CAS CS 640
Artificial Intelligence
Lecture by Margrit Betke

Automated Speech Recognition & Voice Cloning
Learning Outcomes:
Being able to

- Define speech recognition, phoneme, wake word detection, mel scale, spectrogram, encoder, decoder, Short-Time Fourier Transform, voice cloning
- Discuss sources of variability of an acoustic signal and constraints on how a phoneme is realized acoustically
- Explain parsing as a tree search
- Explain the difference between speaker dependent and independent speech recognition
- Explain how HMMs were/are used in speech recognition
- Explain the choice of the wake word and how it can be detected
- Give criteria for evaluation of speech recognition and voice cloning
- Describe the LAS model
- Explain how a language model can be added to a encoder/decoder speech recognition model
- Discuss the state of the art in speech recognition in 2023 (USM)
- Explain a voice cloning model and its connection to the task of speaker identification
- Explain the dangers of voice cloning
- Discuss how to detect voice clones
What is Speech Recognition?

- Speech recognition is the task of transforming an acoustic signal of a speaker talking in a natural language (such as English) into text in that language.

- **words** = a string of words in a given natural language and **signal** = a sequence of observed acoustic data that has been digitized and pre-processed.

- Find the **words** that maximize the probability
  
  \[ P(\text{words} \mid \text{signal}) : \arg\max_{\text{words}} P(\text{words} \mid \text{signal}) \]

- Bayes rule: \( \arg\max_{\text{words}} P(\text{signal} \mid \text{words}) P(\text{words}) \), where \( P(\text{words}) \) represents our **language model** = prior probability of a particular word string and likelihood \( P(\text{signal} \mid \text{words}) = \text{acoustic model} \) (difficult to specify due to high variability of acoustic signal).
Sources of Variability of Acoustic Signal

- Acoustic Variations:
  - Background speech from radio, office mates, TV
  - Background noise at airports, in cars, at home
  - Quality of microphone
  - Position of microphone
Sources of Variability of Acoustic Signal

- **Intra-speaker Variations:**
  - Speaker’s physiological state
    - person may have a cold, may be tired
  - Speaker’s psychological state
    - person may be excited, sad, nervous

Influence speaking rates & style

e.g., voice fillers like “ah”
Sources of Variability of Acoustic Signal

- **Inter-speaker Variations:**
  - Male/female
  - Every voice is unique due to
    - different size and shape of vocal tract
    - speaker’s background (dialect, accent)

- **Coarticulation:**
  Spectral characteristics of a spoken word vary depending on what words surround it
**Phonemes**

**Definition:**
basic distinctive units of speech sound by which words and sentences are represented
different for each language

Example of CMU’s 36 phoneme set for English

phonetic segment = phoneme
phone = smallest perceptible segment
Acoustic Realization of Phonemes Depends on

- **Structural constraints of a language:**
  - limited number of sounds
    - e.g. in English: 60 consonants/consonant clusters can start a word
    - 16 acoustically different vowels

- **Intrinsic characteristics:**
  - Voiced: vocal folds in larynx vibrate by airflow
  - Unvoiced: turbulence in vocal tract
    - e.g. in English: “z” (zoo) and “s” (sing)

- **Coarticulation:**
  - Phoneme /t/ “tea” “tree” “steep” butter” all different
  - Phoneme /s/ “gas station” often deleted
Alexa’s Phonemes?

