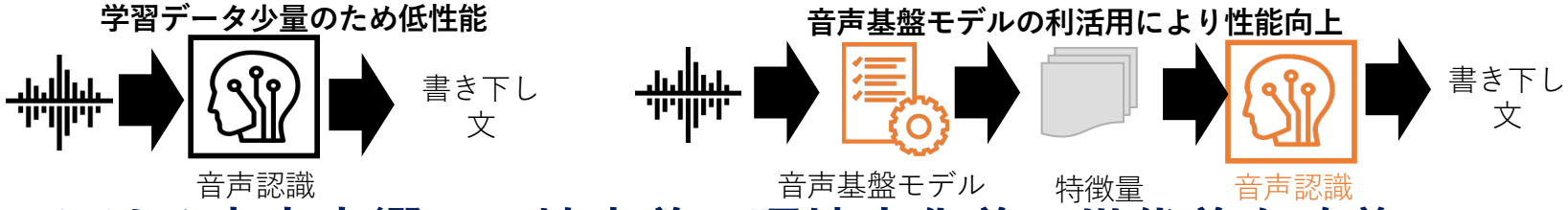
産総研
ともに挑む。つぎを創る。

日本語音声基盤モデルの構築と利活用

国立研究開発法人 産業技術総合研究所 人工知能研究センター
知的メディア処理研究チーム　研究チーム長
深山 覚（ふかやま さとる）

日本語音声基盤モデルの構築と利活用

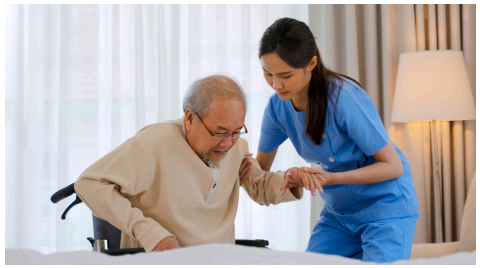
- **音声基盤モデル**
  - **基盤モデル：さまざまなAI構築に活用できる汎用的なモデル**
  - **音声基盤モデル：少量音声データで性能の良い音声認識・合成を可能に**



- **実世界における音声音響AIの地方差・環境文化差・世代差を改善**
  - **方言音声・工場などでの雑音環境下音声・高齢者音声などのAI性能を向上**

地方小売店舗
での音声入力

聴覚付き
産業ロボット



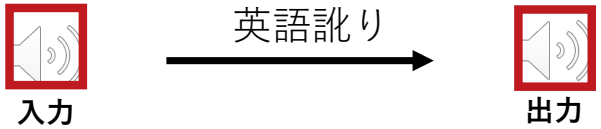
介護施設での
発話促進

実世界音声音響AIの公平性改善

イントネーションの変換

恩田ほか　母語話者音声のみを用いた外国語訛りに頑健な自動音声認識の実現に向けた離散トークンの活用の検討
日本音響学会講演論文集 2025年3月

- **外国語訛りの付与（東京大学等と連携）**
  - **世界の地域ごとの訛りのある音声の音声認識性能改善に活用**
    - **日本語「水をマレーシアから買わなくてはならないのです」**



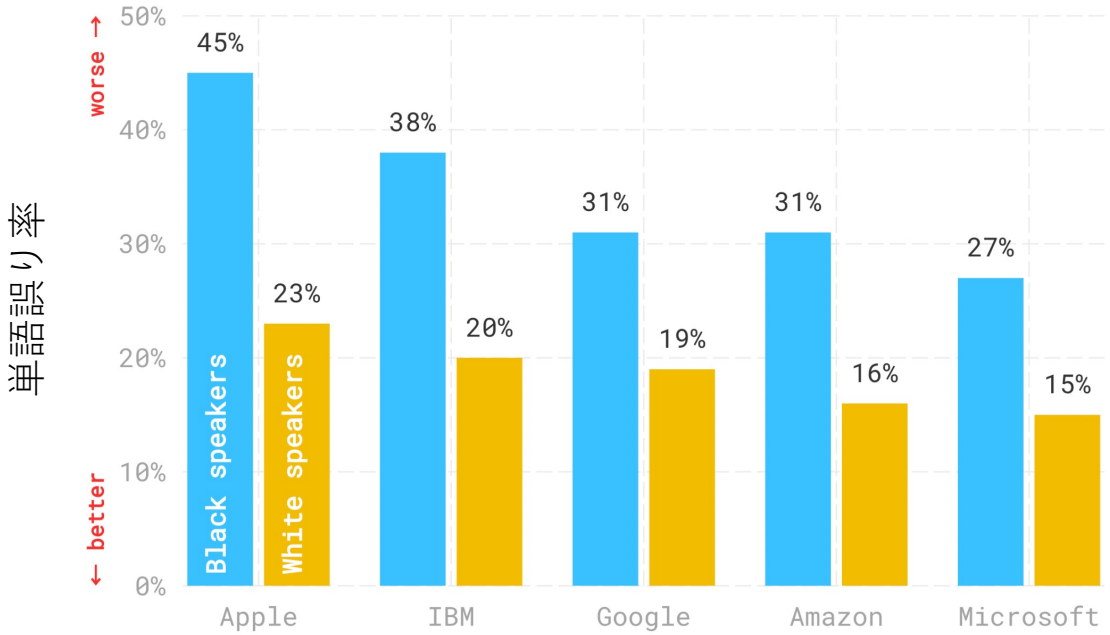
    - **英語 “A lone star shone in the early evening sky.”**

