音声ピッチから類推される発話者の顔表象の画像化 —Classification Image を用いて—

鈴木悠介 立命館大学 OIC 総合研究機構
山﨑大暉 京都大学大学院文学研究科
永井聖剛 立命館大学総合心理学部

人は声から発話者の人物像を推測するが，その際にどのような顔が想像されるのか，その詳細は明らかでない。本研究では Classification Image を用いて，高低ピッチ音声から想像される顔表象を画像化し，その顔が喚起する社会的印象を検討した。基本周波数またはフォルマント周波数を操作した高低ピッチ音声から生成された顔画像を比較した結果，具体的な音響特徴によらず，高い声は女性的で信頼しやすい顔を，低い声は男性的で支配的な顔を想像させることが示され，この傾向は参加者間で一貫していた。さらに想像された顔は，支配性と信頼性という二つの社会的評価軸に沿って評価されており，声からの印象はまず支配性に影響し，そこから信頼性の評価へと波及する可能性が示唆された。声から想像される顔は，発話者の実際の顔の復元というより，音声ピッチと結び付いたステレオタイプにより形成されると考えられる。

Keywords: Classification Image, voice-face matching, voice pitch, fundamental frequency, formant frequency.

問題・目的

声は顔と同様に，発話者のアイデンティティや身体情報など様々な人物情報を伝達する。実際，声と顔は共通の神経生理的基盤を持ち，音声処理領域と顔処理領域は直接的な情報共有を行っていることが示唆されている（von Kriegstein et al., 2005）。しかし，声のみから発話者を同定することは容易ではなく，その照合精度はチャンスレベル程度に留まる（Kamachi et al., 2003）。それにもかかわらず，聴き手は声から発話者の容貌を想像することができる。では，聴き手は声の情報を手がかりにどのような顔表象を形成しているのだろうか。

音声知覚において音声ピッチ情報は最も顕著な特徴であり，発話者の多様な情報を伝達する（Aung & Puts, 2020）。一般に音声ピッチ印象は，基本周波数（Fo）に依存するが，フォルマント周波数（Fn）によっても規定される（Puts et al., 2006）。すなわち，Foの低下も，Fnの低下も「低い声」というピッチ印象をもたらす。しかし，各々が伝達する情報は異なる可能性がある。Fnは声道長という解剖学的構造に規定され，発話者の身体・顔形態を比較的正確に反映する。対して，Foは形態特徴との相関が弱く，むしろ威嚇など社会的支配性の誇張のための社会的信号として機能すると考えられる（Pisanski et al., 2014）。一方で，聴き手はこれら音響特徴を混同し，FoまたはFnの低下をともに，「低い声」として知覚し，男性的で支配的というステレオタイプを形成しやすいことも示されている（Ohala, 1994）。

したがって，高/低く聞こえる声（高低ピッチ音声）から想像される顔表象が，主にFnから得られる顔形態情報に基づいて形成されているのか，それともFoとFnの区別なく寄与する，音声ピッチ印象と紐づいたステレオタイプに基づいているのか，その形成メカニズムは明らかではない。そこで本研究では，Classification Image（Dotsch & Todorov, 2012）を用いて，FoとFnを独立に操作した声から想像される顔の視覚的表象をそれぞれ画像化し，比較を行った。

実験1: 音声から想像される顔の画像化

方法　大学生および大学院生68名（実験1A: M_{age} = 20.882, 女性: 19名, 男性: 15名; 実験1B: M_{age} = 21.176, 女性: 30名, 男性: 4名）が実験に参加した。参加者は，Fo操作群（実験1A）またはFn操作群（実験1B）のいずれかに割り当てられた。

男性8名の平均顔をベースイメージ，平均声/a:/をベース音声とした。ベース音声に対し，±3半音のFo変調（Fo操作）または，スペクトル包絡を周波数軸上で1.2/0.8倍に伸長/短縮（Fn操作）することで，高/低ピッチ音声刺激を作成した。

課題の各試行では，注視点の提示とともに高または低ピッチ音声が再生され，続いて極性反転したノイズパターンがベースイメージに重畳された 2 枚の顔画像が左右に提示された。参加者は，提示された音声の発話者によりふさわしいと感じた方の顔画像を選択するよう求められた。高/低ピッチ音声の提示試行はそれぞれ 320 試行あり，全体で 640 試行であった。高/低ピッチ音声試行毎に，選択された顔画像群を平均し，高/低 Fo または Fn CI とした。また各群の全参加者にわたって平均することで Fo/Fn Group CI を作成した。

結果・考察　Figure 1A に高/低 Fo Group CI および Fn Group CI を示した。参加者ごとに生成された CI のクラスター分析の結果，操作した音響特徴にかかわらず，CI 画像は高低音声ピッチに基づき，明確に異なるカテゴリーを形成した（$\chi^2 = 97.978$, $p < .001$, Cramer's $V = .470$）。また顔特徴に対する線形混合モデルを用いた解析の結果，低ピッチ CI は，高ピッチ CI と比べ，下顎幅が広く（F (1, 86) = 26.092, p_{adj}

