Lecture 1

Introduction/Signal Processing, Part I

Michael Picheny, Bhuvana Ramabhadran, Stanley F. Chen

IBM T.J. Watson Research Center
Yorktown Heights, New York, USA
{picheny,bhuvana,stanchen}@us.ibm.com

10 September 2012

Part I

Introduction

What Is Speech Recognition?

- Converting speech to text (STT).
  - a.k.a. automatic speech recognition (ASR).
- What it's not.
  - *Natural language understanding* — *e.g.*, Siri.
  - *Speech synthesis* — converting text to speech (TTS), *e.g.*, Watson.
  - *Speaker recognition* — identifying who is speaking.

Why Is Speech Recognition Important?

- Demo.

Because It's Fast

modality	*method*	*rate (words/min)*
sound	speech	150–200
sight	sign language; gestures	100–150
touch	typing; mousing	60
taste	covering self in food	<1
smell	not showering	<1

Other Reasons

- Requires no specialized training to do fast.
- Hands-free.
- Speech-enabled devices are everywhere.
  - Phones, smart or dumb.
  - Access to phone $>$ access to internet.
- Text is easier to process than audio.
  - Storage/compression; indexing; human consumption.

Key Applications

- Transcription: archiving/indexing audio.
  - Legal; medical; television and movies.
  - Call centers.
- Whenever you interact with a computer . . .
  - Without sitting in front of one.
  - *e.g.*, smart or dumb phone; car; home entertainment.
- Accessibility.
  - People who can't type, or type slowly.
  - The hard of hearing.

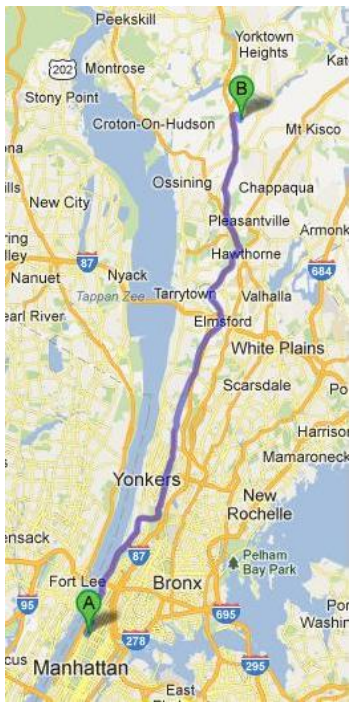
Why Study Speech Recognition?

- Real-world problem.
  - Potential market: ginormous.
- Hasn't been solved yet.
  - Not too easy; not too hard (*e.g.*, vision).
- Lots of data.
  - One of first learning problems of this scale.
- Connections to other problems with sequence data.
  - Machine translation, bioinformatics, OCR, etc.

Where Are We?

Who Are We?

- Michael Picheny: Sr. Manager, Speech and Language.
- Bhuvana Ramabhadran: Manager, Acoustic Modeling.
- Stanley F. Chen: Regular guy.
- IBM T.J. Watson Research Center, Yorktown Heights, NY.



Why Three Professors?

- Too much knowledge to fit in one brain.
  - Signal processing.
  - Probability and statistics.
  - Phonetics; linguistics.
  - Natural language processing.
  - Machine learning; artificial intelligence.
  - Automata theory.

How To Contact Us

- **In E-mail, prefix subject line with "EECS E6870:"!!!.**
  - Michael Picheny — `picheny@us.ibm.com`.
  - Bhuvana Ramabhadran — `bhuvana@us.ibm.com`.
  - Stanley F. Chen — `stanchen@us.ibm.com`.
- Office hours: right after class.
  - Before class by appointment.
- TA: Xiao-Ming Wu — `xw2223@columbia.edu`.
- Courseworks.
  - For posting questions about labs.

Course Outline

week	topic	assigned	due
1	Introduction		
2	Signal processing; DTW	lab 1	
3	Gaussian mixture models		
4	Hidden Markov models	lab 2	lab 1
5	Language modeling		
6	Pronunciation modeling	lab 3	lab 2
7	Finite-state transducers		
8	Search	lab 4	lab 3
9	Robustness; adaptation		
10	Discrim. training; ROVER	project	lab 4
11	Advanced language modeling		
12	Neural networks; DBN's.		
13	Project presentations		project

Programming Assignments

- 80% of grade ($\surd-$, $\surd$, $\surd+$ grading).
- Some short written questions.
- Write key parts of basic large vocabulary continuous speech recognition system.
  - Only the “fun” parts.
  - C++ code infrastructure provided by us.
  - Also accessible from Java (via SWIG).
- Get account on ILAB computer cluster (x86 Linux PC’s).
  - Complete the survey.
- Labs due at Wednesday 6pm.

Final Project

- 20% of grade.
- Option 1: Reading project (individual).
  - Pick paper(s) from provided list, or propose your own.
  - Give 10-minute presentation summarizing paper(s).
- Option 2: Programming/experimental project (group).
  - Pick project from provided list, or propose your own.
  - Give 10-minute presentation summarizing project.

Readings

- PDF versions of readings will be available on the web site.
- Recommended text:
  - *Speech Synthesis and Recognition*, Holmes, 2nd edition (paperback, 256 pp., 2001) **[Holmes]**.
- Reference texts:
  - *Theory and Applications of Digital Signal Processing*, Rabiner, Schafer (hardcover, 1056 pp., 2010) **[R+S]**.
  - *Speech and Language Processing*, Jurafsky, Martin (2nd edition, hardcover, 1024 pp., 2000) **[J+M]**.
  - *Statistical Methods for Speech Recognition*, Jelinek (hardcover, 305 pp., 1998) **[Jelinek]**.
  - *Spoken Language Processing*, Huang, Acero, Hon (paperback, 1008 pp., 2001) **[HAH]**.

Web Site

`www.ee.columbia.edu/~stanchen/fall12/e6870/`

- Syllabus.
- Slides from lectures (PDF).
  - Online by 8pm the night before each lecture.
  - Hardcopy of slides distributed at each lecture?
- Lab assignments (PDF).
- Reading assignments (PDF).
  - Online by lecture they are assigned.
  - Username: *speech*, password: *pythonrules*.

Prerequisites

- Basic knowledge of probability and statistics.
- Fluency in C++ or Java.
- Basic knowledge of Unix or Linux.
- Knowledge of digital signal processing optional.
  - Helpful for understanding signal processing lectures.
  - Not needed for labs.

Help Us Help You

- Feedback questionnaire after each lecture (2 questions).
  - Feedback welcome any time.
- You, the student, are partially responsible . . .
  - For the quality of the course.
- Please ask questions anytime!
- EE's may find CS parts challenging, and vice versa.
- Together, we can get through this.
- Let's go!

Where Are We?

The Early Years: 1950–1960's

- *Ad hoc* methods.
  - Many key ideas introduced; not used all together.
  - *e.g.*, spectral analysis; statistical training; language modeling.
- Small vocabulary.
  - Digits; yes/no; vowels.
- Not tested with many speakers (usually $<$10).

Whither Speech Recognition?

Speech recognition has glamour. Funds have been available. Results have been less glamorous . . .

. . . General-purpose speech recognition seems far away. Special-purpose speech recognition is severely limited. It would seem appropriate for people to ask themselves why they are working in the field and what they can expect to accomplish . . .

. . . These considerations lead us to believe that a general phonetic typewriter is simply impossible unless the typewriter has an intelligence and a knowledge of language comparable to those of a native speaker of English . . .

—John Pierce, Bell Labs, 1969

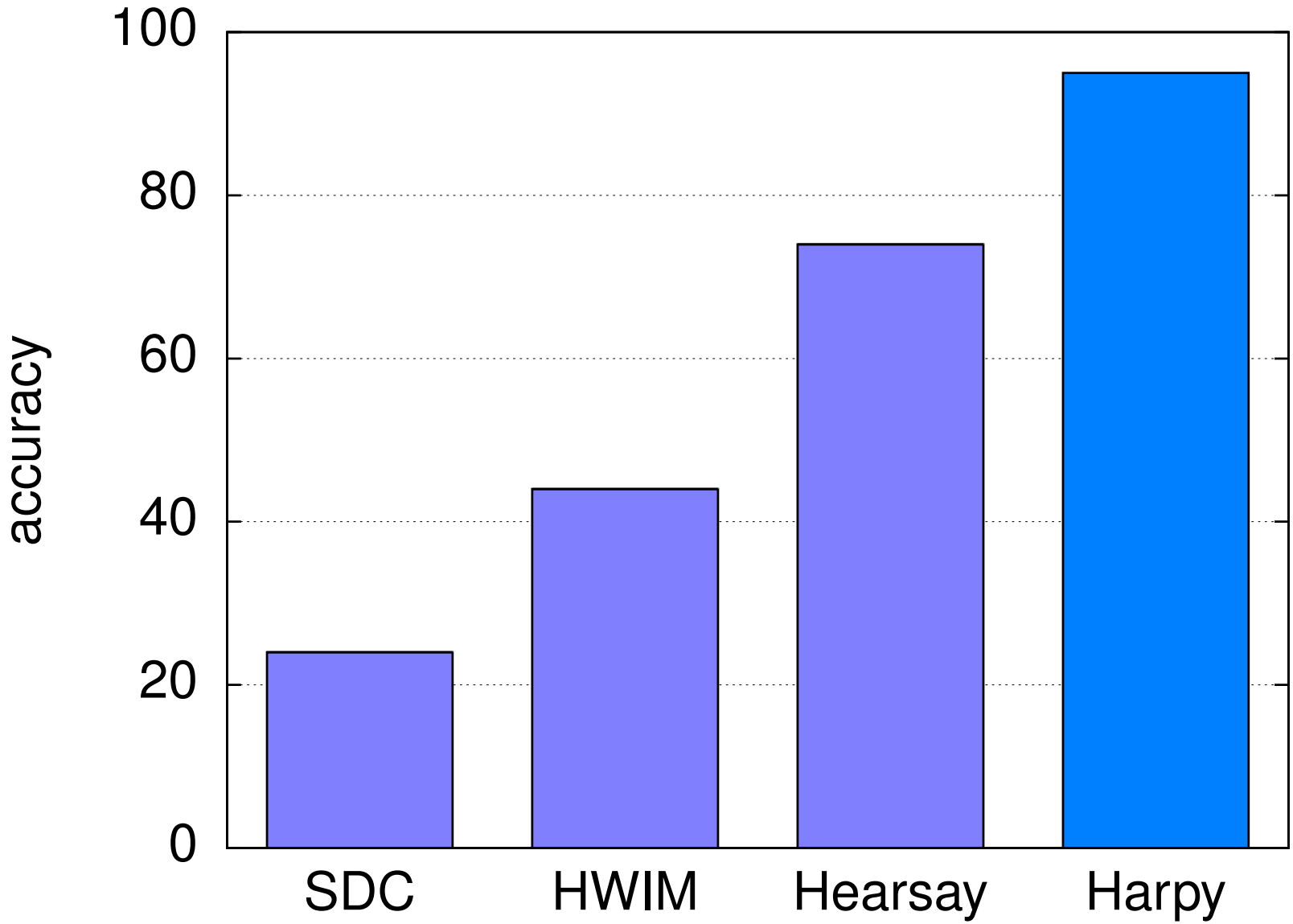
Whither Speech Recognition?

