発話内容に基づく目標印象を付与する話し顔生成

水野沙希　北条伸克　篠田一聡　鈴木啓太
庵愛　佐藤宏　田中智大　増村亮
NTT 株式会社 人間情報研究所
{saki.mizuno, nobukatsu.hojo, kazutoshi.shinoda, keitaxs.suzuki,
mana.ihori, hrs.sato, tomohiro.tanaka, ryo.masumura}@ntt.com

概要

本稿では，目標印象（好感が持てる/配慮がある）を与える話し顔生成を扱う．発話の印象は顔表情だけでなく発話内容にも依存する．しかし，従来の話し顔生成では，音声は主に口唇運動との同期のために用いられており，入力特徴として音韻情報が用いられるため，発話内容を考慮することができない．そこで本稿では，入力発話の音声認識結果から得られる言語情報を導入し，発話内容を考慮可能な話し顔生成法を提案する．評価実験では，各発話に印象ラベルを付与した音声・映像データセットを用いて，提案法と言語情報を用いない比較条件との性能を比較する．その結果，顔表情特徴量の推定誤差が低減し，言語情報の導入が有効であることを示す．

1　はじめに

面談や接客，商談など様々な対話場面で，話者の印象が対人評価や対話の進行に影響する．これまでにオンライン面接や会議を対象に，非言語情報の可視化・フィードバック・コーチングを行う会話支援システムが提案されている [1, 2, 3, 4, 5, 6]．一方で，相手に与えたい印象に合わせた支援は限定されている．このため，会話支援システムが話者の発話映像を変換し，「好感が持てる」といった目標とする印象を与えられる発話動画を作成・提示する支援は，コミュニケーション学習の観点から有用だと考えられる．例えば，利用者がロールプレイ等で話した動画を入力として，システムが目標印象を与えるように顔表情を変換した「手本動画」を生成し，利用者はその手本を参照しながら発話練習を反復する，といった利用形態が考えられる．

発話の印象は，顔表情に加えて発話内容にも依存して形成される．したがって，「好感が持てる」話し顔の場合でも，入力発話がポジティブな内容（例：「今日は楽しかった」）かネガティブな内容（例：「次は君と夕食を共にしたくない」）かによって，適切な出力表情は異なり得る．このため，目標とする印象を表現するには，入力発話の意味を考慮する必要がある [7, 8, 9]．しかし，従来の入力音声に基づいた話し顔生成では，口唇運動と音声の同期を主な目的として MFCC [10] 等の音響特徴や音声埋め込みを用いることが多く，発話内容を考慮することはできない [11, 12, 13, 14, 15]．本研究では，目標とする印象を付与することを目指して，発話内容を考慮した話し顔生成を提案する．

なお，本研究の基本的な枠組みおよび主要結果は ICASSP 2024 で発表済みであり，本稿はその内容を日本語で再整理し，目標印象を付与する話し顔生成タスクとしての位置づけを明確化する（[16]）．

本稿の貢献は以下の 3 点である．

- 入力発話の人物同一性・発話内容を保持しつつ，目標とする印象（好感が持てる／配慮がある）を付与する発話動画生成タスクを提案する．
- データ量が乏しい制約の下で，教師あり学習によって顔表情特徴量を変換することで発話映像を生成する枠組みを提示する．
- 音声認識結果から得た言語特徴を考慮することで，同一印象でも発話内容によって適切な表情が変化する手法を実現する．

2　印象変換技術

2.1　印象変換データセット

本手法の学習・評価のため，3 種類の印象「ニュートラル」「好感が持てる」「配慮がある」からなる印象ラベル付き音声・映像データセットを構築した．本稿では「ニュートラル」を，会話支援システムへ

