ML for Audio Study Group

Session 1:

December 10, 2021, 5 PM CET

hf.co/join/discord

Omar Sanseviero

Vaibhav (VB) Srivastav

JOIN

Suggested readings before this session

- https://heartbeat.comet.ml/the-3-deep-learning-frameworks-for-end-to-end-speech-recognition-that-power-your-devices-37b891ddc380

Organisation

- **Community-led!**
  - We’ll kick off with some basics, but we’ll decide collaboratively where we want to focus
  - Anyone can participate!
  - Members of the HF team and other cool collaborators will join.
- Expectation
  - Before each session: **Read/watch related resources**
  - During each session, you can
    - Ask question in the forum
    - Present a short (~10-15mins) presentation on the topic (agree beforehand)
    - Participate a bit more passively (that’s also ok and you’re welcomed!)
  - Before/after:
    - Keep discussing/asking questions about the topic
    - Share interesting resources

Introduction

Omar Sanseviero (https://twitter.com/osanseviero)

- ML Engineer at Hugging Face
- Previously
  - SWE at Google Assistant
  - Co-founder AI Learners

Vaibhav Srivastav (https://twitter.com/reach_vb)

- MS student @ Uni Stuttgart/ Working Student @ Deloitte Tax
- Previously
  - Strategy @ Deloitte Consulting

Timeline

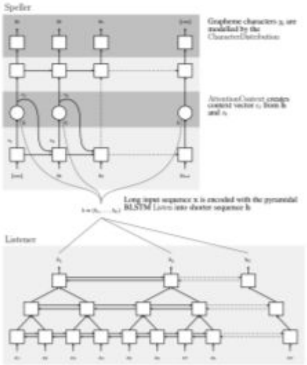
- **Dec 14: Kick off session**
- Dec 21: ASR Deep Dive
- Dec 28: TTS Deep Dive
- Jan 11 and forward:
  - Paper discussions
  - Invited speakers
  - Deep dive into a specific task

An exciting time for spoken language processing

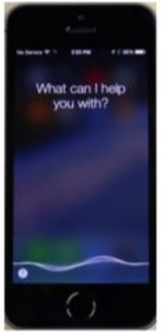
Amazon Alexa + Alexa Prize 2014

THIS IS AN ARTIFICIAL VIDEO CREATED BY LYREBIRD.
IT DOES NOT CONVEY THE OPINION OF BARACK OBAMA.

Neural TTS voice cloning 2017

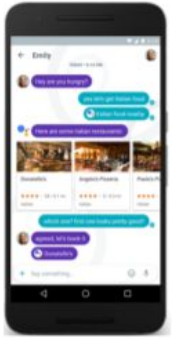

Speller
Grapheme characters y_i are modelled by the CharacterDistribution
AttentionContext creates context vector c_i from h and s_i
Long input sequence x is encoded with the pyramidal BLSTM Listen into shorter sequence h
Listener

End-to-end neural becomes SOTA 2015 - present



What can I help you with?

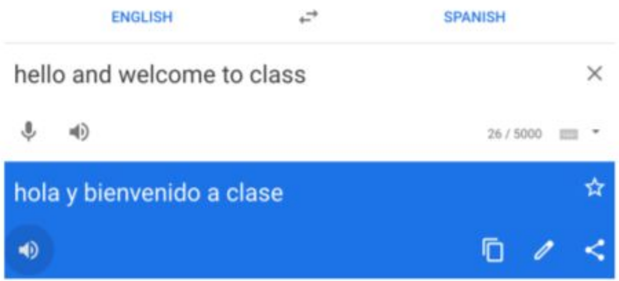
Apple Siri 2011



Google Assistant 2016

Use the area below to ask me anything.