- Killed ASR research at Bell Labs for many years.
- Partially served as impetus for first (D)ARPA program (1971–1976) funding ASR research.
  - Goal: integrate speech knowledge, linguistics, and AI to make a breakthrough in ASR.
  - Large vocabulary: 1000 words.
  - Speed: a few times *real time*.

Knowledge-Driven or Data-Driven?

- Knowledge-driven.
  - People *know* stuff about speech, language,
  - *e.g.*, linguistics, (acoustic) phonetics, semantics.
  - Hand-derived rules.
  - Use expert systems, AI to integrate knowledge.
- Data-driven.
  - Ignore what we think we know.
  - Build dumb systems that work well if fed lots of data.
  - Train parameters statistically.

The ARPA Speech Understanding Project



*Each system graded on different domain.

The Birth of Modern ASR: 1970–1980's

Every time I fire a linguist, the performance of the speech recognizer goes up.

—Fred Jelinek, IBM, 1985(?)

- Ignore (almost) everything we know about phonetics, linguistics.
- View speech recognition as . . . .
  - Finding *most probable* word sequence given audio.
  - Train probabilities automatically w/ transcribed speech.

The Birth of Modern ASR: 1970–1980's

- Many key algorithms developed/refined.
  - Expectation-maximization algorithm; *n*-gram models; Gaussian mixtures; Hidden Markov models; Viterbi decoding; etc.
- Computing power still catching up to algorithms.
  - First real-time dictation system built in 1984 (IBM).
  - Specialized hardware $\approx$ 60 MHz Pentium.

The Golden Years: 1990's–now

	1984	now
CPU speed	60 MHz	3 GHz
training data	$<$10h	10000h+
output distributions	GMM*	GMM
sequence modeling	HMM	HMM
language models	*n*-gram	*n*-gram

- Basic algorithms have remained the same.
- Bulk of performance gain due to more data, faster CPU's.
- Significant advances in adaptation, discriminative training.
- New technologies (*e.g.*, Deep Belief Networks) on the cusp of adoption.

*Actually, 1989.

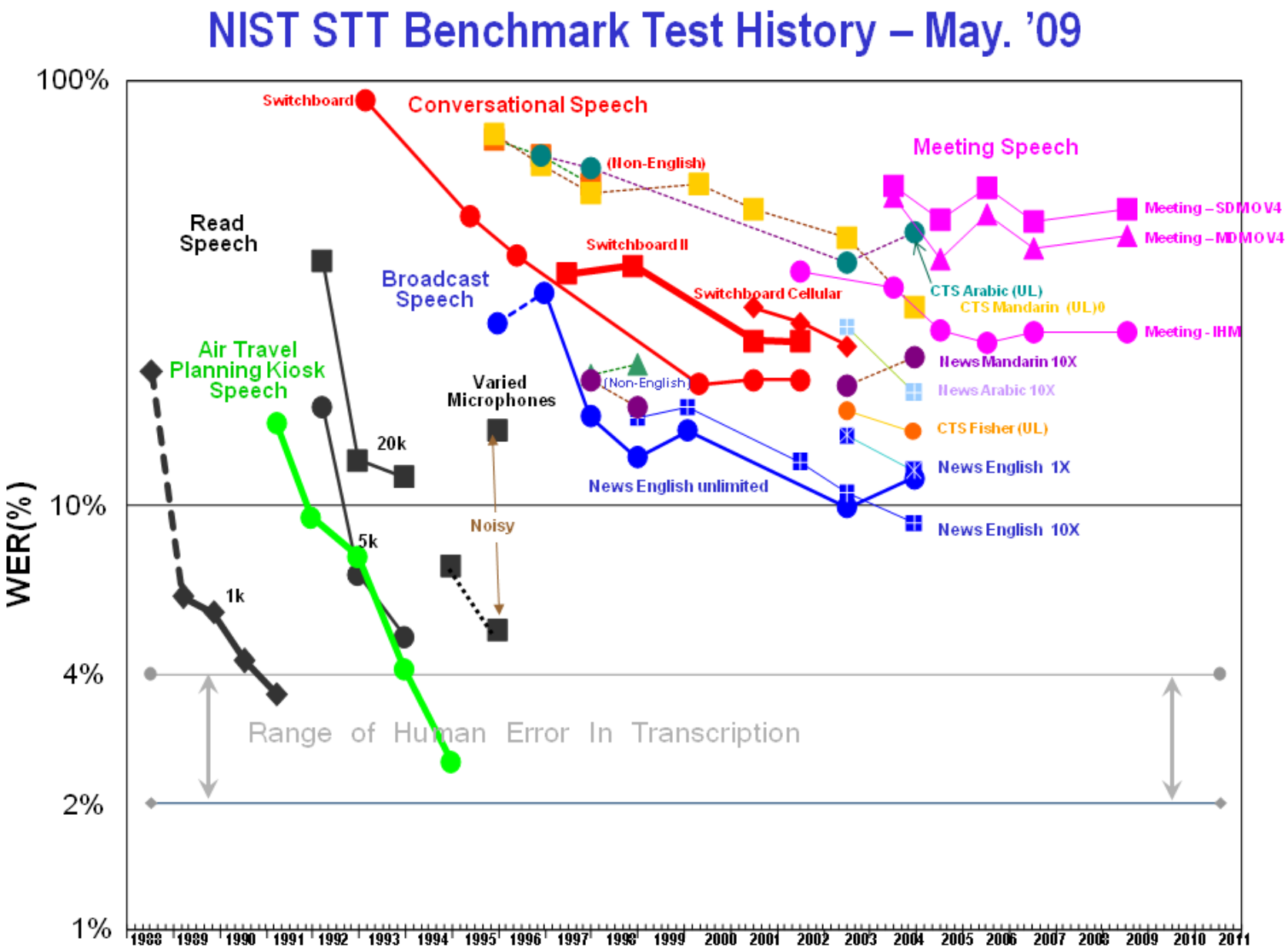
Not All Recognizers Are Created Equal

- Speaker-dependent *vs.* speaker-independent.
  - Need *enrollment* or not.
- Small *vs.* large vocabulary.
  - *e.g.*, recognize digit string *vs.* city name.
- Isolated *vs.* continuous.
  - Pause between each word or speak naturally.
- Domain.
  - *e.g.*, air travel reservation system *vs.* E-mail dictation.
  - *e.g.*, read *vs.* spontaneous speech.

Research Systems

- Driven by government-funded evaluations (DARPA, NIST).
  - Different sites compete on a common test set.
- Harder and harder problems over time.
  - Read speech: TIMIT; resource management (1kw vocab); Wall Street Journal (20kw vocab); Broadcast News (partially spontaneous, background music).
  - Spontaneous speech: air travel domain (ATIS); Switchboard (telephone); Call Home (accented).
  - Meeting speech.
  - Many, many languages: GALE (Mandarin, Arabic).
  - Noisy speech: RATS (Arabic).
  - Spoken term detection: Babel (Cantonese, Turkish, Pashto, Tagalog).

Research Systems



Man *vs.* Machine (Lippmann, 1997)

task	machine	human	ratio
Connected Digits[1]	0.72%	0.009%	80×
Letters[2]	5.0%	1.6%	3×
Resource Management	3.6%	0.1%	36×
WSJ	7.2%	0.9%	8×
Switchboard	43%	4.0%	11×

- For humans, one system fits all; for machine, not.
- Today: Switchboard WER $< 20\%$.

[1] String error rates.

[2] Isolated letters presented to humans; continuous for machine.

Commercial Speech Recognition

- Desktop.
  - 1995 — Dragon, IBM release speaker-dependent isolated-word large-vocabulary dictation systems.
  - Today — Dragon NaturallySpeaking: continuous-word; no enrollment required; “up to 99% accuracy”.
- Server-based; over the phone.
  - Late 1990’s — speaker-independent continuous-word small-vocabulary ASR.
  - Today — Google Voice Search, Dragon Dictate (demo): large-vocabulary; word error rate: top secret.

The Bad News

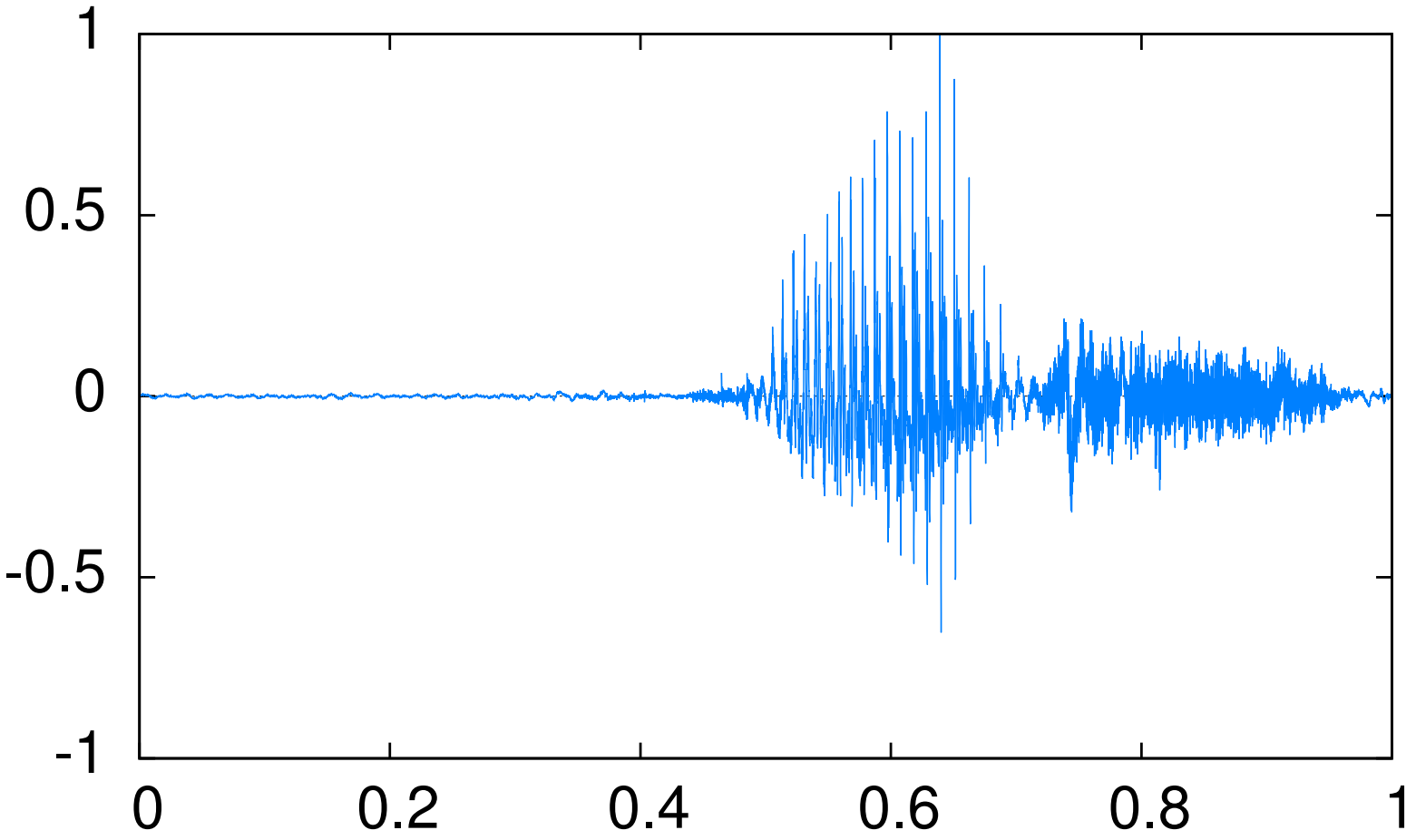
- Demo.
- Still a long way to go.

