慶應科学技術展
シンポジウム④「AIフロンティア：基盤モデルが切り拓く未来」

# 音声の基盤モデルが切り拓く未来

高道 慎之介 (東京大学)

本資料の一部は，東京大学 博士後期課程 佐伯 高明氏の協力を得て作成した．

# 自己紹介


@forthshinji

**名前**

高道 慎之介 (たかみち しんのすけ)

**現職**

東京大学 講師

**経歴**

熊本高専→長岡技大→奈良先端大.
2016年に博士 (工学).

**専門**

音声処理

# 最近の音声技術はスゴイ！

## 音声→音声のリアルタイム変換 (弊研発ベンチャーParakeet)

https://www.youtube.com/watch?v=mc7HXXQ8P7Q.
動画作成者の許可を得て利用しています.

## 英語→ドイツ語への音声翻訳

https://seamless.metademolab.com/expressive を用いて作成

これらを支えるのが**音声の基盤モデル**

# 本講演のテーマ

**音声の基盤モデルとは？ 基盤モデルの現状と未来は？**

本スライドは私の個人ウェブサイトで公開予定です．

# 音声と音声工学技術

# 音声 ＝ “人間が発声器官を通して発する音”

**声道 (フィルタ)**

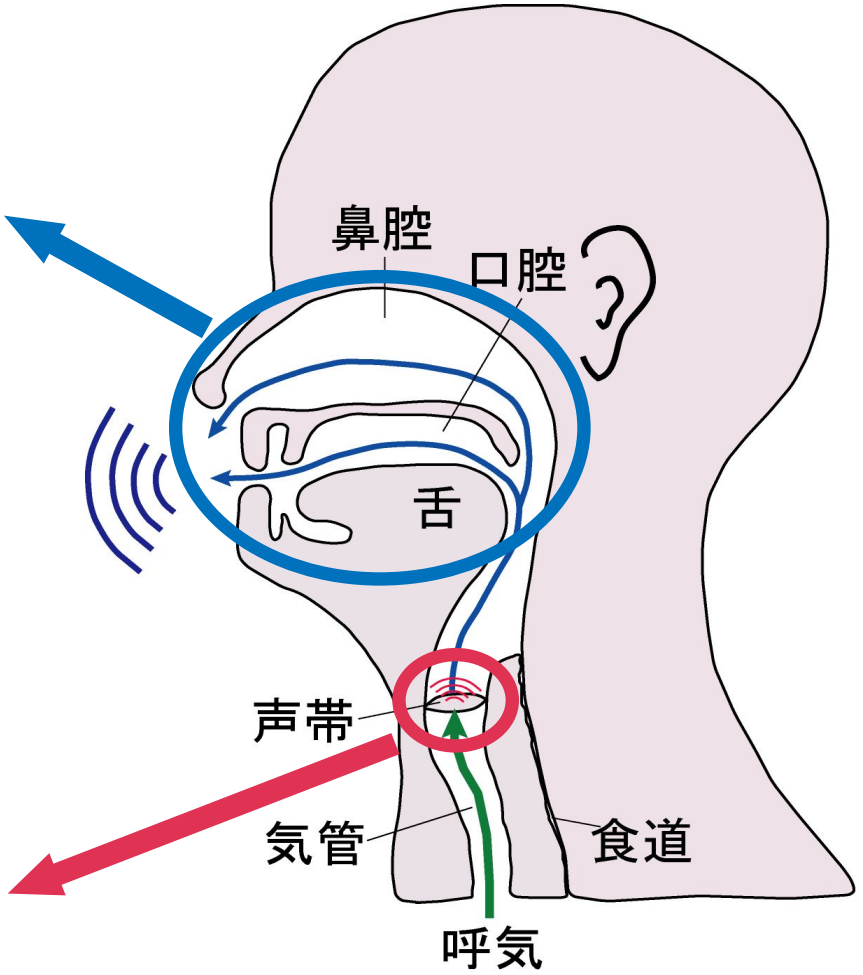
口や喉を動かして
音源信号を音色付け

音声波形

畳込み

**音源 (ソース)**

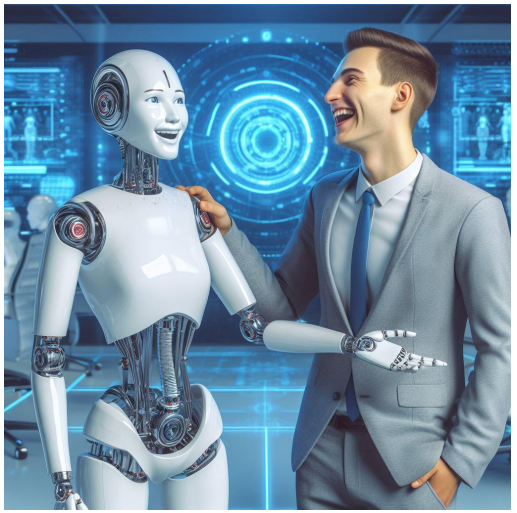
肺からの呼気を
声帯の開閉で制御



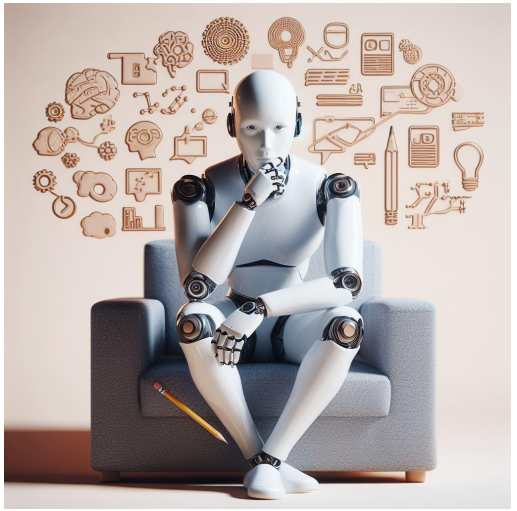

音楽信号(楽音)や環境音は音声に含まないので注意.
(本発表で扱うのはあくまで音声のみ)

# 音声 ＝“マルチモーダル信号の１つ”

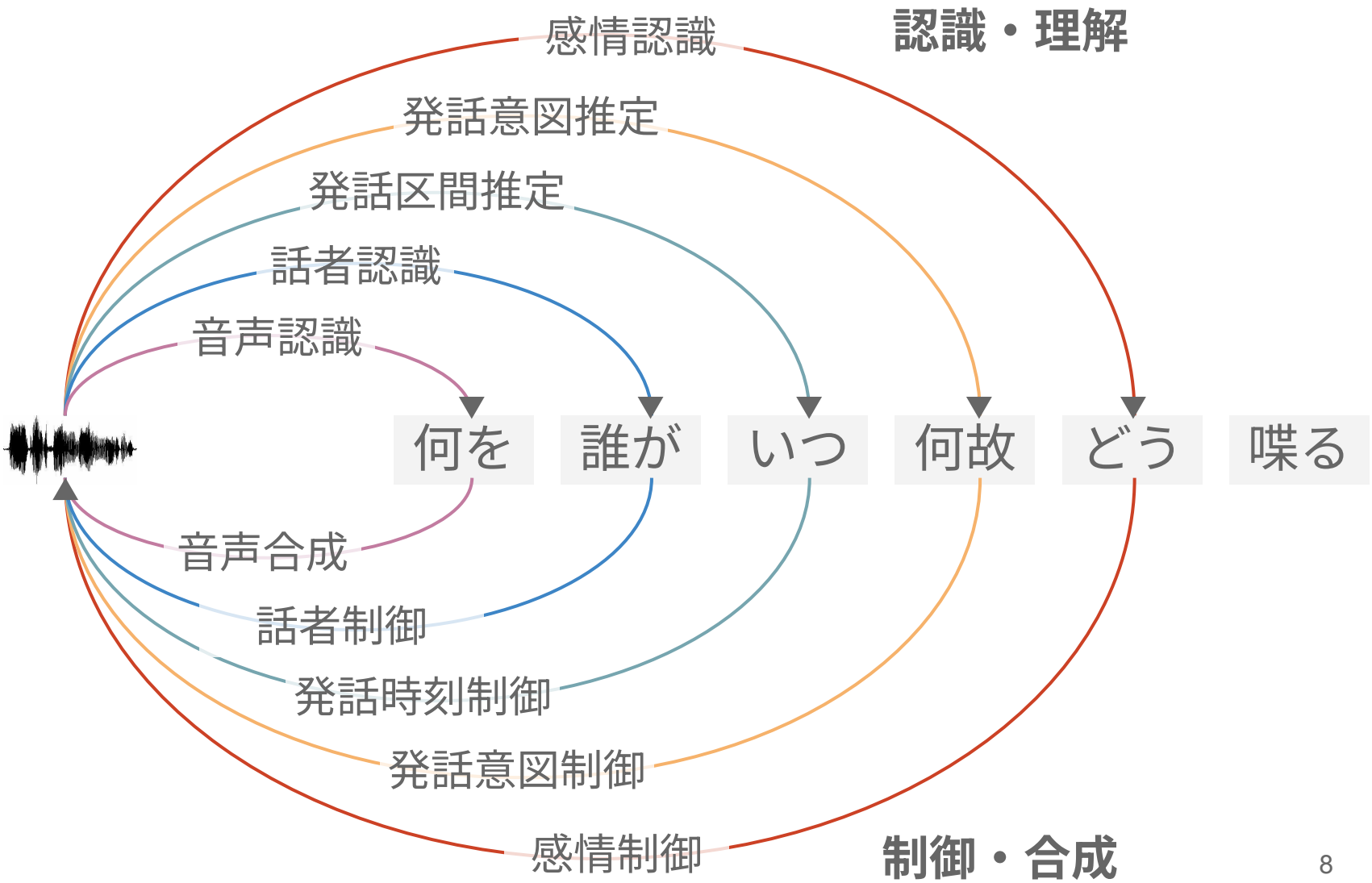
- **画像との相互作用**
  - 音声に加えて相手の顔の動きを見ると発話内容を認識しやすい
  - 音声と顔の動きの組み合わせによって認識される音韻が変わる (マガーク効果)