Microsoft Cortana 2014

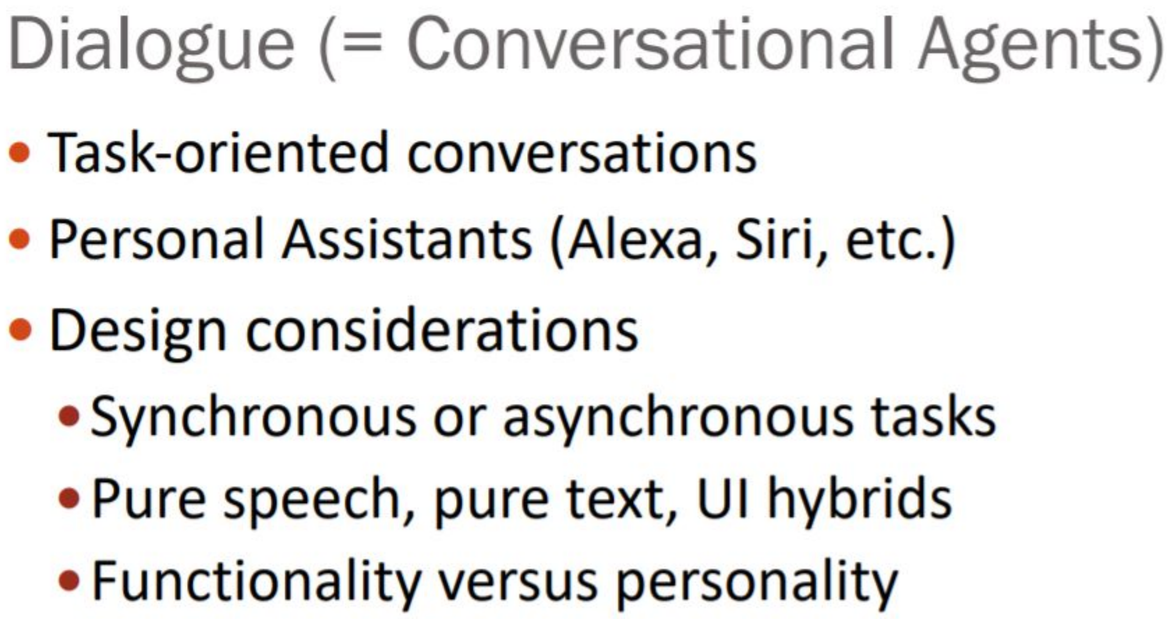


hello and welcome to class

hola y bienvenido a clase

Realtime speech-speech translation 2020

from - stanford CS224S

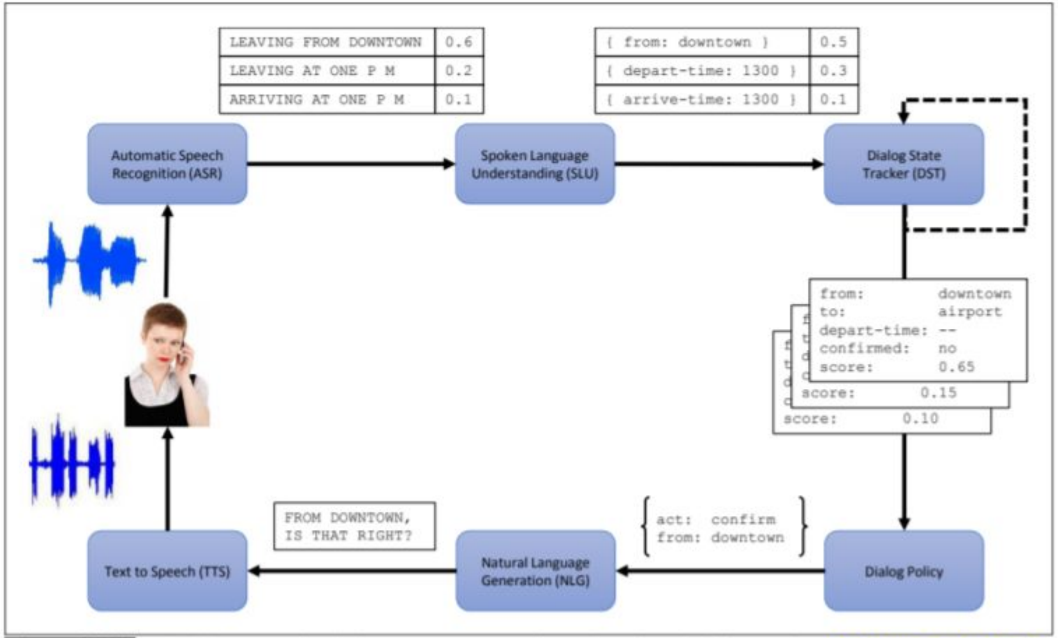


Dialogue (= Conversational Agents)

- Task-oriented conversations
- Personal Assistants (Alexa, Siri, etc.)
- Design considerations
  - Synchronous or asynchronous tasks
  - Pure speech, pure text, UI hybrids
  - Functionality versus personality

from - stanford CS224S

Dialogue (= Conversational Agents)



Automatic Speech Recognition (ASR)
Spoken Language Understanding (SLU)
Dialog State Tracker (DST)
Dialog Policy
Natural Language Generation (NLG)
Text to Speech (TTS)

