

2022年度
市民講座



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

機械はどのように話すのか 音声合成の進展とその課題

Erica Cooper
クーパー・エリカ

日常生活における合成音声

仮想アシスタント



ホームデバイス

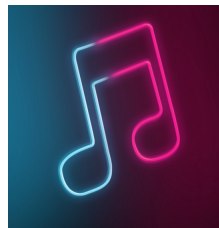


(c)2022 Erica Cooper

GPSデバイス



エンタメ



音声合成でさまざまな壁を取り払う

視覚障がい者向けの
スクリーンリーダー

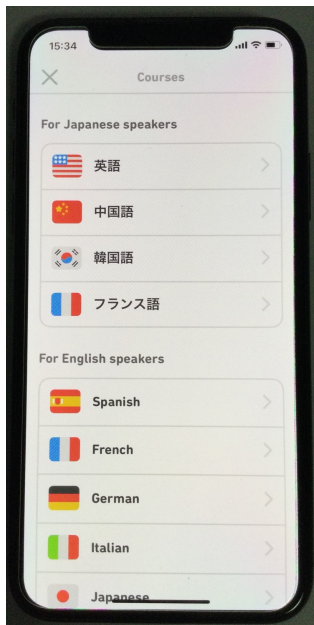


発声障がい者向けの
コミュニケーション機器



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

言語学習アプリ



(c)2022 Erica Cooper

音声翻訳



音声合成の歴史



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

ジョセフ・フェイバーのユーフォニア
1845年

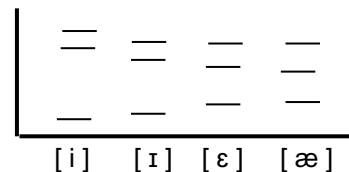


<https://www.youtube.com/watch?v=0rAyrmm7vv0>

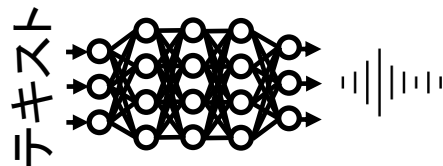
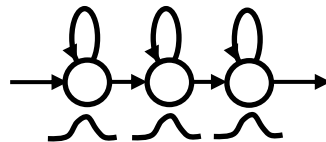
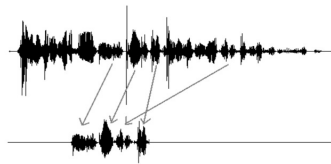
ベル研究所のヴォーダー
1939年

音声合成の歴史

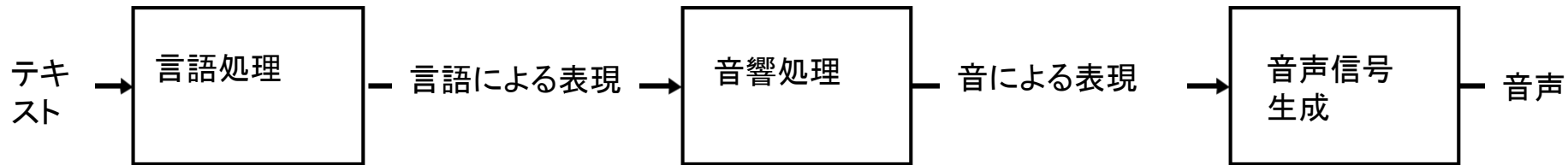
- 1970年代: 言語単位を音声にマッピングする
ルールベースの合成
- 1980年代: 最小の音声単位の録音をつなぎ合わせる
サンプルベースの合成
- 1990年代: 録音をより長くしたサンプルベースの合成
- 2000年代: データから抽出したテキストを音声にマッピング
する方法を学習する機械学習ベースの合成
- 2010年代: ニューラルネットワークを用いた機械学習
ベースの合成



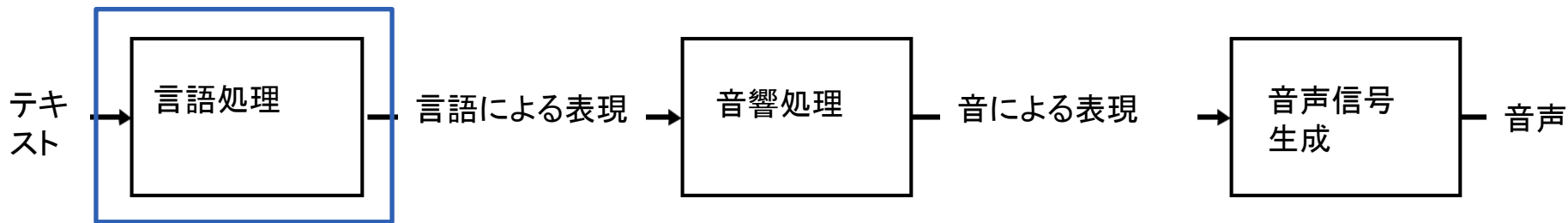
h eh l ow
|---|---|



音声合成の仕組み



音声合成の仕組み



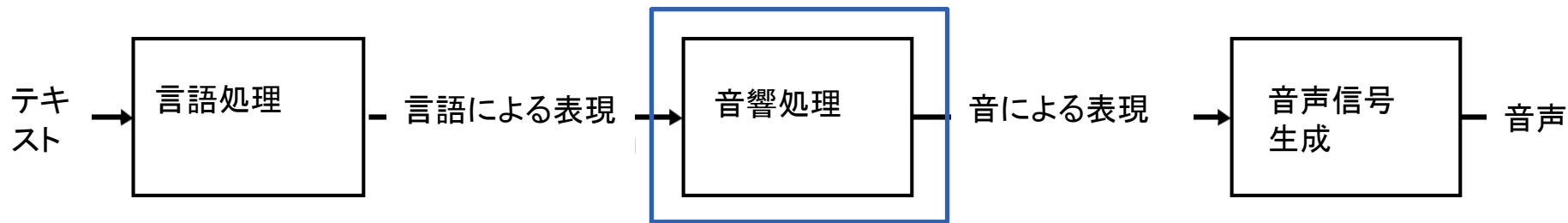
テキストの正規化

- 実世界の知識が必要
- Riverside Dr. → Riverside Drive
- Dr. Smith → Doctor Smith
- 数字と略語をどう展開するか
 - NASA、SAT
 - 128GB

