

2022.7.15

Shimin Koza



大学共同利用機関法人 情報・システム研究機構

国立情報学研究所

National Institute of Informatics

How do Machines Speak?

Progress and Challenges of Speech Synthesis

Erica Cooper

クーパー・エリカ

Synthesized speech in everyday life

Virtual assistants



Home devices

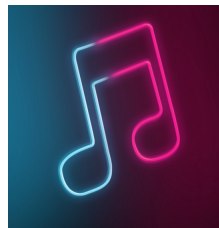


(c)2022 Erica Cooper

GPS devices



Entertainment



https://piapro.net/intl/en_for_creators.html

Speech synthesis can break down barriers

Screen readers for
vision impairments

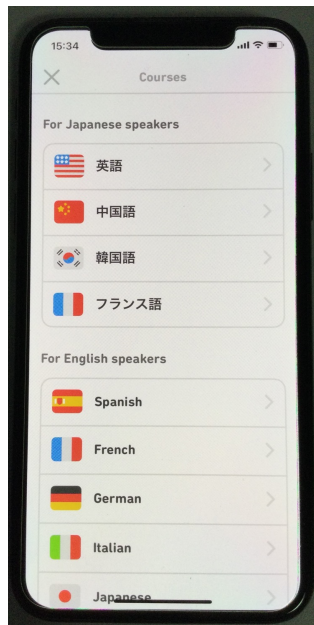


Communication devices
for vocal disabilities



https://en.wikipedia.org/wiki/Stephen_Hawking

Language
learning apps



(c)2022 Erica Cooper

Speech-to-speech
translation



History of speech synthesis



<https://commons.wikimedia.org/wiki/File:Euphoniafaber.jpg>

Joseph Faber's Euphonia
1845

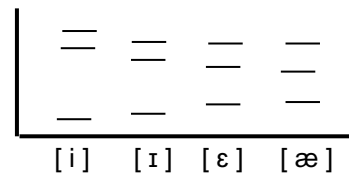


<https://www.youtube.com/watch?v=0rAyrmm7vv0>

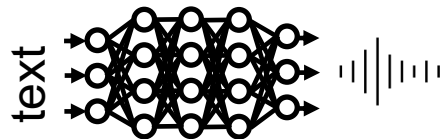
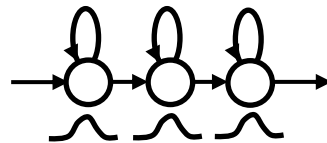
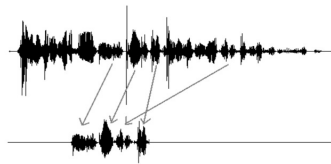
Bell Labs Voder
1939

History of speech synthesis

- **1970s:** Rule-based synthesis to map linguistic units to audio
- **1980s:** Sample-based synthesis to join recordings of the smallest speech units
- **1990s:** Sample-based synthesis using longer recordings
- **2000s:** Machine learning based synthesis learning how to map text into sound from data
- **2010s:** Machine learning based synthesis using neural networks



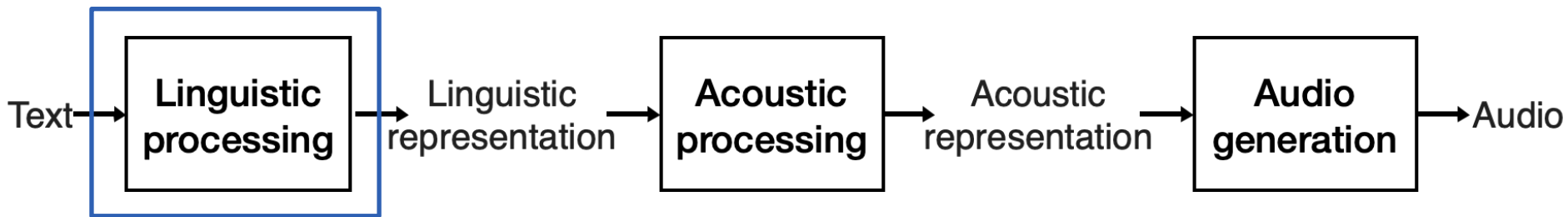
h eh l ow
|---|---|---|



How does it work?



How does it work?



Text normalization

- Requires **real-world** knowledge
- Riverside **Dr.** -> Riverside **Drive**
- **Dr.** Smith -> **Doctor** Smith
- How to expand numbers and abbreviations
 - NASA, SAT
 - 128GB

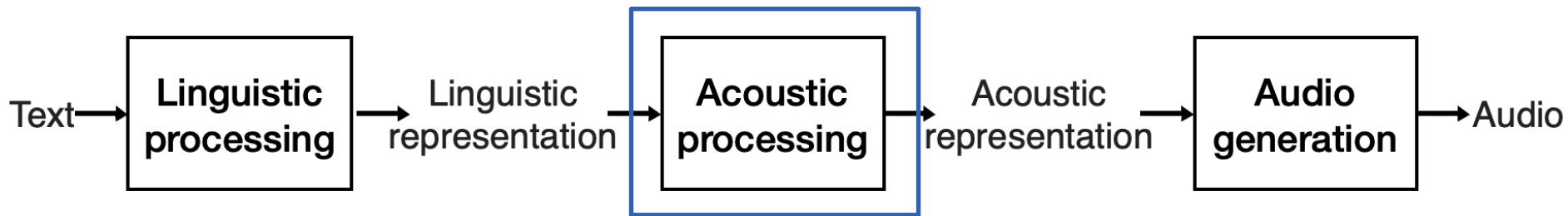
Phonetization

- Requires **linguistic** knowledge

riverside drive	
rɪvərsaɪd draɪv	IPA
R IH V ER S AY D D R AY V	ARPAbet

- Lookup dictionary, handcrafted or learned rules
- Can also include information about stress, timing, and other information in speech

