CS 224S / LINGUIST 285
Spoken Language Processing

Andrew Maas
Stanford University
Spring 2022

Lecture 1: Course Introduction

Original slides by Dan Jurafsky

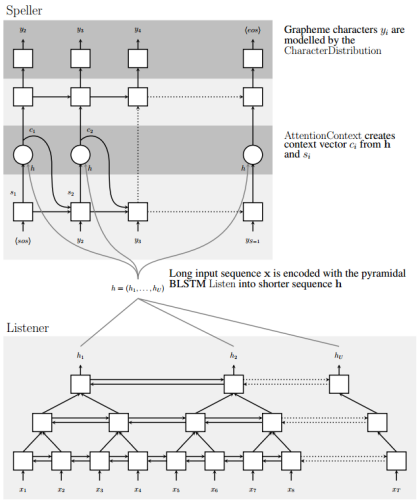
Week 1

- 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

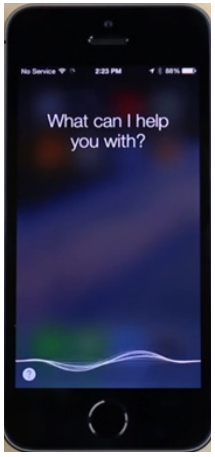
Exciting recent developments have disrupted this field

Amazon Alexa + Alexa Prize
2014

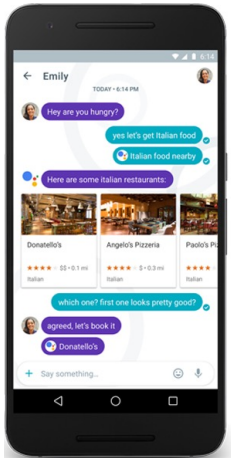
Neural TTS voice cloning
2017



End-to-end neural becomes SOTA
2015 - present

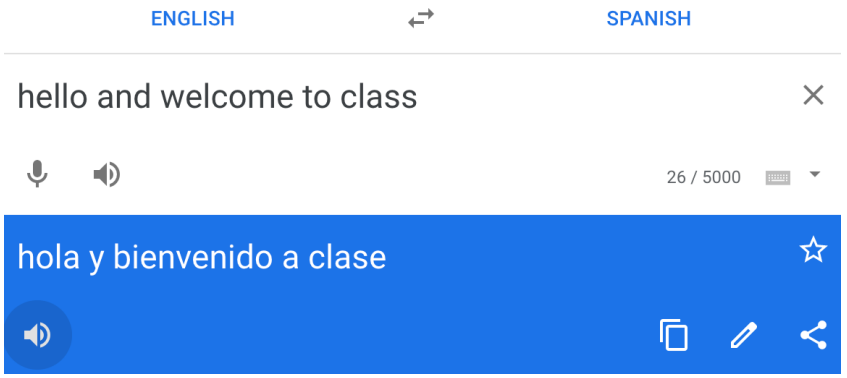


Apple
Siri
2011



Google
Assistant
2016

Microsoft
Cortana
2014



Realtime speech-speech translation
2020

Entering a new era of spoken language applications and impact

EDITORS' PICK | Oct 14, 2021, 07:01am EDT | 86,092 views

Fraudsters Cloned Company Director's Voice In $35 Million Bank Heist, Police Find