LEAVING FROM DOWNTOWN	0.6
LEAVING AT ONE P M	0.2
ARRIVING AT ONE P M	0.1

{ from: downtown }	0.5
{ depart-time: 1300 }	0.3
{ arrive-time: 1300 }	0.1

```
from:          downtown
to:            airport
depart-time:   --
confirmed:     no
score:         0.65
```

```
score:         0.15
```

```
score:         0.10
```

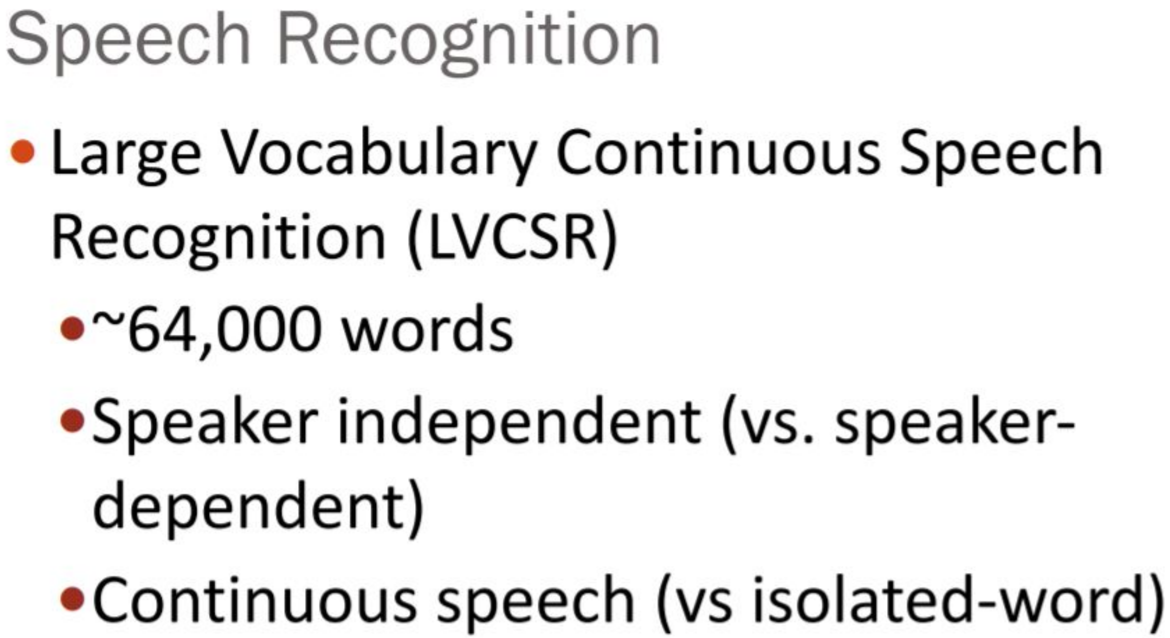
```
{ act:   confirm
  from:  downtown }
```

```
FROM DOWNTOWN,
IS THAT RIGHT?
```

Figure 26.11 Architecture of a dialogue-state system for task-oriented dialogue from Williams et al. (2016).

from - stanford CS224S

Speech Recognition

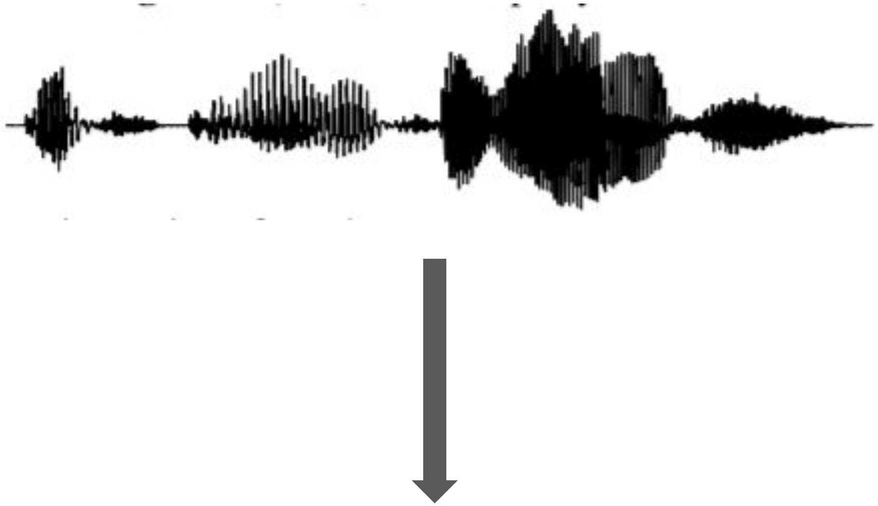


Speech Recognition

- Large Vocabulary Continuous Speech Recognition (LVCSR)
  - ~64,000 words
  - Speaker independent (vs. speaker-dependent)
  - Continuous speech (vs isolated-word)

