CS 224S / Linguist 285
Spoken Language Processing

Andrew Maas | Stanford University | Spring 2025

Lecture 1: Course Introduction

Outline for today

- Course Introduction: Why take this course?
- Course Logistics: How to take this course.
- Course Topics Overview: Here we go!
  - Task-oriented spoken dialogue systems / Conversational Agents
  - Speech Recognition (Speech to Text)
  - Speech Synthesis (Text to Speech)
  - Developing spoken language applications & conversational AI products

Hello class!

Instructor: Andrew Maas

First-year PhD student, 2009

Over 15 years of experience developing and deploying deep learning systems in the real world. Andrew completed a PhD at Stanford advised by Andrew Ng and Dan Jurafsky. His research established key foundations for deep learning applied to natural language, and end-to-end speech recognition (>20,000 academic citations). Andrew co-founded Roam Analytics, an AI platform for extracting information from healthcare text and audio, and developed speech systems as part of Semantic Machines and Wit.ai.

Industry+startup experience

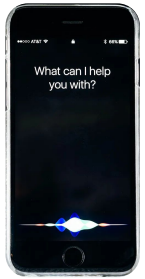
- Currently: Co-founder/CEO @ Pointable. Developing reliable, controllable LLM-based agents for enterprise (mostly stealth mode, not visible online these days)
- 2019 - 2023: Led a team building data-centric ML approaches @ Apple SPG
- 2013 - 2019: Co-founder, healthcare text extraction platform @ Roam Analytics
- 2015 - 2016: Research scientist, speech deep learning framework @ Semantic Machines
- 2014 - Present: Lecturer, Spoken Language Processing @ Stanford
  Researcher / advisor for several startups (Wit.ai, Coursera, UpLimit, StartX)

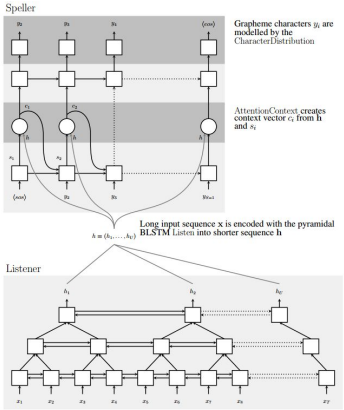
Educational background

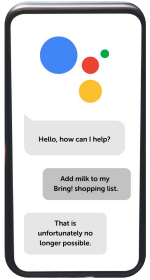
- 2009 - 2015: PhD, Computer Science - Stanford
  - Advised by Andrew Ng and Dan Jurafsky.
  - Topic: Deep learning for spoken + written language
- 2005 - 2009: BSc, Computer Science & Cognitive Science (double major) - Carnegie Mellon
  - Research advisors: Drew Bagnell, Charles Kemp
  - Focus on ML, robotics, and computational neuroscience led to early deep learning research

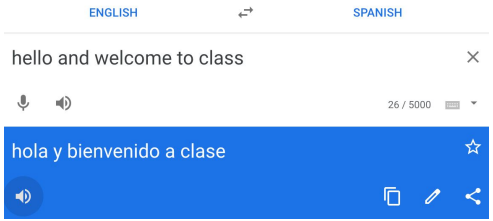
Course Introduction

Exciting ML developments have disrupted this field recently









2011: Apple Siri

2014: Microsoft Cortana, Amazon Alexa, Alexa Prize

2015: End-to-end neural becomes SORA

2016: Google assistant

2017: Neural TTS voice cloning

2020: Realtime speech-speech translation

A New Era of Spoken Language Applications and Impact

The Verge / Tech / Reviews / Science / Entertainment / More +

ARTIFICIAL INTELLIGENCE / TECH / POLITICS

Pakistan's former prime minister is using an AI voice clone to campaign from prison

/ Imran Khan's party crafted a four-minute message using a tool from the AI firm ElevenLabs.

By Amrita Khalid, one of the authors of audio industry newsletter Hot Pod. Khalid has covered tech, surveillance policy, consumer gadgets, and online communities for more than a decade.

Dec 18, 2023, 2:41 PM PST

1 Comment (1 New)

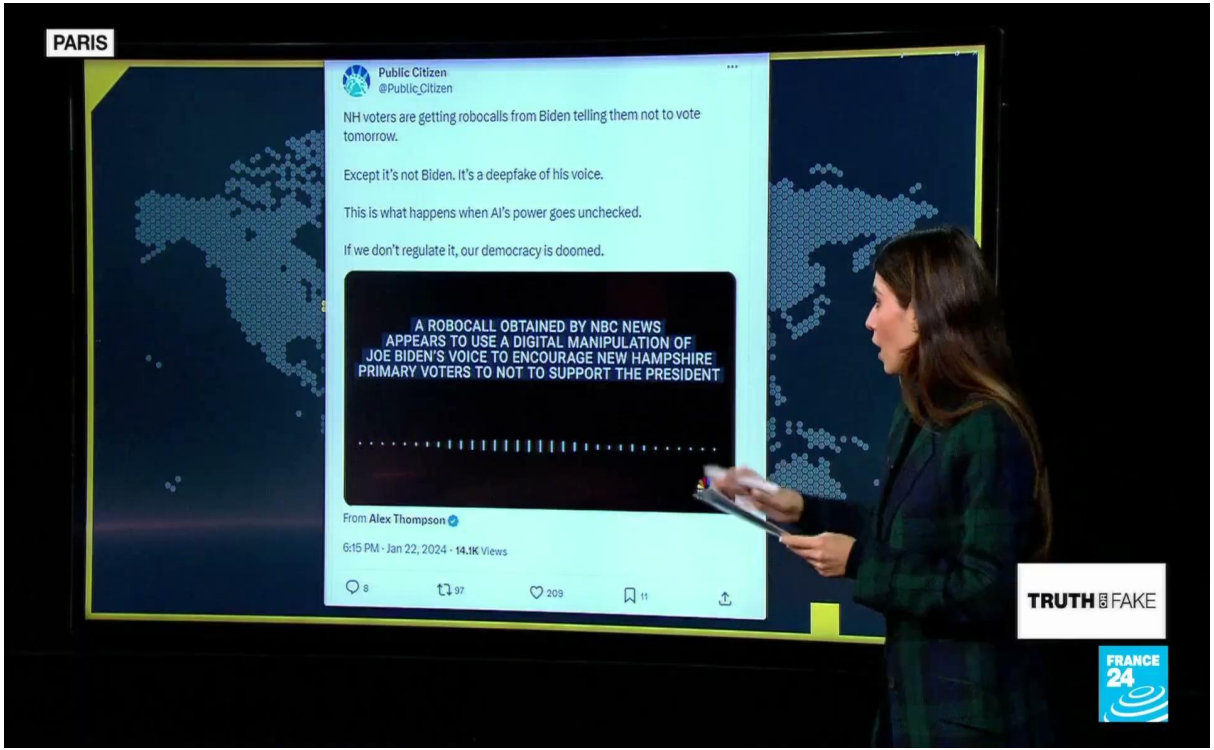
Verge Article

A New Era of Spoken Language Applications and Impact

VITTORIA ELLIOTT | MAKENA KELLY | POLITICS | JAN 23, 2024 12:58 PM

The Biden Deepfake Robocall Is Only the Beginning

An uncanny audio deepfake impersonating President Biden has sparked further fears from lawmakers and experts about generative AI's role in spreading disinformation.