Thomas Brewster Forbes Staff

Cybersecurity

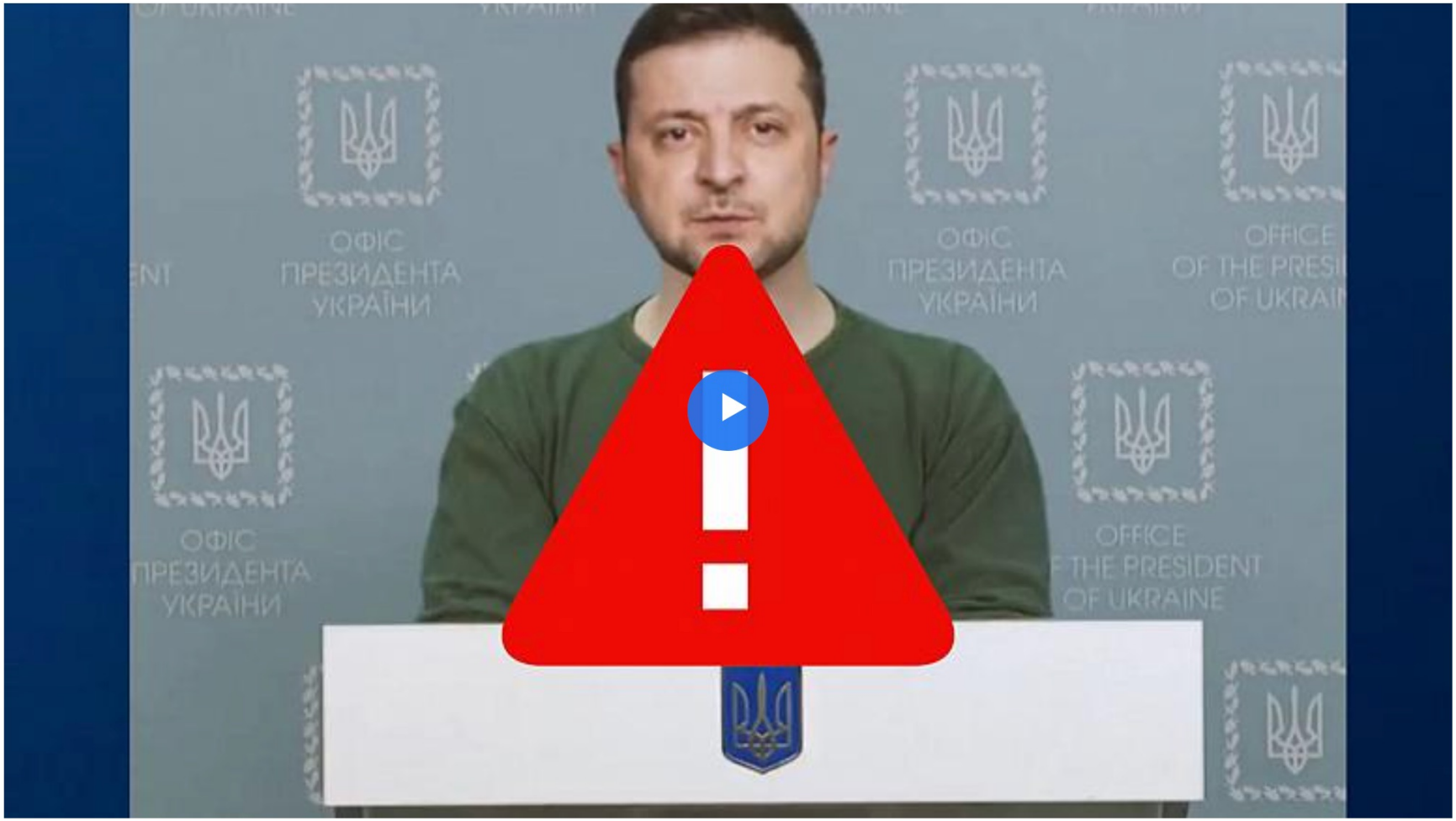
Associate editor at Forbes, covering cybercrime, privacy, security and surveillance.

