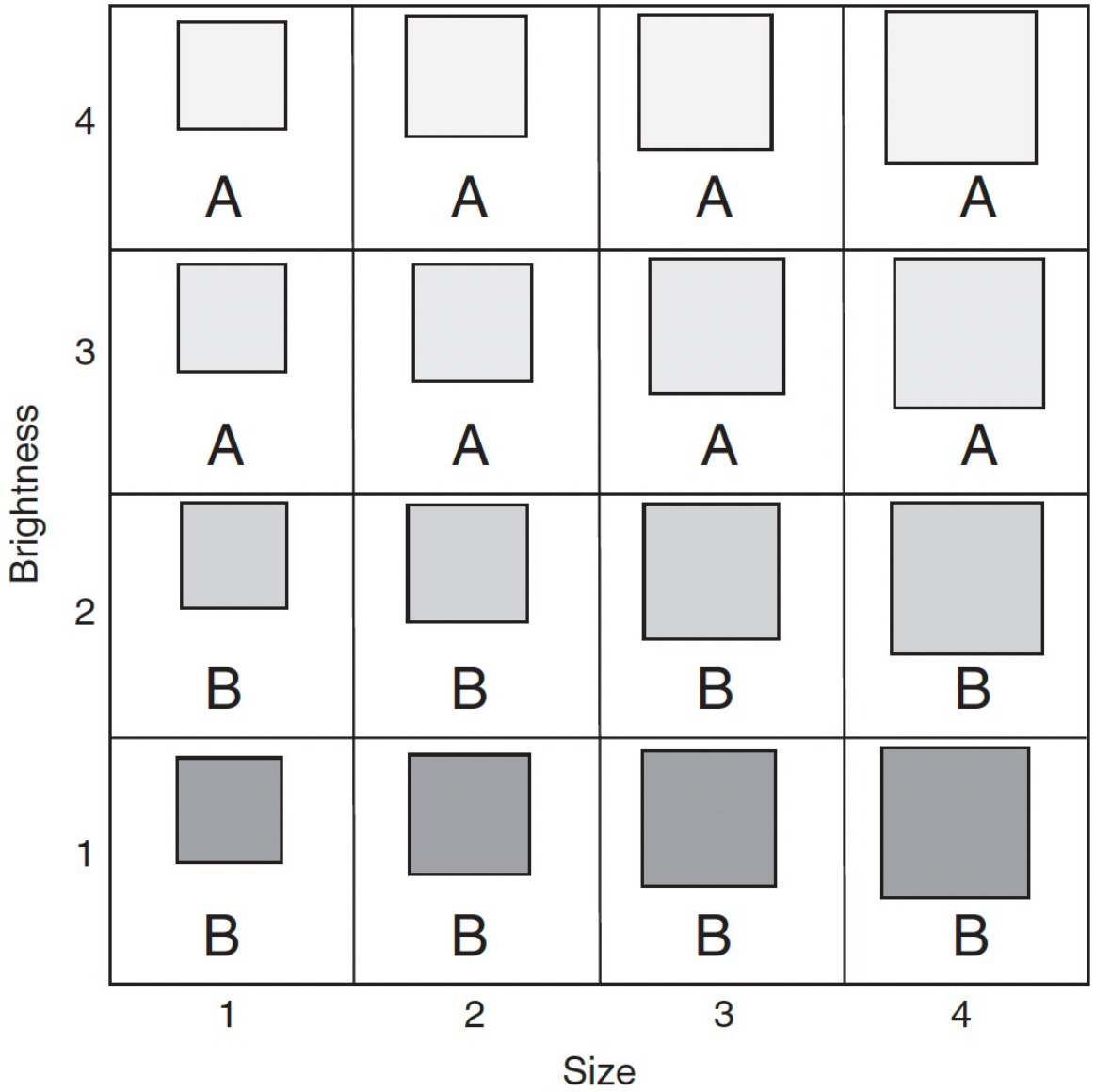


[Brandt & Rosen, 1980](#) © Brain & Language

- 15 synthetic syllables, continuous from /ba/ to /da/ to /ga/
- same label ⇒ hard to tell if they are same sound or different sounds
- different labels ⇒ heard as obviously different

Categorical perception can be learned!

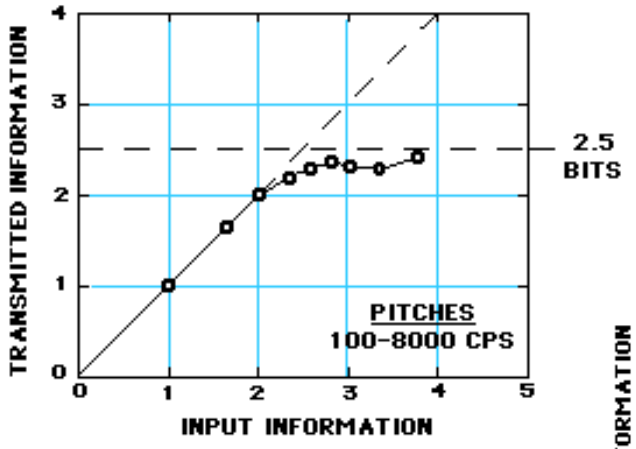
- People trained to categorize based on ***brightness*** show reduced within-category perceptual memory, and greater across-category perceptual memory, for ***brightness***, but ***size*** is perceived on a continuum.
- People trained to categorize based on ***size*** show reduced within-category perceptual memory, and greater across-category perceptual memory, for ***size***, but ***brightness*** is perceived on a continuum.

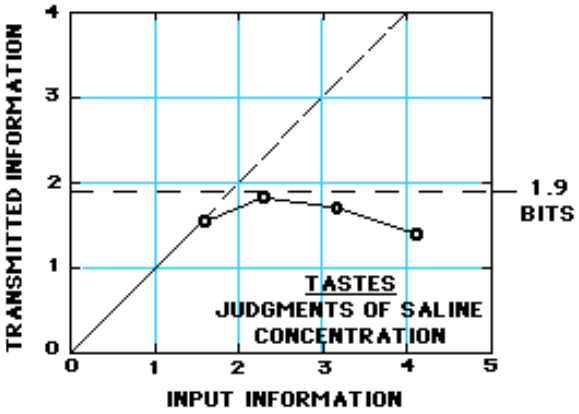


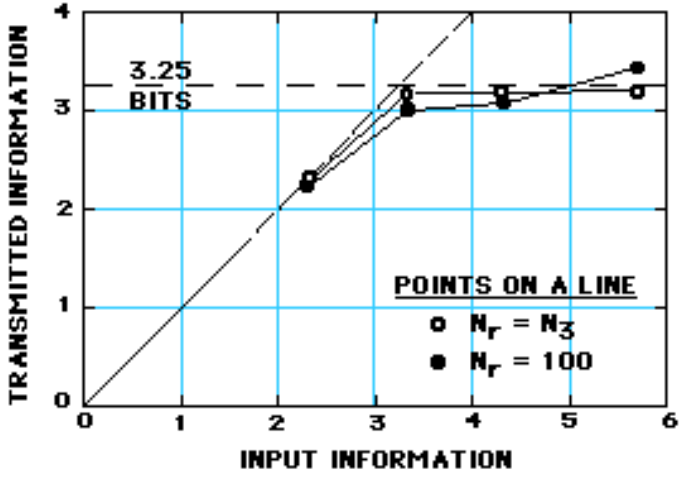
Goldstone & Hendrickson, 2009. © Wiley & Sons

Categorical perception as a cognitive bias

- If we categorize things, maybe we can remember them longer.
- In “The Magic Number Seven,” Miller argued that people can be taught to categorize any continuum (pitch, taste, position, size) into seven categories, but not more.