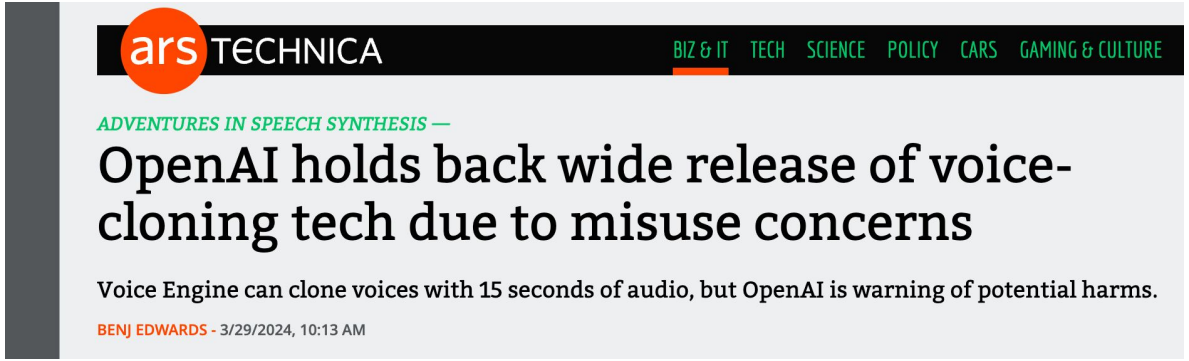


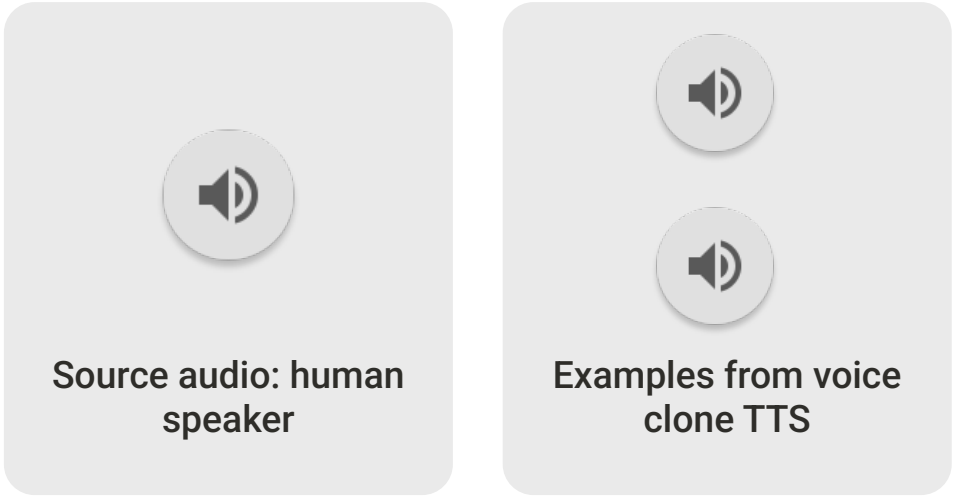
[Wired Article]

Discussion:

Clone anyone's voice with 5 seconds of sample audio

Would you approve releasing this tool publicly?



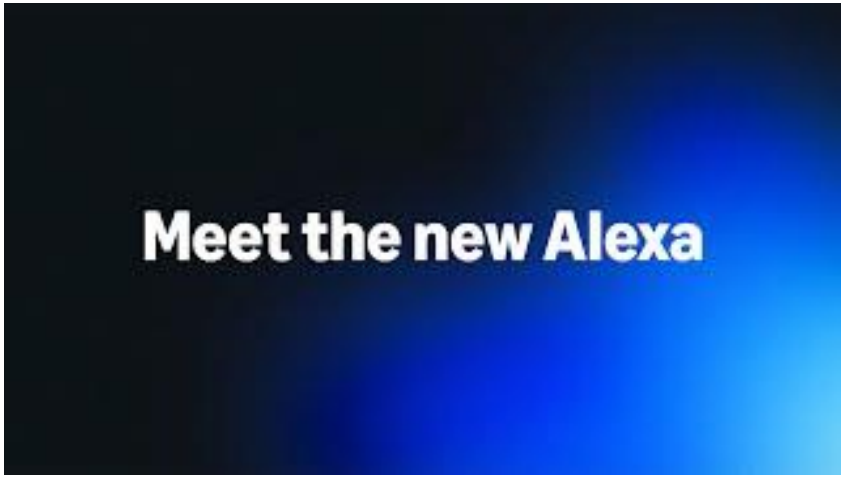


[The Verge article with samples]

A new generation of spoken digital assistants

Talk it out with Gemini Live

With Gemini Live, you can talk back and forth, get help with tasks, brainstorm ideas out loud, rehearse for important moments, and more. All in real-time using just your voice.

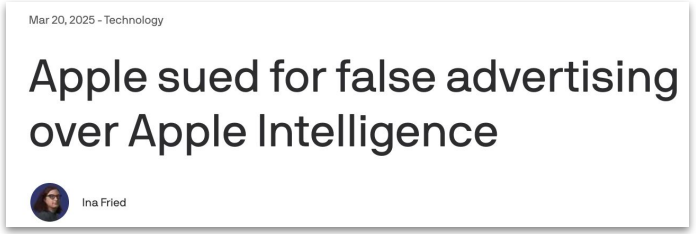


Recent over-promising and high-profile failures

Some AI agent customers say reality doesn't match the hype

BY SAGE LAZZARO
March 20, 2025 at 8:27 AM PDT

"CB Insights surveyed over 40 customers of AI agents and found that they're running into issues with reliability, integration, and security. Other recent headline events have highlighted some of the same issues. For instance, there was a surge of excitement over Manus, which was billed as the first fully autonomous "general agent" and lauded by some as another DeepSeek moment for China—until user tests revealed unreliable performance and questionable outputs." Fortune article



Mar 20, 2025 - Technology

Apple sued for false advertising over Apple Intelligence

Ina Fried

Sam Kohl
@iupdate · Follow

Apple ran this iPhone 16 ad in September 2024 showing off the new AI Siri that understands personal context.

It's now March 2025 and they just delayed the feature until sometime within the next year.

AI Siri is now an iPhone 17 feature that Apple promised for iPhone 16. Show more

Watch on X

12:02 PM · Mar 7, 2025

7K Reply Copy link

Read 432 replies

Lawsuit claims include direct references to Siri feature promises:
"*Apple's advertisements saturated the internet, television, and other airwaves to cultivate a clear and reasonable consumer expectation that these transformative features would be available upon the iPhone's release*"
Axios Article

So, why take this class?

Build skills to participate in the next 10 years of conversational AI invention + re-thinking how humans and computers collaborate

- Understand the challenges, models, and tools to build the next generation of conversational AI systems
- We've crossed a technology threshold – 100x more products and applications feasible to build now compared to even 5-10 years ago
- Design thinking for user experience and technology feasibility are a part of any spoken language system/product. You will learn how to think creatively about tech choices + user considerations
- Publicly-available models are increasingly good and cheap to access/use. Learn the core concepts so you can leverage new tools without getting lost in hype

Course Logistics

- Overview
- Requirements and Grading
- Course Projects
- Necessary Background
- Office Hours and CAs

Learning goals for this course

- **Develop expertise in working with spoken language using modern tools, and create a foundation for contributing to spoken language system development and research**
- **Questions you should be able to answer after this course**
  - How are the data and use cases for spoken language different from written language NLP, computer vision, or robotics tasks?
  - What are modern tools you might use in industry or research to develop a fully capable spoken language application? (e.g. for speech recognition, synthesis, voice cloning, and dialog tasks)
  - What is the most recent research on deep learning models and approaches to spoken language tasks? How do they compare with deep learning architectures for other domains?
  - For a new product or research project, what process and tools would you pursue to effectively work with spoken language? Which tools might you need to modify vs use as-is?

