

Deep Learning for Speech Processing

CSCI-B 659 14328

TUESDAY, AUGUST 27, 2019

LECTURE #1

About me

► **Background:**

- 4th year, Assistant Professor, Computer Science
- Director of the Audio, SPeech, and Information REtrieval (ASPIRE) Research Group

► **Main Research Interests:**

- Speech/Signal processing, Machine/deep learning
- Projects on: noise reduction, human-level speech quality assessment, privacy-preserving speech recognition

► **Teaching Experience**

- Machine learning graduate (3x) and undergraduate (1x) levels
- Deep learning for Speech Processing (1x prior)
- Machine Audition and Perception (1x)

► **Industry Experience**

- 3 years as software/systems engineer @ Lockheed Martin in Moorestown, NJ
- Summer graduate intern @ Audience, Inc. (Knowles) in Mountain View, CA

About you?

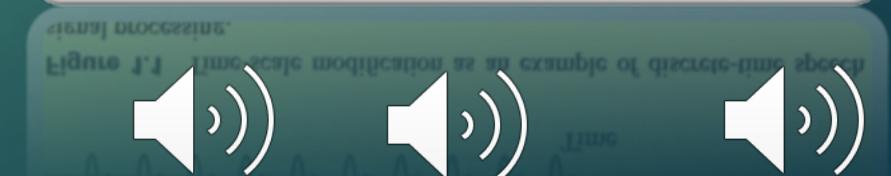
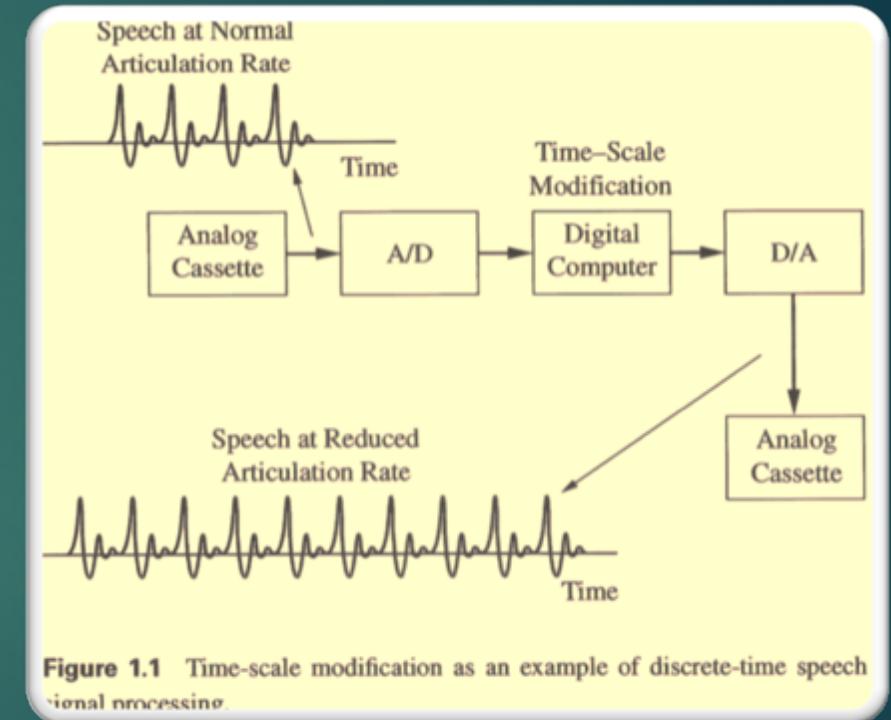
- ▶ Name
- ▶ Degree and program (e.g. M.S. in Data science)
- ▶ Year in program
- ▶ Where are you from?
- ▶ What are your research interests?
- ▶ What do you hope to learn from this course?

Today's agenda

- ▶ Introductions
- ▶ What is speech processing?
- ▶ Why deep learning?
- ▶ Key topics, Course objectives and requirements

What is Speech Processing?

