CS4347/CS5647

Sound and Music Computing (SMC)

L1: Introduction

LECTURER:

Wang Ye

www.comp.nus.edu.sg/~wangye

wangye@comp.nus.edu.sg

Teaching team

WANG Ye (Lecturer), wangye@comp.nus.edu.sg

OU Longshen (TA), oulongshen@u.nus.edu

ZHAO Jingwei (TA), jzhao@u.nus.edu

Right Infringements on NUS Course Materials

All course participants (including permitted guest students) who have access to the course materials on CANVAS or any approved platforms by NUS for delivery of NUS modules are not allowed to re-distribute the contents in any forms to third parties without the explicit consent from the module instructors or authorized NUS officials

Today's topics

1) Practical arrangements of the course
2) Human auditory perception and interaction, and their relationships to other subjects
3) Recap of some key concepts of Digital Signal Processing (DSP)
4) Introduction to a new educational model which guides the course in practice

Topics to Cover (*selective approach*)

Part A: The Core

- Introduction
- Review of DFT, Audio Representation, and Machine Learning
- Music Representation, Analysis and Transcription
- Automatic Music Transcription (AMT)
- Automatic Speech Recognition (ASR)
- Generative Models for Text-to-Speech (TTS) & Singing Voice Synthesis (SVS)

Midterm break

Part B: The Breadth

- Singing voice processing
- Music production audio effects
- Automatic Music Generation
- Synthesis of sound & music – a DSP approach
- Project presentations/demo

CS4347/CS5647 Course Description

- This course introduces the fundamental technologies employed in Sound and Music Computing focusing on three major categories: **speech, music**, and environmental sound. This course introduces the concept of sound and its representations in the analog and digital domains, as well as in time and frequency domains. Moreover, this course provides hands-on instruction on relevant machine learning tools, and an in-depth review of related technologies in sound data **analytics** (Automatic Speech Recognition, Automatic Music Transcription and Sound Event Detection) leading to a group project. Topics in sound **synthesis**, automatic music generation and music information retrieval will be covered for breadth.

Module Learning Outcomes (MLO):

- Understand Discrete Fourier Transform (**DFT**) in the context of audio analysis and synthesis.
- Understand, be able to implement and evaluate the approaches to automatic music transcription (**AMT**)
- Understand the building blocks of automatic speech recognition (**ASR**) systems, be able to implement and evaluate an ASR system with a SOTA toolkit such as Speechbrain
- **Present solutions in both oral and written formats**, and discuss with other students on sound and music computing

Why are oral presentation skills so important?

Let me address this question with a speech communication model
from the Shannon information theory's perspective (*entropy*)

Good speakers keep the ***uncertainty*** as low as possible to ensure accurate decoding of their message. Conversely, ***bad presentation*** makes it a mentally taxing task.

One of the most transferable skills!

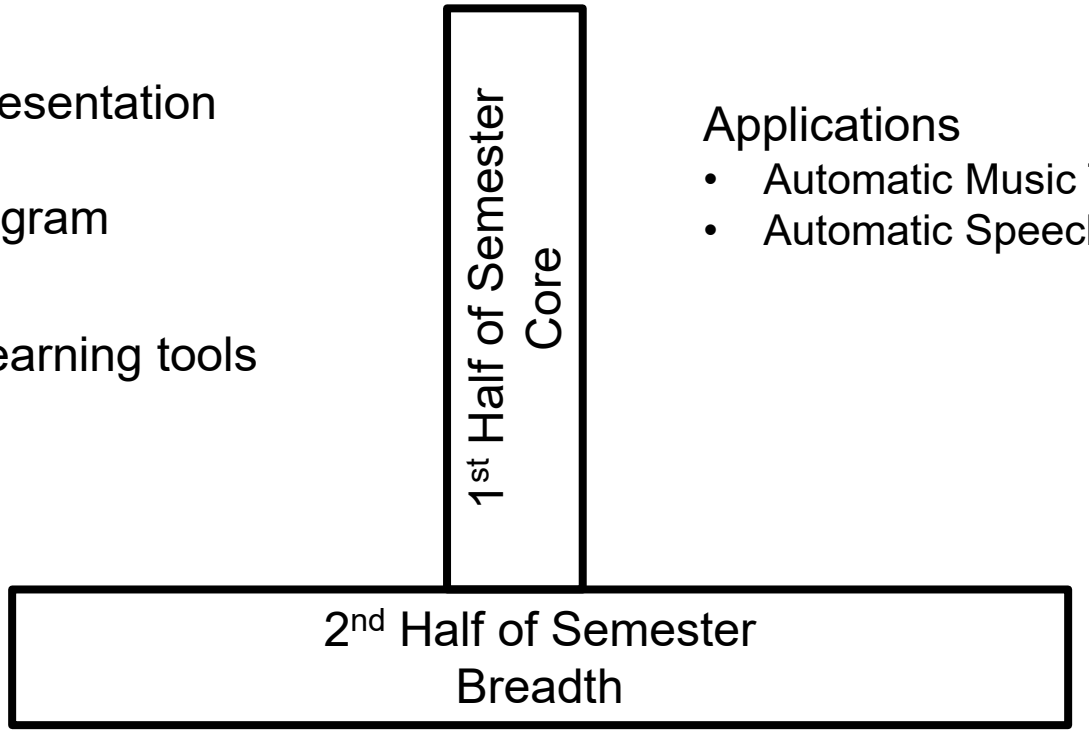
Real life applications:
From a job interview to project or conference presentation

Core and Breadth (Approach)

DSP *and* ***ML*** *techniques to solve real-world problems with* ***speech & music****.*

Audio representation
- DFT
- Spectrogram
- MFCC

Machine learning tools



Applications
- Automatic Music Transcription (AMT)
- Automatic Speech Recognition (ASR)

Assessment (100% CA):

- **Participation effort** **15%**
  - Including lectures, tutorials, survey, *Canvas*, etc.

- **3 Individual assignments** **40%**
  - Week 2 (DFT) 10%
  - Week 4 (AMT) 15%
  - Week 6 (ASR) 15%

- **1 Group project** **45%**
  - Week 10 (Mid-project assessment) 5%
  - Week 13

  1) Presentation 10%
  2) Code 10%
  3) Final report 20%

  You can propose your own project and form your own project team.

Schedule

- **Lectures on Monday 4-6pm LT15**
  - **Lectures will be conducted offline in a f2f manner. Physical attendance is required for the best learning experience.**
- **Tutorials from the third week**
  - **We will use the Canvas forum extensively for discussions, Q/As etc. However, tutorials are essential parts of learning.**

Resource:

- **Course Web-Page:**
  - https://canvas.nus.edu.sg/ and
  - PLEASE check regularly
  - All announcements are posted regularly
- **Consultation with the teaching team:**
  - Canvas will serve as the primary communication channel
  - Appointment by email

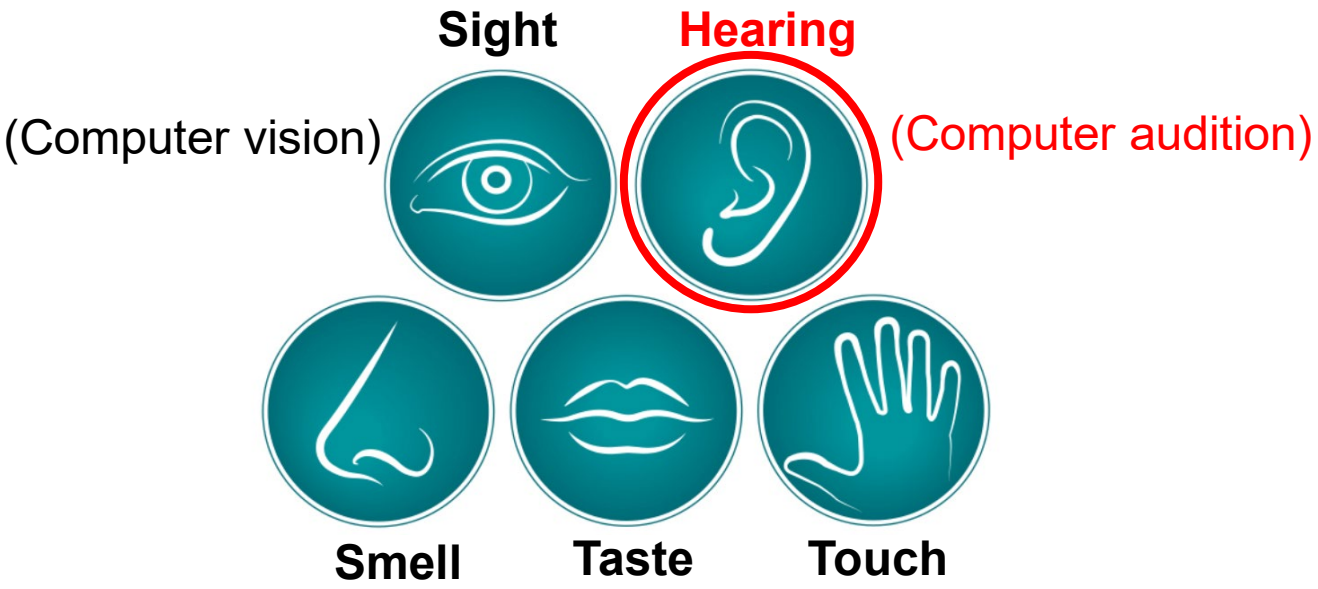
References

- **Reference books + conference and journal papers**
  - Relevant conferences: **ISMIR**, ACMMM, ICME, ICASSP, WASPAA, **InterSpeech**, DCASE (Detection and Classification of Acoustic Scenes and Events)
  - NUS digital library
  - Internet is another great resource!

Today's topics

1) Practical arrangements of the course
2) Human auditory perception and interaction, and their relationships to other subjects
3) Recap of some key concepts of Digital Signal Processing (DSP)
4) Introduction to a new educational model which guides the course in practice

Five basic senses of human

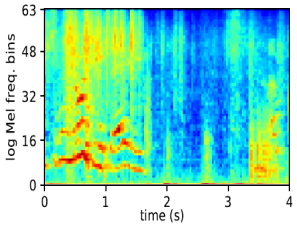


The five basic senses were proposed by Aristotelian [1].