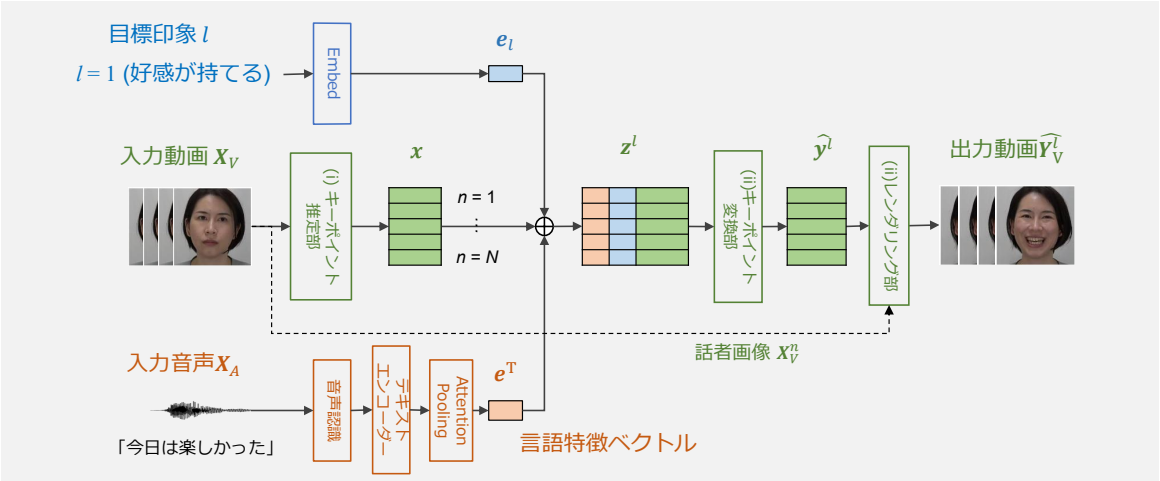

図 1　目標印象を付与する話し顔生成の処理フロー．

の入力発話として想定される，表情変化が比較的小さい発話として定義する．この設定は，実会話データの観察において，印象上の課題を生む発話の多くが「不快な表情」よりも「表情が乏しいこと」に起因していたという知見に基づく．「好感が持てる」「配慮がある」は，自己訓練で有用と考えられる対照的な目標印象として選定した．

印象変換器を学習するためには，同一の発話文で，印象の異なる発話動画のペアを教師データとして使用することが考えられる．このような発話動画を収録するため，本研究では，話者が事前に用意した発話文を指定された印象を与えるよう演技して読み上げる様子を録画する．発話文は，1on1 における上司の発話，お見合いにおける発話を想定し，作成し，各文を 3 種類の印象で演技して収録した．発話意味の多様性を確保するため，挨拶・自己紹介・業務依頼等のニュートラルな内容に加え，賞賛や良い知らせ等のポジティブ内容，否定や悪い知らせ等のネガティブ内容を含めた．自然発話らしさのためフィラーを挿入し，ライターにより計 1200 文（1on1: 620 文，お見合い: 580 文）を作成した．

収録は十分な表現力で，かつ一貫した演技を収録するため，俳優業に携わる 30 代女性 1 名，40 代男性 1 名を選定した．参加者はいずれも日本語話者である．各文を 3 印象で読み上げ，計 7200 クリップ（約 20 時間）を得た．音声は 24-bit/48kHz，映像は 1920×1080/30fps で収録した．

2.2　目標印象を付与する話し顔生成

本研究の目的は，入力話者の人物同一性と発話内容を保持したまま，顔表情を変換して目標印象を付与した発話動画を生成することである．入力を $\boldsymbol{X} = \{\boldsymbol{X}_V, \boldsymbol{X}_A\}$ とし，$\boldsymbol{X}_V$ は入力映像，$\boldsymbol{X}_A$ は入力音声を表す．また，$l \in \{1, 2\}$ をユーザが指定する目標印象ラベルとし，$l = 1$ を「好感が持てる」，$l = 2$ を「配慮がある」とする．本タスクでは，$\boldsymbol{X}_A$ を保持したまま，$\boldsymbol{X}$ と l に基づいて目標印象 l を与える出力映像 $\hat{\boldsymbol{Y}}_V^l$ を推定する：

$$\hat{\boldsymbol{Y}}_V^l = f(\boldsymbol{X}_V, \boldsymbol{X}_A, l), \tag{1}$$

ここで $f(\cdot)$ は印象変換器である．最終的な出力動画は $\hat{\boldsymbol{Y}}^l = \{\hat{\boldsymbol{Y}}_V^l, \boldsymbol{X}_A\}$ とする．

2.3　提案手法

印象は顔表情だけでなく発話内容にも依存するため，目標とする印象に合う表情を生成するには入力発話の意味情報を考慮する必要がある．従来行われてきた入力音声に基づいた話し顔生成で一般的に用いられる音響特徴や音声埋め込みは口唇同期に有効である一方，意味情報を明示的に扱いにくい [11, 12, 13, 14, 15]．そこで本研究では，入力音声の音声認識結果から得られる言語特徴量を導入し，発話内容に応じた顔表情変換を行う目標印象付与モデルを提案する．

