Lecture 1: Course Introduction
Week 1

- Course introduction
- Course Logistics
- Course topics overview
  - Dialog / conversational agents
  - Speech recognition (Speech to text)
  - Speech synthesis (Text to speech)
  - Applications
- Brief history
- Articulatory Phonetics
- ARPAbet transcription
Exciting recent developments have disrupted this field

Amazon Alexa + Alexa Prize 2014

Neural TTS voice cloning 2017

End-to-end neural becomes SOTA 2015 - present

Apple Siri 2011

Google Assistant 2016

Microsoft Cortana 2014

Realtime speech-speech translation 2020
Fraudsters Cloned Company Director’s Voice In $35 Million Bank Heist, Police Find
Deepfake Zelenskyy surrender video is the 'first intentionally used' in Ukraine war

By Matthew Holroyd & Fola Olorunselu • Updated: 16/03/2022

The deepfake video was shared during a hack on Ukrainian television.
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 create a speech synthesizer / voice clone of someone without their consent
  - It might be fun but it’s a little creepy. People get upset
  - Okay to use existing speech datasets (we’ll provide some)

- Consider subgroup and language bias when building real applications
  - Poor performance on subgroups e.g. non-native speakers
  - Many languages are under-served relative to English/Mandarin
Course Logistics

- **Course goal:** *Build something you are proud of*
  - Course project: Research paper? Compelling demo/story for job interviews? Applied system you can use at home/work?

- **Homeworks (2 weeks each):**
  - Introduction to audio analysis and spoken language tools
  - Building a complete dialog system using Amazon Alexa Skills Kit
  - Implementing end-to-end deep neural network approaches to speech recognition using PyTorch
  - Working with advanced deep learning toolkits for speech recognition (SpeechBrain) and voice cloning

- **Homeworks use Colab and PyTorch (AWS for Alexa)**
Course Logistics

- http://www.stanford.edu/class/cs224s
- Homeworks out on Tuesdays and due 11:59pm Monday
- Gradescope for homework submission
- Ed for questions. Use private post for personal/confidential questions
- Final project poster session in person!
Admin: Requirements and Grading

- **Readings:**
  - Jurafsky & Martin. Speech and Language Processing.
    - 3rd edition pre-prints available online
  - A few conference and journal papers

- **Grading**
  - Homework: 45%
  - Course Project: 50%
  - Participation: 5%
    - Attend 3 guest lectures (3%)
    - Ed participation (2%)
Course Projects

- **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

- Project handout + intro lecture / discussion soon. Ideally groups of 2-3
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 intro lecture will adjust to class needs.
- Proficiency in Python
  - Programming heavy homeworks will use Python, Colab Notebooks, and PyTorch
Office hours and CAs

- Andrew: In person after class on Thursdays (projects + other)
- CAs: Zoom with Calendly (homework + projects)
- Meet your teaching staff!
  - Gaurab Banerjee
  - Shreya Gupta
  - Alex Ke
- Questions on logistics?
Week 1

- Course introduction
- Course Logistics

**Course topics overview**

- Dialog / conversational agents
- Speech recognition (Speech to text)
- Speech synthesis (Text to speech)
- Applications

- Brief history
- Articulatory Phonetics
- ARPAbet transcription
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
Dialogue (= Conversational Agents)

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

- **POMDP**
  - Partially-Observed Markov Decision Processes
  - Reinforcement Learning to learn what action to take
  - Asking a question or answering one are just actions
    - “Speech acts”

- **Simple slot filling (ML or regular expressions)**
  - Pre-built frames
    - Calendar
      - Who
      - When
      - Where
  - Filled by hand-built rules
    - (“on (Mon | Tue | Wed...”)”)
Paradigms for Dialogue

- **POMDP**
  - Active research area. Deep learning RL
  - Not quite industry-strength

- **Simple slot filling (ML or regex)**
  - State of the art used most systems

- **Reusing new search engine technology**
  - Intent recognition / semantic parsing

- **Neural network chatbots**
  - Replacing major pieces of dialog systems
Speech Recognition

