

Topics For Today

** project oriented*

- Introduction to the class
- Personal introduction
- Getting started

Instructional Contacts

- Prof. Visar Berisha
- Classroom: COOR2255
 - T 9:00 AM -11:45 AM
- Office hours
 - T 12:00 PM -2:00 PM
 - <https://asu.zoom.us/my/visar> or COOR3472
 - Email: visar@asu.edu
 - Questions about homework/concepts can be difficult to answer over email. Please see me for office hours.

+ assignments → problems
codes

What class is this?

- SHS/EEE 598– Speech analysis
- Textbook
 - There is no official textbook, but the following books are very good to have for reference:
 - DSP First by McClellan, Schafer, and Yolder
 - Discrete-Time Speech Signal Processing by Quatieri *classic & dense*
 - Digital Processing of Speech Signals by Rabiner *maybe this first*
 - (note: You don't have to buy these books for the class. They are useful if you intend to continue working in this field)
- Classroom slides posted on canvas
- Pre/co-requisites
 - Basic signal processing concepts → *1 class review*

Class Topics

- Brief review of signal processing
- Speech production
- Computational models of speech analysis/synthesis (STFT, LPC) *Time → freq domain*
- 1st part*
2nd part • Computing and modifying speech parameters (pitch, formants, envelope, etc.)
- Basic computational psychoacoustics *human perception*
mp3 algorithm
- Applications of speech processing (speech compression, speech recognition, etc.)
- (If we have time)* • Natural language processing *working w/ words & phrases*

Class Grades

- **Homework (50%)**

1. Quantitation, Sampling

- 4 homework assignments. All require Matlab programming
- Due dates are in the syllabus
- Work in groups of two
- Homework assignments will cover concepts discussed in class and will likely require that students spend additional time outside of class further researching these topics.

or python

- **Final Project (50%)**

- A final project is to be submitted and presented (15 min) on the last day of class (November 29)
- Students are free to suggest their own projects. Come see me if you need ideas
- Every student must submit a 1-2 page proposal for their project. Details to follow
- The project will be graded based on the deliverables outlined in the project proposal and on the presentation to be delivered in class by the students

Additional resources

- All HW assignments and the final project require the use of MATLAB
- Mathworks has a very good introductory and interactive tutorial:
https://www.mathworks.com/academia/student_center/tutorials/mltutorial_launchpad.html?confirmation_page
- The sections on “Variables and Expressions” and “Writing Functions” will be particularly useful
- ***If you have no/limited familiarity with Matlab, I strongly recommend that you to complete these tutorials***
- There is no specific textbook for the course. I have borrowed from different sources to put the lectures together. The following will be useful for further information on these topics:
 - *DSP First* by McClellan, Schafer, and Yolder
 - *Discrete-Time Speech Signal Processing* by Quatieri
 - *Digital Processing of Speech Signals* by Rabiner
 - among others
- For each lecture, I will try to make it clear from which resources I have developed the slides
- The lectures and research papers we will study in class (and outside of class for your projects) are a starting point
- Because this is a project-based class, it is very likely that you will have to research other resources at some point *Do more research*

Electrical Engineering → signal processing

Cheating

- Don't cheat
 - It's not worth it
 - I will pass your information on to the school director/department head



- **Please DO study together**
 - Don't just copy though
 - Learn from each other
 - Homework is used to make sure that you understand
 - However, make sure that it's yours if your name is on it

Class Objectives (4)

- Understand basic concepts related to speech/audio perception and speech production * speech analysis
Dolby, Sound Forge
- Implement simple computational models of various aspects of speech/audio production and perception
- Learn to use Matlab for technical programming
- Apply concepts from signal processing, speech science, and hearing science classes to engineering applications Interdisciplinary

Feedback

- Please let me know if something is confusing
 - You won't be alone
 - Ask questions
- This class is usually very diverse in academic background. This is fantastic, but the lectures may be challenging if you don't have experience in signal processing
- If you have suggestions for how to improve the class, please let me know early on
- I will do my best to make sure you know where you stand as class proceeds by posting grades on canvas *be earnest
dedicate time*

Special Needs or Emergencies

- If you have any special needs or constraints let me know
 - Will do my best to work with you
 - Usually easier if I know earlier
- Real emergencies are rare, but they do happen
 - Talk to me

Teaching Style

- Mix of slides and board in class
- In-class Matlab demonstrations
- Will upload some slides, notes, and Matlab scripts

A couple of words about the class

- It's OK if you don't understand every detail of lecture – this is expected given the diversity of backgrounds in the class → *office hour*
- Use your partner's expertise to learn more about concepts that you are having difficulties with
- The homework is designed to make you analyze a problem and provide a solution. There are many “correct” answers
- It's very likely that there will be concepts on the homework that may require addition research on your own. Feel free to reach out to me (or your classmates)

A few words about my background

- Started at ASU in 2013
- Joint appointment between ECEE and CHS
↳ health
- Previously worked at Raytheon Co. and at MIT Lincoln Laboratory
- Machine learning + Information *theory* + signal processing + speech/language analytics
application

Research interests: Speech analysis for health applications



COGNITION

What we say

e.g. 알츠하이머

- Memory
- Semantic/Syntactic
- Coherence
- Affect/Mood/Sentiment

biac → biometric



RESPIRATORY

How we breathe

e.g. asthma

- Capacity
- Pause rate
- Sufficiency
- Time to exhaustion
- Others



MOTOR

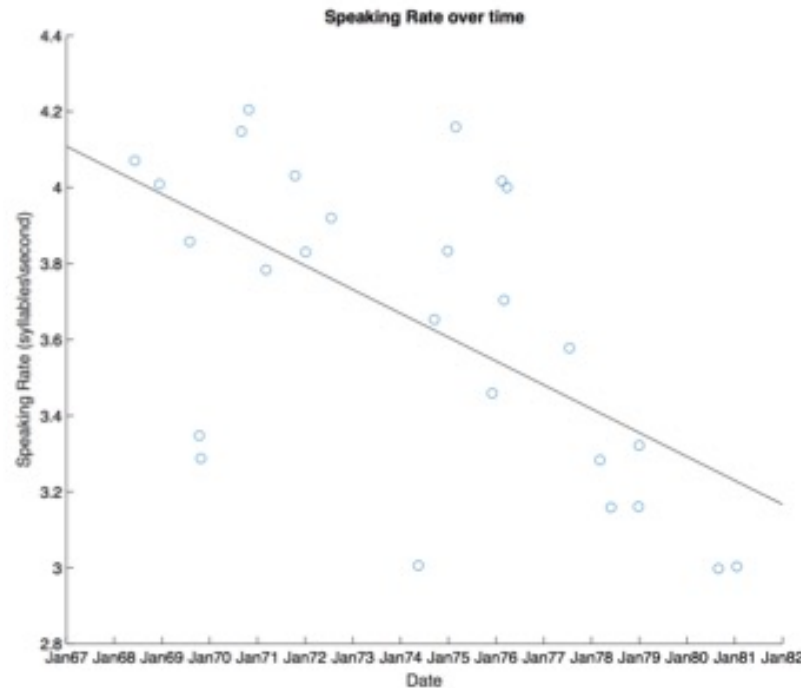
e.g. Parkinson

How we say things

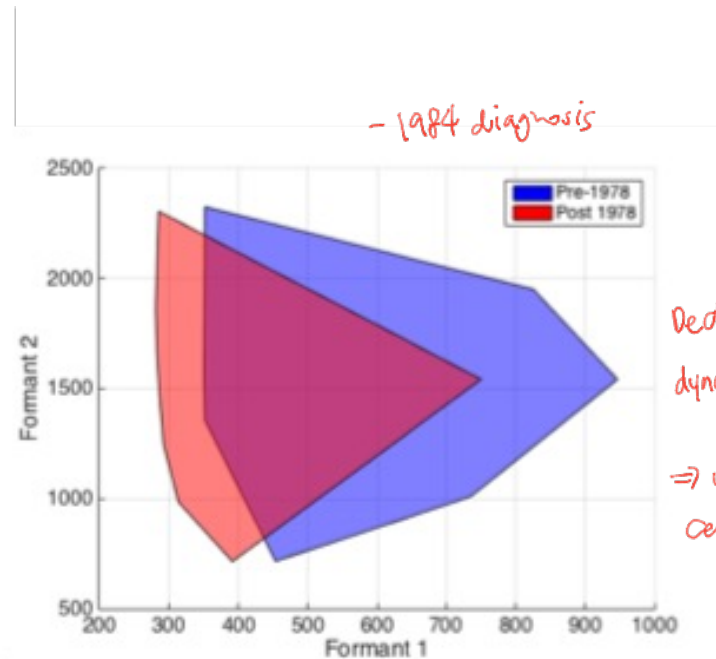
- Phonation
- Articulation
- Velopharyngeal control
- Prosody

