CAS CS 640

Artificial Intelligence

Lecture by Margrit Betke

Automated Speech Recognition & Voice Cloning

Learning Outcomes: Being able to

Computer Science

- 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?

Computer Science

- 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(**words** | **signal**): $\text{argmax}_{\textbf{words}}$ P(**words** | **signal**)
- Bayes rule: $\text{argmax}_{\textbf{words}}$ P(**signal** | **words**) P(**words**),

  where P(**words**) represents our **language model** = prior probability of a particular word string and likelihood P(**signal** | **words**) = **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

Computer Science

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

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

phonetic segment = phoneme

phone = smallest perceptible segment

Acoustic Realization of Phonemes Depends on

Computer Science

- 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?

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

Computer Science

phonetic segment = phoneme

phone = smallest perceptible segment

Alexa's Phonemes?

A l e x a
AH-L-EH-K-S-AH

Rare combination of phonemes (sounds) in English

→ Alexa is a smart "wake up word"

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

Computer Science

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

Computer Science

Why are these funny?

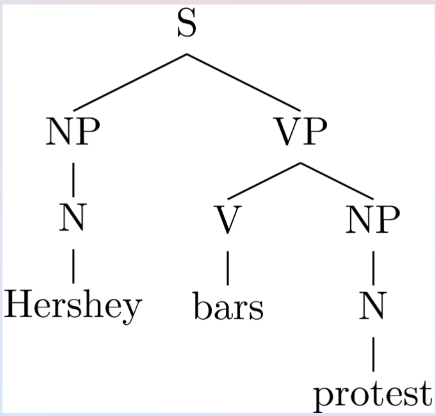
Headlines:

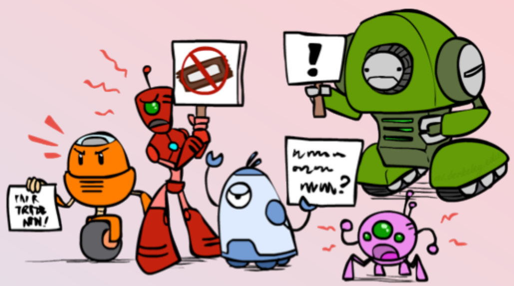
- Enraged Cow Injures Farmer With Ax
- Hospitals Are Sued by 7 Foot Doctors
- Ban on Nude Dancing on Governor's Desk
- Iraqi Head Seeks Arms
- Local HS Dropouts Cut in Half
- Juvenile Court to Try Shooting Defendant
- Stolen Painting Found by Tree
- Kids Make Nutritious Snacks

Slide by D. Klein, P. Abbeel, UC Berkeley

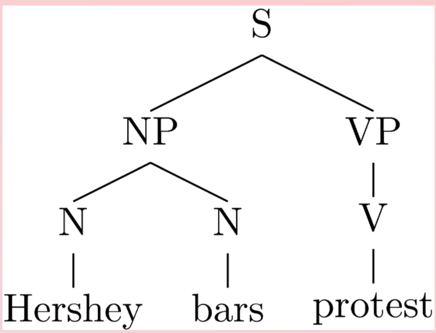
Parsing as Search

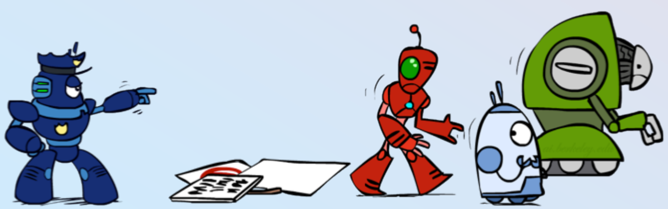
Computer Science



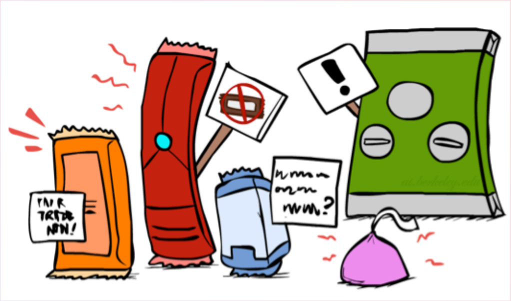


Hershey bars protest





The company Hershey forbids protest.



Chocolate bars are protesting.

Slide by D. Klein, P. Abbeel, UC Berkeley

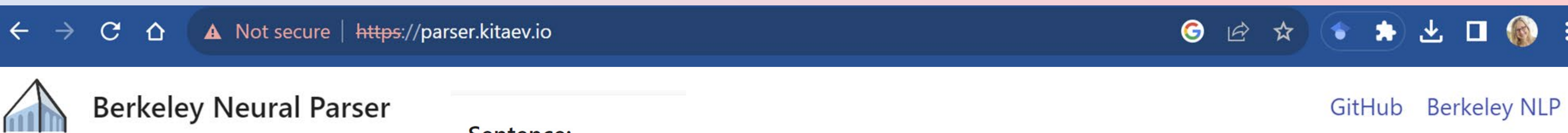
Probabilistic Context-Free Grammars

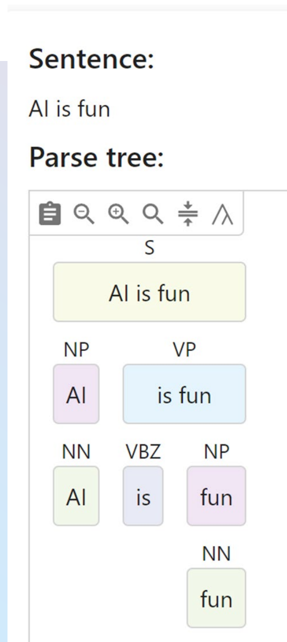
Computer Science

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

https://parser.kitaev.io/

- 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!





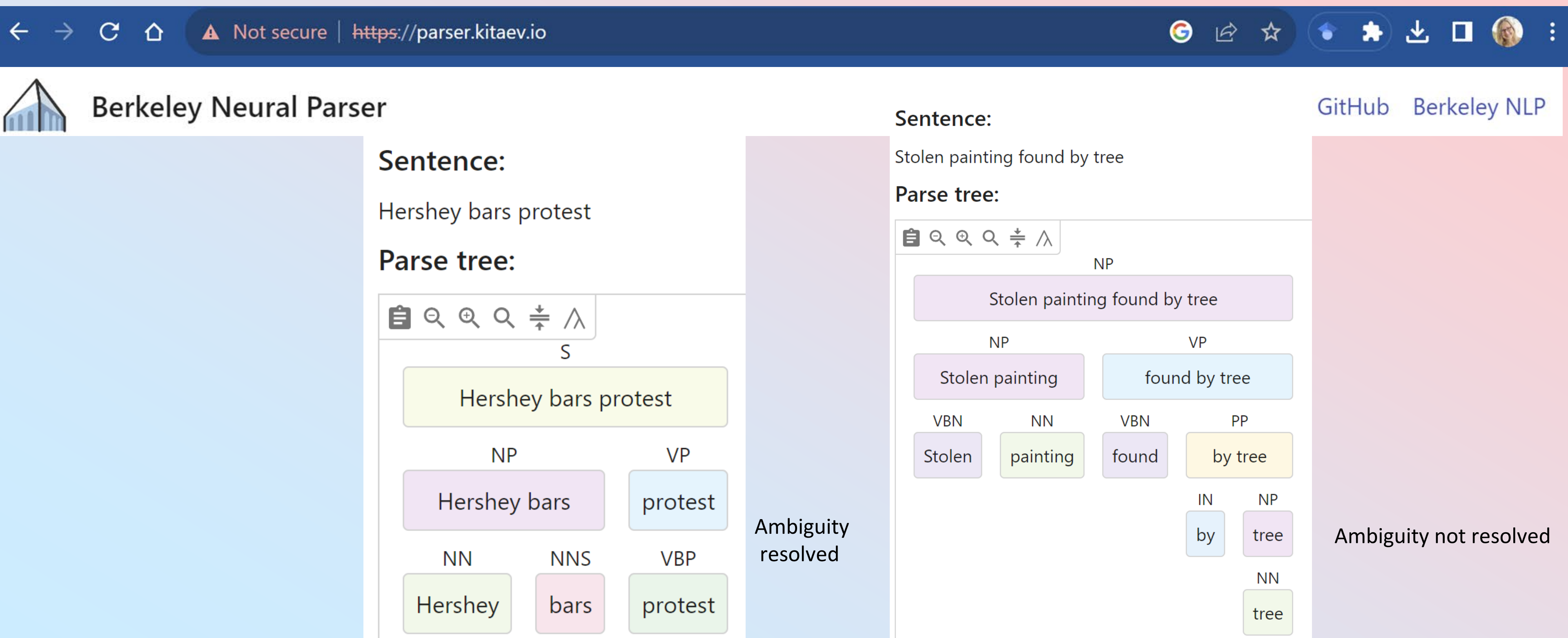
NP: Noun phrase
VP: Verb phrase
NN: Noun singular
VBZ: Verb 3rd person singular present

Probabilistic Context-Free Grammars

Computer Science

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

- 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!



Ambiguity resolved

Ambiguity not resolved

Early Ideas for Automated Speech Recognition (1970s)

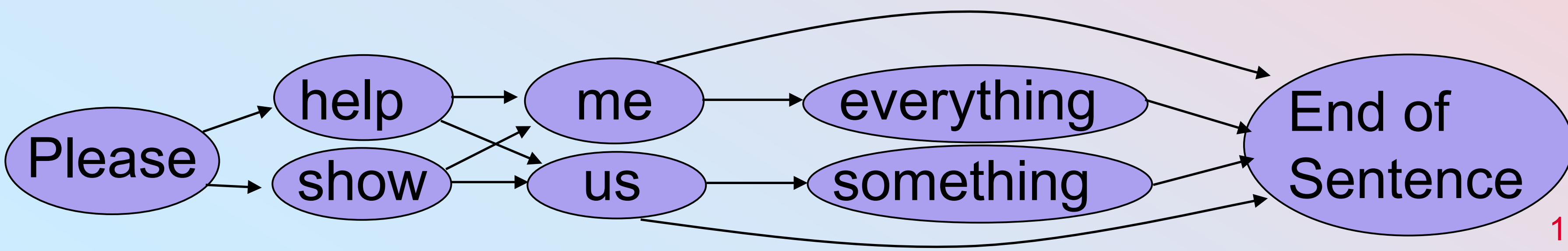
Computer Science

- IBM's "tri-gram model"   Word1   Word2   Word3

  max $P($ **Word3 | Word1 & Word2**$)$

- CMU's Hearsay I   played voice chess
  - top-down, expectation-driven approach
- CMU's Harpy

  sentence = path through network represents sequence of sounds



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

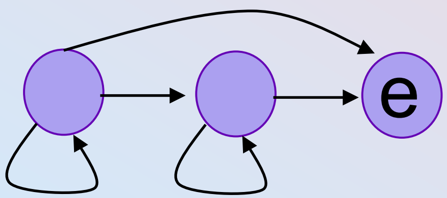
Ability to define “macros”

HMMs in Speech Reognition

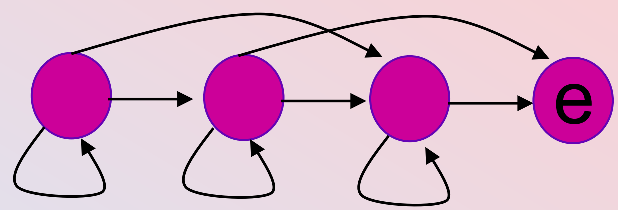
Constructing left-to-right HMM for word sequences:

Concatenate HMMs (with non-emitting end states) for each word in sentence:

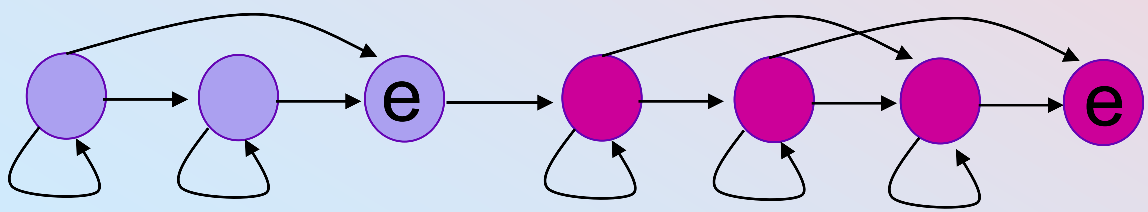
HMM for Word1:



HMM for Word2:



Combined HMM for sequence **Word1 Word2:**



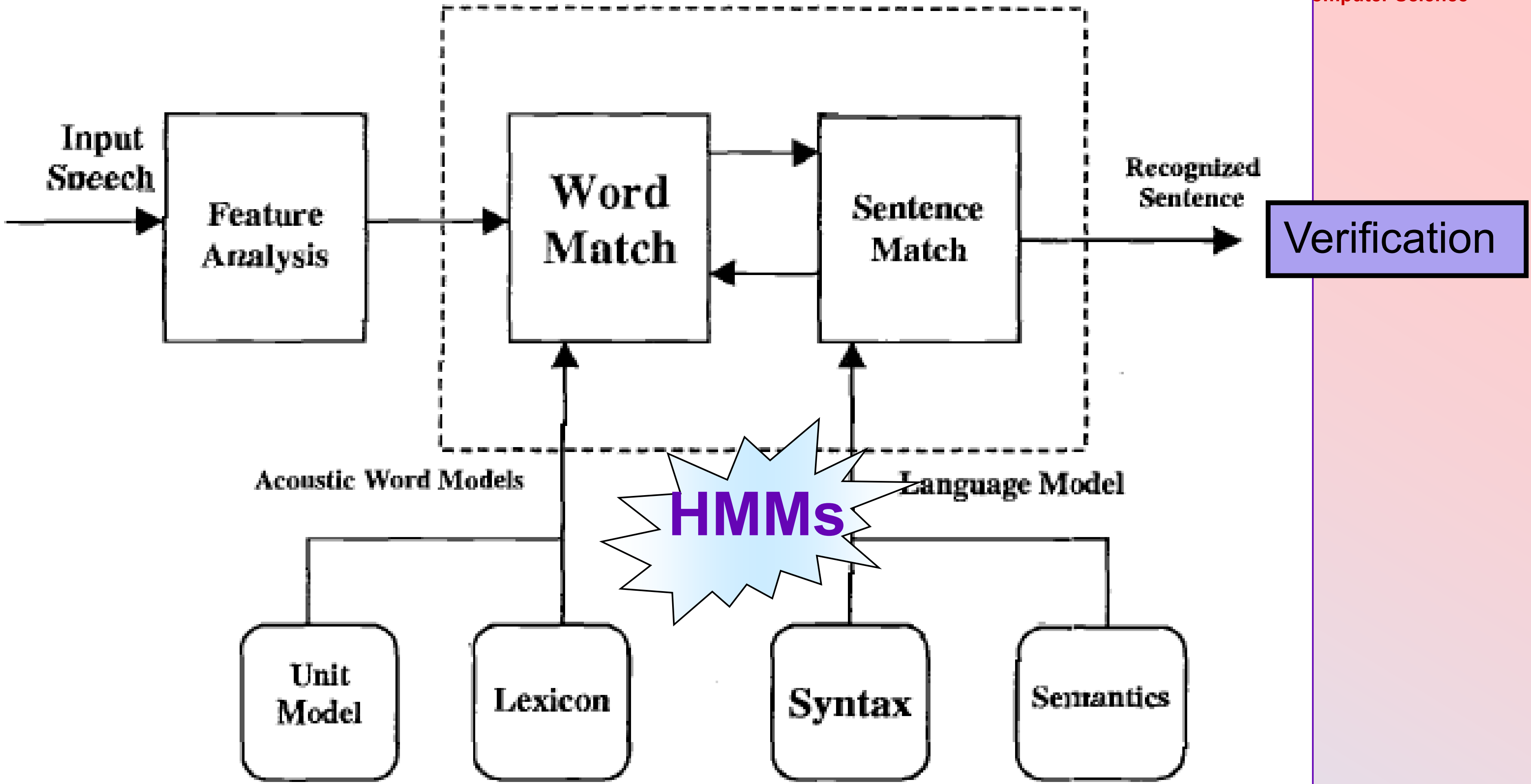
HMMs in Speech Reognition

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 than were never seen in the training data can be constructed from phoneme HMMs and recognized

