話者属性が視聴覚統合判断および統合表象に与える影響 —McGurk 効果を用いた CIMS モデルに基づく検討—

The Effects of Speaker Characteristics on the McGurk Effect: A CIMS-Based Analysis

氏家 悠太†，高橋 康介‡
Yuta Ujiie, Kohske Takahashi
† 立教大学，‡ 立命館大学
Rikkyo University, Ritsumeikan University
yujiie@rikkyo.ac.jp

概要

本研究では, McGurk 効果における話者の顔と声の属性（成人・子供）の影響を，CIMS モデルを用いて検討した。主な結果として, 子供の声が提示された条件では因果推定確率と錯覚率が低下し，視聴覚情報を同一の原因と判断する傾向が弱まることが示された。一方で, 知覚精度には差が見られず，視聴覚統合の判断は話者属性などのより高次な因果的意味づけの影響を受ける可能性が示された。

キーワード：視聴覚統合, CIMS モデル，McGurk 効果

1. 問題と目的

McGurk 効果とは，ある発話映像（ka）に，調音位置を合わせて異なる音声（pa）を重畳し観察者に呈示すると, 観察者には視覚, 聴覚のどちらでもない聞こえ（ta）が生じる錯覚である（McGurk & MacDonald, 1976）。この錯覚は音声知覚における視覚の影響を端的に表す現象であり, 感覚入力時の SN 比のように聴覚情報や視覚情報の明瞭度が変化することで錯覚率に影響することが示されている（e.g. Ujiie & Wakabayshi, 2022）。

近年の McGurk 効果の研究では，話者の顔や声の親近性が錯覚率に影響することが示唆されている（e.g. Ujiie & Takahashi, 2022; Walker et al, 2015）。Ujiie & Takahashi (2022) では，他人種効果を利用して，話者の顔の親近性を操作し，顔の親近性が高い民族の話者刺激では, 親近性が低い民族の話者刺激と比べて, 顔と音声の統合が促進され錯覚が多く生起することを示している。

一方で, 視聴覚統合の因果推定モデルでは, 視覚と聴覚の入力情報が不一致な状況で，知覚がいずれかの感覚情報に引き寄せられるのか，それとも新たな知覚表象として統合されるのかは，感覚入力の物理的な特徴だけでなく，観察者による因果推定（Causal inference）に依存するとしている（e.g. Körding et al., 2007; Shams & Kim, 2010）。とくに，McGurk 効果では,「その音声と口形が同一話者によるものかどうか」という信号源推定に関する主観的判断が，統合の有無に強く関与するとされている（Magnotti & Beauchamp, 2013）。実際，発話映像が男性で音声が女性のように，視覚と聴覚での話者属性が異なる場合には錯覚率が低下することが示されている（e.g. Green et al., 1991）

本研究では，年齢による親近性の効果（The other-age effect, Kuefner et al., 2007）に基づき，成人と児童の話者刺激を用いて話者との親近性を操作することで，McGurk 効果における話者の顔と声の属性の影響を検討した。特に本研究では統合の判断と統合表象への影響を調べるために，視聴覚統合の因果推定モデルの一つである CIMS（Causal Inference in Multisensory Speech）モデル（Magnotti & Beauchamp, 2017）を用いて検討した。CIMS モデルは，McGurk 効果などの視聴覚統合において, 人が複数の感覚情報を「同じ原因に由来する」と推定するかどうかを，複数の処理段階を経て確率的に判断するモデルである（Magnotti & Beauchamp, 2017）。このモデルでは，感覚ごとの信頼性に基づく重みづけに加え，情報が同一の原因に由来すると推定する確率, およびそれらのパラメータをもとに最終的な知覚判断を行う確率が算出でき, McGurk 効果の錯覚が生じる背景にある知覚処理や判断過程を，数理的に予測できる点が特徴である。

本研究では，特に因果推定において，以下の 3 つの仮説を設定した。第一に, 観察者にとって親近性の高い成人話者（顔・声）において視聴覚統合判断の判断が促進されるという「親近性仮説」（e.g., Ujiie & Takahashi, 2022），第二に，子供の声は明瞭度が低いため視覚への依存が高まり統合判断に影響を及ぼすとする「明瞭性仮説」（e.g., Sumby & Pollack, 1954），第三に，顔と声が一致しているときに統合が促進されるとする「一致性仮説」（e.g., Green et al., 1991）とした。

2. 方法

実験参加者　日本人の大学生 26 名が実験に参加した（分析ではデータの不備により 1 名を除外）。

刺激と手続き　実験では成人話者 1 名および子供話者 1 名の発話場面（音素は「ぱ」「た」「か」の 3 種）を撮影し，音声のみの音声単独刺激（成人または子供の声），発話映像のみの視覚単独刺激（成人または子供の発話映像），調音位置をもとに音声と発話映像を組み合わせた 4 種類の視聴覚不一致刺激（音声「ぱ」＋視覚「か」の McGurk 刺激）と視聴覚一致刺激を作成した。実験課題は，音声単独条件，視覚単独条件，視聴覚条件（一致刺激と不一致刺激）の 3 ブロックで構成した。各ブロックでは刺激はランダム順で提示され，刺激の繰り返し数は 10 回とした。各試行において，観察者は提示された刺激に対して知覚した音素を，「ぱ」「た」「か」の選択肢から選択して回答を行った。これに加え，視聴覚刺激の顔と声の一致性判断課題も行った。なお，本研究では，音声単独条件，視覚単独条件，視聴覚不一致条件のデータのみを分析の対象とした。