Research projects: Speech analysis for health applications

Acoustic → motorability



- microphone quality (technology advance) ... unexpected change.
- got hit in the head

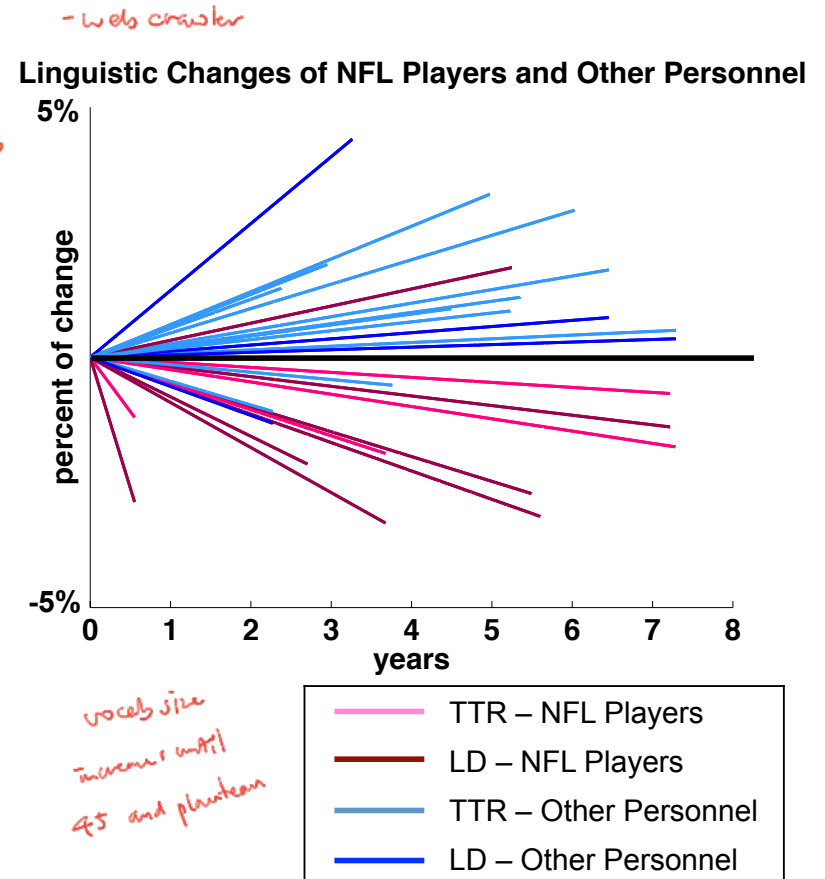
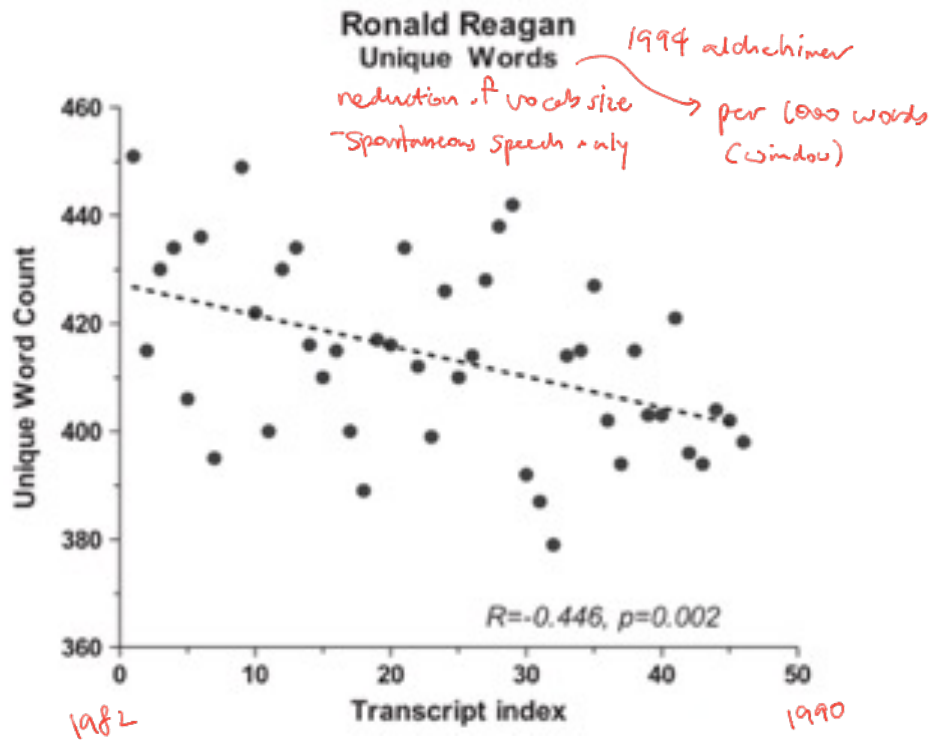


Decline of dynamic range
⇒ vowel centralization

signal to noise ratio

Berisha, Visar, et al. "Float Like a Butterfly Sting Like a Bee: Changes in Speech Preceded Parkinsonism Diagnosis for Muhammad Ali." *INTERSPEECH*. 2017.

Research projects: Language processing for health applications



Berisha, Visar, et al. "Tracking discourse complexity preceding Alzheimer's disease diagnosis: a case study comparing the press conferences of presidents Ronald Reagan and George Herbert Walker Bush." *Journal of Alzheimer's Disease* 45.3 (2015): 959-963.

Berisha, Visar, et al. "Longitudinal changes in linguistic complexity among professional football players." *Brain and language* 169 (2017): 57-63.

Personal introductions

- Go around the room and answer the following questions:
 - Name
 - Degree and major *MA S, PhD I*
 - Describe your research project (if you are doing a thesis-based master's degree or a PhD)
 - Any previous experience with speech/audio
 - Any previous experience with signal processing
 - What do you hope to get out of the class?

Introduction to the course

EEE/SHS 598: Speech Production, Perception and Processing

Speech Processing

Interdisciplinary

- Speech is the most natural form of human-human communications.
- Speech is related to language; **linguistics** is a branch of social science.
- Speech is related to human physiological capability; **physiology** is a branch of medical science.
- Speech is also related to sound and **acoustics**, a branch of physical science.
- Therefore, speech is one of the most intriguing signals that humans work with every day.
- Purpose of speech processing:
 - To understand speech as a means of **communication**;
 - To represent speech for **transmission and reproduction**; *← engineering interest*
 - To analyze speech for **automatic recognition** and extraction of information *(ML application)*
 - To discover some **physiological characteristics** of the talker. *Speaker ID*



Velar pharyngeal Flap

Cross sectional MRI
Sagittal



Beatboxing + singing

Examples of speech applications

Speech Applications — coding, synthesis, recognition, understanding, verification, language translation, speed-up/slow-down

Projects
= Compression
stt
NLP
tes

Building blocks
Speech Algorithms — speech-silence (background), voiced-unvoiced decision, pitch detection, formant estimation