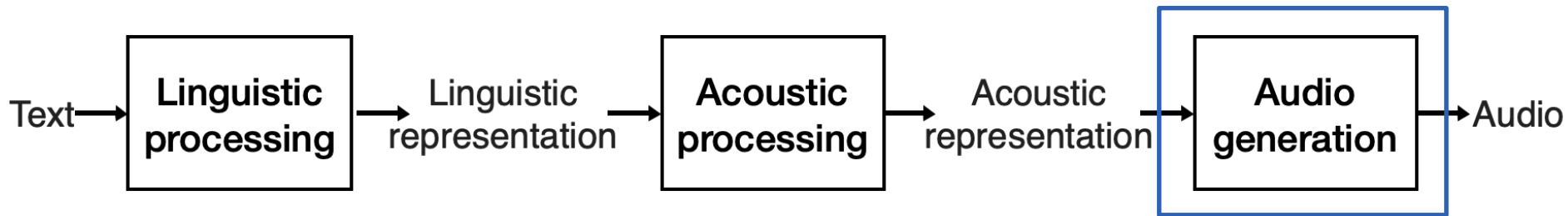
How does it work?



Learn a mapping between the **linguistic representation** created from the text, and an **acoustic representation** of the audio.

The mapping is learned from many examples of **pairs of matching text and audio**.

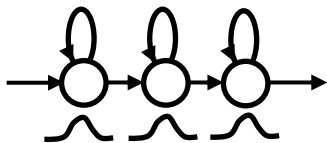
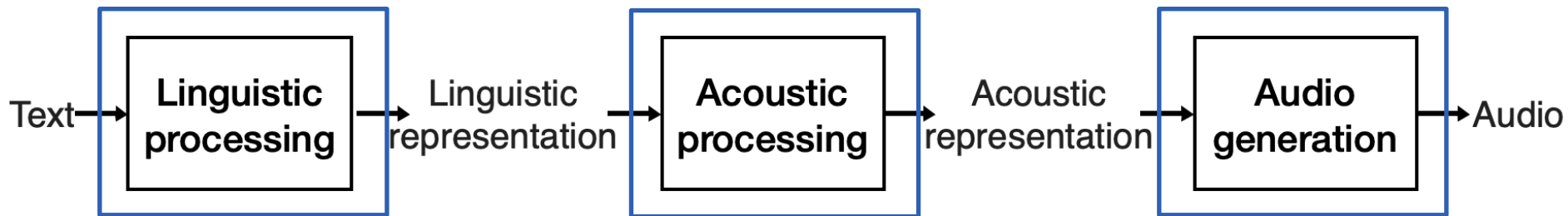
How does it work?



Sometimes an **intermediate representation** of audio is used.

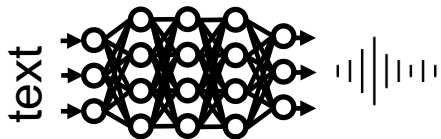
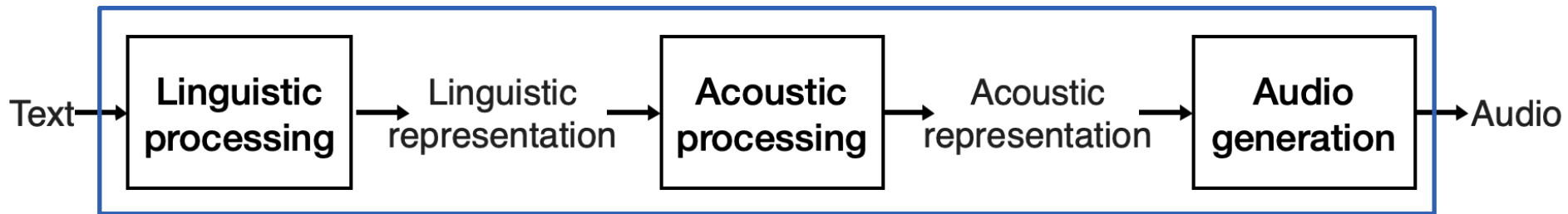
In that case, we finally have to convert from the intermediate representation into **audio**.

How does it work?



In the past, these were all **separately-developed components**.

How does it work?



Modern synthesis methods use **one model** for some or all of these components and learn **multiple steps together**.

Personalized speech synthesis

- Why **personalized** speech synthesis?
 - Assistive technology for the speaking impaired
 - Narration for social media videos and presentations
 - Quick editing and correction of spoken audio

Welcome to my podcast.
Today's episode is a new story.



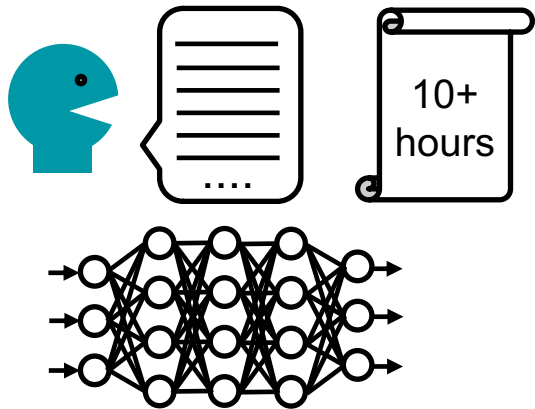
Hello and welcome to my podcast.
Today's episode is **an interesting** story.