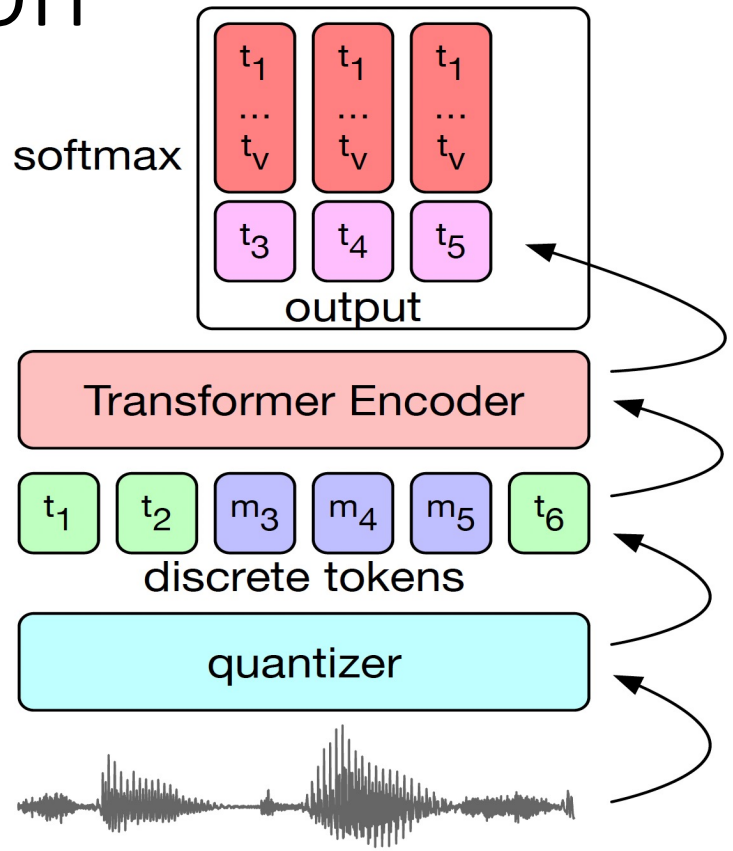






Unsupervised pre-training of transformers based on categorical perception

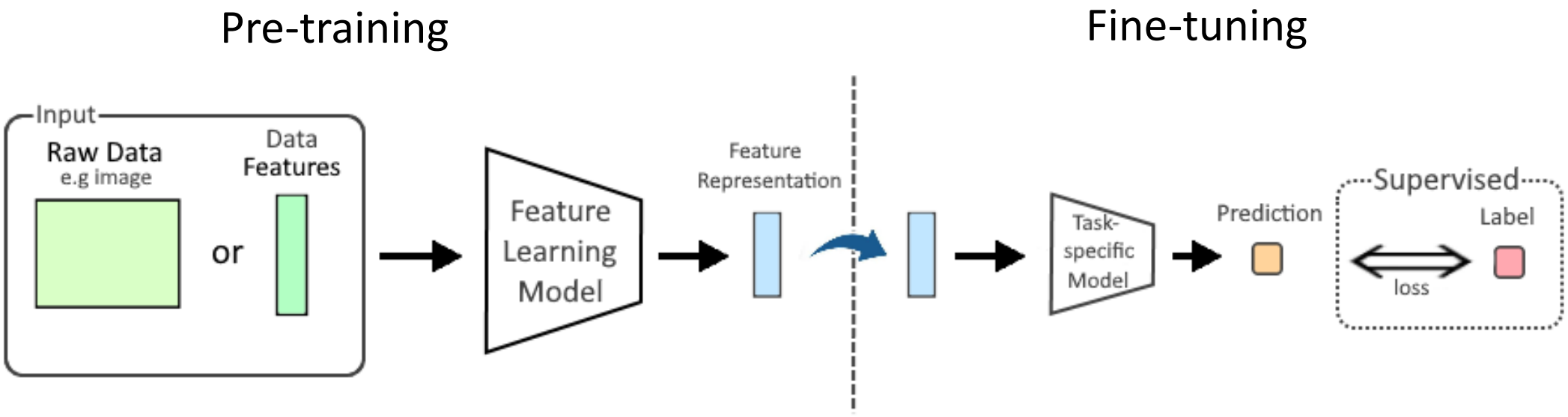
- Given: 60,000 hours of speech, with no associated text.
- Suppose we train the neural network to form its own categories. What would make those categories speech-like?
- **Context-Predictable Speech Categories**: given the context (the quantized units t_1, t_2, and t_6), it should be possible to figure out what phonemes were masked (t_3, t_4, t_5).



Baevski, Auli & Mohamed, 2019

Pre-training and Fine-tuning

- A transformer is pre-trained to create its own context-predictable speech categories using, say, 60,000 hours of speech
- Then it is fine-tuned using a few hours, or a few hundred hours, or labeled speech



Word Error Rates using Pre-Training

Pre-training makes it possible to achieve error rates of

- 4.4% using only 10 minutes of labeled data
- 2.6% using only 1 hour of labeled data

Model	Unlabeled Data	LM	dev-clean	dev-other	test-clean	test-other
		10-min labeled				
DiscreteBERT [52]	LS-960	4-gram	15.7	24.1	16.3	25.2
wav2vec 2.0 BASE [7]	LS-960	4-gram	8.9	15.7	9.1	15.6
wav2vec 2.0 LARGE [7]	LL-60k	4-gram	6.3	9.8	6.6	10.3
wav2vec 2.0 LARGE [7]	LL-60k	Transformer	4.6	7.9	4.8	8.2
HUBERT BASE	LS-960	4-gram	9.1	15.0	9.7	15.3
HUBERT LARGE	LL-60k	4-gram	6.1	9.4	6.6	10.1
HUBERT LARGE	LL-60k	Transformer	**4.3**	7.0	4.7	7.6
HUBERT X-LARGE	LL-60k	Transformer	4.4	**6.1**	**4.6**	**6.8**
		1-hour labeled				
DeCoAR 2.0 [51]	LS-960	4-gram	-	-	13.8	29.1
DiscreteBERT [52]	LS-960	4-gram	8.5	16.4	9.0	17.6
wav2vec 2.0 BASE [7]	LS-960	4-gram	5.0	10.8	5.5	11.3
wav2vec 2.0 LARGE [7]	LL-60k	Transformer	2.9	5.4	2.9	5.8
HUBERT BASE	LS-960	4-gram	5.6	10.9	6.1	11.3
HUBERT LARGE	LL-60k	Transformer	**2.6**	4.9	2.9	5.4
HUBERT X-LARGE	LL-60k	Transformer	**2.6**	**4.2**	**2.8**	**4.8**

Hsu, Boldt, Tsai, Lakhotia, Salakhutdinov & Mohamed,

Outline

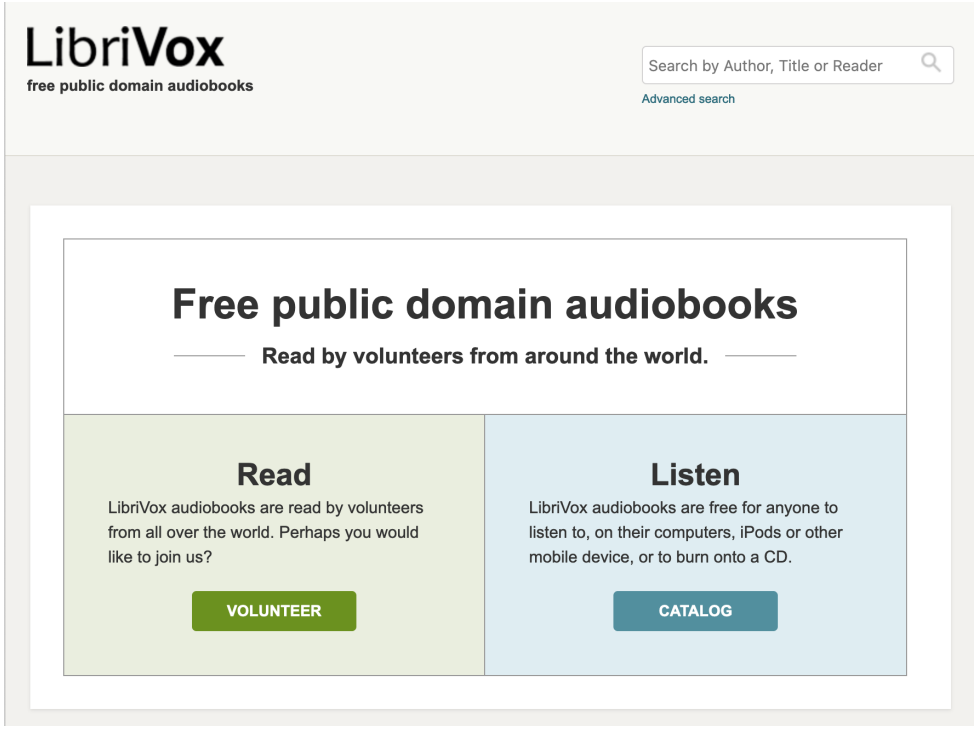
- The many dimensions of ability and disability
- Speech technology: the Transformer
- Speech technology: Pre-training and Fine-tuning
- Speech technology: Audiobooks
- Inclusive speech recognition
- The Speech Accessibility Project
- Disability in the public sphere