Speech Representations — temporal, spectral, homomorphic, LPC

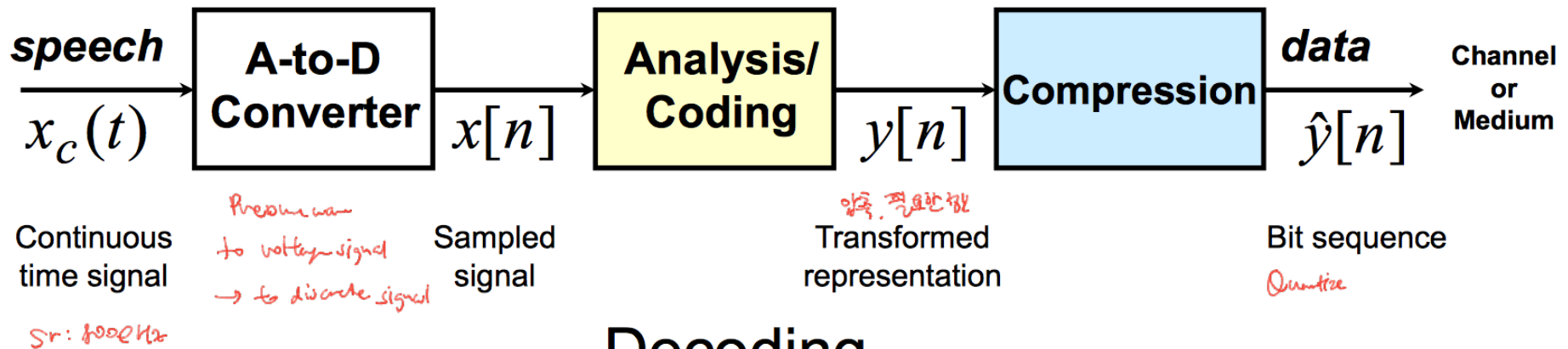
Time domain
Freq. domain

Fundamentals — acoustics, linguistics, pragmatics, speech perception

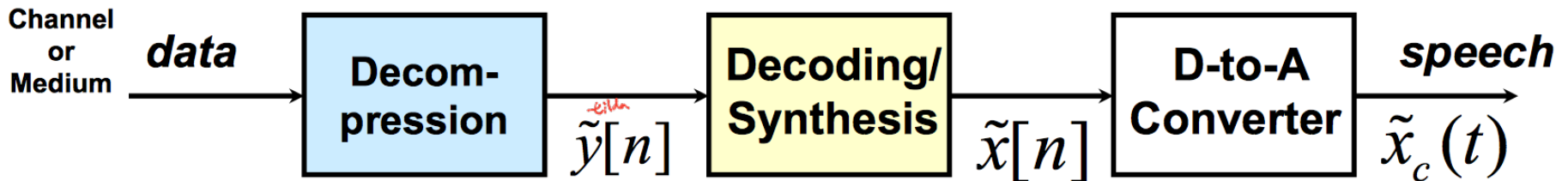
Speech Coding

(compression)

Encoding



Decoding



Demo of Speech Coding

Sampling rate: 8 kHz *Phone cellular network*

- **Narrowband** Speech Coding:

- ❖ 64 kbps PCM
- ❖ 32 kbps ADPCM
- ❖ 16 kbps LDCELP *Low Delay*
- ❖ 8 kbps CELP
- ❖ 4.8 kbps FS1016
- ❖ 2.4 kbps LPC10E

NB Speech *mumbled*



WB Speech



[S] high freq 4-8 kHz
[S] vs. [P] would be difficult

Sampling rate: 16 kHz *Zoom (connection quality 4)*

- **Wideband** Speech Coding: *dropped packets in transmission*
mp3. audio coding storage (eg virtual session)
Male talker / Female Talker

- ❖ 3.2 kHz – uncoded
- ❖ 7 kHz – uncoded
- ❖ 7 kHz – 64 kbps
- ❖ 7 kHz – 32 kbps
- ❖ 7 kHz – 16 kbps

Original



G.722 48 kbps



G.722 56 kbps

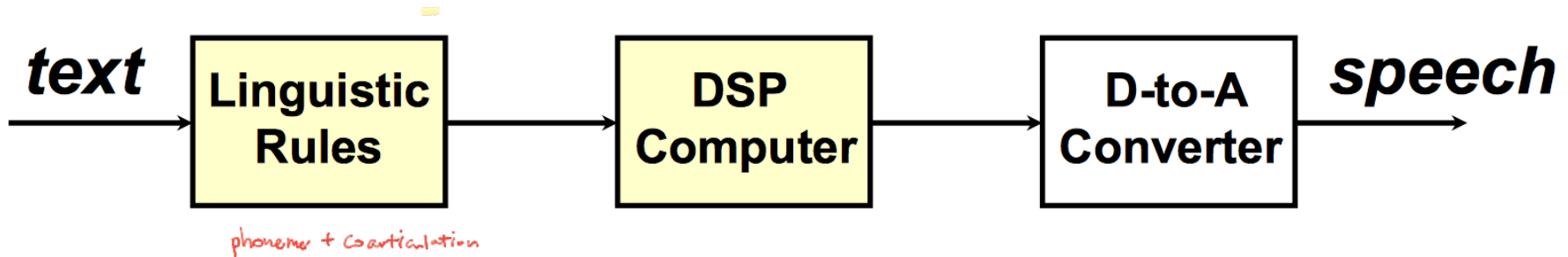


G.722 64 kbps



Speech Synthesis

(traditional method)



TTS Female



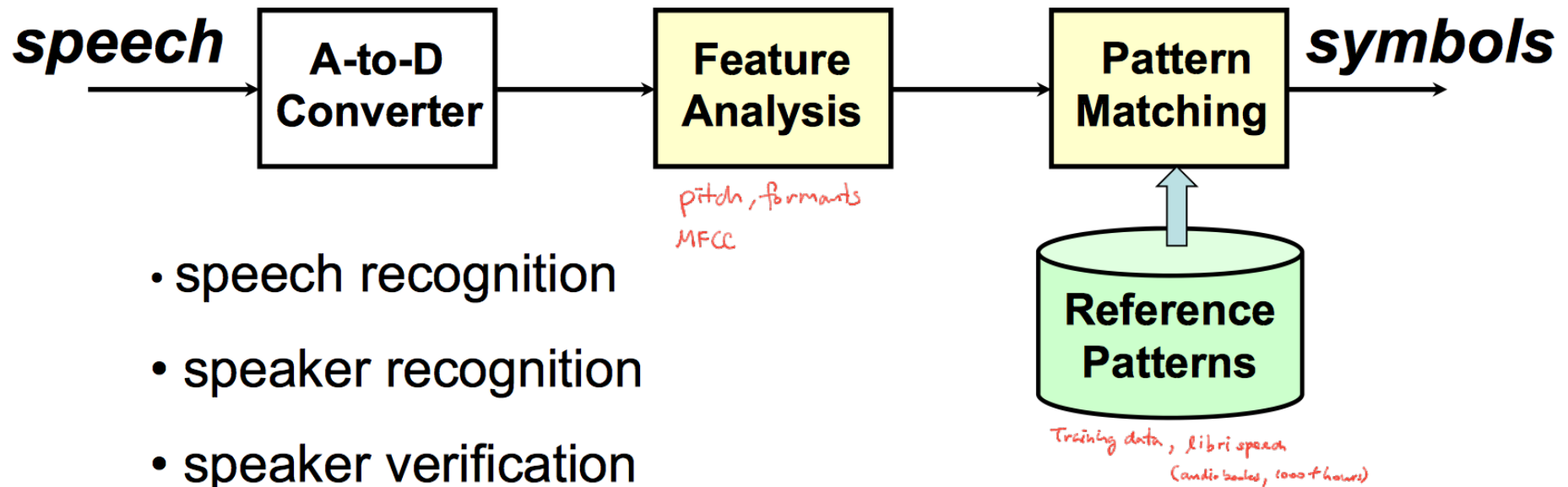
TTS Male



Some very rapid improvement in this field starting with **WavNet**

search on Google
↓

Pattern Matching Problems

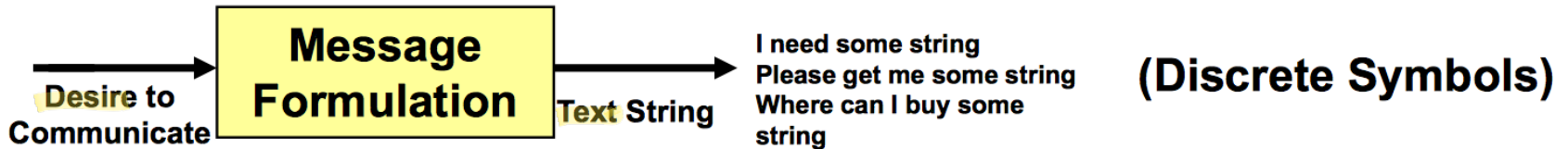


- speech recognition
- speaker recognition
- speaker verification
- word spotting
- automatic indexing of speech recordings