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Example</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>odd</td>
<td>AA D</td>
</tr>
<tr>
<td>AE</td>
<td>at</td>
<td>AE T</td>
</tr>
<tr>
<td>AH</td>
<td>hut</td>
<td>HH AH T</td>
</tr>
<tr>
<td>AO</td>
<td>ought</td>
<td>AO T</td>
</tr>
<tr>
<td>AW</td>
<td>cow</td>
<td>K AW</td>
</tr>
<tr>
<td>AY</td>
<td>hide</td>
<td>HH AY D</td>
</tr>
<tr>
<td>B</td>
<td>be</td>
<td>B IY</td>
</tr>
<tr>
<td>CH</td>
<td>cheese</td>
<td>CH IY Z</td>
</tr>
<tr>
<td>D</td>
<td>dee</td>
<td>D IY</td>
</tr>
<tr>
<td>DH</td>
<td>thee</td>
<td>DH IY</td>
</tr>
<tr>
<td>EH</td>
<td>Ed</td>
<td>EH D</td>
</tr>
<tr>
<td>ER</td>
<td>hurt</td>
<td>HH ER T</td>
</tr>
<tr>
<td>EY</td>
<td>ate</td>
<td>EY T</td>
</tr>
<tr>
<td>F</td>
<td>fee</td>
<td>F IY</td>
</tr>
<tr>
<td>G</td>
<td>green</td>
<td>GR IY N</td>
</tr>
<tr>
<td>HH</td>
<td>he</td>
<td>HH IY</td>
</tr>
<tr>
<td>IH</td>
<td>it</td>
<td>IH T</td>
</tr>
<tr>
<td>IY</td>
<td>eat</td>
<td>IY T</td>
</tr>
<tr>
<td>JH</td>
<td>gee</td>
<td>JH IY</td>
</tr>
<tr>
<td>K</td>
<td>key</td>
<td>K IY</td>
</tr>
<tr>
<td>L</td>
<td>lee</td>
<td>L IY</td>
</tr>
<tr>
<td>M</td>
<td>me</td>
<td>M IY</td>
</tr>
<tr>
<td>N</td>
<td>knee</td>
<td>N IY</td>
</tr>
<tr>
<td>NG</td>
<td>ping</td>
<td>P IH NG</td>
</tr>
<tr>
<td>OW</td>
<td>oat</td>
<td>OW T</td>
</tr>
<tr>
<td>OY</td>
<td>toy</td>
<td>T OY</td>
</tr>
<tr>
<td>P</td>
<td>pee</td>
<td>P IY</td>
</tr>
<tr>
<td>R</td>
<td>read</td>
<td>R IY D</td>
</tr>
<tr>
<td>S</td>
<td>sea</td>
<td>S IY</td>
</tr>
<tr>
<td>SH</td>
<td>she</td>
<td>SH IY</td>
</tr>
<tr>
<td>T</td>
<td>tea</td>
<td>T IY</td>
</tr>
<tr>
<td>TH</td>
<td>theta</td>
<td>TH EY T AH</td>
</tr>
<tr>
<td>UH</td>
<td>hood</td>
<td>HH UH D</td>
</tr>
<tr>
<td>UW</td>
<td>two</td>
<td>T UW</td>
</tr>
<tr>
<td>V</td>
<td>vee</td>
<td>V IY</td>
</tr>
<tr>
<td>W</td>
<td>we</td>
<td>W IY</td>
</tr>
<tr>
<td>Y</td>
<td>yield</td>
<td>Y IY L D</td>
</tr>
<tr>
<td>Z</td>
<td>zee</td>
<td>Z IY</td>
</tr>
<tr>
<td>ZH</td>
<td>seizure</td>
<td>S IY ZH ER</td>
</tr>
</tbody>
</table>

phonetic segment = phoneme

phone = smallest perceptible segment
### Alexa’s Phonemes?

**Alexa**

AH-L-EH-K-S-AH

Rare combination of phonemes (sounds) in English

→ Alexa is a smart “wake up word”