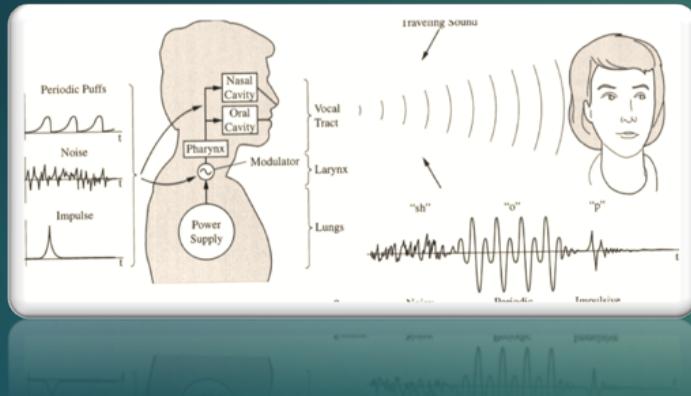
- ▶ How would you define speech processing?
- ▶ Historical Definition
 - ▶ “Manipulation of sampled speech signals ... to obtain a new signal with some desired properties” by T. F. Quatieri
 - ▶ Example: Change a person’s speech rate (e.g. talk faster or slower)
- ▶ This definition does NOT cover all aspects of speech processing



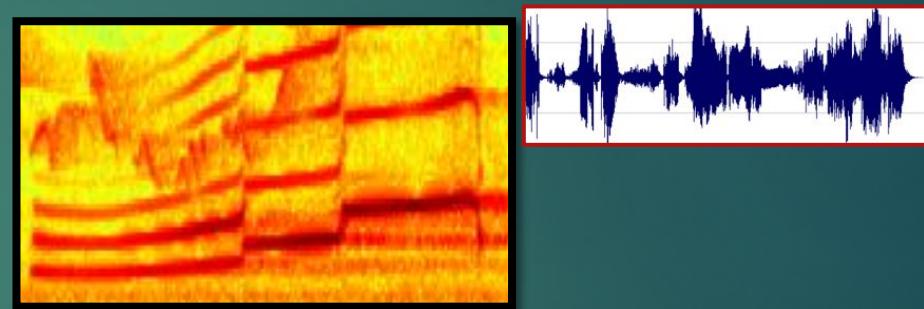
Stages of Speech Processing

► Pre-process stage

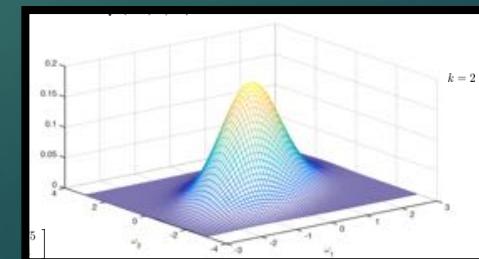
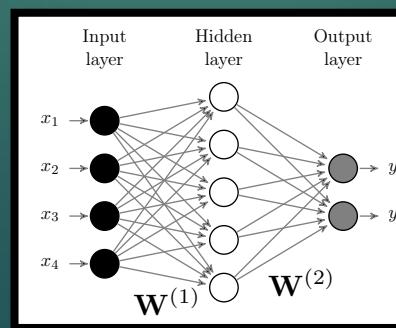
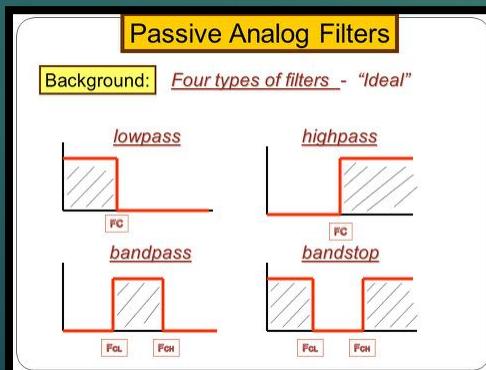
- Speech production and perception
- Sound recording and characteristics