< .001, R_p^2 = .185），口角が下がっている（F (1, 85) = 30.019, p_{adj} < .001, R_p^2 = .422）ことが示された。これらの結果は，参加者が音響特徴の違いによらず，音声ピッチ印象に関連したステレオタイプ的な顔表象を形成している可能性を示唆する。

実験2: CI画像の社会的印象評価

方法　参加者はクラウドワークスで募集され，212名が実験に参加し，155名（M_{age} = 26.290, 女性: 73名, 男性: 79名, 他性別: 3名）を解析対象とした。

参加者は，実験1で作成された高/低FoおよびFn CI（合計136枚）について，顔から推測される声のピッチ，性別，年齢，信頼性，支配性，魅力のいずれか一つをVisual Analogue Scale（0-100）で評価した。各CI画像は2回提示され，合計278試行（CI画像136枚×2回＋注意チェック6試行）実施した。ただし，同一のCI画像が連続して提示されることを防ぐため，全試行を2つのブロックに分割した。各ブロックはCI画像の全評価試行と注意チェック3試行を含み，その提示順はブロック内でランダム化された。

結果・考察　Figure 1Bに各評価項目における平均評価値を示した。評価結果から，声から想像される顔の社会的印象は，操作した音響特徴によらず，音声ピッチ印象によって一貫して決定づけられることが示された（各評価項目: ps < .05）。具体的には，高ピッチ音声からは女性的で信頼性の高い顔が，低ピッチ音声からは男性的で支配的な顔が形成された。またCI画像および評価者間の評価一致度（ICC）を分析したところ，性別や年齢等の身体的特性に加え，支配性において，評価者の個人差を越えて，CI画像間で高い共通性が確認された。これは，音声ピッチ印象と支配性を結びつけるステレオタイプ（周波数コード: Ohala, 1994）が，顔表象の形成に影響することを示唆する。

さらに評価得点に対する探索的因子分析およびパス解析から，声から想像される顔の評価構造は，既存の声および顔の評価モデル（Oosterhof & Todorov, 2008）と同様に，支配性と信頼性の二軸に集約された。加えてパス解析では，音声ピッチ印象が直接的に支配性印象を形成し，信頼性および魅力印象は，支配性印象を媒介して，間接的に形成されるモデルが支持された（$\chi^2(2)$ = 7.290, p = .026, CFI = .990, TLI = .949, SRMR = .019）。つまり，高低ピッチ音声から想像される顔は，単に音声の印象が顔表象へ転移し形成されるのではなく，周波数コードに基づく支配性印象が主導的な役割を果たし，関連して他の社会的印象が構成されている可能性がある。

全体考察

本研究では，Classification Imageを用いて，高低ピッチ音声から想像される顔表象を画像化し，その社会的印象を検討した。その結果，音響特徴の違いにかかわらず，低ピッチCIは，高ピッチCIと比較して，下顎幅が広く口角が下がるといった顔特徴を有し，より男性的かつ支配的な社会的印象を喚起することが示された。さらに，支配性評価はCI画像間で高い一致を示し，個人差を超えて集団レベルで共有されていた。これらの結果は，音響特徴から得られる形態情報よりも，音声ピッチに基づくステレオタイプが顔表象に反映されたことを示唆する。加えて，声から想像される顔表象の印象構造の分析から，支配性と信頼性の二軸からなる既存の二因子モデルと整合する構造が確認された。特に，音声ピッチはまず支配性印象の形成に寄与し，信頼性印象は支配性印象を媒介して間接的に形成する可能性が示唆された。

本研究は，声から想像される顔表象が，個別の音響特徴よりも音声ピッチによって一貫して形成されることを示した。聴き手は，形態特徴を正確に反映する音響特徴（Fn）とそうでない特徴（Fo）を区別せず，声の低さと支配性を対応づけるステレオタイプを優先して顔表象を構築していたと考えられる。本研究の結果から，声に基づく顔表象の形成は，発話者の実際の顔形態特徴を忠実に復元するのではなく，脅威や支配性の要因を重みづけ，発話者の安全性や信頼性を方向づける適応的なプロセスである可能性が指摘される。

引用文献

Aung, T., & Puts, D. (2020). *Curr. Opin. Psychol.*, *33*, 154-161.

Dotsch, R., & Todorov, A. (2012). *Soc. Psychol. Person. Sci.*, *3*, 562-571.

Kamachi, M. et al. (2003). *Curr. Biol.*, *13*, 1709-1714.

Ohala, J. J. (1994). In *Sound symbolism* (eds Hinton et al.), 325–347, CUP.

Oosterhof, N. N., & Todorov, A. (2008). *Proc. Natl. Acad. Sci.*, *105*, 11087-11092.

Pisanski, K. et al. (2014). *Anim. Behav.*, 95, 89-99.

Puts, D. A. et al. (2006). *Evol. Hum. Behav.*, *27*, 283-296.

von Kriegstein, K. et al. (2005). *J. Cogn. Neurosci.*, *17*, 367-376.

Figure 1. 高低 Fo および Fn 音声から想像される顔画像（Group CI）と，社会的印象評価の平均得点