ImTTS：印象推定可視化を用いた多話者音声合成システム

理学専攻　情報科学コース　2240644　岡本 美柚　(指導教員：伊藤 貴之)

1　はじめに

文章を入力として音声を生成する技術であるテキスト音声合成 (Text to Speech: TTS) は，企業や個人による利用の拡大に伴い，その需要も多様化している．深層学習の台頭以降，ひとつのモデルで多数の話者の音声を合成することができる多話者 (Multi-Speaker: MS-TTS) に関する研究も活発に取り組まれている．本研究は，多数の話者音声の中からユーザが抱く印象に近い音声を探索し，テキスト音声合成を実現するためのインタラクティブなシステムを提案するものである．

今後，ユーザのニーズと生成できる合成音声の多様化が進むと，現在利用可能な TTS システムの多くでは，①話者選択のための音声試聴の繰り返し作業がユーザにとって負担になって満足のできる選択ができず，②選択の支援としてシステム開発者によって提供された印象語タグは，必ずしもユーザの感性を反映しないという問題がある．この 2 つの問題によって，ユーザの意図しないコミュニケーションのギャップを引き起こすことが考えられるため，ユーザ個人が声から抱く印象は，TTS において無視できない要素である． 本研究では, ユーザが抱く印象に沿って対話操作によって話者を選択できる MS-TTS システム「ImTTS」を提案する．

2　提案手法

2.1　システムの概要

本研究では，図 1 に示すような TTS システムを提案する. また，本研究で実際に作成したユーザインタフェースを図 2 に示す．画面左部の散布図には，ユーザが事前に構築したモデルにより推定された印象値を可視化する．点は話者を示し，配置は x-vector[1] から得られる．この散布図表現では，類似した特徴量を持つ話者が近くに配置され，ユーザが効率的かつ直感的に話者を選択することに役立つ．話者選択時にクリックすることで合成モデルに話者の ID が渡される．画面右部のテキストボックスには合成に用いるテキストを入力する．テキストボックス下のボタンで，散布図の着色に用いる印象項目を変更できる．

本研究で用いた音声合成モデルは，Takamichi ら提供する YouTube の動画など利用可能性が不明なデータから自動構築された音声コーパス [2] を学習データとし，Seki らによって構築された約 1500 人もの話者の音声を合成可能な MS-TTS モデルを用いた [3].

2.2　声質を考慮した散布図可視化

提案手法では，話者の情報を, 声質を表現する特徴量をもとに 2 次元空間上で可視化することにより，ユーザは声質の特徴を考慮し，多量の話者情報を確認しながら選択することができる．また，ユーザ個人の感性に適応した印象推定モデルを構築し，推定された印象を散布図の点の色の濃淡で可視化することにより，ユーザは自身が抱く印象に紐づけ，多話者の特徴を視覚的に確認しながらな話者選択が可能である．

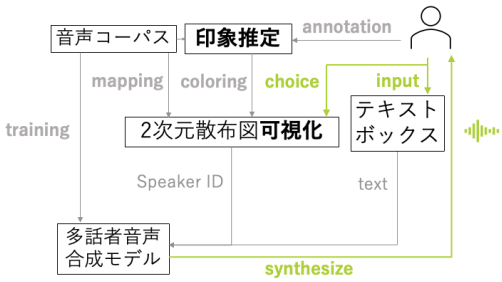

図 1: ImTTS のシステム図

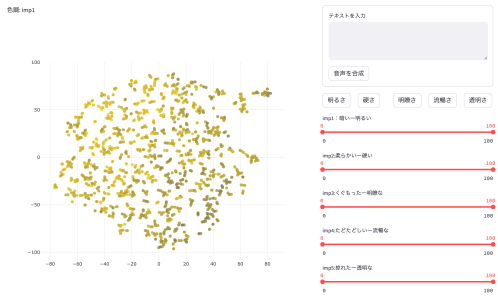

図 2: 音声合成のためのユーザインタフェース

2.3　散布図評価値の算出

散布図評価スコアは，クラスタ分離性を示す客観評価指標「Pseudo.F」[4] と本研究にて定義した近傍話者の類似性を示す主観評価指標「Subjective Similarity」をそれぞれ標準化して足し合わせた式 (1) とした．

$$\text{Scatter Score} = \frac{bcss}{wcss} \cdot \frac{N-k}{k-1} + \frac{\sum_{i=0}^{N} s(i,j)}{N} \quad (1)$$

2.4　実験結果

t-SNE と UMAP，それらの前処理に PCA を行う組み合わせ手法の中から，本研究で算出した散布図評価が上位手法であった手法を，表 1 で降順に掲載している．最も散布図評価スコアが高かったのは，t-SNE の perplexity をデフォルト値である 30 から大きく変更したものであった．

表 1: 散布図評価スコアが上位であった次元削減手法

Method：Parameter	PCA
t-SNE: perplexity=5,early_exaggeration=15	-
t-SNE: perplexity=5,early_exaggeration=15	○
t-SNE: perplexity=5,early_exaggeration=10	-

3　Active Learning を用いた印象推定

3.1　印象推定モデルの概要

提案手法では，Active Learning を導入したニューラルネットワークを用い，ユーザの感性を反映した印象推定モデルを構築する．手順を図 3 に示す．まず,

