Speech Recognition and Audio Processing

Guide to AI and Data Courses · Part III: Signals, Images and Audio

Digitizing sound | Spectrogram and MFCC | Acoustic model | Language model | Speakers and events | Related courses

Sercan KÜLCÜ · Giresun University · Computer Engineering

© Sercan Külcü. All rights reserved. · Interactive HTML version of these slides:
https://sercankulcu.github.io/files/ai_courses/slides/Chapter_19_Speech_and_Audio_Processing.html

INTRODUCTION

What Question Does It Ask?

> **How are words, speakers and events extracted from a sound wave?**

Speech recognition and audio processing study how to analyze the sound wave from a microphone and extract words, the speaker's identity or events in the environment. Sound is first converted to time-frequency features; the **acoustic model** and the language model together find the most likely text.

At a glance

APPROACH	INPUT	OUTPUT
acoustic signal and language model	audio waveform	text, speaker, sound event

SUCCESS MEASURE	MATH LOAD	DATA NEEDS
word error rate (WER)	medium-high	many hours of recordings

A Short History

1952
Bell Labs' Audrey system recognizes digits spoken by a single speaker.

1962
IBM Shoebox is demonstrated, recognizing sixteen English words and the digits.

1970s
In DARPA-funded projects, Carnegie Mellon's Harpy system recognizes about a thousand words.

1980s
Hidden Markov models (HMMs) become the standard method of speech recognition.

2012
Deep neural networks begin replacing Gaussian mixtures in acoustic modeling.

2022
OpenAI's Whisper is trained end to end on about 680,000 hours of multilingual data.

THE COURSE

Core Concepts

Sampling
Measuring the sound wave thousands of times per second.

Spectrogram
A time-frequency image from the short-time Fourier transform.

MFCC
Features adapted to how the human ear perceives frequency.

Acoustic model
Predicts the sound unit (phoneme) from a stretch of audio.

Language model
Says which word sequence is more likely.

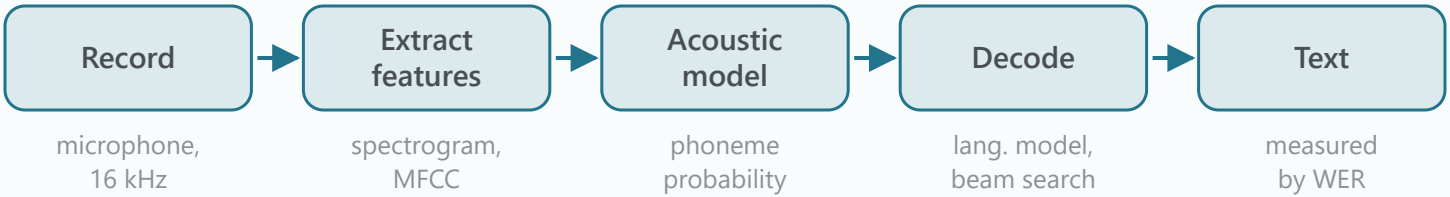
Decoding
Finding the most likely word sequence with beam search.

Speaker recognition
Verifying or identifying whose voice it is.

Word error rate
WER: the rate of insertion, deletion and substitution errors.

THE COURSE

A Typical Workflow



Example: A voice assistant records the question "what's the weather tomorrow" at 16 kHz, computes a spectrogram from 25 ms frames, the acoustic model gives phoneme probabilities for each frame, and the language model turns them into the most likely sentence. The text is passed to NLP for meaning.

THE COURSE

What Does the Course Cover?

1. Speech production and perception; phonemes
2. Digital audio: sampling, framing, windowing
3. Spectrograms, the mel scale and MFCCs
4. Hidden Markov models and Gaussian mixtures
5. Decoding with Viterbi and beam search
6. N-gram and neural language models
7. Deep acoustic models
8. End-to-end recognition: CTC, attention, transducers
9. Evaluation: word error rate
10. Speaker recognition and verification
11. Noise robustness and speech enhancement
12. Speech synthesis (TTS)

Skills you will gain

extracting features with librosa | transcribing and fine-tuning with Whisper | training models with torchaudio | analysing speech in Praat | comparing systems by WER

A typical semester; the order and coverage vary between universities.

