Overview

UCS749: Speech Processing and Synthesis

Raghav B. Venkataramaiyer

CSED TIET Patiala India.

September 16, 2024

Outline

1 ABOUT

2 EVALUATION

3 TOPICS
- Background
- Recognition
- Synthesis

4 PRACTICALS

5 RESOURCES

6 INTRODUCTION
- Parallels b/w NLP and Speech
- Speech Input Processing

OUTLINE

1 ABOUT

2 EVALUATION

3 TOPICS

4 PRACTICALS

5 RESOURCES

6 INTRODUCTION

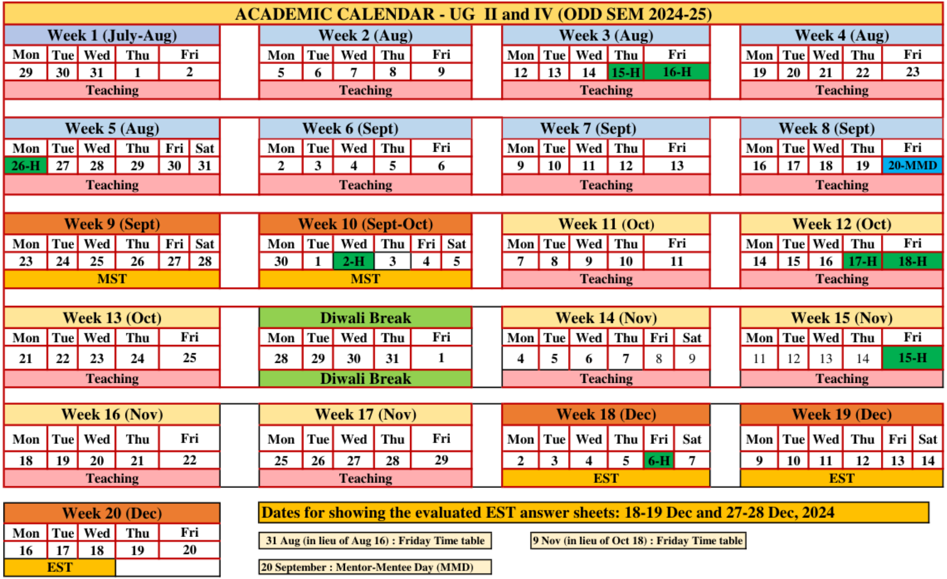
COURSE WEIGHTAGE

L	T	P	Cr
2	0	2	3

SYLLABUS

Link to Syllabus [PDF]

ACADEMIC CALENDAR



ACADEMIC CALENDAR - UG II and IV (ODD SEM 2024-25)

Week 1 (July-Aug)

Mon	Tue	Wed	Thu	Fri
29	30	31	1	2
Teaching				

Week 2 (Aug)

Mon	Tue	Wed	Thu	Fri
5	6	7	8	9
Teaching				

Week 3 (Aug)

Mon	Tue	Wed	Thu	Fri
12	13	14	15-H	16-H
Teaching				

Week 4 (Aug)

Mon	Tue	Wed	Thu	Fri
19	20	21	22	23
Teaching				

Week 5 (Aug)

Mon	Tue	Wed	Thu	Fri	Sat
26-H	27	28	29	30	31
Teaching					

Week 6 (Sept)

Mon	Tue	Wed	Thu	Fri
2	3	4	5	6
Teaching				

Week 7 (Sept)

Mon	Tue	Wed	Thu	Fri
9	10	11	12	13
Teaching				

Week 8 (Sept)

Mon	Tue	Wed	Thu	Fri
16	17	18	19	20-MMD
Teaching				

Week 9 (Sept)

Mon	Tue	Wed	Thu	Fri	Sat
23	24	25	26	27	28
MST					

Week 10 (Sept-Oct)

Mon	Tue	Wed	Thu	Fri	Sat
30	1	2-H	3	4	5
MST					

Week 11 (Oct)

Mon	Tue	Wed	Thu	Fri
7	8	9	10	11
Teaching				

Week 12 (Oct)

Mon	Tue	Wed	Thu	Fri
14	15	16	17-H	18-H
Teaching				

Week 13 (Oct)

Mon	Tue	Wed	Thu	Fri
21	22	23	24	25
Teaching				

Diwali Break

Mon	Tue	Wed	Thu	Fri
28	29	30	31	1
Diwali Break				

Week 14 (Nov)

Mon	Tue	Wed	Thu	Fri	Sat
4	5	6	7	8	9
Teaching					

Week 15 (Nov)

Mon	Tue	Wed	Thu	Fri
11	12	13	14	15-H
Teaching				

Week 16 (Nov)