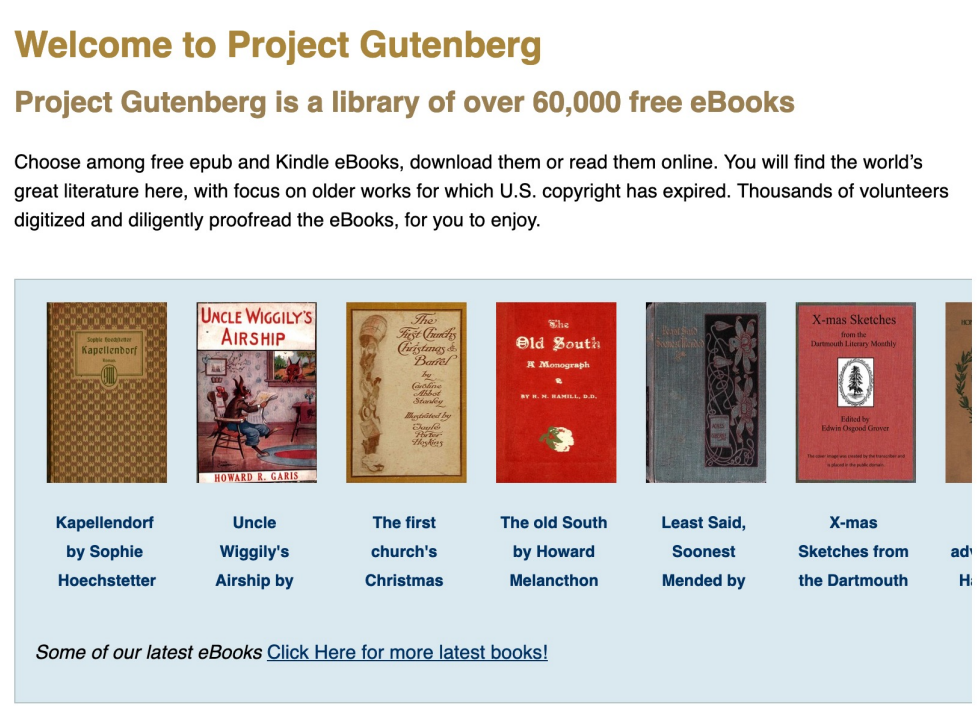
A brief history of speech recognizer training corpora

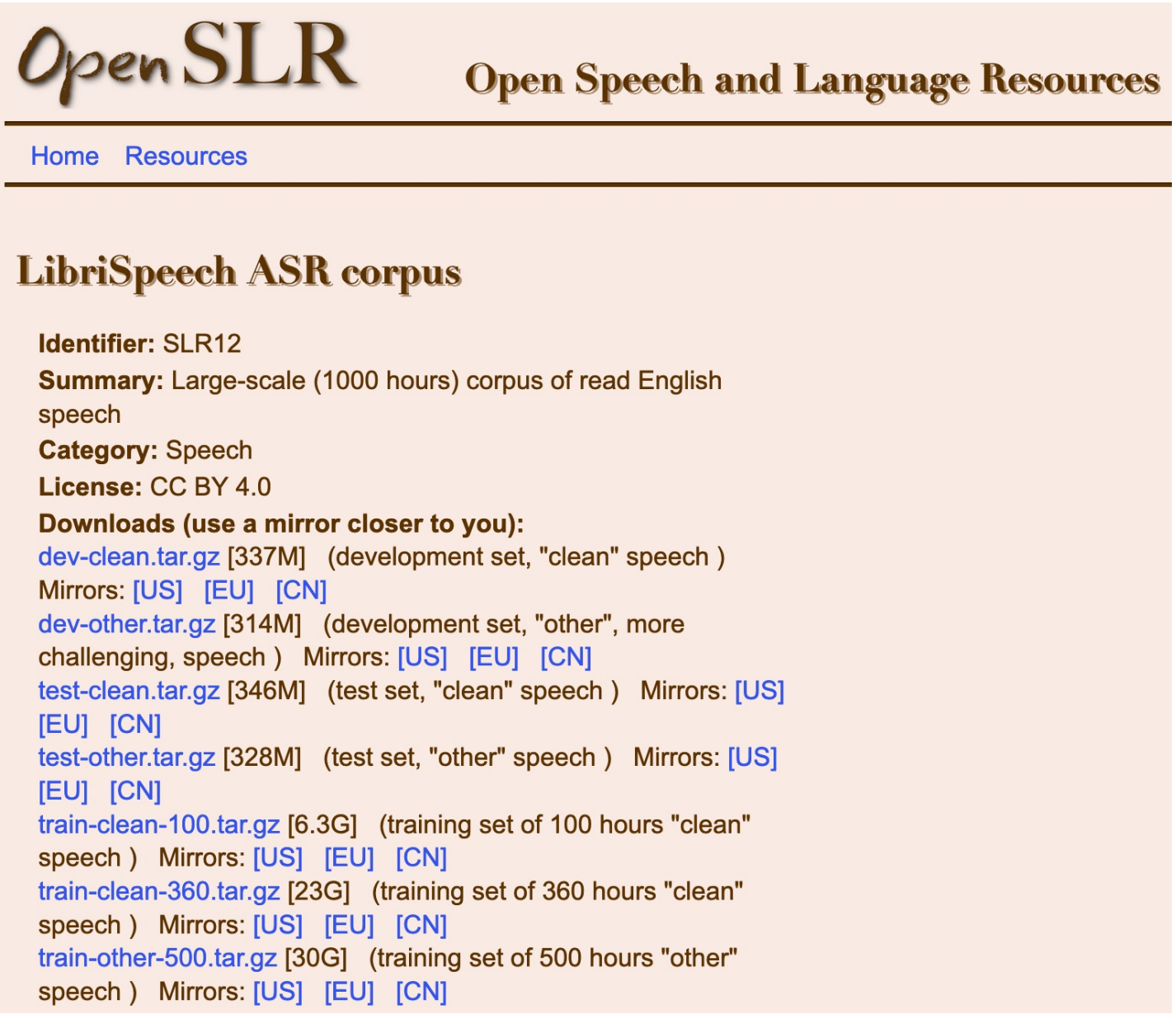
- TIMIT (1993): 4 hours, phonetically transcribed, DARPA-sponsored
- Broadcast News (1996): 104 hours, orthographically transcribed, DARPA-sponsored
- Switchboard (1997): 300 hours, orthographically transcribed, DARPA-sponsored
- Corporate data providers enter the picture. From 2000-2020, high-cost training datasets grow in both size and quality.
- Librispeech (2015): 1000 hours, curated from public domain audiobooks recorded by contributors to librivox.org.
- Multilingual Librispeech (2017): 3000 hours, curated from public domain audiobooks recorded by contributors to librivox.org.

Love, not money

- Audiobooks on librivox.org are usually readings of texts from Gutenberg.org.
- In order to be posted on librivox.org, the audio must be released into the public domain.
  - Portfolio samples of professional audiobook narrators
  - Books contributed by people who love them, and want them to be available
  - Languages contributed by people who love them, and want them to be available
  - Books contributed by people who love audiobooks







Librispeech

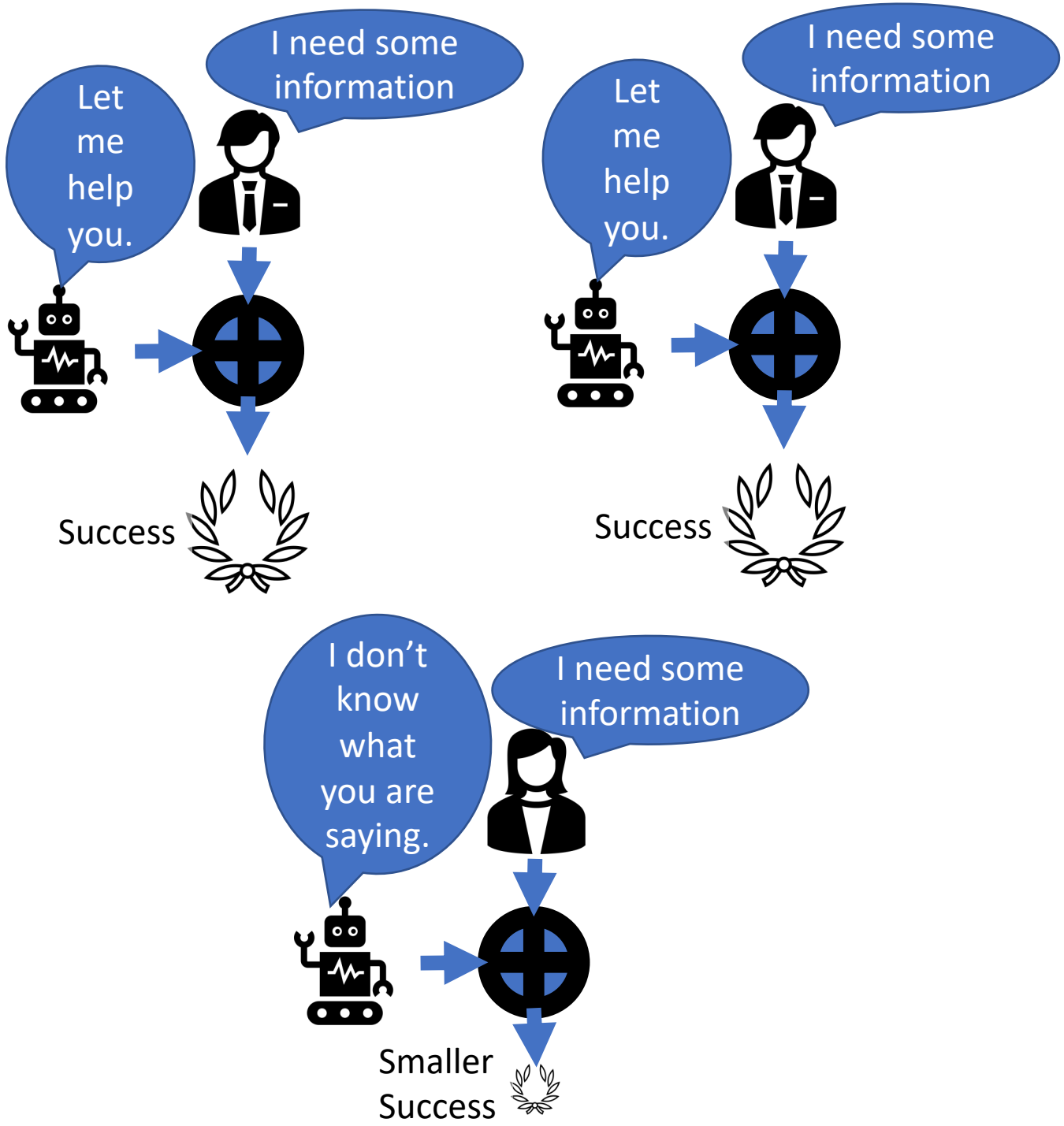
- Librispeech includes 360h clean speech, 500h other speech
- Curated from librivox.org and Gutenberg.org
- Free to download, free to use and redistribute
- Librispeech is the reason for the deep-learning revolution in automatic speech recognition.