► Feature extraction and Signal representation



► Processing and Data Manipulation



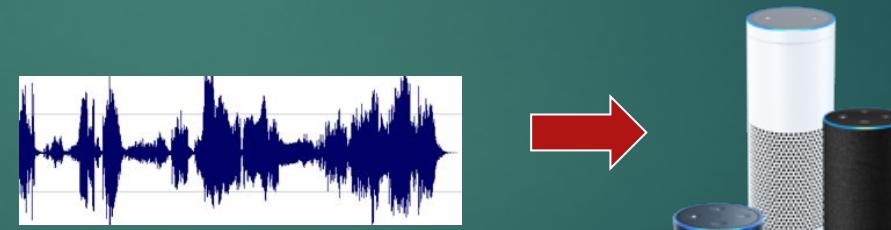
Stages of Speech Processing

▶ Example Applications

Speech
enhancement



Automatic
Speech
Recognition



▶ Performance Assessment:

- ▶ How well does the approach work?
- ▶ In which environments? For which speakers?

What is needed to make Siri work?

8

- ▶ **Demo: Siri**
 - ▶ Me: “Hey Siri... What is today's date?”
 - ▶ Siri: “It is Tuesday, August...”
- ▶ **What does Siri need to do/understand, to answer this question?**
 - ▶ How to capture the signal?
 - ▶ How to process and represent the signal?
 - ▶ Recognize what I'm saying?
 - ▶ Recognize who is speaking? (Demo)
 - ▶ How to reply? What I expect to hear?



The Prevalence of Speech Processing! 9

Traditional



Health



Consumer



Why is this important?