TRY IT

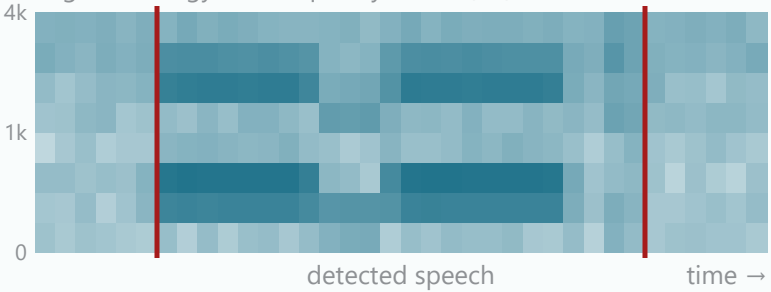
A Smart Home Recognizes Commands

A smart home assistant must recognize four voice commands: yes, no, stop, start. Pick a command, change the speaker, raise the noise; the system compares the sound with stored templates.

Waveform (0.6 s)



Spectrogram: energy in 8 frequency bands (Hz)



yes | no | stop | start

Another speaker

Background noise: **2**

Spoken: “yes” (SNR 13 dB)
Recognized: **yes** · confidence 98% ✓

DTW distance to templates (small = similar)

yes	0.96
no	1.75
stop	2.07
start	2.23

Sound is first turned into a **spectrogram**, a time–frequency picture. The same word can be said fast or slow; DTW matches them by stretching the time axis. As noise grows, confidence drops.

CONTEXT

Prerequisites

Courses worth knowing before this one, and why they are needed.

Course	Why it is needed
Digital Signal Processing	spectrogram, MFCC, filters
Calculus	complex numbers, integrals and an introduction to Fourier series
Programming	array operations and reading audio from files in Python

CONTEXT

Are You Ready for This Course?

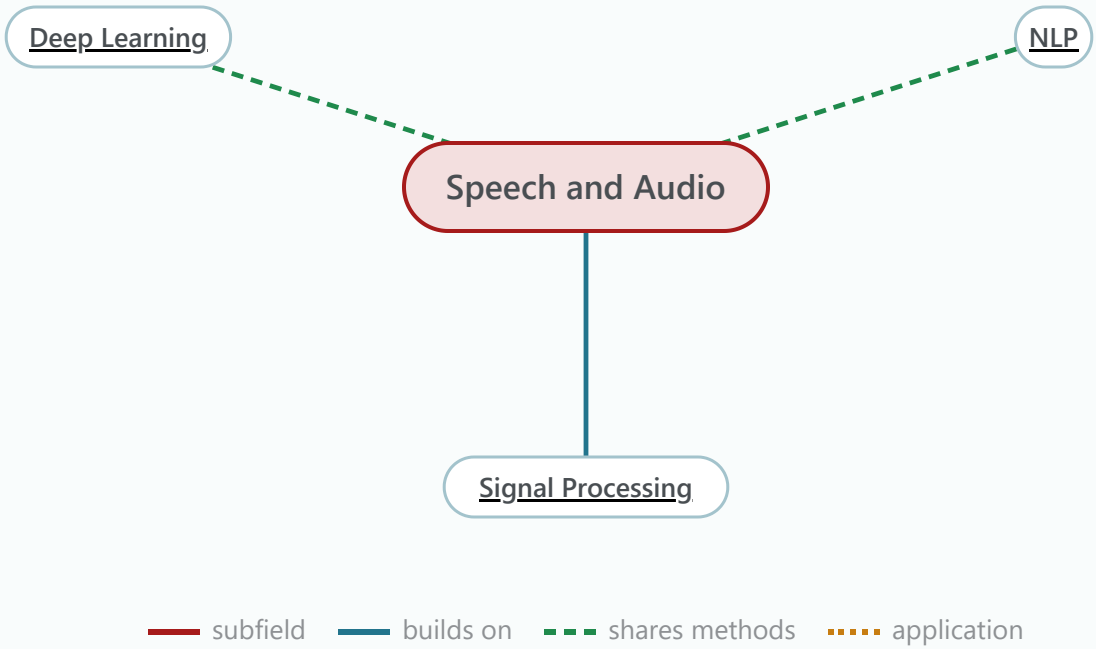
Ask yourself; for every "no", the right-hand column shows where to start.

Ask yourself	If not, start with
Can you explain how the Fourier transform splits a signal into frequencies?	Digital Signal Processing
Can you use conditional probability and Bayes' rule?	Probability and Statistics
Can you explain how a simple neural network is trained?	Artificial Neural Networks
Can you read an audio file in Python and plot its waveform?	An intro programming course, then a short NumPy and librosa tutorial

If you can say "yes" to most of these, you are ready; one or two gaps can be closed during the semester.