Outline

- The many dimensions of ability and disability
- Speech technology: the Transformer
- Speech technology: Pre-training and Fine-tuning
- Speech technology: Audiobooks
- Inclusive speech recognition
- The Speech Accessibility Project
- Disability in the public sphere

Inclusive ASR: why?

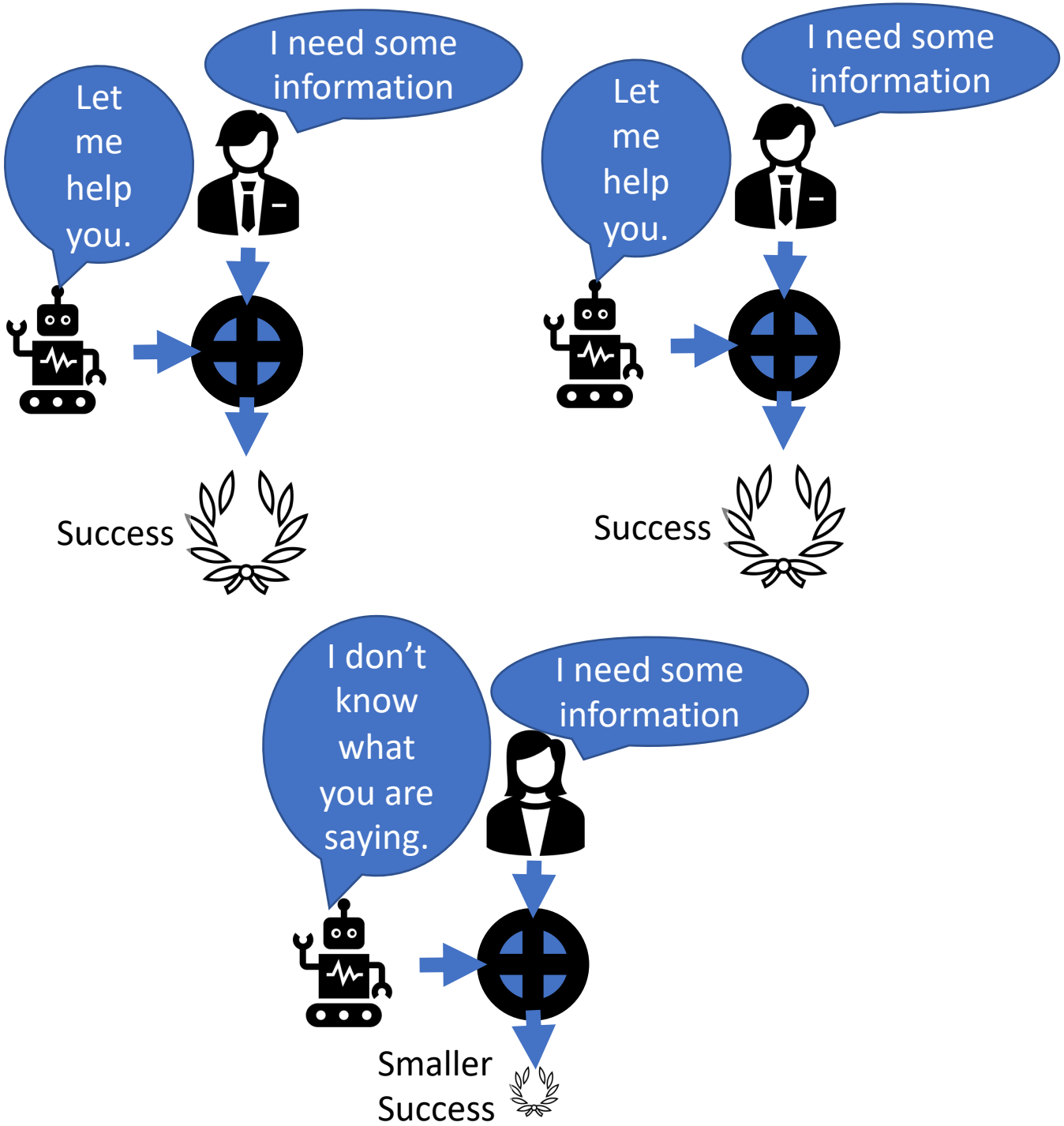
- ASR is a useful productivity tool
- If ASR works for you, it gives you a socioeconomic advantage, compared to somebody for whom ASR fails
- The people for whom ASR fails are often those who are ***already*** socioeconomically disadvantaged



Why does ASR fail?

Reasons it may fail for a group:

- **<u>Under-representation</u>**: The training corpus doesn't have enough examples
- **<u>Inter-group variance:</u>** The group speaks differently from other groups
- **<u>Intra-group variance:</u>** Members of the group all speak differently from one another
- **<u>Intra-individual variance:</u>** Members of the group speak less precisely



Automatic Speech Recognition Word Error Rates (WER)

Gender:

- Women > Men (51% > 38%: Tatman, 2017, YouTube captioning)
- Women > Men (61% > 47%: Garnerin et al., 2019, European broadcast news)
- Black Men > Black Women (41% > 30%: Koenecke et al., 2020)

Dialect:

- Scottish > American (53% > 42%: Tatman, 2017)
- American Deep South > General American (Picone 1991)

Race:

- Black > White (35% > 19%: Koenecke: avg of Amazon, Apple, Google, IBM, Microsoft)

Disability:

- People w/Cerebral Palsy > People w/o (41% > 33%: Issa et al., in review)

Age:

- Teenage (<20)) > Old (>70) > Adult (Feng & Scharenborg, 2020)
- Old adult (50-70) < Young adult (20-30) (Sari et al., 2021)