Personalized speech synthesis

How can I make a synthesized voice that sounds like me?

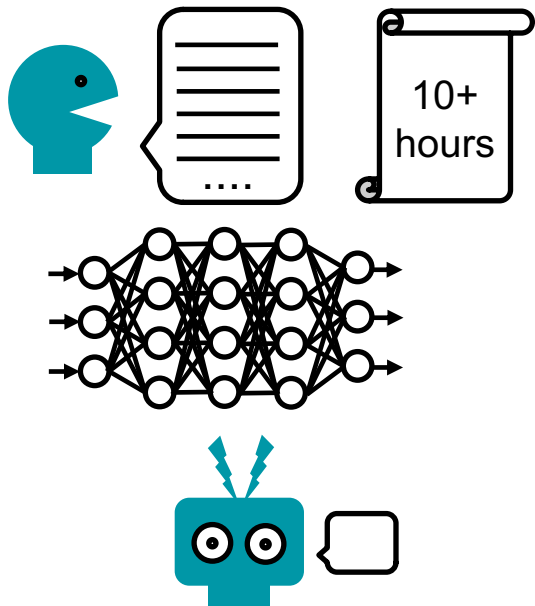
Train a **single speaker** model



Personalized speech synthesis

How can I make a synthesized voice that sounds like me?

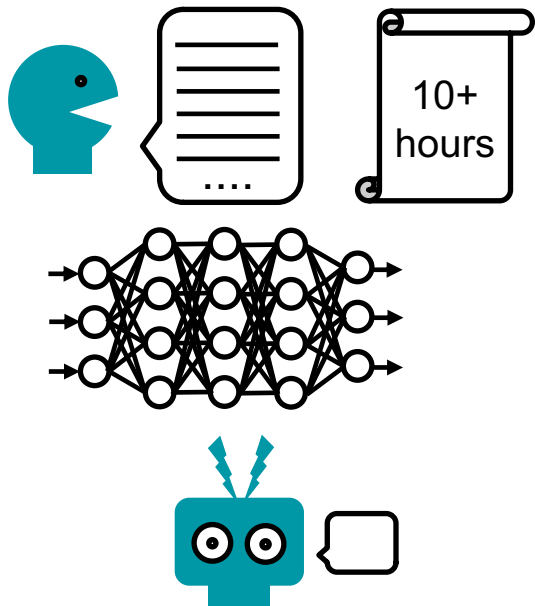
Train a **single speaker** model



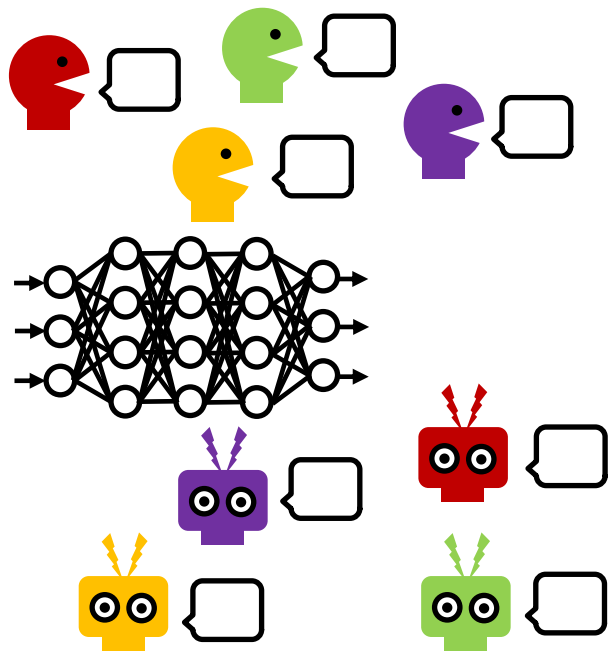
Personalized speech synthesis

How can I make a synthesized voice that sounds like me?