音素による表記

- 言語の知識が必要
 - riverside drive
 - riversaid draiv
 - R IH V ER S AY D D R AY V
 - IPA
 - ARPabet
- 検索性辞書、手作りルールまたは学習済みルール
- 発話時のアクセント、タイミングなどの情報も含まれる

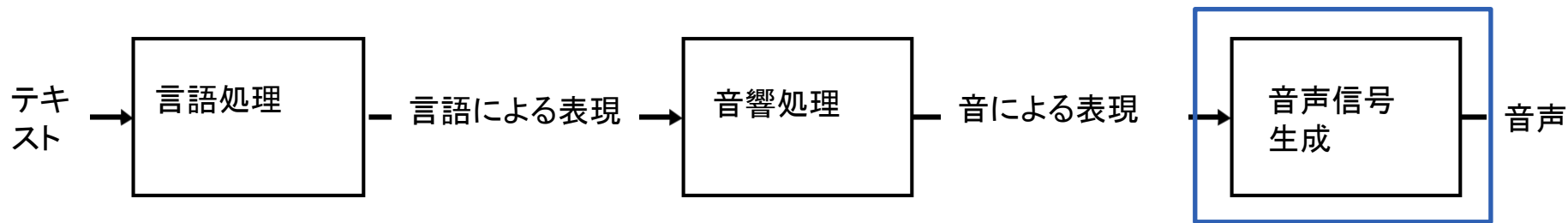
音声合成の仕組み



テキストから作成された言語による表現と
音声の音による表現のマッピングを学習。

このマッピングは、対応するテキストと音声の
ペアの例を多数蓄積することで学習される。

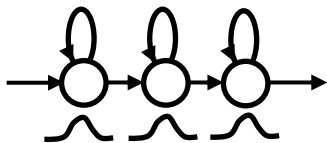
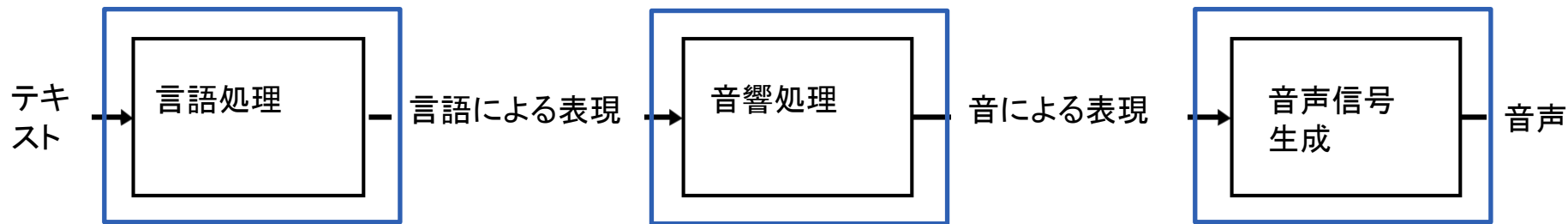
音声合成の仕組み



音声の中間表現が使用されることもある。

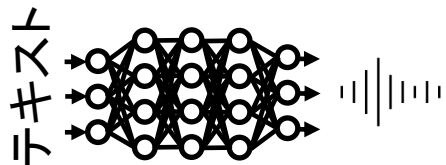
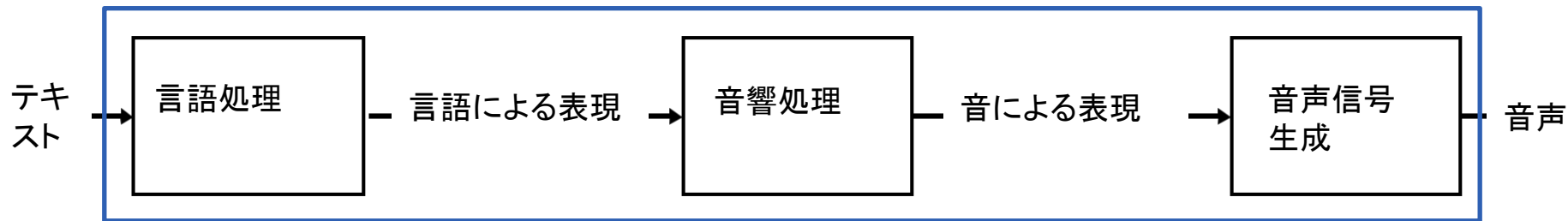
その場合、最終的に中間表現から音声に変換する必要がある。

音声合成の仕組み



従来、これらすべての構成要素は
個別に開発されていた。

音声合成の仕組み



最新の合成方式では、これらの構成要素の一部またはすべてに**1つのモデル**を使用し、**複数のステップをまとめて学習**。

音声合成のパーソナライズ

- なぜ音声合成をパーソナライズするのか？
 - 発話障がい者のための支援技術
 - ソーシャルメディアでの動画やプレゼンテーションのナレーション
 - 話し言葉の音声素早く編集、修正