Generic Fluent-Speech Recognition System

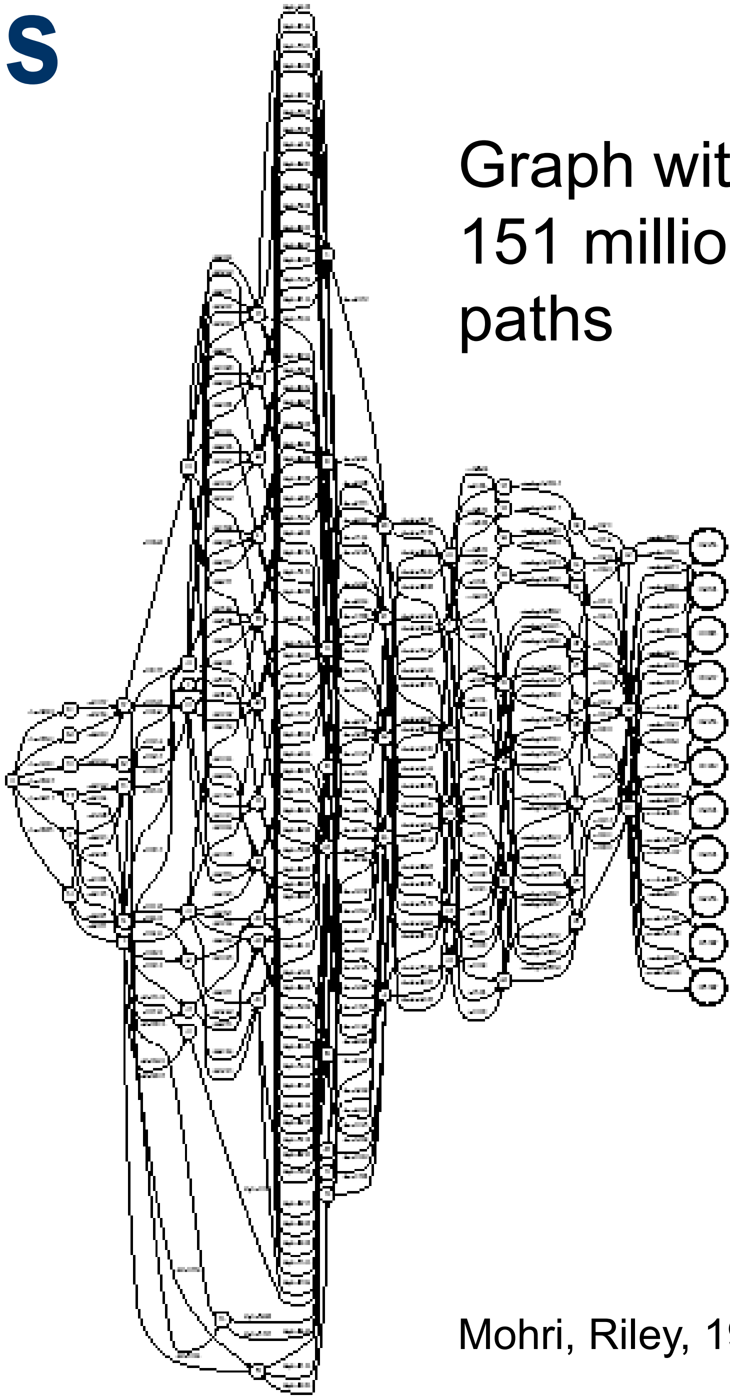


Rabiner 1997

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"



Graph with 151 million paths

Mohri, Riley, 1999

Performance of Speech Recognition Systems

Task	Vocabulary Size	Error Rate
Digits 0-10	11	0.3% per digit
Airline travel info	2,500 words	2% per word
Reading newspaper	64,000 words	8% per word
Radio	64,000 words	27% per word
Conversation over phone	28,000 words	37% per word

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

Computer Science

- 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+

Computer Science

- 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)

New York Times: 1/24/2017

Computer Science

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

Computer Science

- **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'' ($8x8x24cm^3$)
- **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

Computer Science

- Plays all your music from Amazon Music, Spotify, Pandora, iHeartRadio, TuneIn, and more using just your voice
- Fills the room with immersive, 360º 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

Computer Science

- 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?

<u>On device processing:</u>

User: *"Alexa, order flowers for my grandma"*

Signal Processing

beam-formed signal

Wake Word Detection

<u>Cloud Processing:</u>

Alexa: *"I have ordered flowers"*

App Layer, Text to Speech

Recognized Intent: BuyItem ItemName: Flowers

Natural Language Processing

Speech-to-text

Automatic Speech Recognition

Wake Word Detection

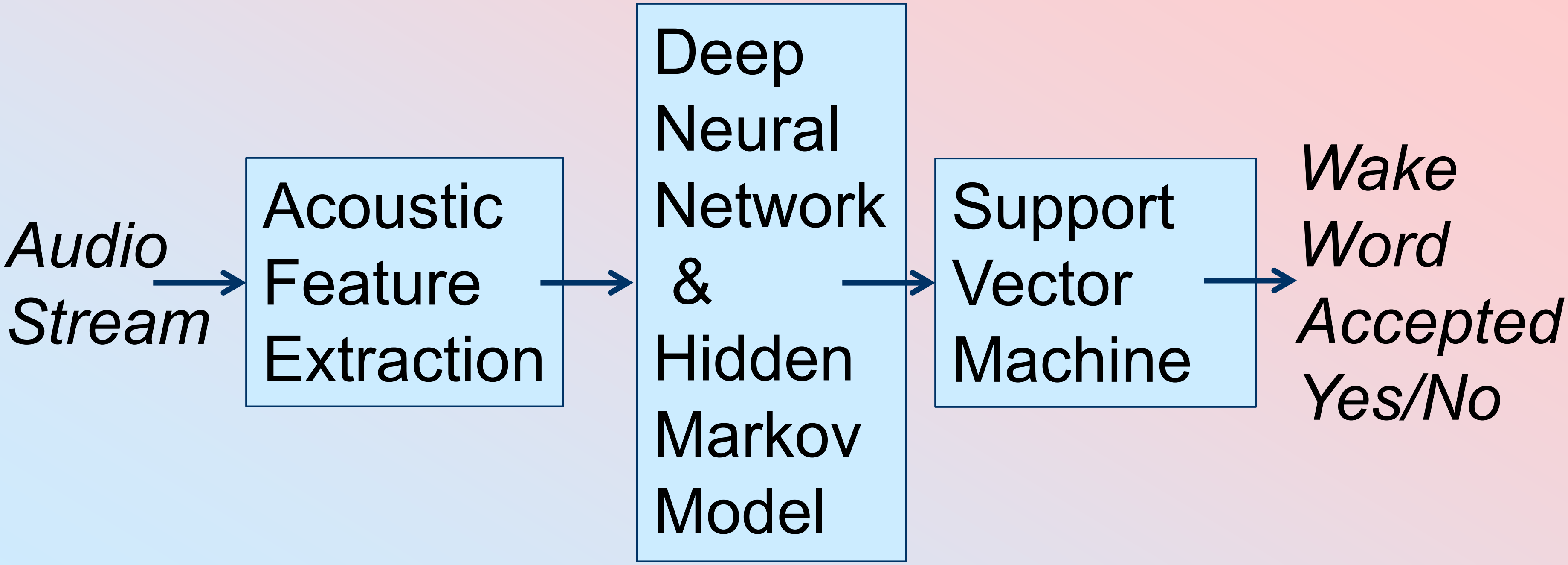
Goal: High “positive” detection rate with **no** false positives

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



Wake Word DNN/HMM Model

Two finite state machines (FSMs):

1. Foreground wake word FSM
2. Background speech/non-speech FSM

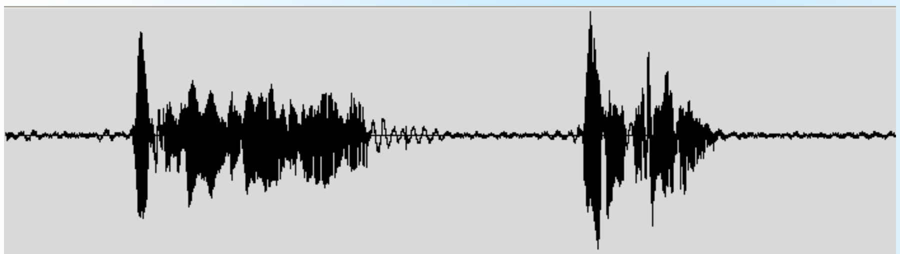
Deep neural network (DNN) produces posterior probabilities p(state | acoustic features)

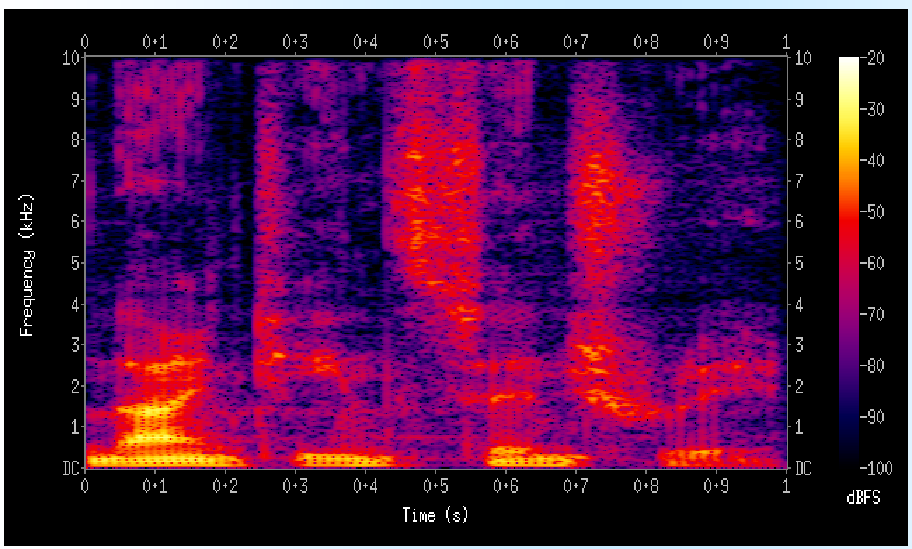
Detection confidence is computed from foreground/background likelihood ratio

Automated Speech Recognition (ASR)

Computer Science

Acoustic signal





Spectrogram

Deep Neural Network → Text

Image credits: CMUSphinx, Aquegg, Wikipedia

Evaluation of Automated Speech Recognition Models

Computer Science

Word Error Rate (WER) =
(S+D+I)/N = (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:

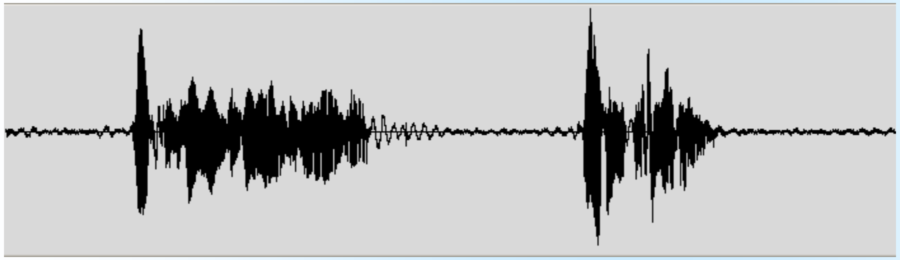
This is example the world error rate calculation for Boston University's see CS 640.

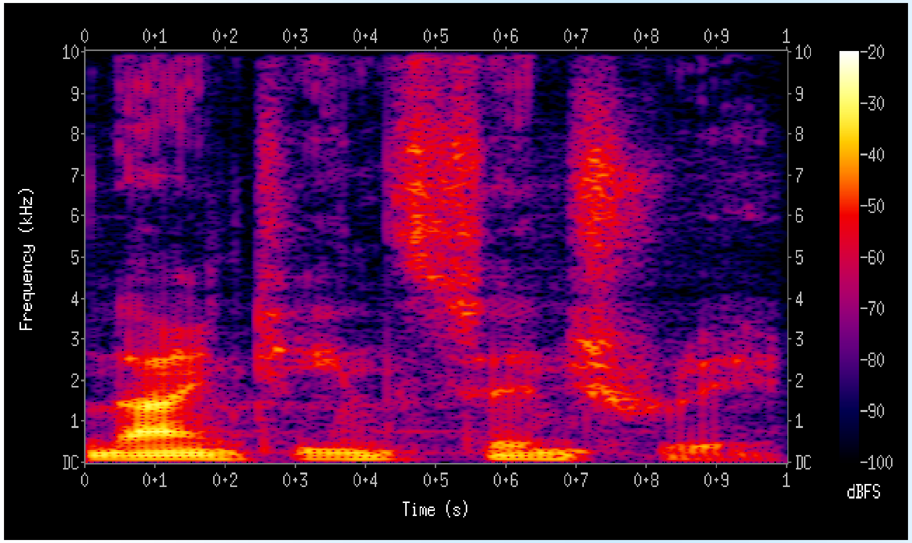
S=1, D=2, I=1, C=12

WER = (1+2+1)/(1+2+12)= 4/15 = 26.6%

Automated Speech Recognition: First Models in 2014, 2015: Google, CMU, UToronto

Acoustic signal





Deep Neural Network

→ Text

Spectrogram

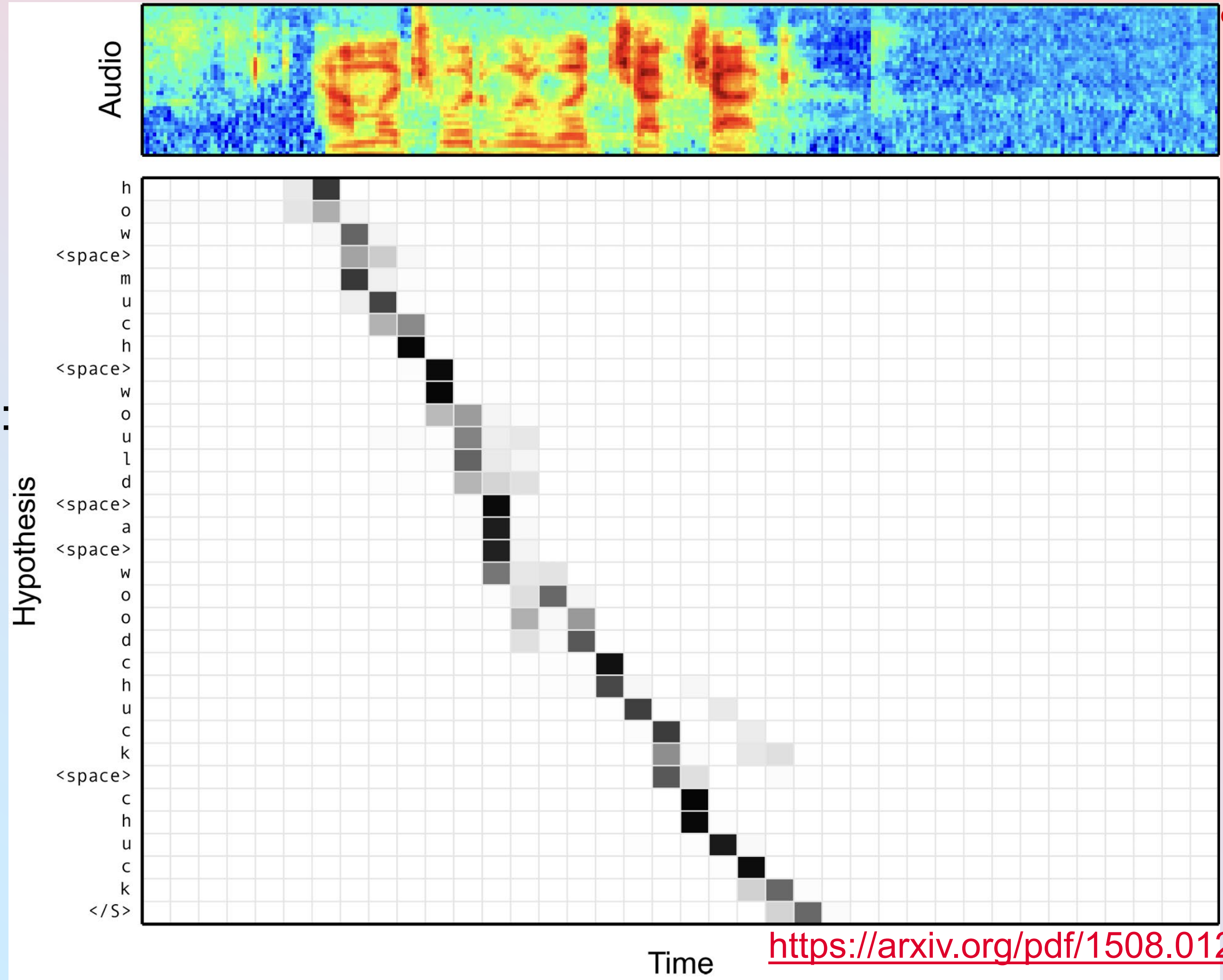
Image credits: CMUSphinx, Aquegg, Wikipedia