Where Are We?

The Data-Driven Approach

- Pretend we know nothing about phonetics, linguistics, ...
  - Treat ASR as just another machine learning problem.
- *e.g.*, *yes/no* recognition.
  - Person either says word *yes* or *no*.
- Training data.
  - One or more examples of each class.
- Testing.
  - Given new example, decide which class it is.

What is Speech?



- *e.g.*, turn on microphone for exactly one second.
- Microphone turns instantaneous air pressure into number.

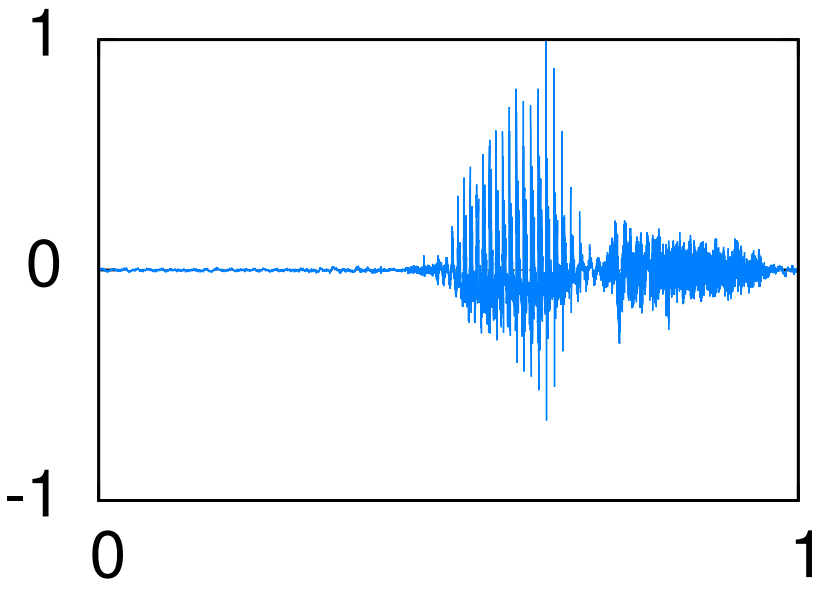
What is (Digitized) Speech?

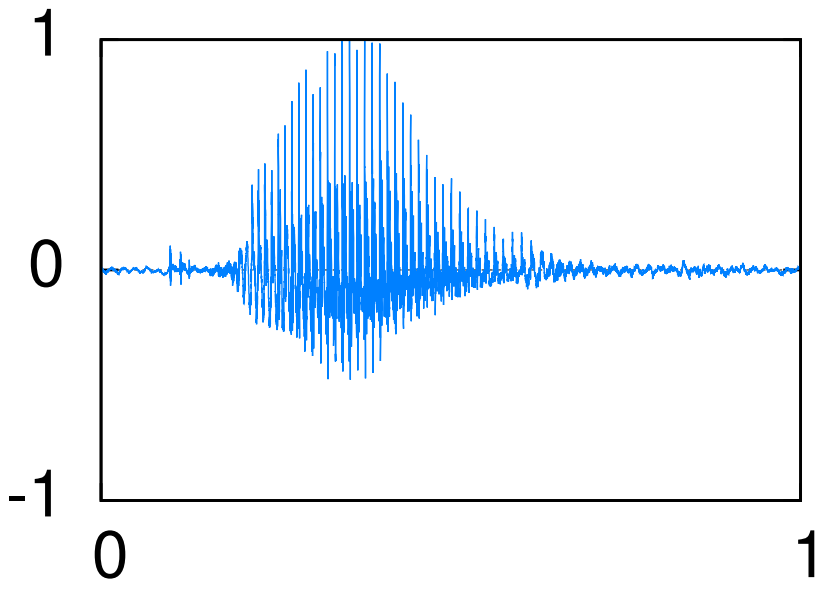
- Discretize in time.
  - Sampling rate, *e.g.*, 16000 samples/sec (Hz).
- Discretize magnitude (A/D conversion).
  - *e.g.*, 16-bit A/D $\Rightarrow$ value $\in [-32768, +32767]$.
- One second audio signal $A \in \mathcal{R}^{16000}$.
  - *e.g.*, [. . . , -0.510, -0.241, -0.007, 0.079, 0.071, . . .].

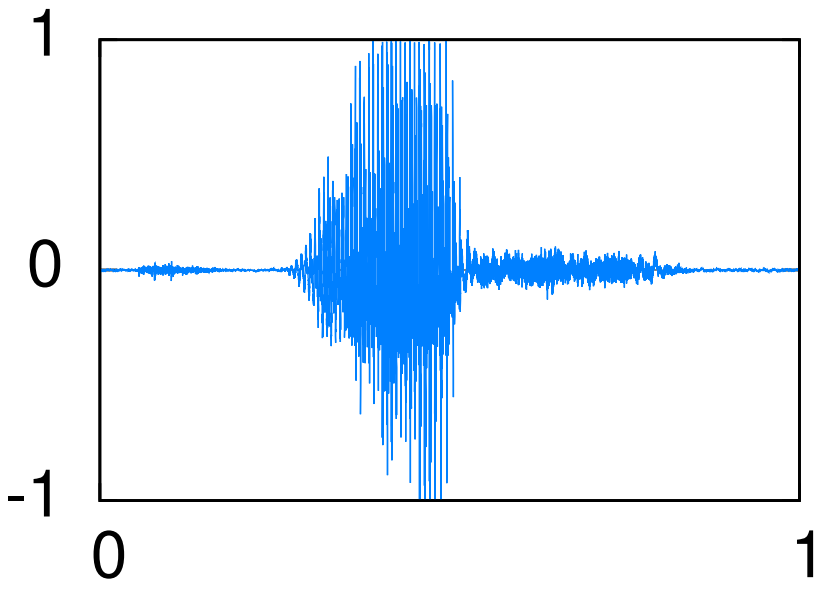
How Much Information Is Enough?

- Regenerate audio from digital signal.
  - If human can still understand, enough information?
- Demo.
  - 16k samples/sec; 16-bits per sample.
  - 2k samples/sec; 16-bits per sample.
  - 16k samples/sec; 1-bit per sample.

Example Training and Test Data







A Very Simple Speech Recognizer

- Audio examples $A_{\text{no}}, A_{\text{yes}}, A_{\text{test}} \in \mathcal{R}^{16000}$.
- Pick class $c^* \in \{\text{yes}, \text{no}\}$ = *vocabulary*:

$$c^* = \arg\min_{c \in \text{vocab}} \text{distance}(A_{\text{test}}, A_c)$$

- Which distance measure? Euclidean?

$$\text{distance}(A_{\text{test}}, A_c) = \sqrt{\sum_{i=1}^{16000} (A_i - A_{c,i})^2}$$

What's the Problem?

- Test set: 10 examples each of *yes*, *no*.
  - Error rate: 50%.
- This sucks.

The Challenge (Isolated Word ASR)

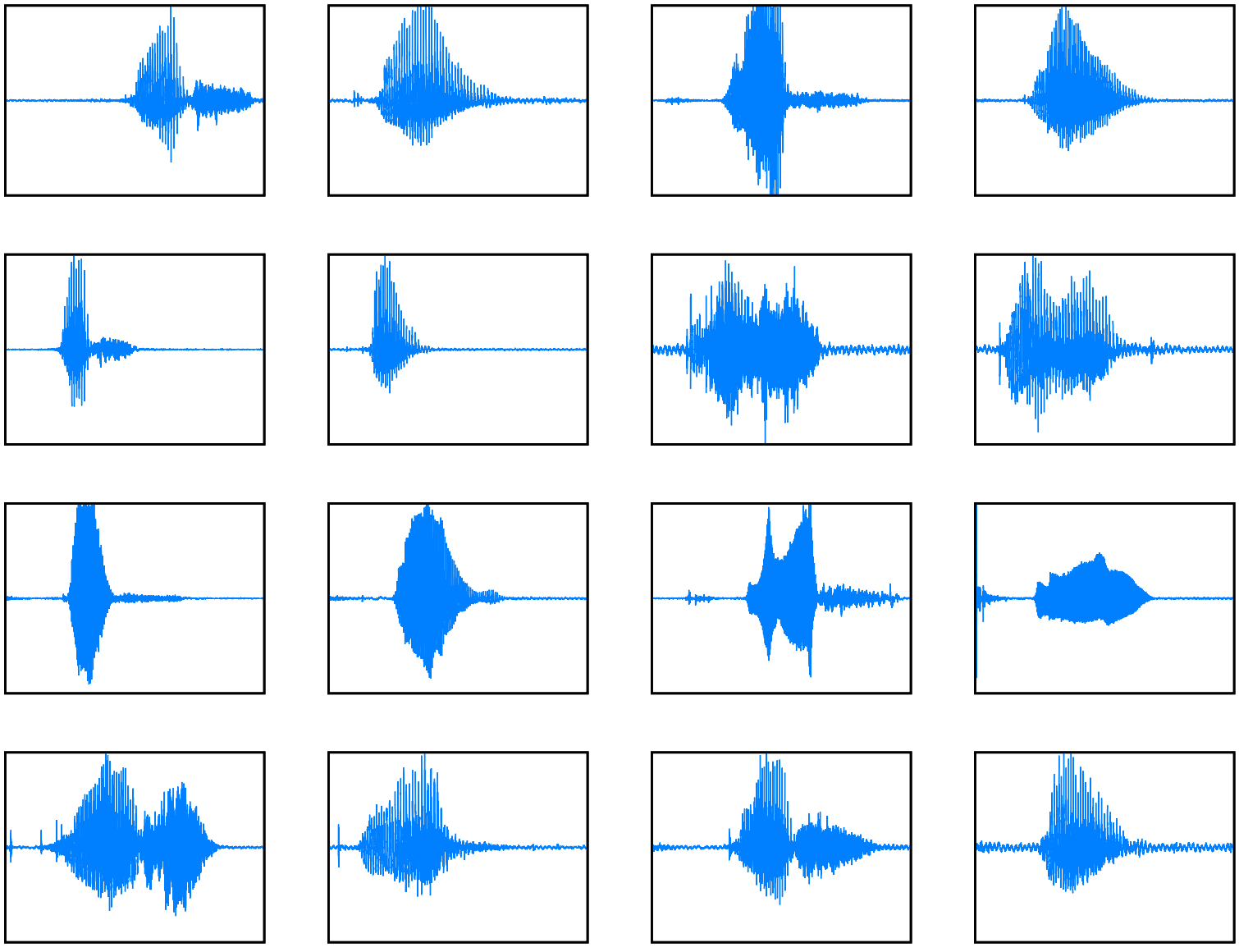
$$c^* = \arg\min_{c \in \text{vocab}} \text{distance}(A_{\text{test}}, A_c)$$

- Find good representation of audio $A \Rightarrow A'$ ...
  - So simple distance measure works.
- Also, find good distance measure.
- This turns out to be remarkably difficult!