Image source: https://www.livescience.com/60752-human-senses.htm
[1] https://en.wikipedia.org/wiki/Sense#Aristotelian_senses

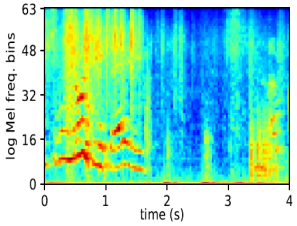
Relation of computer audition (CA) & computer vision (CV)

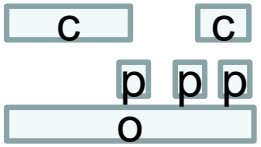
Audio tagging



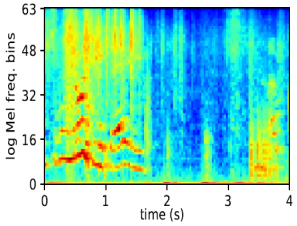
children
percussion
others

Sound event detection





Source separation



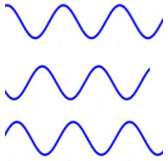
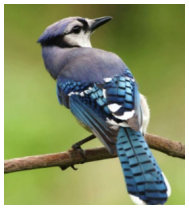


Image classification



Bird

Image localization

Image segmentation

Relation of computer audition (CA) & computer vision (CV)

Speech synthesis

Image generation

Audio 3D reconstruction

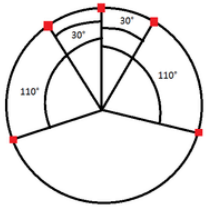


Image 3D reconstruction

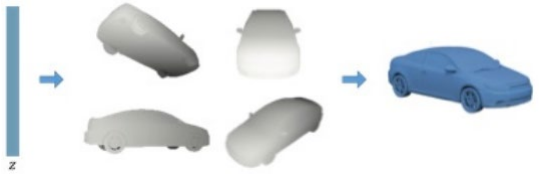
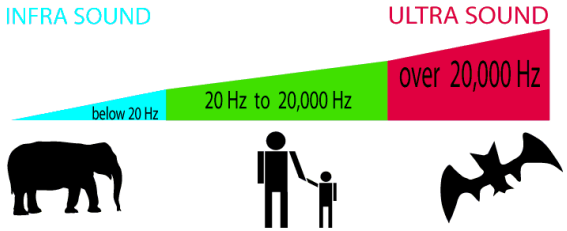
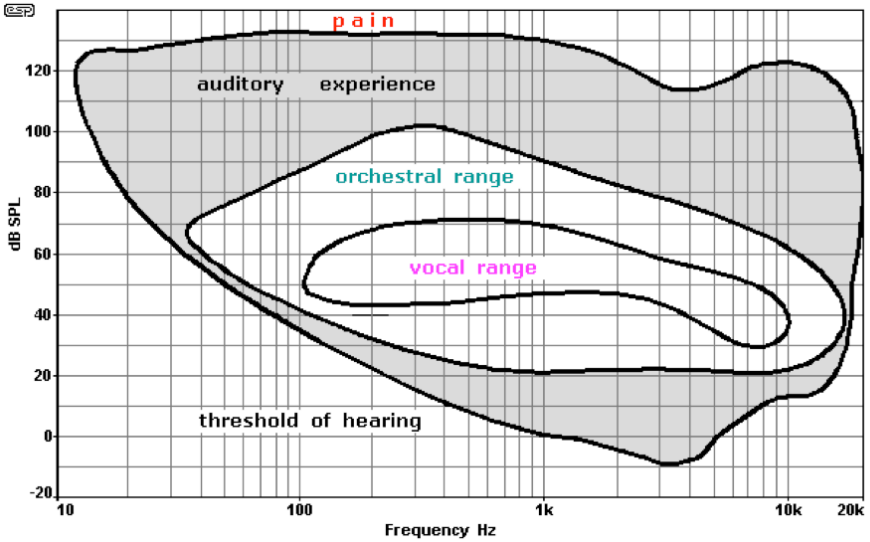


Image generation source: https://www.google.com/url?sa=i&source=images&cd=&ved=2ahUKEwjxtKiWoJzmAhUIohQKHWPrDjoQjRx6BAgBEAQ&url=https%3A%2F%2Fmachinelearningmastery.com%2Fa-gentle-introduction-to-the-biggan%2F&psig=AOvVaw2IjAxRRuWNcn1CHN8UALgT&ust=1575557738531773
Audio 3D reconstruction source: https://www.google.com/url?sa=i&source=images&cd=&ved=2ahUKEwjMoMO0oZzmAhUM0uAKHbedCIsQjRx6BAgBEAQ&url=https%3A%2F%2Fen.wikipedia.org%2Fwiki%2F3D_sound_reconstruction&psig=AOvVaw3vWInEMmIZLwy-FyQ0wOif&ust=1575558042289727
Image 3D reconstruction source: https://upload.wikimedia.org/wikipedia/commons/6/6d/Synthesizing_3D_Shapes_via_Modeling_Multi-View_Depth_Maps_and_Silhouettes_With_Deep_Generative_Networks.png

Human hearing range

- **Sound** is a vibration that typically propagates as an audible wave of pressure, through a transmission medium such as a gas, liquid or solid [1].
- **Infra sound**: lower than 20 Hz.
- **Ultra sound**: higher than 20 kHz.
- Audio is recorded sound in digital format.
- ***Human*** *has a hearing range between 20 Hz and 20 kHz.*
- **Dogs** have a hearing range of 67 Hz - 45 kHz.
- ***Vocal*** *sound has an approximate range between 100 Hz and 10 kHz.*
- **Music** has an approximate range between 50 Hz and 20 kHz.

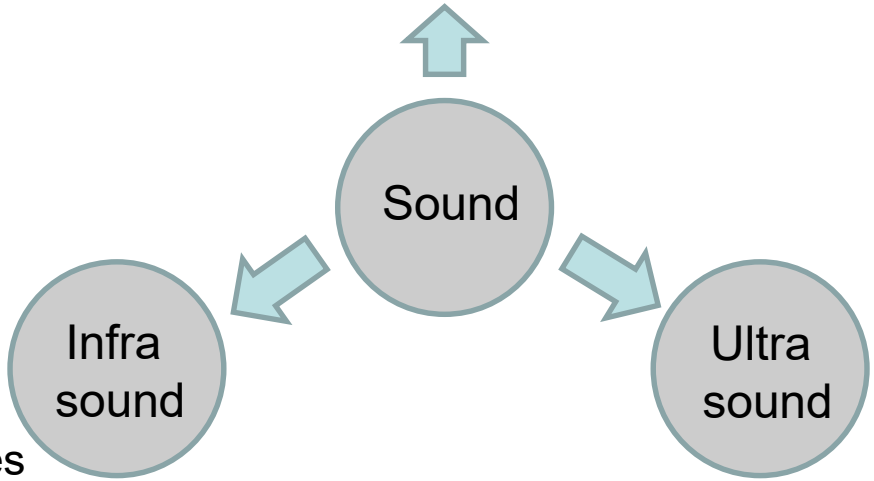




[1] https://en.wikipedia.org/wiki/Sound
Top image source: https://www.google.com/human%20hearable%20sound%20frequency
Bottom image source: https://www.google.com/human%20hearable%20sound%20frequency