- **テキストとの相互作用**
  - 音声の感情と発話テキストの感情は関係しやすい（例えば，怒って喋るときは怒りを伝える単語を使いやすい）



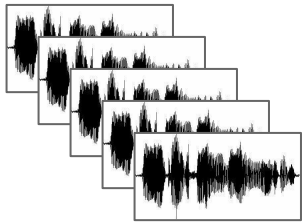
# 音声を扱う工学技術（あくまで一例）



# 音声の基盤モデル (音声分野では self-supervised learning model, SSL モデルと呼ばれます)

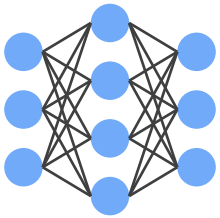
# 音声のSSLモデル

## ラベルなし音声データ



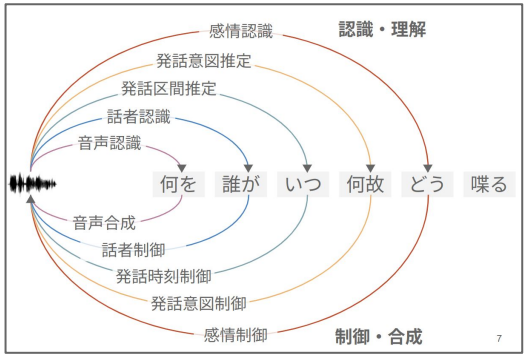
1,000~10,000時間以上．
ラベル = {発話内容, 話者, 感情}など

