第2章 筆者提供の仮想環境で試せる

オープンソース音声生成AI VALL-E-Xのインストール

佐藤 聖

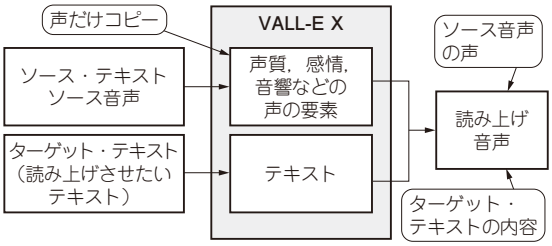

図1 VALL-E Xの処理イメージ

本章と次章で，自分の声や指定の声でテキストを読み上げる音声を生成する実験を行います．本章では環境設定について説明します．

マイクロソフトのVALL-E Xの論文やサンプル集などから作られたオープンソース実装の音声生成AIであるVALL-E-Xを使います．名称が非常に似ていて，「X」の前に「-」があるかないかの違いですが，製作者は異なりますので，本稿ではこの名称で使い分けます．

VALL-E-Xは従来の音声合成と同等もしくは高品質な音声をPC上で作成できます．音声ファイルは，WAV形式で保存されます．

● 使用する音声生成AI

VALL-E Xはマイクロソフトが開発した音声生成AIです．人間そっくりのリアルな音声を生成できます．音声生成や音声合成のサービスやツールはおそらく，最も優れた品質の音声生成AIの1つです．しかし，悪用が懸念され，現在はアプリケーションやソースコードなどを非公開となっています．しかし論文やサンプル集は公開されていています．それを基にPlachtaa氏らが作成したオープンソース実装のVALL-E-Xを本稿では使って音声生成を行います．**図1**にVALL-E Xの処理イメージを示します．

オープンソース実装のVALL-E-Xは，英語・中国語・日本語の音声を生成することができます．日本語に対応した音声生成AIのオープンソース実装は少ないため，とても貴重なツールです注1．

● 実験内容

VALL-E-Xは，ウェブ・ブラウザで利用することも，Pythonプログラムから使うこともできます．今回は後者を行います．

Pythonプログラムからソース音声かソース・テキストと読み上げさせたいテキストをVALL-E-Xに渡して，ソース音声の声でテキストを読み上げた音声ファイルを出力します．読み上げさせたいテキストを変更したい場合でも，Pythonプログラムが読み込む設定ファイルを書き換えるだけで出力する音声を作れます．

例えば，Pythonプログラムを修正して，設定用のテキスト・ファイルの代わりにExcelシートを読み込むようにできます．Excelシートのセルに書き込んだ読み上げさせたいテキストをPythonプログラムで連続的に読んで，その音声ファイルを大量に生成するような使い方ができます．

● VALL-E-Xの実行要件

VALL-E-X公式ページ(1)によると，VALL-E-Xの要件として，**表1**に示すツールを使用します注2．オープンソースのツールですので更新速度が早いため，執筆時点の情報から変わっている可能性があります．基本的にVALL-E-XはPyTorchをベースにしていますので，PyTorchのタスクをCUDA（Compute Unified Device Architecture，エヌビディア）で演算処理する環境が必要になります．本稿でもGPUを使います．

CPUのみで実行する場合，公式ページの

注1：生成可能な音声は，Demo of reproduced VALL-E Xのページ(2)で試聴できます．その中に日本語の音声サンプルがあります．

注2：エヌビディア以外のGPUでのVALL-E-Xの動作は未検証です．VALL-E-Xの音声生成は，PyTorchを使って音声生成モデルを実行し，テキストから音声を推定しています．PyTorch側でエヌビディア製GPU以外で処理できるように環境整備することで，エヌビディア製GPUと同様にVALL-E-Xを実行できるかもしれません．また，一般的にWindowsよりもLinuxの方が，CUDA関連やPyTorchの処理が速くなります．

仮想環境の設定については本誌サポート・ページをご覧ください．
https://interface.cqpub.co.jp/202408_onsei/