  - **外国語訛りに頑健な音声認識の実現に取り組み中**
    - **「母国語訛りの外国語は聞き取りやすい」現象**
    - **訛りのある音声の認識プロセスを音声基盤モデルを用いて模倣**
    - **訛りの元である母国語学習データを用いて訛りありの外国語音声認識性能を向上**

音声基盤モデル研究の意義

- 実世界音声・音響の多様性を考慮した音声処理の基盤を作る
  - 話者の属性（方言・高齢者）等に由来する性能差を抑制した音声処理の構築
  - 特に少量データに由来して性能が低いタスクを音声基盤モデルで改善

データ量の偏りに由来して黒人話者の音声認識誤り率が
白人話者よりも高い旨のスタンフォード大学の報告

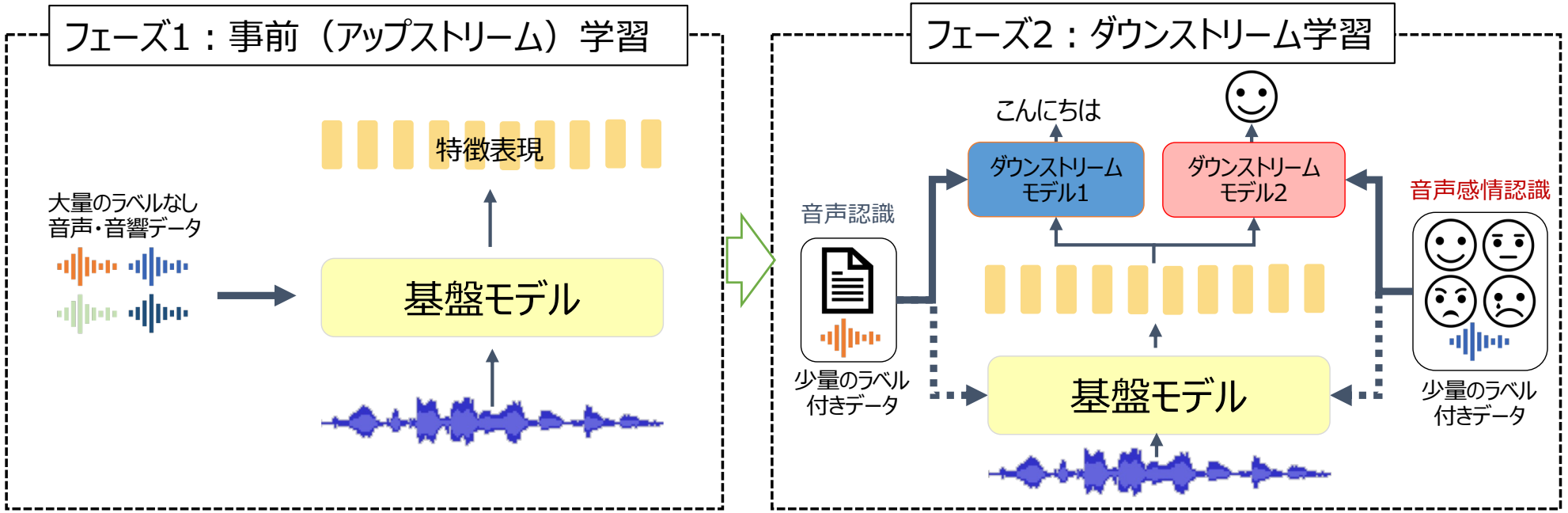


The Race Gap in Speech Recognition Technology
https://fairspeech.stanford.edu/ (viewed 4th Mar. 2024)