Why Is Speech Recognition So Hard?

- There is enormous range of ways a word can be realized.
- Source variation.
  - Volume; rate; pitch; accent; dialect; voice quality (*e.g.*, gender, age); coarticulation; style (*e.g.*, spontaneous, read); . . .
- Channel variation.
  - Microphone; position relative to microphone (angle + distance); background noise; reverberation; . . .
- Screwing with any of these can make accuracy go to hell.

A Thousand Times No!



The First Two Lectures

$$c^* = \arg\min_{c \in \text{vocab}} \text{distance}(A_{\text{test}}, A_c)$$

- *signal processing* — Extract *features* from audio $A \Rightarrow A'$ . . .
  - That discriminate between different words.
  - Normalize for volume, pitch, voice quality, noise, ...
- *dynamic time warping* — Handling time/rate variation in the distance measure.

Where Are We?

Data-Driven *vs.* Knowledge-Driven

- Don't ignore *everything* we know about speech, language.

- Knowledge/concepts that have proved useful.
  - Words; phonemes.
  - A little bit of human production/perception.
- Knowledge/concepts that haven't proved useful (yet).
  - Nouns; vowels; syllables; voice onset time; . . .

Finding Good Features

- Extract features from audio . . .
  - That help determine word identity.
- What are good types of features?
  - Instantaneous air pressure at time t?
  - Loudness at time t?
  - Energy or phase for frequency ω at time t?
  - Estimated position of speaker's lips at time t?
- Look at human production and perception for insight.
  - Also, introduce some basic speech terminology.
- Diagrams from **[R+J]**, **[HAH]**.

Speech Production

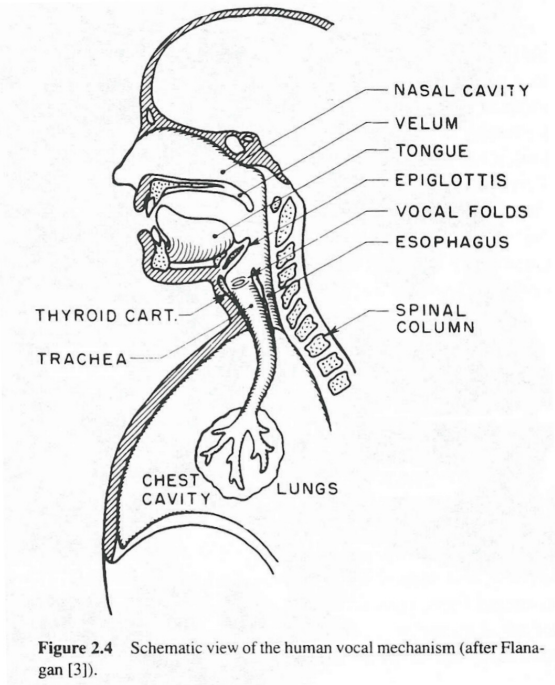


Figure 2.4 Schematic view of the human vocal mechanism (after Flanagan [3]).

- Air comes out of lungs.
- Vocal cords tensed (vibrate $\Rightarrow$ voicing) or relaxed (unvoiced).
- Modulated by vocal tract (glottis $\rightarrow$ lips); resonates.
  - Articulators: jaw, tongue, velum, lips, mouth.

Speech Consists Of a Few Primitive Sounds?

- Phonemes.
  - 40 to 50 for English.
  - Speaker/dialect differences.
  - *e.g.*, do MARY, MARRY, and MERRY rhyme?
  - Phone: acoustic realization of a phoneme.
- May be realized differently based on context.
  - *allophones*: different ways a phoneme can be realized.
  - *e.g.*, P in SPIN, PIN are two different allophones of P.

spelling	phonemes
SPIN	S P IH N
PIN	P IH N

  - *e.g.*, T in BAT, BATTER; A in BAT, BAD.

Classes of Speech Sounds

- Can categorize phonemes by how they are produced.
- Voicing.
  - *e.g.*, F (unvoiced), V (voiced).
  - All vowels are voiced.
- Stops/plosives.
  - Oral cavity blocked (*e.g.*, lips, velum); then opened.
  - *e.g.*, P, B (lips).

Classes of Speech Sounds

- Spectogram shows energy at each frequency over time.
- Voiced sounds have pitch (F0); formants (F1, F2, F3).
- Trained humans can do recognition on spectrograms with high accuracy (*e.g.*, Victor Zue).

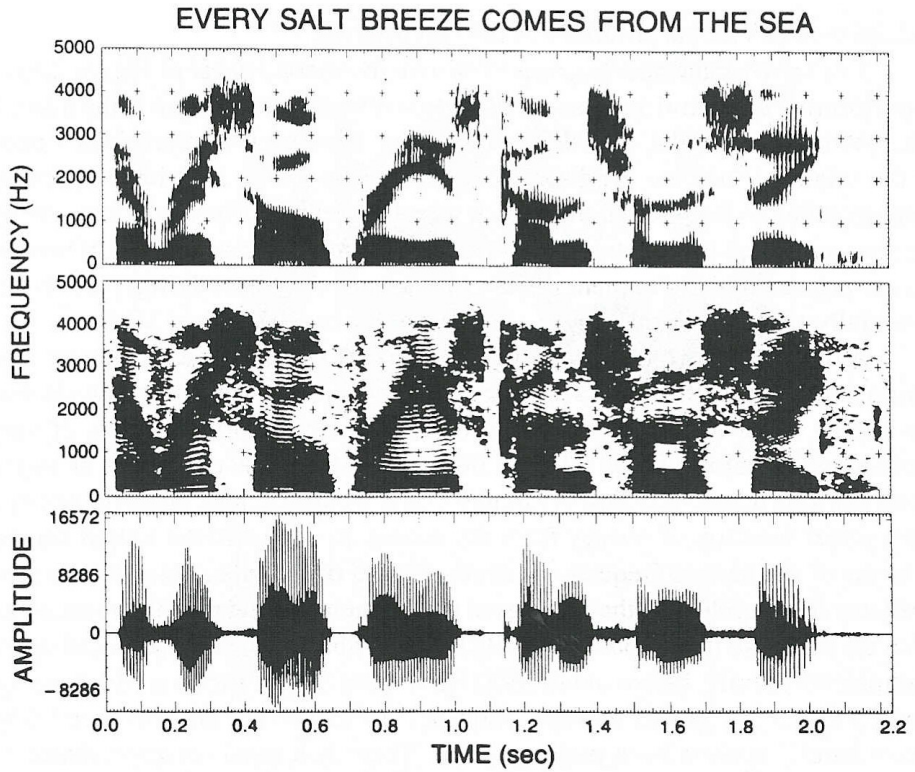
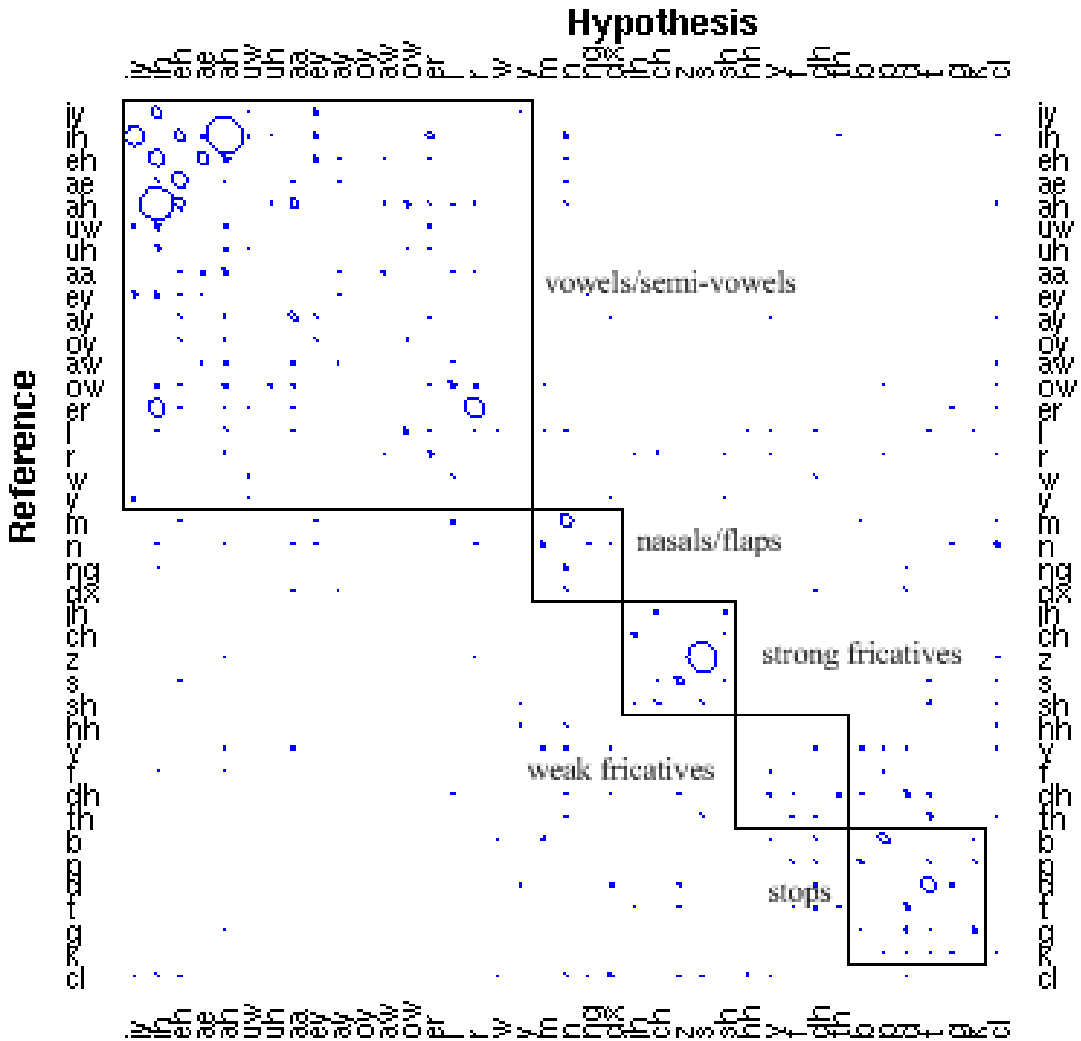


Figure 2.8 Wideband and narrowband spectrograms and speech amplitude for the utterance "Every salt breeze comes from the sea."

Classes of Speech Sounds

- What can the machine do? Here is a sample on TIMIT:



Classes of Speech Sounds

- Vowels — EE, AH, etc.
  - Differ in locations of formants.
  - Dipthongs — transition between two vowels (*e.g.*, COY, COW).
- Consonants.
  - Fricatives — F, V, S, Z, SH, J.
  - Stops/plosives — P, T, B, D, G, K.
  - Nasals — N, M, NG.
  - Semivowels (liquids, glides) — W, L, R, Y.

Coarticulation

- Realization of a phoneme can differ very much depending on context (allophones).
- Where articulators were for last phone affect how they transition to next.