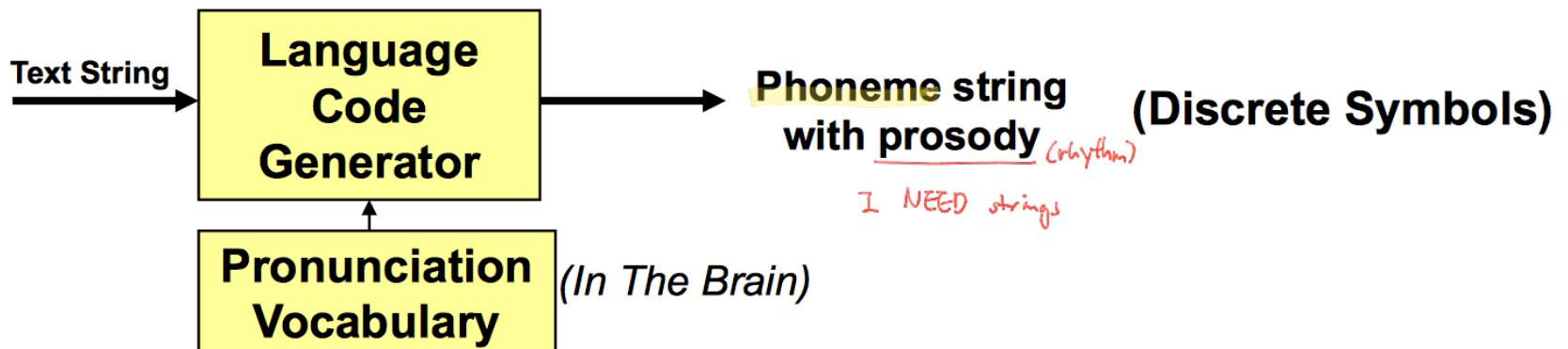
Speech production and speech perception

Speech Production/Generation Model

- **Message Formulation** → desire to communicate an idea, a wish, a request, => express the message as a sequence of words



- **Language Code** → need to convert chosen text string to a sequence of sounds in the language that can be understood by others; need to give some form of emphasis, prosody (tune, melody) to the spoken sounds so as to impart non-speech information such as sense of urgency, importance, psychological state of talker, environmental factors (noise, echo)

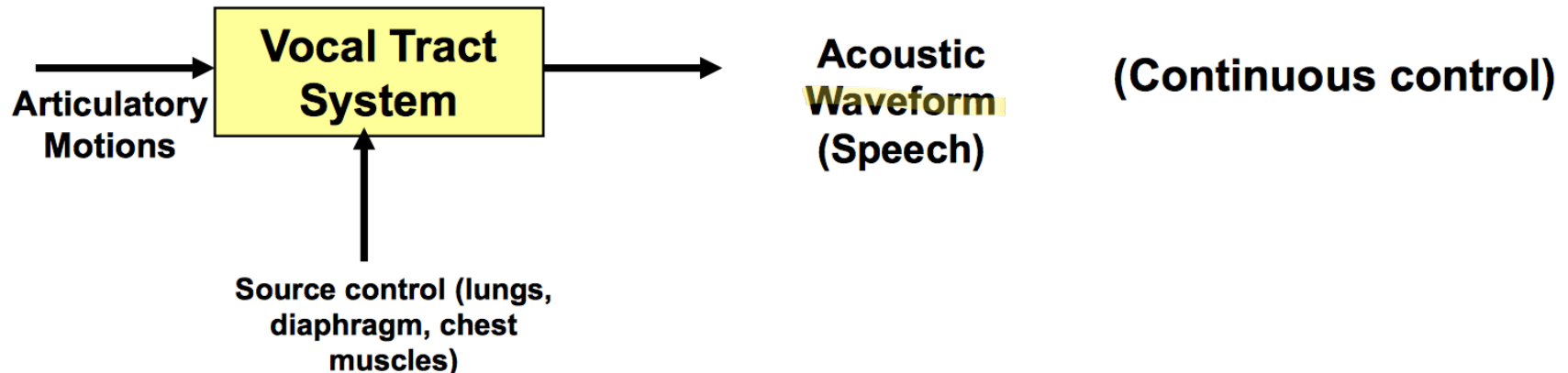


Speech Production/Generation Model

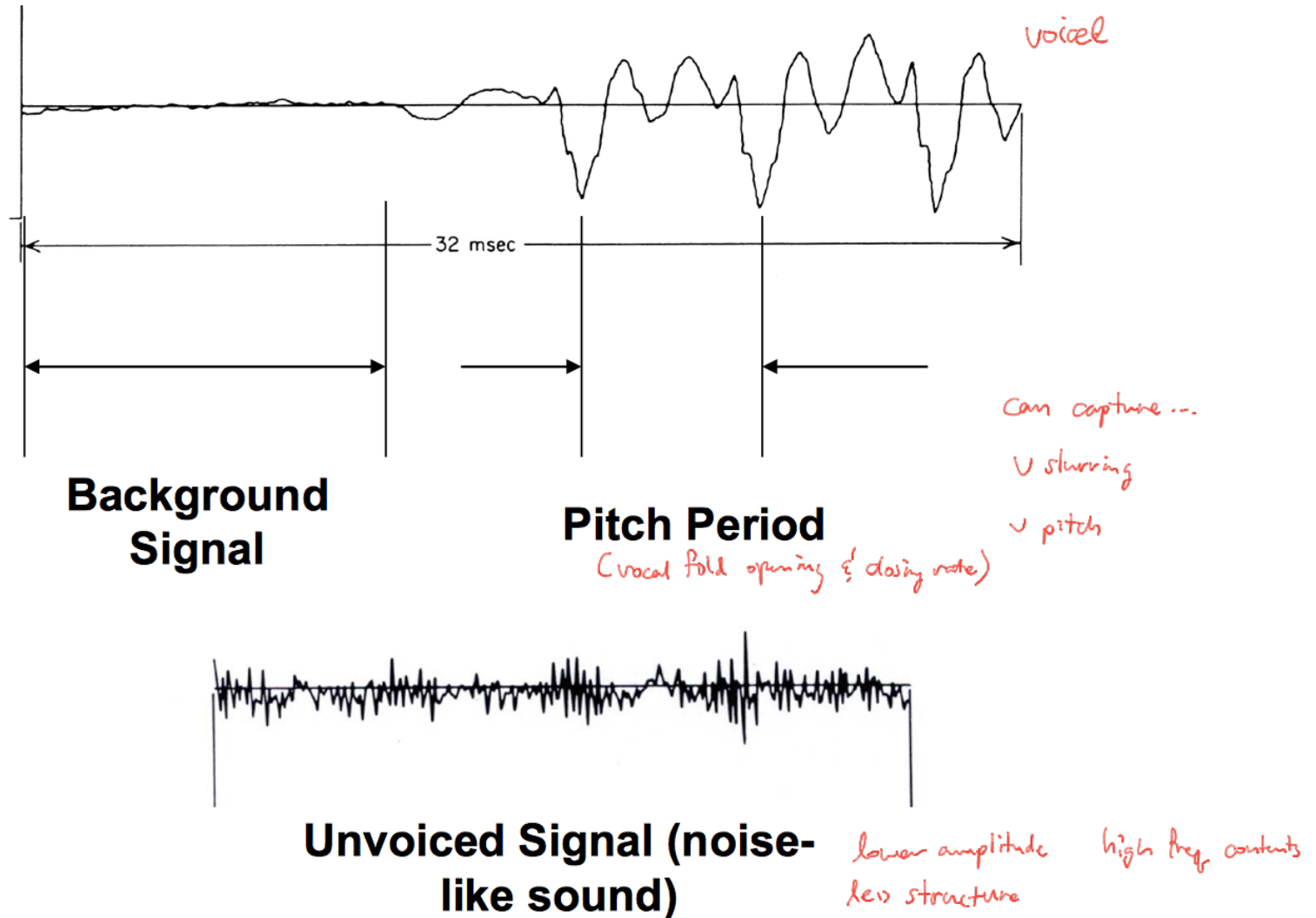
- **Neuro-Muscular Controls** → need to direct the neuro-muscular system to move the articulators (tongue, lips, teeth, jaws, velum) so as to produce the desired spoken message in the desired manner



- **Vocal Tract System** → need to shape the human vocal tract system and provide the appropriate sound sources to create an acoustic waveform (speech) that is understandable in the environment in which it is spoken