- ▶ Speech is the primary form of communication between humans

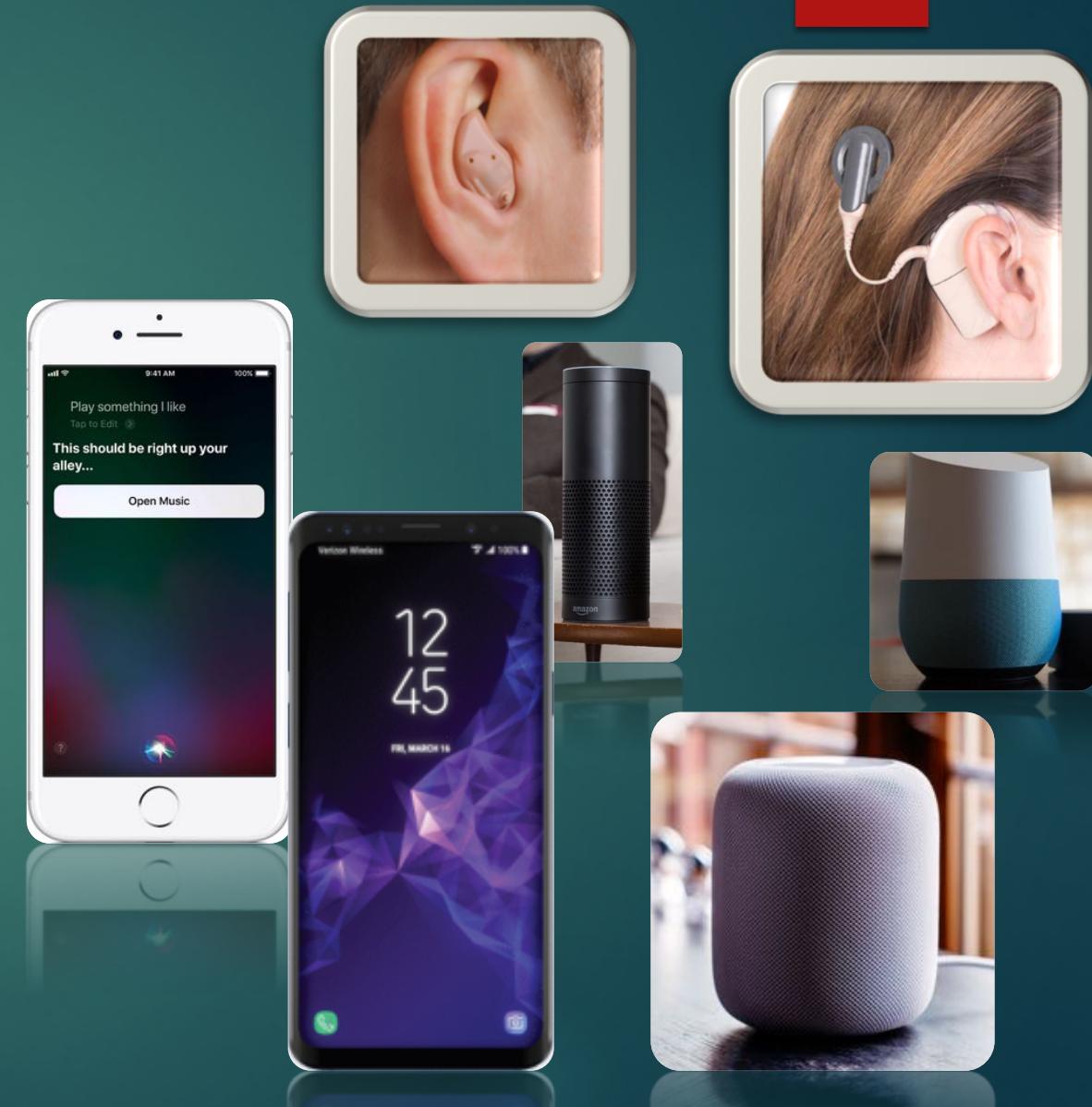


- ▶ Now, speech is often (or will be) used to interface with machines

Why is this important?

- ▶ **A lot of information is contained in speech**
 - ▶ Speaker? age? gender? nationality?
 - ▶ Language? Number of speakers?
 - ▶ Location? left? right? close? Inside? Outside?

- ▶ **More R&D is needed**
 - ▶ Current devices really aren't that smart!
 - ▶ There are many things that they cannot do or do not do well



Commercial Interest

12



Today's agenda

- ▶ Introductions
- ▶ What is speech processing?
- ▶ Why deep learning?
- ▶ Key topics, Course objectives and requirements

Why deep learning? It Works!

14

► Advancements in speech applications

New Advancements in Spoken Language Processing

May 6, 2019 | By Xuedong Huang, Technical Fellow, Speech and Language



Deep learning algorithms, supported by the availability of powerful Azure computing infrastructure training data, constitutes the most significant driving force in our AI evolution journey. In the Microsoft reached several historical AI milestones being the first to achieve human parity in the benchmark tasks that have been broadly used in the speech and language community:

- 2017: [Speech Recognition](#) on the conversational speech transcription task (Switchboard)
- 2018: [Machine Translation](#) on the Chinese to English news translation task (WMT17)
- 2019: [Conversational QA](#) on the Stanford conversational question and answering task (CoQA)

These breakthroughs have a profound impact on numerous spoken language applications from smart loudspeakers. While smart speakers are commercially available today, most of them

Home > Technology > Microsoft demonstrates "breakthrough" speech translation technology | Nov. 09, 2012 at 2:00 am

TECHNOLOGY

Microsoft demonstrates "breakthrough" speech translation technology

The technology demonstrated by Rashid uses a technique called Deep Neural Networks, developed by Microsoft Research and the University of Toronto two years ago. Deep Neural Networks is a technique "patterned after human brain behaviour" and provides the ability to "train more discriminative and better speech recognisers" than older methods.