Listen, Attend, and Spell (LAS) Model

Computer Science

Input:

Output:



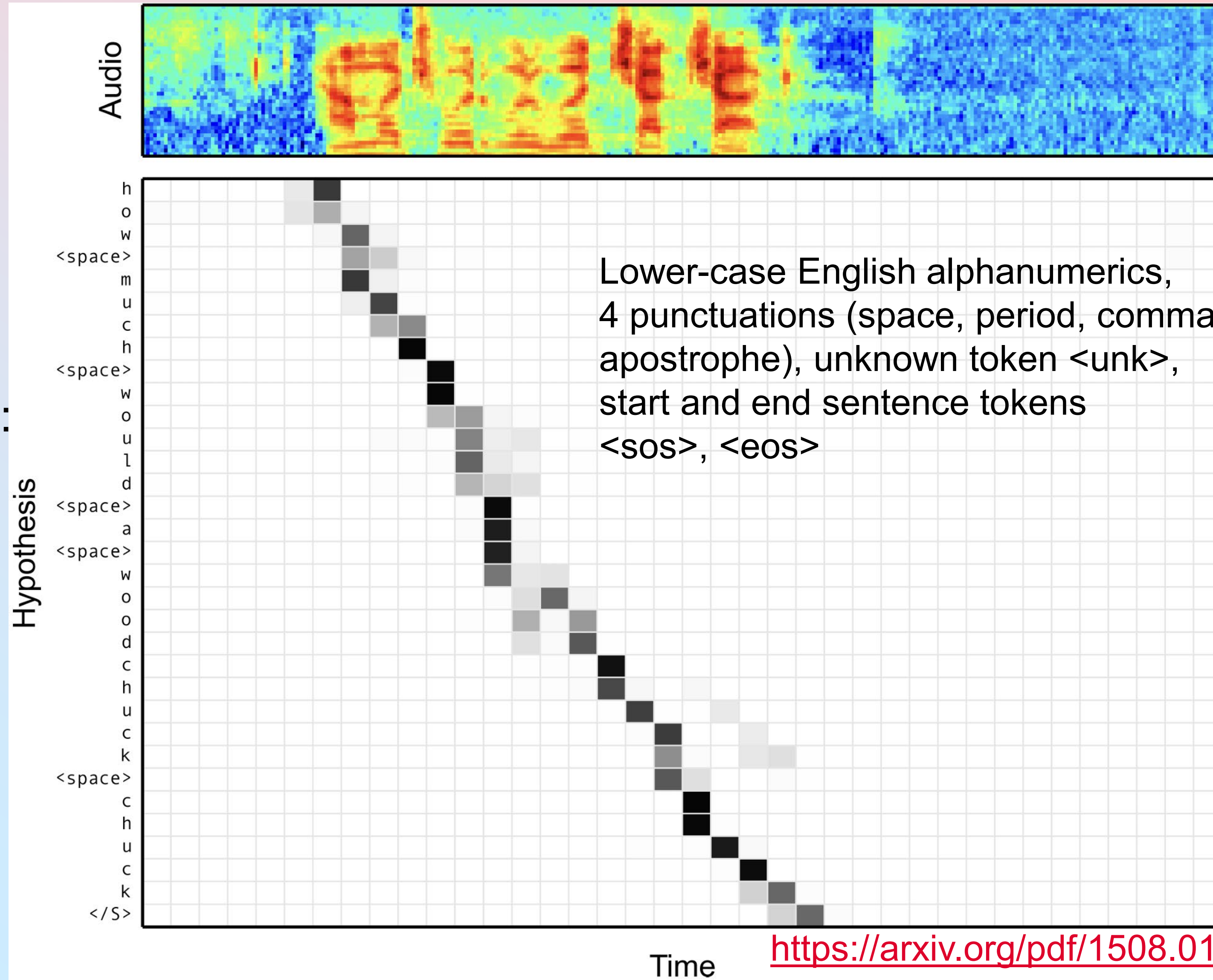
https://arxiv.org/pdf/1508.01211.pdf

Listen, Attend, and Spell (LAS) Model

BOSTON UNIVERSITY
Computer Science

Input:

Output:



Mel-log spectrogram

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

https://arxiv.org/pdf/1508.01211.pdf

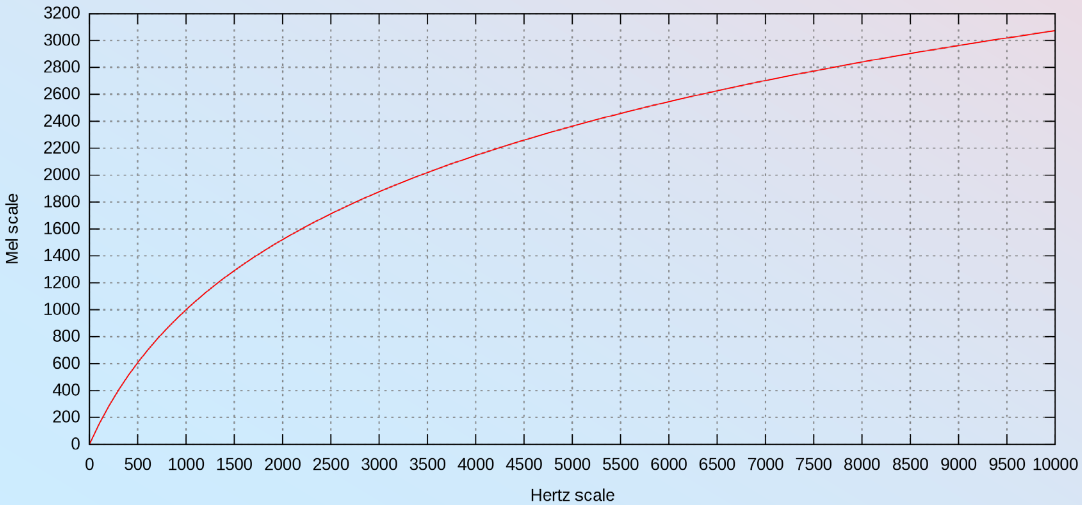
What is a mel log spectrogram?

Computer Science

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.,

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$



Source: Wikipedia

What is a mel log spectrogram?

Computer Science

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 ω is continuous.

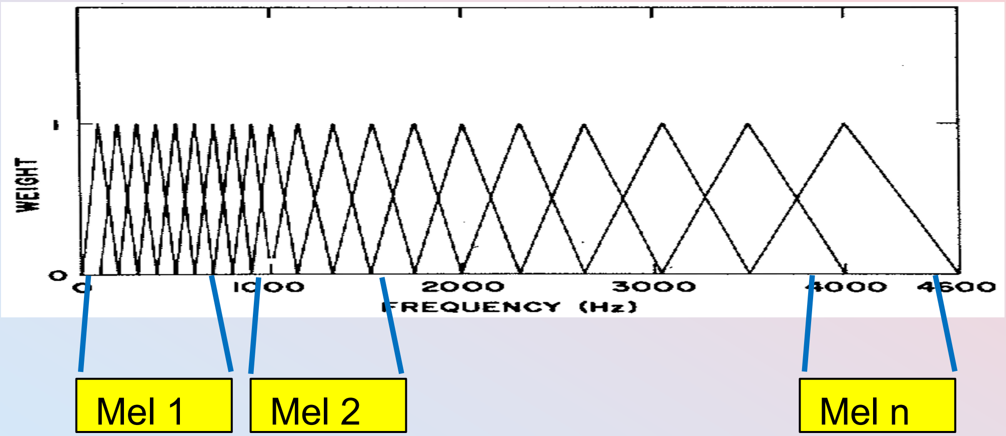
$$\mathbf{STFT}\{x[n]\}(m,\omega) \equiv X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-i\omega n}$$

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

Source: Wikipedia

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

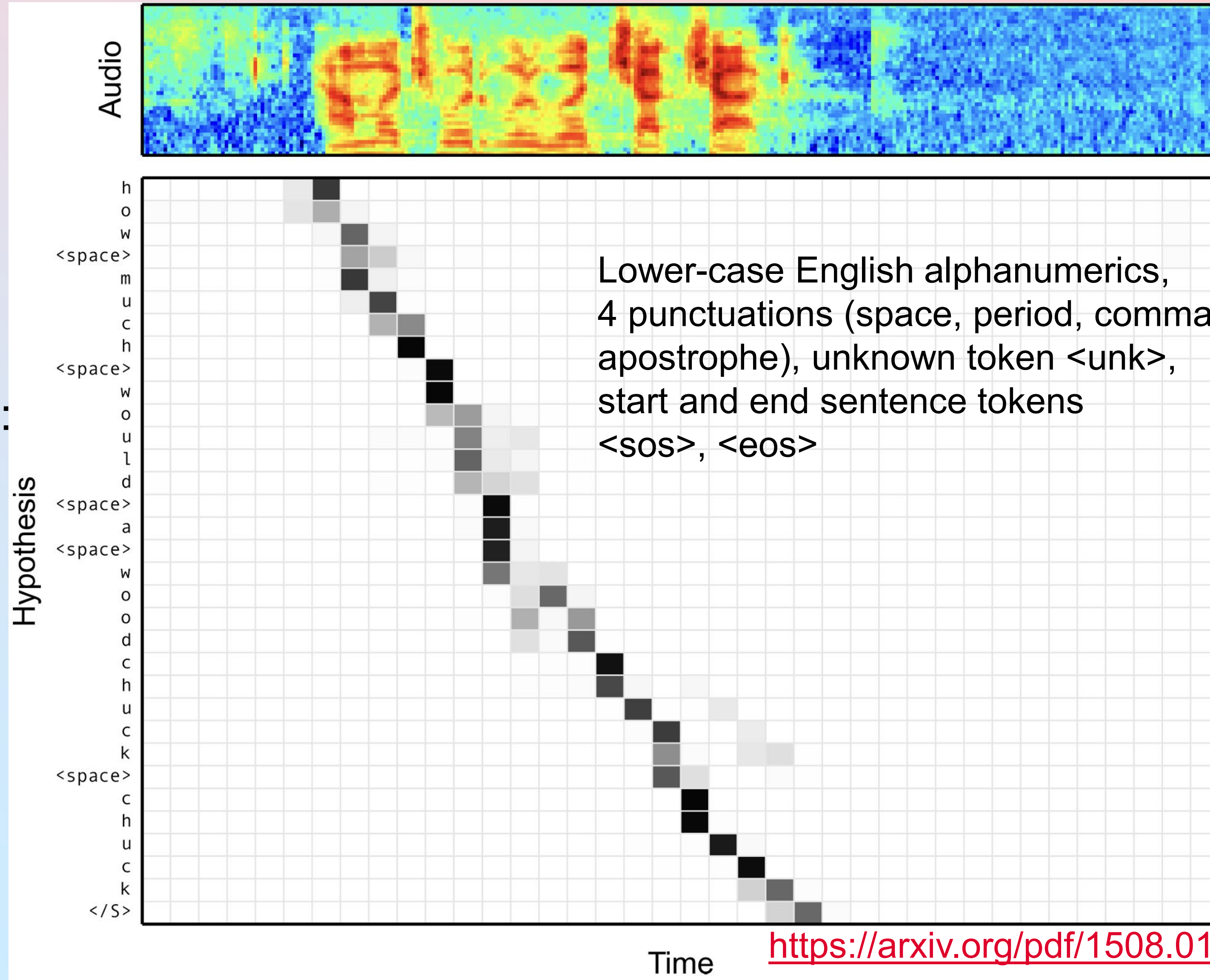


Source of plot: Davis and Mermelstein, 1980

Listen, Attend, and Spell (LAS) Model

Input:

Output:



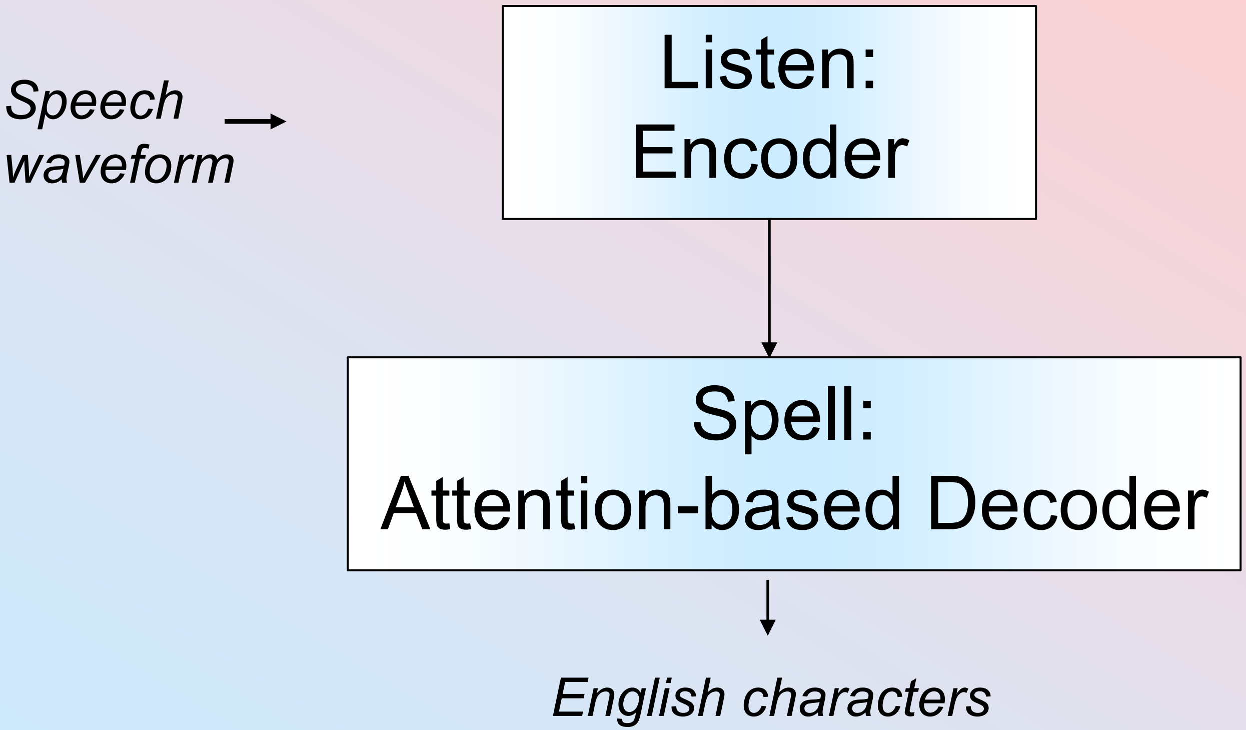
Mel-log spectrogram

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

https://arxiv.org/pdf/1508.01211.pdf

Listen, Attend, and Spell (LAS) Model

Computer Science



https://arxiv.org/pdf/1508.01211.pdf

Listen, Attend, and Spell (LAS) Model

Computer Science

Speech waveform →

Listen:
Encoder

↓

Spell:
Attention-based Decoder

↓

English characters

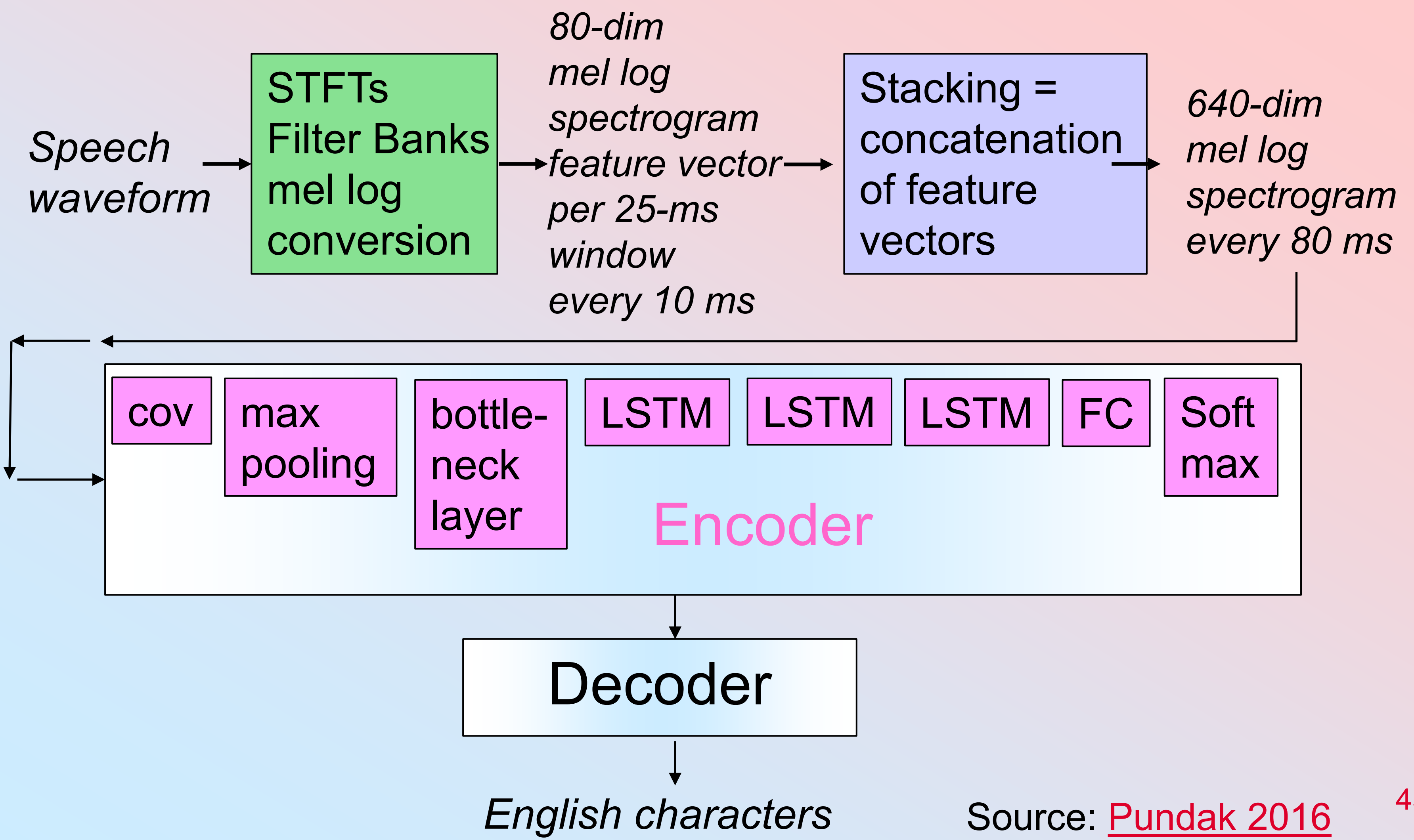
Huge number of parameters

Much smaller number of possible characters

https://arxiv.org/pdf/1508.01211.pdf

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

Computer Science



Source: Pundak 2016

Add a Language Model

Computer Science

- 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:

$$\text{Score}(\text{character}|\text{acoustic}) = \log p(\text{character}|\text{acoustic}) + \alpha \log p_{LLM}(\text{character})$$

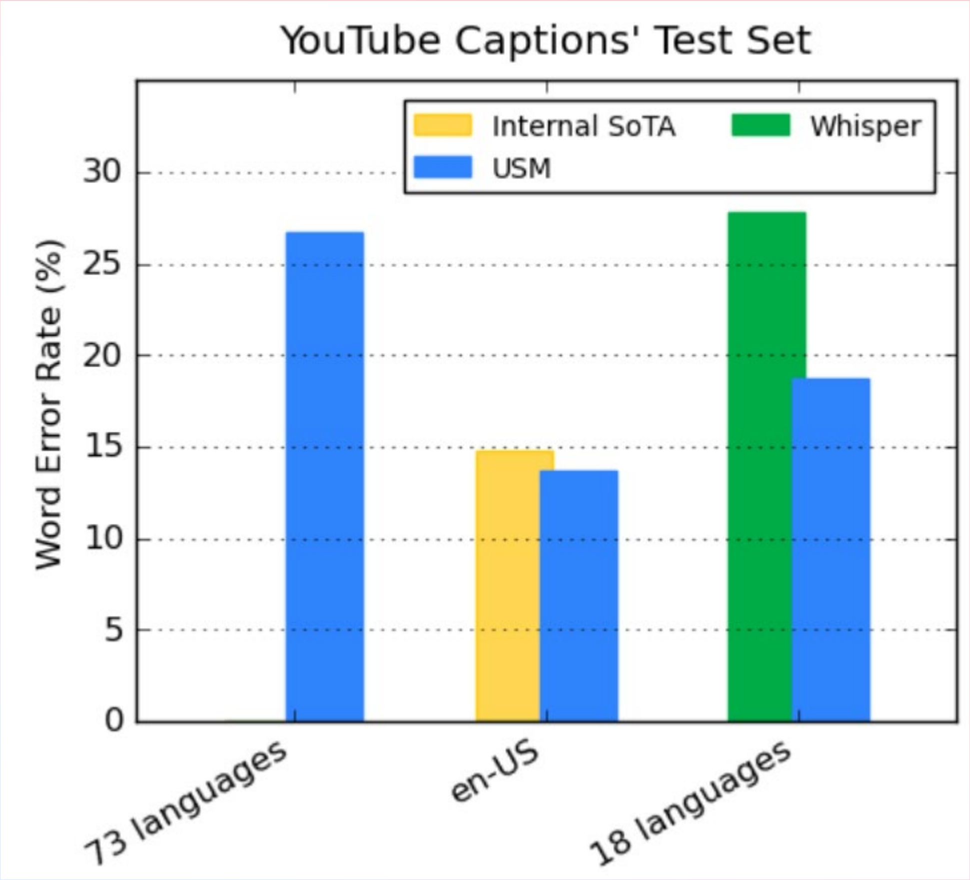
2023: Speech Recognition in 100+ Languages: Google

Computer Science

- Universal Speech Model (USM):

Google blog

- Encoder/Decoder Architecture
- Self-supervised learning with fine-tuning
- https://arxiv.org/pdf/2303.01037.pdf



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

Computer Science

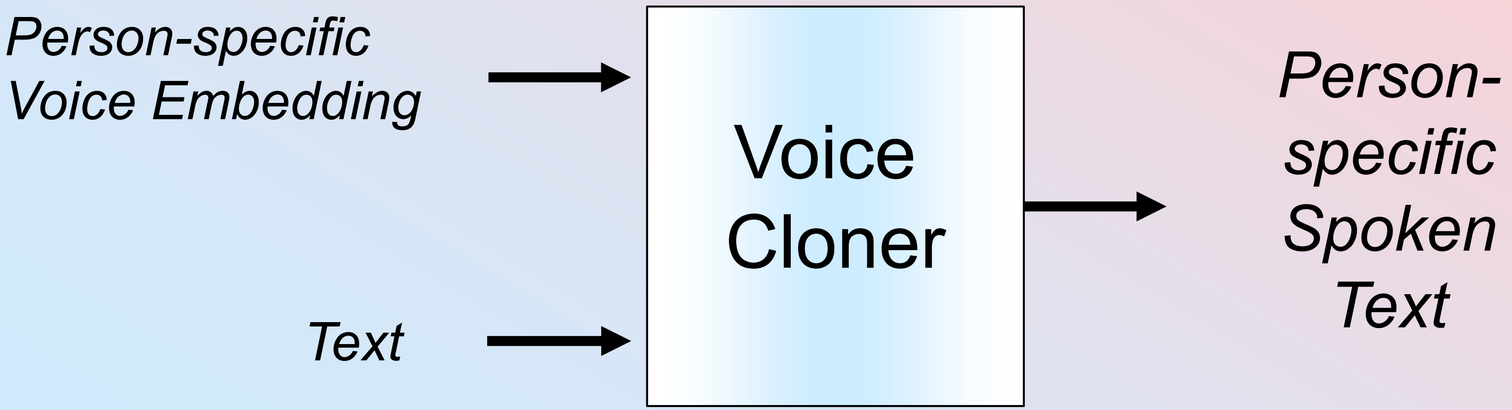
- Encoder: Conformer (convolution-augmented transformer), Gulati et al., 2020. Subsamples mel-log spectrograms and sends resulting feature vectors to attention, feed-forward, and convolutional modules, to produce final embedding.

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

Voice Cloning

Definition:

Artificial simulation of a person’s voice



Evaluation of Voice Cloning

Two criteria evaluated by humans:

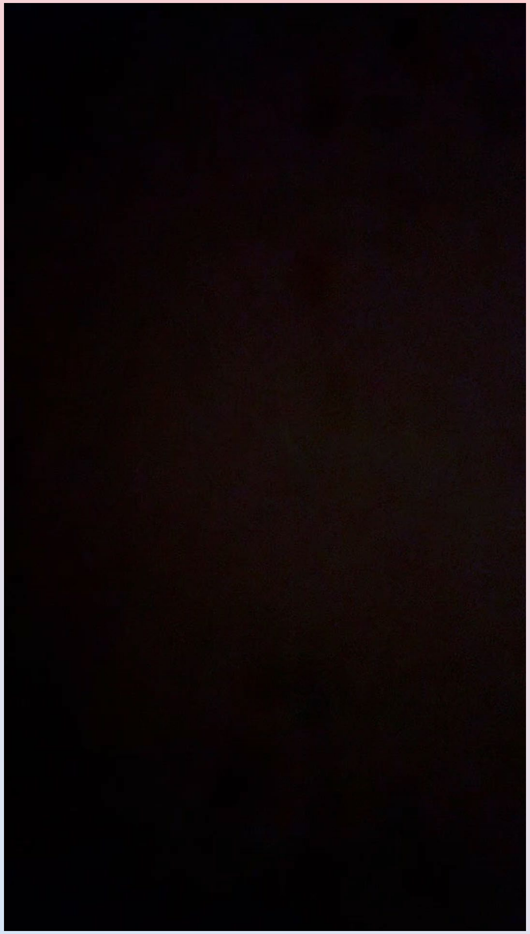
- Naturalness of voice
- Similarity of voice

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?

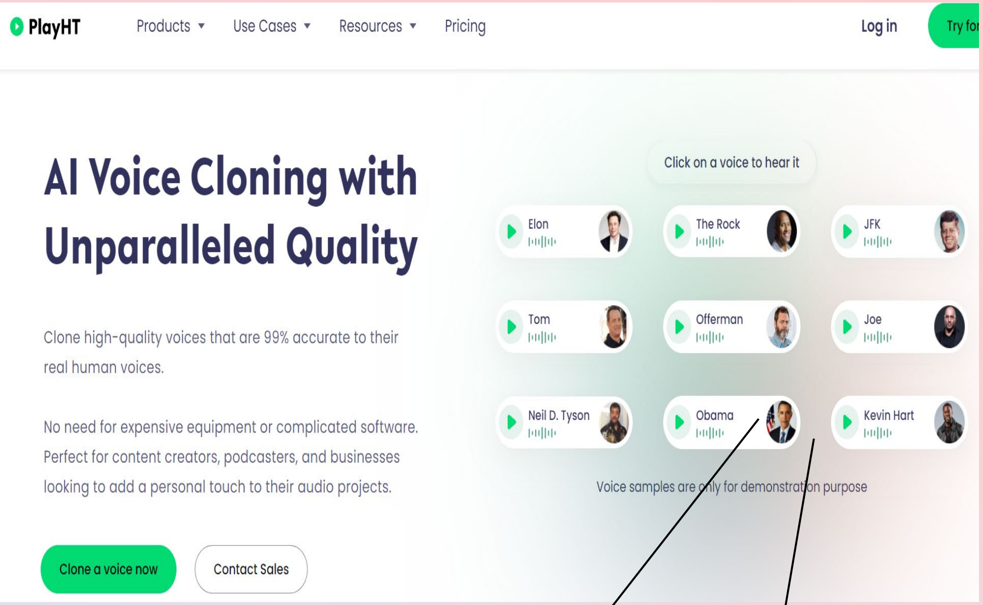


Dangers of Voice Cloning

Computer Science

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

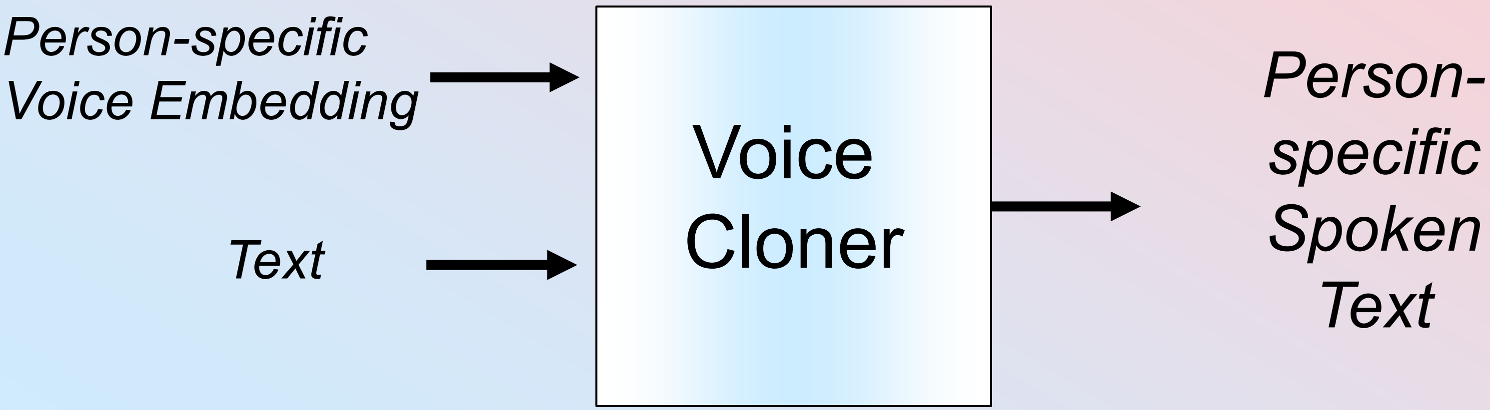
Before a generative disease takes away a person’s ability to speak, the person could train a neural network to compute a speaker-specific voice embedding. This embedding could later be used to control a person-specific text-to-speech voice synthesizer.