Train a **single speaker** model



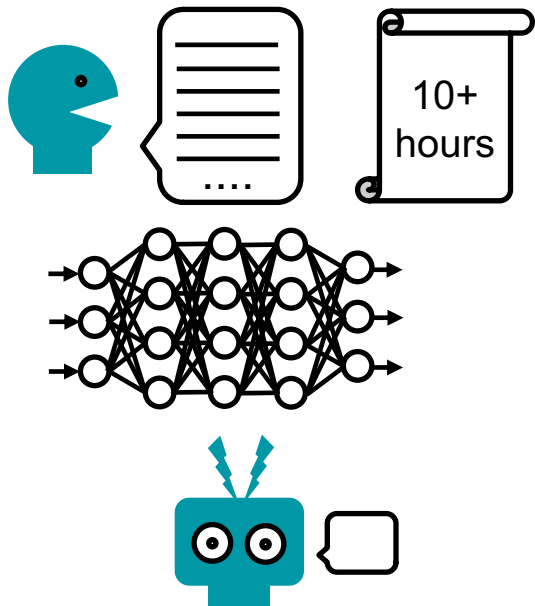
Adapt a **multi-speaker** model



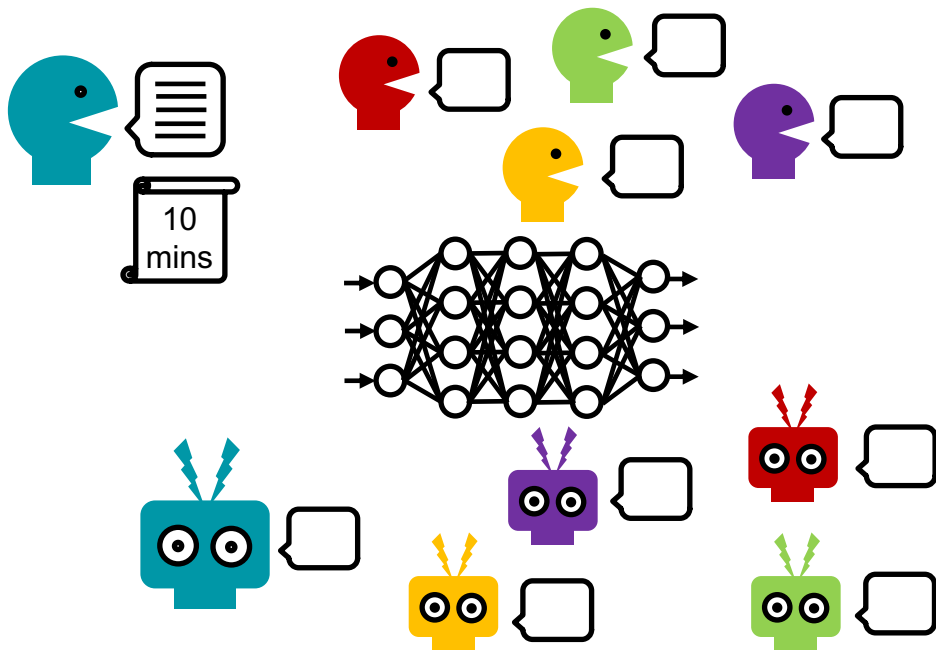
Personalized speech synthesis

How can I make a synthesized voice that sounds like me?

Train a **single speaker** model



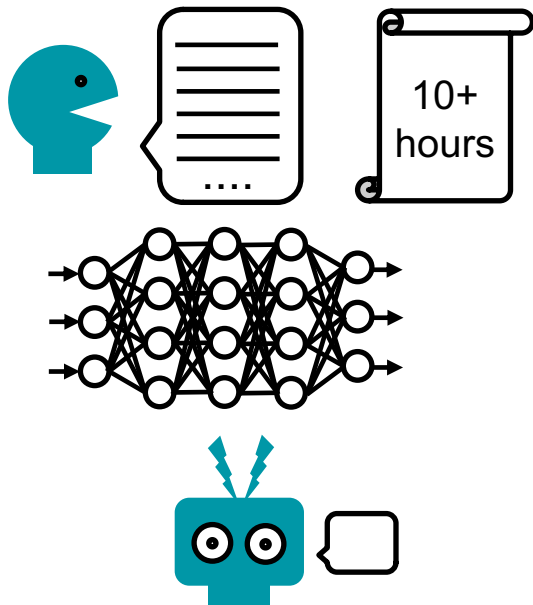
Adapt a **multi-speaker** model



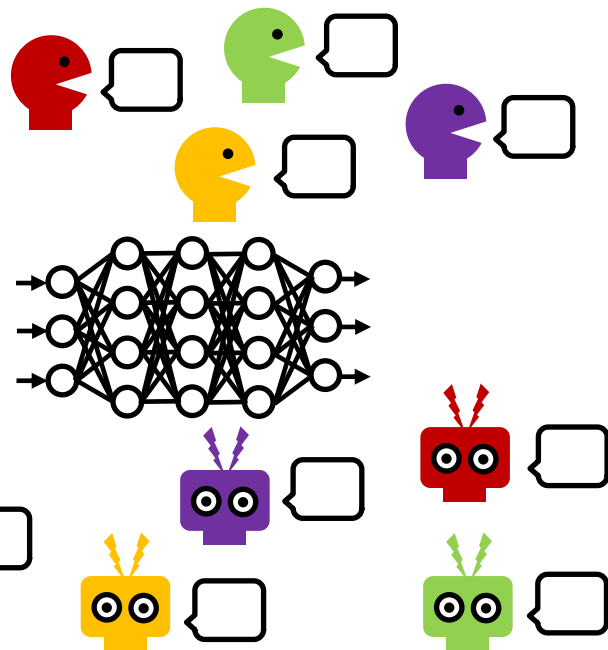
Personalized speech synthesis

How can I make a synthesized voice that sounds like me?