## SSLモデル



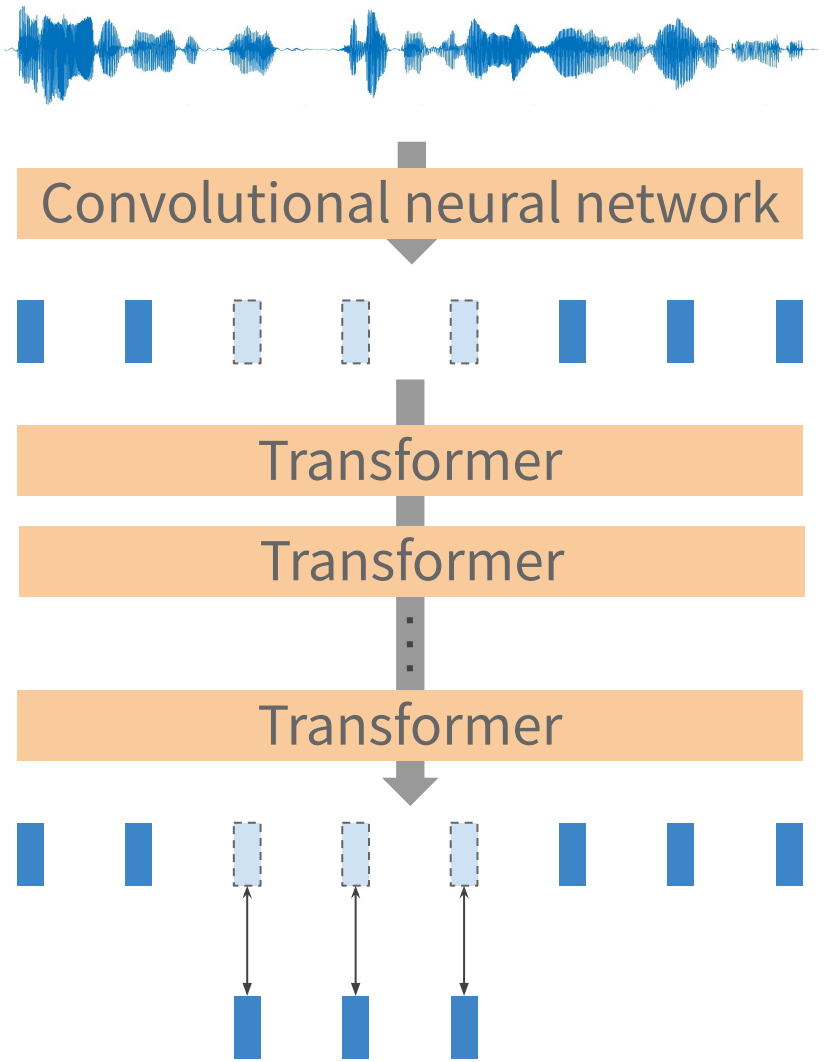
Transformer 等に基づくニューラルネットワーク
次のスライドで例示．

## 下流タスク




様々なタスクでSSLモデルを利用．具体的な利用方法は後述．

# 有力なSSLモデルの一例：WavLM




音声波形 (16,000サンプル/秒～)

時間の解像度を粗くする

時間量子化特徴量 (50 フレーム/秒).
一部をマスクする.

様々な研究分野で使われる
Transformer が音声でも使われる

音声特徴量 (50 フレーム/秒).