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

<table>
<thead>
<tr>
<th>English Tasks</th>
<th>WER%</th>
</tr>
</thead>
<tbody>
<tr>
<td>LibriSpeech audiobooks 960 hour clean</td>
<td>1.4</td>
</tr>
<tr>
<td>LibriSpeech audiobooks 960 hour other</td>
<td>2.6</td>
</tr>
<tr>
<td>Switchboard telephone conversations between strangers</td>
<td>5.8</td>
</tr>
<tr>
<td>CALLHOME telephone conversations between family</td>
<td>11.0</td>
</tr>
<tr>
<td>Sociolinguistic interviews, CORAAL (AAVE)</td>
<td>27.0</td>
</tr>
<tr>
<td>CHiMe5 dinner parties with body-worn microphones</td>
<td>47.9</td>
</tr>
<tr>
<td>CHiMe5 dinner parties with distant microphones</td>
<td>81.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chinese (Mandarin) Tasks</th>
<th>CER%</th>
</tr>
</thead>
<tbody>
<tr>
<td>AISHELL-1 Mandarin read speech corpus</td>
<td>6.7</td>
</tr>
<tr>
<td>HKUST Mandarin Chinese telephone conversations</td>
<td>23.5</td>
</tr>
</tbody>
</table>

**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.
Why is conversational speech harder?

- A piece of an utterance without context
- The same utterance with more context
### HSR versus ASR

<table>
<thead>
<tr>
<th>Deletions</th>
<th>Insertions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SWB</strong></td>
<td><strong>CH</strong></td>
</tr>
<tr>
<td>ASR</td>
<td>Human</td>
</tr>
<tr>
<td>20: i</td>
<td>17: it</td>
</tr>
<tr>
<td>17: that</td>
<td>16: and</td>
</tr>
<tr>
<td>16: a</td>
<td>14: that</td>
</tr>
<tr>
<td>14: you</td>
<td>12: the</td>
</tr>
<tr>
<td>12: %bcack</td>
<td>11: a</td>
</tr>
<tr>
<td>11: to</td>
<td>9: have</td>
</tr>
</tbody>
</table>

Table 3: Most frequent deletion and insertion errors for humans and ASR system on SWB and CH.

<table>
<thead>
<tr>
<th>Substitution Errors</th>
<th>SWB</th>
<th>CH</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ASR</strong></td>
<td>Human</td>
<td>ASR</td>
</tr>
<tr>
<td>11: and / in</td>
<td>16: (%hes) / oh</td>
<td>21: was / is</td>
</tr>
<tr>
<td>9: was / is</td>
<td>12: was / is</td>
<td>16: him / them</td>
</tr>
<tr>
<td>7: it / that</td>
<td>7: (i-) / %hes</td>
<td>15: in / and</td>
</tr>
<tr>
<td>6: (%hes) / oh</td>
<td>5: (%hes) / a</td>
<td>8: a / the</td>
</tr>
<tr>
<td>6: him / them</td>
<td>5: (%hes) / hmmm</td>
<td>8: and / in</td>
</tr>
<tr>
<td>6: too / to</td>
<td>5: (a-) / %hes</td>
<td>8: is / was</td>
</tr>
<tr>
<td>5: (%hes) / i</td>
<td>5: could / can</td>
<td>8: two / to</td>
</tr>
<tr>
<td>5: then / and</td>
<td>5: that / it</td>
<td>7: the / a</td>
</tr>
<tr>
<td>4: (%hes) / %bcack</td>
<td>4: %bcack / oh</td>
<td>7: too / to</td>
</tr>
<tr>
<td>4: (%hes) / am</td>
<td>4: and / in</td>
<td>6: (%hes) / a</td>
</tr>
</tbody>
</table>

Table 2: Most frequent substitution errors for humans and ASR system on SWB and CH.

(Saon et al, 2017)
Why accents are hard

- A word by itself

- The word in context
So is speech recognition solved? Why study it vs use some API?

- In the last ~10 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
Speech Recognition Design
Intuition

• Build a statistical model of the speech-to-words process
• Collect lots and lots of speech, and transcribe all the words.
• Train the model on the labeled speech
• Paradigm: Supervised Machine Learning + Search
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
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
Figure 1: 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.
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

- New area: Self-supervised foundation models
Extraction of Social Meaning from Speech

- Detection of student uncertainty in tutoring
  - Forbes-Riley et al. (2008)
- Emotion detection (annoyance)
  - Ang et al. (2002)
- Detection of deception
  - Newman et al. (2003)
- Detection of charisma
  - Rosenberg and Hirschberg (2005)
- Speaker stress, trauma
Conversational style

- 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:
Week 1

- Course introduction
- Course Logistics
- Course topics overview
  - Dialog / conversational agents
  - Speech recognition (Speech to text)
  - Speech synthesis (Text to speech)
  - Applications
- Brief history
- Articulatory Phonetics
- ARPAbet transcription