Course Logistics Overview

- http://www.stanford.edu/class/cs224s
- **Four homework assignments:**
  - 1 - Introduction to audio analysis tools, phonetic transcription, and modern spoken dialog systems
  - 2 - Working with speech tools, transcripts, and synthesizng audio.
  - 3 - Building a neural speech recognition model on conversational speech dataset.
  - 4 - Fine tuning speech recognition base models and working with non-English speech.
  - First homework will go out this Wednesday. Approximately 2 weeks for each homework (first one is shorter)
- **Optional self-directed project: Students can submit a project proposal to submit a research report instead of homework 4**
- **Gradescope for homework submission. Ed for questions. Use private Ed posts for personal/confidential questions**

Requirements and Grading

- **Readings:**
  - Jurafsky & Martin. Speech and Language Processing. 3rd edition pre-prints available online. *Purchasing textbook not required*
  - We expect you to read all readings posted to the course website syllabus. We may have occasional participation quizzes to check for attendance + reading background knowledge.
- **Reserve $60 for compute budget to complete assignments. We recommend using Colab (cost estimate based on this) but any GPU compute is fine provided you can complete the work.**
- **Grading:**
  - Homework 1: 20%
  - Homework 2: 22%
  - Homework 3: 23%
  - Homework 4 (or course project with approved proposal): 25%
  - Participation: 10%:
    - Occasional lecture + guest lecture attendance quizzes: 8%
    - Ed participation: 2%

Optional: course project in place of homework 4

- **Optional! Recommended only for those who are pursuing research or want to build an ambitious demo system**

- **You may choose to pursue work an a project instead of homework 4**
  - Submit proposal and receive staff approval to do this. You must get approval to opt-in to project work.
  - We will only approve projects with a solid starting point and framing of experiment goals.
- **What can you do as a project?**
  - First priority: *Build something you are proud of*
  - Full systems / demos, research papers on individual components, applying spoken language analysis to interesting datasets, etc. are all great projects
- **Combining projects with other courses is great!**
  - CS236G (GANs), CS224N, CS329S, CS229 all relevant
  - Need instructor permission to combine

- **See course website for more ideas and logistics. Project proposals not due until Week 5+ of the quarter.**

Necessary Background

- **Foundations of machine learning and natural language processing**
  - CS 124, CS 224N, CS 229, or equivalent experience
- **Mathematical foundations of neural networks**
  - Understand forward and back propagation in terms of equations
  - Deep learning architectures and basics of training models
  - We won't do a full deep learning tutorial during lecture. Use office hours or use CS224N / CS229 materials if you need additional background
- **Proficiency in Python**
  - Programming heavy homeworks will use Python, Colab Notebooks, and PyTorch

Office Hours and CAs

- Supporting both in-person and remote / online education students this quarter
- Andrew office hours: In person after class on Wednesdays (projects + other)
  - Zoom/remote office hours time TBD
- CAs: Office hour times TBD (homework questions here)
- Ed announcements for office hour times + posted to course website
- Questions on logistics? Use Ed first. Staff email list if needed,
  - DO NOT email individual staff – response times are slower in a large course if so. Fine to tag staff members on Ed or use private Ed posts as needed

Course Topics Overview

- **Task-oriented spoken dialogue systems / Conversational Agents**
- Speech Recognition (Speech to Text)
- Speech Synthesis (Text to Speech)
- Developing spoken language applications & conversational AI products

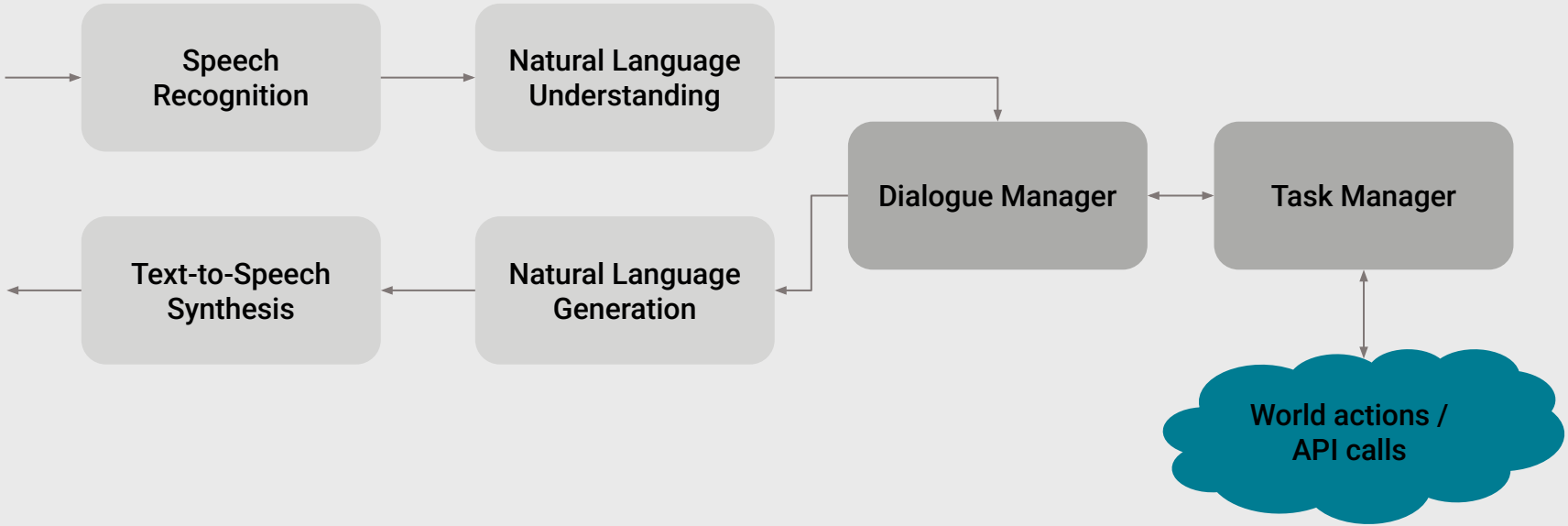
Task-oriented Spoken Dialogue Systems

- AKA spoken conversational agents. Many applications
- Personal Assistants (Alexa, Siri, etc.) is the big, obvious use case.
  - Many apps / spoken interfaces are actually small dialogue systems underneath

- Design considerations
  - Synchronous or asynchronous tasks
  - Pure speech, pure text, various multi-model user interaction types
  - Functionality versus personality. Task completion vs social chat

- Our focus in this class
  - Understand the problems. What do we need to solve when building spoken interfaces for tasks?
  - Define a language of designing a dialogue system similar to designing a software product
  - Evaluate some of the many technology solutions and components to develop and deploy dialogue systems. There is no single best approach, and techniques are evolving rapidly these days

Dialogue systems (= *Task-oriented conversational agents*)
Central dogma of dialog: Connecting language and actions