from - stanford CS224S

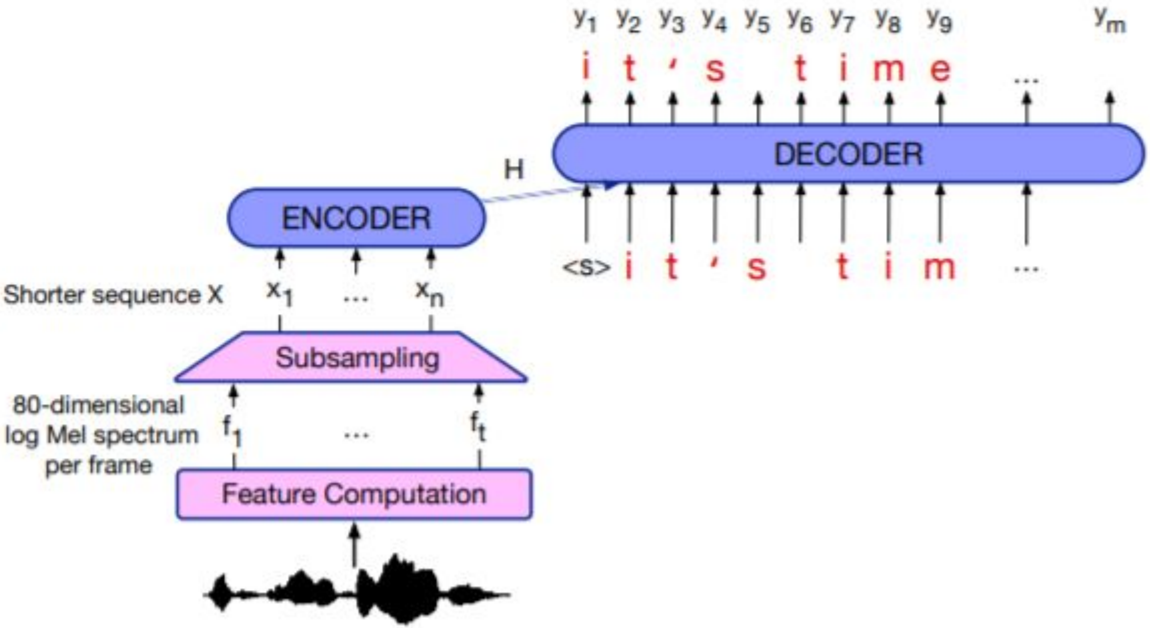
A slide explaining ASR architecture



It’s time for lunch!