"We have been able to reduce the word error rate for speech by over 30% compared to previous methods. This means that rather than having one word in 4 or 5 incorrect, now the error rate is one word in 7 or 8. While still far from perfect, this is the most dramatic change in accuracy since the introduction of hidden Markov modeling in 1979, and as we add more data to the training we believe that we will get even better results."

Nuance Advances Text-to-Speech Technology through Deep Learning

Nuance - Friday, February 23rd, 2018 [Print](#) | [Email](#)



In a recent press release from Nuance Communications, Inc. they announced that it has advanced its text-to-speech (TTS) technology with deep neural networks (DNN) to deliver a new standard of quality, reducing errors by 40 percent compared to previous speech synthesis techniques.

Combining advancements in deep learning with knowledge-based developments, Nuance's Vocalizer suite of TTS solutions – including Vocalizer Embedded for embedded platforms, Vocalizer Server for cloud applications and the Vocalizer Studio development tool – enables speech output that is nearly indistinguishable from human speech, enriching user experiences across automotive, enterprise, healthcare, IoT and smart home offerings and resulting in a more intuitive and conversational interaction between people and machines. The application of artificial intelligence (AI) techniques gives Vocalizer the ability to quickly learn new words, phrases and pronunciations and communicate with more expressivity and personality across more than 50 languages.

Why deep learning? It's Available

15

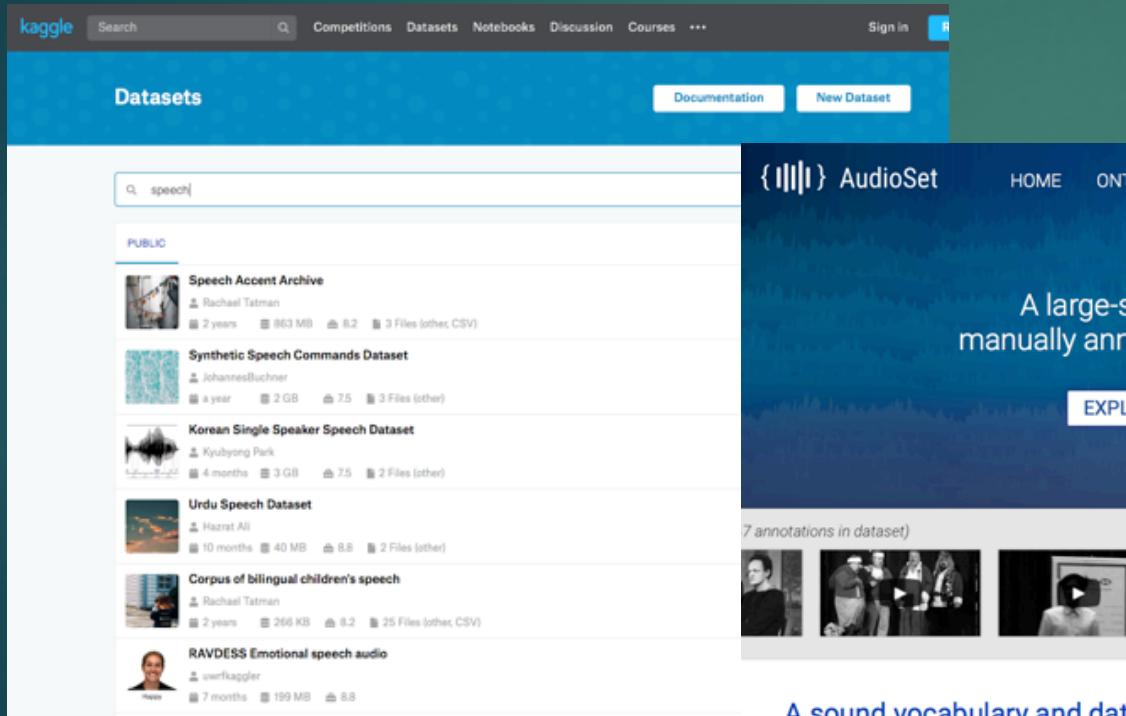
- ▶ There are many deep (machine) learning libraries



Why deep learning? It's Available

16

- ▶ There is an abundance of data!



