CS 224S / Linguist 285
Spoken Language Processing

Andrew Maas | Stanford University | Spring 2024

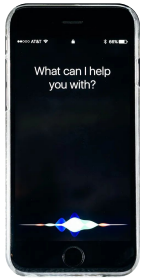
Lecture 1: Course Introduction

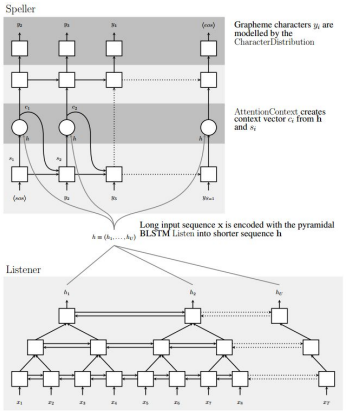
Outline

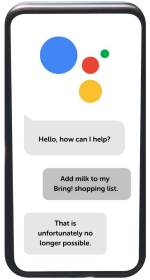
- **Course Introduction**
- **Course Logistics**
- **Course Topics Overview**
  - Dialogue / Conversational Agents
  - Speech Recognition (Speech to Text)
  - Speech Synthesis (Text to Speech)
  - Applications

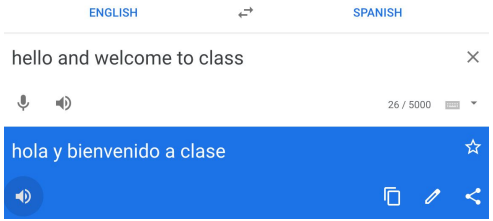
Course Introduction

Exciting recent developments have disrupted this field









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)

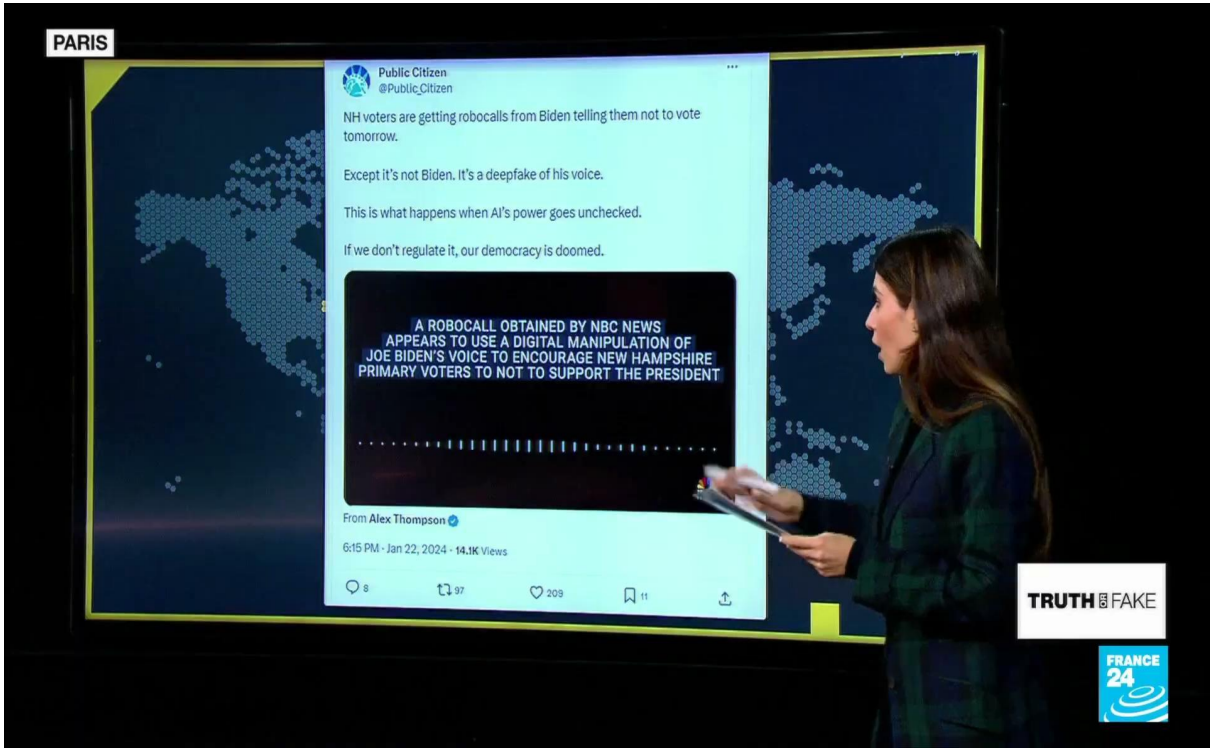
Verge Article

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.