Taxonomy of sound

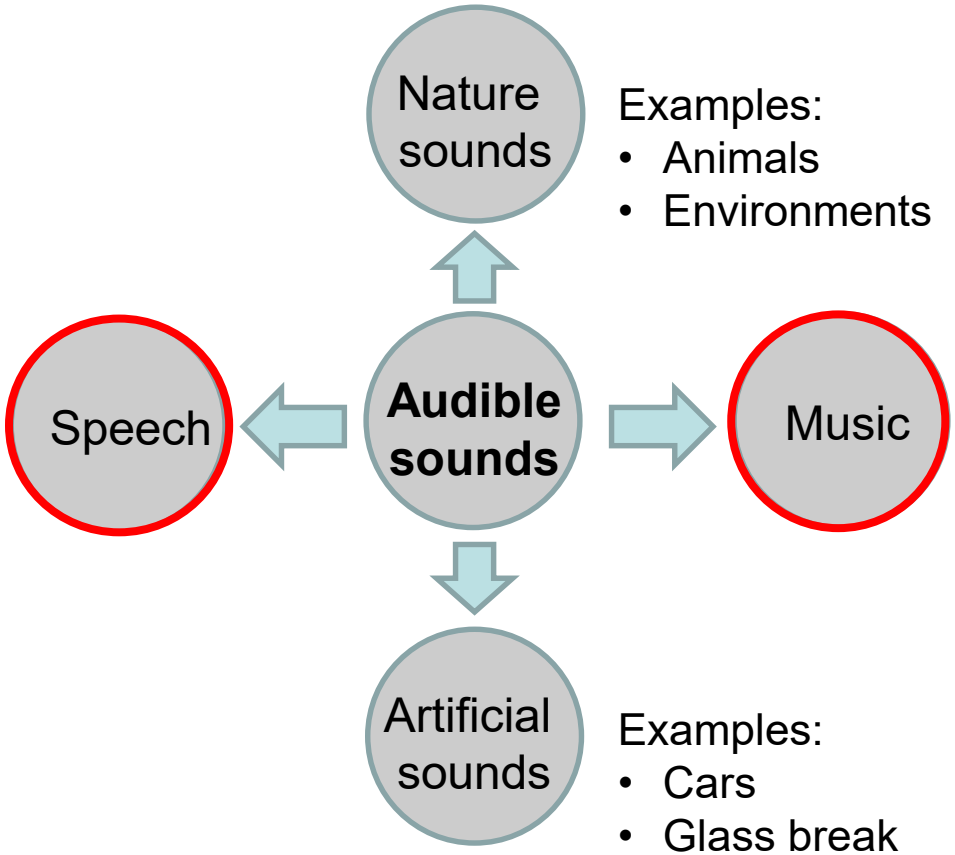
Focus of this course



Examples:
- Elephants
- Whale
- Earthquakes
- Volcano

Examples:
- Bats
- Dolphins
- Teeth cleaning
- Acoustic microscopy

Taxonomy of sound

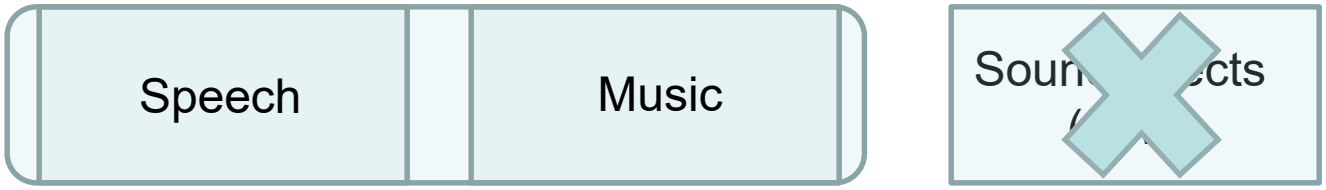


Evolution of CS4347

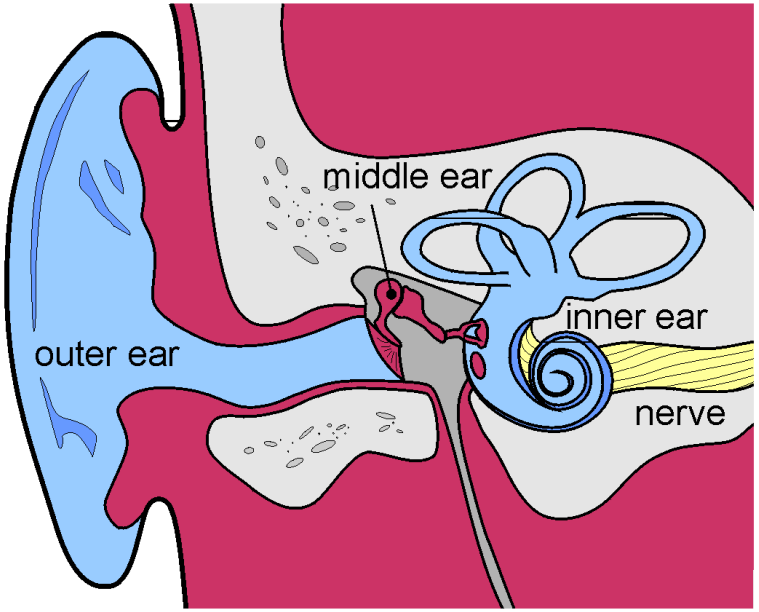
CS4347: Sound and Music Computing	CS5241: Speech Processing

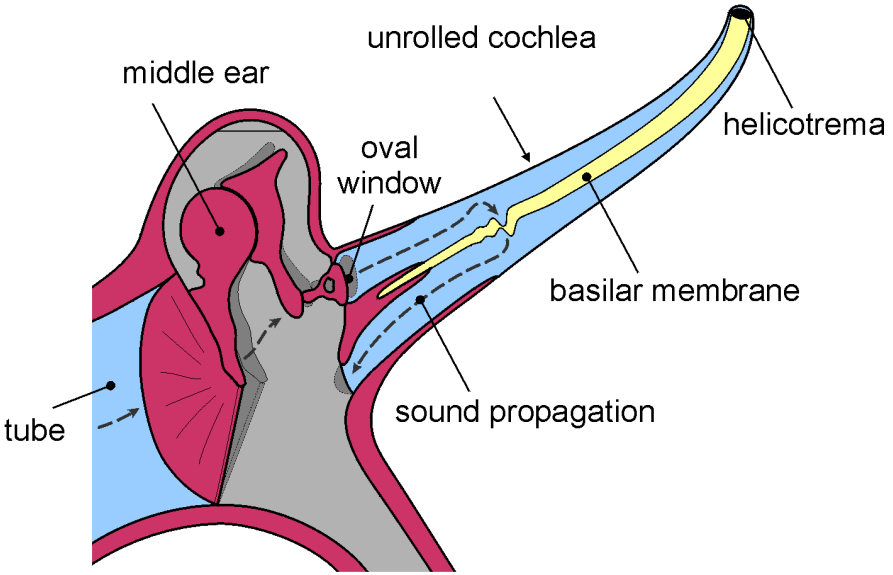
Music analysis and synthesis are becoming hot AI topics

Speech analysis and synthesis are core AI problems



Our Peripheral Auditory System

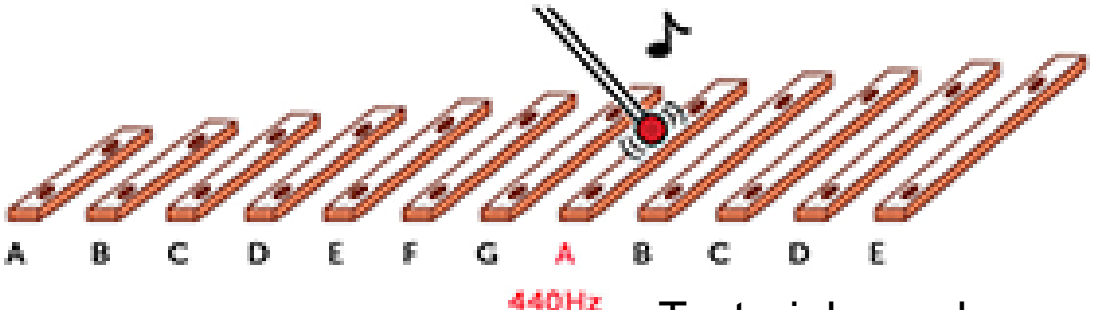




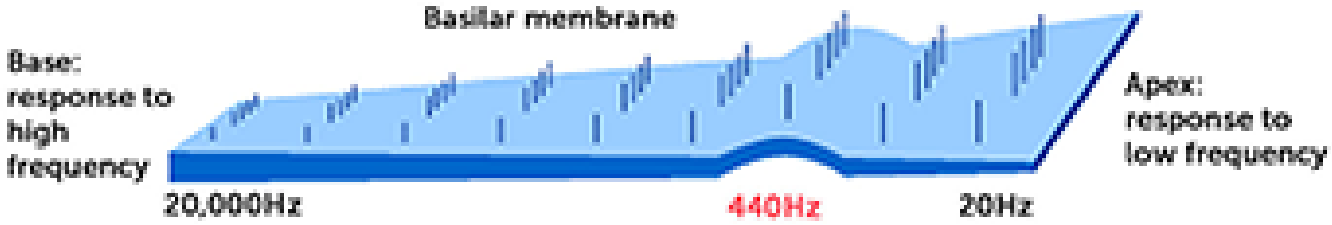
Bandpass filter

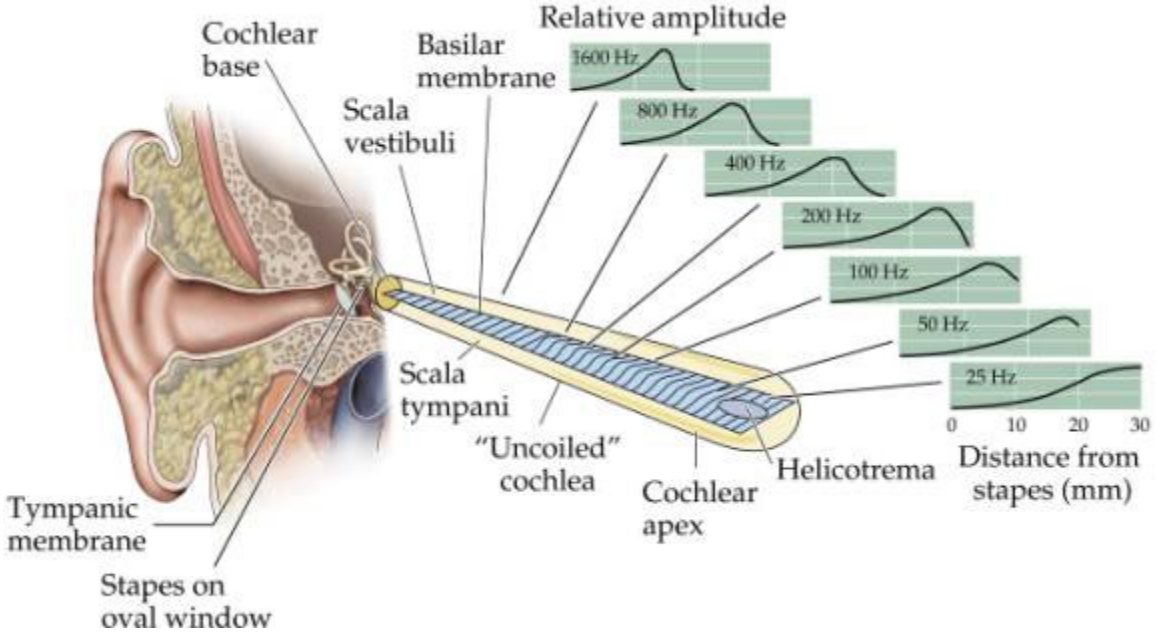
Dynamic range
compressor (log)

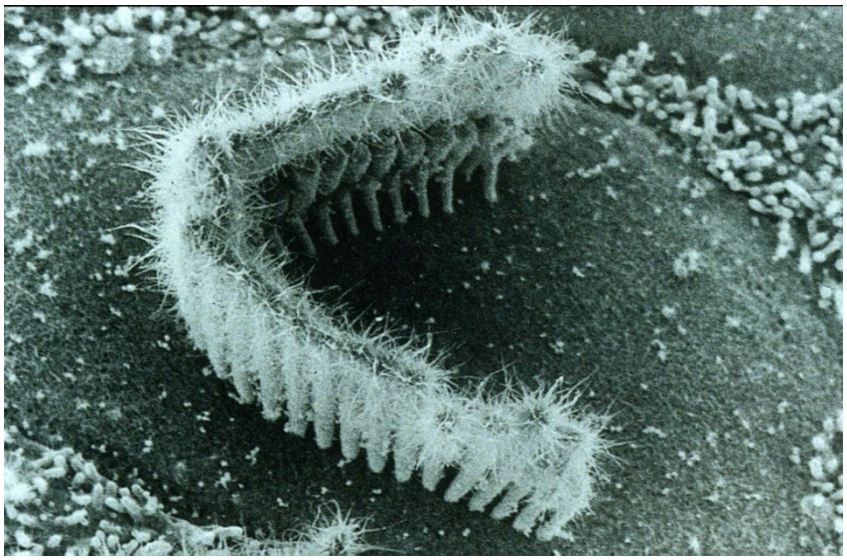
Frequency analyzer
(DFT)