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Example</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>odd</td>
<td>AA D</td>
</tr>
<tr>
<td>AE</td>
<td>at</td>
<td>AE T</td>
</tr>
<tr>
<td>AH</td>
<td>hut</td>
<td>HH AH T</td>
</tr>
<tr>
<td>AO</td>
<td>ought</td>
<td>AO T</td>
</tr>
<tr>
<td>AW</td>
<td>cow</td>
<td>K AW</td>
</tr>
<tr>
<td>AY</td>
<td>hide</td>
<td>HH AY D</td>
</tr>
<tr>
<td>B</td>
<td>be</td>
<td>B IY</td>
</tr>
<tr>
<td>CH</td>
<td>cheese</td>
<td>CH IY Z</td>
</tr>
<tr>
<td>D</td>
<td>dee</td>
<td>D IY</td>
</tr>
<tr>
<td>DH</td>
<td>thee</td>
<td>DH IY</td>
</tr>
<tr>
<td>EH</td>
<td>Ed</td>
<td>EH D</td>
</tr>
<tr>
<td>ER</td>
<td>hurt</td>
<td>HH ER T</td>
</tr>
<tr>
<td>EY</td>
<td>ate</td>
<td>EY T</td>
</tr>
<tr>
<td>F</td>
<td>fee</td>
<td>F IY</td>
</tr>
<tr>
<td>G</td>
<td>green</td>
<td>G R IY N</td>
</tr>
<tr>
<td>HH</td>
<td>he</td>
<td>HH IY</td>
</tr>
<tr>
<td>IH</td>
<td>it</td>
<td>IH T</td>
</tr>
<tr>
<td>IY</td>
<td>eat</td>
<td>IY T</td>
</tr>
<tr>
<td>JH</td>
<td>gee</td>
<td>JH IY</td>
</tr>
<tr>
<td>K</td>
<td>key</td>
<td>K IY</td>
</tr>
<tr>
<td>L</td>
<td>lee</td>
<td>L IY</td>
</tr>
<tr>
<td>M</td>
<td>me</td>
<td>M IY</td>
</tr>
<tr>
<td>N</td>
<td>knee</td>
<td>N IY</td>
</tr>
<tr>
<td>NG</td>
<td>ping</td>
<td>P IH NG</td>
</tr>
<tr>
<td>OW</td>
<td>oat</td>
<td>OW T</td>
</tr>
<tr>
<td>OY</td>
<td>toy</td>
<td>TOY</td>
</tr>
<tr>
<td>P</td>
<td>pee</td>
<td>P IY</td>
</tr>
<tr>
<td>R</td>
<td>read</td>
<td>R IY D</td>
</tr>
<tr>
<td>S</td>
<td>sea</td>
<td>S IY</td>
</tr>
<tr>
<td>SH</td>
<td>she</td>
<td>SH IY</td>
</tr>
<tr>
<td>T</td>
<td>tea</td>
<td>T IY</td>
</tr>
<tr>
<td>TH</td>
<td>theta</td>
<td>TH EY T AH</td>
</tr>
<tr>
<td>UH</td>
<td>hood</td>
<td>HH UH D</td>
</tr>
<tr>
<td>UW</td>
<td>two</td>
<td>T UW</td>
</tr>
<tr>
<td>V</td>
<td>vee</td>
<td>V IY</td>
</tr>
<tr>
<td>W</td>
<td>we</td>
<td>W IY</td>
</tr>
<tr>
<td>Y</td>
<td>yield</td>
<td>Y IY L D</td>
</tr>
<tr>
<td>Z</td>
<td>zee</td>
<td>Z IY</td>
</tr>
<tr>
<td>ZH</td>
<td>seizure</td>
<td>S IY ZH ER</td>
</tr>
</tbody>
</table>

**phonetic segment** = **phoneme**

**phone** = smallest perceptible segment
Acoustic Realization of Phonemes
Depends on

- **Impact of prosodics:**
  - Fluctuation of stress and intonation

- **Syntax:**
  - Grammar constraints the number of possible sentences
  - Phonemes often lengthened before boundaries

- **Semantics:**
  - Constraints on number of sentences:
    - Unlikely speech: “The snow was loud”
Problem: Ambiguities

Why are these funny?

Headlines:
- Enraged Cow Injures Farmer With Ax
- Hospitals Are Sued by 7 Foot Doctors
- Ban on Nude Dancing on Governor’s Desk
- Iraqi Head Seeks Arms
- Local HS Dropouts Cut in Half
- Juvenile Court to Try Shooting Defendant
- Stolen Painting Found by Tree
- Kids Make Nutritious Snacks
The company Hershey forbids protest.

Chocolate bars are protesting.
Probabilistic Context-Free Grammars

- Natural language grammars are very ambiguous!
- PCFGs are a formal probabilistic model of trees
  - Each “rule” has a conditional probability (like an HMM)
  - Tree’s probability is the product of all rules used
- Parsing: Given a sentence, find the best tree – search!

NP: Noun phrase
VP: Verb phrase
NN: Noun singular
VBZ: Verb 3rd person singular present

Material from D. Klein, P. Abbeel, UC Berkeley

https://parser.kitaev.io/
Probabilistic Context-Free Grammars

- Natural language grammars are very ambiguous!
- PCFGs are a formal probabilistic model of trees
  - Each “rule” has a conditional probability (like an HMM)
  - Tree’s probability is the product of all rules used
- Parsing: Given a sentence, find the best tree – search!
Early Ideas for Automated Speech Recognition (1970s)

- IBM’s “tri-gram model”
- CMU’s Hearsay I played voice chess
  - top-down, expectation-driven approach
- CMU’s Harpy
  - sentence = path through network represents sequence of sounds

\[ \text{max } P(\text{Word3} | \text{Word1} \& \text{Word2}) \]
Speaker-dependent Speech Recognition (1980s and 1990s)

- **Isolated Word Recognition**
  - Words: 10 ms
  - Pauses: 200 ms
  - Speech signal = sequence of spectra matched with stored templates of words of vocabulary