ポッドキャストへようこそ。
今日は新しい話をします。



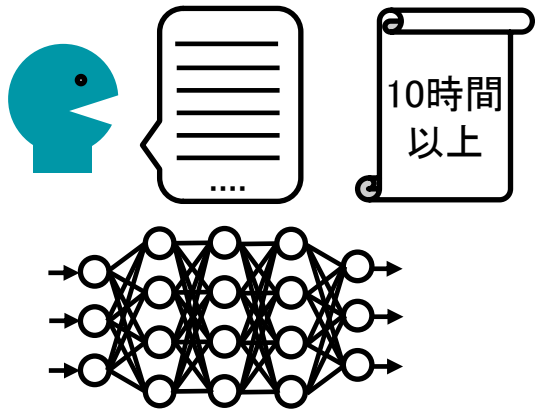
こんにちは、ポッドキャストへようこそ。
今日は面白い話をします。



音声合成のパーソナライズ

自分が話しているような合成音声を作るにはどうすればよいか？

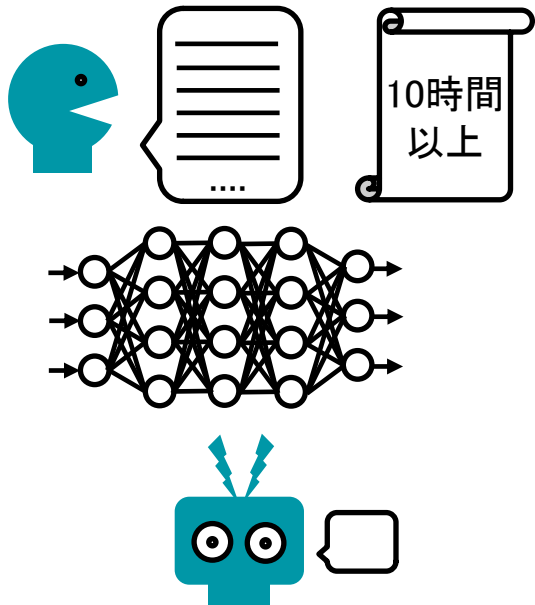
一人の話者モデルを学習



音声合成のパーソナライズ

自分が話しているような合成音声を作るにはどうすればよいか？

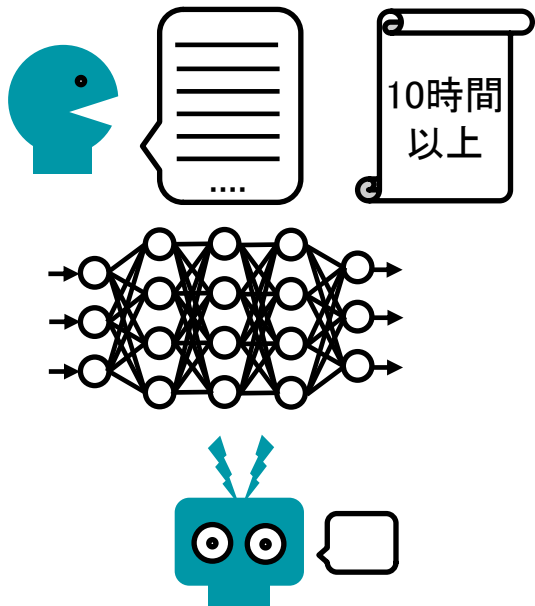
一人の話者モデルを学習



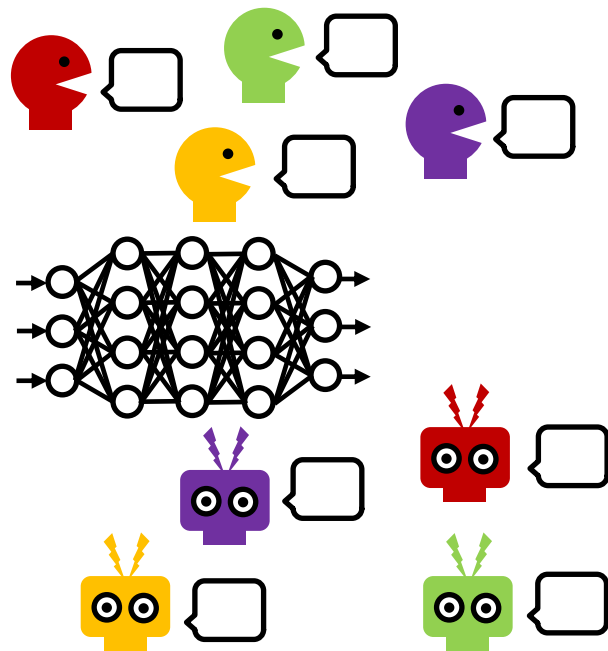