A screenshot of the Kaggle Datasets interface. At the top, there is a search bar with the query "speech". Below the search bar, a list of datasets is displayed under the heading "PUBLIC". The datasets listed are:

- Speech Accent Archive (Rachael Tatman, 2 years, 863 MB, 8.2, 3 Files (other, CSV))
- Synthetic Speech Commands Dataset (Johannes Buchner, 1 year, 2 GB, 7.5, 3 Files (other))
- Korean Single Speaker Speech Dataset (Kyubyong Park, 4 months, 3 GB, 7.5, 2 Files (other))
- Urdu Speech Dataset (Hazrat Ali, 10 months, 40 MB, 8.8, 2 Files (other))
- Corpus of bilingual children's speech (Rachael Tatman, 2 years, 266 KB, 8.2, 25 Files (other, CSV))
- RAVDESS Emotional speech audio (uerkaggle, 7 months, 199 MB, 8.8)

On the right side of the screen, there is a separate window for the "AudioSet" dataset. The AudioSet website has a dark blue background with white text. It features a logo with three vertical bars and the text "AudioSet". Below the logo, it says "A large-scale dataset of manually annotated audio events" and "EXPLORE THE DATA". At the bottom, there are two examples of annotations: "7 annotations in dataset" with three thumbnail images, and "Explosion (2,274 annotations in dataset)" with five thumbnail images.

A sound vocabulary and dataset

AudioSet consists of an expanding ontology of 632 audio event classes and a collection of 2,084,320 human-labeled 10-second sound clips drawn from YouTube videos. The ontology is specified as a hierarchical graph of event categories, covering a wide range of human and animal sounds, musical instruments and genres, and common everyday environmental sounds.

By releasing AudioSet, we hope to provide a common, realistic-scale evaluation task for audio event detection, as well as a starting point for a comprehensive vocabulary of sound events.



The screenshot shows the homepage of the "The 5th CHiME Speech Separation and Recognition Challenge". The header includes the CHiME logo and the title "The 5th CHiME Speech Separation and Recognition Challenge". Below the header, there is a navigation menu with links to Home, Overview, Data, Software, Instructions, Download, Submission, Results, Contact, Forum, and FAQ.

The 5th CHiME Speech Separation and Recognition Challenge

News - 10th September
Results of the CHiME-5 challenge are now available.

Following the success of the 1st, 2nd, 3rd, and 4th CHiME challenges we are pleased to announce the 5th CHiME Speech Separation and Recognition Challenge (CHiME-5). The new challenge will consider the problem of distant multi-microphone **conversational speech recognition** in everyday home environments. Speech material was elicited using a **dinner party scenario** with efforts taken to capture data that is representative of natural conversational speech.

All approaches are welcome, whether they are **emerging or established**, and whether they rely on **speech processing, signal separation or machine learning**.

THE CHiME-5 WORKSHOP
The results of CHiME-5 will be presented at a dedicated workshop that will take place on **September 7th** in Hyderabad **following** Interspeech 2018. Further details are available on the [workshop website](#).

The workshop will be held at the Microsoft Campus close to the Interspeech conference centre.

MATERIALS
CHiME-5 features two tracks depending on the number of microphone arrays available for testing:

- single-array track
- multiple-array track

Participants are encouraged to submit results for all tracks. Submitting results for one track is also allowed.