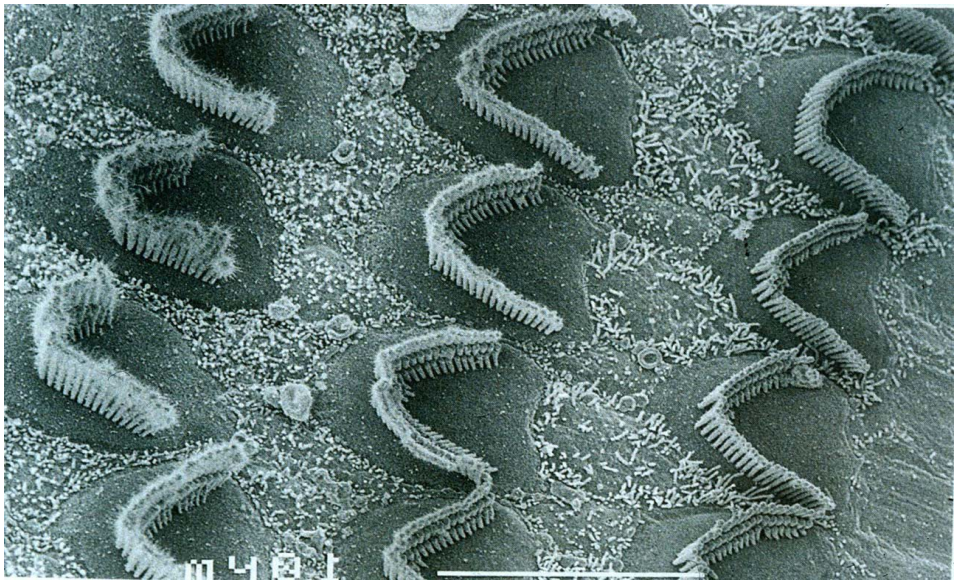


Tectorial membrane



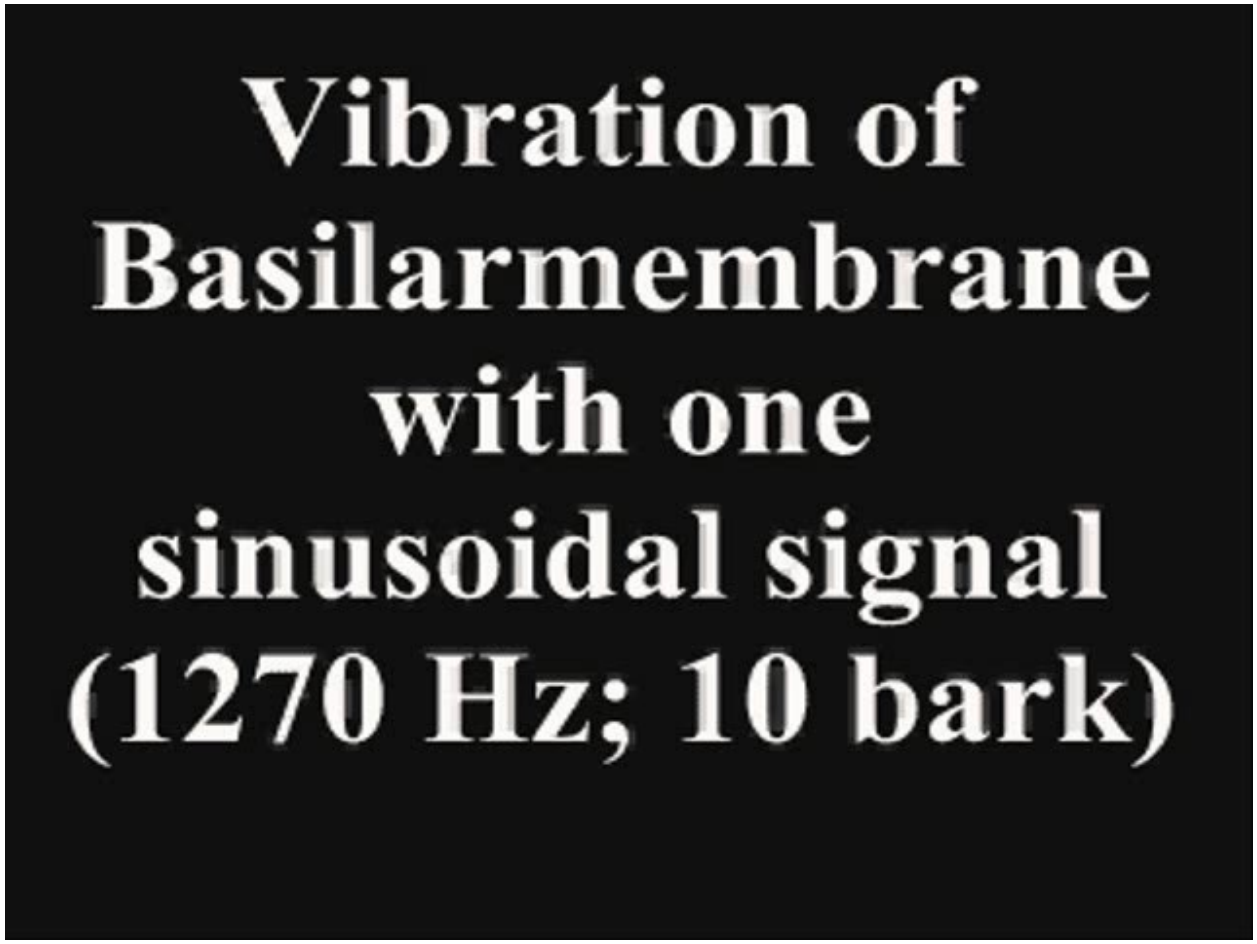






Hair Cells in the Ear

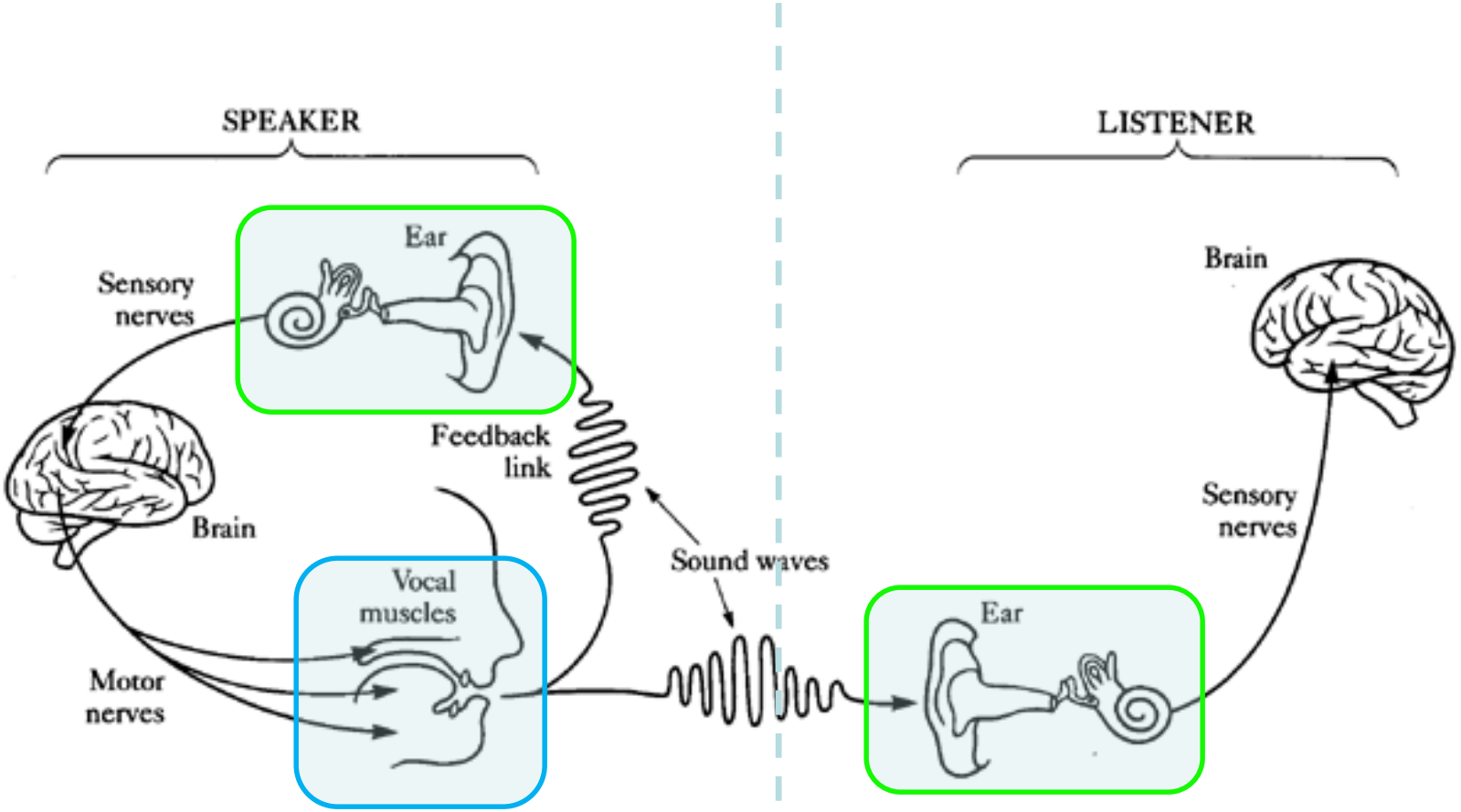
Fourier Transform in Our Ear!



It is a physiological justification for the short-time Fourier transform (STFT)!

Human auditory communication

The core theme of this semester is
singing voice analysis which bridges ASR and AMT!