音声合成のパーソナライズ

自分が話しているような合成音声を作るにはどうすればよいか？

一人の話者モデルを学習



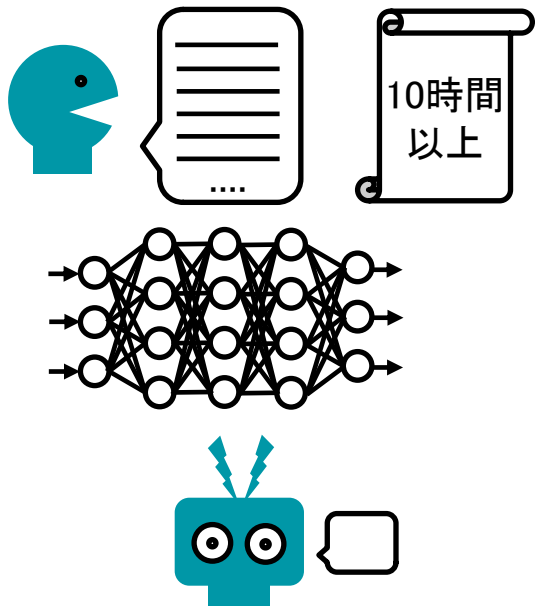
複数人の話者モデルを適用



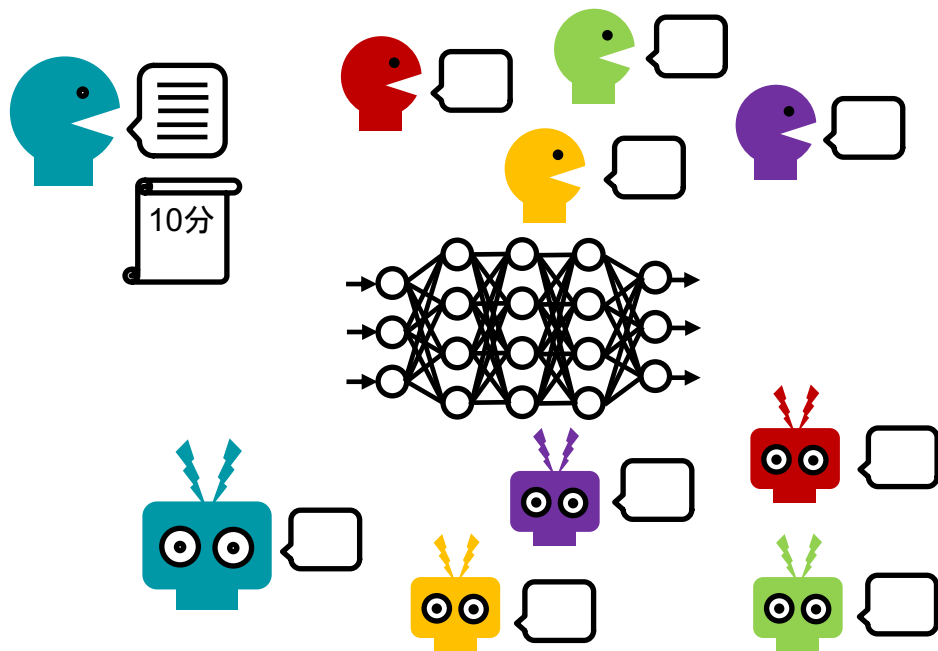
音声合成のパーソナライズ

自分が話しているような合成音声を作るにはどうすればよいか？

一人の話者モデルを学習



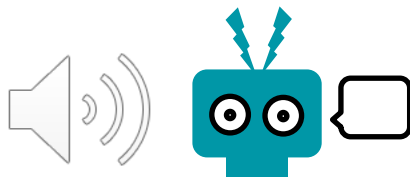
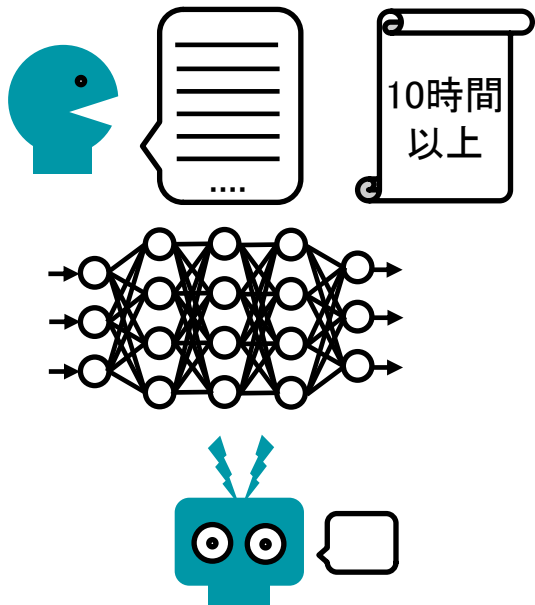
複数人の話者モデルを適用