Database Collection Efforts

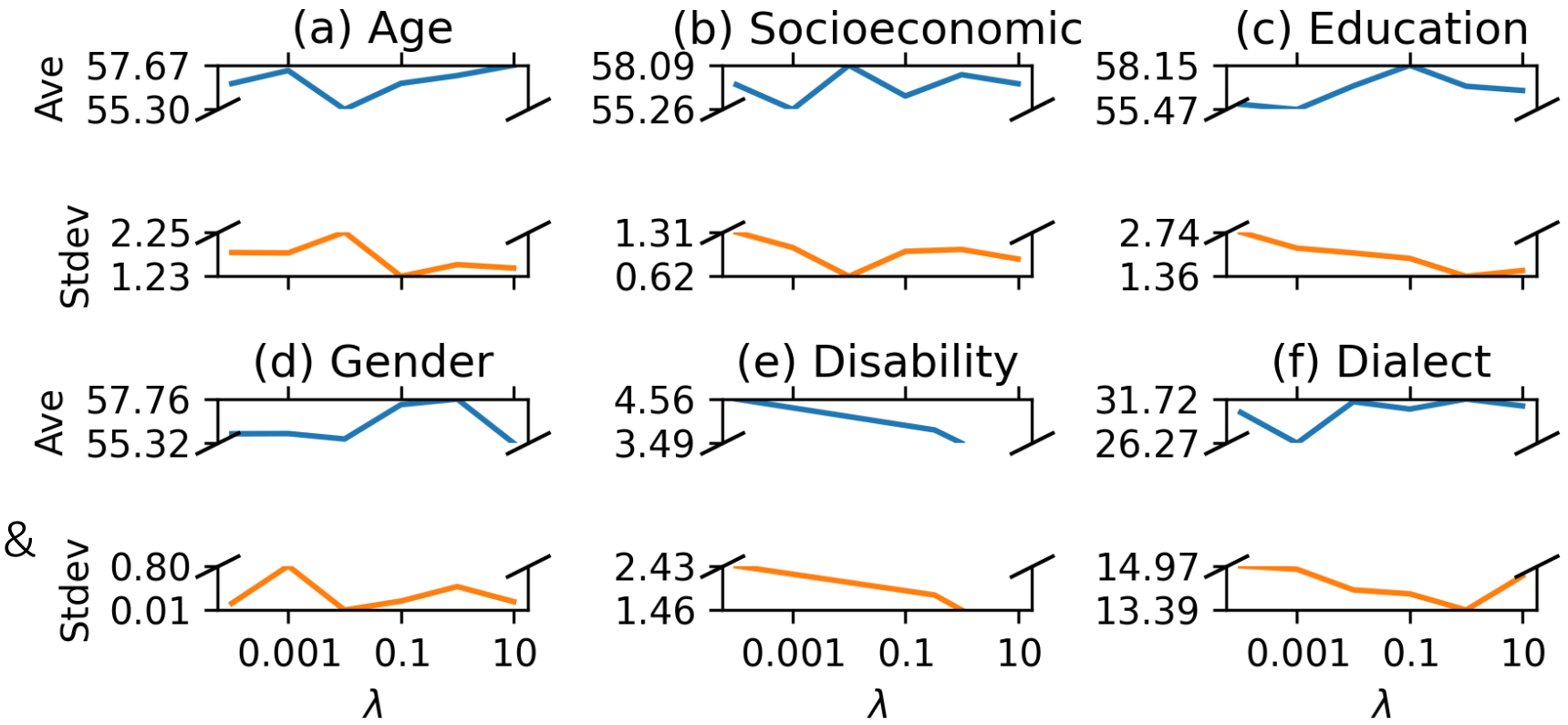
- UASpeech: dysarthria as a symptom of Cerebral Palsy (Kim et al., 2008)
  - Perhaps 3.2 hours
- CORAAL (Corpus of Regional African American Languages, Kendall et al., 2018)
  - Perhaps 200 hours

… but …

- Librispeech: largest publicly available corpus of General American English speech:
  - 1000 hours
- Librivox Complete (Baevski et al., 2020):
  - 10,000 hours

Seamless equal accuracy ratio for inclusive CTC speech recognition

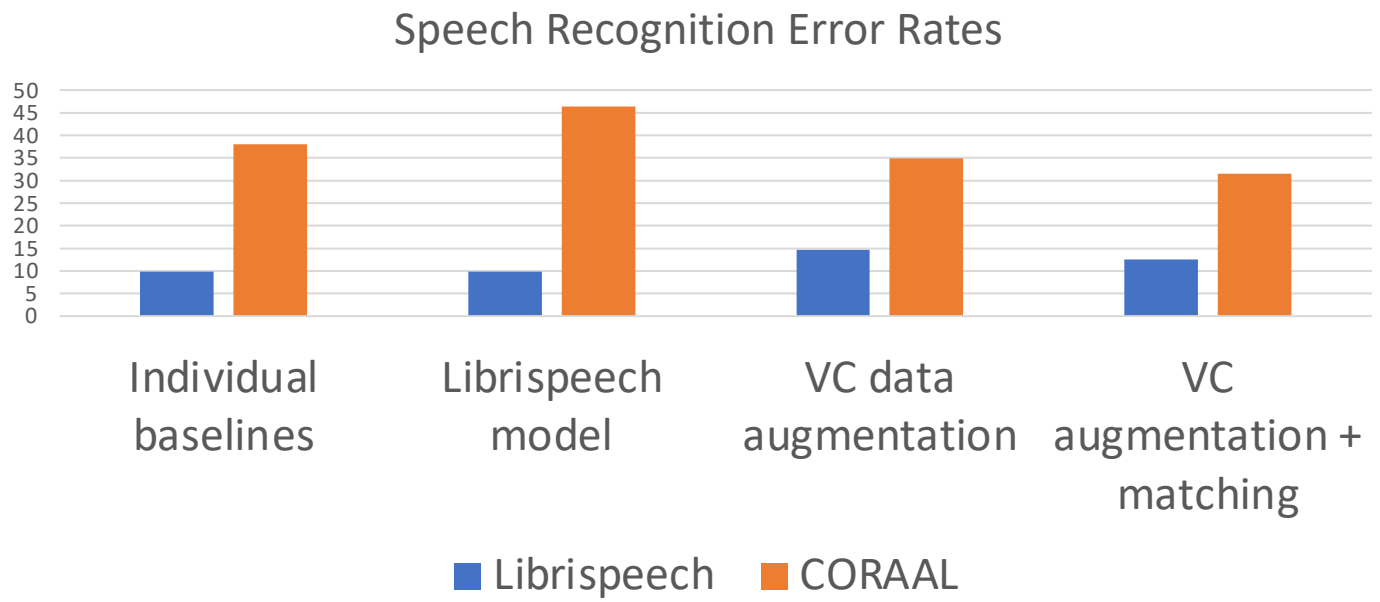
Gao, Wang, Kang, Mina, Issa, Harvill, Sarı, Hasegawa-Johnson & Yoo, 2020



- $\lambda = 0$: train speech recognizer to reduce average error rate
- $\lambda = 1$: train it to reduce the maximum error rate incurred by any group
- As λ gets larger, standard deviation of error rate (among speakers) decreases. This was expected.
- As λ gets larger, the average error rate (averaged across speakers) also decreases. This was unexpected. Apparently, focusing on the most poorly-served group can help to improve the error rate for all groups.

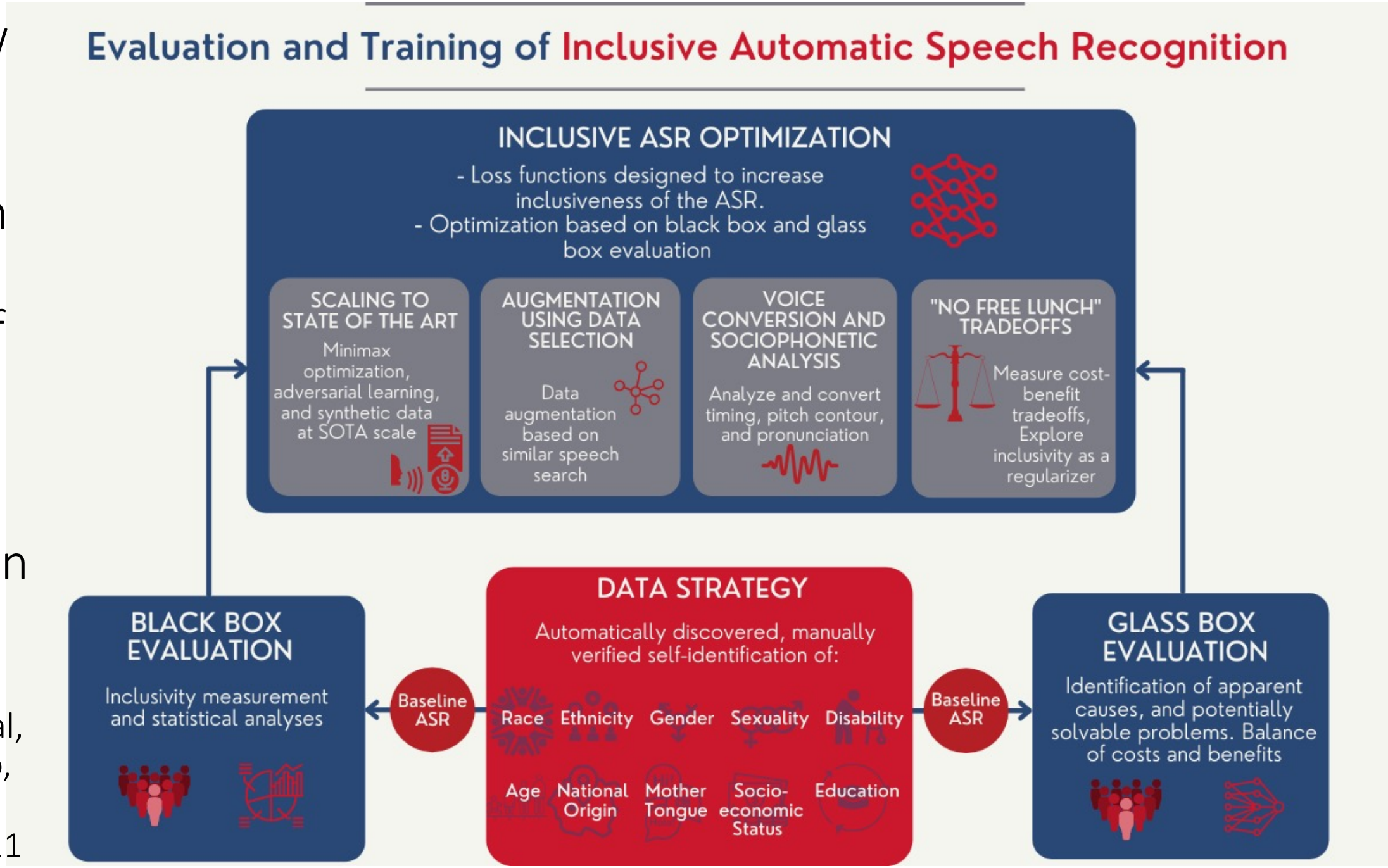
Counterfactually fair automatic speech recognition