The Speech Signal



Speech Perception Model

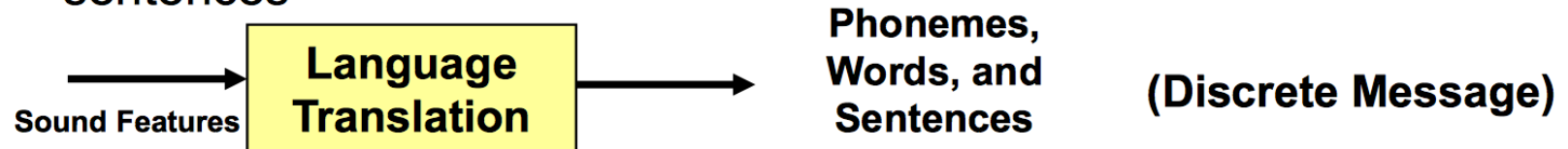
- The acoustic waveform impinges on the **ear** (the basilar membrane) and is spectrally analyzed by an equivalent filter bank of the ear



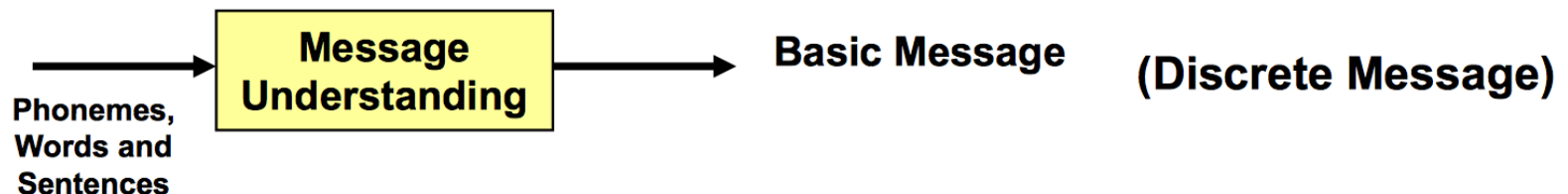
- The signal from the basilar membrane is neurally transduced and coded into features that can be decoded by the brain



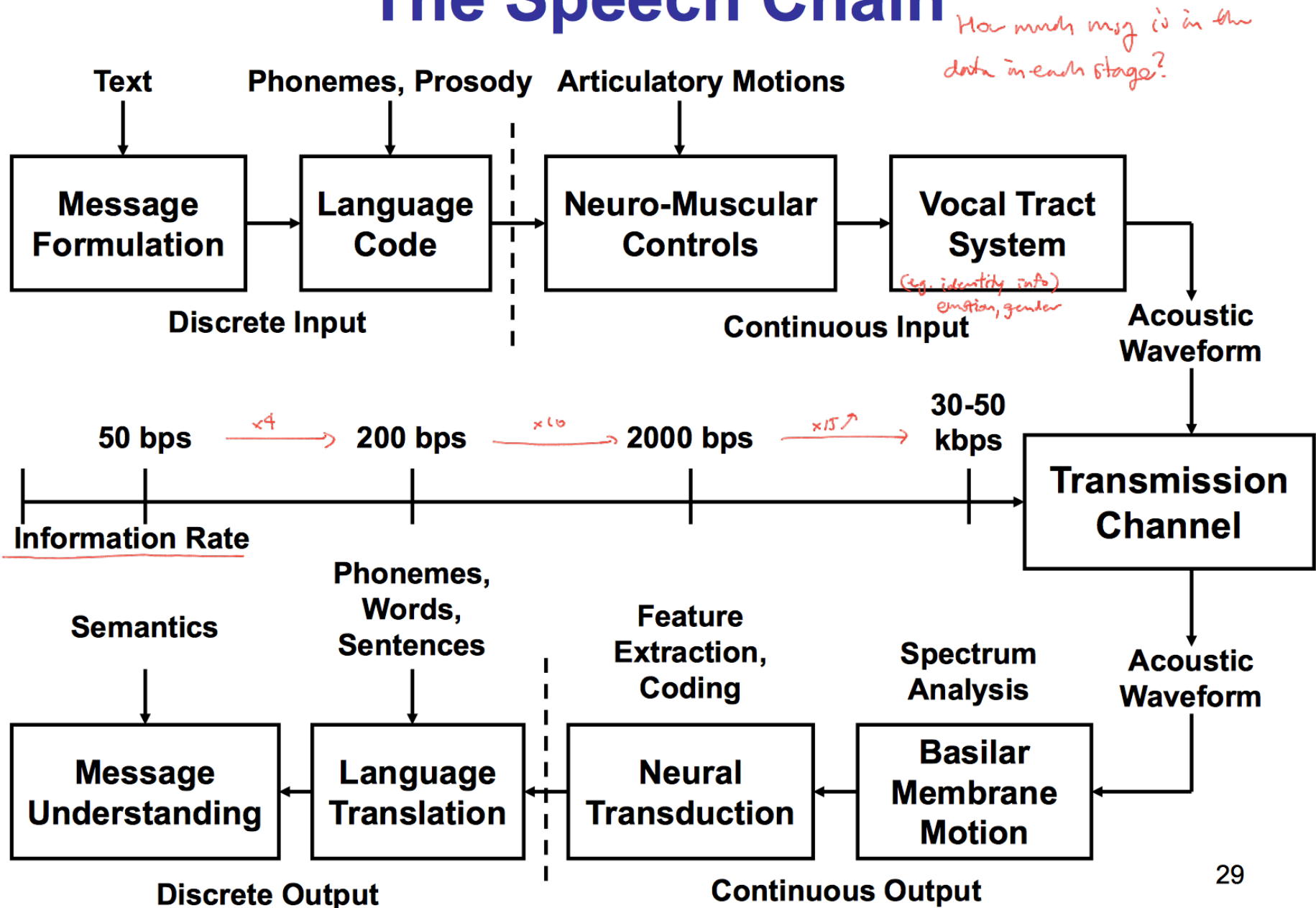
- The brain decodes the feature stream into sounds, words and sentences



- The brain determines the meaning of the words via a message understanding mechanism



The Speech Chain



What We Will Be Learning

- Review of DSP concepts
- Acoustic categorization of speech
- time* (• Time-domain methods in speech analysis
- Estimating speech parameters (pitch, formants, energy, etc.)
- freq* (• Spectral representations of speech (Fourier, Linear prediction, cepstral representation)
- Speech applications (pattern matching, compression)
- Computational psychoacoustics
- This is a project-based course
- I expect that a good deal of the learning will happen outside the classroom
- You will be working on projects that require you to understand not just the theory behind speech analysis, but also practical implementation issues
- I will be available outside of class for guidance

Homework and Final Project

- Your grade for the course is based on your performance on the four homework assignments and the final project.
- The homework assignments will consist of implementation and evaluation of concepts discussed during class.
- They are purposefully open-ended and will likely require further research into the concepts discussed in class.
- The first assignment has been uploaded and it will give you an idea of the type of projects we will be working on.
- There is a reason that you have 3 weeks to work on it. Don't wait until the last minute.
- I will be available outside of class for any questions (see syllabus).
- The final project will be one of your choosing. Please be sure to get an OK from me prior to starting. I would like to meet with each of the groups and discuss the specifics of the project, including deliverables and implementation details.

Possible Projects

- Music Applications
 - Virtual Jam sessions with ultra-low-delay audio compression (Opus interactive Audio Codec)
 - Music for cochlear implant users *- Guitar tuner*
- Speech Modification:
 - time-scale manipulations:
 - Speeding up speech –
 - Message playback
 - Voice mail
 - Reading machines and books for the blind
 - Slowing down speech –
 - Learning a foreign language
 - Voice transformations using Pitch and spectral changes of speech signal:
 - Voice disguise *(gender change)*
 - Entertainment
 - Speech synthesis
 - Spectral change of frequency compression and expansion:
 - may be useful in transforming speech as an aid to the partially deaf
 - Many methods can be applied to music and special effects.
 - Processing “whale” speech*

Possible Projects

- Speech Coding *A lot has been done already*
 - Goal is to reduce the information rate measured in bits per second while maintaining the quality of the original waveform.
 - Waveform coders:
 - Represent the speech waveform directly and do not rely on a speech production model.
 - Operate in a high range of 16-64 kbps
 - Vocoders:
 - Largely are speech model-based and rely on a small set of model parameters.
 - Operate at the low bit range of 1.2-4.8 kbps
 - Lower quality than waveform coders.
 - Hybrid coders:
 - Partly waveform based and partly speech model-based
 - Operate in the 4.8 – 16 kbps range
 - Applications of speech coders include:
 - Digital telephony over constrained bandwidth channels
 - Cellular
 - Satellite
 - Voice over IP (Internet)
 - Video phones
 - Storage of Voice messages for computer voice mail applications.