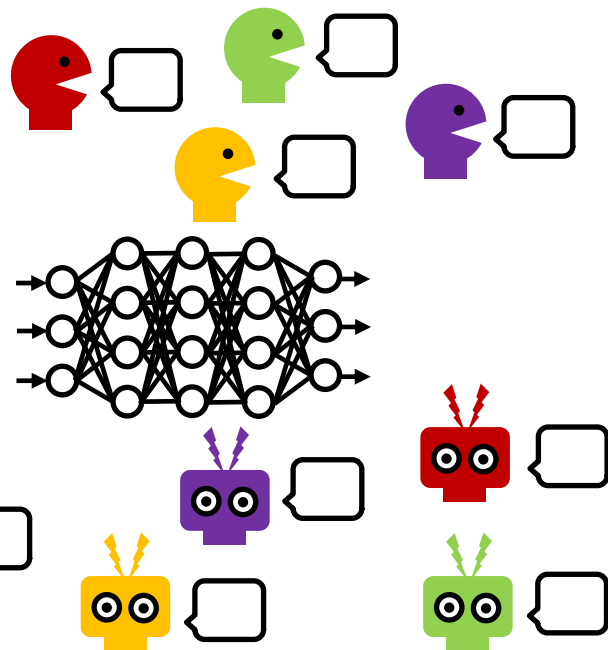
音声合成のパーソナライズ

自分が話しているような合成音声を作るにはどうすればよいか？

一人の話者モデルを学習

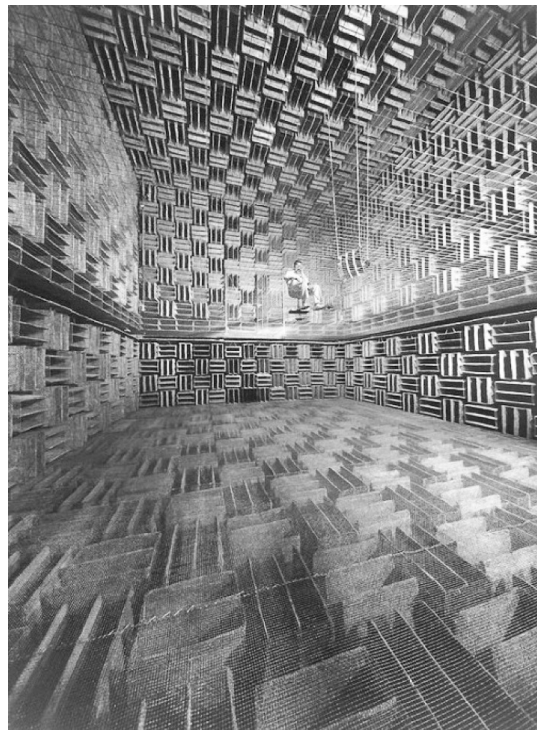


複数人の話者モデルを適用



リソースの少ない言語向けの音声合成

人間らしく聞こえる音声合成には、専門的に録音した数十時間もの音声が必要。



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

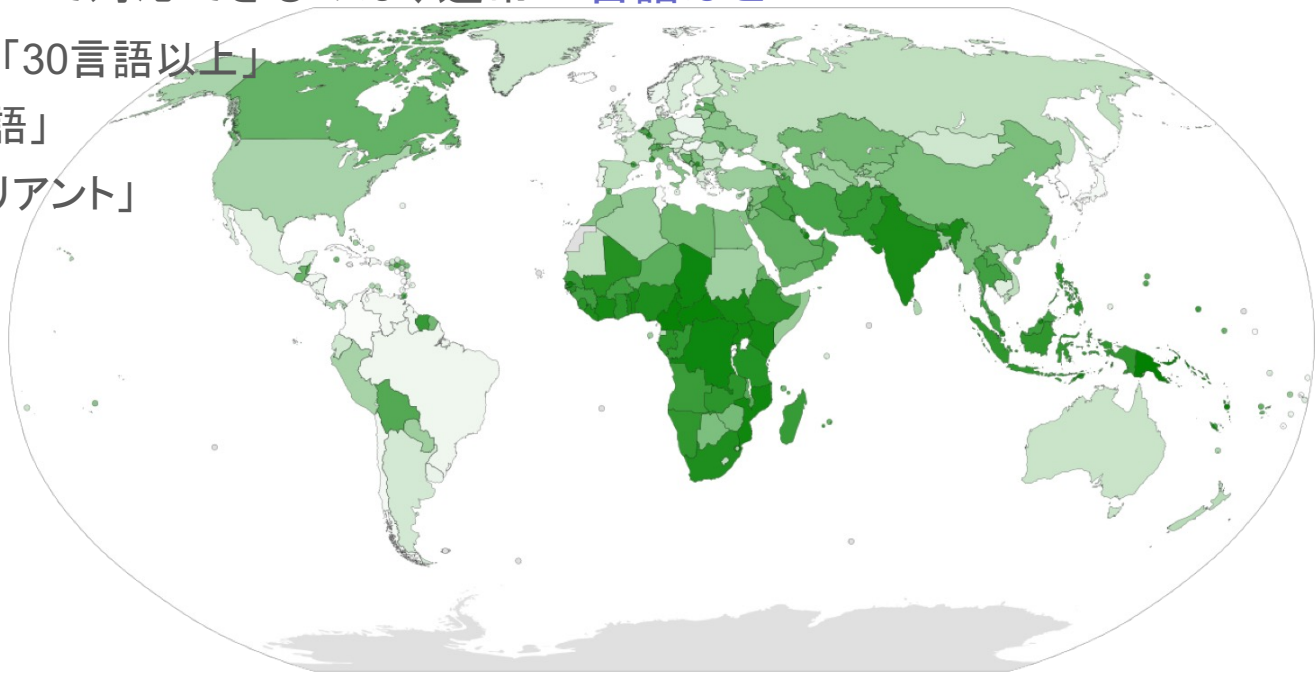
Bell Labs anechoic chamber, 1947

Photo credit: Eric Schaal

ベル研究所の無響室、1947年
写真提供: エリック・シャール