Train a **single speaker** model

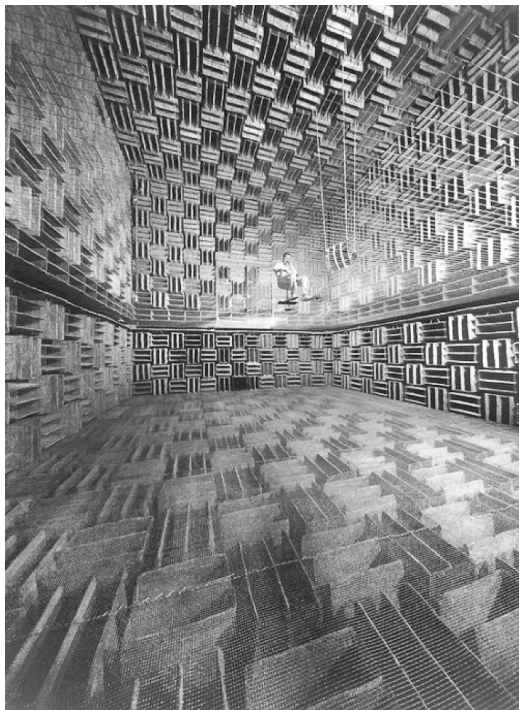


Adapt a **multi-speaker** model



Speech synthesis for low-resource languages

Human-sounding speech synthesis requires dozens of hours of professionally-recorded speech.



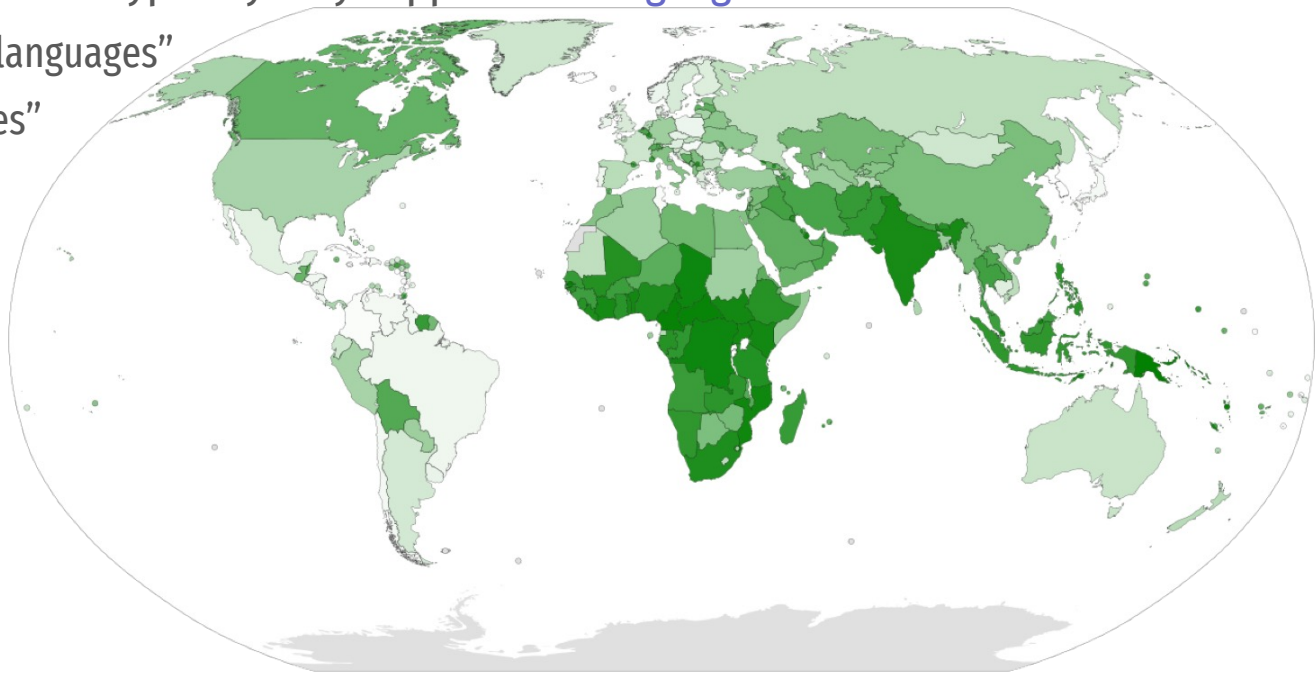
<https://www.curiousspeckle.net/2013/05/13/silent-room-room-for-silence/>

Bell Labs anechoic chamber, 1947

Photo credit: Eric Schaal

Speech synthesis for low-resource languages

- There are more than **7,000 languages** spoken in the world !
- Popular commercial systems typically only support **~30 languages**
 - Apple MacinTalk: “over 30 languages”
 - Amazon Polly: “31 languages”
 - Google: “32 languages and variants”