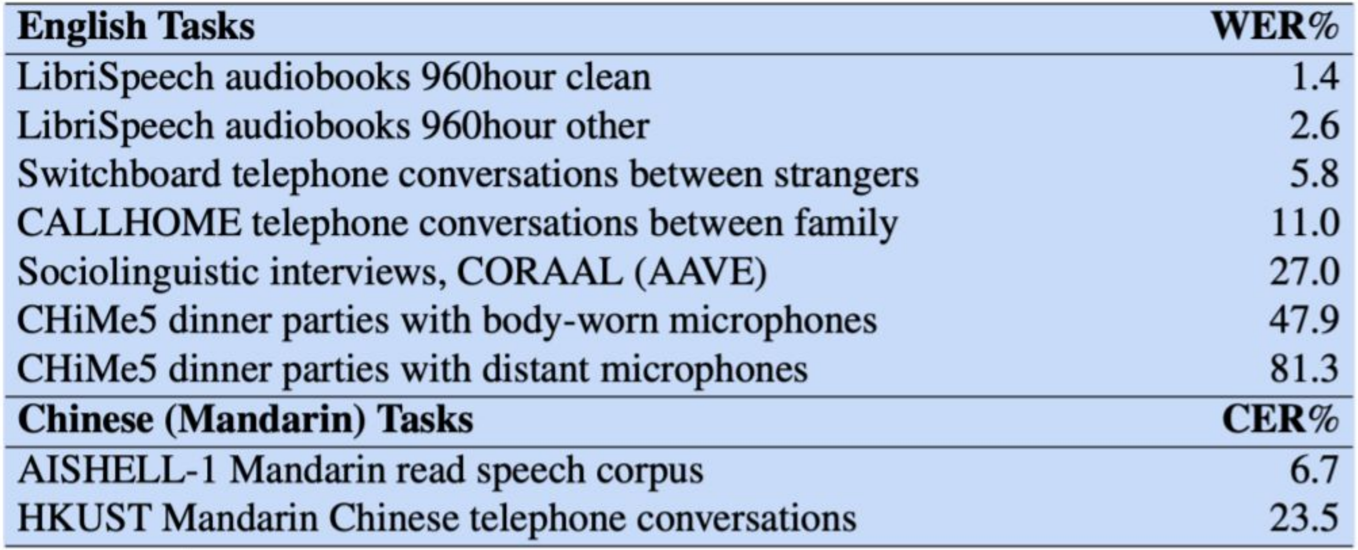
from - SLP3, Ch 26

Basic architecture for ASR



from - SLP3, Ch 26

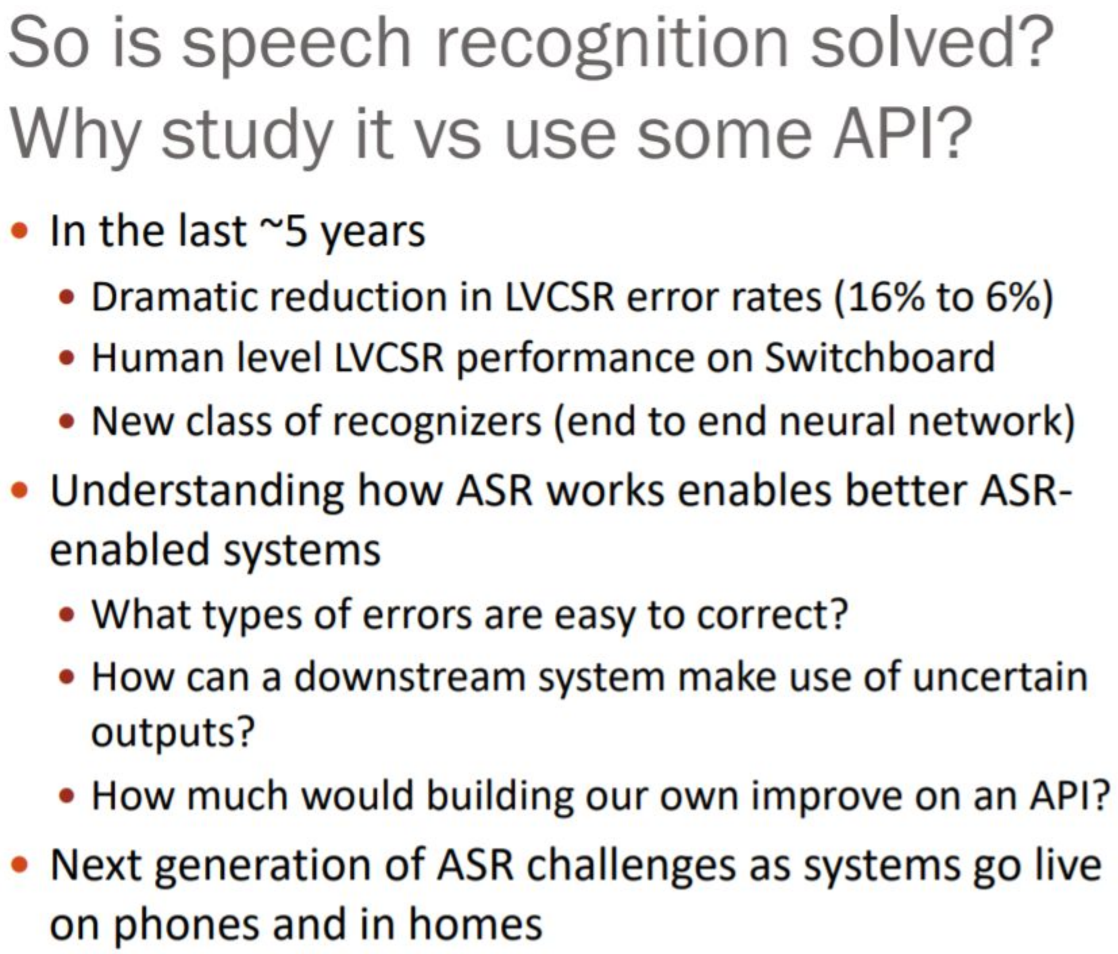
Current error rates



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 27.1 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.