映像を入出力とする変換器 $f(\cdot)$ を教師あり学習するには大量の教師データが必要である．多くの話し顔生成モデルは大規模音声映像データセットで学習される [17, 18, 19, 20] が，本稿で作成した印象変換データセットは小規模であるため，同様の手法での学習は困難である．そこで本研究では，印象変換を (i) キーポイント（顔表情特徴量）推定部，(ii) キーポイント変換部，(iii) レンダリング部の 3 段階に分解し，学習を低次元なキーポイント変換に集中させ

表 1　言語情報の有無による推定キーポイント特徴量と正解キーポイント特徴量間の MSE の比較．提案法（言語あり）は全条件で比較条件（言語なし）の推定精度を上回った（好感：好感が持てる，配慮：配慮がある）．

話者	話者 1				話者 2				平均
会話シーン	お見合い		1on1		お見合い		1on1		
与える印象	好感	配慮	好感	配慮	好感	配慮	好感	配慮	
入力映像	1.472	1.209	1.588	1.575	1.307	1.336	1.647	1.649	1.473
比較条件（言語なし）	0.598	0.564	0.661	0.683	0.707	0.692	0.722	0.764	0.674
提案法（言語あり）	**0.579**	**0.553**	**0.645**	**0.646**	**0.648**	**0.653**	**0.697**	**0.704**	**0.640**
改善率（%）	3.2 %	2.0 %	2.4 %	5.4 %	8.3 %	5.6 %	3.5 %	7.9 %	5.0 %

ることで少量データでの学習を可能にする．手法の全体像を図 1 に示す．

(i) キーポイント推定部　入力映像データ $\boldsymbol{X}_V$ からキーポイント特徴量時系列を推定する．キーポイント推定部では，既存技術（First Order Motion Model [21] など）を利用する．

(ii) キーポイント変換部　入力動画のキーポイント特徴量時系列と印象ラベルに加えて，入力音声の音声認識結果に基づく言語特徴を用いて，出力動画のキーポイント特徴量時系列を推定する．具体的には，印象ラベルを埋め込みベクトルに変換し，各フレームのキーポイント特徴に連結する．さらに，入力音声に音声認識より得た発話テキストをテキストエンコーダで符号化し，attention pooling により固定長の言語特徴ベクトルへ集約した上で，同様に連結する．これにより，発話内容に応じた適切な顔表情への変換が可能となる．

学習は，印象ラベル付きデータセットから作成した入出力動画ペアを教師データとして行う．この際，入出力動画のキーポイント特徴量時系列は，同じ時間フレームで同じ音素を発話するなど，時間的な対応関係が取られている必要がある．そこで本研究では，距離関数に音声のメルケプストラム特徴量間の二乗誤差を用いた DP マッチングを適用し，入出力動画間の時間対応を推定する．さらに，得られた対応に基づいて出力側動画を時間伸縮し，入力側と整合させる．時間整合後の入出力動画から推定したキーポイント特徴量ペアと印象ラベル，音声認識により得られた発話テキストを教師データとし，平均二乗誤差（mean squared error; MSE）の最小化によりキーポイント変換器を学習する．

(iii) レンダリング部　推定されたキーポイント特徴量時系列と，入力動画の話者の顔画像から，キーポイントが表す表情情報と人物同一性を併せ持つ出力映像を生成する．レンダリング部では，キーポイント推定部と同様に既存手法（First Order Motion Model [21] など）を利用する．

3　実験設定

3.1　比較条件

提案法（言語あり）では，入力音声の音声認識結果から抽出した言語特徴をキーポイント変換部の条件として導入する．これに対し，比較条件（言語なし）では，キーポイント特徴量と印象ラベルのみを入力としてキーポイント変換を行い，言語特徴を導入しない．言語特徴を導入しない条件を比較条件とした理由は，従来の話し顔生成が口唇運動と音声の同期を主目的とし，言語特徴を考慮しない設計が一般的であるためである [11, 12, 13, 14, 15]．比較条件（言語なし）は言語特徴を入力しない点のみが提案法（言語あり）と異なり，キーポイント抽出・レンダリング・変換器構成・学習データおよび学習条件は同一とした．

3.2　前処理

データセット中の各動画に対し，先行研究のレンダリング設定 [21] に従い，face-alignment ライブラリ [22] を用いて顔領域を切り出した．

3.3　入力特徴量

顔表情特徴（キーポイント特徴量）　顔表情特徴量として，First Order Motion Model [21] により各フレームから推定される 10 個の 2 次元キーポイント座標（20 次元）と，それらの Jacobian（40 次元）を連結した 60 次元ベクトルを用いた．学習および評価では，入出力のキーポイント特徴量を標準正規化して使用した．