Compound AI systems: Module scope is a design choice

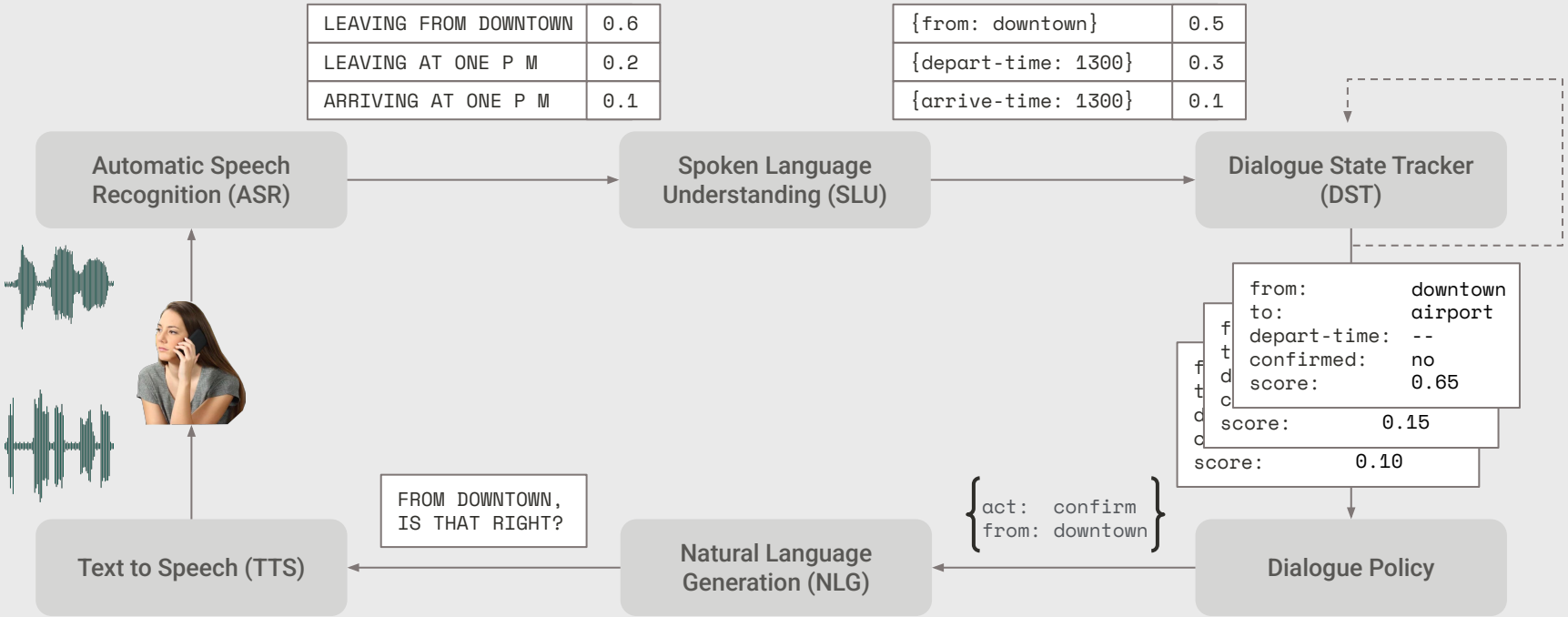


Figure: Architecture of dialogue-state system for task-oriented dialogue (William et al, 2016)

LLMs unlock new options for task-oriented dialog

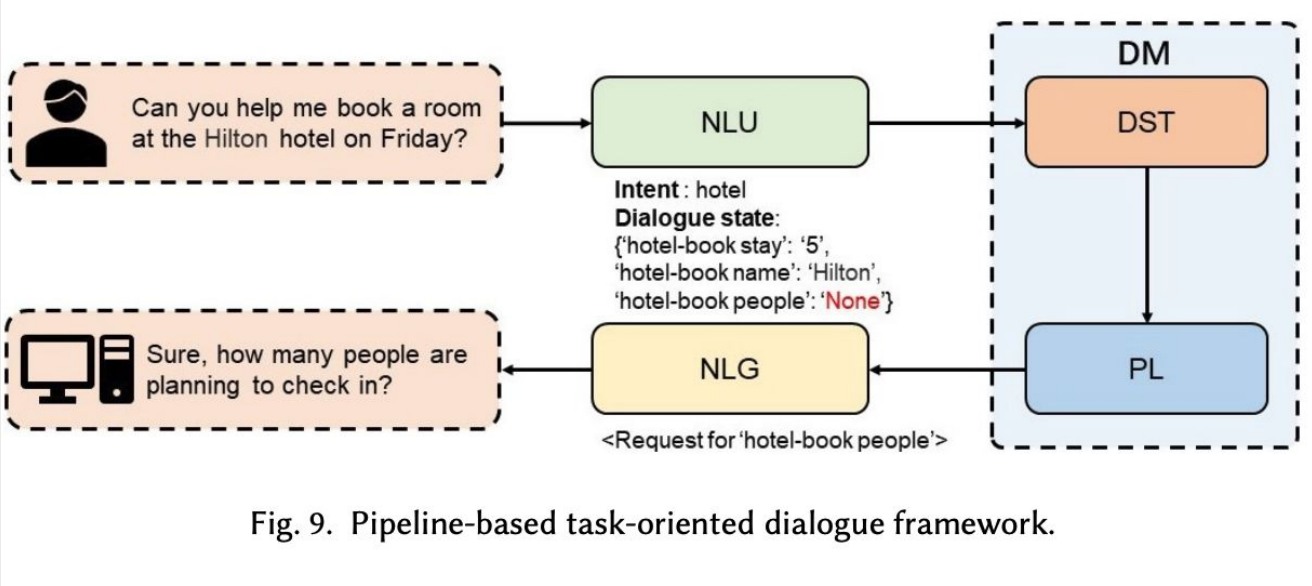


Fig. 9. Pipeline-based task-oriented dialogue framework.

LLMs allow dialog state to closely track action/task needs. Example state for flights:

```
DOMAIN: flight_book
ARGS: origin: SFO,
destination: _, type: one_way,
date: Feb_12_2026

DIALOG MOVES:
User_recent: info_provided:
date,
Internal_state: gathering_info
```

Figure: Architecture of LLM-based system with powerful natural language understanding (**NLU**) producing a semi-structured output for dialog state tracking (**DST**) via specialized tags for LLM-based tracking of dialogue state.

The policy learning (**PL**) module chooses actions (e.g. API calls) or responses in a semi-structured representation drawing on dialog state. The dialog policy produces high-level descriptions of responses for an LLM-based natural language generation (**NLG**) module to convert into a (possibly multi-modal) response (Yi et al. 2024).