自己教師あり学習モデル as 音声基盤モデル

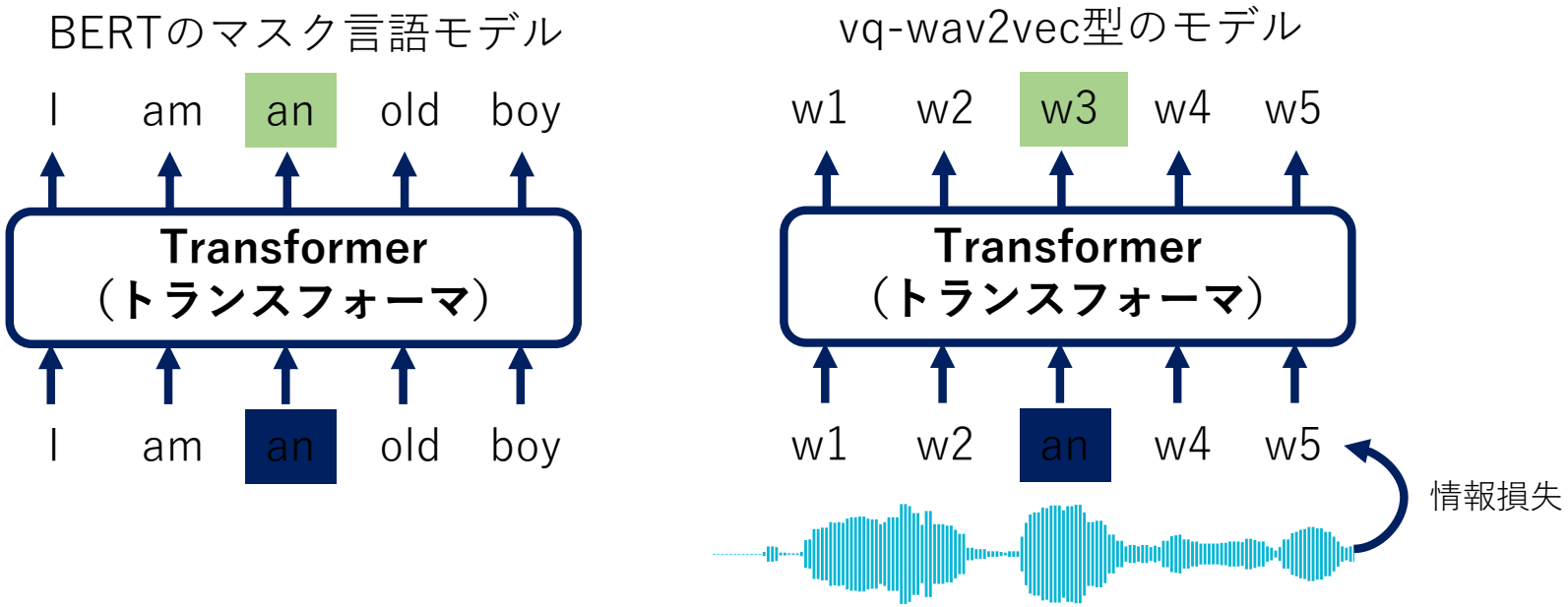
- 深層学習モデルの中間層表現が音声処理の複数タスクに活用可能
  - 中間表現を用いて音声認識・音声感情認識・音声合成
  - 少量データセットからの音声処理システム構築に有用
    - カンボジア特別法廷のクメール語音声認識 [1]
      - 自己教師あり学習モデル（XLS-R）-> 10時間クメール語音声データ 文字誤り率 11.1%
      - クメール語音声データのみで音声認識器を構築：45時間データで文字誤り率 12%