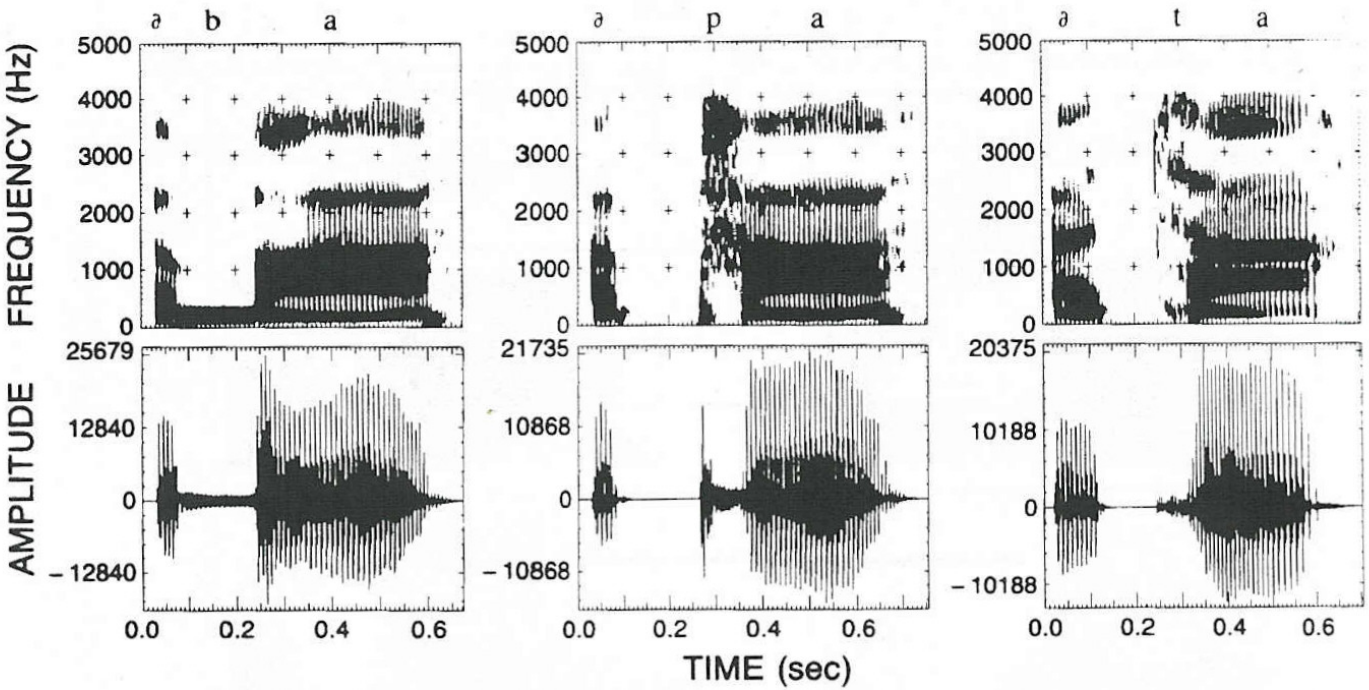


Figure 2.28 Spectrogram comparisons of the sequences of voiced (/ə-b-a/) and voiceless (/ə-p-a/ and /ə-t-a/) stop consonants.

Speech Production and ASR

- Directly use features from acoustic phonetics?
  - *e.g.*, (inferred) location of articulators; voicing; formant frequencies.
  - In practice, doesn't help.
- Still, influences how signal processing is done.
  - Source-filter model.
  - Separate excitation from modulation from vocal tract.
  - *e.g.*, frequency of excitation can be ignored (English).

Speech Perception and ASR

- As it turns out, the features that work well . . . .
  - Motivated more by speech perception than production.
- *e.g.*, Mel Frequency Cepstral Coefficients (MFCC).
  - Motivated by human perception of pitch.
  - Similarly for perceptual linear prediction (PLP).

Speech Perception — Physiology

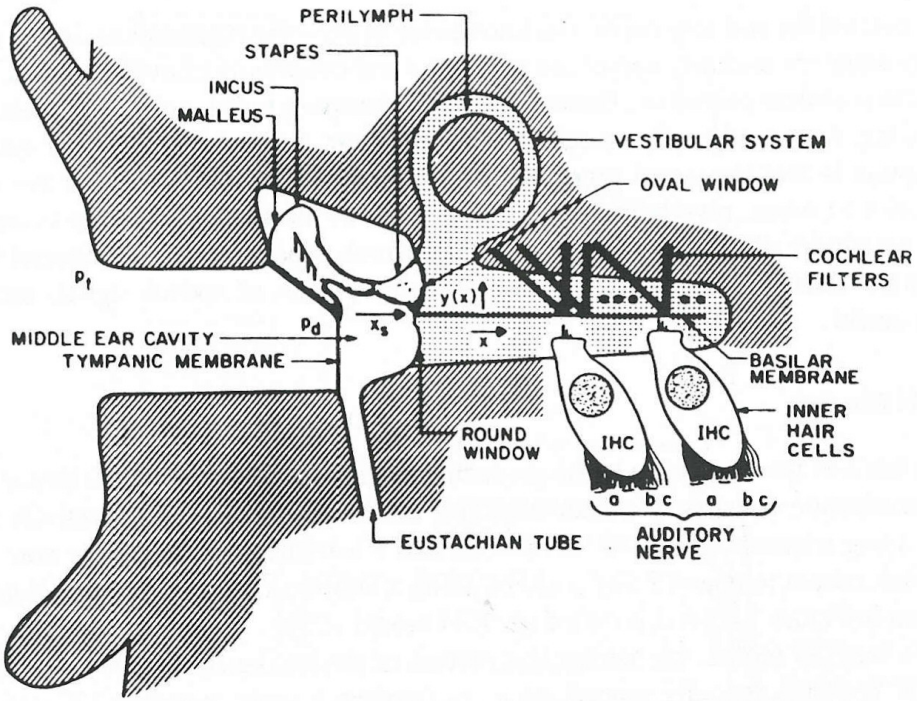


Figure 3.48 Expanded view of the middle and inner ear mechanics.

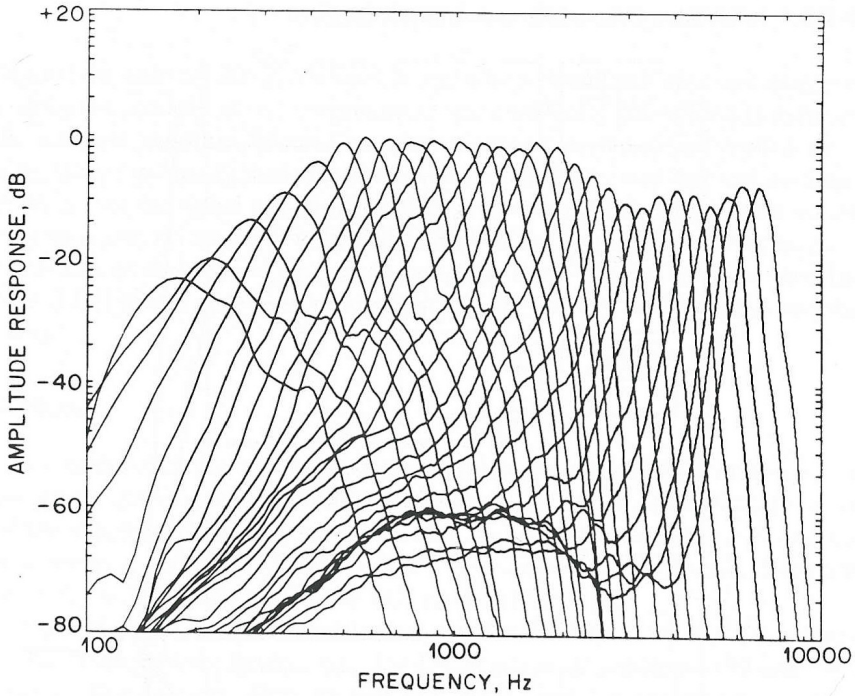


Figure 3.50 Frequency response curves of a cat's basilar membrane (after Ghitza [13]).

- Sound enters ear; converted to vibrations in cochlear fluid.
- In fluid is basilar membrane, with $\sim$30,000 little hairs.
  - Sensitive to different frequencies (band-pass filters).

Speech Perception — Physiology

- Human physiology used as justification for frequency analysis ubiquitous in speech processing.
- Limited knowledge of higher-level processing.
  - Can glean insight from psychophysical experiments.
  - Look at relationship between physical stimuli and psychological effects.

Speech Perception — Psychophysics

- Threshold of hearing as a function of frequency.
- 0 dB sound pressure level (SPL) $\Leftrightarrow$ threshold of hearing.
  - +20 decibels (dB) $\Leftrightarrow$ 10$\times$ increase in loudness.
- Tells us what range of frequencies people can detect.

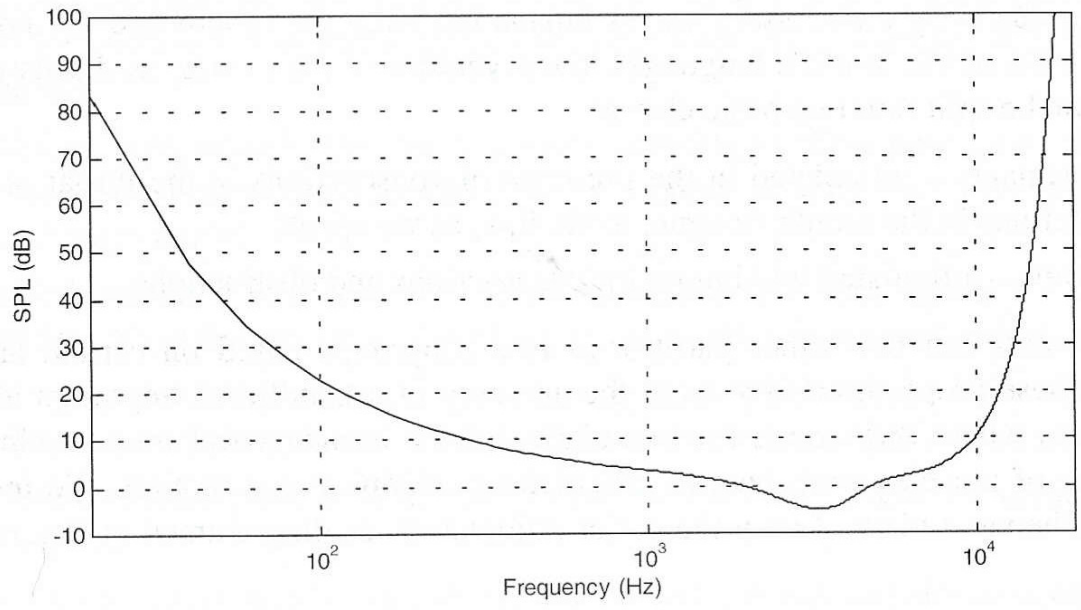
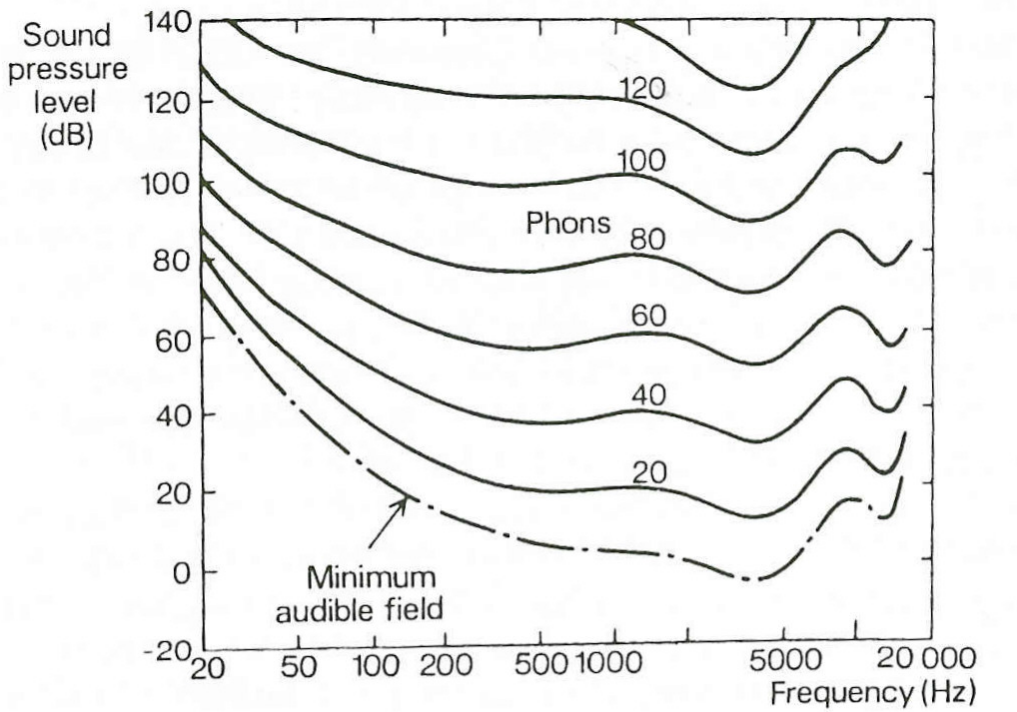