Sarı, Hasegawa-Johnson & Yoo, 2021



- Librispeech = audiobooks, 1000h; plurality are “General American English” (GAE)
- CORAAL = corpus of regional African American Languages (AAL), 200h, autobiographical narratives from Atlanta, DC, Detroit, Manhattan, Princeville, Rochester, Valdosta
- Applying the Librispeech model to CORAAL data is worse than using the CORAAL-only baseline
- Synthesizing new data using voice conversion (VC) reduces error rates on CORAAL
- Counterfactual matching forces the transcriptions for a given speaker to remain unchanged even after the voice is converted from AAL to GAE or vice versa

FAI: A new paradigm for the evaluation and training of inclusive automatic speech recognition



Hasegawa-Johnson, Fagyal, Dehak, Zelasko, and Moro-Velazquez, 2021

Outline

- The many dimensions of ability and disability
- Speech technology: the Transformer
- Speech technology: Pre-training and Fine-tuning
- Speech technology: Audiobooks
- Inclusive speech recognition
- The Speech Accessibility Project
- Disability in the public sphere

The Speech Accessibility Project

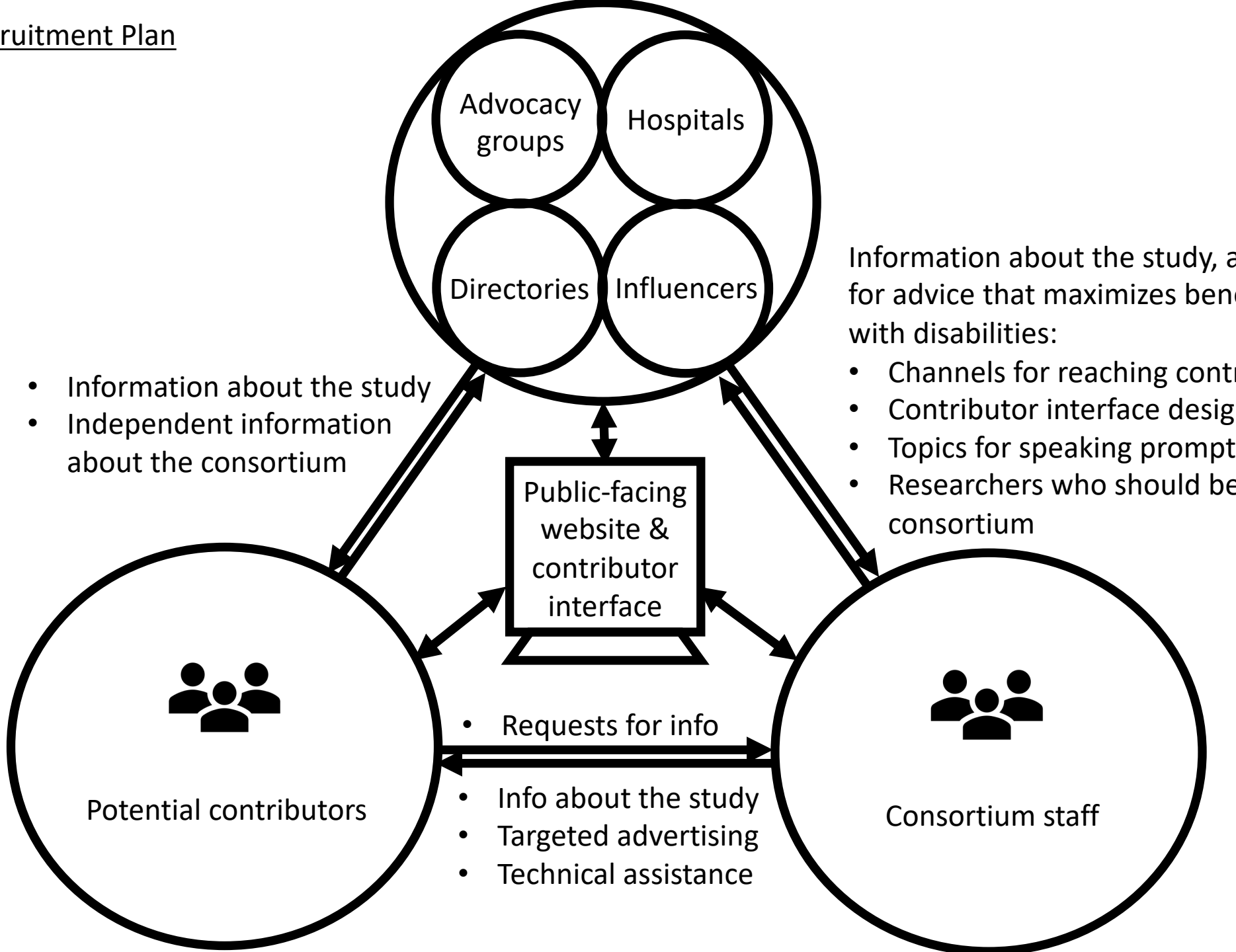
- Speech technology is now good enough to be useful for people speaking “General American English” without disabilities.
- To make it useful for people with disabilities, we need about 1000 hours of transcribed speech (~1.2 million sentences).

The Speech Accessibility Project is intended as a communication channel that will connect people with disabilities with the engineers and computer scientists who create speech technology.

Etiologies

- Etiology = cause of the disability
- Different etiologies have different speech patterns
- Having many different etiologies is desirable (corpus includes speech representative of different etiologies)
- Having many speakers for each etiology is also desirable (speech patterns within each group may be quite variable)
- Our compromise: we intend to recruit 400 speakers from each of 5 different etiologies, starting with Parkinson's, ALS, and Cerebral Palsy.

Outreach & Recruitment Plan



Human subject protection principles

Contributors should know how their data will be used; Researchers should commit to the same terms.

- The purpose of the study and permitted uses, as specified in the participant consent form,
- …should be the same as the Permitted Purpose and permitted uses in the data use agreement.

IRB APPROVED
IRB Number: 23183
IRB Approval Date: 03/07/2023
IRB Expiration Date: 03/02/2028

University of Illinois Urbana-Champaign
Online Consent Form

Speech Accessibility Project: Individuals with disabilities helping researchers to improve technology

You are being asked to participate in a voluntary research study that is called the "Speech Accessibility Project." The purpose of this study is to help researchers at universities and companies to develop spoken-language user interfaces that work for people with atypical speech. Software programs that understand speech are developed using machine learning. Machine learning is a software development method that imitates the way human
behavior whenever it makes a mistak
mistakes in the future. By providing e
possible for the software to learn hov
the future. Participating in this study
your audio recordings to a secure enc
samples or about 5 hours of your tim
convenience. Risks associated with th
expected benefit of this research to y
and organizations using your speech
with speech and motor disabilities, w