マスクされたフレームの特徴量を
推定するようにネットワークを学習

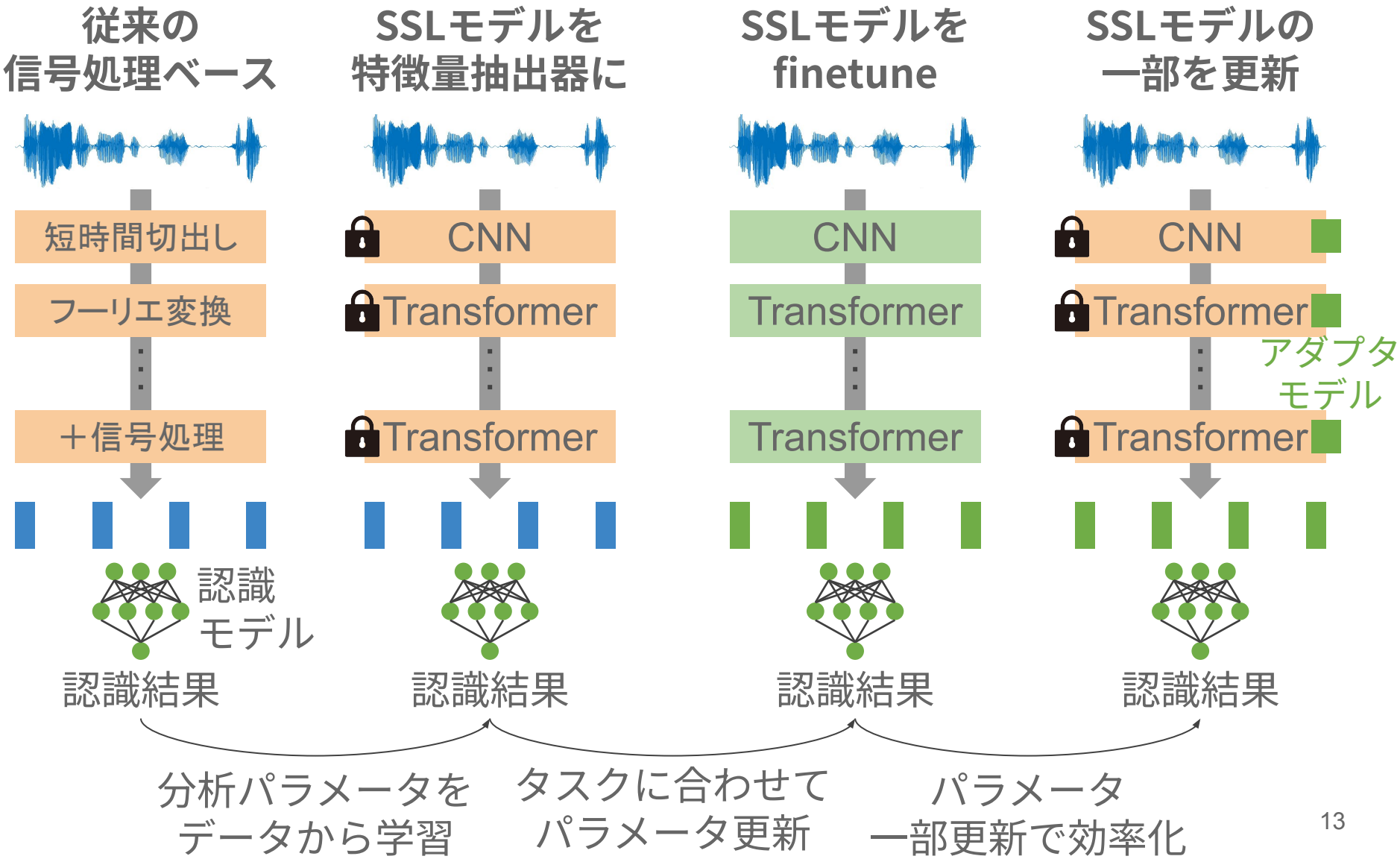


[Chen21] “WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing,” 2021.

# SSLモデルの使い方，現状，未来

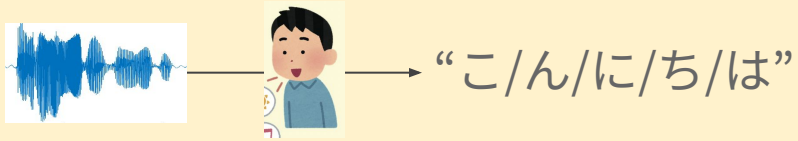
1. 特徴量抽出器（符号化器）としての音声SSL
2. 音声以外も扱うマルチモーダルSSL
3. 音声SSLに潜むリスクとバイアス