from - stanford CS224S

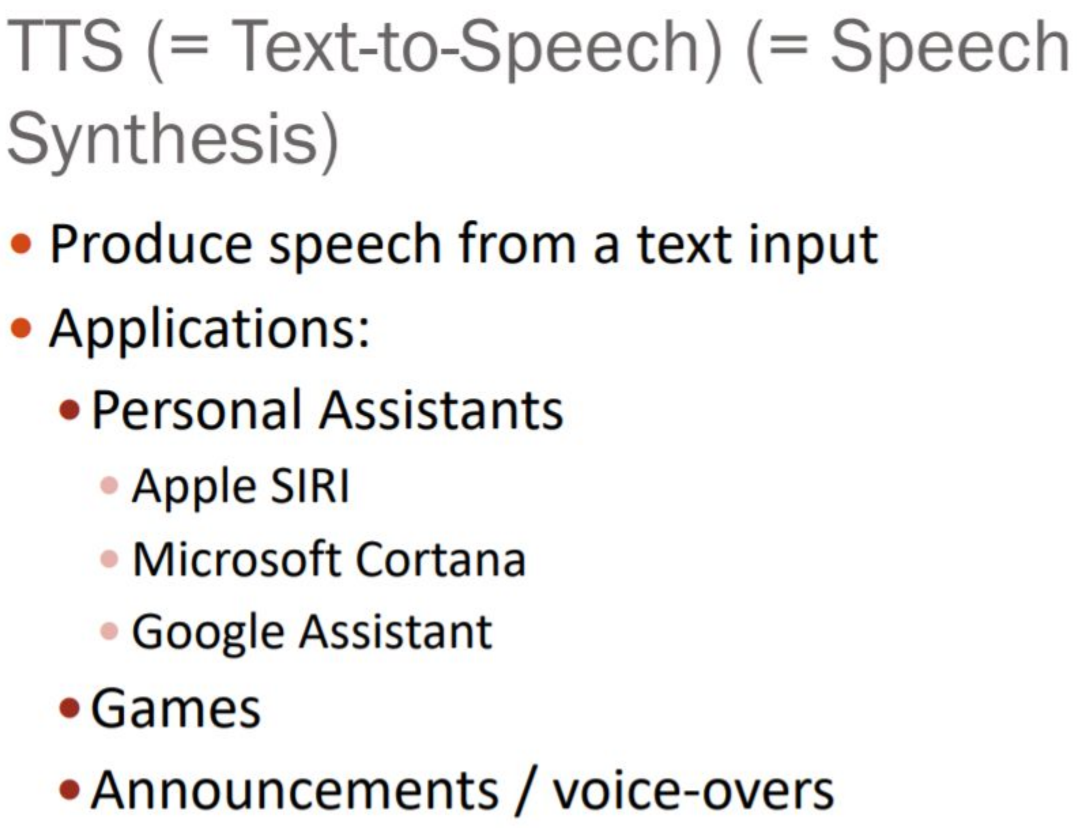


So is speech recognition solved? Why study it vs use some API?

- In the last ~5 years
  - Dramatic reduction in LVCSR error rates (16% to 6%)
  - Human level LVCSR performance on Switchboard
  - New class of recognizers (end to end neural network)
- 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

from - stanford CS224S

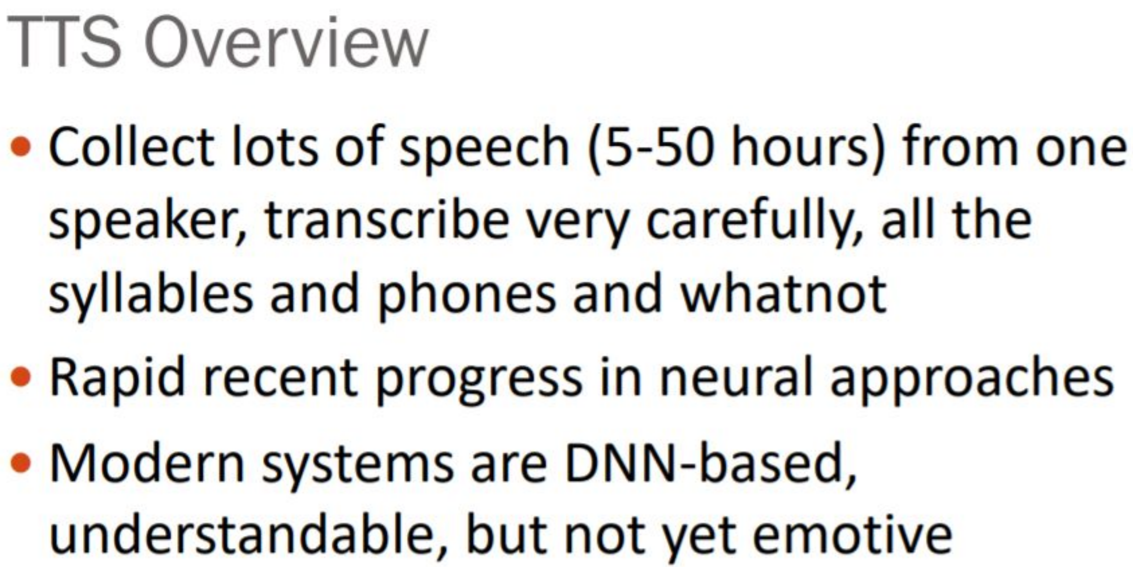
Speech Synthesis



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

- Produce speech from a text input
- Applications:
  - Personal Assistants
    - Apple SIRI
    - Microsoft Cortana
    - Google Assistant
  - Games
  - Announcements / voice-overs