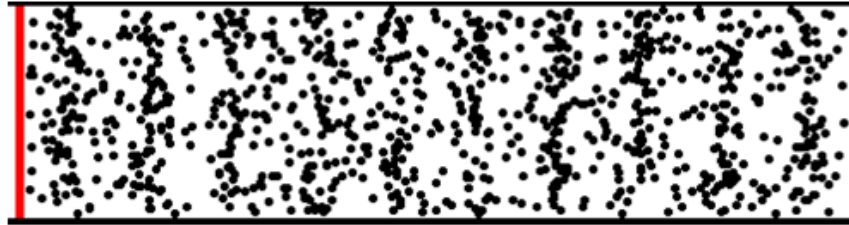
Possible Projects

- Speech Enhancement
 - Goal is to improve the quality of degraded speech.
 - Preprocess speech before is degraded:
 - Increasing the broadcast range of transmitters constrained by a peak power transmission limits (e.g., AM radio and TV transmissions).
 - Enhancing the speech waveform after it is degraded.
 - Reduction of additive noise in
 - » (Digital) telephony
 - » Vehicle and aircraft communications
 - Reduction of interfering backgrounds and speakers for the hearing impaired,
 - Removal of unwanted convolutional channel distortion and reverberation
 - Restoration of old phonograph recordings degraded by:
 - » Acoustic horns
 - » Impulse-like scratches from age and wear

Possible Projects

- Speaker Recognition
 - Speech signal processing exploits the variability of speech model parameters across speakers.
 - Verifying a person's identity (Biometrics)
 - Voice identification in forensic investigation.
 - Understanding of the speech model features that cue a person's identity is also important in speech modification where model parameters can be transformed for the study of specific voice characteristics:
 - Speech modification and speaker recognition can be developed synergistically.
- Marmoset monkey vocalization
 - We have an excellent dataset of monkey calls. We would like to build a classifier for identifying specific types of calls.

What is sound?



Sound is oscillation of air pressure (pressure wave).

high pressure: air molecules bunched up

low pressure: air molecules spread out

Air molecules do **not** travel through
space to carry sound

Sound is a SIGNAL

Tuning fork

- A tuning fork is a device made of steel or magnesium that is used to tune musical instruments or by singers to obtain certain pitches
- It emits a pure tone at a particular frequency
- The frequency depends on the size and shape of the device
- Originally used by audiologists to evaluate hearing

Tuning fork



FIGURE 2.3 Several tuning forks. The larger forks vibrate at lower frequencies (produce lower-pitched tones) than the smaller forks.

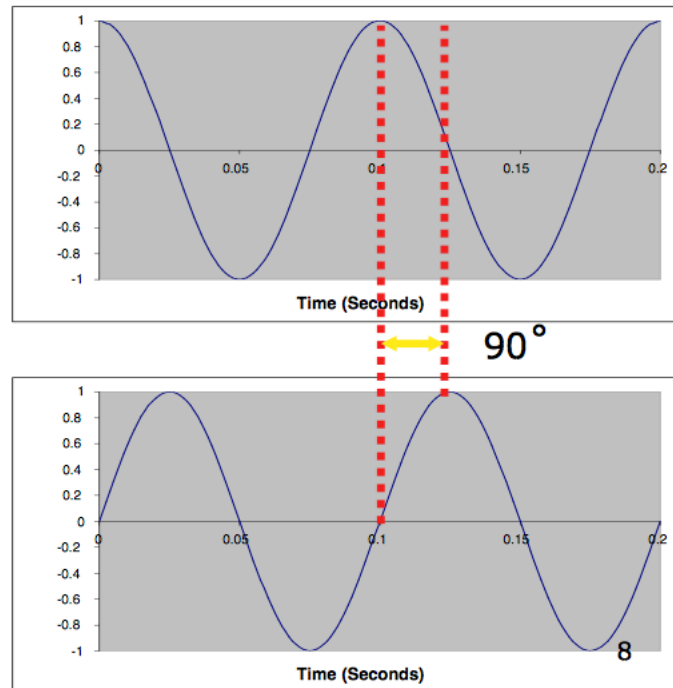


FIGURE 2.4 Vibration pattern of tuning forks.

What is sound?

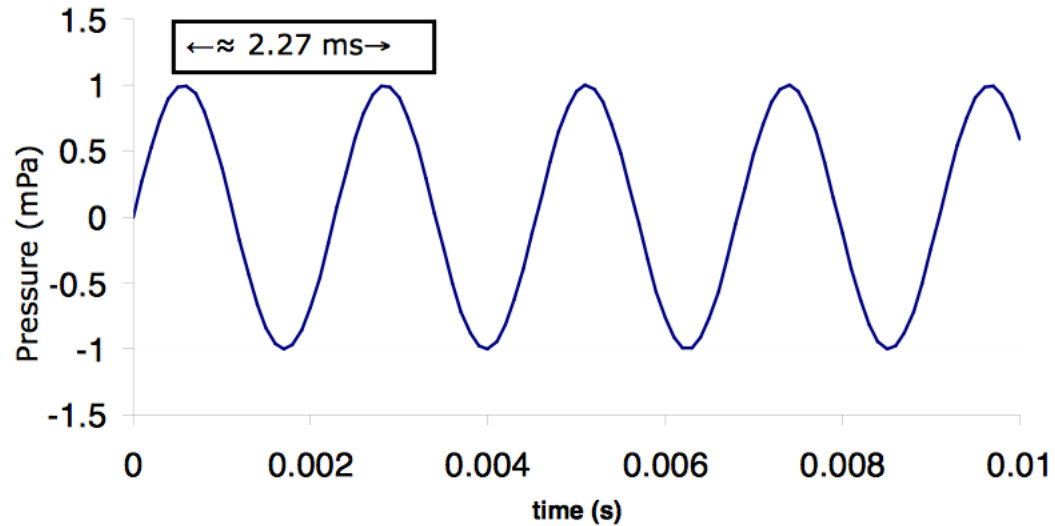
I: Phase

- Where a sinewave *starts* at some arbitrary time
- Measured in cycles or degrees (or radians)
 - $360^\circ = 1 \text{ period}$
 - $180^\circ = \frac{1}{2} \text{ period}$
 - $90^\circ = \frac{1}{4} \text{ period}$
- Equivalent to a shift in time
- Relatively little effect on perception but still important in many situations



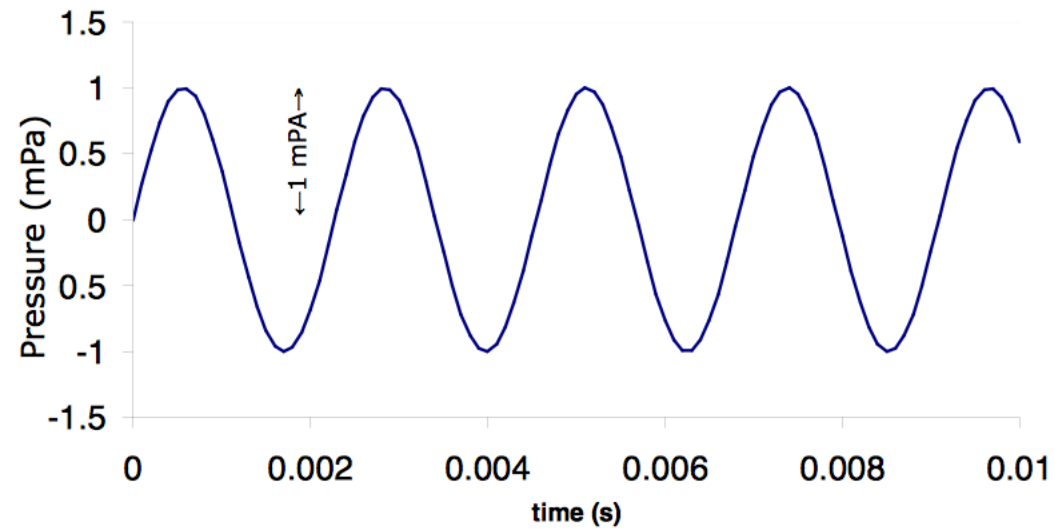
What is sound?