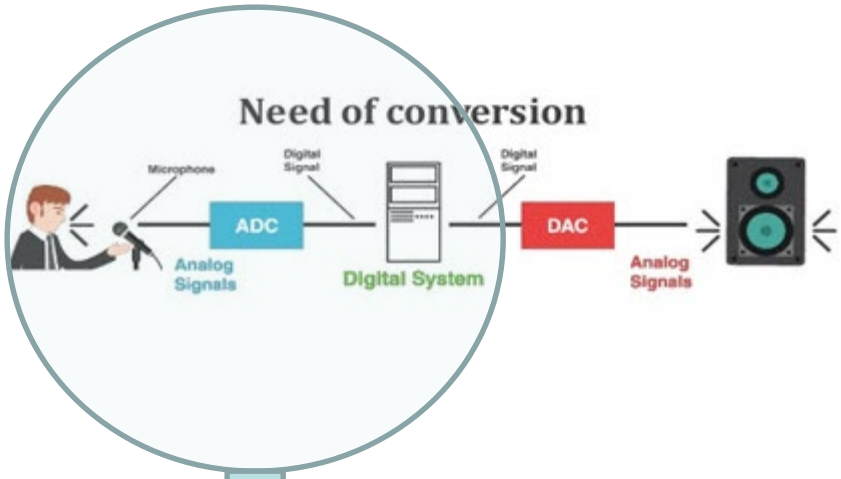
synthesis

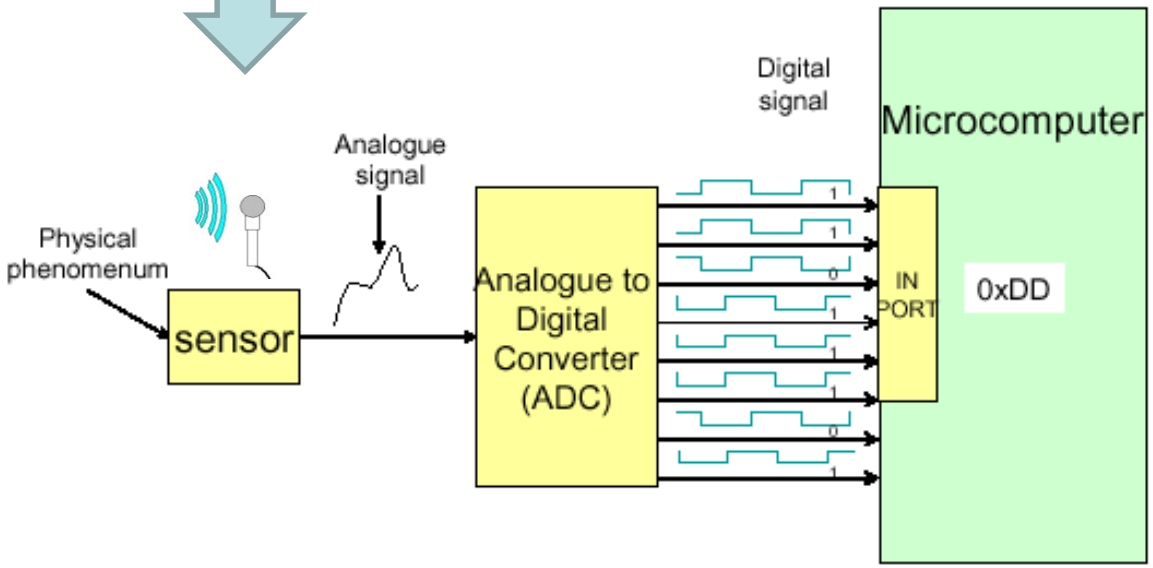
analysis

Today's topics

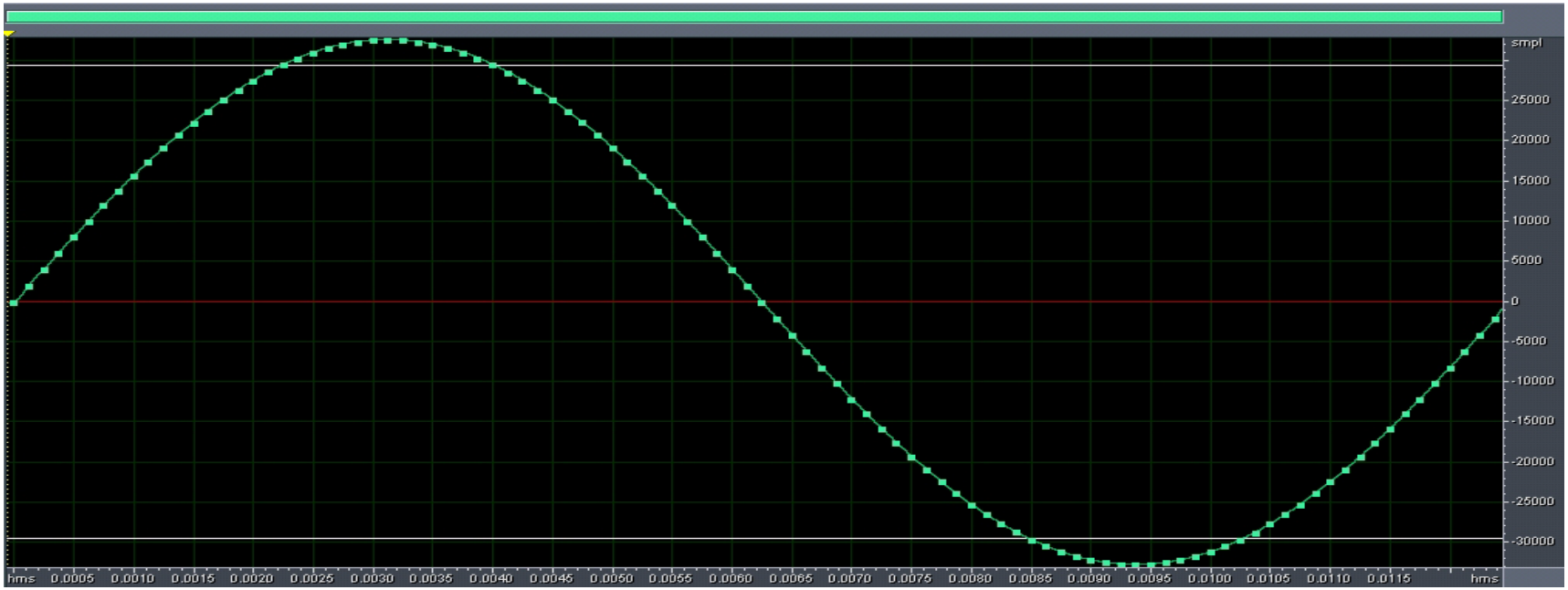
1) Practical arrangements of the course
2) Human auditory perception and interaction, and their relationships to other subjects
3) Recap of some key concepts of Digital Signal Processing (DSP)
4) Introduction to a new educational model which guides the course in practice

Digitization of Audio Signals





Digitization of Audio Signals



Sampling rate – samples per second (8kHz, 22.05kHz, 44.1kHz, 88.2kHz, 96kHz)

Bit depth – number of bits per sample 8, 16, 24

What happens if we change the sampling frequency (e.g., from 8k to 16k) of a music signal during playback?

What happens if we change the Fs during the playback?

A few Fundamental Concepts

Let us review a few fundamental concepts when converting from analogue to digital representations:

1. Discretization: Consider the process of digitization. What is the "**sampling**" rate? How many bits are recorded per sample? (This is called "**quantization**".)
2. **Discrete Fourier Transform (DFT):** a common technique to convert a time domain signal to a frequency domain representation.
3. Filtering: Real-time techniques to process a signal (e.g., lowpass filtering).

Summary of the Sampling Theorem

Nyquist rate - For lossless discretization, the sampling rate should be *at least twice* the maximum frequency responses.

- **i.e., Fs > 2B where Fs is the sampling rate and B is the expected bandwidth.**

Demo (Aliasing and bit depth)

- **Question: What happens to audio when it is sampled at less than 2B?**

a) Original audio (**Fs=44100**), b) Down-sampled without prefiltering (Fs=5512.5), c) Even more down-sampled (Fs=2756.25).



What Type of Numbers?

- how PRECISE are the numbers?
  - bit depth: generally signed 16-bit integers (-32,768 to 32,767)
  - usually interpreted as being from -1.0 to 1.0, but not always
- how MANY numbers are there per unit time?
  - sampling rate: generally 44,100 Hz (samples/second)
- Are these numbers always accurate?
  - not really: signal + noise
  - applications: noise removal, separating audio into distinct audio sources, etc.

Bit Depth

- Occasionally 8-bit in the past, and sometimes 24-bit ints or floats.
- Dynamic range: humans can perceive ~140 dB, normal concert less than 80 dB, human speech around 40 dB.

- Bit depth b to dynamic range R: $R = 20 \cdot \log 2^b$
  - 16-bit audio ≈ 96 dB
  - 24-bit audio ≈ 144 dB
- These are *theoretical* dynamic ranges - other considerations are the playback system (e.g., tiny ipod speakers, “home theatre" speakers), environmental noise, dithering.

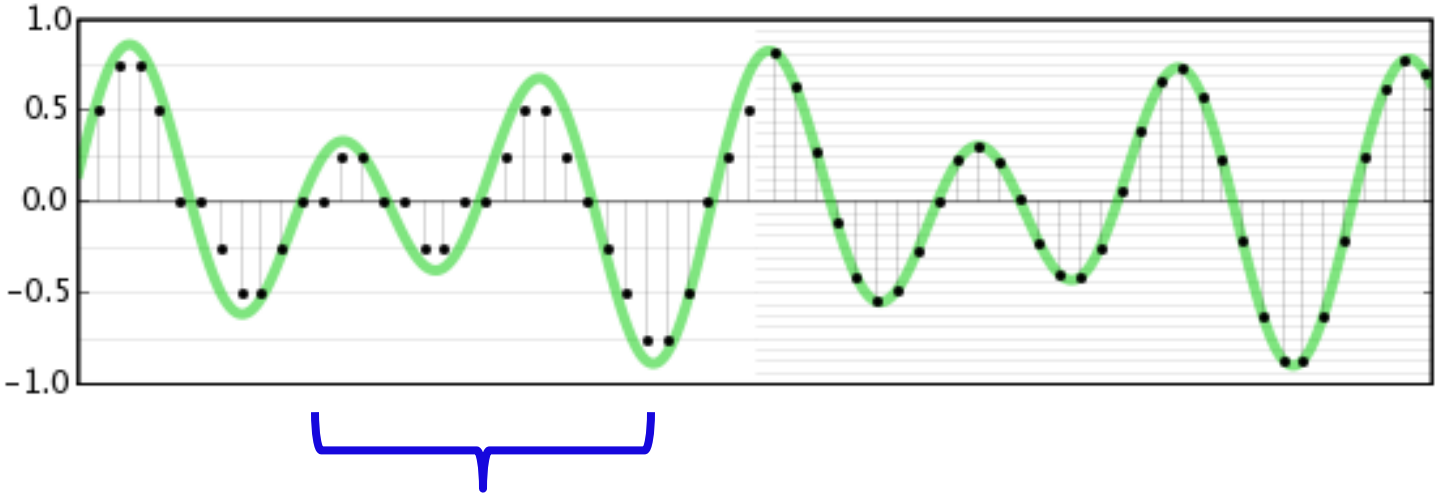
Sampling Rate

- 44.1 kHz most of the time; occasionally 48 kHz or 96 kHz.

- Human perception rule of thumb: 20 - 20,000 Hz
  - Adults (20-30 years) probably limited to 16 or 17 kHz
  - Loss of upper frequencies perception due to age and exposure to loud music (rock concerts, high volume thru headphones, etc.)

- **Nyquist-Shannon sampling theorem**: highest frequency that can be represented by a digital signal is **HALF** of the sampling rate (we will show you why later)
  - 20kHz → 40kHz; a bit of extra for low-pass filter roll-off → 44.1kHz

Mimicking Human Hearing



Rationale for block-wise signal processing and DFT?