Project Name: The Speech Accessibili
https://speechaccessibilityproject.be

LICENSED SPEECH SIGNAL DATA
DATA USE TERMS AND CONDITIONS

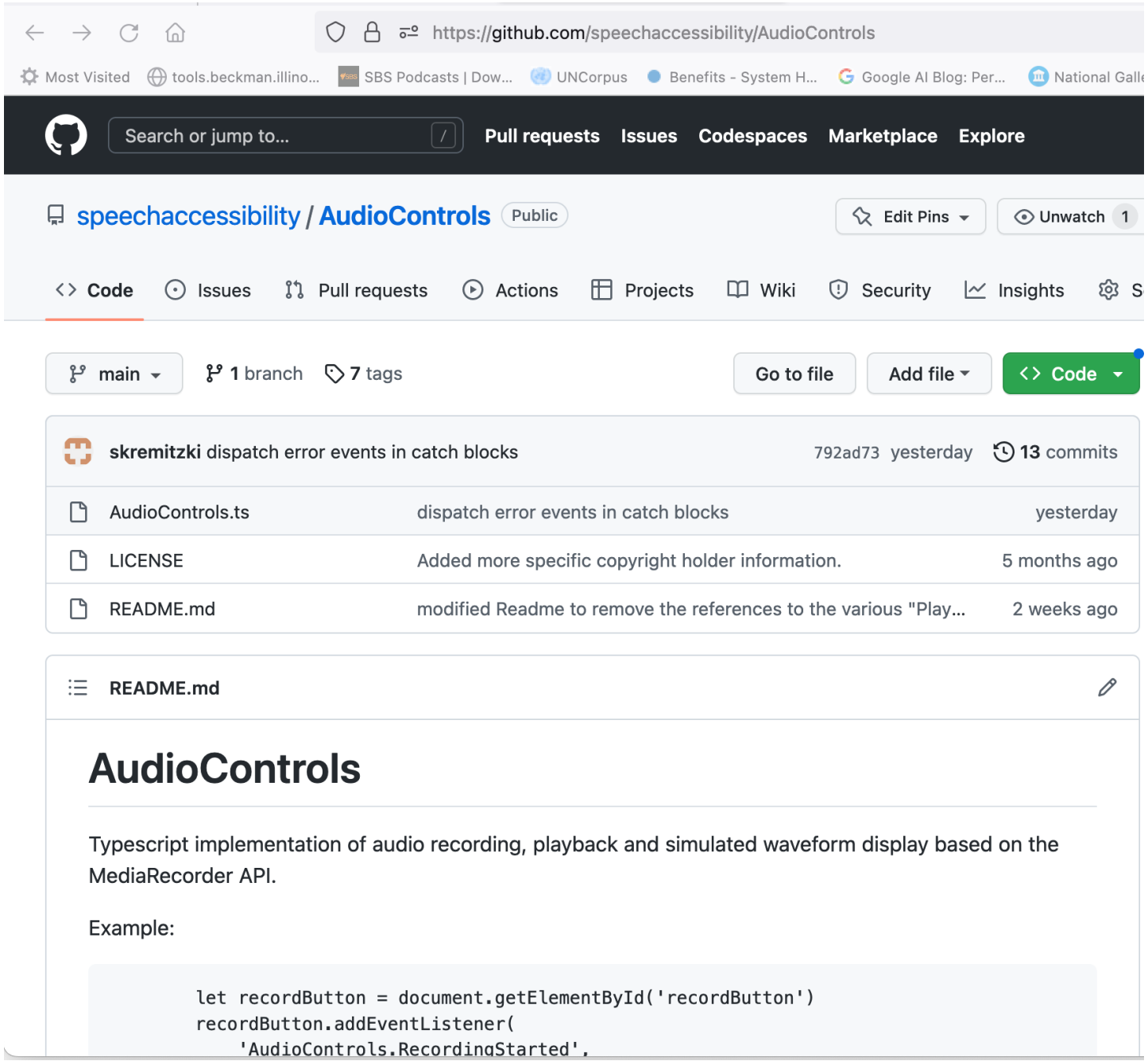
The licenses to Licensed Speech Signal Data granted to each Member pursuant to Section 9.b. of the Agreement are subject to the following Terms and Conditions. Each Member receiving such a license is referred to herein as the "**Data User**". Any capitalized terms used but not otherwise defined in this Exhibit C shall have the meanings given to them elsewhere in the Agreement.

1. Provision and Use of Data

1.1 Effective as of the Effective Date, Data User may make, have made, use, load, access, store, copy, reproduce, destroy, modify, transmit, display, make derivative works of and otherwise use the Data made available to it by Organization under the Agreement on a non-exclusive, non-assignable, non-sublicensable (except as provided in Section 2.1 below) basis solely for the Permitted Purpose subject to the terms of this Exhibit C and the Agreement. Any such modifications or derivative works made by Data User are "**Modification(s)**".

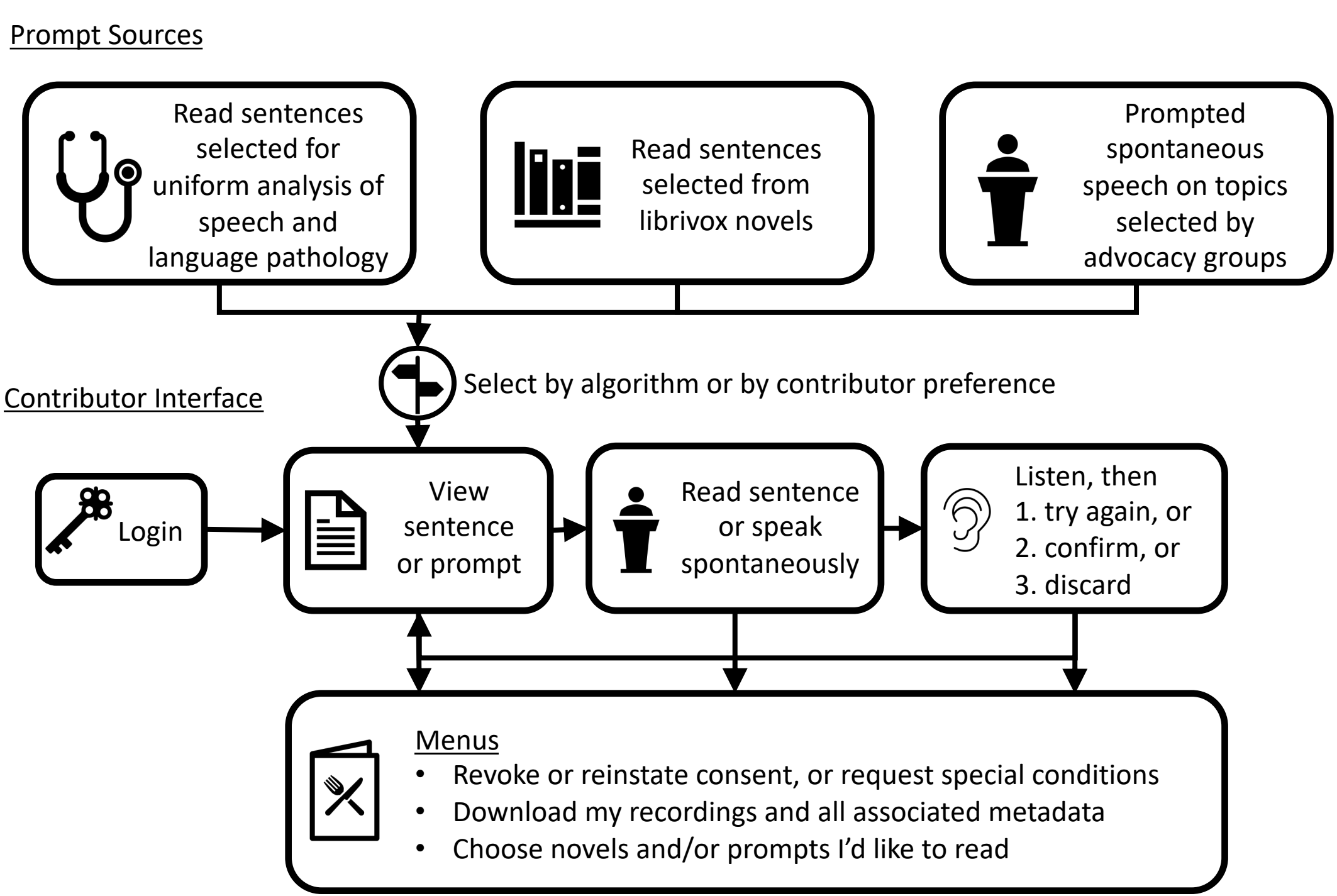
1.2 This Agreement does not restrict Data User's use or modification of any portions of the Data that Organization makes publicly available under a more permissive license, if applicable, to the extent Data User uses or modifies the Data under the terms of such public license, or that otherwise become public.

1.3 The rights and restrictions set forth in this Exhibit C as it relates to the use, modification, distribution, disclosure, and privacy of the Data may only be modified by a written agreement between Data User and Organization specifically indicating it is amending these Terms and Conditions, e.g., these Terms and Conditions may not be amended by or superseded by any general terms and conditions that are presented or included as part of access to or use of data storage, data processing, platform, or computing resources.



https://github.com/speechaccessibility/AudioControls

- Software design principles:
  - Contributor safety
  - Data recoverability
  - Contributor ease of use
  - Annotator ease of use
- Contributor and Annotator user interfaces go through dev and test environments before being pushed to production.
- All finished products are released open-source.


Prompt Sources
Read sentences selected for uniform analysis of speech and language pathology
Read sentences selected from librivox novels
Prompted spontaneous speech on topics selected by advocacy groups
Select by algorithm or by contributor preference
Contributor Interface
Login
View sentence or prompt
Read sentence or speak spontaneously
Listen, then
1. try again, or
2. confirm, or
3. discard
Menus
• Revoke or reinstate consent, or request special conditions
• Download my recordings and all associated metadata
• Choose novels and/or prompts I'd like to read

Types of speech, Transcription quality

- Speech recognition is most robust if training samples are as natural and spontaneous as possible.

…however…

- Read speech is easier to transcribe: We know what the person was trying to say, so “transcribing” just means correcting the prompt text so it matches what they actually said.
- Our compromise:
  - 300 command & control sentences (“find my phone,” “What’s on my shopping list?”)
  - 100 phonetically diverse sentences (“When I was a young boy, father always said I was a born business manager.”)
  - 200 spontaneous sentences, in response to 50 prompts for spontaneous speech ("Tell us about one of your favorite bands or singers.”)