from - stanford CS224S



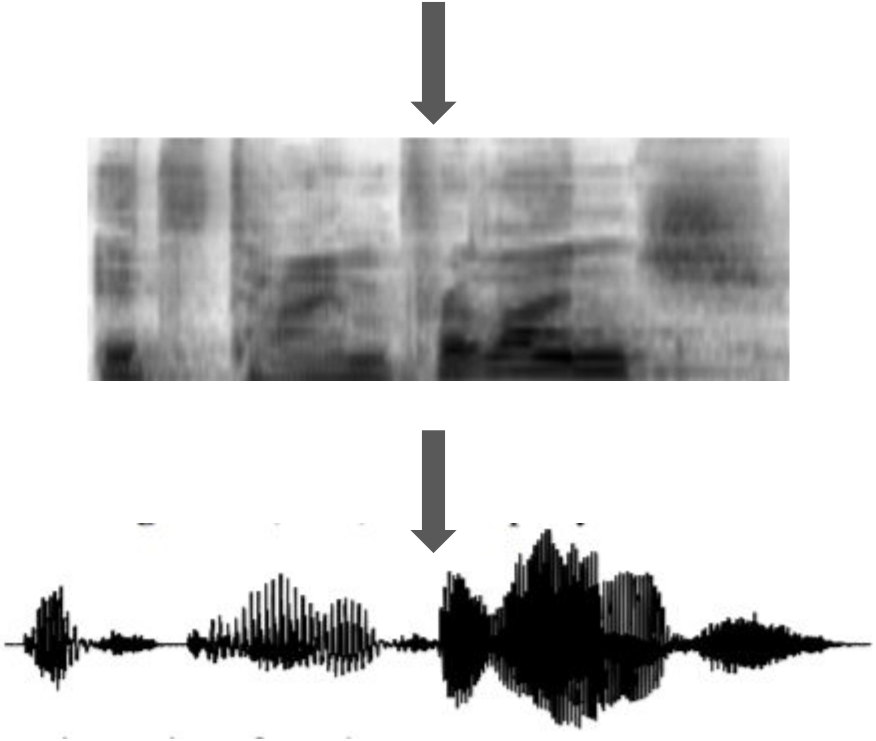
TTS Overview

- Collect lots of speech (5-50 hours) from one speaker, transcribe very carefully, all the syllables and phones and whatnot
- Rapid recent progress in neural approaches
- Modern systems are DNN-based, understandable, but not yet emotive