# 特徴量抽出器（符号化器）としての音声SSL①: 信号処理を機械学習で置き換える

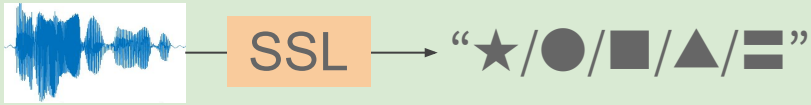


# 特徴量抽出器（符号化器）としての音声SSL②: 自然言語以外の「音の塊」を見つける

**自然言語シンボル(文字，単語)**:
音や意味を表すように人間が定義



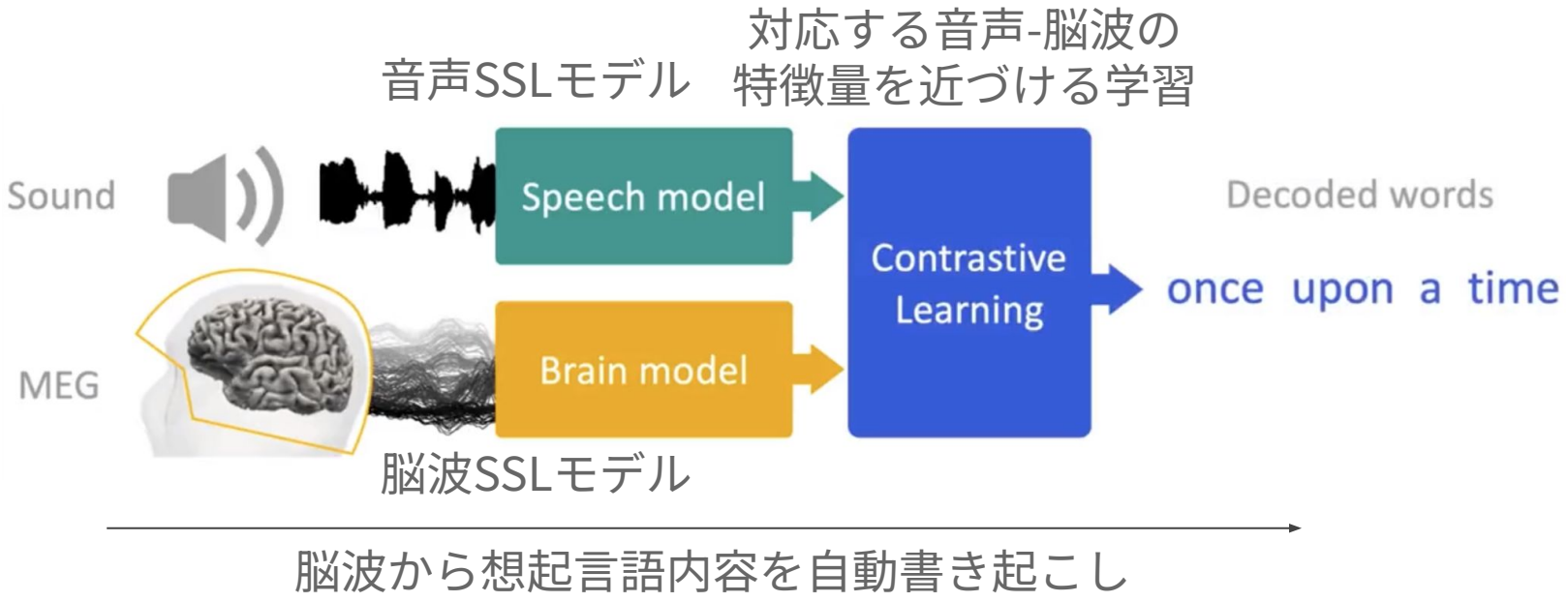

**SSLシンボル**:
音や意味を表すようにSSLが獲得




- **人間は音声とシンボルを対応付けられる（言語獲得）**
  - 人間とコンピュータで獲得シンボルは違うのだろうか(記号接地)？
  - コンピュータによる獲得を分析すると人間の理解に繋がるか？

- **全ての音声が自然言語シンボルで表せるわけではない**
  - 例えば，笑い声，泣き声，叫び声などの非言語音声
  - 自然言語シンボルのかわりにSSLシンボルを用いることで，非言語音声の認識・理解・合成も可能に．(例えば人間のように笑うコンピュータ)