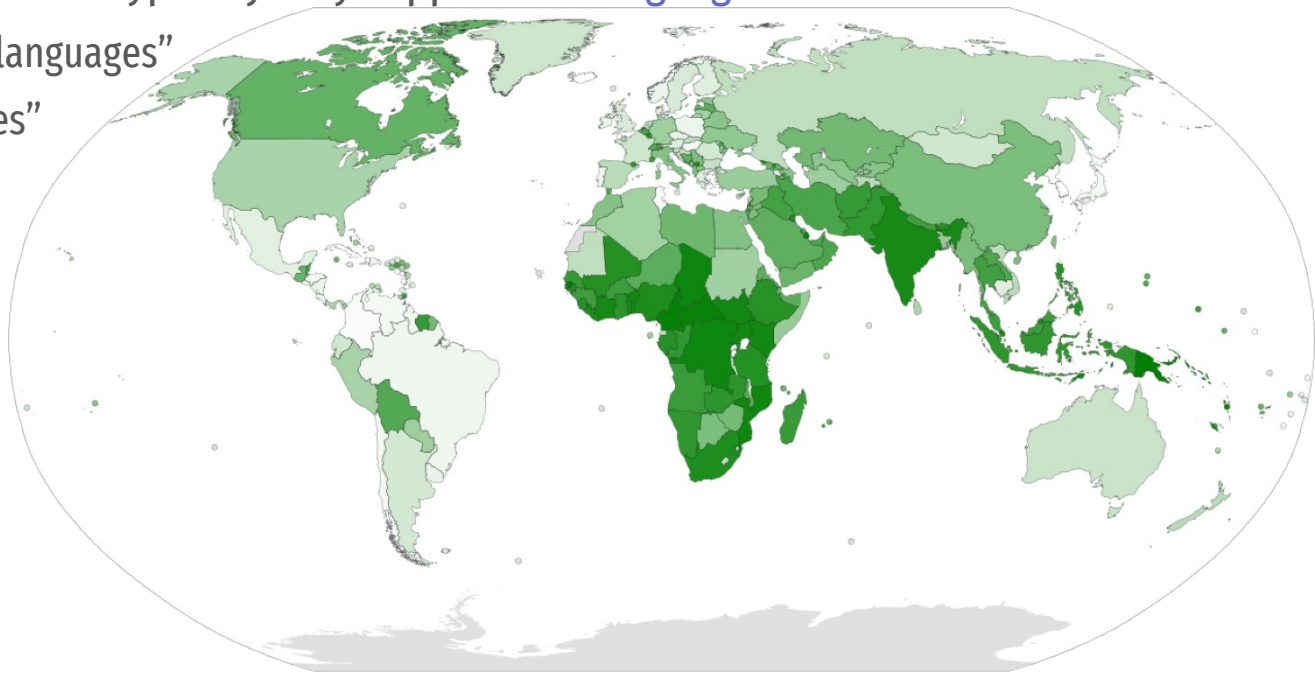


Source: https://en.wikipedia.org/wiki/Linguistic_diversity_index

Speech synthesis for low-resource languages

- There are more than **7,000 languages** spoken in the world !
- Popular commercial systems typically only support **~30 languages**
 - Apple MacinTalk: “over 30 languages”
 - Amazon Polly: “31 languages”
 - Google: “32 languages and variants”

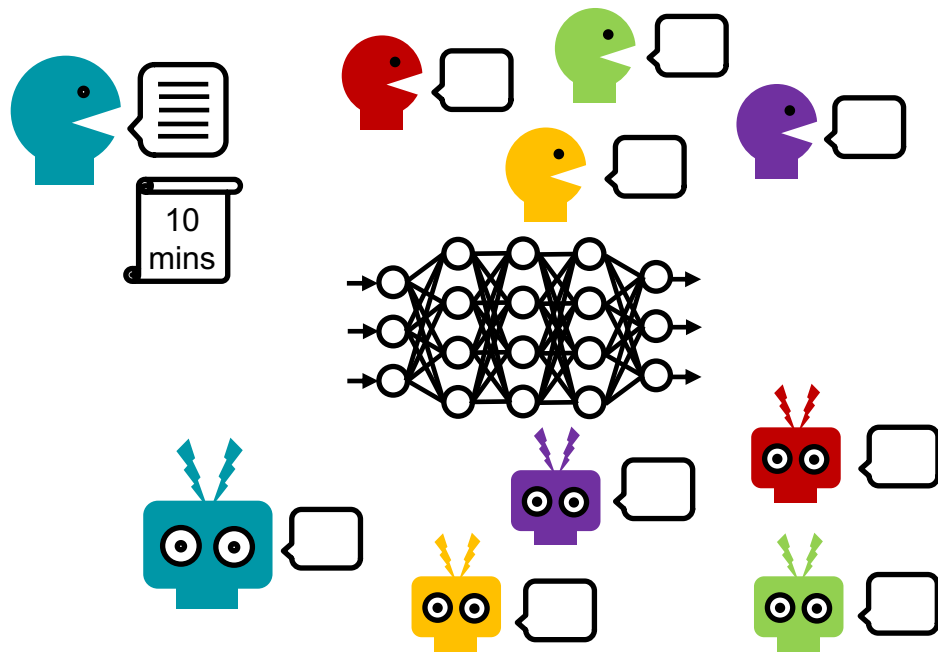
Can we build a speech synthesizer for a new language even if we don't have large amounts of high-quality speech recordings?



Source: https://en.wikipedia.org/wiki/Linguistic_diversity_index

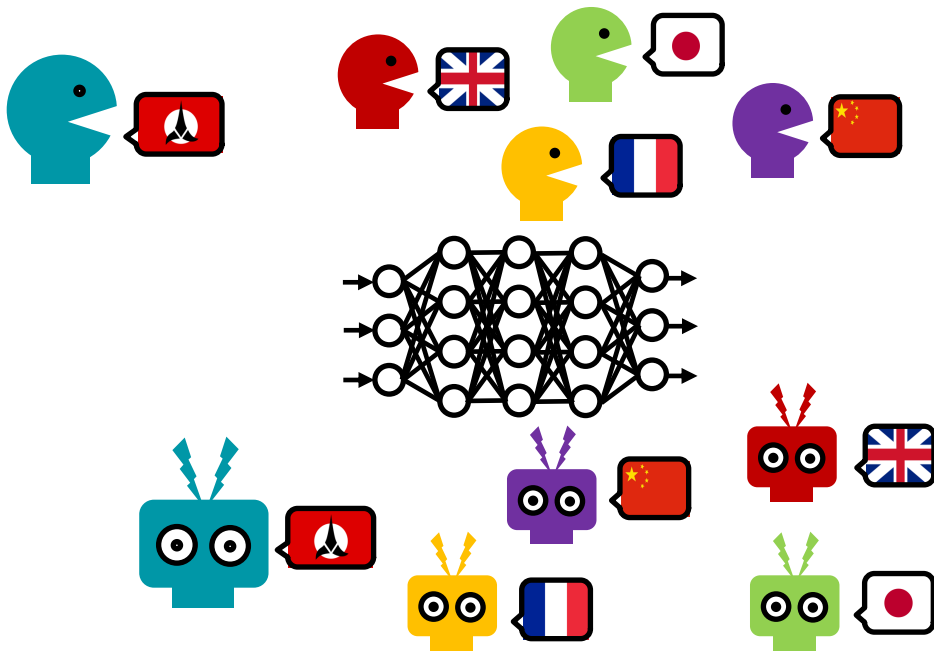
Speech synthesis for low-resource languages