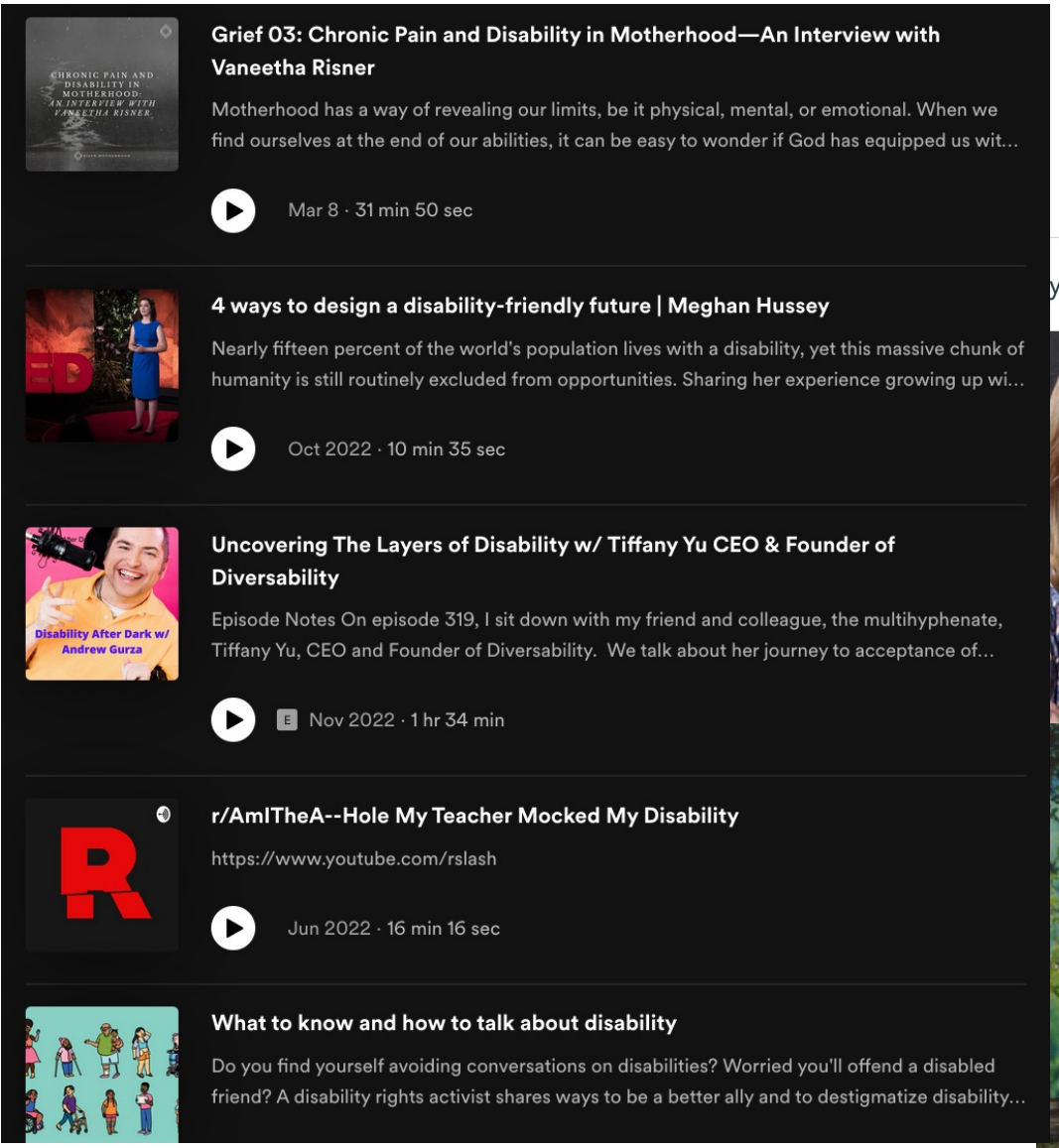
Distribution to researchers

- The data distribution includes:
  - Contributed audio
  - Prompt text, Corrected text,
  - Annotations of dysarthria or dysphonia type and severity
  - Start time & End time of speech, signal to noise ratio, clipping severity
- Data is packaged into an encrypted tar file
- Each consortium member (each organization that has signed a data use agreement) can specify one person who gets a decryption key
- Consortium members will begin testing data quality & attempting to train ASR as soon as we begin distributing the data.

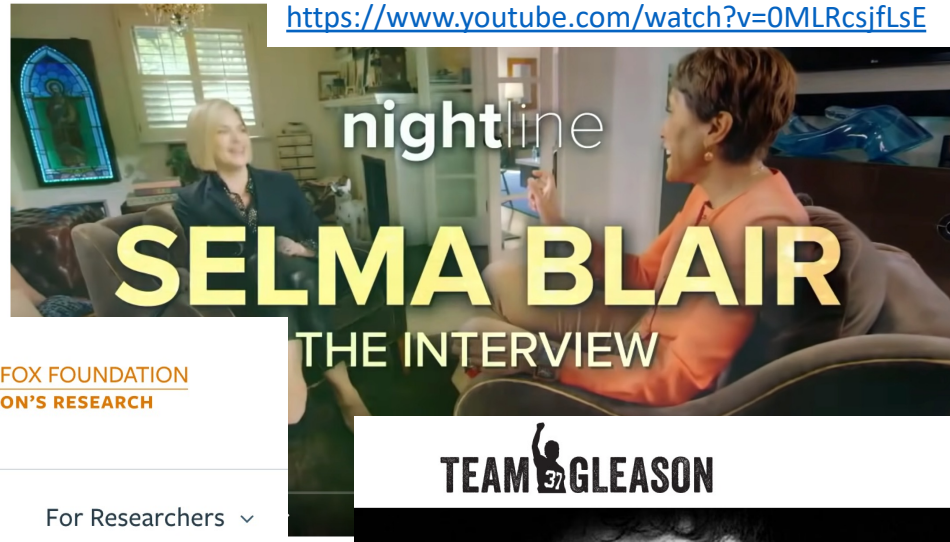
Outline

- The many dimensions of ability and disability
- Speech technology: the Transformer
- Speech technology: Pre-training and Fine-tuning
- Speech technology: Audiobooks
- The Speech Accessibility Project
- Inclusive speech recognition
- Disability in the public sphere

Disabled in public

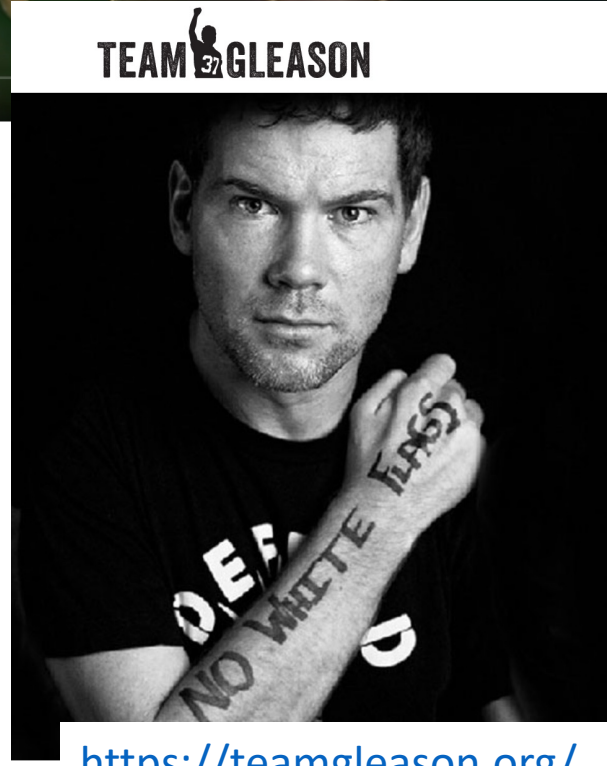


https://www.youtube.com/watch?v=0MLRcsjfLsE



https://www.michaeljfox.org/our-promise

https://commons.wikimedia.org/wiki/File:Davis_Phinney_1991_Thrift_Drug_Classic.jpg



https://teamgleason.org/

The growing population of Americans with disabilities

- The number of people diagnosed with Parkinson's Disease in the United States is expected to double by 2040. – NINDS
- The estimated number of people with Multiple Sclerosis … is 30% higher than in 2013. – Walton et al., 2020
- By our estimates, the number of cases of ALS in the world will increase from 222,801 in 2015 to 376,674 in 2040 – Arthur et al., 2016
- In the United States, about 85% of stroke victims now survive. – Donkor, 2018
- In 1900 the life expectancy of people with Down Syndrome was 9 years. By 1984, [it] had increased to 28 years. Since then, the life expectancy of people with DS has increased to about 60 years. – Chicoine, 2021

Summary

- Speech technology benefits from psychologically-inspired innovations in machine learning.
- Speech technology benefits from community involvement.
- Physical differences mean disability only if the technology fails to work.
- The Speech Accessibility Project aims to make the technology work.