from - stanford CS224S

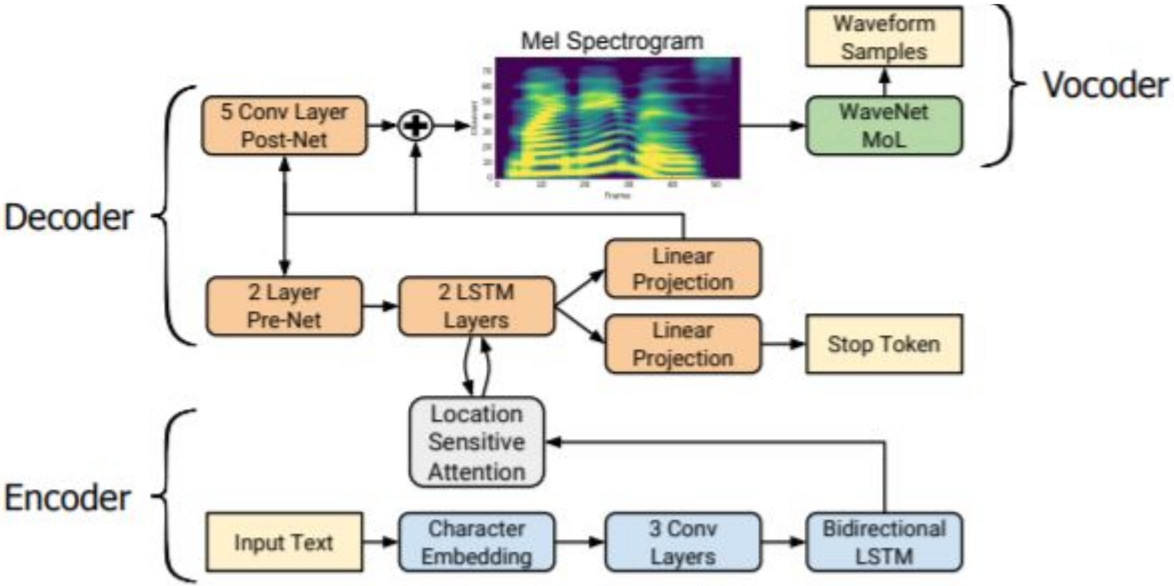
Text to Speech

It’s time for lunch!

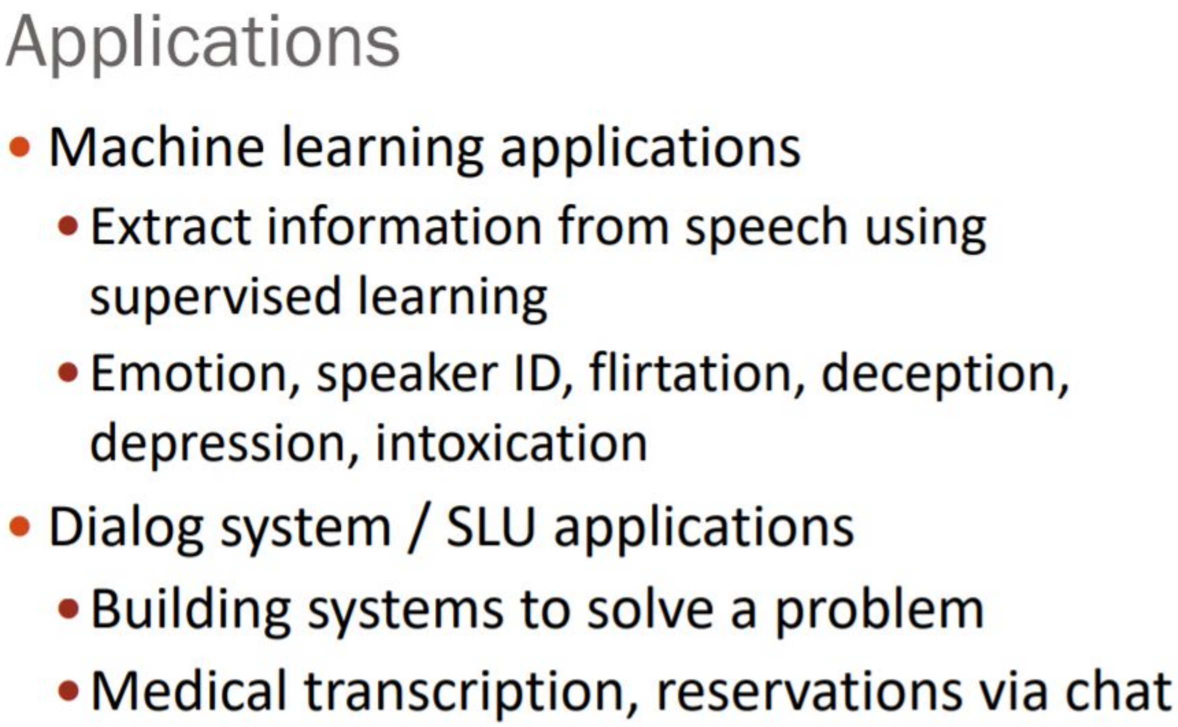


from - [SLP3, Ch 26]

Text to Speech



from - SLP3, Ch 26



Applications

- Machine learning applications
  - Extract information from speech using supervised learning
  - Emotion, speaker ID, flirtation, deception, depression, intoxication
- Dialog system / SLU applications
  - Building systems to solve a problem
  - Medical transcription, reservations via chat

from - stanford CS224S

Other speech related tasks

- **wake word** - to detect a word or short phrase, usually in order to wake up a voice-enable assistant
- **speaker diarization** - determining ‘who spoke when’ in a long speaker diarization multi-speaker audio recording
- **speaker recognition** - task of identifying the speaker
- **language identification** - identify which language is being spoken

Next steps

- Next week: 2 short (10-20min presentations + discussion
  - Presentation 1: Intro to Audio (Omar Sanseviero)
  - Presentation 2: ASR Deep Dive (Vaibhav Srivastav)
- Recommended resources
  - Intro to Audio for FastAI sections 1 and 2
  - SLP Chapter 26.1-26.5

Thanks for tuning in!