EuroNews Article

Deepfake Zelenskyy surrender video is the 'first intentionally used' in Ukraine war

COMMENTS

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



The deepfake video was shared during a hack on Ukrainian television. - Copyright Euronews via Twitter

EuroNews Article
Youtube video

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)

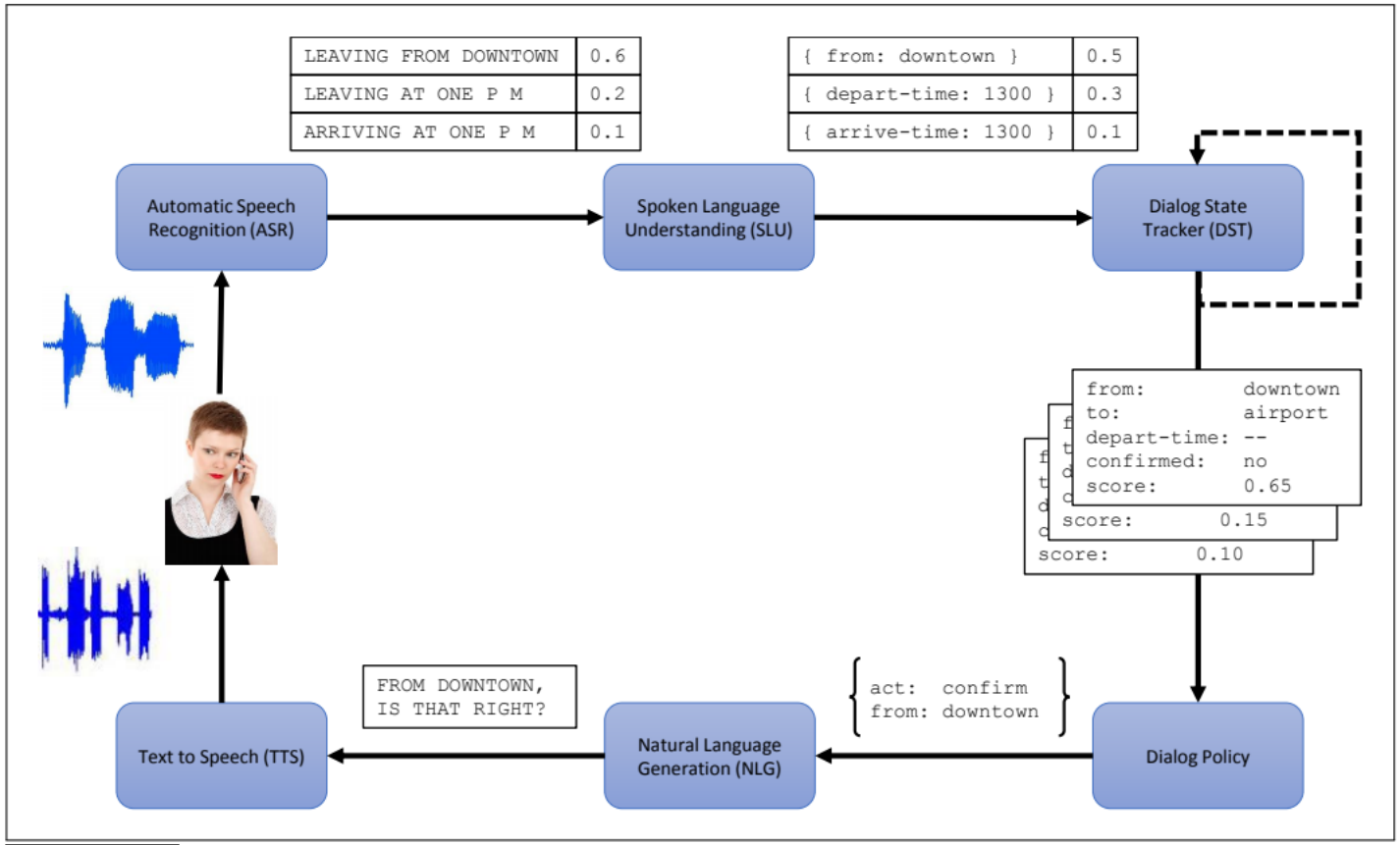


Figure 26.11 Architecture of a dialogue-state system for task-oriented dialogue from Williams et al. (2016).