リソースの少ない言語向けの音声合成

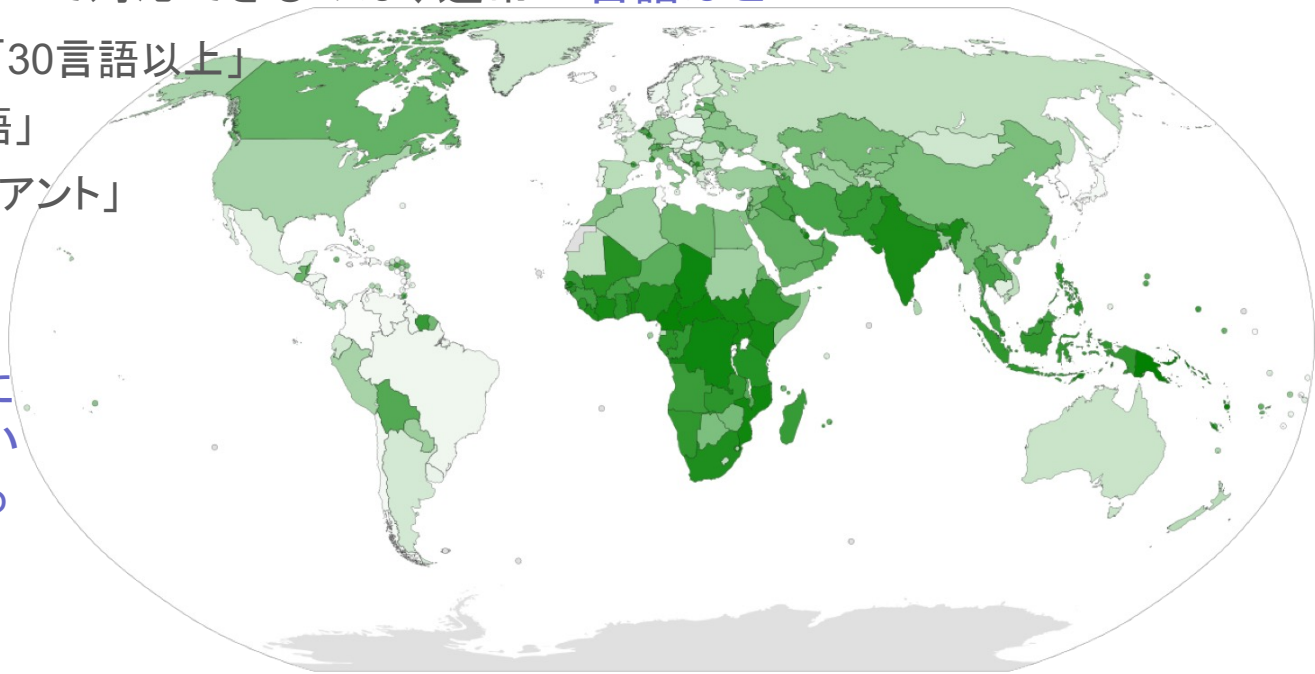
- 世界で話されているのは**7,000言語**以上！
- 一般向け商用システムで対応できるのは、通常**30言語**ほど
 - アップル社MacinTalk:「30言語以上」
 - Amazon Polly:「31言語」
 - Google:「32言語とバリエーション」



リソースの少ない言語向けの音声合成

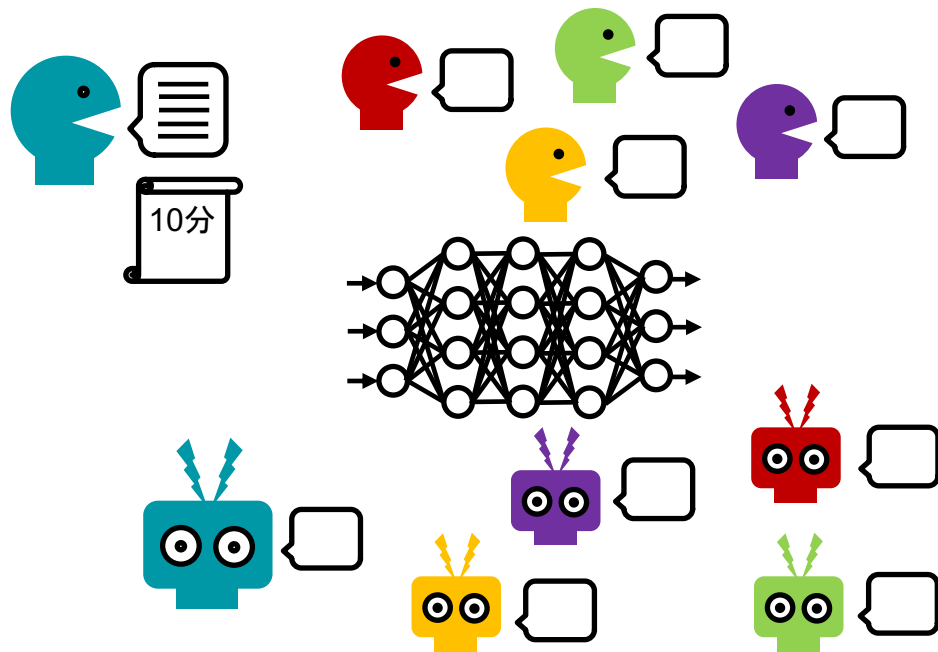
- 世界で話されているのは**7,000言語**以上！
- 一般向け商用システムで対応できるのは、通常**30言語**ほど
 - アップル社MacinTalk:「30言語以上」
 - Amazon Polly:「31言語」
 - Google:「32言語とバリエーション」

