複数の生成 AI を組み合わせた語学学習支援ツールの開発

白水 美歌† 大坪 竜馬† 勝瀬 郁代†
† 近畿大学産業理工学部情報学科

1. はじめに

本研究では，様々な生成 AI の学習済みモデルを利用した語学学習支援 Web アプリケーションを開発した．

語学学習者は，自分の好みの外見や声，性格を持つバーチャル教師と対話することができる．バーチャル教師なら，余計な気遣いが必要なく，学習者の都合に合わせていつでも利用できる．また，学習者各人の嗜好を反映させた教師を相手とした語学学習は，学習意欲の維持に良い影響を与えると考えられる．

バーチャル教師の口の動きは，音声発音と対応付けて生成されるため，語彙や文法の技能習得だけでなく，動画を参考にして，発音の学習もできる．また，語学学習者側は，音声入力だけでなく，テキスト入力も可能にしているため，外国語でのスピーチ原稿の読み上げをバーチャル教師に依頼し，その動画を参考にスピーチの練習を行うなど，会話以外の用途にも活用できる．

なお，現時点ではまだ語学学習者側の音声入力は実現できておらず，テキスト入力のみである．

2. 利用した学習済みモデル

本研究では，現在，次の4つの生成 AI の学習済みモデルを利用している．

(1) Latent Diffusion Model (LDM)[1]:プロンプトを与えて顔画像を生成する．
(2) ESPnet2-TTS [2]:音声合成．108 種類の声質を選べる Multi-speaker Model を利用する．
(3) Sad Talker [3]:顔画像と音声からスピーチ動画を生成する．また，リップシンク機能により，音声に合った口の動きを生成できる．
(4) OpenAI ChatGPT API:語学学習者との対話におけるバーチャル教師の返答をテキストで生成する．

3. Web アプリケーションの実装とモデルの連携

Web アプリケーションの実装には Flask を利用した.図 1(a)は，最初の入力画面である．図のように，入力欄は2つある．語学学習者はまず，上段の入力欄に，好みのバーチャル講師の人物描写に関する記述を英語で行う．年齢，性別，容姿，性格などを記入すると，それらをプロンプトとして LDM により顔画像が生成される．

ESPnet2-TTSの声質選択については，現在，プロンプトから声質を呼び出せる機能がない．そこでまず，108 種類の声質に，自分たちで，性別や声の印象などの言語ラベルを付与した．人物描写のプロンプトとこれら言語ラベル間の意味的な距離を Word Mover's Distance [4]を用いて測り，距離が最小だった声質が選択されるようにした．このようにして，語学学習者の嗜好を反映させたバーチャル講師を生成できる．

次に語学学習者は，下段の入力欄に，バーチャル講師への対話文を入力する．これがプロンプトとして ChatGPT に送られて返答文が生成される．ESPnet2-TTS は，この返答文を，先ほど選択された声質の音声に変換する．この合成音声と先ほど LDM が生成した顔画像が Sad Talker に送られ，バーチャル講師のスピーチ動画が生成されて，図 1(b)のように，Web アプリケーション上で再生される．

(a)プロンプトと質問文の入力 (b)生成された動画の表示
図1. 生成 AI を利用した Web アプリケーション

4. 今後の課題

音声入力の実装と，非同期通信の導入によるアプリケーションの応答性を向上させたい．また，バーチャル講師の人物描写を日本語でも入力できるようにしたい．

5. むすび

本研究では,自然言語処理,音声合成,画像生成,動画生成の学習済みモデルを用いた語学学習支援 Web アプリケーションを開発した．生成 AI の進化が多彩な応用可能性を示す今，語学学習などの教育分野に新たな展望をもたらすと期待している．

参考文献

[1] R. Rombac, et al., "High-Resolution Image Synthesis with Latent Diffusion Models," arXiv:2112.10752, 2021.
[2] T. Hayashi, et al., "ESPnet2-TTS: Extending the Edge of TTS Research," arXiv:2110.07840, 2021.
[3] W. Zhang, et al., "SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation," arXiv:2211.12194, 2022.
[4] M. Kusner, et al., "From Word Embeddings to Document Distances," ICML2015, pp. PMLR 37:957-966, 2015.