- **Connected Word Recognition**
  - Challenge: Acoustic signal altered at word boundaries

- **Fluent Speech Systems**
  - First commercial successes: Dragon Dictate (out of CMU), IBM
  - Used heavily for dictation by lawyers and doctors, for example, radiology reports

Speakers needed to train systems carefully
Ability to define “macros”
HMMs in Speech Recognition

Constructing left-to-right HMM for word sequences:
Concatenate HMMs (with non-emitting end states) for each word in sentence:

HMM for Word1:

HMM for Word2:

Combined HMM for sequence Word1 Word2:
HMMs in Speech Recognition

HMMs representing words are themselves constructed by concatenating phonemes

Advantage of this approach:

- Fewer phonemes than words (e.g. 36 versus tens of thousands)
- Phonemes occur more frequently in training data than words: often difficult to find a sufficient number of examples per word in training data, even if data set is large
- Words that were never seen in the training data can be constructed from phoneme HMMs and recognized
Generic Fluent-Speech Recognition System

HMMs

Input Speech

Feature Analysis

Word Match

Sentence Match

Recognized Sentence

Acoustic Word Models

Language Model

Unit Model

Lexicon

Syntax

Semantics

Verification

Rabiner 1997
HUGE Models Are Used

1,500 word Air Travel Information System

Graph represents utterance of the sentence
“Show me the flights from Charlotte to Minneapolis on Monday”

Mohri, Riley, 1999

Graph with 151 million paths
# Performance of Speech Recognition Systems

<table>
<thead>
<tr>
<th>Task</th>
<th>Vocabulary Size</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digits 0-10</td>
<td>11</td>
<td>0.3% per digit</td>
</tr>
<tr>
<td>Airline travel info</td>
<td>2,500 words</td>
<td>2% per word</td>
</tr>
<tr>
<td>Reading newspaper</td>
<td>64,000 words</td>
<td>8% per word</td>
</tr>
<tr>
<td>Radio</td>
<td>64,000 words</td>
<td>27% per word</td>
</tr>
<tr>
<td>Conversation over phone</td>
<td>28,000 words</td>
<td>37% per word</td>
</tr>
</tbody>
</table>

Rabiner, 1997
Automated Speech Recognition in the Telecommunications Industry

- Automation of operator services:
  Collect calls, 3rd-party billing, calling cards, automated acceptance/rejection of reverse calls
- Automation of directory assistance:
  Front-end city name recognition (general)
  Recognition of employee name (corporate environment)
- Voice dialing:
  spoken commands such as “call home,” “call office”

Rabiner, 1997
Automated Speech Recognition Provided by the Telecommunications Industry

- **Voice banking services:**
  Access to customer accounts, balances, transactions
  First created in Japan by NTT

- **Interactive voice response systems:**
  Speak touch-tone position (AT&T introduced it first in Spain)

- **Directory assistance call completion:**
  Interface speech recognition system with speech synthesis system that dials for user (due to fragmentation of industry)

- **Reverse directory assistance:**
  Speak telephone number, receive address (NYNEX, Bellcore)

- **Information services:**
  Access to scores of sporting events, traffic reports, theater reservations

Rabiner, 1997
Speech Recognition Technology in last decade+

- User-specific fluent speech systems – 99% accurate
  e.g., Dragon Naturally Speaking
    - Medical 10.1 (80 medical specialties) $1,599
    - Legal 10 (30,000 legal terms) $1,199
    - Professional 10 $ 899

- Customer care
  Dialogue-type interaction, e.g. AT&T’s system: HowMayIHelpYou

- Google Voice: 2009
  e.g., 2011: voice transcription: Your voice mail is automatically converted into an email, available in US only

- Siri: Oct. 2011: intelligent personal assistant with Nuance speech recognition interface

- Google Now (2012), Facebook (Jan. 2015)
How Alexa Fits Into Amazon’s Prime Directive On Technology
By JENNA WORTHAM
JAN. 24, 2017

It took a team of 1,000 engineers to write its code, and when the device was finished, Amazon decided to call it Alexa, shorthand for Alexandria, as in the ancient Library of Alexandria in Egypt
Amazon Echo & Alexa