# 音声以外も扱うマルチモーダルSSL①：音声と音声以外で同じ表現を獲得する



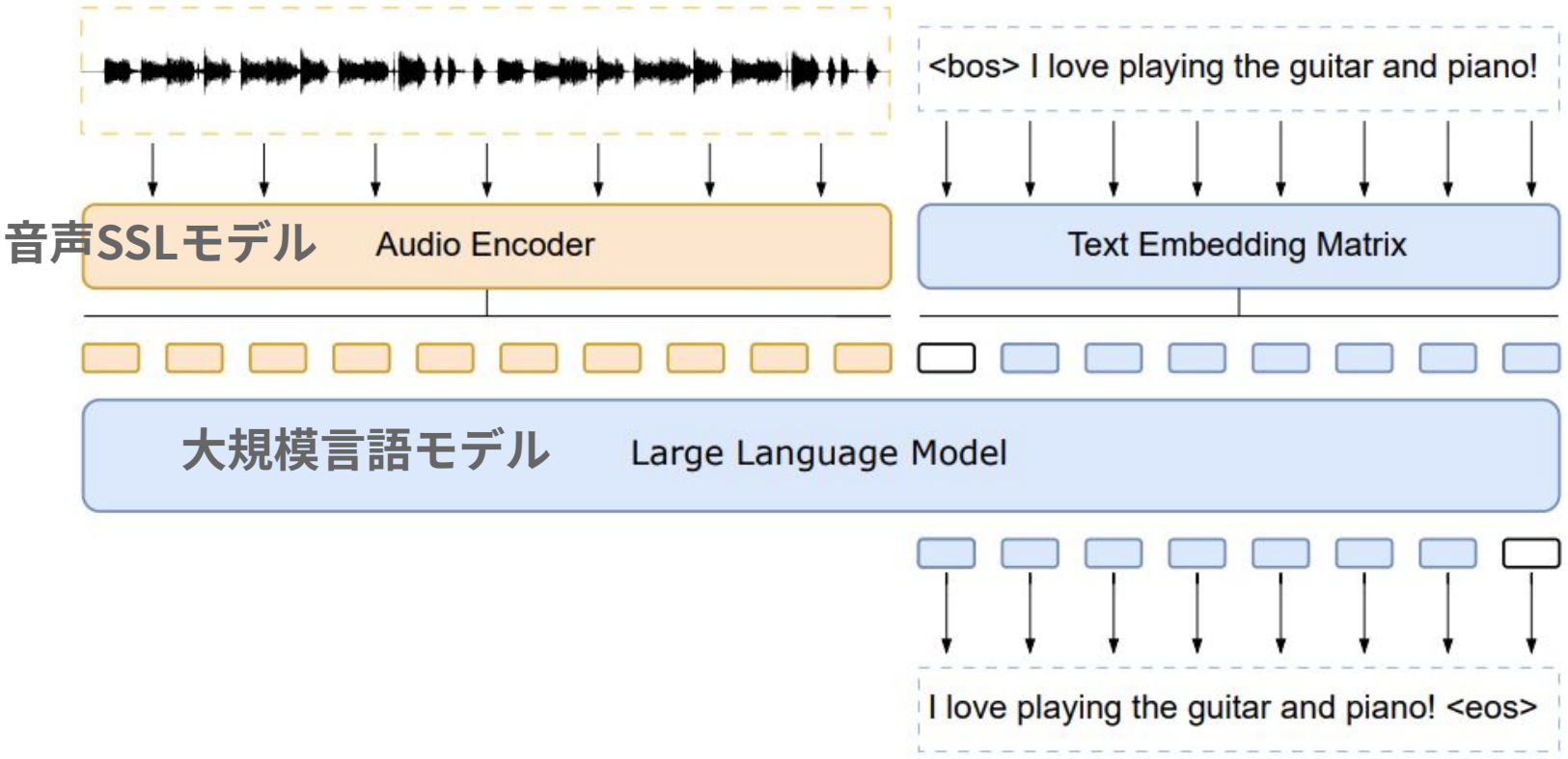

同様の考え方は他のモーダルでも可能．
例えば，音声とテキストの contrastive learning .



https://x.com/JeanRemiKing/status/1709957806030311798?s=20 から画像を引用．一部トリミング

# 音声以外も扱うマルチモーダルSSL②：音声を音声以外の基盤モデルに入力する

テキスト特徴量だけでなく音声特徴量も大規模言語モデルに入力できるようにする




音声を説明したり理解できるLLM (Large “audio” model?) はもうすぐ？



https://arxiv.org/pdf/2307.11795.pdf から図を引用

# 音声SSLに潜むリスクとバイアス①：
# 個人情報に紐づく音声の利用を前提にしている

現代の音声生成AIは音声SSLモデルを利用する．AIモデル及びSSLモデルは
実在人物の音声を用いて学習され，また，実在人物の声を再現できてしまう


(法整備などによる保護は別途必要な上で)
実在人物の音声を利用しない学習，音声を忘れる学習が進められている



https://www.nhk.jp/p/gendai/ts/R7Y6NGLJ6G/episode/te/9V5355JGJ6/ から図を引用

# 音声SSLに潜むリスクとバイアス②：音声に潜む言語差・文化差

- **既存の学習済みSSLモデルの殆どは英語である一方で，音声には言語差と文化差がある**
  - 例えば，英語と日本語では音分布が異なる．日本人と中国人では感情の表現方法が異なる
  - → 英語＋αの音声で学習されたSSLモデル，また，大規模言語モデル（LLM）の音声版もこれから必要になる

- **日本語に特化したSSLモデルを学習するためのデータは？**
  - テレビ放送に基づくデータセット reasonspeech
  - YouTubeデータに基づくデータセット jtubespeech，YODAS

# まとめ

# まとめ

- **音声と音声工学技術**
  - 音声は，人間が口で出す音である．
  - 音声は，画像やテキストと相互作用する．
  - 音声工学技術は，音声の種々の情報を認識合成する．

- **音声の基盤モデル（SSL）**
  - 特徴量抽出器（符号化器）としての音声SSL
  - 音声以外も扱うマルチモーダルSSL
  - 音声SSLに潜むリスクとバイアス
    - 日本語SSLの必要性．