Figure 2.3 The sound pressure level (SPL) level in dB of the absolute threshold of hearing as a function of frequency. Sounds below this level are inaudible. Note that below 100 Hz and above 10 kHz this level rises very rapidly. Frequency goes from 20 Hz to 20 kHz and is plotted in a logarithmic scale from Eq. (2.3).

Speech Perception — Psychophysics

- Sensitivity of humans to different frequencies.
- Equal loudness contours.
  - Subjects adjust volume of tone to match volume of another tone at different pitch.
- Tells us what range of frequencies may be good to focus on.



Speech Perception — Psychophysics

- Human perception of distance between frequencies.
- Adjust pitch of one tone until twice/half pitch of other tone.
- Mel scale — frequencies equally spaced in Mel scale are equally spaced according to human perception.

$$\text{Mel freq} = 2595 \log_{10}(1 + \text{freq}/700)$$

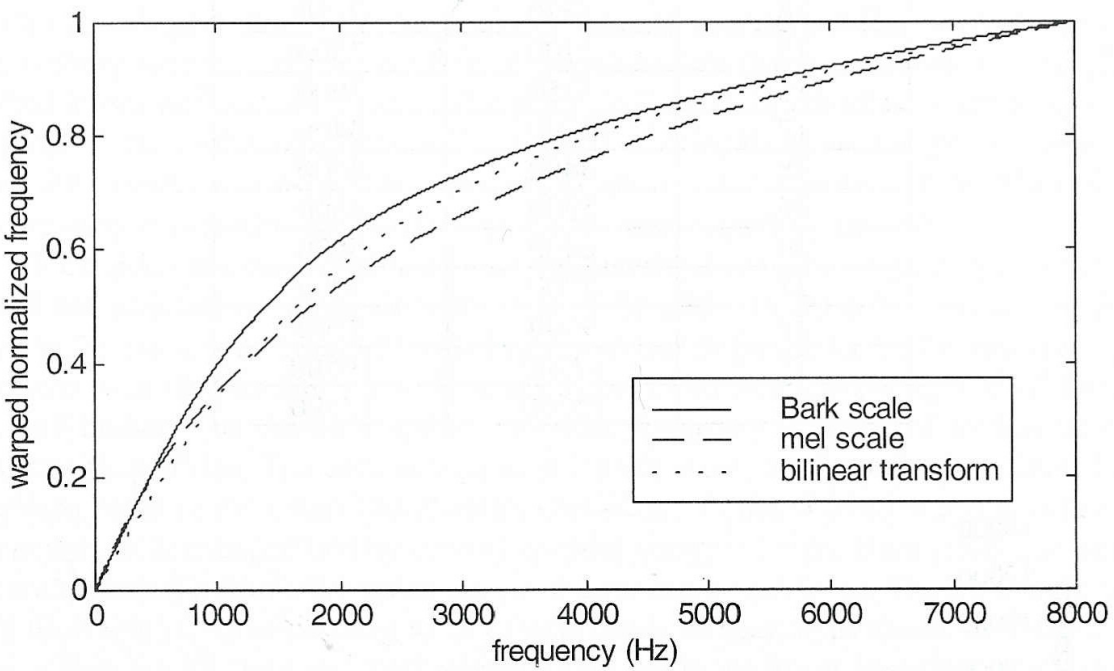


Figure 2.13 Frequency warping according to the Bark scale, ERB scale, mel-scale, and bilinear nsform for $\alpha = 0.6$: linear frequency in the x-axis and normalized frequency in the y-axis.

Speech Perception — Psychoacoustics

- Use controlled stimuli to see what features humans use to distinguish sounds.
- Haskins Laboratories (1940's); Pattern Playback machine.
  - Synthesize sound from hand-painted spectrograms.
- Demonstrated importance of formants, formant transitions, trajectories in human perception.
  - *e.g.*, varying second formant alone can distinguish between B, D, G.

`www.haskins.yale.edu/featured/bdg.html`

Speech Perception — Machine

- Just as human physiology has its quirks . . .
  - So does machine “physiology”.
- Sources of distortion.
  - Microphone — different response based on direction and frequency of sound.
  - Sampling frequency — *e.g.*, 8 kHz sampling for landlines throws away all frequencies above 4 kHz.
  - Analog/digital conversion — need to convert to digital with sufficient precision (8–16 bits).
  - Lossy compression — *e.g.*, cellular telephones, VOIP.

Speech Perception — Machine

- Input distortion can still be a significant problem.
  - Mismatched conditions between train/test.
  - Low bandwidth — telephone, cellular.
  - Cheap equipment — *e.g.*, mikes in handheld devices.
- Enough said.

Segue

- Now that we see what humans do.
- Let's discuss what signal processing has been found to work well empirically.
  - Has been tuned over decades.
- Start with some mathematical background.

Part II

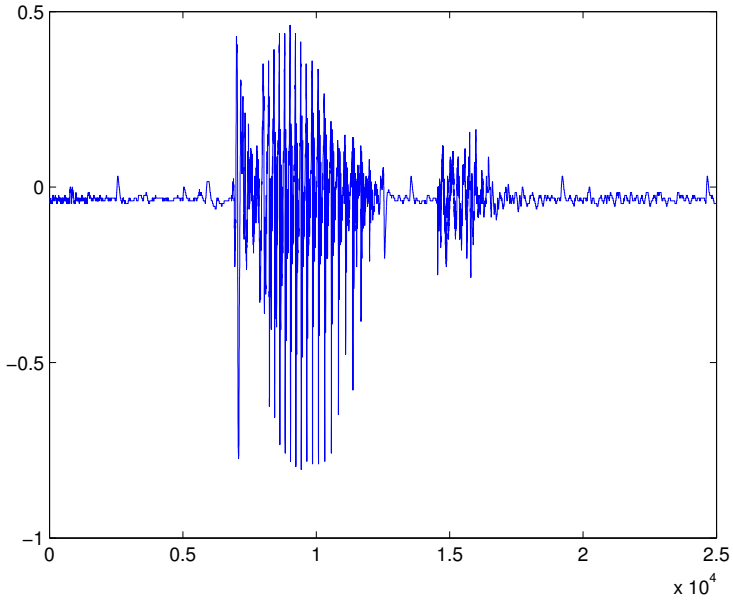
Signal Processing Basics

Overview

- Background material: how to mathematically model/analyze human speech production and perception.
  - Introduction to signals and systems.
  - Basic properties of linear systems.
  - Introduction to Fourier analysis.
- Next week: discussion of actual features used in ASR.
- Recommended readings: **[HAH]** pg. 201-223, 242-245. **[R+J]** pg. 69-91. All figures taken from these texts.

Signals and Systems

- Signal: a function $x(t)$ over time (continuous or discrete).
  - *e.g.*, output of A/D converter is a digital signal $x[n]$.



- A digital *system* (or *filter*) H takes an input signal $x[n]$ and produces a signal $y[n]$:

$$y[n] = H(x[n])$$

Speech Production

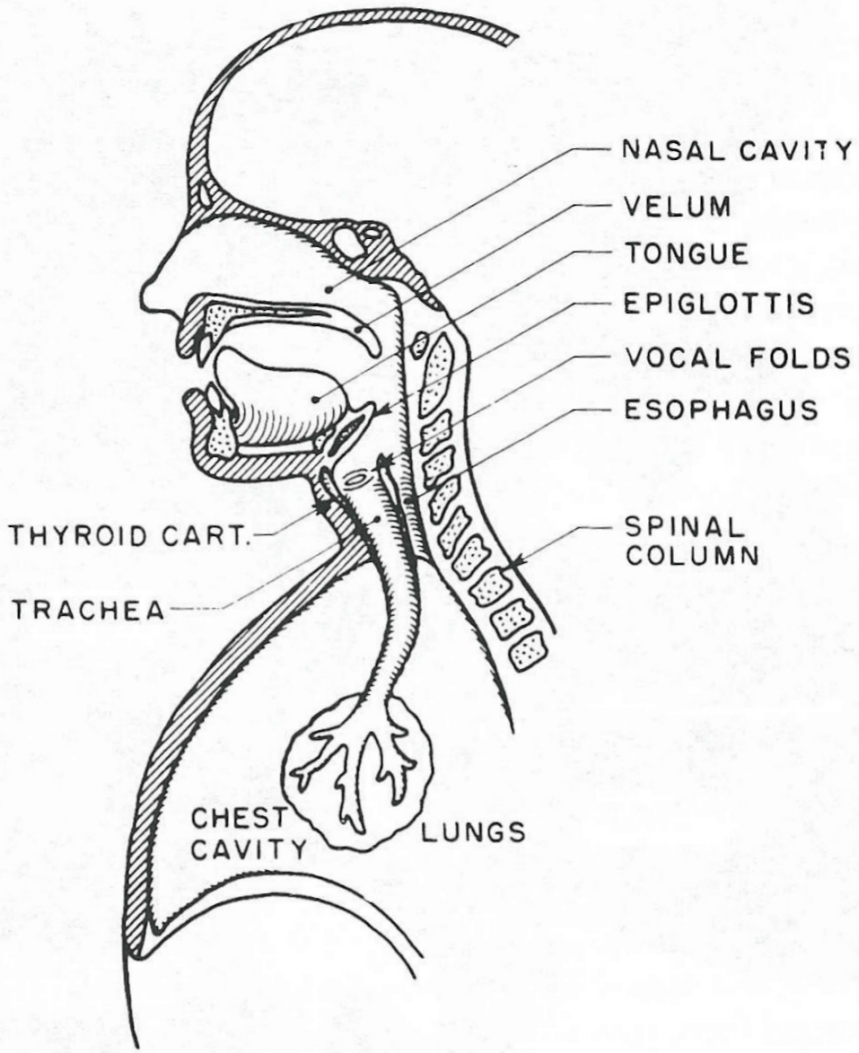
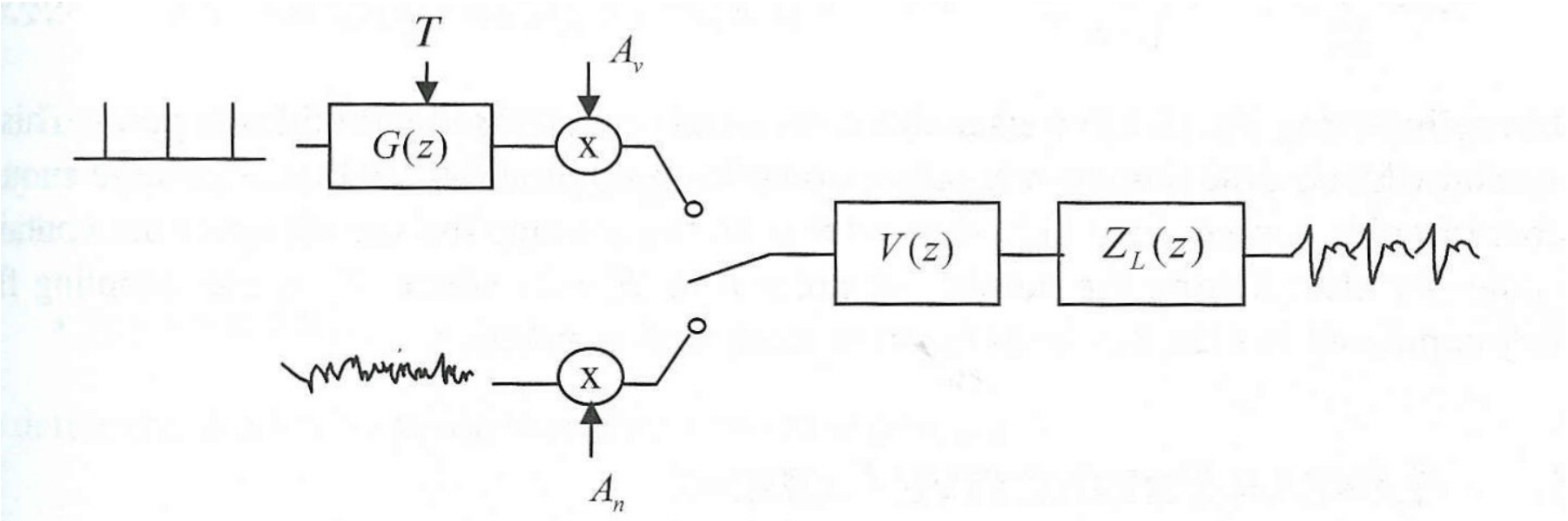


Figure 2.4 Schematic view of the human vocal mechanism (after Flanagan [3]).

The Source-Filter Model

- Vocal tract is modeled as sequence of filters.



- $G(z)$ — glottis (low-frequency emphasis).
- $V(z)$ — vocal tract; linear filter w/ time-varying resonances.
- $Z_L(z)$ — radiation from lips; high-frequency pre-emphasis.
- Interspeaker variation: glottal waveform; vocal-tract length.

Linear Time-Invariant Systems

- Calculating output of H for input signal x becomes very simple if digital system H satisfies two basic properties.
- H is *linear* if

$$H(a_1 x_1[n] + a_2 x_2[n]) = a_1 H(x_1[n]) + a_2 H(x_2[n])$$

- H is *time-invariant* if

$$y[n - n_0] = H(x[n - n_0])$$

i.e., a shift in the time axis of x produces the same output, except for a time shift.

Linear Time-Invariant Systems

- Let $h[n]$ be the response of an LTI system H to an impulse $\delta[n]$ (a signal which is 1 at $n = 0$ and 0 otherwise).
- Then, response of system to arbitrary signal $x[n]$ will be weighted superposition of impulse responses:

$$y[n] = \sum_{k=-\infty}^{\infty} x[k]h[n-k] = \sum_{k=-\infty}^{\infty} x[n-k]h[k]$$

The above is also known as *convolution* and is written as

$$y[n] = x[n] * h[n]$$

- *i.e.*, an LTI system H can be characterized completely by its *impulse reponse* $h[n]$.

Fourier Analysis

- Moving towards more meaningful features.
  - Time domain: $x[n] \sim$ air pressure at time n.
  - Frequency domain: $X(\omega) \sim$ energy at frequency ω.
  - This is what cochlear hair cells measure?
- Can express (almost) any signal $x[n]$ as sum of sinusoids.
  - Coefficient for sinusoid w/ frequency ω is $X(\omega)$.
- Given $x[n]$, can compute $X(\omega)$ efficiently, and *vice versa*.
  - Time and frequency domain representations are equivalent.
- *Fourier transform* converts between representations.

Review: Complex Exponentials

- Math is simpler using complex exponentials.
- Euler's formula.

$$e^{j\omega} = \cos\omega + j\sin\omega$$

- Sinusoid with frequency ω, phase ϕ.

$$\cos(\omega n + \phi) = \mathrm{Re}(e^{j(\omega n+\phi)})$$

The Fourier Transform

- The discrete-time Fourier transform (DTFT) is defined as

$$X(\omega) = \sum_{n=-\infty}^{\infty} x[n] e^{-j\omega n}$$

  Note: this is a *complex* quantity.
- The inverse Fourier transform is defined as

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

- Exists and is invertible as long as $\sum_{-\infty}^{\infty} |x[n]| < \infty$.
- Can apply DTFT to system/filter as well: $h[n] \Rightarrow H(\omega)$.

The Z-Transform

- One can generalize the discrete-time Fourier Transform to

$$X(z) = \sum_{n=-\infty}^{\infty} x(n) z^{-n}$$

  where z is any complex variable. The Fourier Transform is just the z-transform evaluated at $z = e^{-j\omega}$.
- The z-transform concept allows us to analyze a large range of signals, even those whose integrals are unbounded. We will primarily just use it as a notational convenience, though.

The Convolution Theorem

- Apply system H to signal x to get signal y: $y[n] = x[n] * h[n]$.

$$
\begin{aligned}
Y(z) &= \sum_{n=-\infty}^{\infty} y[n] z^{-n} = \sum_{n=-\infty}^{\infty} \left(\sum_{k=-\infty}^{\infty} x[k] h[n-k] \right) z^{-n} \\
&= \sum_{k=-\infty}^{\infty} x[k] \left(\sum_{n=-\infty}^{\infty} h[n-k] z^{-n} \right) \\
&= \sum_{k=-\infty}^{\infty} x[k] \left(\sum_{n=-\infty}^{\infty} h[n] z^{-(n+k)} \right) \\
&= \sum_{k=-\infty}^{\infty} x[k] z^{-k} H(z) = X(z) \cdot H(z)
\end{aligned}
$$

The Convolution Theorem (cont'd)

- Duality between time and frequency domains.

$$\begin{aligned} \mathrm{DTFT}(x[n] * y[n]) &= \mathrm{DTFT}(x) \cdot \mathrm{DTFT}(y) \\ \mathrm{DTFT}(x[n] \cdot y[n]) &= \mathrm{DTFT}(x) * \mathrm{DTFT}(y) \end{aligned}$$

- *i.e.*, convolution in time domain is same as multiplication in frequency domain, and *vice versa*.

Another Perspective

- If feed complex sinusoid $x[n] = e^{j\omega n}$ with frequency ω into LTI system H, then

$$y[n] = \sum_{k=-\infty}^{\infty} e^{j\omega(n-k)} h[k] = e^{j\omega n} \sum_{k=-\infty}^{\infty} e^{-j\omega k} h[k] = H(\omega) e^{j\omega n}$$

  Hence, if the input is a complex sinusoid, the output is a complex sinusoid with the same frequency, scaled (and phase-adjusted) by $H(\omega)$. In other words, H acts on each frequency independently.
- If $x[n] = \int X(\omega) e^{-j\omega n} d\omega$ is a combination of complex sinusoids, then by the LTI property

$$y[n] = \int H(\omega) X(\omega) e^{-j\omega n} d\omega$$

  This is another way to show $Y(\omega) = H(\omega) \cdot X(\omega)$.

Some Useful Quantities

- The *autocorrelation* of $x[n]$ with lag j is defined as

$$R_{xx}[j] = \sum_{n=-\infty}^{\infty} x[n+j]x^*[n] = x[j] * x^*[-j]$$

  where x^* is the complex conjugate of x. Can be used to help find pitch/F_0.
- The Fourier transform of $R_{xx}[j]$, denoted as $S_{xx}(\omega)$, is called the *power spectrum* and is equal to $|X(\omega)|^2$
- The *energy* of a discrete-time signal can be computed as:

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

The Discrete Fourier Transform (DFT)

- Preceding analysis assumes *infinite* signals: $n = -\infty, \ldots, +\infty$.
- In reality, can assume signals $x[n]$ are finite and of length N ($n = 0, \ldots, N-1$). Then, we can define the DFT as

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\omega n} = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi kn}{N}}$$

  where we have replaced ω with $\frac{2\pi k}{N}$
- The DFT is equivalent to a Fourier series expansion of a periodic version of $x[n]$.

The Discrete Fourier Transform (cont'd)

- The inverse of the DFT is

$$\begin{aligned}\frac{1}{N}\sum_{k=0}^{N-1} X[k]e^{j\frac{2\pi kn}{N}} &= \frac{1}{N}\sum_{k=0}^{N-1}\left[\sum_{m=0}^{N-1} x[m]e^{-j\frac{2\pi km}{N}}\right] e^{j\frac{2\pi kn}{N}} \\ &= \frac{1}{N}\sum_{m=0}^{N-1} x[m]\sum_{n=0}^{N-1} e^{j\frac{2\pi k(n-m)}{N}}\end{aligned}$$

- The last sum on the right is N for $m = n$ and 0 otherwise, so the entire right side is just $x[n]$.

The Fast Fourier Transform

- Note that the computation of

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi kn}{N}} \equiv \sum_{n=0}^{N-1} x[n] W_N^{nk}$$

  for $k = 0, \ldots, N-1$ requires $O(N^2)$ operations.
- Let $f[n] = x[2n]$ and $g[n] = x[2n+1]$. Then, we have

$$\begin{aligned} X[k] &= \sum_{n=0}^{N/2-1} f[n] W_{N/2}^{nk} + W_N^k \sum_{n=0}^{N/2-1} g[n] W_{N/2}^{nk} \\ &= F[k] + W_N^k G[k] \end{aligned}$$

  when $F[k]$ and $G[k]$ are the $N/2$ point DFT's of $f[n]$ and $g[n]$. To produce values for $X[k]$ for $N > k \geq N/2$, note that $F[k + N/2] = F[k]$ and $G[k + N/2] = G[k]$.
- The above process can be iterated to compute the DFT using only $O(N \log N)$ operations.

