発話訓練経験による文章発話時の顔ランドマーク変位の違い*

☆安田 奈央，北村 達也 (甲南大)

1　はじめに

発話時の調音運動と顔の動きの間には関連があることが知られている．発話中の顔ランドマークの動きと舌運動をそれぞれOPTOTRAK等の光学的装置と磁気センサシステムにて同時計測した研究により，これらの相互作用が報告されている [1, 2, 3]．また，Tang ら [4] は，Clear speech と通常の発話における顔ランドマークの動きを計測し，これらの 2 つの発話スタイルにおいて顔の動きに差異があることを示した．

近年では，画像認識 AI の性能向上が著しく，先行研究にて用いられていた光学的装置を用いずとも動画からリアルタイムに顔ランドマークの座標が得られるようになった．これを利用すれば，発話スタイルと顔の動きの関係を従来より手軽に調査したり，発話訓練のフィードバックとして活用したりできると考えられる．

本研究では，プロのナレーター，発話訓練経験のある大学生，発話訓練経験のない大学生の 3 群を対象に，発話時の顔画像を収録し，これらの間の顔ランドマークの動きの差異を調査する．そして，顔ランドマークの動きを発話訓練に利用する可能性を検討する．

2　方法

2.1　話者

発話訓練の訓練や経験が異なる以下の 3 群の話者，計 18 名が参加した．

A 群 プロダクションに所属するプロのナレーター 2 名 (男性 1 名，女性 1 名)

B 群 発話訓練経験のある放送サークルに所属する大学生 (男性 1 名，女性 5 名)

C 群 発話訓練経験のない大学生 (男性 5 名，女性 5 名)

本研究において話者内の発話スタイルの違いを比較しなかったのは，発話訓練経験のある話者は普段の発話も明瞭性が高く，発話訓練経験のない話者の発話とは異なるケースがあるためである．

2.2　発話資料

イソップ童話「北風と太陽」の文章を発話資料とした．読み上げに要する平均的な時間は約 30 秒であった．

2.3　収録方法

実験に先立ち，書面により実験の説明を行い，同意書に署名を得た．

話者は椅子に座り，タブレットアームで固定したタブレットに表示された発話資料を読み上げた．A 群，B 群の話者には「滑舌よくアナウンサー風の発話で」と指示し，C 群の話者には「友達と話すような普段通りの発話で」と指示した．話者の顔の正面から約 0.4 m の位置にビデオカメラ (Panasonic HC-V360MS) を固定し，30 fps，画像サイズ 1080×1920 にて動画を撮影した．

動画とは別に音声のみの収録も行った．コンデンサマイク (RODE NT2-A)，オーディオインタフェース (Roland Rubix24) を用いて標本化周波数 48 kHz，量子化ビット数 24 bit にて録音した．

2.4　顔ランドマークの抽出および分析

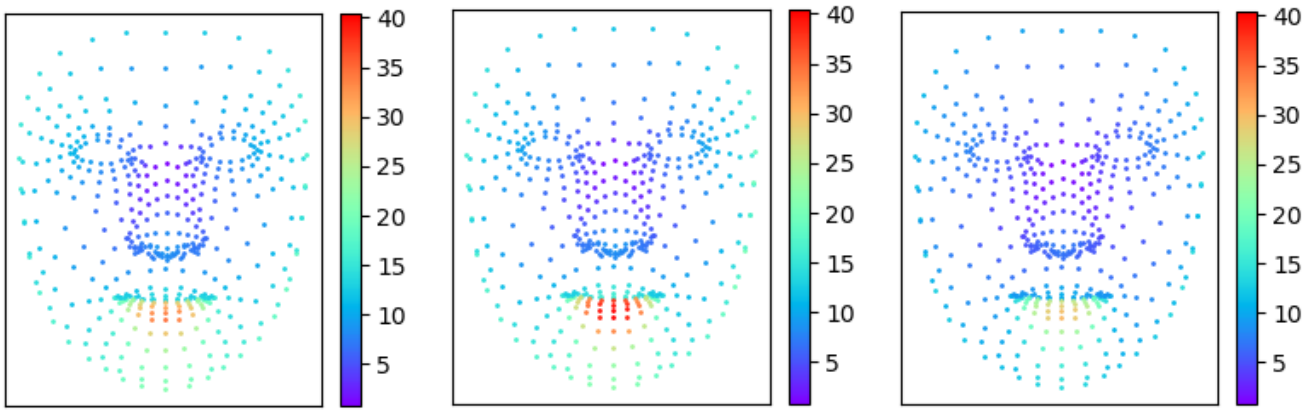
MediaPipe Face Mesh [5] は，画像認識クラウド AI の一種で，静止画あるいは動画から高速に 468 点の顔ランドマークの 3 次元座標を推定することができる．動画中の発話区間を抽出し (文と文との間の非発話区間は除去)，Face Mesh にて各フレームから顔ランドマークを求め，各点の移動距離を求めた．その際，発話中の頭部の動きを除去するため，鼻根点の座標を原点とした．また，話者ごとの動画内の顔の相対的サイズの差異に対応するため，左右の外眼角 (目尻) の距離を 100 として正規化した．そして，各点の 1 s あたりの移動距離を求めた．

