

合成データを使用した日本語音声LLMの開発

堤 歩斗, 大城 治城(都立大)

日本語合成音声データを作成し、音声理解可能なマルチモーダルLLMを構築

- 音声対話・音声認識・音声キャプション生成などの複数の音声タスクを学習
- 日本語の対話性能で既存オープンモデルを上回る
- 音声認識ではWhisper-large-v3を上回る

ポスターではデモも準備しています！

Team 031 合成データを使用した日本語音声LLMの開発
 ○堤 歩斗, 大城 治城 (都立大)

1. 概要

意義

- Gemma3n や Voxtral のような Large Audio Language Model がオープンウェイトで公開され始めた
- 日本語中心で学習されたモデルに音声入力対応のモデルはない
- **日本語に強い音声入力可能なモデルを構築したい**
 - LLM-JP-4の持つ日本の知識を活用した対話システム
 - 日本語の音声言語タスクにおいて他モデルより高い性能を示す可能性

貢献

- 高い日本知識・日本語能力をもつ音声LLMを開発
- 2種類のモデルを構築
 - 対話モデル : 音声の指示に 대응するモデル
 - 音声認識モデル : 日本語音声認識に特化したモデル
- 対話モデルではADU-Benchで**Voxtral, gemma3nを上回る**
- 音声認識特化モデルでは**Whisper-large-v3を上回る**
- 提供された計算資源を活用した合成データの作成
- 音声合成により作成した合計約1075時間の日本語音声対話データセットを公開
- Qwen3-Omniを使用して作成したCCライセンスの環境音キャプションデータ (43k) も公開

モデルアーキテクチャと学習戦略

2. 合成データの作成

使用可能なデータの要件

商用利用可能な日本語の音声リソースは限られている

- **Pretraining**
 - ASR (音声認識) データ → ReazonSpeechを使用
- **SFT**
 - 指示-応答の形式の音声データ → TTSで合成
 - 環境音キャプションデータ → Qwen3-Omniで合成
- **DPO**
 - 指示-応答の形式の音声の嗜好データ → TTSで合成

TTSによる音声データの生成

使用モデル: CosyVoice2 TTS
 参照音声を20語者分用意し、**話者多様性のあるデータ**を生成
 音声データの合成に600GPU時間 (H200) 程度を使用

元データセット	rows	時間
SFT illn-jp/magpie-sft-v1.0	132k	700 時間
SFT kaitani/keiyama/AutoMakiTurndByCalm3022B	59k	265 時間
DPO illn-jp/ih-hit-12k-ja	12k	32 時間
DPO cybersgmt/chatbot-arena-ja-clan2-7b-chat-experimental	29k	78 時間

Qwen3-Omniによるキャプションデータの生成

- 環境音のキャプションデータも使用したい
- 商用利用可能なテキストラベル付きデータが存在しない
- FSD50kのうちCCライセンスのデータ (43k) を Qwen3-Omni-Captionerでキャプションを生成 (英語)

3. 学習

Voxtralのテクニカルレポートを参考に3ステップで学習を実施
Pretraining
 PretrainingではLLMもフリーズしAdapterのみをASRタスクで学習する
 これによりAudio Encoderの出力をLLMが理解できるように変換するAdapterが学習できる

音声認識特化モデルは、LLMのフリーズを解除しASRの学習を継続する

SFT
 SFTではLLMのフリーズを解除し、複数のタスクで学習を行う

- 日本語ASR
- 英語ASR
- シングルターンSFT
- マルチターンSFT
- 環境音キャプション生成

DPO
 音声化したDPOデータでSFT済みモデルでDPO
 DPOはとりあえずやってみたって感じて効果があつたかは微妙

4. 評価

対話性能評価
 ADU-Bench (Large Audio Language Modelの対話評価ベンチ) の日本語subsetで評価。evaluatorはGPT-4

音声認識評価
 CommonVoice 8 (ja) の CER(文字誤り率)で評価

まとめ・今後の展望

- 音声認識でWhisper-large-v3を上回る性能
- MistralやGoogleのオープンソースモデルと同程度の対話性能
- より詳細な評価と分析が必要。Speech-to-SpeechやDuplex全体

Model	対話性能(ADU-Bench)と音声認識性能(CER%)の評価	
	ADU-Bench (ja) ↑	CER (%) ↓
Whisper-large-v3	-	8.51
SALMONN	1.37	-
Qwen-Audio-Chat	1.08	-
Voxtral Mini-3B-2507	5.181	15.65
Gemma-3n E4B-it	5.143	51.23
JSpeechLLM-SB-Transcribe	-	8.36
MistralLLM-SB-SFT	5.335	10.25
JSpeechLLM-SB-DPO	5.165	10.42