[1] 大規模事前学習モデルに基づく音声認識, 河原達也, 三村正人, 日本音響学会誌79巻9号(23), 2023



音声データの自己教師あり学習

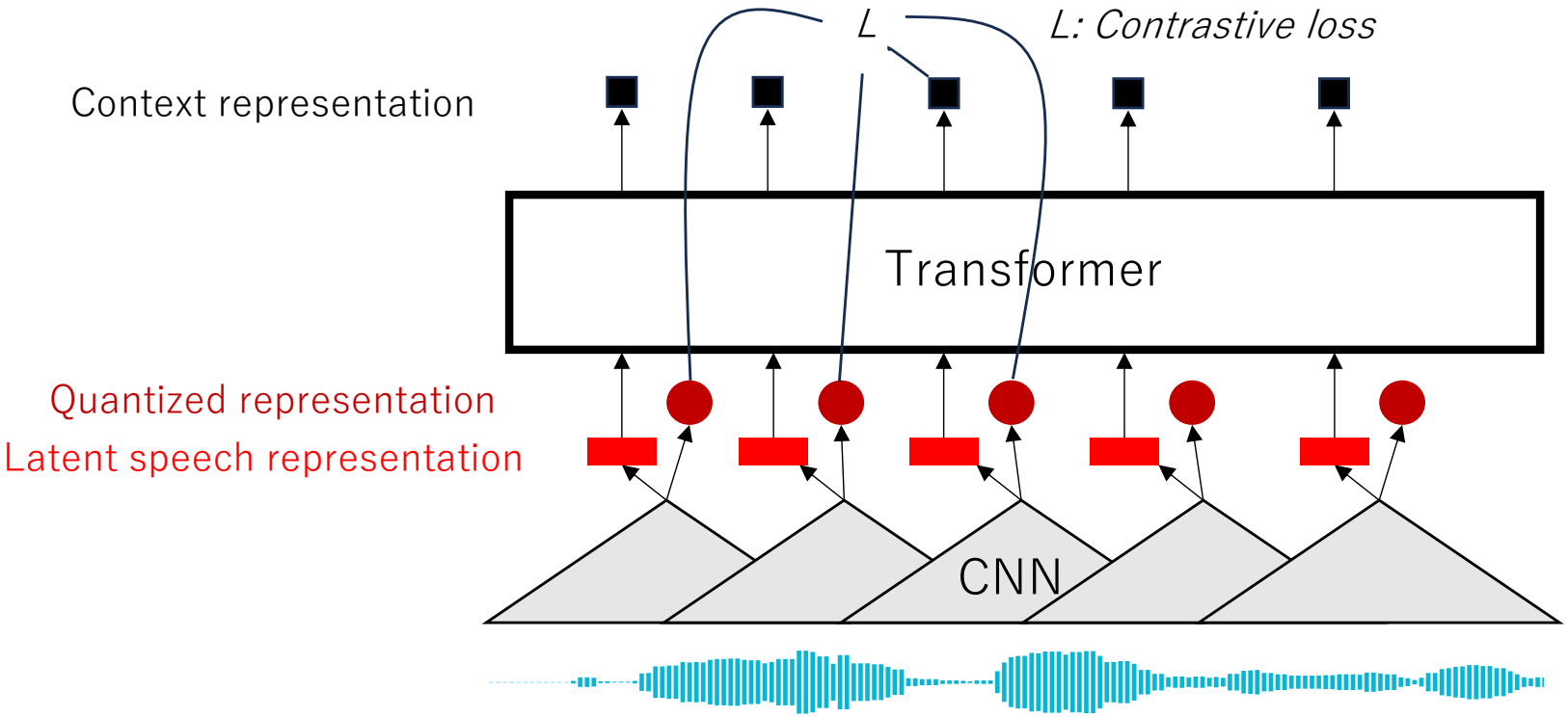
- 言語データの自己教師あり学習の影響
  - BERT: Transformerを用いたmasked language model [2]
  - vq-wav2vec: 音声の各フレームを符号化してBERTに入力 [3]
    - 事前の音声符号化における情報損失に由来する限界あり



[1] 大規模事前学習モデルに基づく音声認識, 河原達也, 三村正人, 日本音響学会誌79巻9号(23), 2023 （上記図を引用・改変）
[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proc. NAACL-HLT 2019.
[3] A. Baevski, S. Schneider, M. Auli. VQ-wav2vec: Self-Supervised Learning of Discrete Speech Representations, arXiv:1910.05453, 2019.