System Design Process: Design Phase

1

Define overall system goal

2

Define set of task actions system can perform

3

Create example interactions

System Design Process: Technology Choices

1. Define overall system goal
2. Define set of task actions system can perform
3. Create example interactions
4. Define dialog manager approach (actions + dialog acts/state of system)
5. Choose NLU approach matching complexity of tasks and approach to initiative + dialog acts
6. Define NLG approach and dialog state -> NLG interface
7. Create a dialog policy (choosing next dialog action and sending to NLG)
8. Choose ASR/TTS approach. Update NLU/NLG if needed

Pre-LLM technology paradigms for dialogue

- **POMDP: Reinforcement learning, but difficult to scale up to practical system scope**
  - Partially-Observed Markov Decision Processes
  - Reinforcement Learning to learn what action to take and explore dialog policy space
  - Asking a question or answering one are just actions / dialog acts

- **Simple slot filling (ML or regular expressions)**
  - Pre-built frames / intents with information slots. Robust representation and maps to API / agent tools
    - Calendar: Who, When, Where
    - Flight: From, To, Date, flight number

- **NLU and NLG being brittle and narrow made dialogue systems slow, difficult to iteratively build**

- **Reusing search engine technology: Recognize command + issue search into existing NL search system (Google assistant did this for a while**

Course Topics Overview

- Task-oriented spoken dialogue systems / Conversational Agents
- **Speech Recognition (Speech to Text)**
- Speech Synthesis (Text to Speech)
- Developing spoken language applications & conversational AI products

Automatic Speech Recognition (ASR)

- **Large Vocabulary Continuous Speech Recognition (LVCSR) basic requirements**
  - Need to hand at least~64,000 lexemes (meaningful word-like units)
  - Speaker independent (vs. speaker-dependent). Should work for anyone with minimal set up
  - Continuous speech (vs isolated-words or short commands)
  - Robust to background noise and capable of handling multi-party conversations

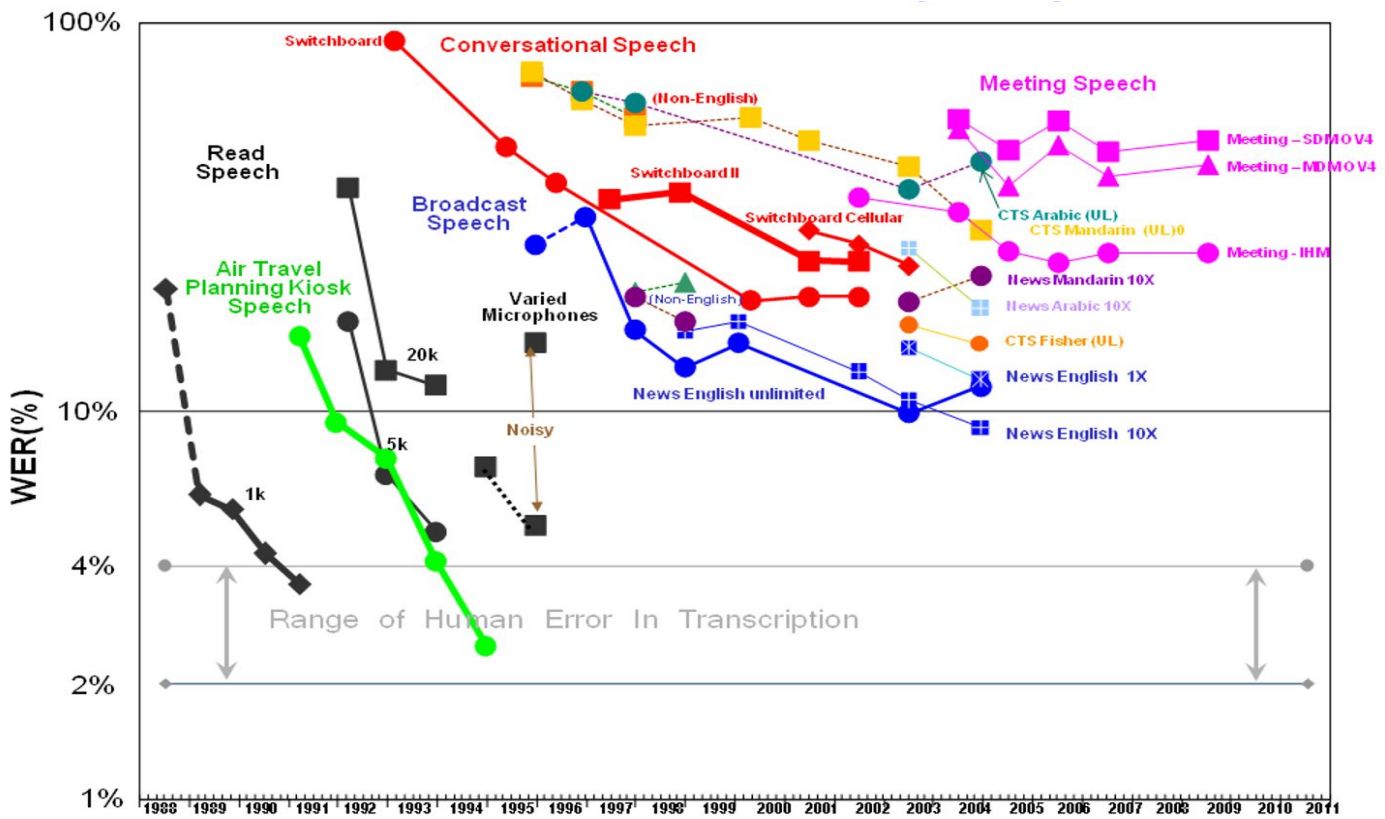
- **ASR systems widely available for major languages in cloud-connected settings. Current focus areas and open challenges:**
  - On-device and compute-constrained ASR systems. Actively re-thinking major components to integrate new deep learning modules
  - Accents, dialects, and support for non-native language speakers. ASR performance can fall off rapidly for various speaker sub-groups
  - Low-resource languages and developing multi-lingual systems
  - Integrating ASR deeply with other modalities for audio-visual or audio-text joint processing

Modern word/character error rates on high-resource langs

English Tasks	WER%
LibriSpeech audiobooks 960hour clean	1.4
LibriSpeech audiobooks 960hour other	2.6
Switchboard telephone conversations between strangers	5.8
CALLHOME telephone conversations between family	11.0
Sociolinguistic interviews, CORAAL (AAVE)	27.0
CHiMe5 dinner parties with body-worn microphones	47.9
CHiMe5 dinner parties with distant microphones	81.3
Chinese (Mandarin) Tasks	**CER%**
AISHELL-1 Mandarin read speech corpus	6.7
HKUST Mandarin Chinese telephone conversations	23.5

Figure 2: Rough Word Error Rates (WER = % of words misrecognized) reported around 2020 for ASR on various American English recognition tasks, and character error rates (CER) for two Chinese recognition tasks

Large-scale deep learning helped break accuracy barriers



Conversational speech is harder than you realize

A piece of utterance without context

A piece of utterance with context

Human vs machine speech recognition. What mistakes?

Deletions				Insertions			
SWB		CH		SWB		CH	
ASR	Human	ASR	Human	ASR	Human	ASR	Human
30: it	19: i	46: i	20: i	13: i	16: is	23: a	17: is
20: i	17: it	46: it	18: and	10: a	14: %hes	14: is	17: it
17: that	16: and	39: and	15: it	7: and	12: i	11: i	16: and
16: a	14: that	32: is	15: the	7: of	11: and	10: are	14: have
14: and	14: you	26: oh	14: is	6: you	9: it	10: you	13: a
14: oh	12: is	25: a	13: not	5: do	6: do	9: the	13: that
14: you	12: the	20: to	10: a	5: the	5: have	8: have	12: i
12: %bcack	11: a	19: that	10: in	5: yeah	5: yeah	8: that	11: %hes
12: the	10: of	19: the	10: that	4: air	5: you	7: and	10: not
11: to	9: have	18: %bcack	10: to	4: in	4: are	7: it	9: oh

Table 1: Most frequent deletion and insertion errors for humans and ASR system on SWB and CH. (Saon et al, 2017)

SWB		CH	
ASR	Human	ASR	Human
11: and / in	16: (%hes) / oh	21: was / is	28: (%hes) / oh
9: was / is	12: was / is	16: him / them	22: was / is
7: it / that	7: (i-) / %hes	15: in / and	11: (%hes) / %bcack
6: (%hes) / oh	5: (%hes) / a	8: a / the	10: bentsy / benji
6: him / them	5: (%hes) / hmm	8: and / in	10: yeah / yep
6: too / to	5: (a-) / %hes	8: is / was	9: a / the
5: (%hes) / i	5: could / can	8: two / to	8: is / was
5: then / and	5: that / it	7: the / a	7: (%hes) / a
4: (%hes) / %bcack	4: %bcack / oh	7: too / to	7: the / a
4: (%hes) / am	4: and / in	6: (%hes) / a	7: well / oh

Table 2: Most frequent substitution errors for humans and ASR system on SWB and CH (Saon et al, 2017)

Why are Accents Hard?