II: Periodicity (frequency)

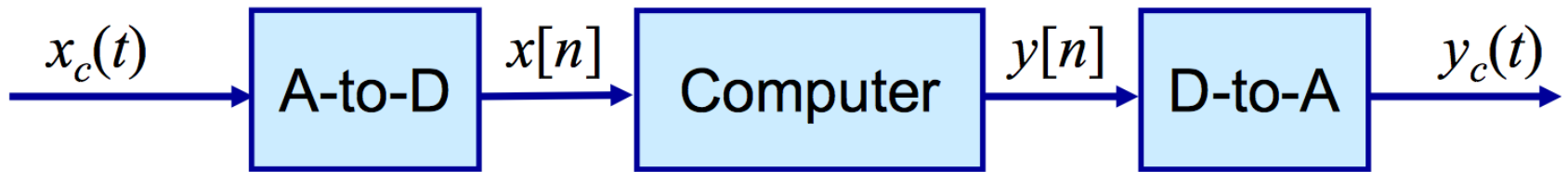


What is sound?

III: Amplitude



Digital Processing of Analog Signals



- **A-to-D conversion:** bandwidth control, sampling and quantization
- **Computational processing:** implemented on computers or ASICs with finite-precision arithmetic
 - **basic numerical processing:** add, subtract, multiply (scaling, amplification, attenuation), mute, ...
 - **algorithmic numerical processing:** convolution or linear filtering, non-linear filtering (e.g., median filtering), difference equations, DFT, inverse filtering, MAX/MIN, ...
- **D-to-A conversion:** re-quantification* and filtering (or interpolation) for reconstruction

Discrete-Time Signals

- A sequence of numbers
- Mathematical representation:

$$x = \{x[n]\}, \quad -\infty < n < \infty$$

- Sampled from an analog signal, $x_a(t)$, at time $t = nT$,

$$x[n] = x_a(nT), \quad -\infty < n < \infty$$

- T is called the **sampling period**, and its reciprocal, $F_s = 1/T$, is called the **sampling frequency**

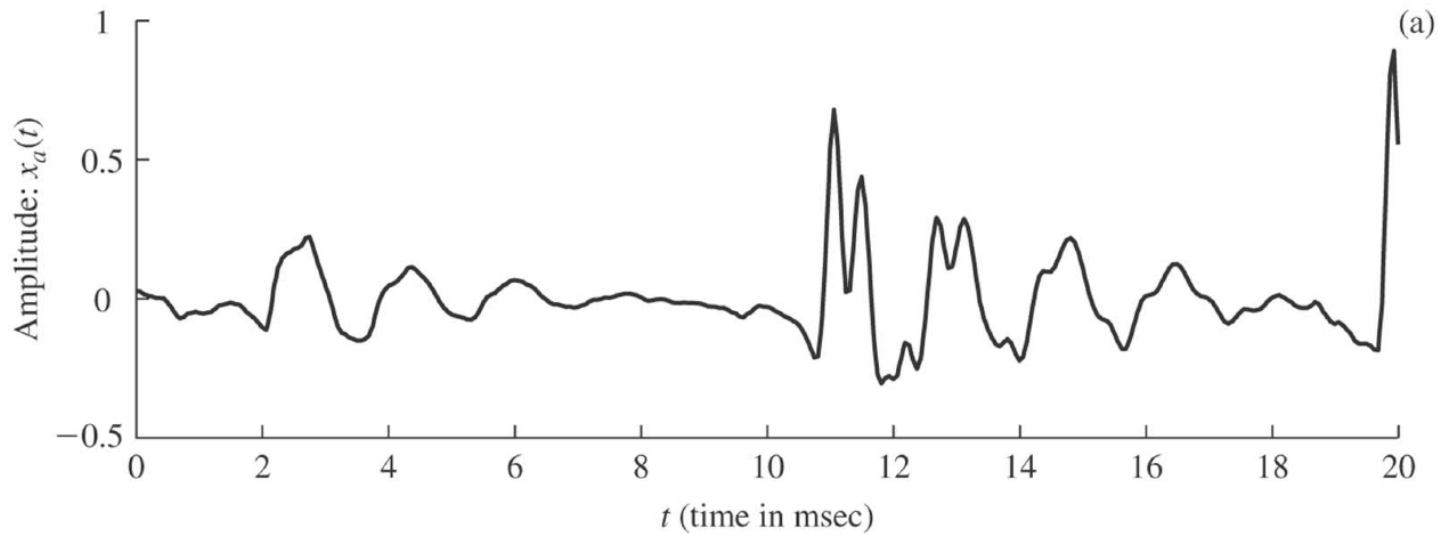
$$F_s = 8000 \text{ Hz} \leftrightarrow T = 1/8000 = 125 \mu\text{sec}$$

$$F_s = 10000 \text{ Hz} \leftrightarrow T = 1/10000 = 100 \mu\text{sec}$$

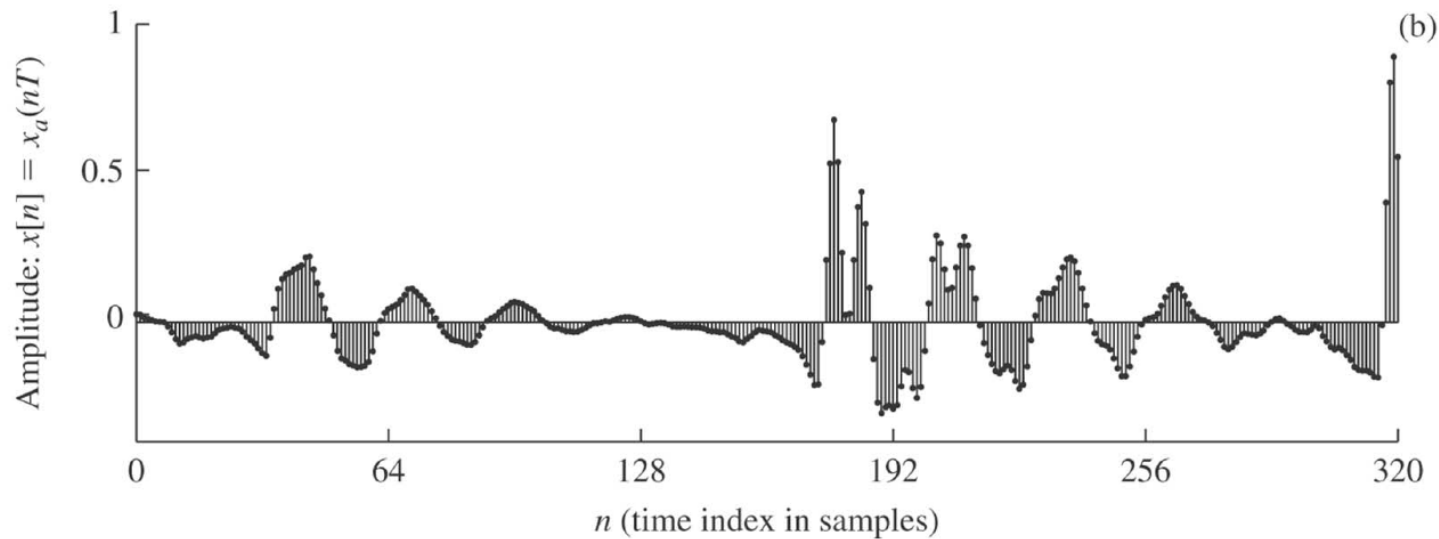
$$F_s = 16000 \text{ Hz} \leftrightarrow T = 1/16000 = 62.5 \mu\text{sec}$$