音声符号化を含めて自己教師あり学習

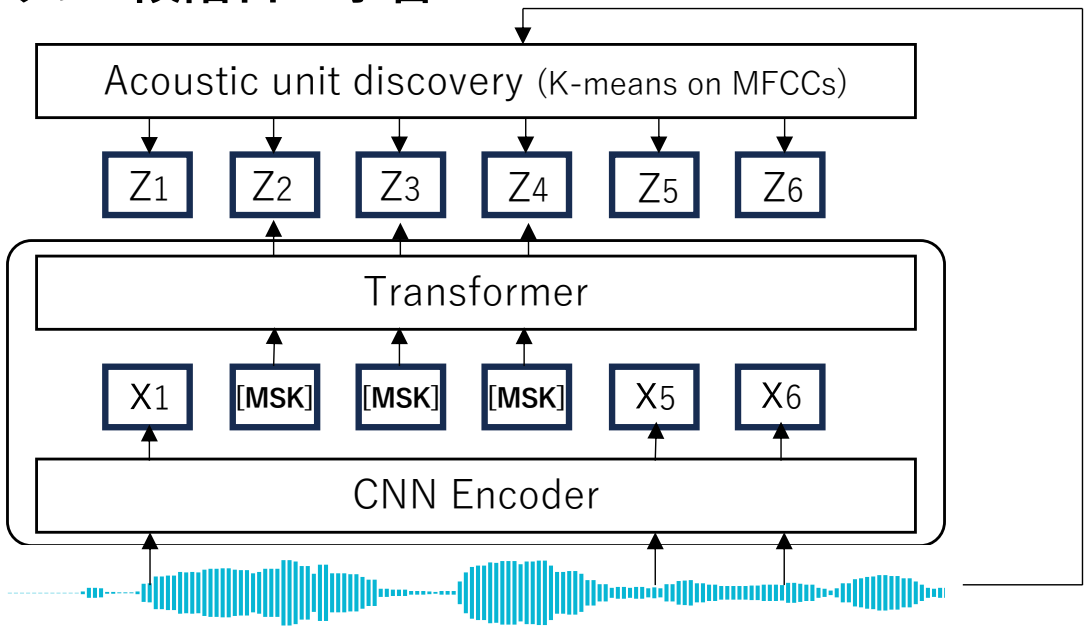
- **wav2vec 2.0** [4]
  - **CNNの出力を量子化（Gumbel-softmax）して符号化**
  - **Transformerの出力と符号のペアを作り対照学習**
  - **対象学習とともに多くの符号が一様に用いられるようなロスを導入**



[4] A. Baevski, H. Zhou, A. Mohamed, and M. Auli. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Proc. NIPS 2020.

音声符号化を含めて自己教師あり学習

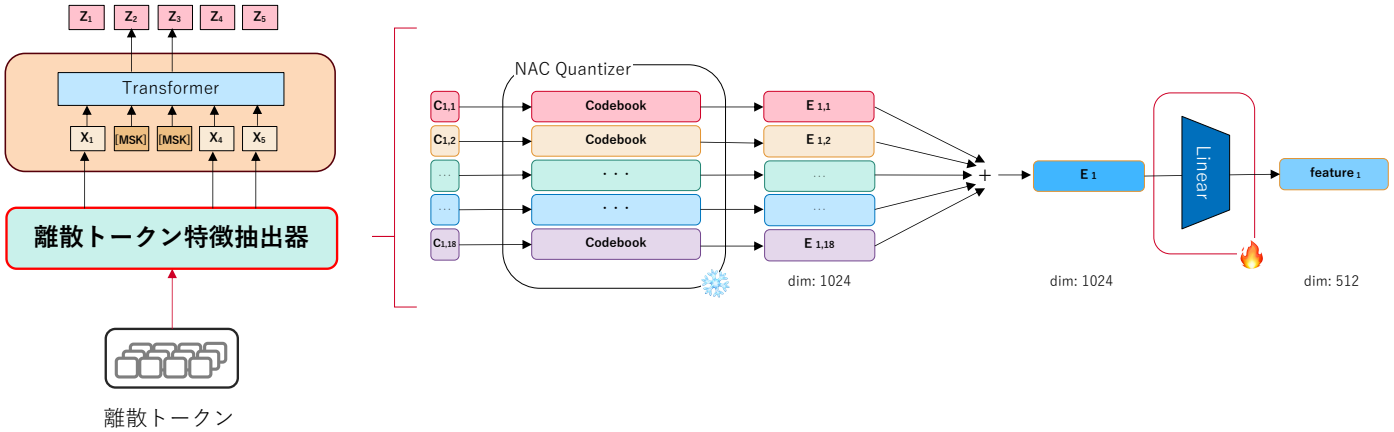
- **HuBERT** [5]
  - wav2vec から量子化と対照学習を削除
  - トランスフォーマの出力する符号が音声認識の古典的な特徴量(MFCC)のクラスタリング（k-means）結果と一致するように学習
  - さらに学習されたTransformerの中間層の特徴量をクラスタリングした結果と一致するように2段階目の学習