Voice Cloning

Definition:

Artificial simulation of a person's voice

Use Case, Inference:

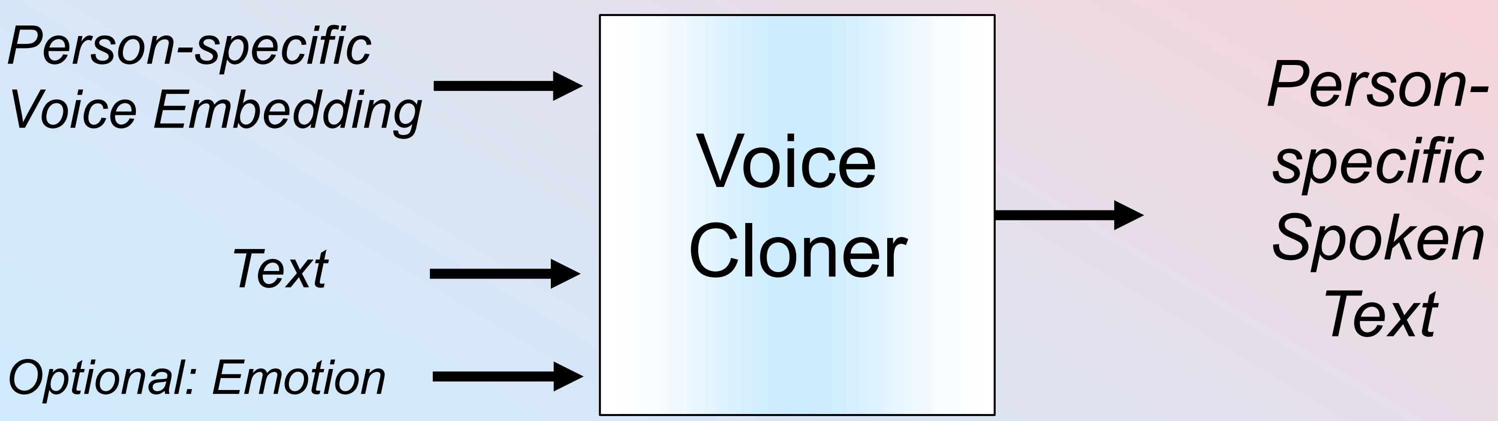


Voice Cloning

Definition:

Artificial simulation of a person's voice

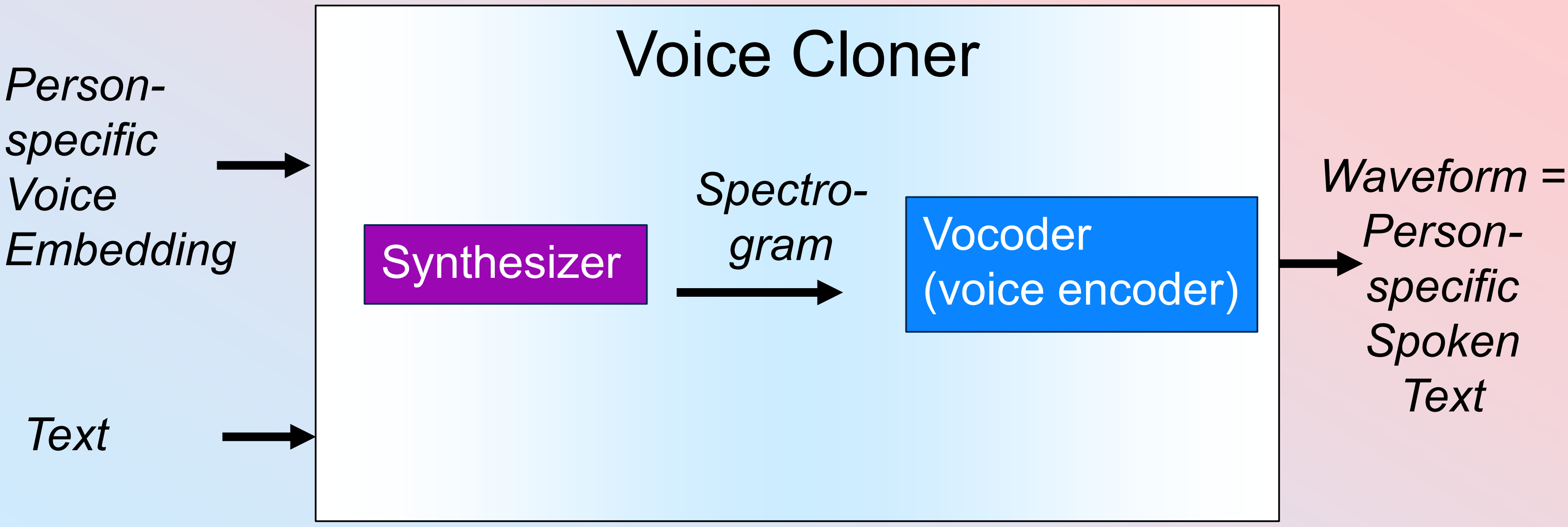
Use Case, Inference:



Voice Cloning

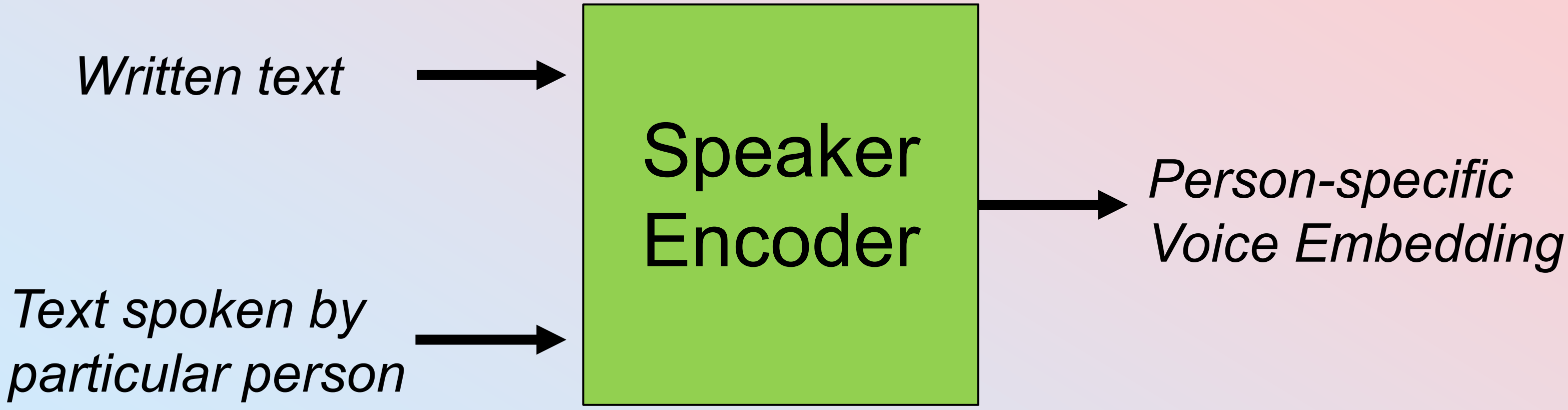
Computer Science

Use Case, Inference:



Voice Cloning

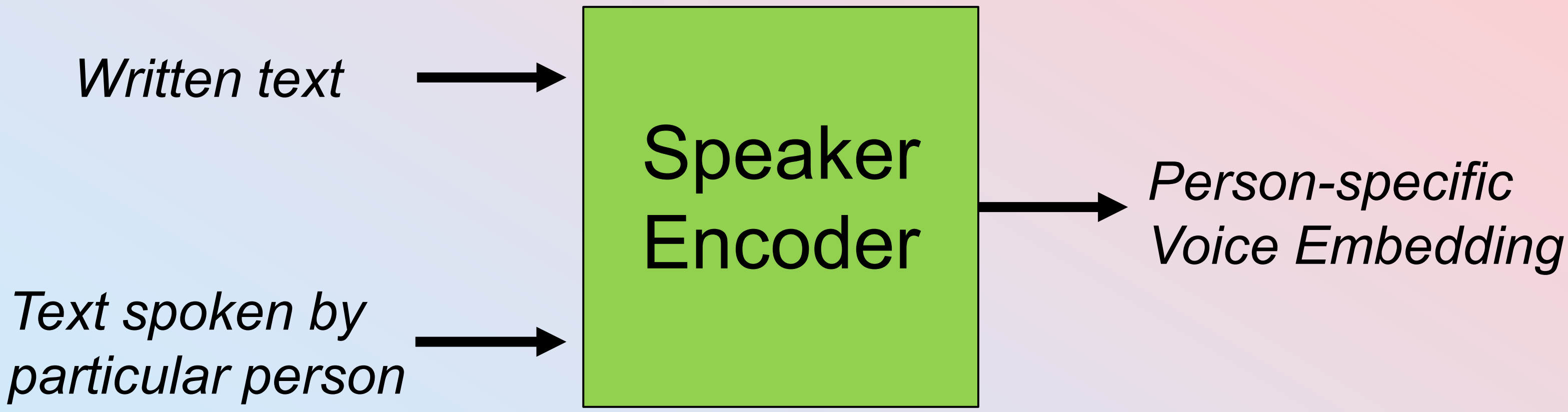
How to obtain a person-specific voice embedding:



Voice Cloning

How to obtain a person-specific voice embedding:

3-4 Hours of



Voice Cloning

How to obtain a person-specific voice embedding:

Short Utterance of

Written text →

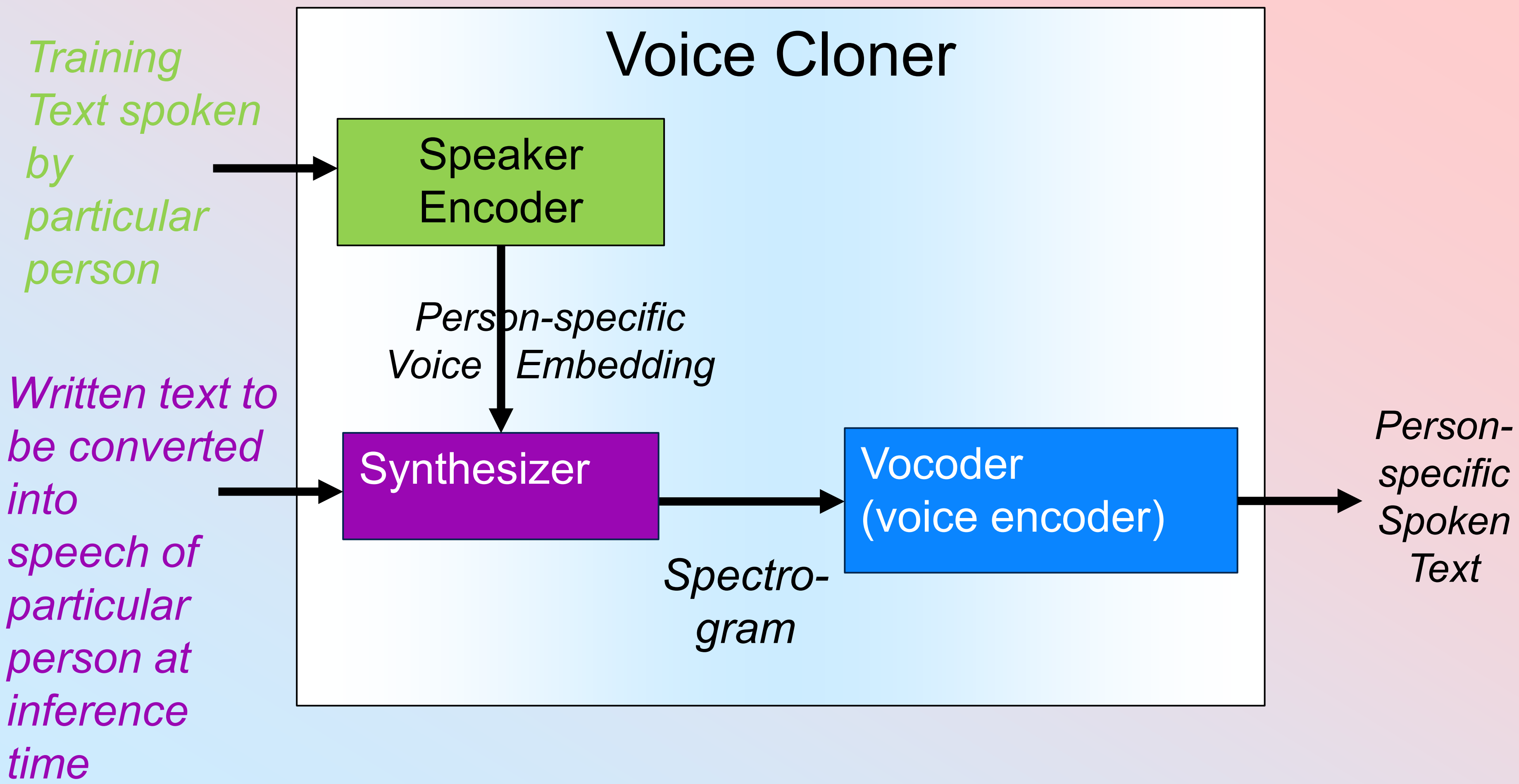
Text spoken by particular person →

Pretrained Speaker Encoder

→ *"Pretty Representative" Person-specific Voice Embedding*

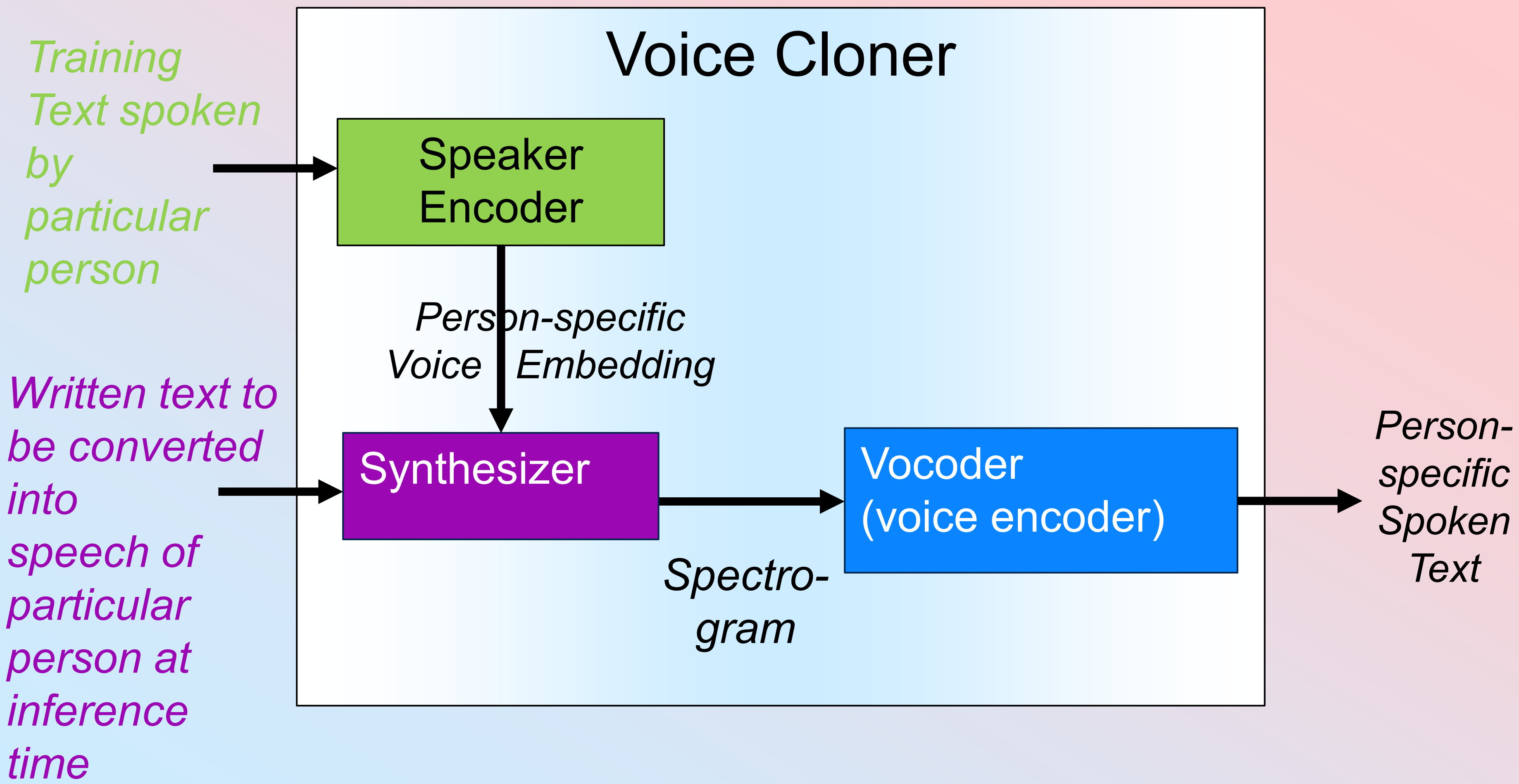
Voice Cloning

Computer Science



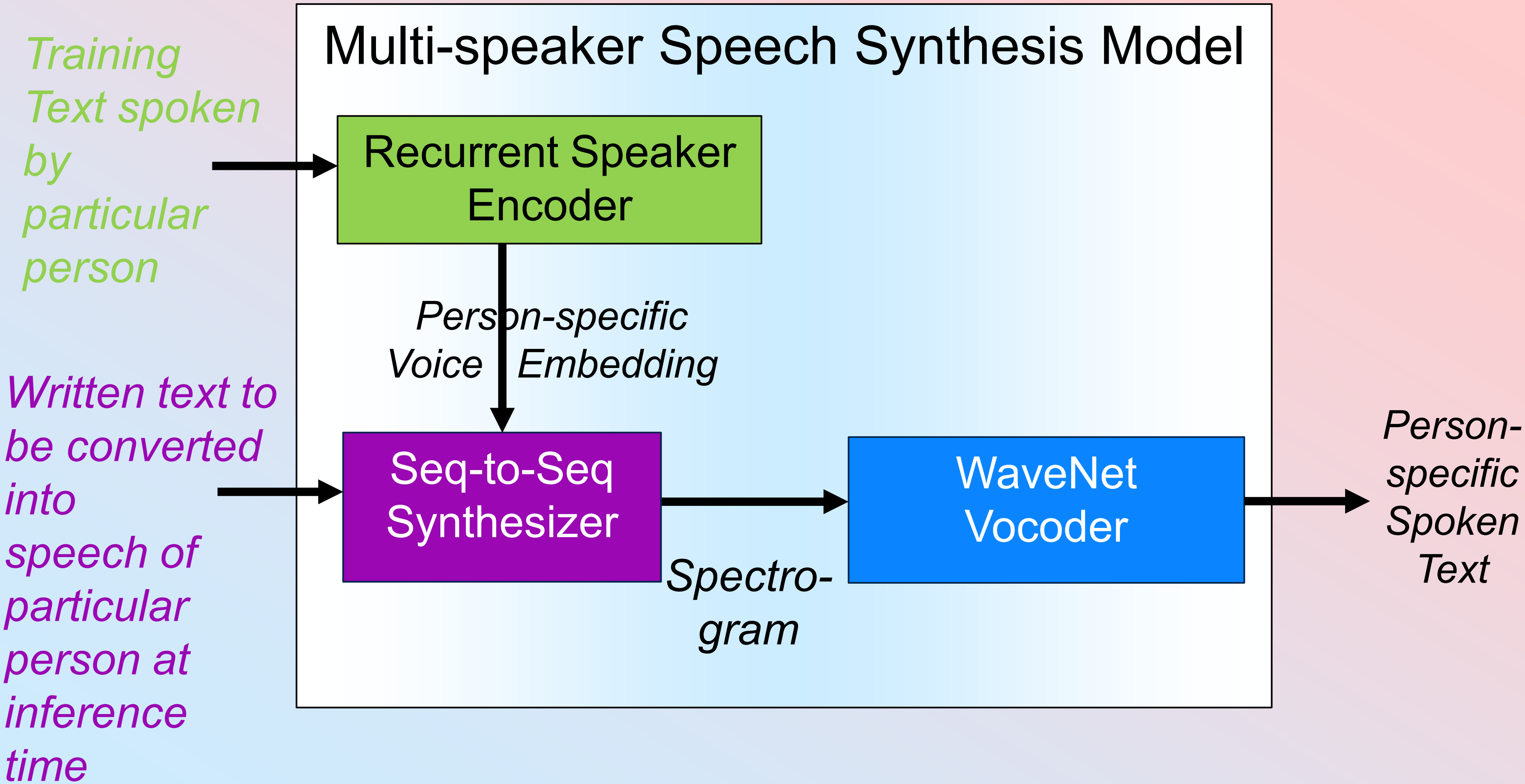
Voice Cloning:
3 Independently trained neural nets

Computer Science



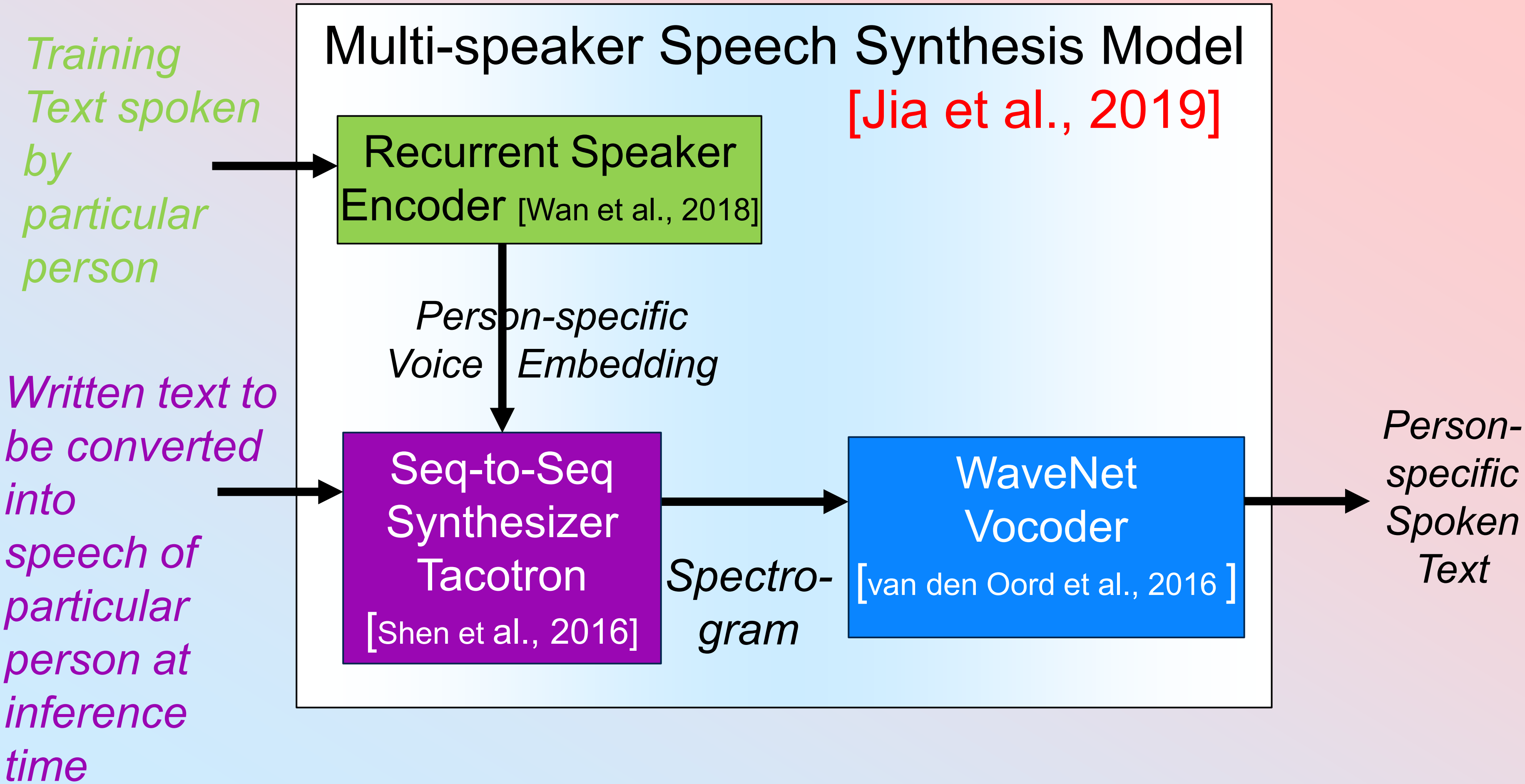
Voice Cloning by Google
3 Independently trained neural nets

Computer Science



Voice Cloning by Google
3 Independently trained neural nets

Computer Science



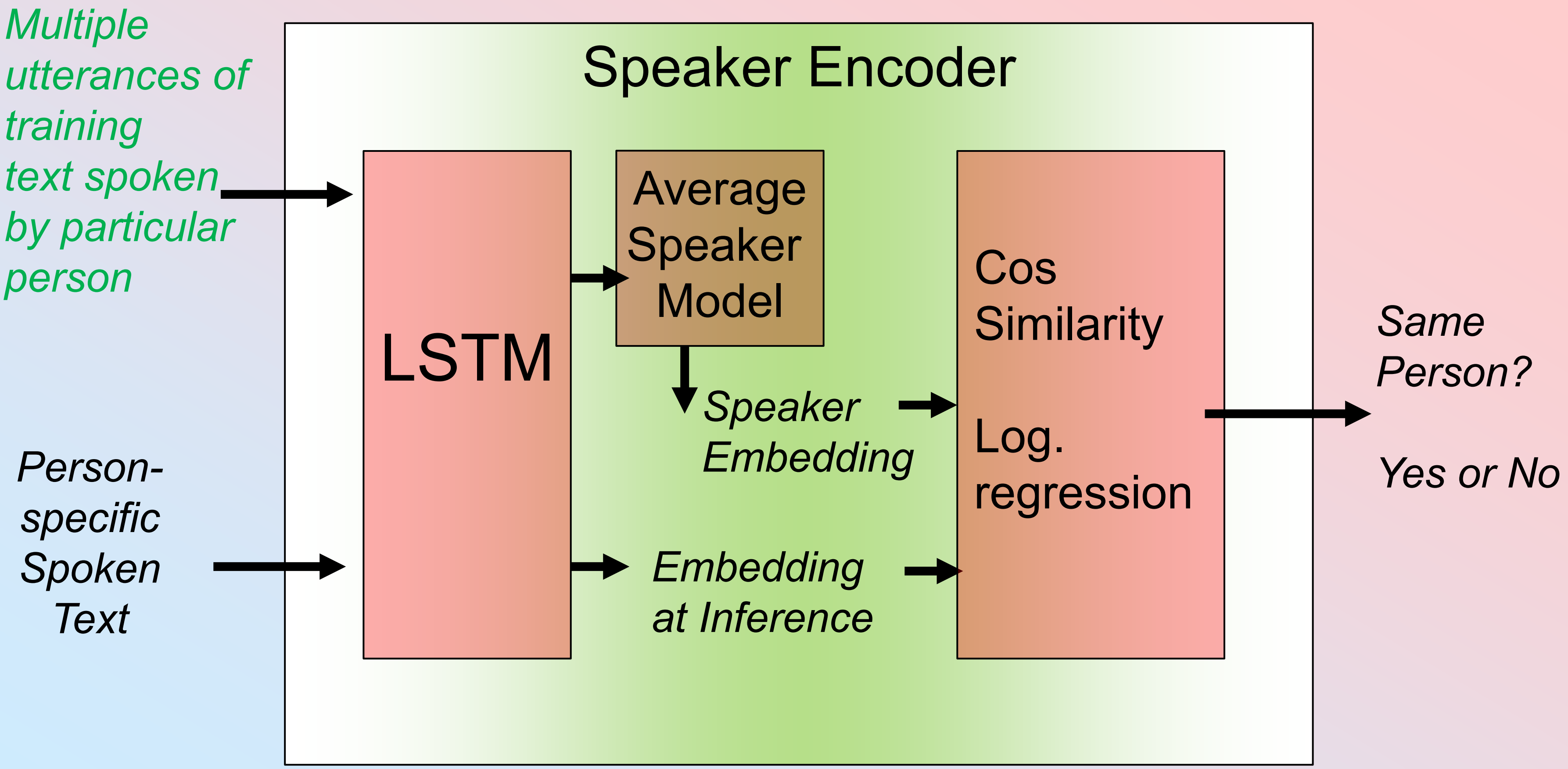
Wan et al., 2018

- 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:

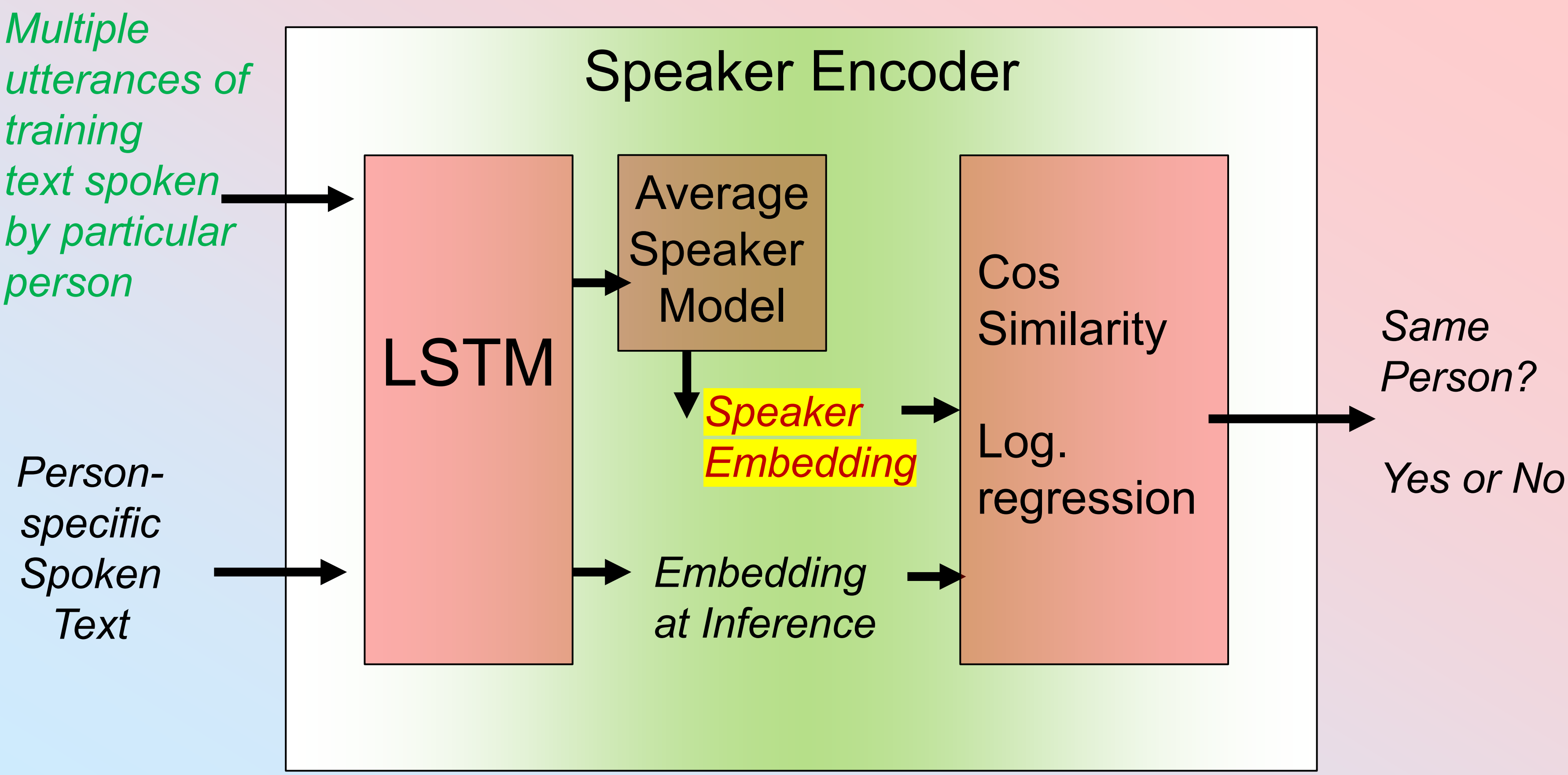
Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2018. https://arxiv.org/pdf/1710.10467.pdf

Previous State-of-the Art of Speaker Verification

Computer Science



Previous State-of-the Art of Speaker Verification



Wang et al.'s Contribution: GE2E Loss function

GE2E uses a similarity matrix $\mathbf{S}_{ji,k}$ that defines the similarities between each embedding $\mathbf{e}_{ji}$ (jth speaker, ith word) and all centroids $\mathbf{c}_k$ (kth speaker) to computed the contrast loss

$$L(\mathbf{e}_{ji}) = 1 - \sigma(\mathbf{S}_{ji,j}) + \max_{\substack{1 \le k \le N \\ k \ne j}} \sigma(\mathbf{S}_{ji,k}),$$

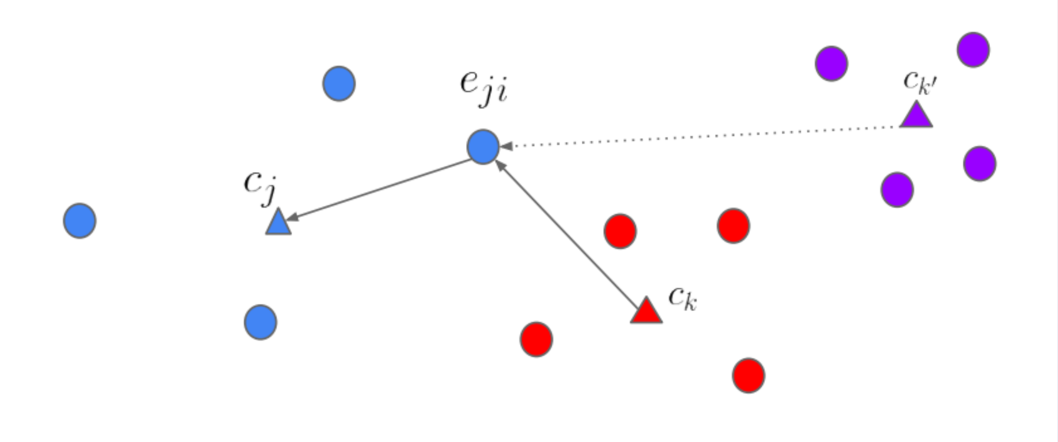
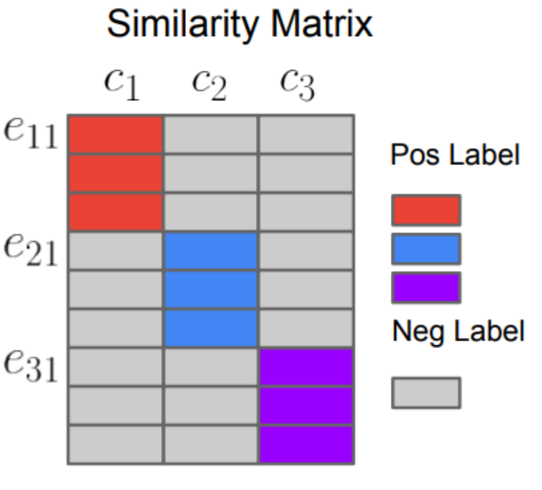


Fig. 2. GE2E loss pushes the embedding towards the centroid of the true speaker, and away from the centroid of the most similar different speaker.



Wang et al.'s Contribution: GE2E Loss function

Computer Science

GE2E uses a similarity matrix $\mathbf{S}_{ji,k}$ that defines the similarities between each embedding $\mathbf{e}_{ji}$ (jth speaker, ith word) and all centroids $\mathbf{c}_k$ (kth speaker) to computed the contrast loss

$$L(\mathbf{e}_{ji}) = 1 - \sigma(\mathbf{S}_{ji,j}) + \max_{\substack{1 \leq k \leq N \\ k \neq j}} \sigma(\mathbf{S}_{ji,k}),$$

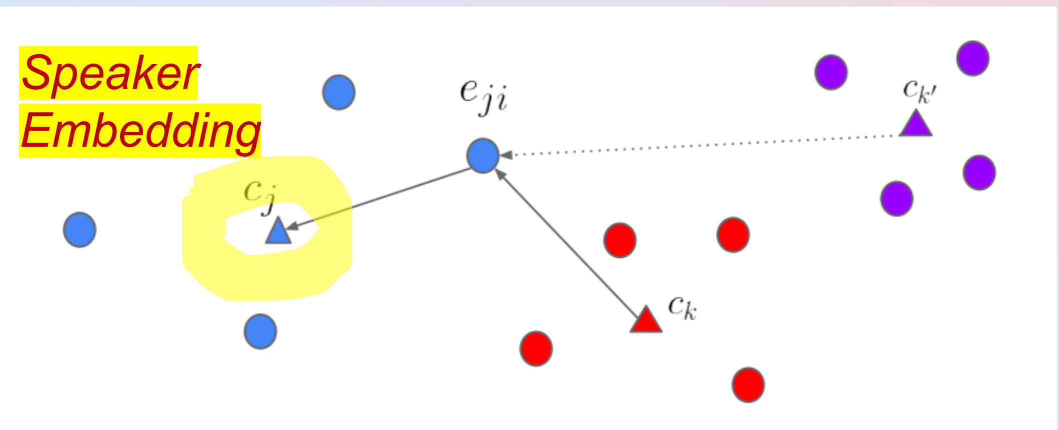
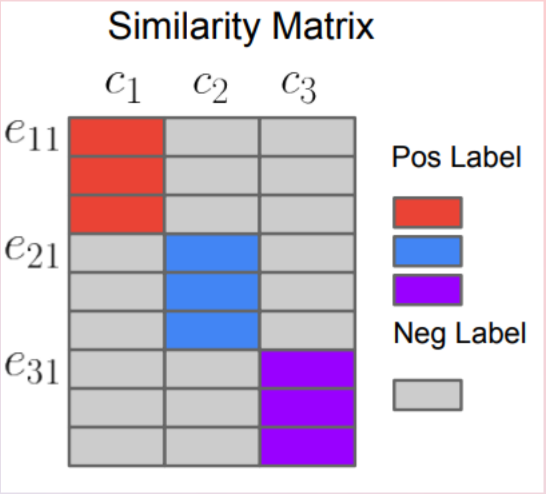
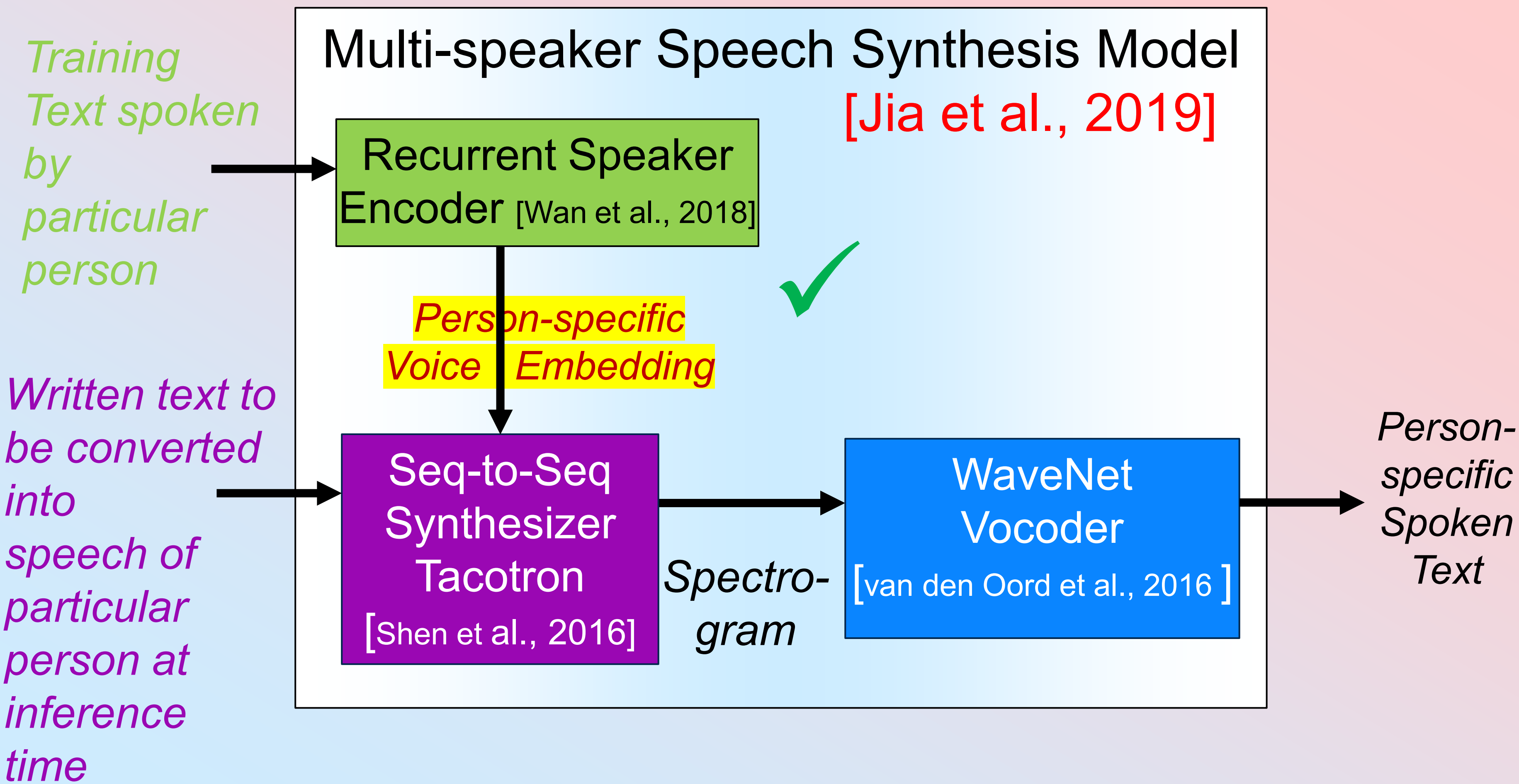


Fig. 2. GE2E loss pushes the embedding towards the centroid of the true speaker, and away from the centroid of the most similar different speaker.



Voice Cloning by Google
3 Independently trained neural nets

Computer Science



van den Oord et al., 2016

Computer Science

- Task: Convert spectrogram into natural-sounding speech signal
- Input: *Spectrogram*
- Output: *Waveform*
- Contribution: Network architecture based on "dilated causal convolutions"
- Publication:

Ääron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. CoRR abs/1609.03499, 2016