A word by itself

A word in context

Do multi-modal LLMs / large transformers “just solve” speech?

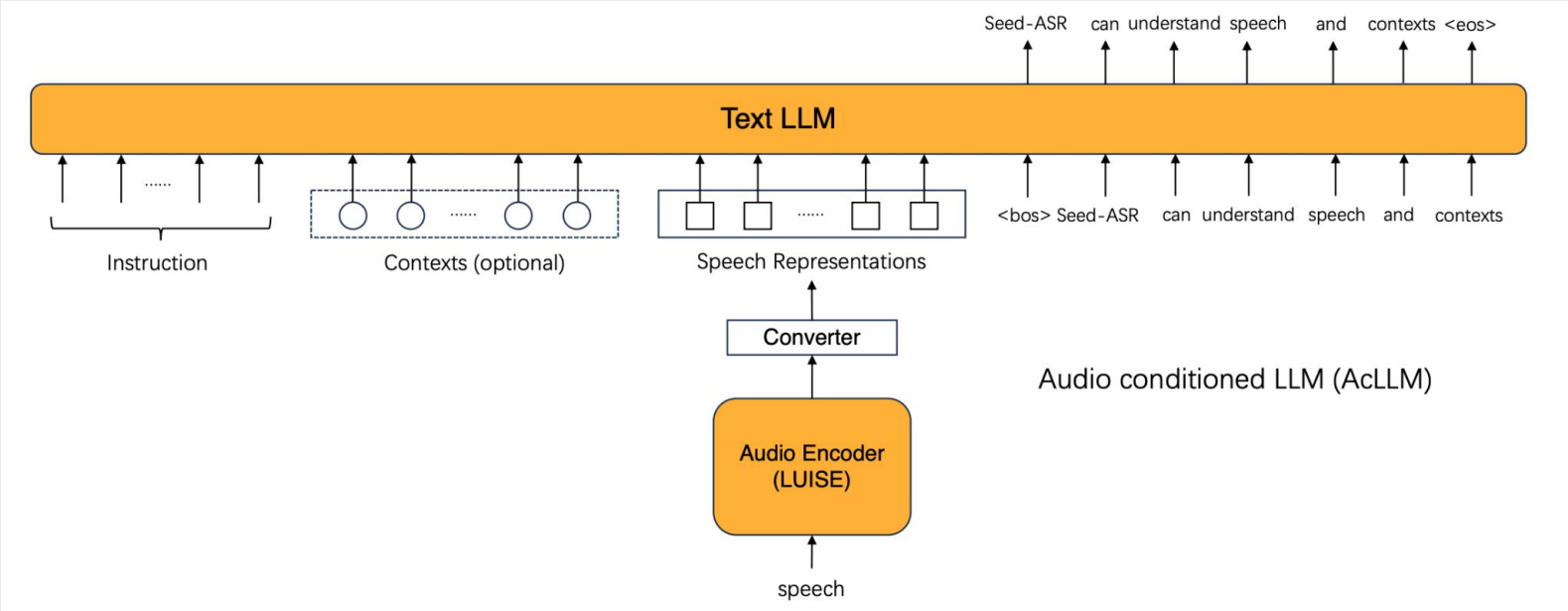
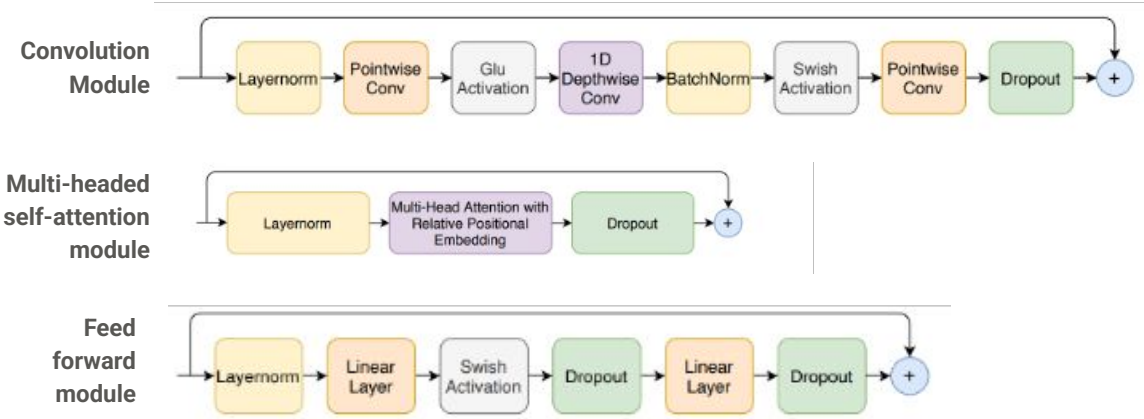


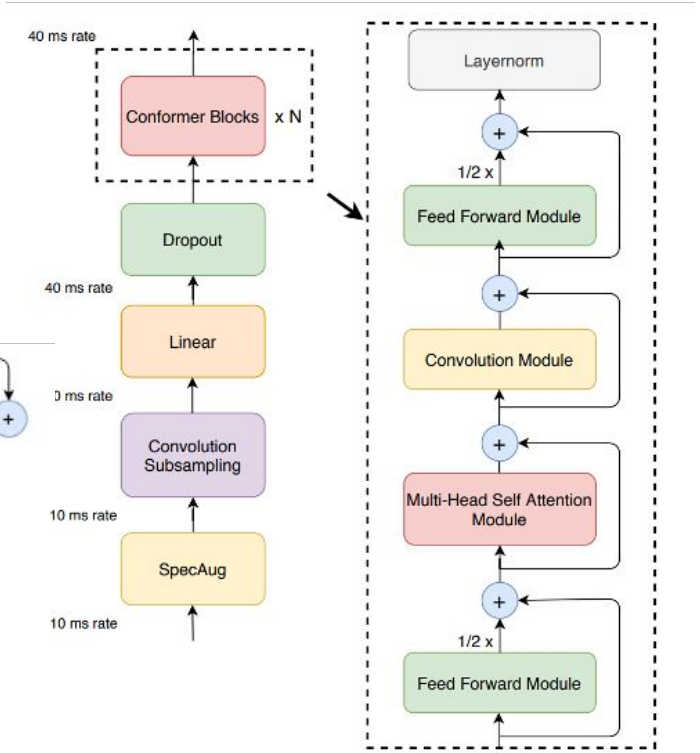
Figure: Audio-conditioned LLM (AcLLM) speech recognition approaches re-use text-based LLMs and add an audio input representation via a specialized encoder.