Mon	Tue	Wed	Thu	Fri
18	19	20	21	22
Teaching				

Week 17 (Nov)

Mon	Tue	Wed	Thu	Fri
25	26	27	28	29
Teaching				

Week 18 (Dec)

Mon	Tue	Wed	Thu	Fri	Sat
2	3	4	5	6-H	7
EST					

Week 19 (Dec)

Mon	Tue	Wed	Thu	Fri	Sat
9	10	11	12	13	14
EST					

Week 20 (Dec)

Mon	Tue	Wed	Thu	Fri
16	17	18	19	20
EST				

Dates for showing the evaluated EST answer sheets: 18-19 Dec and 27-28 Dec, 2024

31 Aug (in lieu of Aug 16) : Friday Time table

9 Nov (in lieu of Oct 18) : Friday Time table

20 September : Mentor-Mentee Day (MMD)

NUM LECTURES

	W	L	P
Prior to MST	8	16	7/8
MST – Diwali	3	6	2/3
Diwali – EST	4	8	4

OUTLINE

1 ABOUT

2 EVALUATION

3 TOPICS

4 PRACTICALS

5 RESOURCES

6 INTRODUCTION

EVALUATION SCHEDULE

	Date	MM
MST	TBA	30
EST	TBA	40
Quiz 1	~~12-Sep 05:30pm~~	5
Quiz 2	21-Nov 05:30pm	5
Lab Eval 1	~~9-Sep 13-Sep~~	10
Lab Eval 2	18-Nov – 22-Nov	10
		100

ABOUT LAB EVAL

All exercise(s) shall be solved in (Colab) python notebook(s), committed to Github using @thapar.edu account. Only a Github Repo link and commit id shall be submit using the Google Form. Any attachments are not allowed. [Read more...]

OUTLINE

1 ABOUT

2 EVALUATION

3 TOPICS

4 PRACTICALS

5 RESOURCES

6 INTRODUCTION

OUTLINE

1 ABOUT

2 EVALUATION

3 TOPICS

- Background
- Recognition
- Synthesis

4 PRACTICALS

5 RESOURCES

6 INTRODUCTION

PRE-REQUISITES

1. Linear Algebra: Vector Spaces/ Linear Maps/ Singularity/ Matrix Decomposition/ Null Space/ Span/ Markov Chains;
2. Probability and Statistics: Central Limit Theorem/ Conditionals & Marginals/ Bayes Theorem/ Markov Assumption/ Stochastic Process
3. Information Theory: Cross Entropy
4. Neural Network: Perceptron Model/ Hidden Layers/ Convolution/ Activation/ Pooling/ Atrous/ Padding/ Backpropagation☒
5. Optimisation: Stochastic Gradient Descent/ Momentum/ Dropout/ RMSProp/ Adam
6. Deep Learning: Sequential Model/ Residual Model/ Adversarial Model/ Attention Model/ Encoder-Decoder Model

INTRODUCTION

1. NLP: Lexeme/ Grapheme
2. Speech: Phoneme
3. Statistical Models: Noise/ Pattern/ Characterisation
4. Language Model: N-Grams/ TFIDF/ Word2Vec/ BERT
5. Speech Models: Wav2Vec/ HuBERT

OUTLINE

1 ABOUT

2 EVALUATION

3 TOPICS

- Background
- Recognition
- Synthesis

4 PRACTICALS

5 RESOURCES

6 INTRODUCTION

AUTOMATED SPEECH RECOGNITION

Hidden Markov Model PDF (Concise), More literature from Google, Duck,Duck,Go; Rabiner's Tutorial.

Time Delay DNN (TDNN)

- Time-delay Networks (TDNN),
- Connectionist Temporal Classification (CTC),
- Jasper,
- QuartzNet,
- Citrinet

Speech Command Recognition MatchboxNet: [PwC] [Colab] (Implementation: here and here uses AvgPool after blocks)

OUTLINE

1 ABOUT

2 EVALUATION

3 TOPICS

- Background
- Recognition
- Synthesis

4 PRACTICALS

5 RESOURCES

6 INTRODUCTION

synthesis (text-to-speech; tts)

Spectrogram Generators Tacotron2, GlowTTS

Audio Generators WaveGlow, SqueezeWave

OUTLINE

1 ABOUT

2 EVALUATION

3 TOPICS

4 PRACTICALS

5 RESOURCES

6 INTRODUCTION

LAB 1: GETTING FAMILIAR WITH SPEECH PROCESSING

1. Getting familiar with the pipeline of Speech Recognition: Speech Recognition with Wav2Vec2 (Pytorch)
2. Perform a simple command classification task with a sequential model:
   - (Tensorflow) Simple Audio Recognition :Recognising keywords; or if you prefer
   - (Pytorch) Speech Command Classification with M5.