CONTEXT

Neighbouring Courses



OVERLAP

Overlapping Courses

Relation	Course	Shared ground
builds on	This course builds on Digital Signal Processing	spectrogram, MFCC, filters
shares methods	Shares methods with Deep Learning	end-to-end speech recognition
shares methods	Shares methods with Natural Language Processing and Text Mining	language model, speech to text

Where It Differs (1/2)

Criterion	Speech and Audio	Signal Processing	NLP
Core question	How are words, speakers and events extracted from a sound wave?	How are sampled signals analyzed, filtered and transformed?	How does a computer analyze, understand and generate human language?
Approach	acoustic signal and language model	mathematical transforms and filters	linguistics, statistics and deep learning
Input	audio waveform	sampled signal	text
Output	text, speaker, sound event	filtered signal, spectrum	label, translation, summary, answer
Success measure	word error rate (WER)	signal-to-noise ratio, frequency response	accuracy, BLEU, human evaluation
Math load	medium-high	high: complex numbers, Fourier	medium
Data needs	many hours of recordings	no learning; the signal suffices	large text corpora

- **Signal Processing:** Signal processing filters and transforms any kind of signal; this course focuses on features and models specific to audio and speech.
- **NLP:** NLP assumes it already has text; this course obtains that text from the sound wave and also handles audio events.

Where It Differs (2/2)

Criterion	Speech and Audio	Deep Learning	Biometrics
Core question	How are words, speakers and events extracted from a sound wave?	How do many-layered networks learn useful representations from raw data?	How is a person recognized from their face, fingerprint or voice?
Approach	acoustic signal and language model	multilayer representation learning	matching and verification
Input	audio waveform	raw data: images, text, audio	face, fingerprint, iris, voice
Output	text, speaker, sound event	learned representation and prediction	identity decision
Success measure	word error rate (WER)	accuracy, loss, scaling	FAR, FRR, EER
Math load	medium-high	high: derivatives, matrices, probability	medium
Data needs	many hours of recordings	large; needs GPUs	few samples per person

- **Deep Learning:** Deep learning offers general network architectures; this course adapts them to the time-frequency structure of sound and the alignment problem.
- **Biometrics:** Biometrics uses the voice as proof of identity, like a face or fingerprint; this course also analyzes what is said.

PRACTICE

Applications

Voice assistants
Understanding commands to control a phone or home.

Captioning
Turning video, meeting and lecture recordings into text.

Call centers
Analyzing and routing calls.

Speaker verification
Voice-based banking and access control.

Sound event detection
Detecting breaking glass, sirens, coughs.

Speech synthesis
Turning text into natural-sounding speech.

PRACTICE

Risks and Ethics

Voice cloning

A voice imitated from a few seconds of recording is used in fraud and to defeat voice authentication.

Always-on listening

Devices waiting for a wake word can record by mistake, and a voice also carries health and emotion information.

Accent and dialect gaps

Error rates are clearly higher for some accents and dialects, so these speakers gain less from captions and assistants.

Hallucinated transcripts

End-to-end models can write sentences nobody said during silence or noise, which is dangerous in medical transcription.

In depth: Explainable AI and AI Ethics.

PRACTICE

Tools and Datasets

Tools and libraries

librosa | Kaldi | Praat | torchaudio | ESPnet | SpeechBrain | Whisper

Well-known datasets

TIMIT | LibriSpeech | Mozilla Common Voice | Switchboard | VoxCeleb | AudioSet

PRACTICE

Common Misconceptions

Common belief	In fact
Speech recognition maps sound directly to words.	The same word sounds very different across accents, speeds and noise; the language model resolves the ambiguity.
The louder the recording, the better.	Clipping and reverberation hurt recognition; a clean, balanced recording is what matters.
A model that works well in English works in Turkish too.	Each language needs enough data; agglutinative Turkish also has a vocabulary problem to solve.
Audio processing is only about speech.	Music, environmental sounds and medical sounds (heart, cough) are also part of the course.

NEXT STEPS

After This Course

Courses that build on or extend this one

This course is mostly an end point: it builds on the others.

Related careers

Speech recognition engineer | Audio processing engineer | Researcher | Embedded audio developer

After this course you can move on to natural language processing for the meaning of the text, to embedded AI for on-device voice assistants, or to biometric systems for voice authentication. Digital signal processing deepens the mathematics behind the subject.

NEXT STEPS

Further Reading

- Rabiner and Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.
- Huang, Acero and Hon, *Spoken Language Processing*, Prentice Hall, 2001.
- Jurafsky and Martin, *Speech and Language Processing*, 2nd ed., Prentice Hall, 2009.
- Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE* 77(2), 1989.

The End

Chapter 19: Speech Recognition and Audio Processing

← Previous chapter · Map of the field · Next chapter →