Seed-ASR uses an audio encoder and prompted LLM to transcribe speech. When contexts are provided, the instruction is "There are relevant contexts, transcribe the speech into text:". Otherwise, the instruction is "Transcribe the speech into text:". (Bai et al, 2024)

Conformer: Convolution-augmented Transformer for Speech Recognition

- Sequence-to-sequence transformer with multi-headed self attention. Directly optimizes target word sequence
- Combines attention (global context) with convolution (local invariance)





(Gulati *et al*, 2020)

Is speech recognition solved? Why study it vs use some API?

- **In the last ~14 years**
  - Dramatic reduction in LVCSR error rates (16% to 3%) in ideal conditions
  - Human-level LVCSR performance on conversational speech in realistic conditions
  - New deep learning paradigms for recognizers (end-to-end neural networks and foundation models)
- **Understanding how ASR works enables better ASR-enabled systems**
  - What types of errors are easy to correct?
  - How can a downstream system make use of uncertain outputs?
  - How much would building our own improve on an API?
- **Next generation of ASR challenges as systems go live on phones and in homes**
- **Now we can finally hold ourselves accountable for developing speech technology that works well for everyone! Regardless of language, accent, disability, or other variations in speech**

Course Topics Overview

- Task-oriented spoken dialogue systems / Conversational Agents
- Speech Recognition (Speech to Text)
- **Speech Synthesis (Text to Speech)**
- Developing spoken language applications & conversational AI products

TTS (= Text-to-Speech) (= Speech Synthesis)

- Produce speech from a text input. Ideally controllable, accurate words and emotive styles
- Applications:
  - Virtual Assistants
    - Amazon Alexa
    - Apple Siri
    - Google Gemini / Assistant
  - Games
  - Announcements / voice-overs
- New variation: Voice cloning. TTS systems that mimic particular speakers with minimal training data. Modern systems can learn multi-speaker models and even interpolate voices
- General recipe: Collect lots of speech (5-50 hours) from one speaker, transcribe very carefully. Supervised ML + careful audio signal processing create high quality voices
- Rapid recent progress in neural approaches. Modern systems are DNN-based, understandable, but not yet fully emotive

Text to speech with end-to-end neural models

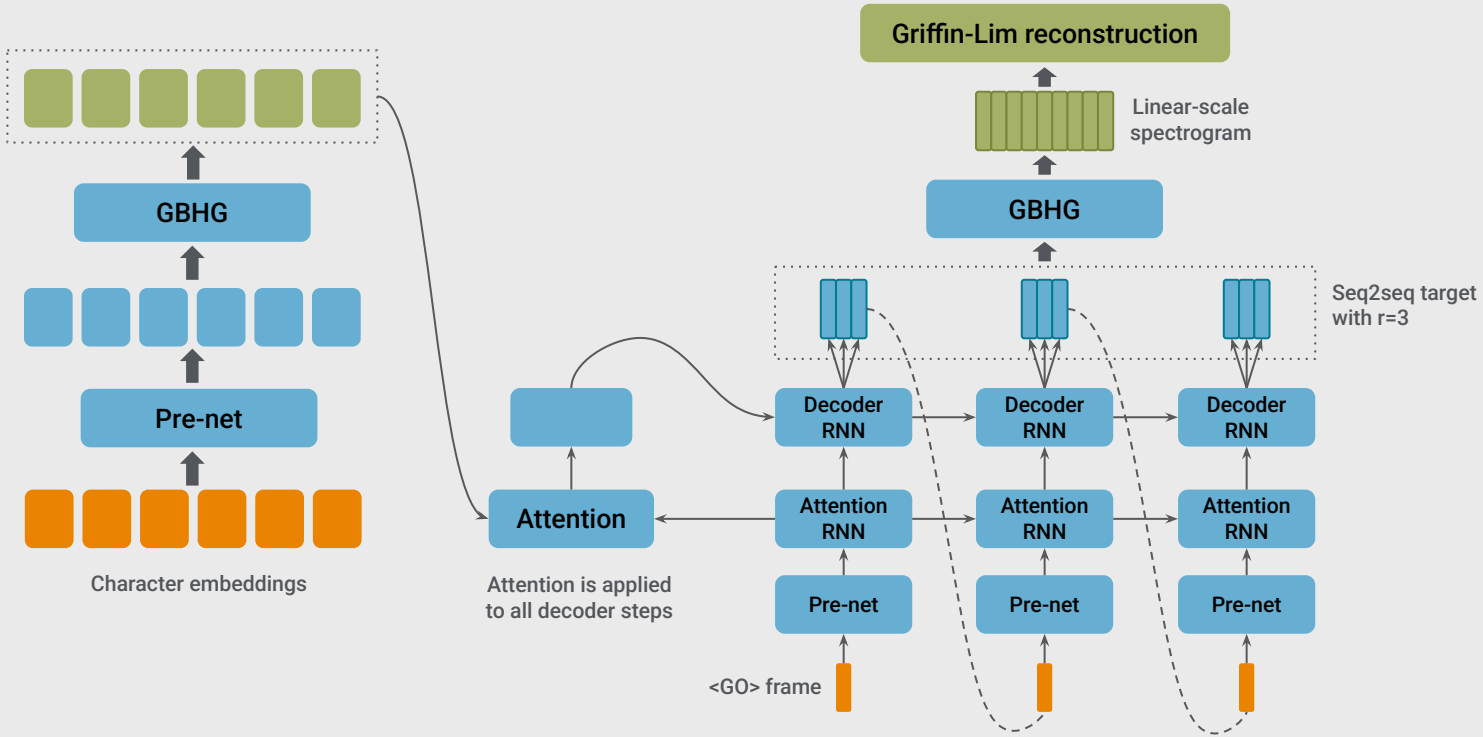
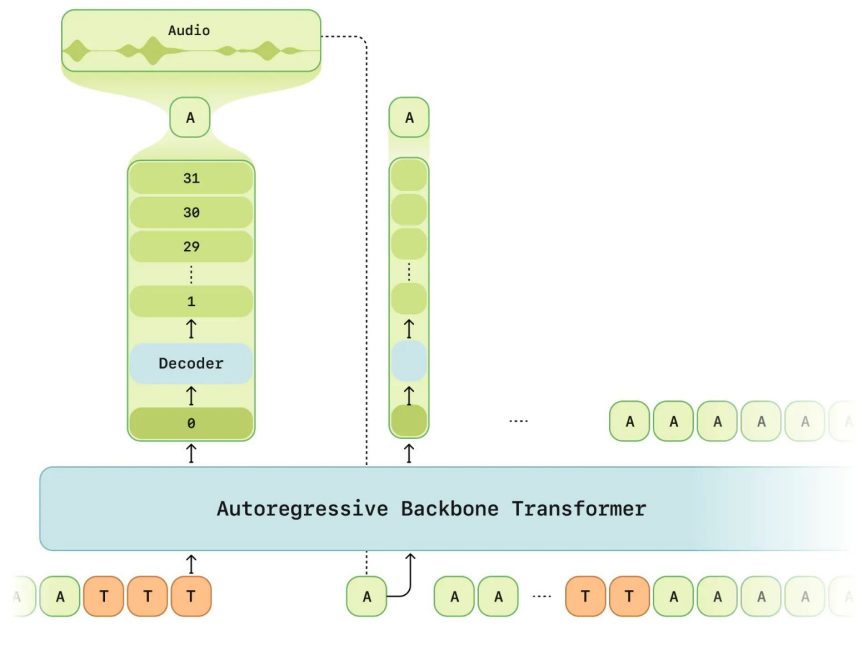