Why deep learning? It's Useful

17

- ▶ Many problems are being successfully addressed with deep learning

The image displays a grid of news cards from various sources, each illustrating a different application of deep learning:

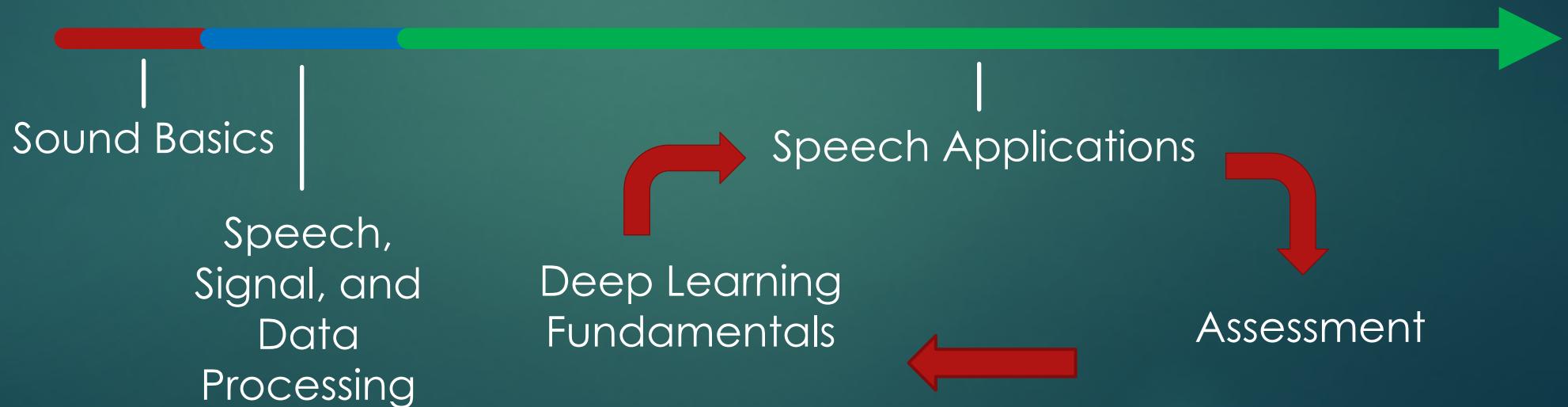
- New Atlas**: Learning from photos, a deep neural network identifies deepfakes. They're known as deepfakes – photos or videos that have been very convincingly manipulated to depict people saying or doing things that they ...
1 month ago
- Nature.com**: Deep learning detects impending organ injury in the clinic. Organ damage is often detected late, when treatment options are limited. The use of artificial intelligence to continuously monitor a patient ...
3 weeks ago
- MarTech Series**: Innovative Deep Learning Solution for Face Detection Developed by ...
Results for Sightcorp's own solution, the newly minted "Deep Learning face detector," on the other hand, returned a 10x10 grid of nearly a full ...
1 day ago
- Medical Xpress**: Deep learning model detects diabetic eye diseases accurately
According to the research findings published in Scientific Reports, the deep learning model detects the severity grade of diabetic retinopathy ...
1 day ago
- Futurity: Research News**: Deep learning speeds chemical test for food, blood, and more
Because this type of analysis is used in a variety of ways, the new method will have a major impact on quality, efficiency, and cost when ...
1 day ago
- Xtelligent Healthcare Media**: Deep Learning Tool Detects Cancer in Radiology Reports
"Our hypothesis was that deep learning algorithms could use routinely generated radiology text reports to identify the presence of cancer and ...
1 day ago
- Healio**: Deep learning: Getting better information to clinicians faster, cheaper ...
By developing a proper pipeline and applying deep learning neural networks, clinicians can save time and money for patients while delivering ...
1 day ago

Today's agenda

- ▶ Introductions
- ▶ What is speech processing?
- ▶ Why deep learning?
- ▶ Key topics, Course objectives and requirements