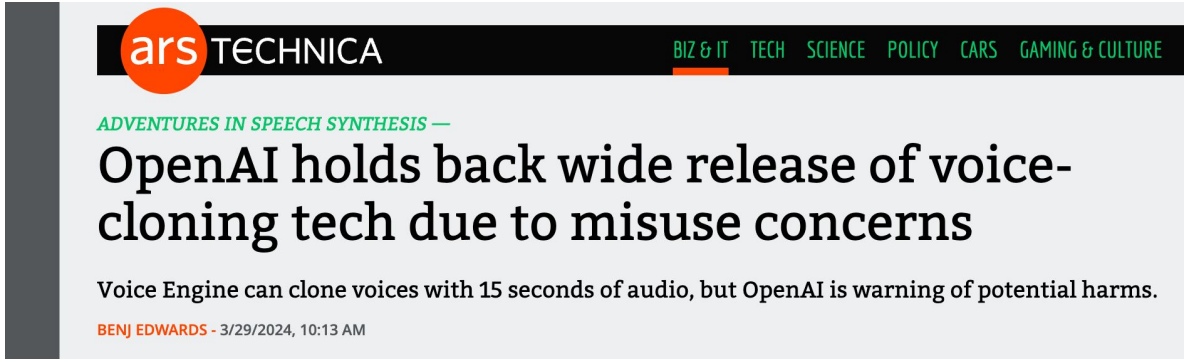


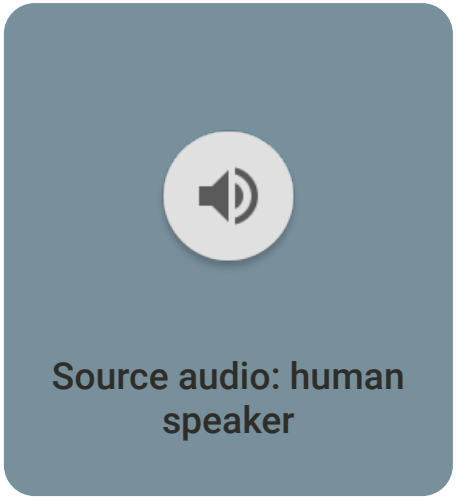
Wired Article

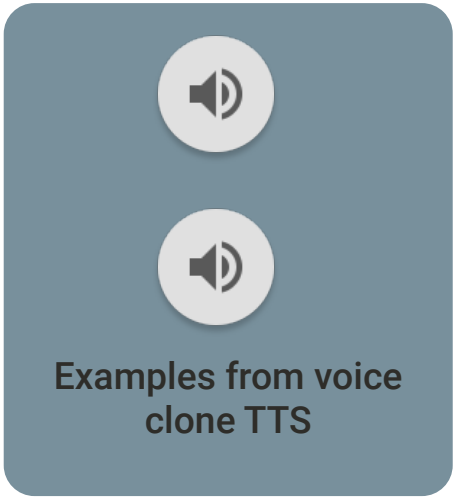
Discussion:

OpenAI just announced a voice cloning tool.

Would you approve releasing this tool publicly?







[The Verge article with samples]

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).

Do consider subgroup and language bias when building systems

Poor performance on subgroups e.g. non-native speakers

Many languages are under-served relative to English/Mandarin.

Course Logistics

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

Learning goals for this course

- **You will develop expertise on 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?

What you will build in this course

- Course project. **We allow different types of project goals:**
  - Research paper: Novel, focused research contribution using or improving speech technology
  - Product/demo development. Create a working system with available speech tools
  - Add spoken language elements to your existing research or product development
- **Three Homeworks:**
  - Introduction to audio analysis and speech synthesis tools
  - Working with speech recognition toolkits and APIs
  - Leveraging audio foundation models and working with non-English speech tasks
- **Homeworks use Colab and PyTorch**

Course Logistics Overview

- http://www.stanford.edu/class/cs224s
- **Homeworks**
  - First homework out today.
  - Approximately 2 weeks for each homework
- **Projects**
  - Written proposal, written milestone, final poster session & final paper
  - In-class project check-in presentations (~2 minutes per group)
- **Gradescope for homework submission**
- **Ed for questions. Use private post for personal/confidential questions**

Requirements and Grading

- **Readings:**
  - Jurafsky & Martin. Speech and Language Processing.
    - 3rd edition pre-prints available online
  - A few conference and journal papers
- **Grading:**
  - Homework: 35%
  - Course Project: 60%
  - Participation: 5%
    - Attend 6 guest lectures (3%)
    - Ed participation (2%)

Course Projects

- 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
- Project handout posted by April 8.
- Start thinking of groups and topics now. Ideally groups of 2-3. We will circulate some suggested projects from people around campus and industry.