- **Price:** 1/24/2017: $179.99.
- **3rd Generation:** 12/10/2019: $79.99
- **4th Generation:** 11/17/2020: $99.99
- 11/2/2023: Echo Dot $49.99, Echo Studio $199.99
- **Release Date:** November 2014
- **Dimensions:** ~3”x3”x9” (8x8x24cm³)
- **Feature:** Bluetooth, Wireless, Smart Speaker
- **Supported Host Device OS:** iOS, Android
- **Initial Features:** Compatible with Belkin WeMo WiFi, compatible with Philips Hue smart lighting, built-in 7 microphones
Amazon Echo & Alexa in 2017

- Plays all your music from Amazon Music, Spotify, Pandora, iHeartRadio, TuneIn, and more using just your voice
- Fills the room with immersive, 360° omni-directional audio
- Allows hands-free convenience with voice-control
- Hears you from across the room with far-field voice recognition, even while music is playing
- Answers questions, reads audiobooks and the news, reports traffic and weather, gives info on local businesses, provides sports scores and schedules, and more using the Alexa Voice Service
- Controls lights, switches, and thermostats with compatible WeMo, Philips Hue, Samsung SmartThings, Wink, Insteon, Nest, and ecobee smart home devices
- Always getting smarter and adding new features, plus thousands of skills like Uber, Domino's, and more
Sources of Variability of Acoustic Signal

- Acoustic Variations:
  - Background speech from radio, office mates, TV
  - Background noise at airports, in cars, at home
  - Quality of microphone
  - Position of microphone

Amazon Echo is often placed in a cubby shelf instead of in the middle of the room, even if manufacturer recommends against it, causing reverberations making it difficult for Alexa to “wake up”
How do Amazon Echo and Alexa Work?

On device processing:
User: “Alexa, order flowers for my grandma”

Cloud Processing:
Alexa: “I have ordered flowers”

App Layer, Text to Speech

Signal Processing
beam-formed signal

Wake Word Detection

Natural Language Processing

Recognized Intent: BuyItem
ItemName: Flowers

Speech-to-text

Automatic Speech Recognition
Goal: High “positive” detection rate with no false positives

Challenges:
- Low signal-to-noise ratio, reverberation, competing speech, music playback
- Pronunciation differences
- Achieving high accuracy and low latency with limited on-device processing power

Solution: Classifiers trained on positive and negative samples of the wake word
Wake Word Model

Audio Stream → Acoustic Feature Extraction → Deep Neural Network & Hidden Markov Model → Support Vector Machine → Wake Word Accepted Yes/No
Wake Word DNN/HMM Model

Two finite state machines (FSMs):
1. Foreground wake word FSM
2. Background speech/non-speech FSM

Deep neural network (DNN) produces posterior probabilities $p(\text{state} | \text{acoustic features})$

Detection confidence is computed from foreground/background likelihood ratio
Automated Speech Recognition (ASR)

Acoustic signal

Deep Neural Network

Spectrogram

Image credits: CMUSphinx, Aquegg, Wikipedia
Evaluation of Automated Speech Recognition Models

Word Error Rate (WER) = 
\[
\frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}
\]

where
- \( S \) is the number of substitutions,
- \( D \) is the number of deletions,
- \( I \) is the number of insertions,
- \( C \) is the number of correct words,
- \( N \) is the number of words in the reference (\( S + D + C \))

Ground-truth speech (= Reference): \( N=15 \)
This is an example of the word error rate calculation for Boston University’s CS 640.

Model output:
This is example the word error rate calculation for Boston University’s see CS 640.

\( S=1, \ D=2, \ I=1, \ C=12 \)

\[
\text{WER} = \frac{1+2+1}{1+2+12} = \frac{4}{15} = 26.6\%
\]
Automated Speech Recognition: First Models in 2014, 2015: Google, CMU, UToronto

Acoustic signal

Spectrogram

Deep Neural Network

→ Text

Image credits: CMUSphinx, Aquegg, Wikipedia
Listen, Attend, and Spell (LAS) Model

Input:

Output:

Listen, Attend, and Spell (LAS) Model

Input:
- Mel-log spectrogram

Output:
- Lower-case English alphanumerics,
- 4 punctuations (space, period, comma, apostrophe),
- unknown token <unk>,
- start and end sentence tokens <sos>, <eos>

What is a mel log spectrogram?

The mel scale (after the word melody) is a perceptual scale of pitches judged by listeners to be equal in distance from one another. The reference point between this scale and a frequency measurement $f$ is defined by assigning a perceptual pitch of 1000 mels to a 1000 Hz tone. Above about 500 Hz, increasingly large intervals are judged by listeners to produce equal pitch increments.