This mimics human auditory perception of sound/music!

This is the most important equation that you have learnt in *CS2108 Introduction to Media Computing (prerequisite)*.

DFT Formula:

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

What are your key insights into the Discrete Fourier Transform (DFT) formula?

Or what is the physical interpretation of X[k]?

Today's topics

1) Practical arrangements of the course
2) Human auditory perception and interaction, and their relationships to other subjects
3) Recap of some key concepts of Digital Signal Processing (DSP)
4) Introduction to a new educational model which guides the course in practice

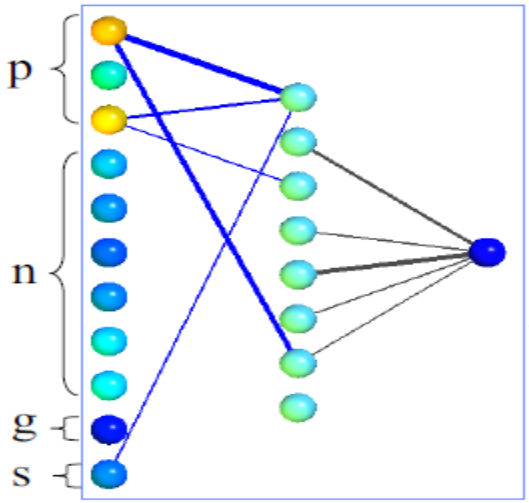
Active and Joyful Learning

Diet *Sleep* *Exercise* *Social interaction*

Ingredients to sustain the **flame** (healthy & happy life)!

How to achieve this?

Learning when exhausted

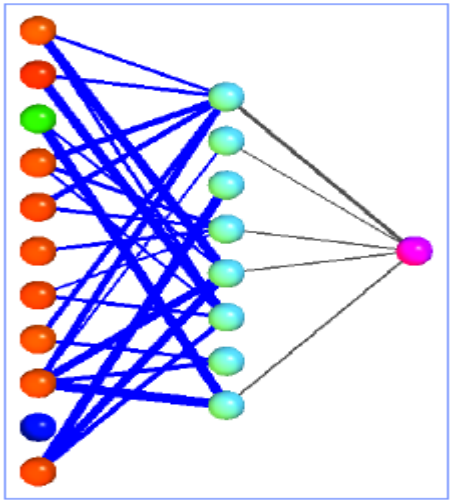


A job to teach

Socrates (c. 470 BC - 399 BC, Athens):

Education is the kindling of a **flame**, not the filling of a vessel.

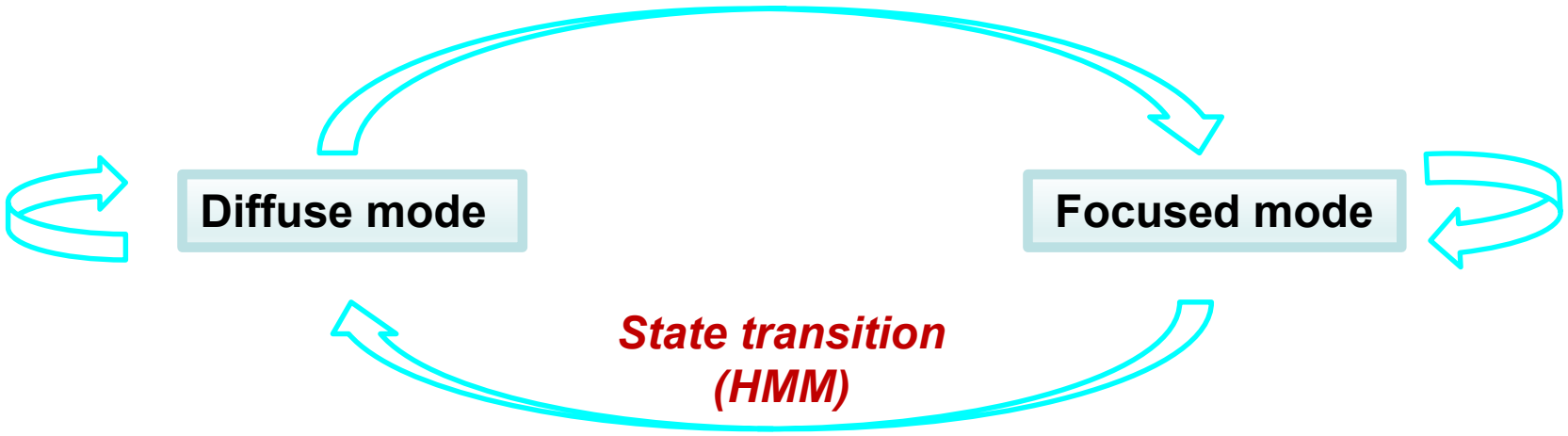
Learning when fully energized



A joy to teach

Take care of your brain (stress management)! But how?

My educational philosophy:
To cultivate active and joyful learning



Learning How to Learn | Barbara Oakley (neuroscience based methods)
https://www.youtube.com/watch?v=O96fE1E-rf8

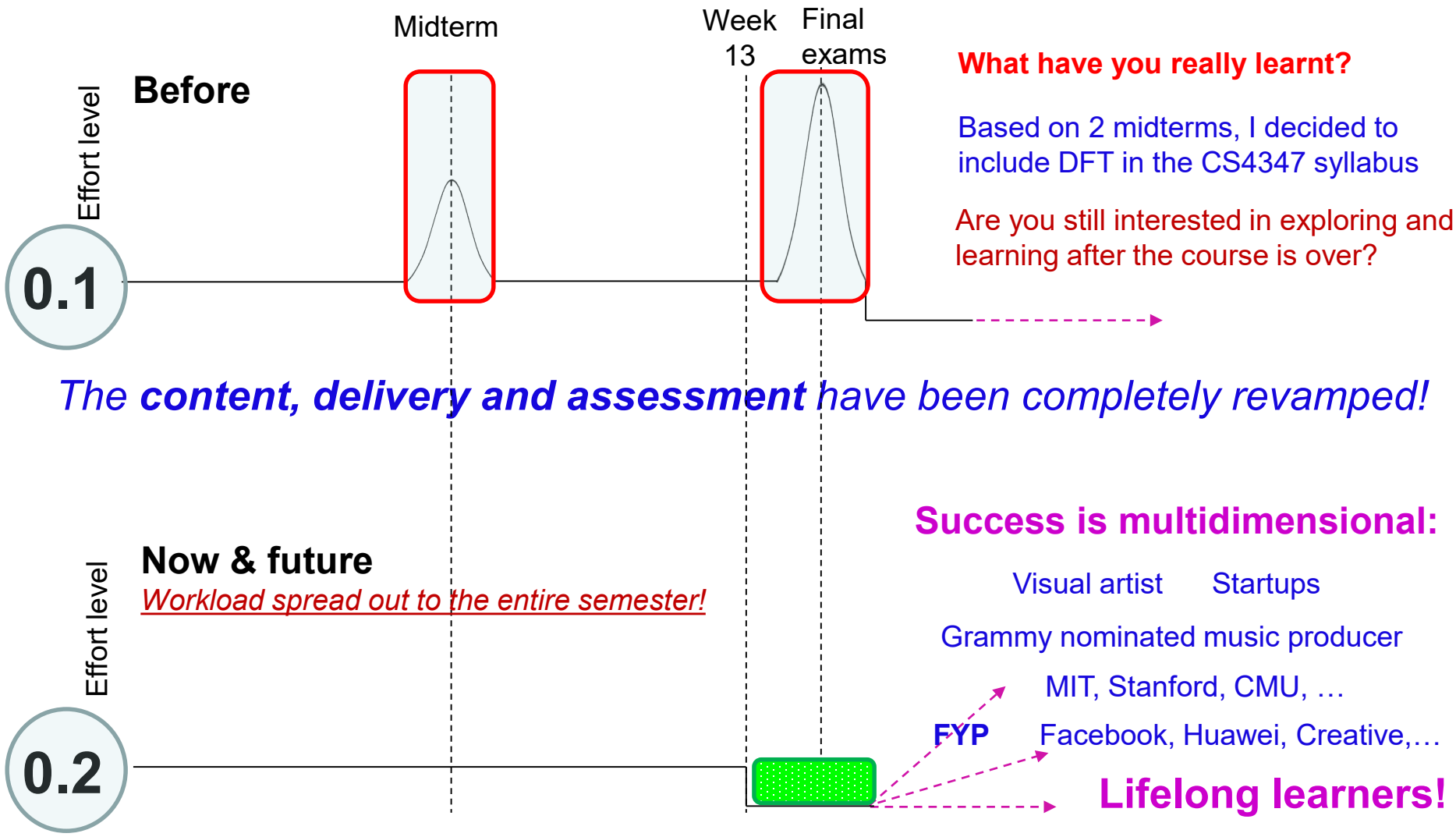
Leverage on neuroscience principles to achieve
"Teach Less, Learn More"

Curiosity, Creativity and Connecting the Dots

CS4347: Computing & Music

Computing + **Music** The only course @NUS

Whole-brain approach of learning!



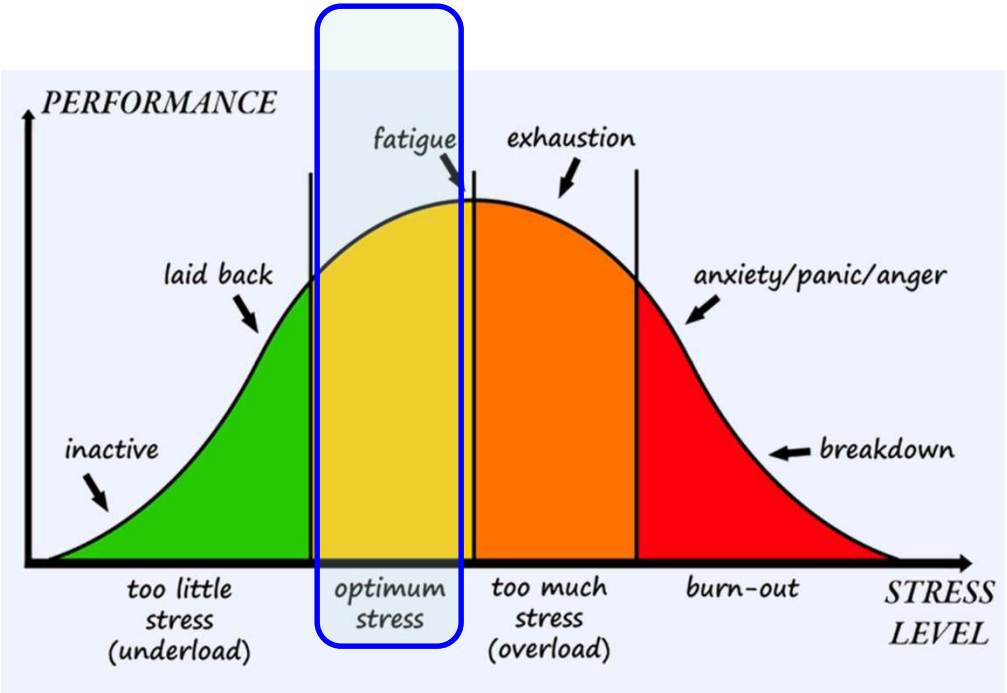
Hope this course will offer you a different learning experience so that you could enjoy the learning journey and would continue exploring *after* graduation.