Figure: Tacotron Model Architecture. The model takes characters as input and outputs the corresponding raw spectrogram, which is then fed to the Griffin-Lim reconstruction algorithm to synthesize speech. (Wang et al, 2017)

Very recent: Speech input+output with audio LLMs



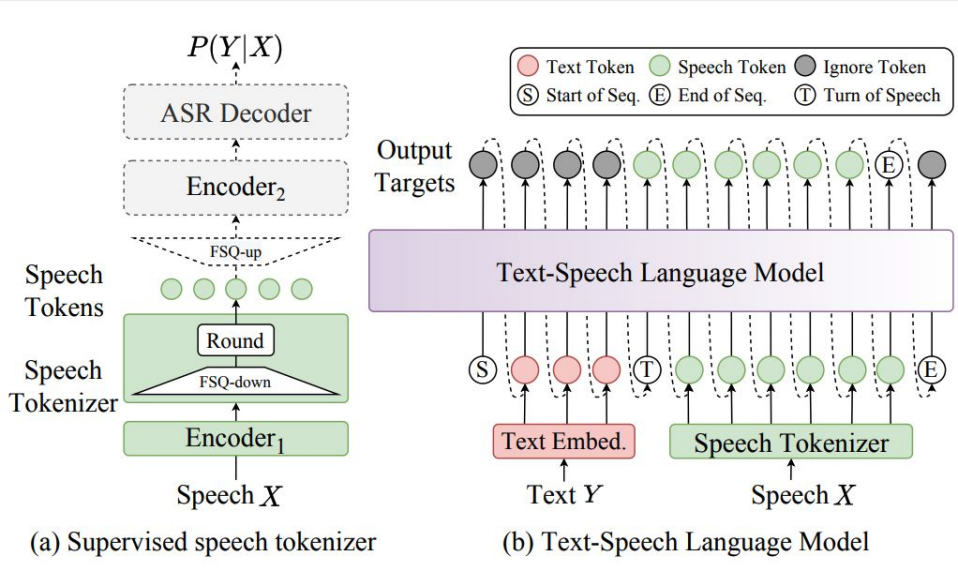


Figure: Speech in / speech out models attempt to fuse NLU, NLG, ASR, and TTS in various ways. Sesame (left) conversational speech model (CSM) which interleaves audio (A) and text (T) tokens in a multimodal, autoregressive token transformer (Iribe *et al*, 2025).

CosyVoice2 (right) uses a text-speech LM with specialized audio encoder and speech audio output generator from speech tokens to achieve interruptible, full-duplex spoken conversations (Du *et al*, 2024).

Course Topics Overview

- Task-oriented spoken dialogue systems / Conversational Agents
- Speech Recognition (Speech to Text)
- Speech Synthesis (Text to Speech)
- **Developing spoken language applications & conversational AI products**

Extraction of Social Meaning from Speech

- Given speech and text from a conversation
- Can we tell if a speaker is
  - Awkward?
  - Flirtatious?
  - Friendly?
- Dataset:
  - 1000 4-minute “speed-dates”
  - Each subject rated their partner for these styles
  - The following segment has been lightly signal-processed

- Caveat: Meaning extraction is largely based on supervised machine learning these days
  - Training dataset breadth and accuracy/consistency of labels is critical
  - Easy to create biased, inaccurate systems due to training data inconsistencies or lack of data coverage

Supervised meaning extraction + shared encoder features

Model	Speech	Input format	Framework	Encoder	Loss	Inspired by
LIM [36]	✓	raw waveform	(d)	SincNet	BCE, MINE or NCE loss	SimCLR
COLA [36]	✗	log mel-filterbanks	(d)	EfficientNet	InfoNCE loss	SimCLR
CLAR [33] (semi)	✗	raw waveform log mel-spectrogram	(d)	1D ResNet-18 ResNet-18	NT-Xent + cross-entropy	SimCLR
Fonseca et al. [36]	✗	log mel-spectrogram	(d)	ResNet, VGG, CRNN	NT-Xent loss	SimCLR
Wang et al. [88]	✗	raw waveform + log mel-filterbanks	(d)	CNN ResNet	NT-Xent loss + cross-entropy	SimCLR
BYOL-A [89]	✗	log mel-filterbanks	(b)	CNN	MSE loss	BYOL
Speech2Vec [48]	✓	mel-spectrogram	(a)	RNN	MSE loss	Word2Vec
Audio2Vec [91]	✓✗	MFCCs	(a)	CNN	MSE loss	Word2Vec
Carr [67]	✓	MFCCs	(a)	Context-free network	Fenchel-Young loss	-
Ryan [68]	✗	constant-Q transform spectrogram	(a)	AlexNet	Triplet loss	- -
Mockingjay [92]	✓	mel-spectrogram	(a)	Transformer	L1 loss	BERT
TERA [93]	✓	log mel-spectrogram	(a)	Transformer	L1 loss	BERT
Audio ALBERT [94]	✓	log mel-spectrogram	(a)	Transformer	L1 loss	BERT
DAPC [95]	✓	spectrogram	(a)	Transformer	Modified MSE loss + orthogonality penalty	BERT
PASE [96]	✓	raw waveform	(a)	SincNet + CNN	L1, BCE loss	BERT
PASE+ [97]	✓	raw waveform	(a)	SincNet + CNN + QRNN	MSE, BCE loss	BERT
CPC [40]	✓	raw waveform	(a)	ResNet + GRU	InfoNCE loss	-
CPC v2 [59]	✓	raw waveform	(a)	ResNet + Masked CNN	InfoNCE loss	-
CPC2 [98]	✓	raw waveform	(a)	ResNet + LSTM	InfoNCE loss	-
Wav2Vec [84]	✓	raw waveform	(a)	1D CNN	Contrastive loss	-
VQ-Wav2Vec [85]	✓	raw waveform	(a)	1D CNN + BERT	Contrastive loss	BERT
Wav2Vec 2.0 [81]	✓	raw waveform	(a)	1D CNN + Transformer	Contrastive loss	BERT
HuBERT [99]	✓	raw waveform	(c)	1D CNN + Transformer	Contrastive loss	BERT

Table: An overview of the recent audio self-supervised learning methods. The "speech" column distinguishes whether a method addresses speech tasks or for general purpose audio representations. (Liu *et al*, 2022)

Some basic ethics when working on speech technologies

Don’t record someone without their consent

In California, all parties to any confidential conversation must give their consent to be recorded. For calls occurring over cellular or cordless phones, all parties must consent before a person can record, regardless of confidentiality.

Don’t use someone’s speech data without consent. Especially with speech synthesizers / voice cloning

It might be fun, but it’s a little creepy. People get upset.

Okay to use existing speech datasets (we’ll provide some).

Do consider subgroup and language bias in systems

Poor performance on subgroups e.g. non-native speakers. Many languages are under-served relative to English/Mandarin.

Social meaning extraction task supervision labels can create bias

Thank you. Questions?

- Homework 1 released Wednesday. Start course readings (see syllabus on site)
- Join Ed. Office hours announced later this week.