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

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.

Why is conversational speech harder?

- A piece of an utterance without context

- The same utterance with more context

HSR versus ASR

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

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 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

TTS Overview: End-to-end neural

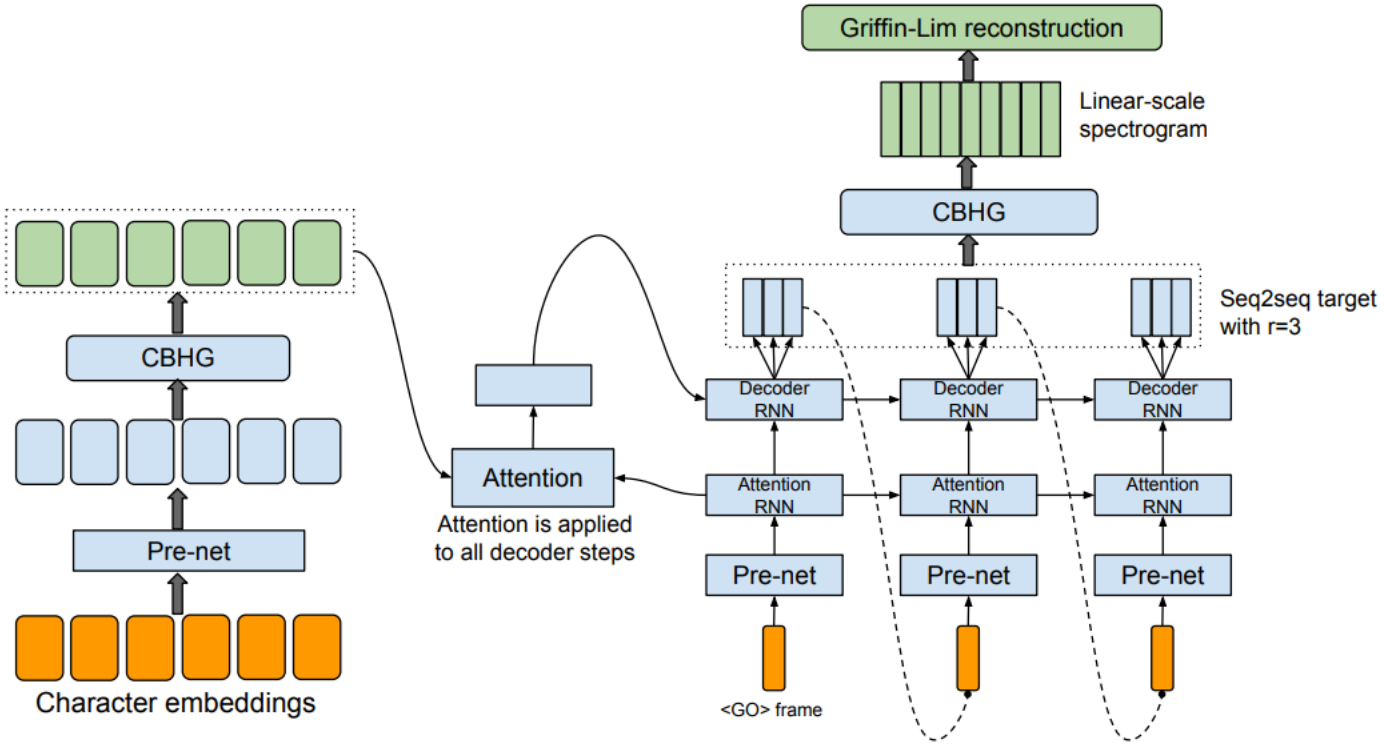


Figure 1: *Model architecture. The model takes characters as input and outputs the corresponding raw spectrogram, which is then fed to the Griffin-Lim reconstruction algorithm to synthesize speech.*

Tacotron. (Wang et al. 2017)

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
  - Rude et al. (2004), Pennebaker and Lay (2002)

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*