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 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

- Andrew: In person after class on Wednesdays (projects + other)
- CAs: Office hour times TBD (homework + projects)
- Meet your teaching staff!
  - Head CA: Tolúlọpẹ́ Ogunremi
  - Abhinav Garg
  - Fahad Nabi
  - Gautham Raghupathi
- Questions on logistics?

Course Topics Overview

- Dialogue / Conversational Agents
- Speech Recognition (Speech to Text)
- Speech Synthesis (Text to Speech)
- Applications

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)

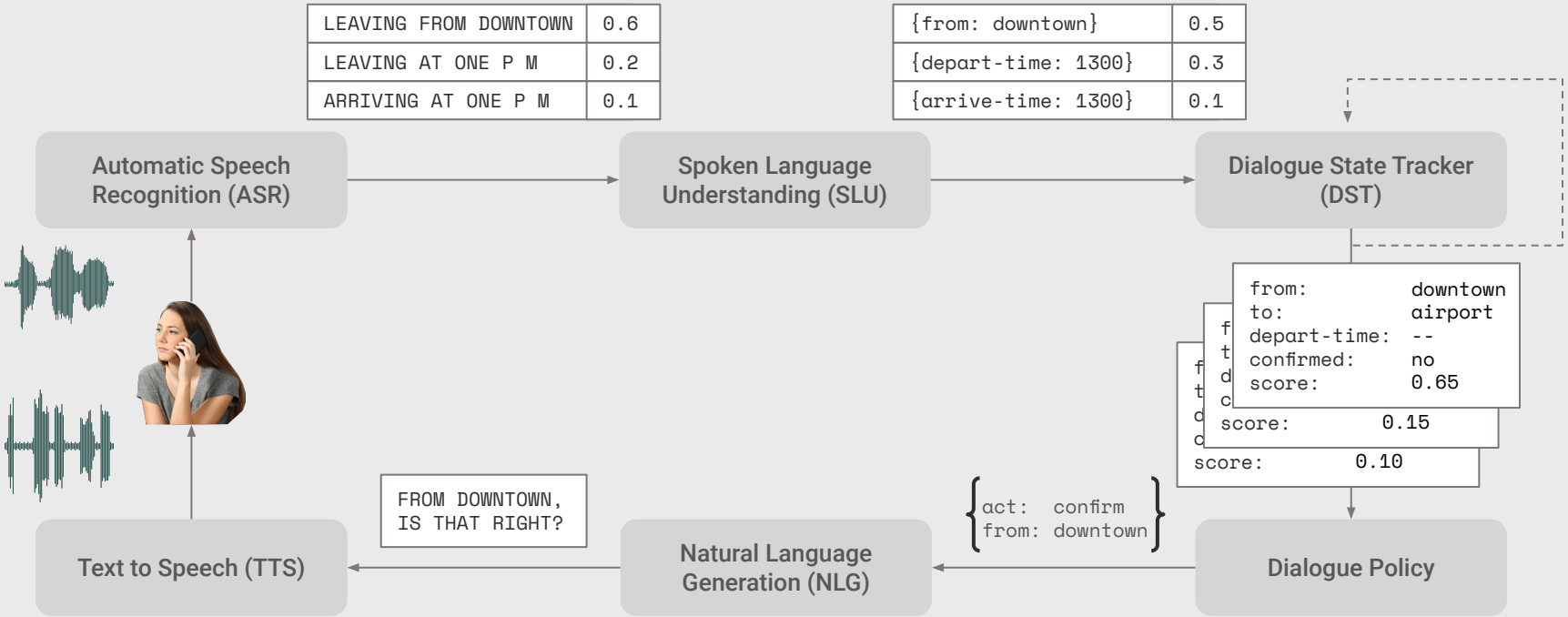


Figure: Architecture of dialogue-state system for task-oriented dialogue (William et al, 2016)

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**
  - Deep learning RL
  - Not quite industry-strength historically, but now combining with LLM-based approaches
- **Simple slot filling (ML or regex)**
  - State of the art used most systems
- **Reusing new search engine technology**
  - Intent recognition / semantic parsing
- **Neural network (LLM) chatbots and dialog systems**
  - Great for ungrounded chat (e.g. ChatGPT)
  - Active R&D on *task-oriented dialog* and chat-based agents that can take concrete actions

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 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

Why is Conversational Speech Harder?