This may be the only course @ SoC where you can earn up to 5 marks (survey) in Week 1☺

Types of Stress

Rationale for the revamp

- Good stress
  - Within our control
  - Generates energy, drive & excitement
  - Meaningful stress
- Bad stress
  - Chronic Stress
  - Damages physical, mental & emotional health



This slide is borrowed from Dr Andrew Epaphroditus Tay & Jeanie Chu
Health & Wellbeing
Office of the President, NUS

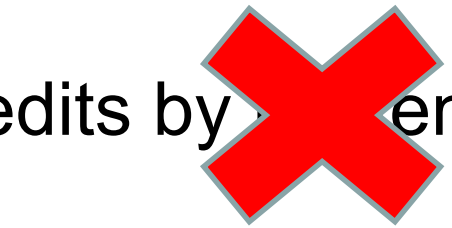
CS4347/CS5647 is about

DSP and ML methods
for speech and music processing
(analysis and synthesis)

It is designed to help you find your own resonance!

CS4347/CS5647 is NOT about

Easy credits by enjoying music in class



Spoon-feeding is NOT education, and is completely out of date!

Zero tolerance for plagiarism

- *Why is it a bad idea to cheat?*
- *Remember to add a reference when using other's ideas*
- *We will help manage/avoid the free rider problem in a group project*
  - *Put down the nature of individual tasks, who does what, who shares what material with the group and when, include the agreed upon tasks in the final submission of the project*
  - *Employ peer review*
- *Don't take this course if you have problems with my educational philosophy and requirements!*

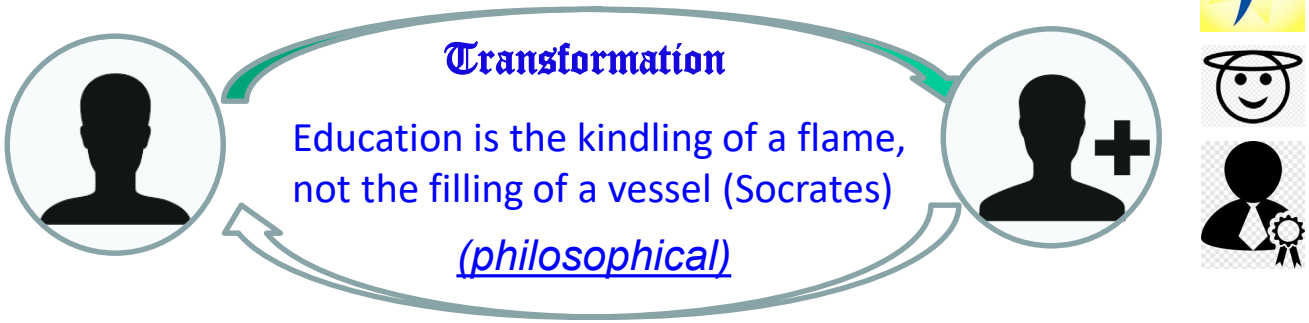
Cognitive Neuroscience-informed Educational Model *(Operational)*

F • T = P+

Multi-**I**ntelligences
Imagination
Initiative
Integrity

Excite
Energize
Engage
Enable
Collaboration
Connect the dots

Person
*__P__roject
*__P__aper
*__P__atent
*__P__roduct
*__P__hD



Ability, curiosity & desire to learn

Inspire students to find their eigenfrequencies Ignite a fire within!

Happy lifelong learner with a strong honor code!

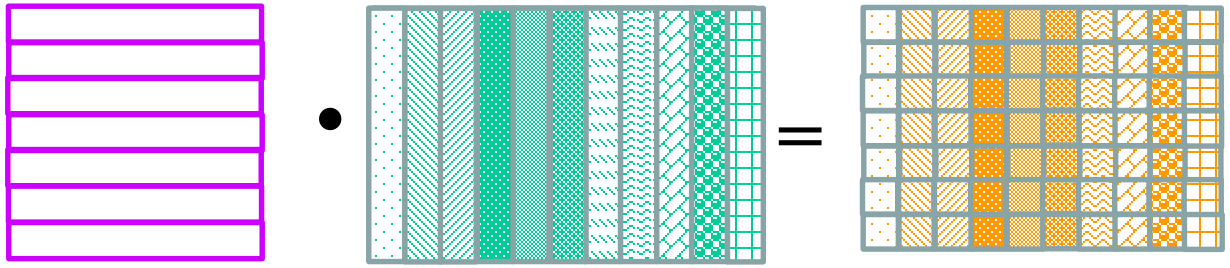
Neuroscience-inspired & DL-based Educational Model

F • T = P+

Multi-**I**ntelligences
Imagination
Initiative
Integrity

Excite
Engage
Energize
Enable
Collaboration
Connect the dots

Person
*__P__roject
*__P__aper
*__P__atent
*__P__roduct
*__P__hD



students courses selections

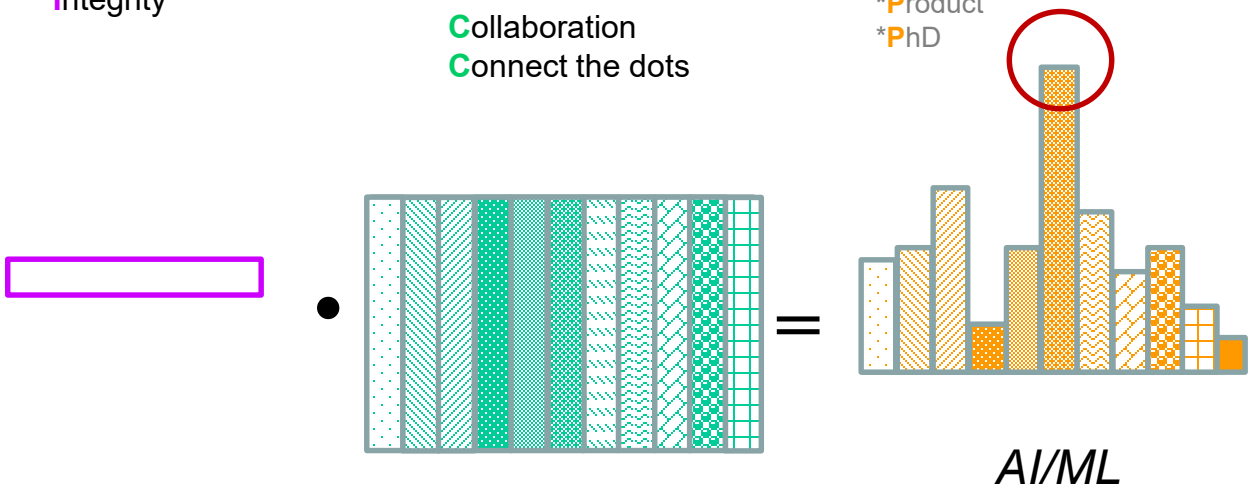
Neuroscience-inspired & DL-based Educational Model

F • T = P+

Multi-**I**ntelligences
Imagination
Initiative
Integrity

Excite
Engage
Energize
Enable
Collaboration
Connect the dots

Person
*__P__roject
*__P__aper
*__P__atent
*__P__roduct
*__P__hD



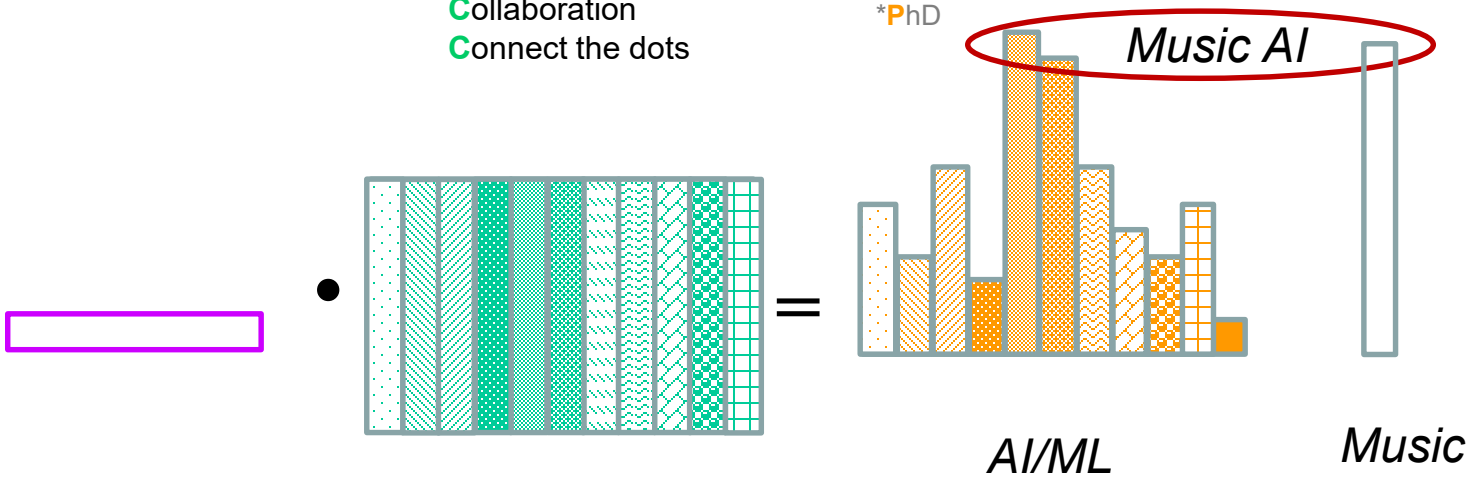
Neuroscience-inspired & DL-based Educational Model

