4S-05

対話時の心理的距離を縮めることを目的とした発話内容変換の提案とその検証

滝田巧平† 青柳西蔵† 平井辰典†
駒澤大学†

1. はじめに

本研究では，発話者のスピーチスタイル（話し方）を声質を維持したまま別のスピーチスタイルへと変換するための枠組みである発話内容変換技術を提案する．本技術は音声認識によって音声をテキストに変換し， データベースを用いて別の話し方へと変換した後，x-vector という話者認証モデルによって抽出した話者埋め込みを条件付けとしてテキストから話者の声質を維持して音声を合成する．

発話内容変換技術の提案に加えて本研究では，提案した発話内容変換を用いて敬体と常体の話し方に注目する実験を行う．先行研究において，松本ら[1]は，日本語学習者支援を目的に常体から敬体への敬語変換を行うシステムを提案しているが，本研究では，敬体から常体への変換をすることでスピーチスタイルシフトを発生させ，対話時における話者間のスピーチスタイルに関する変化や心理的な距離を縮めることができるかを調査する．また，一対一の遠隔通話を想定した実験を実施し，その結果を考察する．

2. 発話内容変換の提案

2.1 発話内容変換技術の詳細

本研究で提案する発話内容変換は，一つのスピーチスタイルから敬語や方言，役割語などあらゆる話し方へ発話者の声質を維持したまま変換を行う枠組みである．開発した発話内容変換は，音声認識・データベース・音声合成の三つの技術で構成されている．

最初に，音声認識により発話者が入力した音声を文字起こしする．音声認識には Watanabe らによる音声処理ツールキット[2]の学習済みモデルを短いコードで利用できる ESPnet Model Zoo を利用した．この実装では，ESPnet で提供されている Transformer の音声認識モデルを利用した．

認識後，テキストを別のスピーチスタイルへと変換する．言語情報の変換に関してはデータベースを用いて辞書に登録した形態素及び文節から変換を行う．MeCab によってテキストの分かち書きを行った後に各形態素及び文節をユーザが登録したデータベースから検索し，該当する変換情報があればテキストを変換する．

テキストを変換した後，処理されたテキストはテキスト音声合成手法によって合成して出力する．図 1 にテキスト音声合成モデルの外観を示す．音素からメルスペクトログラムを生成するための音響モデルには Tacotron2 [3]に用いられる音響モデルを使用した．この時，発話者の声質で音声を合成するために，話者認証モデルである x-vector [4]によって音声から抽出した話者埋め込みを音響モデルの Encoder の出力に結合することで学習時に存在しない発話者の声を合成した．そして，メルスペクトグラムから音声波形を生成するニューラルボコーダには Parallel WaveGAN [5]を用いた．Parallel WaveGAN によって高速に品質の高い音声を合成できる．

音響モデルと Vocoder は JSUT コーパス[6]及び JVS コーパス[7]を用いて学習をおこない，x-vector は濵田らが提供する学習済みモデル*を用いる．

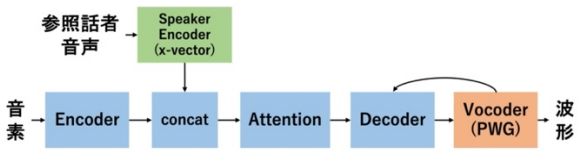

図 1 音声合成モデルの外観

2.2 発話内容変換により期待される対話支援

日本語では相手に礼節を示す為に敬語が用いられる．また，仲が良い間柄では砕けた会話表現が用いられることも多い．しかし，必ずしも全ての会話で一つのスピーチスタイルが用いられるわけではない．対話時にスピーチスタイルシフトが起こる場合がある．スピーチスタイルシフトとは，敬体で話していた発話者が話し方を常体へと変えること，またはその逆のシフトを表す．これが起こる要因の一つとして，生田らは，心的態度の変化の表明があると論じる[8]．常体から敬体へのシフトは心的距離の短縮（親しみ・同調等）を表し，常体から敬体へは心的距離の伸長（プライベートに踏み込まない意思表示等）を表す．

以上のように心理的距離が短縮した時，そのシフトが発生するが，距離が短縮していない時に提案アプリによって意図的にシフトを発生させることで対話者が心理的距離の短縮を錯覚させられるのではないかと考えられる．

3. 敬体から常体へのシフト実験

3.1 実験に使用する発話内容変換アプリ

敬体から常体へのスピーチスタイルシフトを人為的に起

The propose and verification of Speech Content Conversion for reducing psychological distance during conversation
†KOHEI TAKITA, SAIZO AOYAGI, TATSUNORI HIRAI, Komazawa University
* https://github.com/sarulab-speech/xvector_jtubespeech