Various experimentally-determined f-to-mel conversion formulas exist, e.g.,

$$m = 2595 \log_{10} \left( 1 + \frac{f}{700} \right)$$
A spectrogram is an intensity plot, usually on a log scale, so the term “log spectrogram” is also used. The plotted intensity is the squared magnitude of a Short-Time Fourier Transform (STFT) of audio data. The STFT is a sequence of Fast Fourier Transforms $X(m, \omega)$ of overlapping data windows $x[n]$ (overlap 25-50%).

Three important parameters:

- Window width $L$ (also called frame size), e.g., 25 milliseconds, long enough to encode part of a phoneme
- Frame stride (also called shift or offset) between successive windows, e.g., 10 ms
- Shape of window, e.g., Hamming Window $w[n]=0.54-0.46 \cos(2\pi n/L)$, between 0 and $L-1$, $w[n]=0$ otherwise.

The frequency $\omega$ is continuous.

$$\text{STFT}\{x[n]\}(m, \omega) \equiv X(m, \omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-i\omega n}$$

Spectrogram$x(t)$(m,\omega) = | X(m,\omega)|^2

Mel log Spectrograms

- Human hearing is more sensitive at lower frequencies and less sensitive at higher frequencies
- For speech recognition, we use a bank of filters

Source of plot: Davis and Mermelstein, 1980
Listen, Attend, and Spell (LAS) Model

Input: Mel-log spectrogram

Output: Lower-case English alphanumerics, 4 punctuations (space, period, comma, apostrophe), unknown token <unk>, start and end sentence tokens <sos>, <eos>

Listen, Attend, and Spell (LAS) Model

Listen: Encoder

Spell: Attention-based Decoder

Speech waveform

English characters

Listen, Attend, and Spell (LAS) Model

Listen:
Encoder

Spell:
Attention-based Decoder

Speech waveform

Huge number of parameters

English characters

Much smaller number of possible characters

DNN for Speech Recognition -- First Models: Pundak & Sainath’s Frame Rate Reduction

Speech waveform → STFTs → Filter Banks → mel log conversion → 80-dim mel log spectrogram feature vector per 25-ms window every 10 ms

Stacking = concatenation of feature vectors → 640-dim mel log spectrogram every 80 ms

cov max pooling bottle-neck layer LSTM LSTM LSTM FC Soft max

Encoder

Decoder

English characters

Source: Pundak 2016
Add a Language Model

- Encoder/Decoder models implicitly learn a language model from training with speech & character labels (e.g., 3 million utterances = 2000 hr of Google voice search traffic were used by Pundak & Sainath)

- Instead of text paired with speech, we can also use text alone, using a very large language model (LLM):
  - Get list of n-best hypotheses, i.e., beam search
  - Use LLM to rescore hypotheses in beam:

    \[
    \text{Score}(\text{character}|\text{acoustic}) = \log p(\text{character}|\text{acoustic}) + \alpha \log p_{\text{LLM}}(\text{character})
    \]
2023: Speech Recognition in 100+ Languages: Google

- **Universal Speech Model (USM):**
  - Google blog

- Encoder/Decoder Architecture

- Self-supervised learning with fine-tuning

2023: Speech Recognition in 100+ Languages: Google’s USM

- **Encoder:** Conformer (convolution-augmented transformer), Gulati et al., 2020. Subsamples mel-log spectrograms and sends resulting feature vectors to attention, feed-forward, and convolutional modules, to produce final embedding.

- **Decoder:** CTC, RNN-T, or LAS (see Google blog for links to relevant papers)
Voice Cloning

Definition:
Artificial simulation of a person’s voice

Person-specific Voice Embedding -> Voice Cloner -> Person-specific Spoken Text

Text
Evaluation of Voice Cloning

Two criteria evaluated by humans:
- Naturalness of voice
- Similarity of voice

Two evaluation methodologies:
1. Likert scale: On a scale from 1 to 5, evaluate the criterium.
2. A/B testing: Listen to 2 voices, created by model or person A and B respectively, and give preference according to the criterium. Best practice is to “blind” human tester to which voice is produced by A or B.
Voice Cloning Example

Whose voice is this?
Dangers of Voice Cloning

Use of voices without permission of speaker e.g. : https://play.ht/voice-cloning/

Cyberbullying

Warfare with Deep Fakes

Screenshot used for educational purposes without permission by President Obama or PlayHT
One Useful Application:

Help Users with ALS or Multiple Sclerosis to “keep” their voice

Before a generative disease takes away a person’s ability to speak, the person could train a neural network to compute a speaker-specific voice embedding. This embedding could later be used to control a person-specific text-to-speech voice synthesizer.
Voice Cloning

Definition:
Artificial simulation of a person’s voice

Use Case, Inference:

Person-specific Voice Embedding → Voice Cloner

Text → Person-specific Spoken Text
Voice Cloning

Definition:
Artificial simulation of a person’s voice

Use Case, Inference:

Person-specific Voice Embedding → Voice Cloner → Person-specific Spoken Text

Text
Optional: Emotion
Voice Cloning

Use Case, Inference:

Person-specific Voice Embedding

Text

Voice Cloner

Synthesizer

Spectrogram

Vocoder (voice encoder)

Waveform = Person-specific Spoken Text
Voice Cloning

How to obtain a person-specific voice embedding:

1. **Text spoken by particular person**
2. Input into Speaker Encoder
3. **Person-specific Voice Embedding**

- **Written text**

Diagram representation:

- **Written text** ➔ **Speaker Encoder** ➔ **Person-specific Voice Embedding**
Voice Cloning

How to obtain a person-specific voice embedding:

3-4 Hours of

Written text

Text spoken by particular person

Speaker Encoder

Person-specific Voice Embedding
Voice Cloning

How to obtain a person-specific voice embedding:

**Short Utterance of**

- **Written text**
- **Text spoken by particular person**

**Pretrained Speaker Encoder**

“Pretty Representative” Person-specific Voice Embedding
Voice Cloning

Training
Text spoken by particular person

Written text to be converted into speech of particular person at inference time

Voice Cloner

Speaker Encoder

Person-specific Voice Embedding

Synthesizer

Vocoder (voice encoder)

Spectrogram

Person-specific Spoken Text
Voice Cloning:
3 Independently trained neural nets

**Diagram:**
- **Voice Cloner**
  - **Speaker Encoder**
    - Person-specific Voice Embedding
  - **Synthesizer**
  - **Vocoder** (voice encoder)

**Text:**
- Training Text spoken by particular person
- Written text to be converted into speech of particular person at inference time
- Person-specific Spoken Text
Voice Cloning by Google
3 Independently trained neural nets

Multi-speaker Speech Synthesis Model

- Recurrent Speaker Encoder
- Seq-to-Seq Synthesizer
- WaveNet Vocoder

Person-specific Spoken Text

Written text to be converted into speech of particular person at inference time

Training Text spoken by particular person
Voice Cloning by Google
3 Independently trained neural nets

Multi-speaker Speech Synthesis Model

- Recurrent Speaker Encoder [Wan et al., 2018]
- Person-specific Voice Embedding
- Seq-to-Seq Synthesizer Tacotron [Shen et al., 2016]
- WaveNet Vocoder [van den Oord et al., 2016]

Written text to be converted into speech of particular person at inference time

Training Text spoken by particular person

Person-specific Spoken Text
Task: Text-independent Speaker Verification on specific text, e.g., “OK Google”

Input: Text spoken by a particular person

Output: Person-specific Voice Embedding

Contribution: New Loss Function “GE2E”

Publication:
Multiple utterances of training text spoken by particular person

Previous State-of-the-Art of Speaker Verification

Speaker Encoder

Person-specific Spoken Text

LSTM

Average Speaker Model

Same Person?

Yes or No

Log. regression

Cos Similarity

Embedding at Inference

Speaker Embedding
Previous State-of-the Art of Speaker Verification

- Multiple utterances of training text spoken by particular person

Person-specific Spoken Text

LSTM

Speaker Encoder

- Average Speaker Model

- Speaker Embedding

- Cos Similarity

- Log. regression

- Same Person?

Yes or No

Embedding at Inference

Computer Science
Wang et al.’s Contribution: GE2E Loss function

GE2E uses a similarity matrix $S_{ji,k}$ that defines the similarities between each embedding $e_{ji}$ (jth speaker, ith word) and all centroids $c_k$ (kth speaker) to compute the contrast loss. The formula for $L(e_{ji})$ is given by:

$$L(e_{ji}) = 1 - \sigma(S_{ji,j}) + \max_{1 \leq k \leq N, k \neq j} \sigma(S_{ji,k}),$$

where $\sigma$ is the ReLU function. The figure illustrates the similarity matrix and the contrastive loss, showing how the embedding $e_{ji}$ is pulled towards the centroid $c_j$ of the true speaker and away from the centroids $c_k$ of similar different speakers.