$$F_s = 20000 \text{ Hz} \leftrightarrow T = 1/20000 = 50 \mu\text{sec}$$

Speech Waveform Display

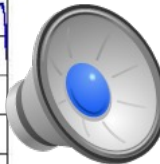
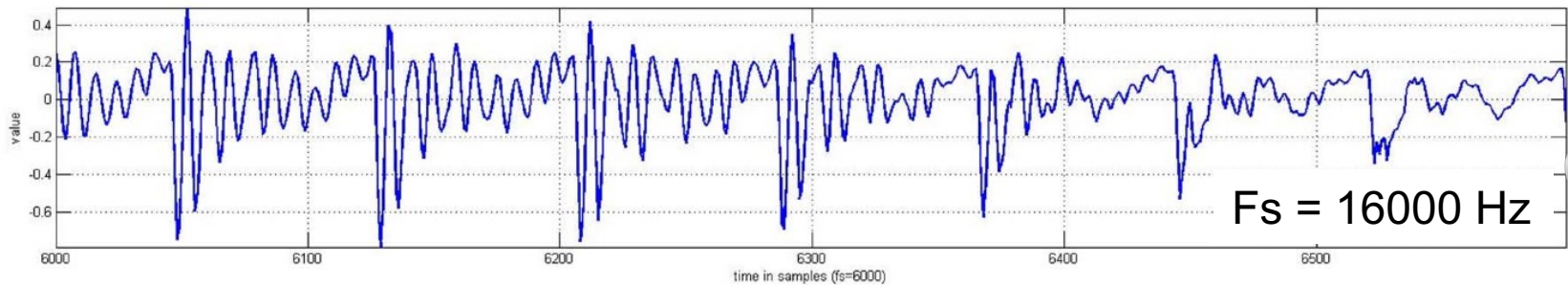
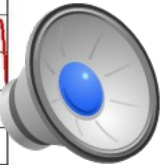
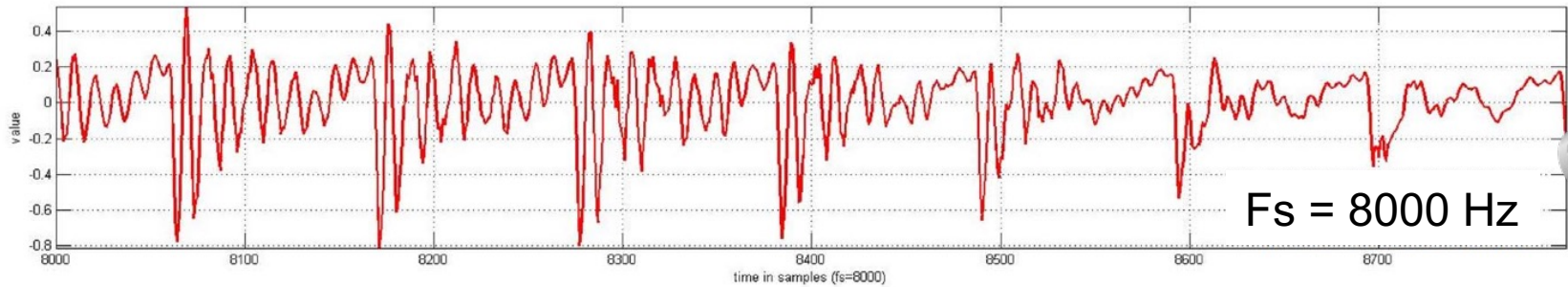


`plot();`



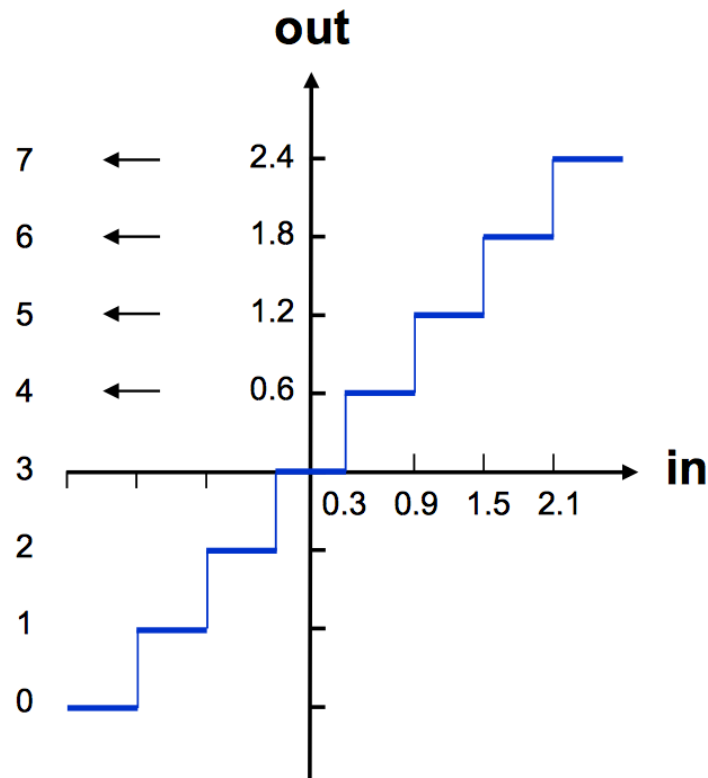
`stem();`

Varying Sampling Rates



Quantization

□ $x[n]$ can be quantized to one of a finite set of values which is then represented digitally in bits, hence a truly digital signal; the course material mostly deals with discrete-time signals (discrete-value only when noted).



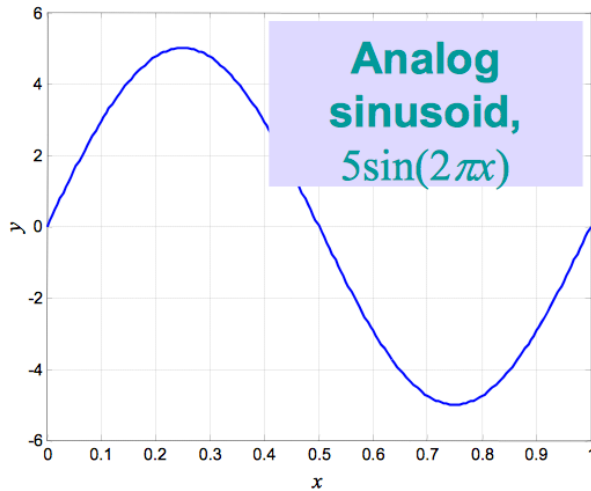
A 3-bit uniform quantizer

Quantization:

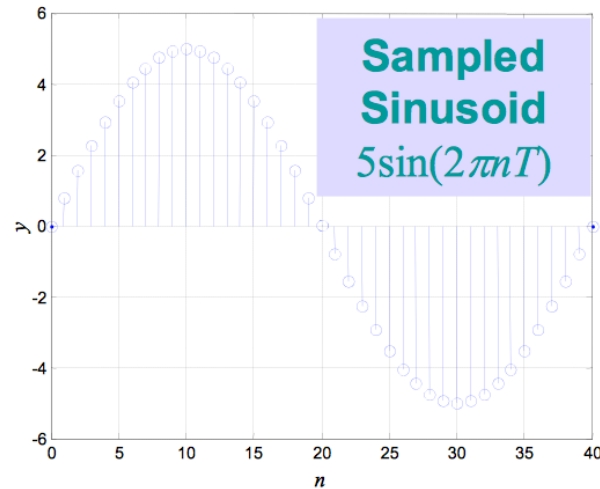
- Transforming a continuously-valued input into a representation that assumes one out of a finite set of values
- The finite set of output values is indexed; e.g., the value 1.8 has an index of 6, or $(110)_2$ in binary representation
- Storage or transmission uses binary representation; a quantization table is needed

What is sound?

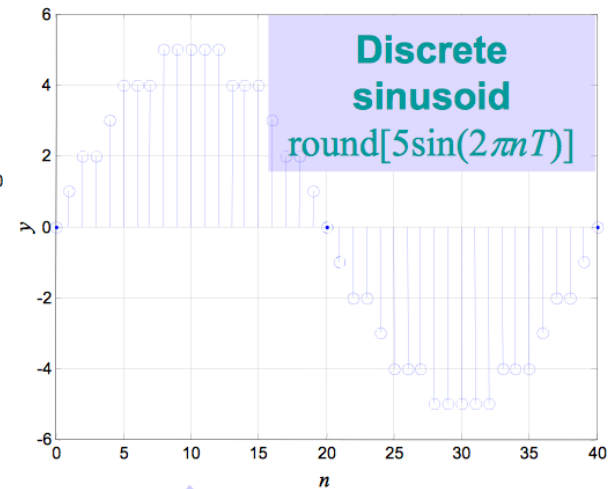
sample



quantize



quantize



sample