高品質の音声録音が大量に蓄積されていなくても、新しい言語の音声合成装置を作ることにはできるのか？



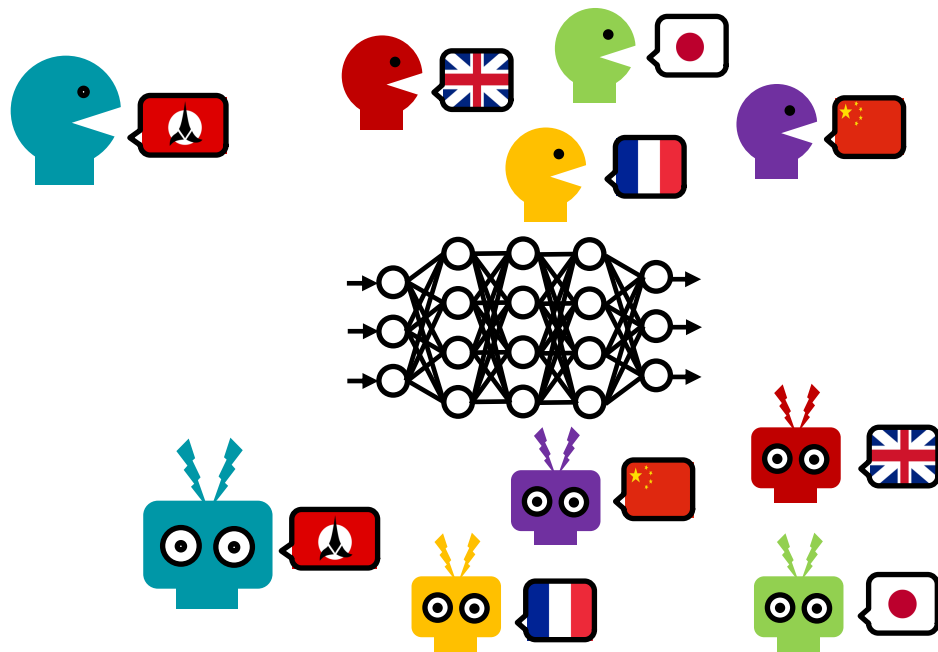
リソースの少ない言語向けの音声合成

複数人の話者モデルを適用



リソースの少ない言語向けの音声合成

多言語モデルを適用



ディープフェイクと音声合成のパーソナライズとの比較

- **技術は同じだが、目的が異なる**
- **音声合成のパーソナライズ:**
 - 個人的な用途
 - 日常生活をサポートするため
 - 独自のコンテンツを作成するため
 - エンタメ用
- **ディープフェイク:**
 - 知らない間に作成される
 - 目的は人をだますこと
 - ある種の詐欺
 - 誤った情報を社会に広める

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

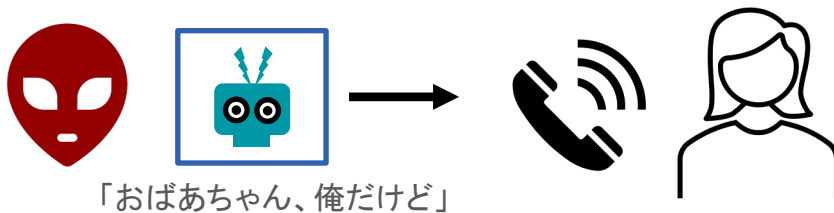
BY JENNIFER ELSEVER

May 5, 2021 8:30 AM GMT+9

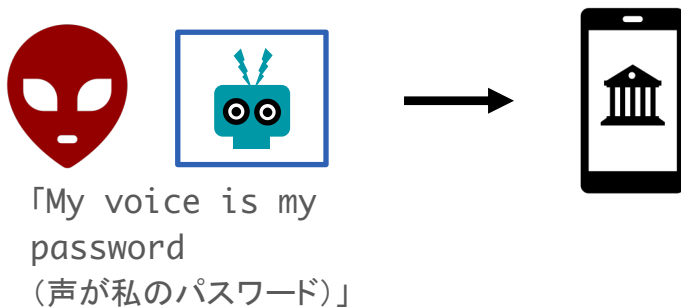
<https://fortune.com/2021/05/04/voice-cloning-fraud-ai-deepfakes-phone-scams/>