3　結果と考察

3.1　XY 平面上の動き

「北風と太陽」を読み上げた時の群 A，B，C の話者の各顔ランドマークの XY 平面上の平均変位 (絶対値) を図 1 に示す．各点の変位は，各群内で平均したものである．この図から群 A，B

*Effects of voice training experiences on facial landmark movements during reading sentences. by YASUDA, Nao and KITAMURA, Tatsuya (Konan Univ.)



(a) Group A (b) Group B (c) Group C

Fig. 1 Mean displacement of each facial landmarks on the XY-plane for Groups A, B, and C.

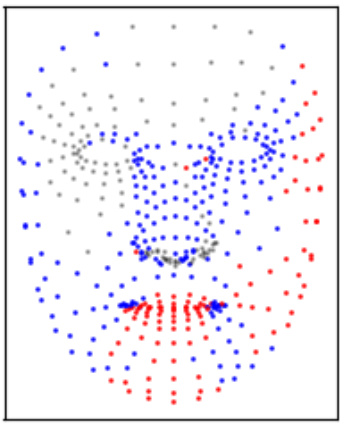


Fig. 2 Results of Mann-Whitney *U*-test between Groups A and B. Red: significant at the 1% level. Blue: significant at the 5% level. Gray: non-significance.

では群 C と比較して口唇および下顎の動きが大きいことがわかる．これは，Tang ら [4] の結果から予想されたことである．

図 2 は，各点について群 B と群 C の間で Mann-Whitney U 検定を行い，有意水準 1 %で有意差があった点を赤，有意水準 5 %で有意差があった点を青，有意差があるとはいえなかった点をグレーで示したものである[1]．この図から，発話訓練の有無によって口唇周辺，下顎，頬の領域の動きが有意に大きくなることがわかった．

なお，図 2 では鼻梁に有意水準 5 %の有意差が見られる．図 1 に示すように鼻梁における変位が元々小さいことから，一部の話者における顔の動き等に起因するランドマークの認識誤差が原因ではないかと考えている．

[1]群 A は話者が 2 名のみのため，統計処理の対象としなかった．

3.2 Z 軸上の動き

Face Mesh により推定された顔ランドマークの Z 軸方向の平均変位 (絶対値) を群間で比較したが，有用な情報は得られなかった．

4 おわりに

本研究では，発話訓練の訓練や経験の異なる話者が文章を読み上げた際の顔ランドマークの動きの差異を調査した．その結果，発話訓練の有無や発話スタイルの違いにより，発話時の口唇周辺，下顎，頬の領域の変位に差異が表れることを示した．今後，話者を増やして精度を高めるとともに，発話訓練への応用を検討する．

謝辞 本研究は JSPS 科研費基盤研究 (A)「ポップアウト・ボイスの生成・知覚基盤の解明に基づく高性能拡声音技術の開発」(JP20H00291) の助成を受けた．収録にご協力いただいた広島大学 山根典子先生，牧野桃子様に感謝します．

参考文献

[1] Yehia *et al.*, *Speech Commun.*, 26, 23–43 (1998).
[2] Jiang *et al.*, *EURASIP J. Advances in Signal Processing*, 1174–1188 (2002).
[3] Beskow *et al.*, *Proc. ICPhS 2003*, 431–434 (2003).
[4] Tang *et al.*, *Speech Commun.*, 75, 1–13 (2015).
[5] Kartynnik, *et al.*, arXiv:1907.06724 (2019).