図 2　発話内容変換アプリ

こすことで対話者間の心理的な距離が縮まるかどうかを調査する実験を行った．本項では，実験に用いたアプリについて記述する．

本実験で用いたアプリを図 2 に示す．本アプリは左上のマイクボタンを押すことで録音が始まり，ボタンを離すことで認識・変換・合成の一連の処理が行われる．発話者が喋り終わってから音声が合成されるまでには，文章の長さに応じて 3 秒から 6 秒ほどの遅延が生じる．

3.2　実験の詳細

前述のアプリを用いた実験は，駒澤大学に在籍する面識のない二人の学生に協力してもらって行った．実験は，ビデオ会議システムを使って約 8 分間会話をしてもらい，発話者の一人に前述したアプリを使用してもらい，もう一人の発話者にはアプリを用いず会話をしてもらった．本アプリでは遅延が発生するため，アプリ未使用者にも 3 秒から 6 秒ほどの遅延を付与した．被験者の二人に対しては発話内容が敬体から常体へ変換されることを実験前には開示せず，音声合成を使った実験であるとだけ説明を行った．図 3 に実際に実験で交わされた会話の一部を示す．図中の赤太字はシステムを通じて会話時に相手に伝わった言葉を表す．

実験後，二人の心理的距離が短縮したと思うかについてアンケート及び聞き取り調査を実施した．「友達口調になった時に親しみやすさを感じたか」というアンケートに対してアプリ未使用者は親しみやすさを感じたと答えた．また，「馴れ馴れしさを感じたかという」質問に対して「感じなかった」と答えた．心理的距離の短縮に本技術の影響もあるのではないかと考えられる．また，聞き取りを行うと「敬体と常体の変化に会話の最中では気付かなかった」との回答もあった．気付かなかった理由の一つとしてシステムの遅延が大きく，遅延を考慮した会話に夢中になってしまったことが挙げられた．また，アプリを使用していた話者で『自分の声が合成音になることによって友達口調で話してみようかなと思った』という回答も見られた．合成音による発言の代替は，心理的距離の短縮に寄与するのではないかと考えられる．

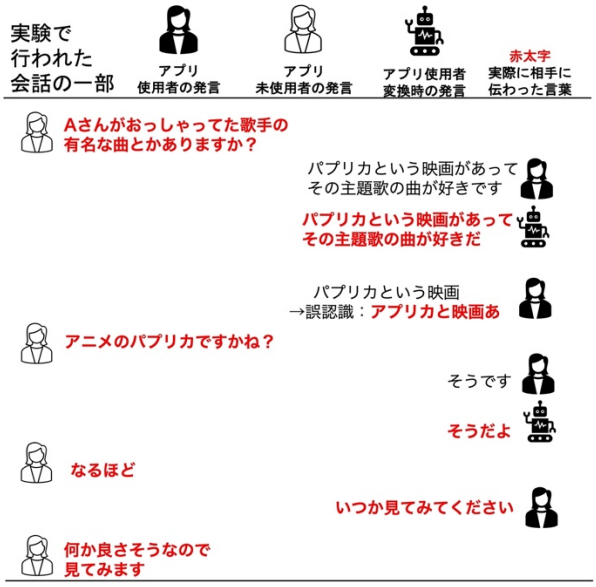

図 3　実験で行われた会話の一部

4.　おわりに

本研究では，あるスピーチスタイルを別のスピーチスタイルへと変換する発話内容変換を提案し，敬体から常体へのスピーチスタイルシフトに着目した実験を実施した．

結果として，アプリの遅延が大きく，対話への集中力が削がれてしまった．遅延の短縮は今後の課題である．しかし，親しみやすさを感じるという回答や合成音を挟むことによって友達口調で話してみようと思ったという回答が得られたことで心理的距離が縮まることへの可能性が見えた．今後はアプリの改良とさらなる調査を行なっていきたい．

参考文献

[1] 松本悠太ほか：“日本語学習者支援のための敬語変換タスクの提案.” 第 36 回人工知能学会全国大会 (2022)
[2] Watanabe, S., et al.: ESPnet : End-to-End Speech Processing Toolkit, Interspeech, pp. 2207-2211 (2018)
[3] Shen, J., et al.: Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions, ICASSP, pp.4779-4783 (2018)
[4] Snyder, D., et al.: X-Vectors: Robust DNN Embeddings for Speaker Recognition, ICASSP, pp. 5329-5333 (2018)
[5] Yamamoto, R., et al.: Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram, ICASSP, pp.6199-6203 (2020)
[6] Sonobe, R., et al.: JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis, arXiv preprint, 1711.00354 (2017)
[7] Takamichi, S., et al.: JVS corpus.: free Japanese multi-speaker voice corpus, arXiv preprint, 1908.06248 (2019)
[8] 生田少子ほか: “社会言語学における談話研究”, 月刊言語 12(12), pp.77-84, 大修館書店 (1983)