言語特徴（提案法のみ）　提案法（言語あり）では，入力音声に対して音声認識を行い，得られた認

図 2　生成された動画の例（お見合い，好感が持てる）．

識結果をもとに言語特徴量を抽出する．音声認識器として MediaGnosis [23, 24] を使用した．言語エンコーダとして，文字トークンに基づく Transformer 4 層，出力ベクトルが 512 次元の BERT-like な事前学習済みモデルを使用した．テキストエンコーダの出力系列に attention pooling を適用し，固定長の言語特徴ベクトルを得る．印象ラベルの埋め込みベクトルおよび言語特徴ベクトルの次元は，それぞれ 2 とした．なお，音声認識およびテキストエンコーダのパラメータは学習中に更新せず，attention pooling 以降のパラメータのみを学習した．

3.4　モデル構成

キーポイント変換器は，1 層，ユニット数 256 の bidirectional LSTM とした．比較条件（言語なし）では，各フレームの入力としてキーポイント特徴量と印象ラベルを連結したものを用いる．提案法（言語あり）では，これに加えて固定長の言語特徴ベクトルを全フレームに共有して連結し，キーポイント変換を行う．キーポイント推定およびレンダリングには First Order Motion Model [21] を用いた．

3.5　学習条件

目的関数として MSE を使用し，Adam アルゴリズム [25] で学習した．提案法（言語あり）では印象ラベルの embedding，attention pooling 層，キーポイント変換器を学習し，比較条件（言語なし）では印象ラベルの embedding とキーポイント変換器を学習した．また，本研究では訓練セット 2000 ペア，検証セット 104 ペア，評価セット 232 ペアを用いてモデル学習を行った．評価は，会話シーン（1on1 ／お見合い）および目標印象（好感が持てる／配慮がある）ごとに実施した．

3.6　評価指標

客観評価指標として，標準正規化したキーポイント特徴量に対する MSE を用いた．MSE は推定キーポイント特徴量と正解キーポイント特徴量との差をフレームごとに算出し，それらを平均することで求めた．得られた MSE を，話者・会話シーン（1on1 ／お見合い）・各目標印象（好感が持てる／配慮がある）ごとに集計した．

4　実験結果

4.1　客観評価（MSE）

表 1 に，推定キーポイント特徴量と正解動画のキーポイント特徴量の MSE を示す．比較条件（言語なし）および提案法（言語あり）はいずれも入力映像より MSE が小さく，本研究の枠組みにより目標印象に対応する表情変換が可能であることが示される．さらに，提案法（言語あり）は全ての条件で比較条件（言語なし）の推定精度を上回り，特に話者 2 において最大 8.3%の改善が見られた．発話内容に基づく言語特徴量を導入することで，目標印象に対応するキーポイント推定精度が改善することが確認された．

4.2　生成例

図 2 に，同一入力に対する生成例を示す．提案手法では，特に目元や口元の動きにおいて不自然さが緩和され，発話内容に整合した表情が得られる傾向が観察された．

5　おわりに

本稿では，自己訓練に用いる手本生成を目的として，入力発話の人物同一性と発話内容を保持したまま目標印象（好感が持てる／配慮がある）を付与する発話動画生成タスクとして印象変換を定義した．また，小規模なデータセットでの話し顔変換手法として，キーポイント推定・キーポイント変換・レンダリングの三段階に分ける枠組みを提示し，少量データでのモデル学習を可能にした．さらに，印象が発話内容にも依存する点に着目し，音声認識結果に基づく言語特徴を条件として導入することで，発話内容に応じた表情変換を行う目標印象付与モデルを提案した．評価実験により，言語特徴を導入することでキーポイント推定誤差が一貫して低減し，言語情報が印象変換に有効であることを示した．今後は，話者・状況の多様性を拡張したデータ収集と主観評価の実施により，実運用における印象付与効果を検証する．

参考文献

[1] Keith Anderson, Elisabeth André, Tobias Baur, Sara Bernardini, Mathieu Chollet, Evi Chryssafidou, Ionut Damian, Cathy Ennis, Arjan Egges, Patrick Gebhard, et al. The tardis framework: intelligent virtual agents for social coaching in job interviews. In **International Conference on Advances in Computer Entertainment Technology**, pp. 476–491. Springer, 2013.