A Wholistic Viewpoint

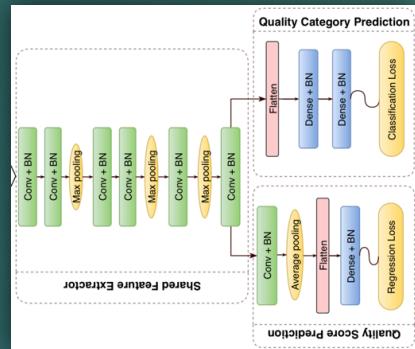
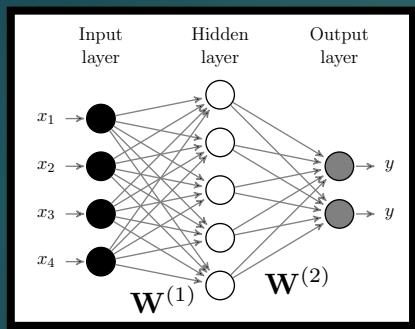
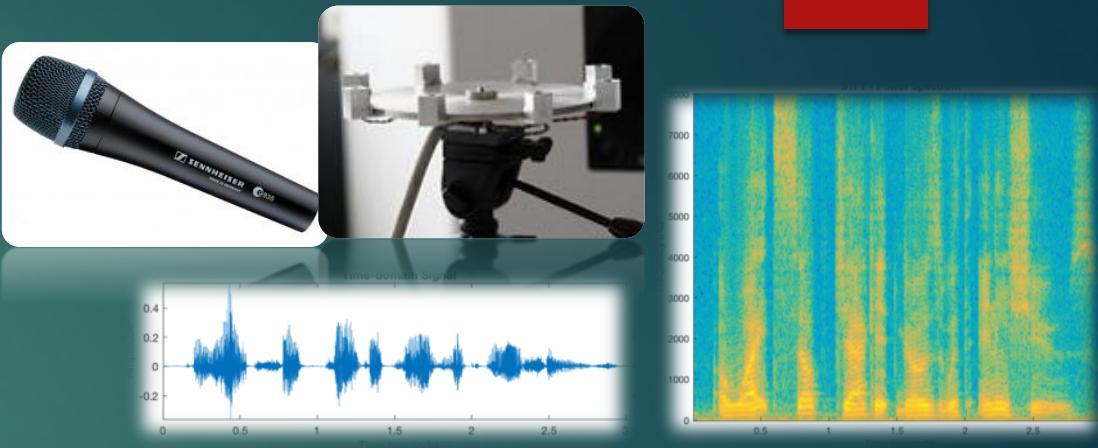
- ▶ This class will introduce you to all stages of deep learning and speech processing research
- ▶ Programming assignments (in Python) will reinforce concepts at every stage
- ▶ Classes will toggle between lectures and practical experience



Key Topics

► Sound and Speech Processing

- What is sound? How do we capture speech?
How is it represented?
- Signal power, energy, intensity; Sound level



► Deep Learning

- Types of networks? DNN? RNN? CNN? LSTM?
- How are networks optimized?
- How much and what types of data?
- Deep learning in Python

► Deep Learning for Speech Processing

- How is deep learning used for different applications?
- How do they perform? Assessment?
- What work still needs to be done?

Course Description, Objectives,...

21

- ▶ See syllabus

My Expectations of You

- ▶ You are hard working
 - ▶ You will read course material and review class notes
 - ▶ You will spend at least 9 hours per week outside of class on this course
 - ▶ You will complete all post-class assignments
- ▶ You will participate in class (Ask questions)
- ▶ You are motivated to learn and succeed in class

Your Expectations of Me

- ▶ I will be transparent in grading
- ▶ I will be prepared for each class
- ▶ I am here to help you learn and answer your questions
- ▶ I will treat you with respect
- ▶ I am available, outside of class, to answer questions

Closing Items

- ▶ Course requirements
 - ▶ Come to class!
 - ▶ Participate!
 - ▶ Read material before class!
 - ▶ Ask questions!
- ▶ Next class:
 - ▶ Topic: Basics of sound and audio
 - ▶ Complete Homework #0 before next class
 - ▶ **Bring headphones**