The Discrete Cosine Transform

- Instead of decomposing a signal into a sum of complex sinusoids, it can also be useful to decompose a signal into a sum of *real* sinusoids.
- The Discrete Cosine Transform (DCT) (a.k.a. DCT-II) is defined as

$$C[k] = \sum_{n=0}^{N-1} x[n] \cos\left(\frac{\pi}{N}(n + \frac{1}{2})k\right) \quad k = 0, \ldots, N-1$$

The Discrete Cosine Transform (cont'd)

- We can relate the DCT and DFT as follows. If we create a signal

$$y[n] = x[n] \qquad n = 0, \ldots, N-1$$
$$y[n] = x[2N-1-n] \qquad n = N, \ldots, 2N-1$$

then $Y[k]$, the DFT of $y[n]$, is

$$Y[k] = 2e^{j\frac{\pi k}{2N}} C[k] \qquad k = 0, \ldots, N-1$$
$$Y[2N-k] = 2e^{-j\frac{\pi k}{2N}} C[k] \qquad k = 1, \ldots, N-1$$

- By creating such a signal, the overall energy will be concentrated at lower frequency components (because discontinuities at the boundaries will be minimized). The coefficients are also all real. This allows for easier truncation during approximation and will come in handy later when computing MFCCs.

Long-Term *vs.* Short-Term Information

- Have infinite (or long) signal $x[n]$, $n = -\infty, \ldots, +\infty$.
  - Take DTFT or DFT of whole damn thing.
  - Is this interesting?
- Point: we want short-term information!
  - *e.g.*, how much energy at frequency ω over span $n = n_0, \ldots, n_0 + k$?
- Going from long-term to short-term analysis.
  - Windowing.
  - Filter banks.

Windowing: The Basic Idea

- Excise N points from signal $x[n]$, $n = n_0, \ldots, n_0 + (N-1)$ (*e.g.*, 0.02s or so).
- Perform DFT on truncated signal; extract some features.
- Shift n_0 (*e.g.*, by 0.01s or so) and repeat.

What's the Problem?

- Excising N points from signal $x \Leftrightarrow$ multiplying by rectangular window y.
- Convolution theorem: multiplication in time domain is same as convolution in frequency domain.
  - Fourier transform of result is $X(\omega) * Y(\omega)$.
- Imagine original signal is periodic.
  - Ideal: after windowing, $X(\omega)$ remains unchanged $\Leftrightarrow$ $Y(\omega)$ is delta function.
  - Reality: short-term window cannot be perfect.
  - How close can we get to ideal?

Rectangular Window

$$h[n] = \begin{cases} 1 & n = 0, \ldots, N-1 \\ 0 & \text{otherwise} \end{cases}$$

- The FFT can be written in closed form as

$$H(\omega) = \frac{\sin \omega N/2}{\sin \omega/2} e^{-j\omega(N-1)/2}$$

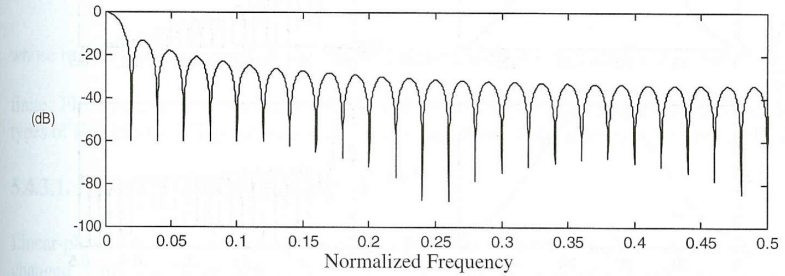


Figure 5.19 Frequency response (magnitude in dB) of the rectangular window with $N = 50$, which is a digital sinc function.

- Note the high sidelobes of the window. These tend to distort low energy components in the spectrum when there are significant high-energy components also present.

Hanning and Hamming Windows

- Hanning: $h[n] = .5 - .5\cos 2\pi n/N$
- Hamming: $h[n] = .54 - .46\cos 2\pi n/N$

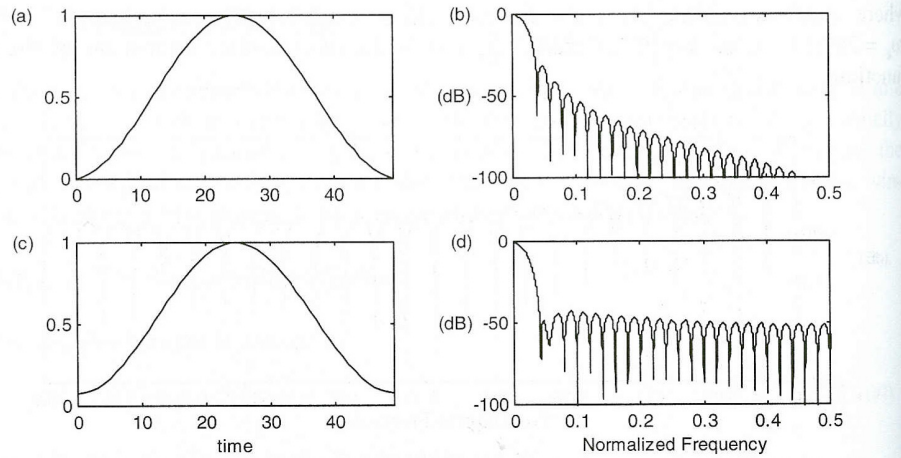


Figure 5.20 (a) Hanning window and (b) the magnitude of its frequency response in dB; (c) Hamming window and (d) the magnitude of its frequency response in dB for $N = 50$.

- Hanning and Hamming have slightly wider main lobes, much lower sidelobes than rectangular window.
- Hamming window has lower first sidelobe than Hanning; sidelobes at higher frequencies do not roll off as much.

Human Perception and the FFT

- Each cochlear hair acts like band-pass filter?
  - Input signal: air pressure; output: hair displacement.
  - Each hair responds to different frequency.
  - Cochlea is a *filter bank*?
- Implementing filter bank via brute force convolution.
  - For each output point n, computation for ith filter is on order of L_i (length of impulse response).

$$x_i[n] = x[n] * h_i[n] = \sum_{m=0}^{L_i-1} h_i[m]x[n-m]$$

Filter Terminology

- A filter H acts on each input frequency ω independently.
  - Scales component with frequency ω by $H(\omega)$.
- *Low-pass* filter.
  - "Lets through" all frequencies below cutoff frequency.
  - Suppresses all frequencies above.
- *High-pass* filter; *band-pass* filter.

Implementation of Filter Banks

- Given low-pass filter $h[n]$, can create band-pass filter $h_i[n] = h[n]e^{j\omega_i n}$ via *heterdyning*.
  - Multiplication in time domain $\Rightarrow$ convolution in frequency domain $\Rightarrow$ shift $H(\omega)$ by ω_i.

$$\begin{aligned} x_i[n] &= \sum h[m]e^{j\omega_i m}x[n-m] \\ &= e^{j\omega_i n}\sum x[m]h[n-m]e^{-j\omega_i m} \end{aligned}$$

- The last term on the right is just $X_n(\omega)$, the Fourier transform of a windowed signal, where now the window is the same as the filter. So, we can interpret the FFT as just the instantaneous filter outputs of a uniform filter bank whose bandwidths corresponding to each filter are the same as the main lobe width of the window.

Implementation of Filter Banks (cont'd)

- Notice that by combining various filter bank channels we can create non-uniform filterbanks in frequency.

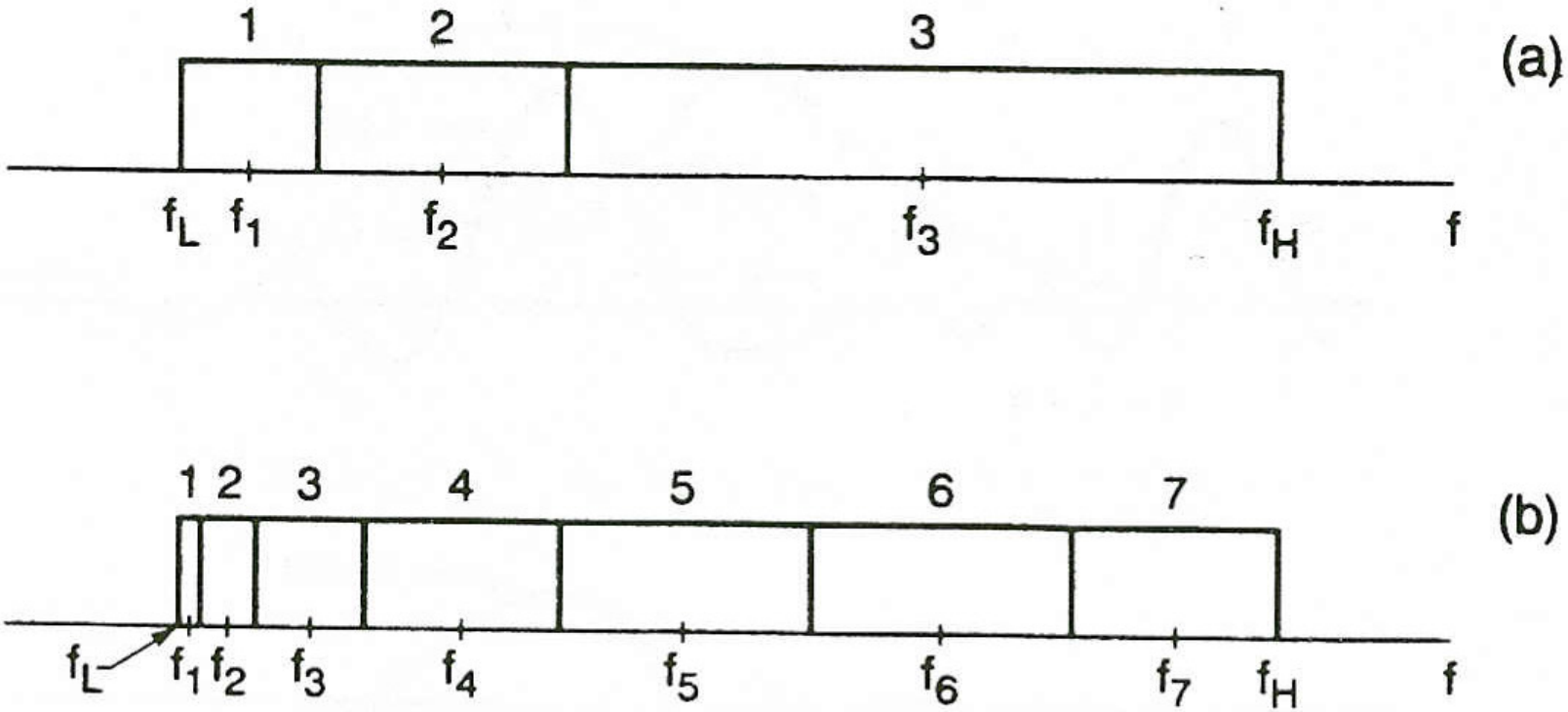


Figure 3.18 Two arbitrary nonuniform filter-bank ideal filter specifications consisting of either 3 bands (part a) or 7 bands (part b).