400 個のデータに対して複数人によるアノテーションを実施し，初期モデルを作成する．次に，全印象の初期モデルから出力される印象値の平均が中間的であったデータをデータプールからランダムで 20 個選出し，次モデルのためのアノテーションを行う．データプール内には，全発話の音声と 10 次元に削減した x-vector と音声データのペアが格納されている．

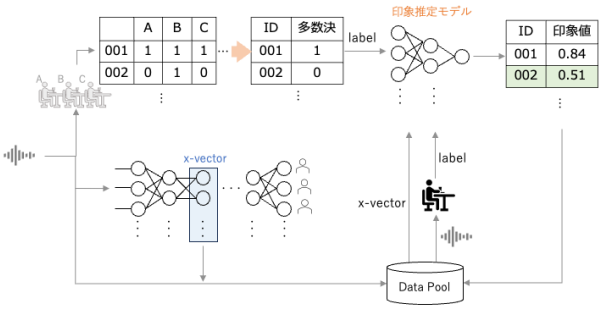

図 3: 印象推定モデル構築までの手順

3.2 著者のデータを用いた実験結果

図 4 には，Active Learning により構築されたモデルから得られた印象値と真値との二乗誤差を可視化したものを示す．部分的に，濃い青で表示されている誤差の小さい話者が固まって存在している．しかし，赤色で表示されている大きな誤差がところにより存在してしまっている．このことから，精度にばらつきがあることがわかる．

また，図 5 に、モデルの重みについて EMD(Earth mover's distance) を用いて距離算出し，構築された距離行列より，Active Learning の試行回数ごとのモデル間の距離を可視化したものを示す．試行回数を重ねるごとにモデル間の距離が近づいていくことが望ましい．可視化結果より，7-8 は近づいたのにも関わらず 8-9 で遠ざかってしまっていることがわかる．しかし，可視化の左上部に前半の番号，右下部に後半の番号が表示されており，全体的には試行回数によってモデルが一定の方向に変化していることが読み取れる．

3.3 被験者のデータを用いた実験結果

20 代男女 2 名ずつの印象推定モデルを構築し，3.2 節と同様にユーザごとのモデル間の距離を可視化した結果を図 6 に示す．赤点が女性被験者，青点が男性被験者のモデルを示す．男女で左右に分かれて可視化されていることから，提案手法によって構築されるモデルは男女の感性の違いを考慮できていることがわかる．

4 まとめ

本研究では，実験にて検討した結果より，散布図表現を用いた話者表示とユーザが話者の声質に対して持つ印象値の推定可視化により話者選択を行うユーザインタフェース，MS-TTS モデルを用いた提案システムを構築した．今後は，印象推定精度を改善するため，アルゴリズムを再検討する必要がある．

謝辞： 本研究にあたり，多くの助言を賜った東京大学大学院の関健太郎様，高道慎之介講師，齋藤佑樹助教に感謝いたします．

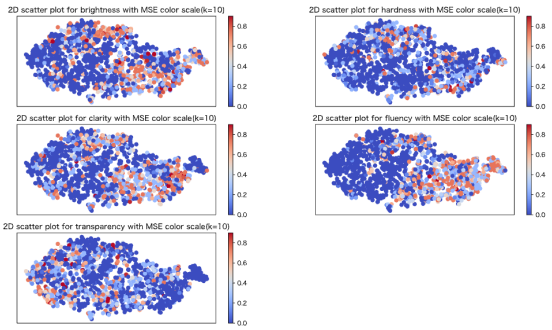

図 4: ActiveLearning 後モデルによる話者ごとの印象推定誤差可視化

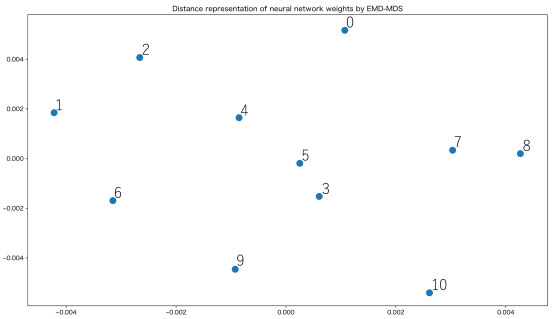

図 5: Active Learning 試行回数ごとのモデル間距離関係の散布図可視化

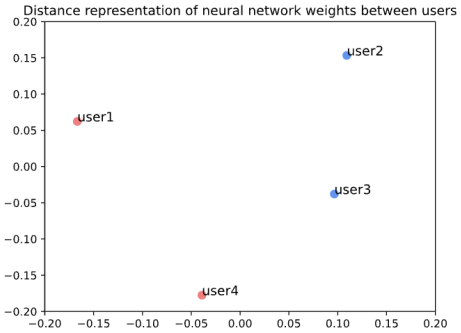

図 6: ユーザごとのモデル間距離関係の散布図可視化

参考文献

[1] D. Snyder et al., X-vectors: Robust DNN enbeddings for speaker recognition, ICASSP, pp. 5329-5333, 2018.

[2] S. Takamichi et al., JTubeSpeech: Corpus of Japanese speech collected from YouTube for speech recognition and speaker verification, arXiv:2112.09323, 2021.

[3] K. Seki et al., Text-to-speech synthesis from dark data with evaluation-in-the-loop data selection, ICASSP, 2023.

[4] T. Calinski et al., A Dendrite Method for Cluster Analysis, Communication in statistics, 3, pp. 1-27, 1974.