[5] W.-N. Hsu, B. Bolte, Y.-H. Hubert Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units, IEEE/ACM Trans. Audio, Speech & Language Proc., Vol. 29, pp. 3451–3460, 2021.

省データ容量・省計算量の自己教師あり学習

- **NACHuBERT [6]**
  - **音声符号化（ニューラルオーディオコーデック; NAC）を用いHuBERT構築**
  - **学習データ最大96.5%・学習GPU時間最大18.8%・GPUメモリ使用量最大15.4%削減しつつ、ほぼ同等の性能を達成**



SSL	データ	英語 WER [%] ↓	日本語 CER [%] ↓		
		Libri-Light (test_clean / test_other)	LaboroTVSpeech	CSJ (eval 1 / 2 / 3)	COJADS
HuBERT	オリジナル音声	10.7 / 18.6	14.1	5.1 / 3.9 / 4.3	34.7
	再構成音声	10.9 / 19.4	14.7 ※	5.5 / 3.8 / 4.2	36.9
NAC HuBERT	**離散トークン**	**10.4 / 20.8**	**14.5**	**5.2 / 3.6 / 3.9**	**37.5**

[6] 瀧澤 他, Neural Audio Codecを用いた自己教師あり学習モデルにおける事前学習データの言語が下流タスクに与える影響の検証, 日本音響学会講演論文集, 2026.

プレス発表報告

- 日本語音声基盤モデル「いざなみ」「くしなだ」を公開（2025年3月10日発表）
  - 少量の日本語音声データで高性能な音声処理を構築可能に
    - 日本経済新聞ほか複数新聞・メディア取材
    - 企業・自治体より活用の問い合わせ

日本語特有な感情表現に有用な音声基盤モデル

- 少量の自然発話感情音声を用いて高性能な音声感情認識を行う
  - 日本語と他言語では言語特性・感情表現に差異 [Russell91]
  - 日本語特有な感情表現に有用な音声感情認識があるとよい
  - 感情ラベルを持つ日本語自然発話データは小規模（学習データ10時間程度）
    - 大規模データによる性能向上を目論む機械学習アプローチが不得手
- 日本語音声基盤モデルを構築し利活用
  - 基盤モデルに自己教師あり学習モデルを使用（ラベル無しで学習）
  - 産総研大規模GPUクラスタであるABCI2.0および3.0を大規模利用