van den Oord et al., 2016:
Dilated Causal Convolutional Layers

Computer Science

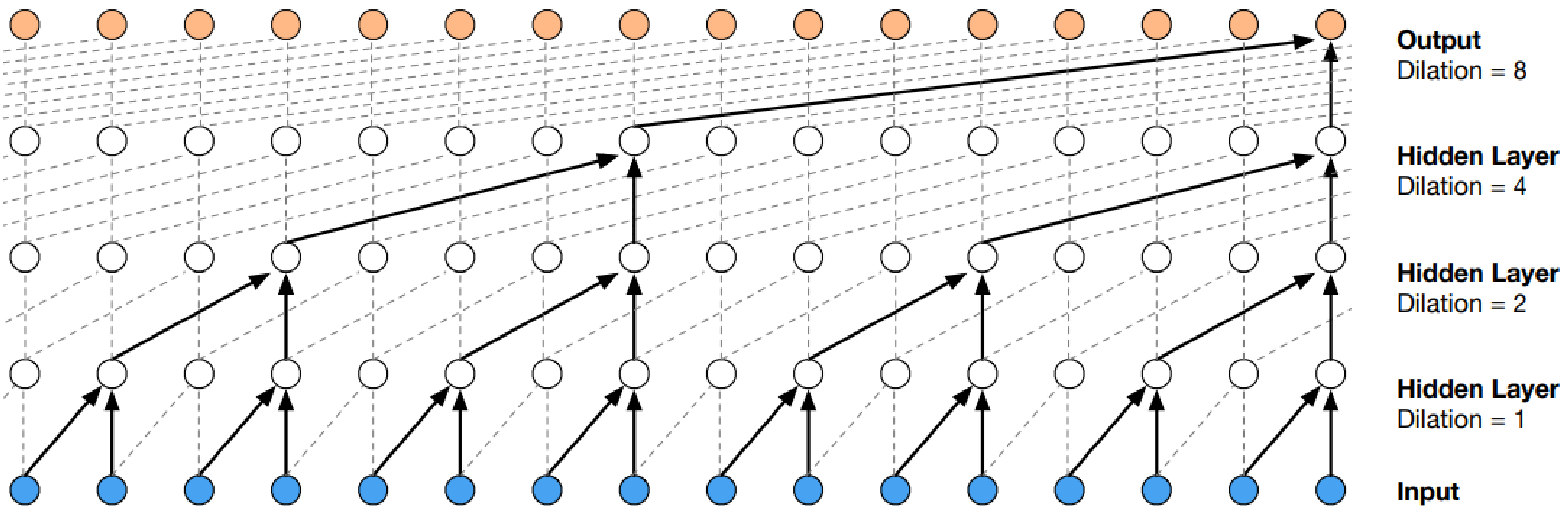
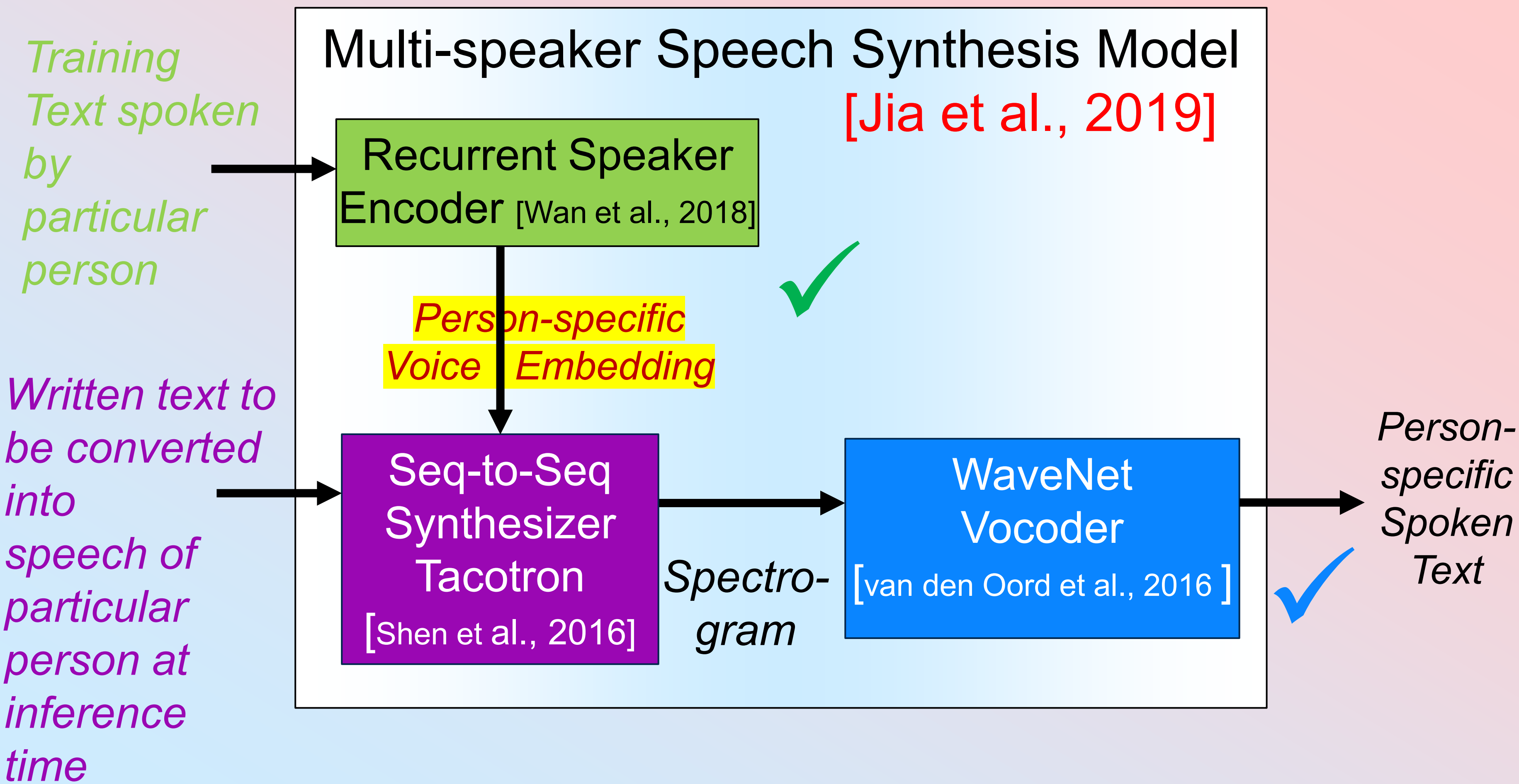


Figure 3: Visualization of a stack of *dilated* causal convolutional layers.

Voice Cloning by Google
3 Independently trained neural nets

Computer Science



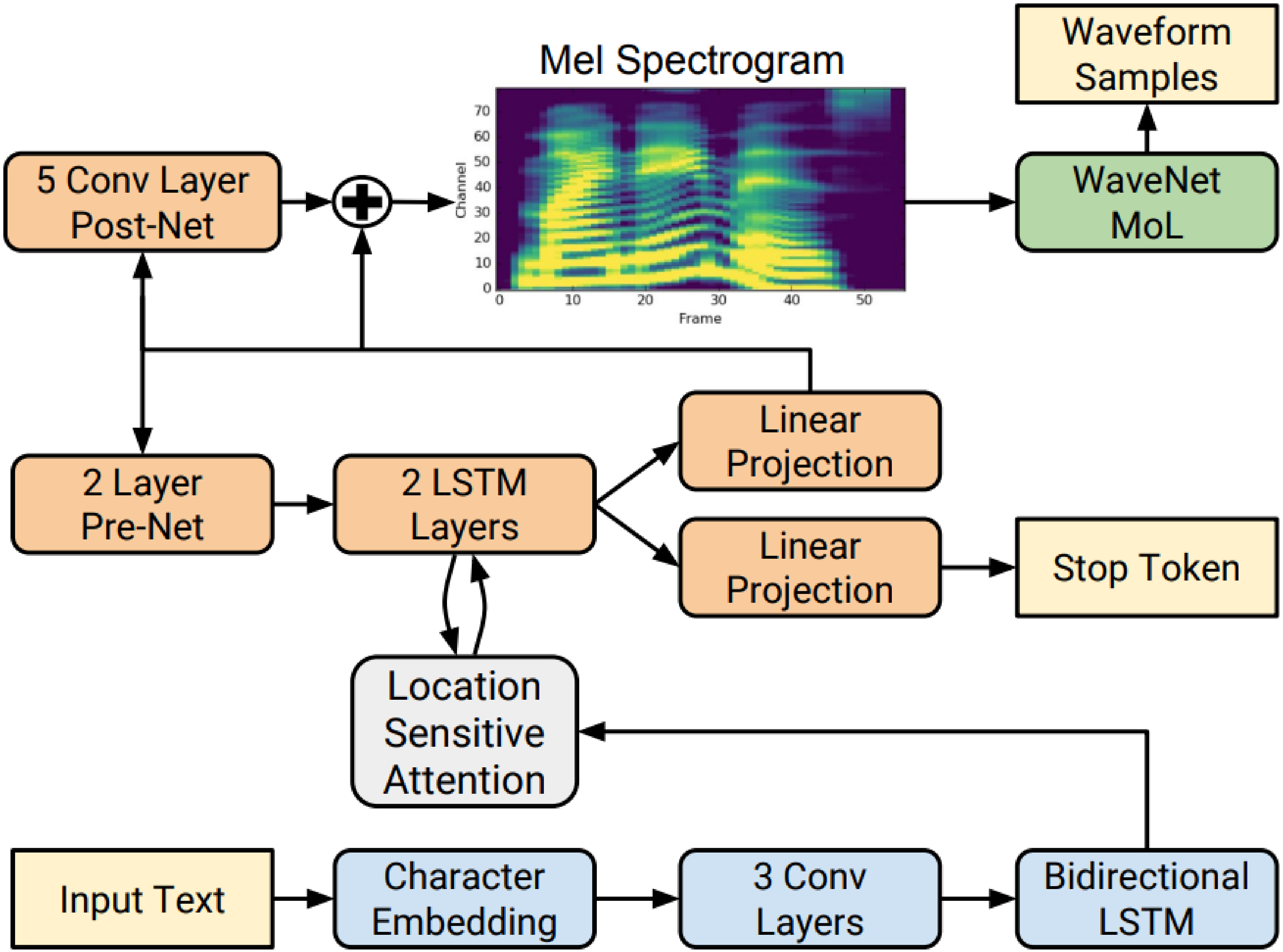
Shen et al., 2016

Computer Science

- 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:

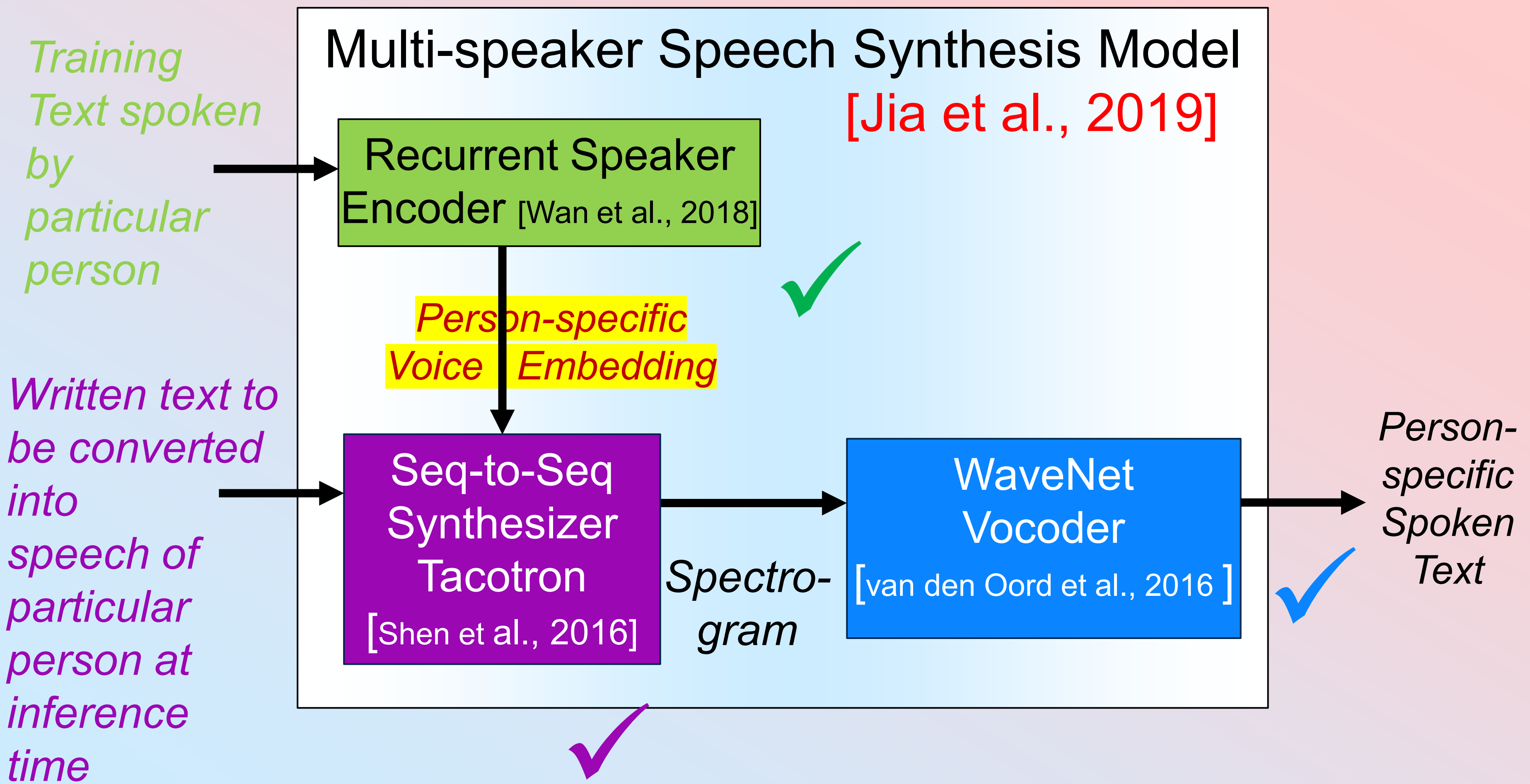
Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui. Wu. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2018.

Shen et al., 2016



Voice Cloning by Google
3 Independently trained neural nets

Computer Science



Jia et al.'s Voice Cloner

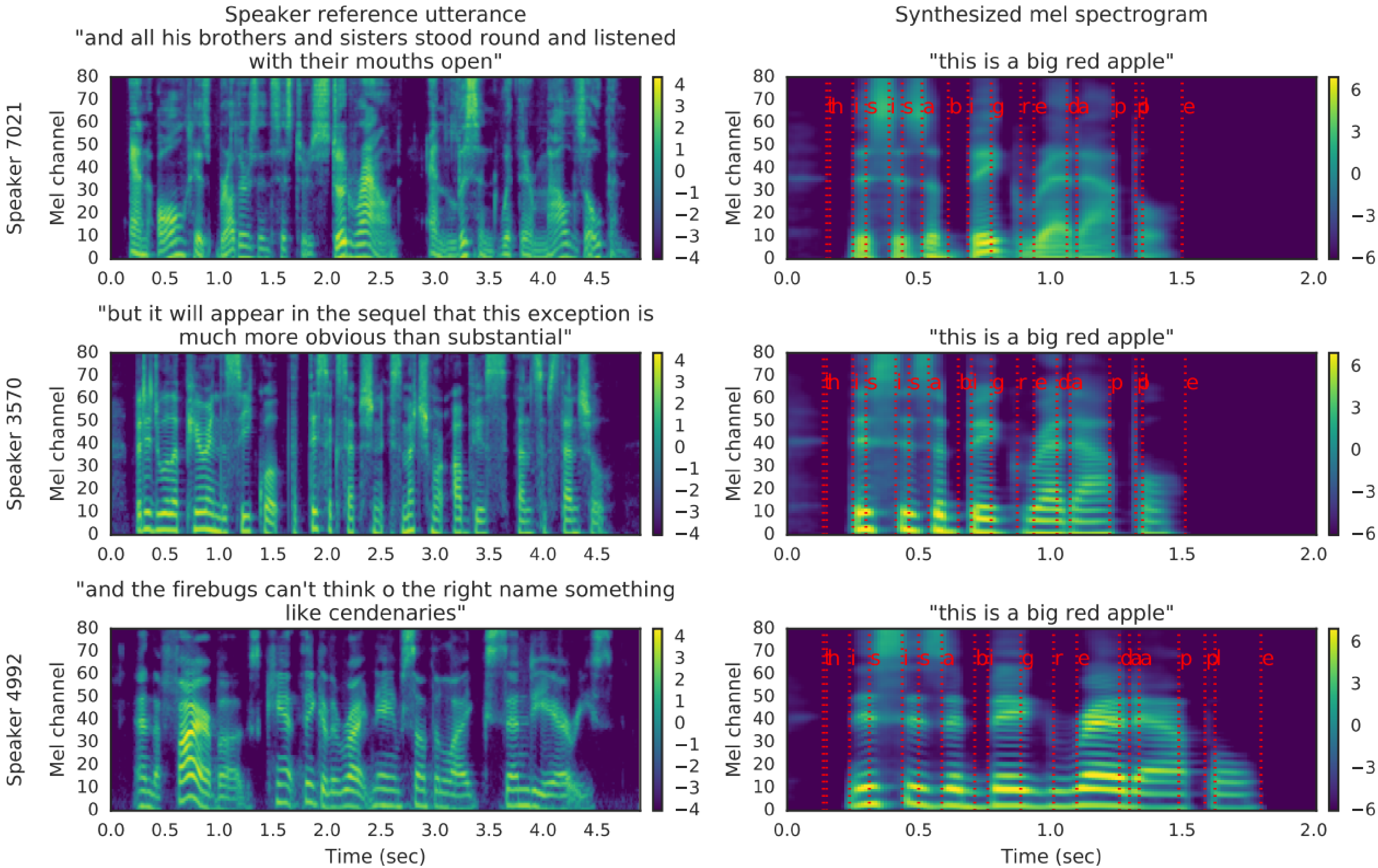


Figure 2: Example synthesis of a sentence in different voices using the proposed system. Mel spectrograms are visualized for reference utterances used to generate speaker embeddings (left), and the corresponding synthesizer outputs (right). The text-to-spectrogram alignment is shown in red. Three speakers held out of the train sets are used: one male (top) and two female (center and bottom).

How to detect voice clones?

Computer Science

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:

- Wang et al., 2020: ASVspoof 2019: a large-scale public database of synthetized, converted and replayed speech. Computer Speech and Language, Vol. 64:101114, https://doi.org/10.1016/j.csl.2020.101114
- Yamagishi et al., 2021: ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. arXiv preprint arXiv:00537

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