End-to-End 音声基盤モデルの開発・事業成果概要

実施者	株式会社Kotoba Technologies Japan
事業概要	End-to-Endで音声を汎用的に入力・出力し処理することのできる音声基盤モデルを1.データの準備、2. 基盤モデルの開発、3. 学習モデル評価のステップに分けて行う。

1.データセットの構築: インターネット上から50万時間分の高品質、日本語・英語混合データを収集する。
2. 基盤モデルの開発: 上記のデータセットを活用して、{1.3B, 7B} の音声基盤モデルの分散並列学習を行う。下記の評価セットで特に日本語のドメインで最先端の性能を達成する。最終的に7Bのモデル学習を成功させ、TTS Arenaなどで汎用性において日本語最高性能を目指す。
3. 学習モデル評価構築: 音声基盤モデルの評価を整え、日本語音声下流評価タスクを新たに作成しコミュニティに貢献する。東北大学と連携して、今後の基盤音声モデル開発の礎となる評価データセットを作る。

社会実装イメージ

当事業で開発された音声基盤モデルは、APIやB2B音声アプリケーションとして、社会実装をすすめる。本モデルは、リアルタイムで音声出力から入力まで一貫して行えるモデルであることを重視している。
これが実現されれば、コールセンター、広告、教育、福祉、医療、エンターテインメントなど、音声を用いるアプリケーションにおいて、大規模な音声AIの利活用が期待できる。

事業成果

1. データセットの構築: 日英合算で50万時間のハイクオリティ訓練データを構築した。特に日本語においては、従来にないスケールで、データ収集及びクリーニングのパイプラインを確立し、HuggingFace DatasetsなどのAI分野のスタンダードとなるツールも駆使しながら、データ構築の基盤技術を完成させた。
2. 基盤モデルの開発: 1.3B, 7Bのモデルの訓練は完了し、7Bモデルにおいては国内外の最大級の日英音声基盤モデルとなり、TTS Arena、S2S Translation Arena, 音声Retrieval、音声Perplexityにおいても最高ベースラインを遥かに上回る性能を記録した。日本語音声生成においてOpenAIやGoogle TTSを上回る評価を獲得した。また日英の音声翻訳においても非常に優れた精度を発揮してデモビデオとしてシステムを公開した。
3. 学習モデル評価構築: 日本語音声において世界初の試みであるTTS Arena Japaneseをリリースし、1200件以上の評価ポイントが集まった。また、コールセンターを意識したSpeaker Clusteringデータセットも構築した。