[2] Mohammed Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W Picard. Mach: My automated conversation coach. In **Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing**, pp. 697–706, 2013.

[3] Markus Langer, Cornelius J König, Patrick Gebhard, and Elisabeth André. Dear computer, teach me manners: Testing virtual employment interview training. **International Journal of Selection and Assessment**, Vol. 24, No. 4, pp. 312–323, 2016.

[4] Samiha Samrose, Ru Zhao, Jeffery White, Vivian Li, Luis Nova, Yichen Lu, Mohammad Rafayet Ali, and Mohammed Ehsan Hoque. Coco: Collaboration coach for understanding team dynamics during video conferencing. **Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies**, Vol. 1, No. 4, pp. 1–24, 2018.

[5] Samiha Samrose, Daniel McDuff, Robert Sim, Jina Suh, Kael Rowan, Javier Hernandez, Sean Rintel, Kevin Moynihan, and Mary Czerwinski. Meetingcoach: An intelligent dashboard for supporting effective & inclusive meetings. In **Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems**, pp. 1–13, 2021.

[6] Samiha Samrose and Ehsan Hoque. MIA : Motivational Interviewing Agent for Improving Conversational Skills in Remote Group Discussions. **PACM on Human-Computer Interaction**, Vol. 6, No. GROUP, pp. 1–24, 2022.

[7] Lucía Teijeiro-Mosquera, Joan-Isaac Biel, José Luis Alba-Castro, and Daniel Gatica-Perez. What your face vlogs about: expressions of emotion and big-five traits impressions in youtube. **IEEE Transactions on Affective Computing**, Vol. 6, No. 2, pp. 193–205, 2014.

[8] Amirmohammad Kazameini, Samin Fatehi, Yash Mehta, Sauleh Eetemadi, and Erik Cambria. Personality trait detection using bagged svm over bert word embedding ensembles. **arXiv preprint arXiv:2010.01309**, 2020.

[9] Dipika Jain, Akshi Kumar, and Rohit Beniwal. Personality bert: A transformer-based model for personality detection from textual data. In **Proceedings of International Conference on Computing and Communication Networks: ICCCN 2021**, pp. 515–522. Springer, 2022.

[10] Steven B Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, Vol. 28, No. 4, pp. 357–366, 1980.

[11] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 3661–3670, 2021.

[12] Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. One-shot talking face generation from single-speaker audio-visual correlation learning. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 36, pp. 2531–2539, 2022.

[13] Se Jin Park, Minsu Kim, Joanna Hong, Jeongsoo Choi, and Yong Man Ro. Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 36, pp. 2062–2070, 2022.

[14] Sanjana Sinha, Sandika Biswas, Ravindra Yadav, and Brojeshwar Bhowmick. Emotion-controllable generalized talking face generation. **arXiv preprint arXiv:2205.01155**, 2022.

[15] Jianrong Wang, Yaxin Zhao, Li Liu, Tianyi Xu, Qi Li, and Sen Li. Emotional Talking Head Generation based on Memory-Sharing and Attention-Augmented Networks. In **Proc. INTERSPEECH 2023**, pp. 2–6, 2023.

[16] Saki Mizuno, Nobukatsu Hojo, Kazutoshi Shinoda, Keita Suzuki, Mana Ihori, Hiroshi Sato, Tomohiro Tanaka, Naotaka Kawata, Satoshi Kobashikawa, and Ryo Masumura. Talking face generation for impression conversion considering speech semantics. In **ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 8411–8415. IEEE, 2024.

[17] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. VoxCeleb: a large-scale speaker identification dataset. **Telephony**, Vol. 3, pp. 33–039, 2017.

[18] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep speaker recognition. **Proc. Interspeech 2018**, pp. 1086–1090, 2018.

[19] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. **IEEE transactions on affective computing**, Vol. 5, No. 4, pp. 377–390, 2014.

[20] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In **European Conference on Computer Vision**, pp. 700–717. Springer, 2020.

[21] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. **Advances in Neural Information Processing Systems**, Vol. 32, , 2019.

[22] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In **Proceedings of the IEEE international conference on computer vision**, pp. 1021–1030, 2017.

[23] Ryo Masumura, Mana Ihori, Akihiko Takashima, Takafumi Moriya, Atsushi Ando, and Yusuke Shinohara. Sequence-level consistency training for semi-supervised end-to-end automatic speech recognition. In **ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 7054–7058. IEEE, 2020.

[24] MediaGnosis. https://www.rd.ntt/mediagnosis/.

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, 2014.