Adapt a **multi-speaker** model



Speech synthesis for low-resource languages

Adapt a **multi-language** model



Deepfakes vs. personalized speech synthesis

- **Same technology, different purposes**
- **Personalized speech synthesis:**
 - For your own personal use
 - To assist your everyday life
 - To create your own content
 - For entertainment
- **Deepfakes:**
 - Created without the person's knowledge
 - Intended to deceive
 - A type of fraud
 - Spread misinformation in society

TECHNOLOGY

An artificial-intelligence first: Voice-mimicking software reportedly used in a major theft



By [Drew Harwell](#)

September 4, 2019 at 6:27 p.m. EDT

<https://www.washingtonpost.com/technology/2019/09/04/an-artificial-intelligence-first-voice-mimicking-software-reportedly-used-major-theft/>



Public Service Announcement

FEDERAL BUREAU OF INVESTIGATION



June 28, 2022

Deepfakes and Stolen PII Utilized to Apply for Remote Work Positions

<https://www.ic3.gov/Media/Y2022/PSA220628>

TECH • SCAMS

Beware: Phone scammers are using this new sci-fi tool to fleece victims

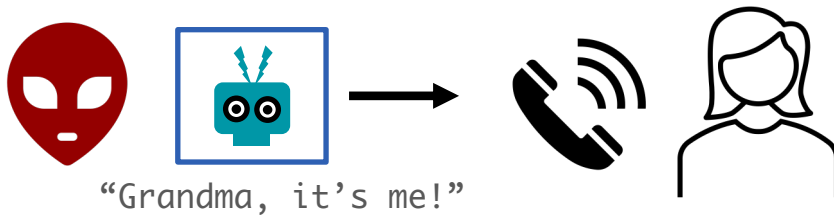
By [JENNIFER ALSEVER](#)

May 5, 2021 8:30 AM GMT+9

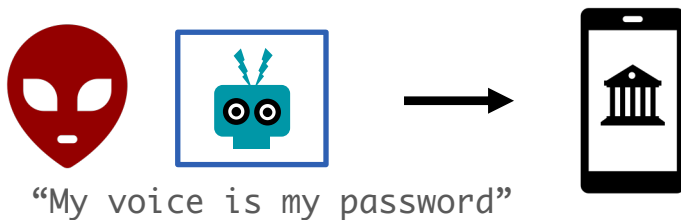
<https://fortune.com/2021/05/04/voice-cloning-fraud-ai-deepfakes-phone-scams/>

Deepfakes and spoofing

- Synthesized speech can fool **humans**



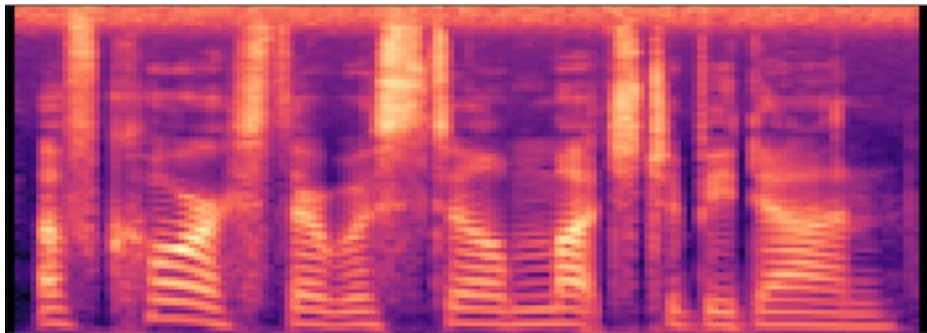
- Synthesized speech can fool **machines**