ディープフェイクとなりすまし

- 合成音声は人を騙せる



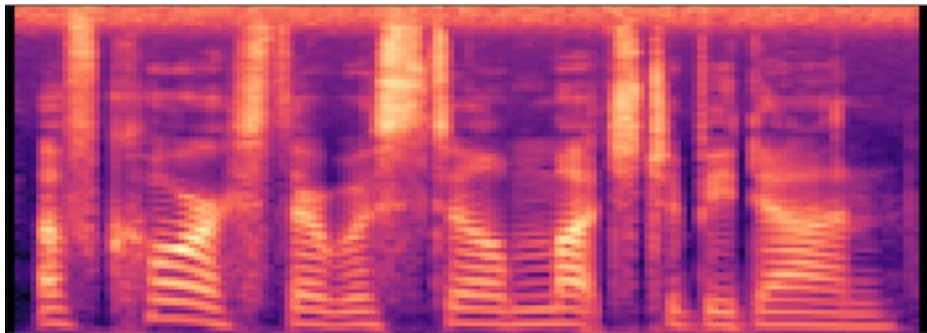
- 合成音声は機械も騙せる



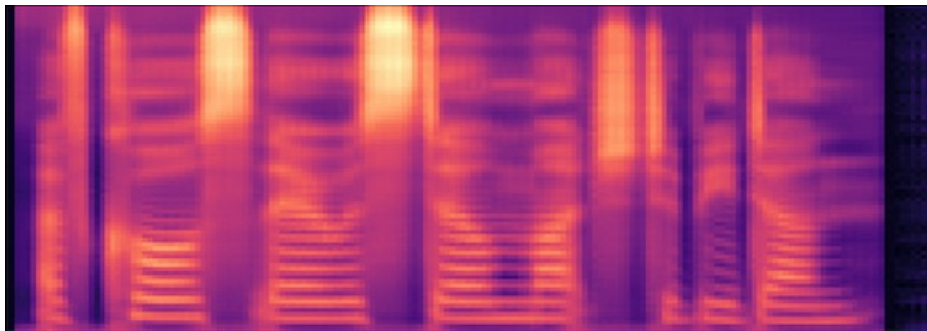
ディープフェイクとなりすまし

- 合成音声は検出できる

実際の音声

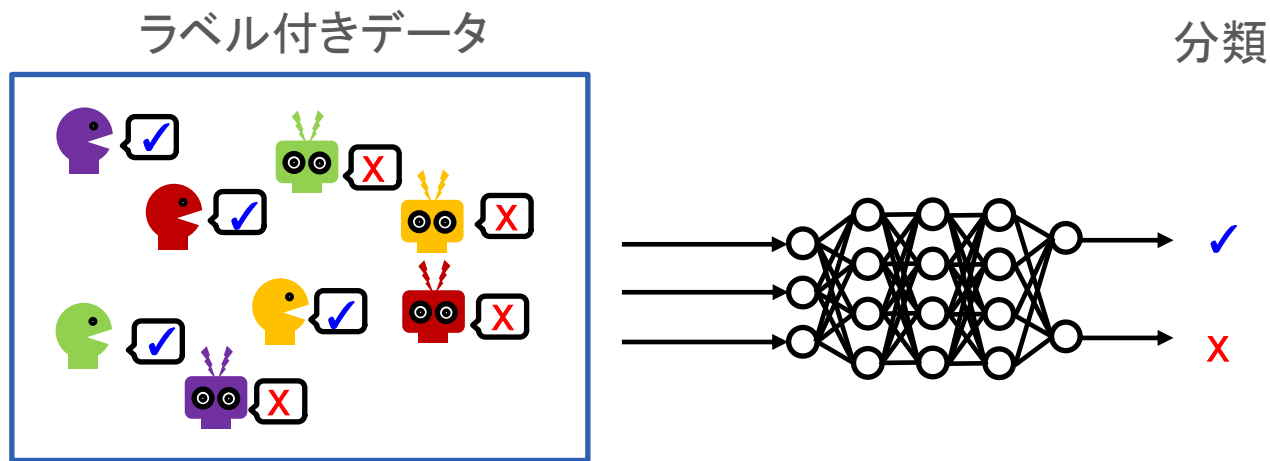


合成音声



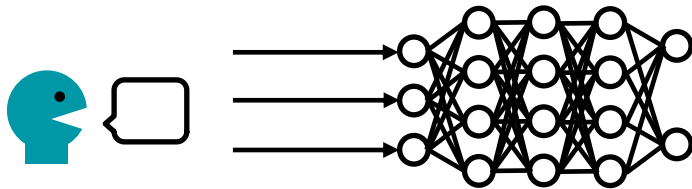
ディープフェイクとなりすまし

- 合成音声なのか、実際の音声なのかは、機械学習によってデータから分かる



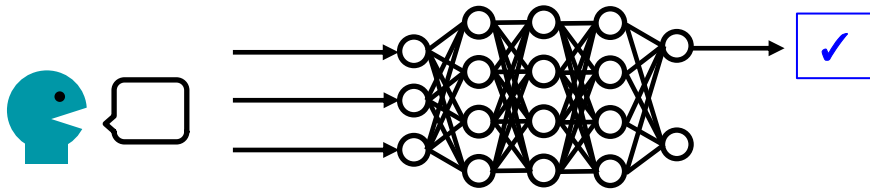
ディープフェイクとなりすまし

- 合成音声なのか、実際の音声なのかは、機械学習によってデータから分かる



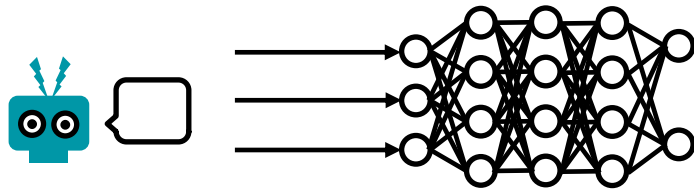
ディープフェイクとなりすまし

- 合成音声なのか、実際の音声なのかは、機械学習によってデータから分かる



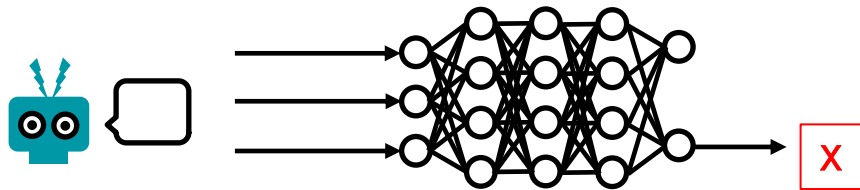
ディープフェイクとなりすまし

- 合成音声なのか、実際の音声なのかは、機械学習によってデータから分かる



ディープフェイクとなりすまし

- 合成音声なのか、実際の音声なのかは、機械学習によってデータから分かる



音声合成:オープンリサーチにおける課題

- なりすましとディープフェイクの検出
 - 人間と機械では音声の知覚が異なるため、エラーの種類も異なる
 - 音声合成は常に改良されており、検出器も追従する必要がある
 - まだ発見されていない新しい合成装置の検出
- リソースの少ない言語向けの音声合成
 - 新しい言語に対応した音声合成装置をできるだけ簡単に作成する方法
- 表現力豊かな音声合成
 - ニュートラル以外のスタイルに合成する方法
 - 聞いて楽しい合成音声を作る方法
- 音声合成の高速化および軽量化
 - ニューラルネットワークモデルには、大量の計算リソースが必要