アプローチ

利用可能な日本語音声資源が少ないため、提供頂いた計算資源を活用し合成データを作成し使用

ねらい

- 日本語に強い音声入力対応モデルは少ない
- LLM-JP-4の日本知識を活かした音声対話を実現したい
- 対話モデルとASR特化モデルを構築

1075h

合成した日本語
音声対話データ

43k

音声
キャプション

多話者

話者多様性のある
合成音声

学習の流れ

Pretraining

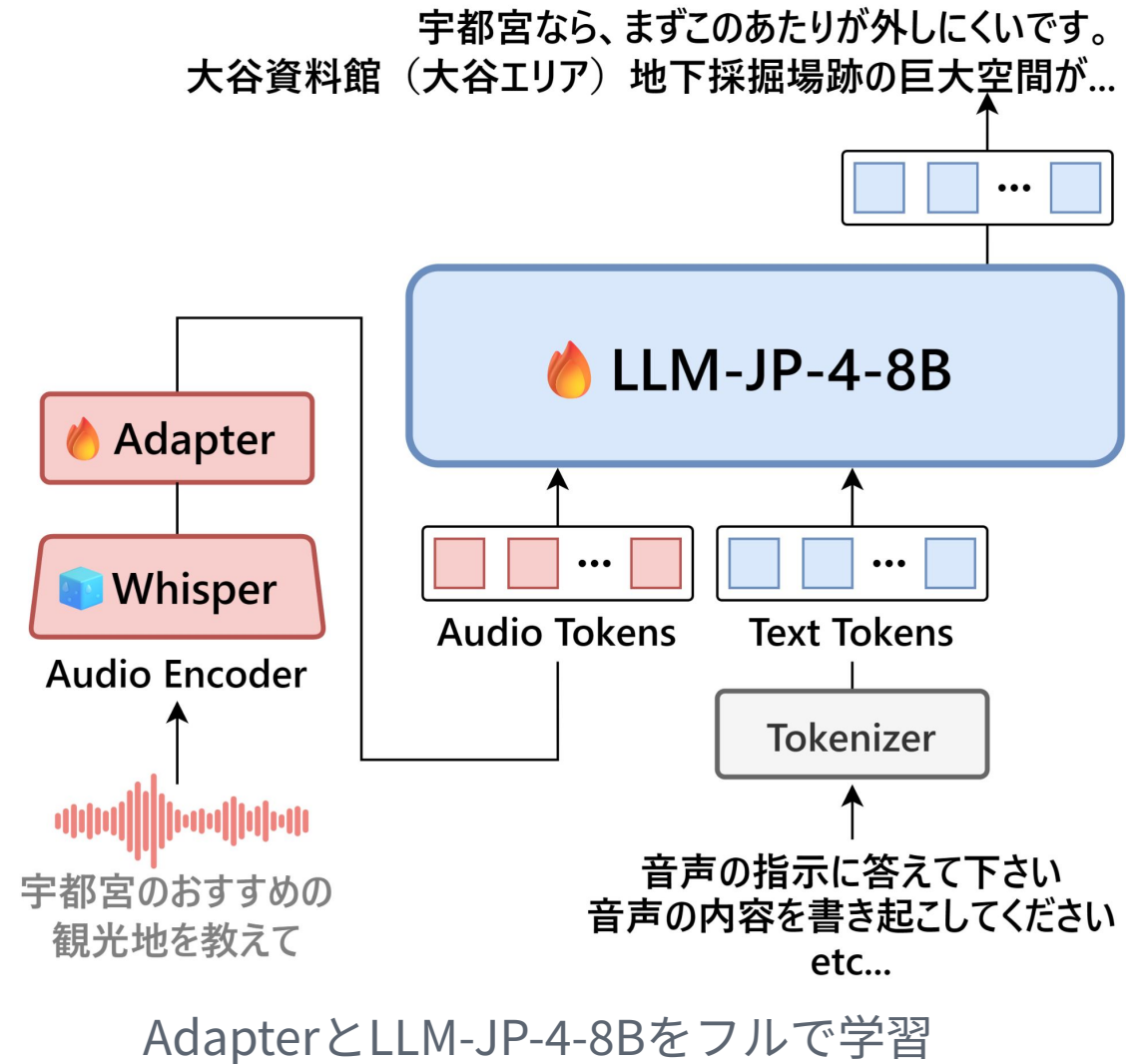
Adapterを
音声認識で学習

SFT

対話・音声キャプ
ション等を学習

DPO

音声化した選好
データで調整



デモ

A5000で推論しているデモ

対話

カスタムプロンプト

テンプレートから選択
(選択してください)

カスタム指示文
任意の指示文を入力...

推論パラメータ

サンプリングを有効化

Temperature 1

Top-p 1

Max tokens 1024

クリア

Model: Atotti/LlamaForSpeechLM-ja-DPO-Full-step8000

見どころ

- 日本語の音声指示に有益な応答ができる
- LLM-JP-4由来の日本知識を活用した応答
- 1つのモデルでASRと対話のタスクの双方が可能

データセット公開済み
モデルはApache2.0で公開予定