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

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

- **In the last ~12 years**
  - Dramatic reduction in LVCSR error rates (16% to 3%) in ideal conditions
  - Human level LVCSR performance on Switchboard
  - 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**

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
- New variation: Voice cloning. TTS systems that mimic particular speakers with minimal training data

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 fully emotive
- No longer require extensive data from a single speaker to make a TTS voice for them

TTS Overview End-to-end Neural

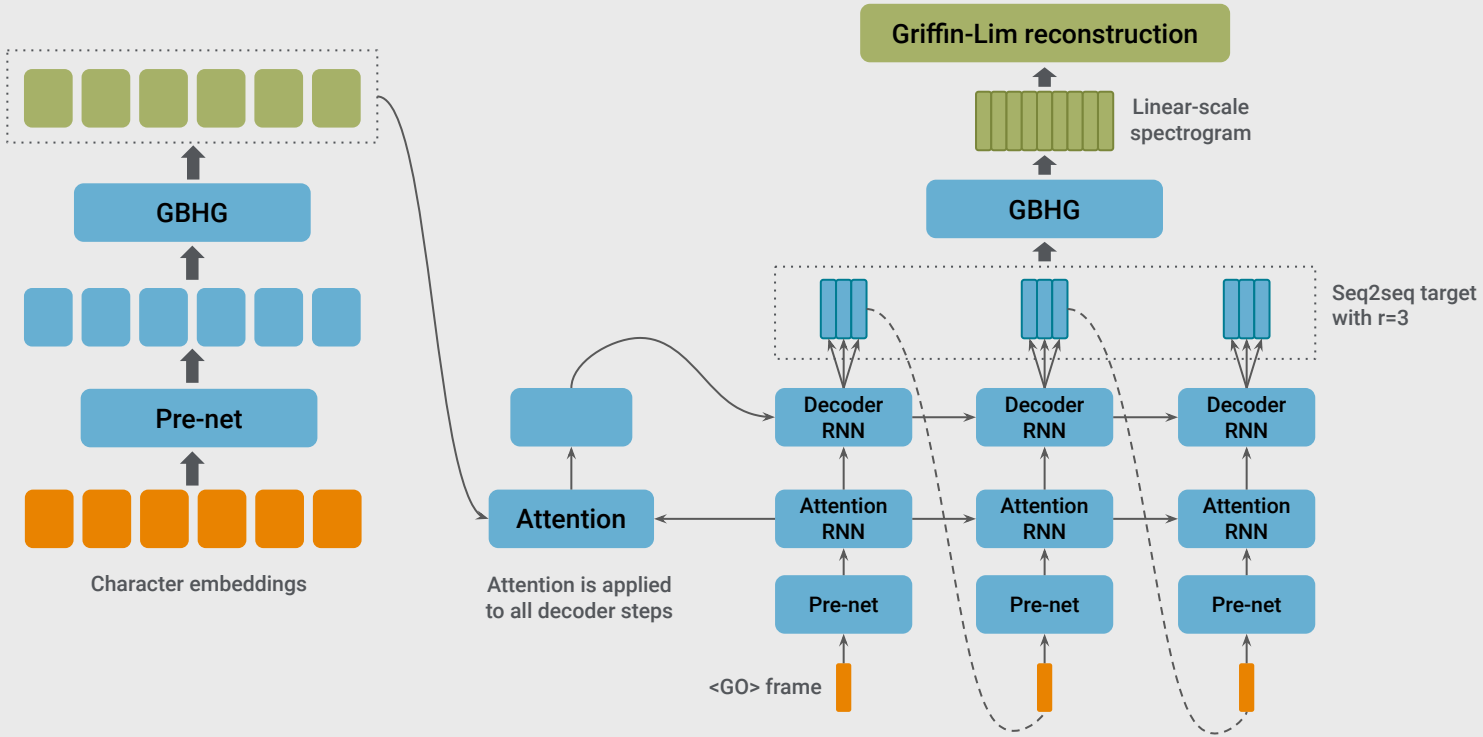


Figure 3: Model Architecture. The model takes characters as input and outputs the corresponding raw spectrogram, which is then fed to the Griffin-Lim reconstruction algorithm to synthesize speech. Tacotron (Wang et al, 2017)

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

- 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

Thank You