演技感情音声データによる評価実験結果

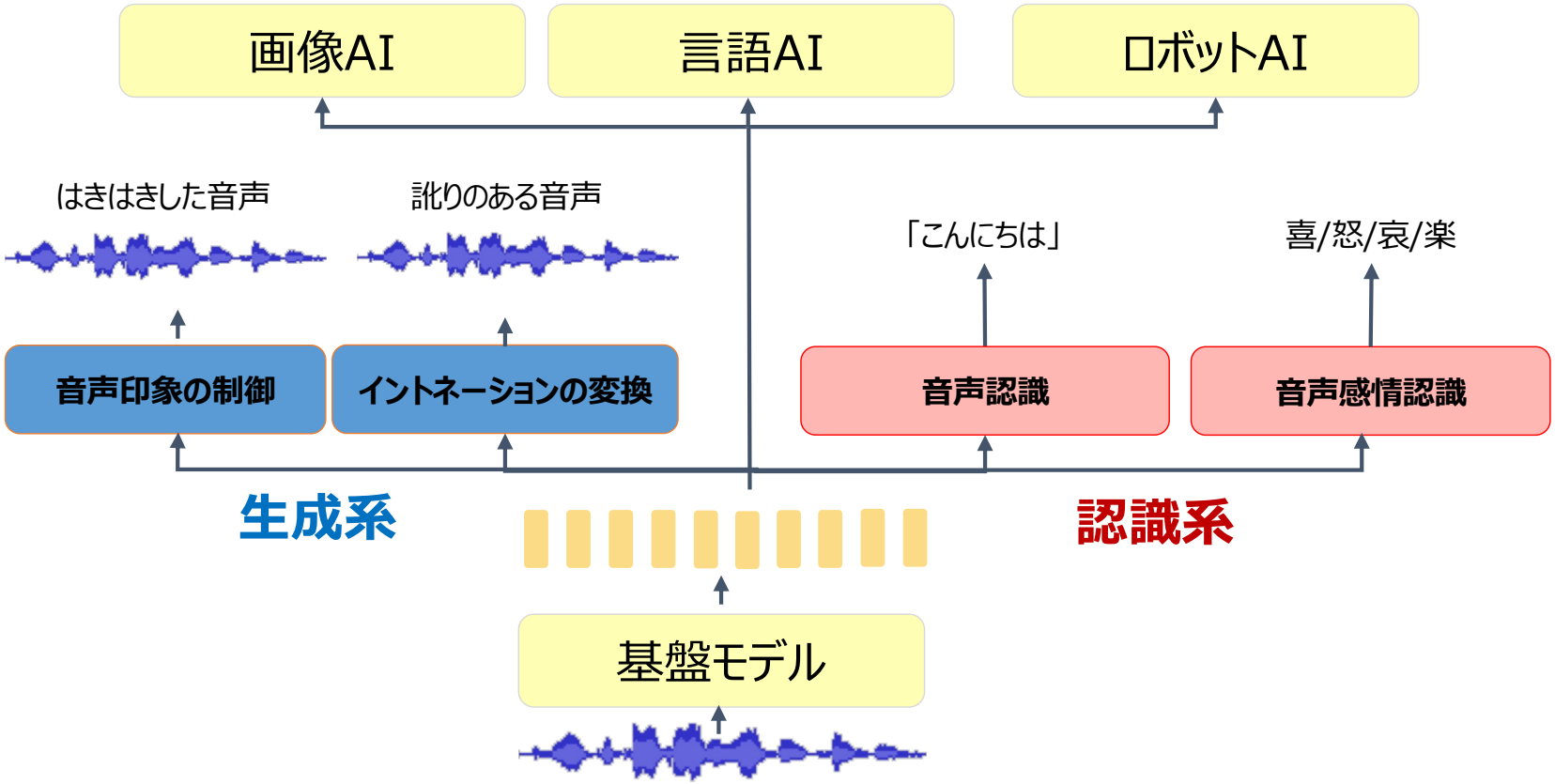
- 少量データによる感情認識器構築に日本語音声基盤モデルが有用
  - 11時間程度の学習データであっても高い感情認識率を実現
  - 日本語以外で学習された基盤モデルに比べ約10-15ポイント改善

基盤モデル	言語	基盤モデル構築用データセット	基盤モデル構築用データ量(h)	JTES (4感情)
wav2vec2*	英語	LibriSpeech	960	70.10
xls_r_300m*	128言語	VoxPopuli etc.	43600	72.90
wavlm_large*	英語	GigaSpeech etc.	94000	71.70
くしなだ	**日本語**	**テレビ音声データ**	**62000**	**84.67**

* SUPERB公開モデル

音声基盤モデルの応用

- 音声基盤モデルは汎用的な音声処理の事前学習モデル
  - 基盤モデルにより得られる特徴量を活用することで様々な音声AIを実現
  - 画像AI・言語AI・ロボットAIに繋ぐ中間的な表現として活用可能



まとめ：日本語音声基盤モデルの構築と利活用

- 少量データ活用のための大規模基盤モデル研究開発
  - 音声合成・変換などの音声生成の基盤技術
  - 自己教師あり学習による日本語音声基盤モデルの構築
  - 少量の日本語感情音声を用いた音声感情認識の性能改善
- 国産日本語音声基盤モデルの構築
  - 構築されたモデルを公開（プレスリリース）
    - 広く音声・音響基盤モデルを利用してもらい音声・音響処理の構築を容易化
- 音声AIを強みとするマルチモーダルAIの構築へ
  - 音声・音響：発話のみならず感情認識・空間把握などに重要
    - 音声基盤モデルを利活用して画像・言語・ロボットのAIと情報を受け渡し