F • T = P+

Multi-**I**ntelligences
Imagination
Initiative
Integrity

Excite
Engage
Energize
Enable
Collaboration
Connect the dots

Person
*__P__roject
*__P__aper
*__P__atent
*__P__roduct
*__P__hD



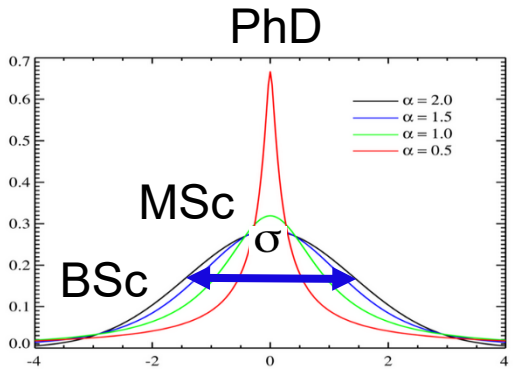
Ma Xichu, Selective advances in music computing for human health and potential, PhD thesis, 2023

Some advices

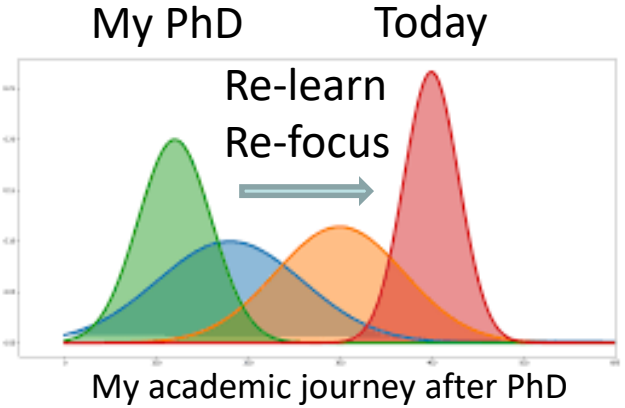
My educational model is to help you identify your eigenfrequency - focus more on smartworking than hardworking! Take good care of your brain which is a delicate organ – don't abuse it!

4 key ingredients for a healthy and happy life from the neuroscience perspective:

Diet	*Sleep*	*Exercise*	*Social interaction*
e.g., background music	*e.g., lullaby for kids*	*e.g., energetic music*	*e.g., party, karaoke*



Breadth or depth



My academic journey after PhD

Broaden your horizon (undergrats + masters)!

Apply Dijkstra's algorithm (PhD students)

NUS gives students more choices for their education

It is an important ability to make the right choices!

My educational model seeks to help you work smarter than harder!

Singaporeans are amongst the most sleep deprived nation worldwide.

Two projects from last year

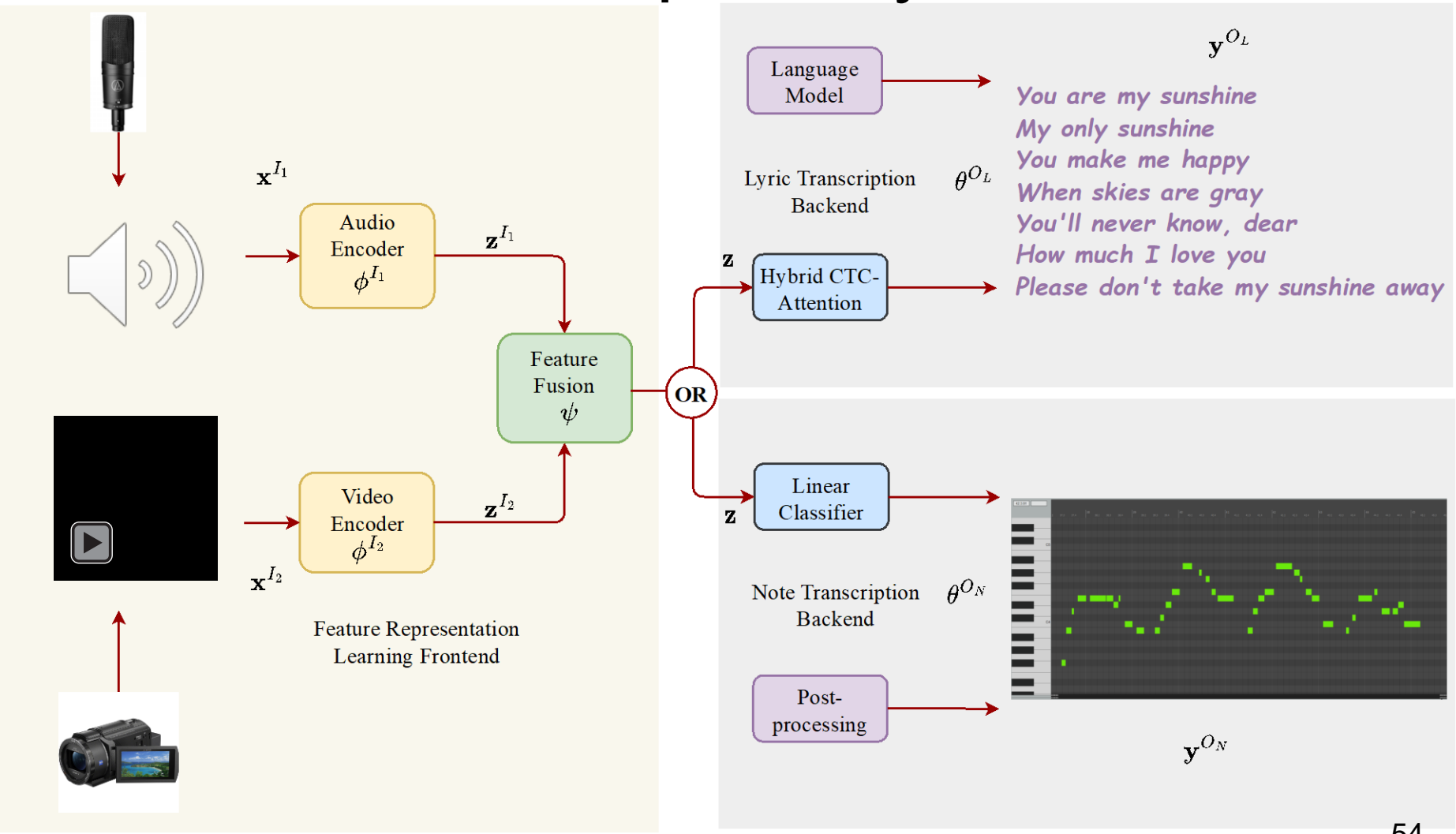
- Singing voice transcription

  A possible default group project
  (real-time singing transcription?)

- Singing voice synthesis

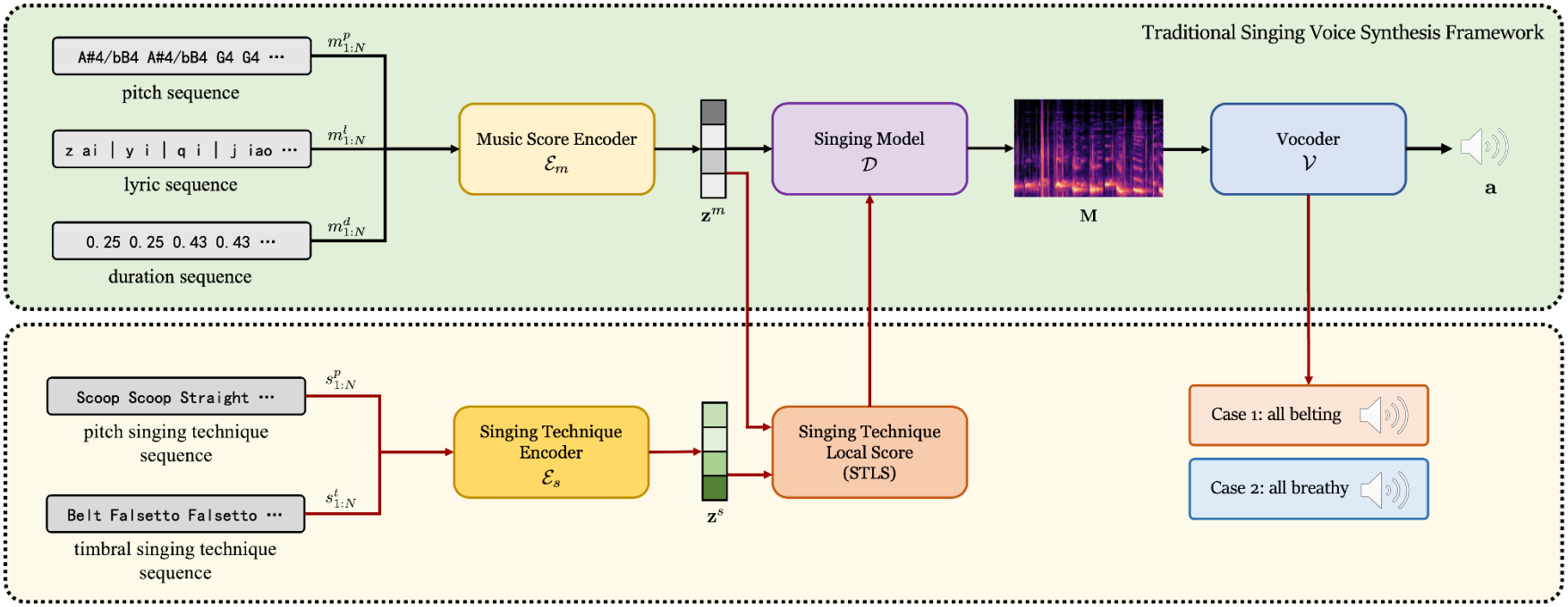
  A possible default group project

Multimodal Lyric and Note Transcription Systems



Real-time singing transcription?

Singing Technique Controllable Singing Voice Synthesis System



On diversity of team composition

Xiangming Gu	Longshen Ou	Danielle Ong	Sng Jia Ming Fadi Faris
CV/EE/ISEP/PhD	AMT/CS/PhD	Linguist/MComp	CS/FYP

Association for Computing Machinery

SIGMM

ACM Multimedia 2022
Lisbon, Portugal | 10-14 October

The ACM Multimedia 2022

Top Paper Award

is presented to

Xiangming Gu, Longshen Ou, Danielle Ong, Ye Wang

for the contribution

MM-ALT: A Multimodal Automatic Lyric Transcription System

Xavier Alameda-Pineda, Qin Jin
Laura Toni and Vincent Oria

ACM Multimedia 2022 Program co-Chairs

A possible SMC4HHP Concert + H e a l t h & W e l l b e i n g Workshop

Q/A

https://smcnus.comp.nus.edu.sg/seminar_concert_2022