LAB 2: HIDDEN MARKOV MODEL

Using MFCCs as features from this example:
MFCC Example [Colab] by Raghav B. Venkataramaiyer;
along with the following dataset:
Free Spoken Digit Dataset (10 digits x 6 speakers x 50 repeats) [Github];
and using hmmlearn as in this tutorial to fit the model
HMM Learn [ReadTheDocs]

1. Compute the probability of occurrence of a given sequence, say $\{3, 2, 5, 4, 0\}$. (Encode the Forward Algorithm)
2. Predict the most likely sequence, given an audio sequence. (Encode the Viterbi algorithm)

LAB 2: HIDDEN MARKOV MODEL (CONTD...)

Theory PDF (Concise), More literature from Google, Duck,Duck,Go; Rabiner's Tutorial.

More Datasets hmm-speech-recognition [Google Code]

More Feature Descriptors CMVN, i-vectors

See Also

- HMM Tutorial [Colab] by BAMB School 2023
- Bean-Machine based Tutorial [Colab]
- HMM Predicting Gold Prices [Medium]
- Single Speaker Word Recognition with HMM [Colab]
- ASR using HMM from scratch [Colab]

LAB 3: ASR IN ENGLISH

ASR with NeMo (Colab)

Additional references:

- `amp_level`”O1”= : the argument used in `PytorchLightning.Trainer` instance;
- But Apex deprecated out of PL v2.0;

For Starters :

NeMo Installation and Getting Started Guide with Citrinet ASR Evaluation

LAB 4: ASR IN INDIC LANGUAGE

Use the method from Lab 3, but use Indic Dataset.

LAB 5: SPEECH COMMANDS

Speech Command Recognition with MatchboxNet

LAB 6: TTS WITH TACOTRON 2

Training with Tacotron 2

Lab 7: TTS in Indic Language

Use the method from Lab 6, but along with Indic Dataset for TTS.

OUTLINE

1 ABOUT

2 EVALUATION

3 TOPICS

4 PRACTICALS

5 RESOURCES

6 INTRODUCTION

SPEECH

1. Directory Listing of SoTA
2. Another Directory Listing of SoTA
3. Jasper (2019)
4. QuartzNet (2019)
5. Citrinet (2021)
6. NVidia NeMo Framework
7. Speech Synthesis Model Zoo (NeMo)
8. Mel Spectrogram

LINEAR ALGEBRA

1. 3B1B
2. Gilbert Strang

PROBABILITY AND STATISTICS

1. Bertsekas & Tsitsiklis: Introduction To Probability; Probabilistic Systems Analysis And Applied Probability
2. 3B1B

NEURAL NETWORK CONCEPTS

1. Andrew Ng on Coursera
2. Andrej Karpathy on Youtube; also on Stanford

1 David McKay

DATASETS

1. Torch Audio (Pytorch)
2. Speech & Speech Recognition Datasets (Tensorflow)
3. ASR Datasets (NeMo)
4. Speech Classification Datasets (NeMo)
5. Lhotse Speech and its use with NeMo
6. Speaker Recognition Datasets (NeMo)
7. Public TTS Datasets (NeMo)
8. Indic ASR Dataset
9. Indic Dataset for TTS

CODE

1. OpenSeq2Seq
2. AI4Bharat
3. NeMo Tutorials

OUTLINE

1 ABOUT

2 EVALUATION

3 TOPICS

4 PRACTICALS

5 RESOURCES

6 INTRODUCTION

OUTLINE

1 ABOUT

2 EVALUATION

3 TOPICS

4 PRACTICALS

5 RESOURCES

6 INTRODUCTION

- Parallels b/w NLP and Speech
- Speech Input Processing

NLP

Natural Language Processing Lexeme/ Grapheme

Language Models Statistical Models N-Grams/ TFIDF

Recently Word2Vec/ BERT etc.