**Fig. 2.** GE2E loss pushes the embedding towards the centroid of the true speaker, and away from the centroid of the most similar different speaker.
Wang et al.’s Contribution: GE2E Loss function

GE2E uses a similarity matrix $S_{ji,k}$ that defines the similarities between each embedding $e_{ji}$ (jth speaker, ith word) and all centroids $c_k$ (kth speaker) to compute the contrast loss:

$$L(e_{ji}) = 1 - \sigma(S_{ji,j}) + \max_{1 \leq k \leq N} \sigma(S_{ji,k}),$$

where $\sigma$ is the softmax function.

**Fig. 2.** GE2E loss pushes the embedding towards the centroid of the true speaker, and away from the centroid of the most similar different speaker.
Voice Cloning by Google
3 Independently trained neural nets

Multi-speaker Speech Synthesis Model
[Jia et al., 2019]

- Recurrent Speaker Encoder [Wan et al., 2018]
  - Person-specific Voice Embedding
  - Seq-to-Seq Synthesizer Tacotron [Shen et al., 2016]
  - Spectrogram

- WaveNet Vocoder [van den Oord et al., 2016]

Training
Text spoken by particular person

Written text to be converted into speech of particular person at inference time

Person-specific Spoken Text
Task: Convert spectrogram into natural-sounding speech signal
Input: Spectrogram
Output: Waveform
Contribution: Network architecture based on “dilated causal convolutions”
Publication:
van den Oord et al., 2016: Dilated Causal Convolutional Layers

Figure 3: Visualization of a stack of *dilated* causal convolutional layers.
Voice Cloning by Google
3 Independently trained neural nets

Multi-speaker Speech Synthesis Model
[Jia et al., 2019]

Recurrent Speaker Encoder [Wan et al., 2018]

Seq-to-Seq Synthesizer Tacotron
[Shen et al., 2016]

WaveNet Vocoder
[van den Oord et al., 2016]

Written text to be converted into speech of particular person at inference time

Training Text spoken by particular person

Person-specific Spoken Text

Person-specific Voice Embedding

Spectrogram
Shen et al., 2016

- Task: Convert text into spectrogram that can be passed into WaveNet Vocoder
- Input: Text
- Output: Spectrogram
- Contribution: Improved Naturalness of Voice, Reduction of size of WaveNet
- Publication:

Voice Cloning by Google
3 Independently trained neural nets

Multi-speaker Speech Synthesis Model
[Jia et al., 2019]

Recurrent Speaker Encoder [Wan et al., 2018]

Seq-to-Seq Synthesizer Tacotron [Shen et al., 2016]

Person-specific Voice Embedding

WaveNet Vocoder [van den Oord et al., 2016]

Written text to be converted into speech of particular person at inference time

Training Text spoken by particular person

Person-specific Spoken Text
Jia et al.’s Voice Cloner
Figure 2: Example synthesis of a sentence in different voices using the proposed system. Mel spectrograms are visualized for reference utterances used to generate speaker embeddings (left), and the corresponding synthesizer outputs (right). The text-to-spectrogram alignment is shown in red. Three speakers held out of the train sets are used: one male (top) and two female (center and bottom).
How to detect voice clones?

Two types of approaches:
1) Handcraft features, 2) Learn features
that NNs then use to distinguish real speech and synthesized speech

Handcrafted features include acoustic features, inverse Fourier transform coefficients, correlation of audio signal frames, etc.

Dataset to train/test:

Comprehensive journal paper on deep fake generation & detection (up to 2022): Masood et al., 2023
Learning Outcomes:

Being able to

- Define speech recognition, phoneme, wake word detection, mel scale, spectrogram, encoder, decoder, Short-Time Fourier Transform, voice cloning
- Discuss sources of variability of an acoustic signal and constraints on how a phoneme is realized acoustically
- Explain parsing as a tree search
- Explain the difference between speaker dependent and independent speech recognition
- Explain how HMMs were/are used in speech recognition
- Explain the choice of the wake word and how it can be detected
- Give criteria for evaluation of speech recognition and voice cloning
- Describe the LAS model
- Explain how a language model can be added to a encoder/decoder speech recognition model
- Discuss the state of the art in speech recognition in 2023 (USM)
- Explain a voice cloning model and its connection to the task of speaker identification
- Explain the dangers of voice cloning
- Discuss how to detect voice clones