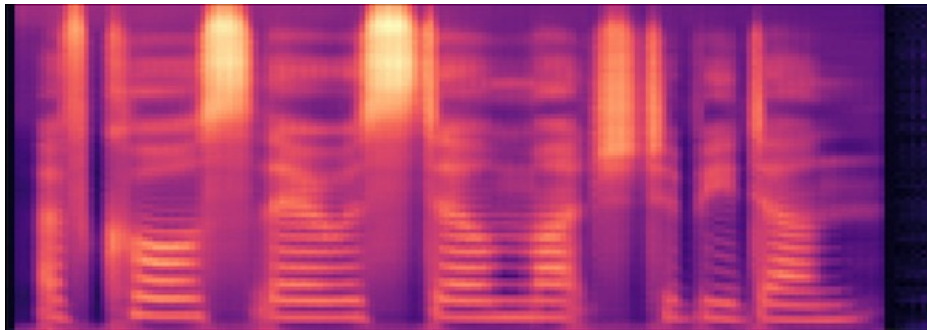
Deepfakes and spoofing

- Synthesized speech can be **detected**

Real speech

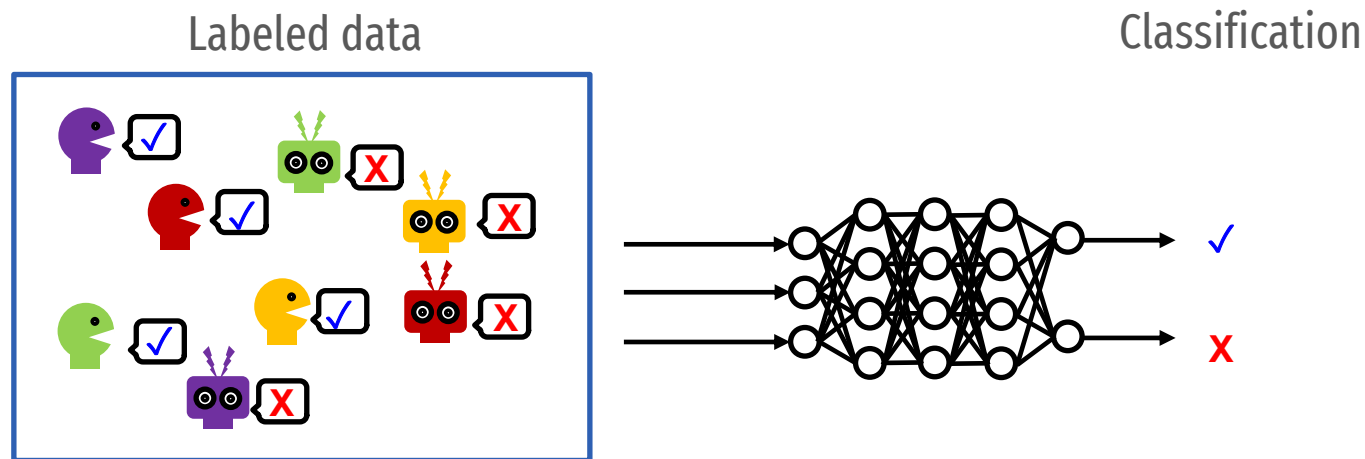


Synthesized speech



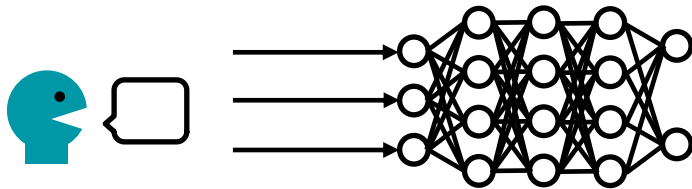
Deepfakes and spoofing

- Synthesized vs. real speech can be **learned from data by machine learning**



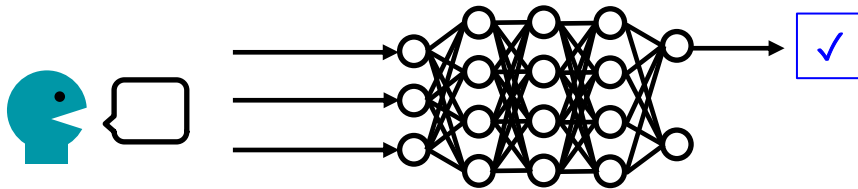
Deepfakes and spoofing

- Synthesized vs. real speech can be **learned from data by machine learning**



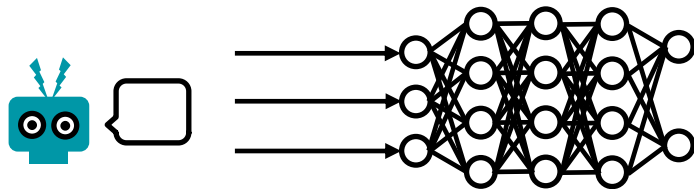
Deepfakes and spoofing

- Synthesized vs. real speech can be **learned from data by machine learning**



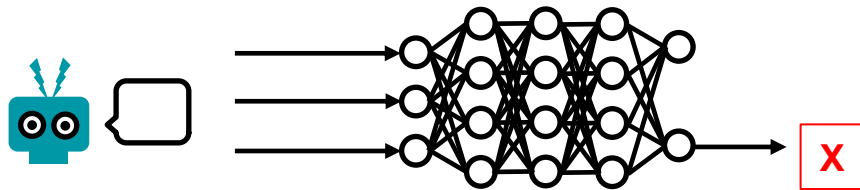
Deepfakes and spoofing

- Synthesized vs. real speech can be **learned from data by machine learning**



Deepfakes and spoofing

- Synthesized vs. real speech can be **learned from data by machine learning**



Speech Synthesis: open research problems

- Spoofing and deepfake detection
 - Humans and machines perceive speech differently and make different kinds of errors
 - Speech synthesis is always improving, so detectors need to keep up
 - Detection of new, unseen synthesizers
- Speech synthesis for low-resource languages
 - How to make it as easy as possible to build a speech synthesizer for a new language
- Expressive speech synthesis
 - How to synthesize in styles other than neutral
 - How to make synthesized speech that is entertaining to listen to
- Faster and more lightweight speech synthesis
 - Neural network models require large computational resources