SPEECH

- **Speech Processing** Phoneme
- **Speech Models**
  - **Statistical Models** Noise/ Pattern/ Characterisation; Spectrograms
  - **Recently** Wav2Vec/ HuBERT

OUTLINE

1 ABOUT

2 EVALUATION

3 TOPICS

4 PRACTICALS

5 RESOURCES

6 INTRODUCTION

- Parallels b/w NLP and Speech
- Speech Input Processing

WAVEFORM

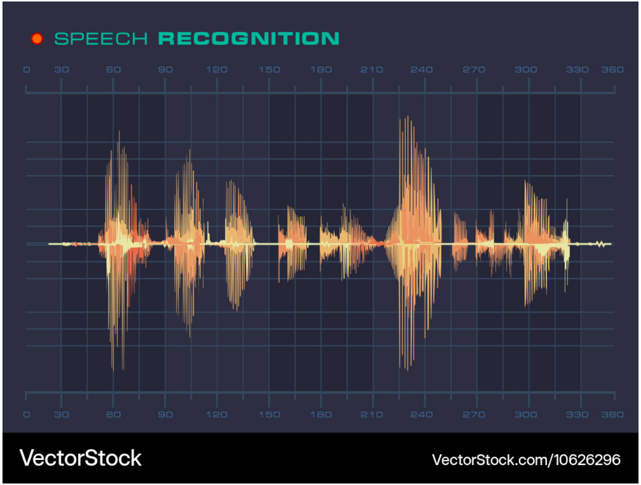


FIGURE: Image Courtesy: [Stock Images on Web]

SPECTRAL ANALYSIS

- Waveform is a time series data.
- Fourier Transform is a function that maps the information in time domain to frequency domain.
- Energy intensity histogram drawn against frequency bands (or spectral bands), is called a spectrum.
- Time domain information may be too dense to make meaning of; hence frequency domain may be favoured.
- Analysis in frequency domain is called spectral analysis.

SPECTROGRAM

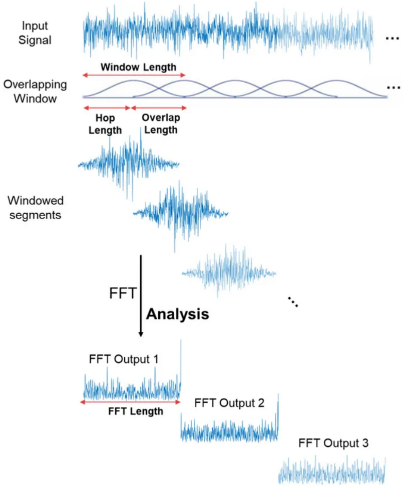


FIGURE: Image Courtesy: MathWorks

Spectrogram is a Short-Time Fourier Transform of the input waveform; or “short-term power spectrum” of sound.

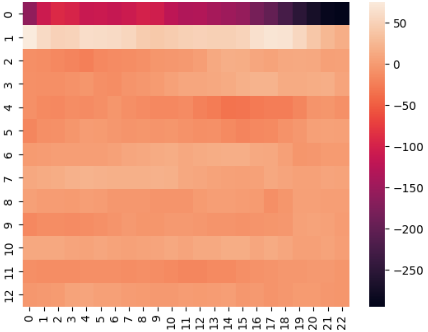


FIGURE: Spectrogram with 12 freq bands and 22 short-time windows. Adapted from Lab 2: MFCC Example [Colab].

MEL SCALE

Mel (named after the word melody) is a non standard perceptual scale of frequency, that is judged by listeners to be equidistant from one-another.

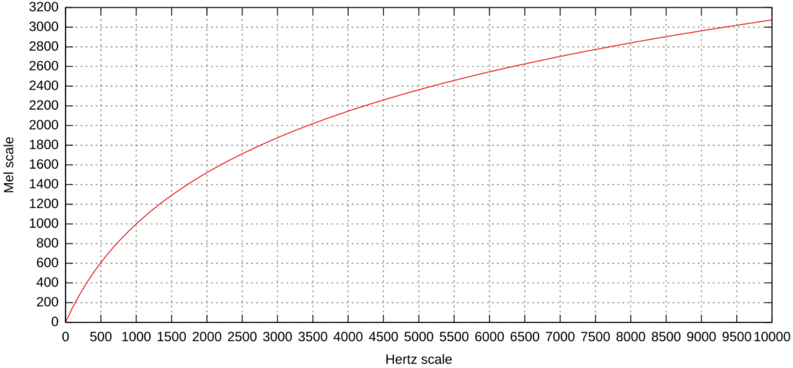


FIGURE: Image Courtesy: [Wikimedia]

MEL SCALE

Mathematically, one of the linear+log fit looks like:

$$m(f) = \begin{cases} \frac{3f}{200}, & f < 1000; \\ 15 + 27\log_{6.4}\left(\frac{f}{1000}\right), & f \geqslant 1000. \end{cases}$$

This was popularised by MATLAB Auditory Toolbox of Slaney

MEL CEPSTRUM

Recall, that Spectrogram is a “short-term power spectrum.”
Mel-frequency cepstrum (MFC) is

- a short-term power spectrum,
- based on linear cosine transform
- of log-power-spectrum
- on a non-linear mel scale of frequency.

Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC.

MFCC

Read More [Medium]