分析方法　音声単独および視覚単独条件のデータから混同マトリクスを作成し，下記の式に基づき，音声および視覚単独刺激に対する知覚ノイズ（σ_A，σ_V）を，それぞれについて最尤推定法により算出した。ここで $\phi(x;\mu,\sigma)$ は平均μ，標準偏差がσの正規分布の確率密度関数を表し，x_sとx_r はそれぞれ刺激と反応カテゴリに割り当てた知覚空間上の値とした（例：「ぱ」＝-1，「た」＝0，「か」＝+1）。

$$P(r \mid s,\sigma) = \frac{\phi(x_r;x_s,\sigma)}{\sum_{r'}\phi(x_{r'};x_s,\sigma)}$$

視覚と聴覚に対して推定された知覚ノイズ（σ_A, σ_V）をもとに統合表象（感覚の重みづけ; X_{AV}）を下記の式に基づき算出した。

$$X_{AV} = \frac{x_A/\sigma_A^2 + x_V/\sigma_V^2}{1/\sigma_A^2 + 1/\sigma_V^2}$$

次に，視聴覚情報が同一の原因に由来すると判断する確率$P(C=1 \mid x_A, x_V)$を，視覚と聴覚の入力の差分をもとに，以下の式により推定を行った。

$$P(C=1 \mid x_A, x_V) = \frac{p_{\text{common}}\cdot\phi\left(\Delta_{AV};0,\sigma_A^2+\sigma_V^2+\sigma_C^2\right)}{p_{\text{common}}\cdot\phi\left(\Delta_{AV};0,\sigma_A^2+\sigma_V^2+\sigma_C^2\right)+(1-p_{\text{common}})\cdot\phi(0;0,\sigma_A^2)\cdot\phi(0;0,\sigma_V^2)}$$

最後に，上記の因果推定に基づき，以下の式を用いて，視聴覚統合が生じた場合と生じなかった場合のカテゴリ判断確率を加重平均して錯覚の報告確率を算出した。

$$P(\text{ta}) = P(C=1)\cdot P(\text{ta} \mid X_{AV}) + \big(1-P(C=1)\big)\cdot P(\text{ta} \mid x_A)$$

また，上記の指標（感覚表象，因果推定，錯覚率）における話者属性の影響を検討するため，話者の声（成人・子供）と顔の属性（成人・子供），および交互作用を固定効果，被験者をランダム効果とした線形混合効果モデル（linear mixed-effects model; LMM）を用いて，各指標（統合表象，因果推定値，錯覚率）の条件間差を分析した。なお，知覚ノイズに関しては，各条件の話者属性（成人 vs. 子供）の違いについて，対応のある t 検定により検討を行った。

3. 結果

音声単独条件および視覚単独条件のデータから推定された知覚ノイズ（σ_A，σ_V）の平均値を図 1 に示した。各条件において，知覚ノイズを指標として t 検定を行った結果，いずれの条件においても話者属性による有意差は示されなかった[音声単独条件, $t(24) = .44, p = .66, d = .12$，視覚単独条件，$t(24) = 1.16, p = .25, d = .30$].

次に，視覚と聴覚の知覚ノイズをもとに統合表象（感覚の重みづけ; X_{AV}）を刺激条件別に算出し，図 2 に示した。話者属性の影響を検討するため，話者の顔と声の属性，およびその交互作用を固定効果，被験者をランダム効果とする線形混合効果モデルにより解析を行った。その結果，話者の顔や声の効果は有意でなく（顔: $\beta = -.15$，95% CI [−.58, .28]，$p = .501$，声: $\beta = -.16$，95% CI [−.60, .27]，$p = .454$），交互作用も有意でないことが示された（$\beta = .04$，95% CI [−.57, .65]，$p = .906$）。

図 1. 刺激条件別の知覚ノイズ（σ_A，σ_V）

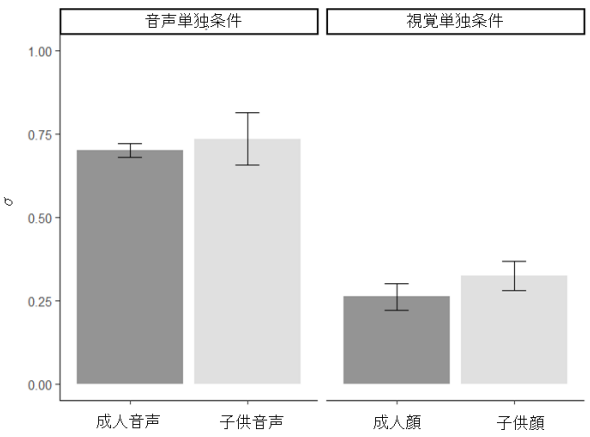


図 2. 刺激条件別の統合表象（感覚の重みづけ; X_{AV}）

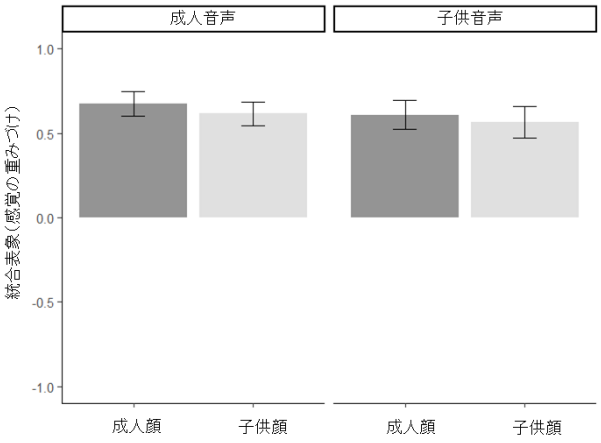


図 3. 刺激条件別の因果推定の確率(P(C=1))

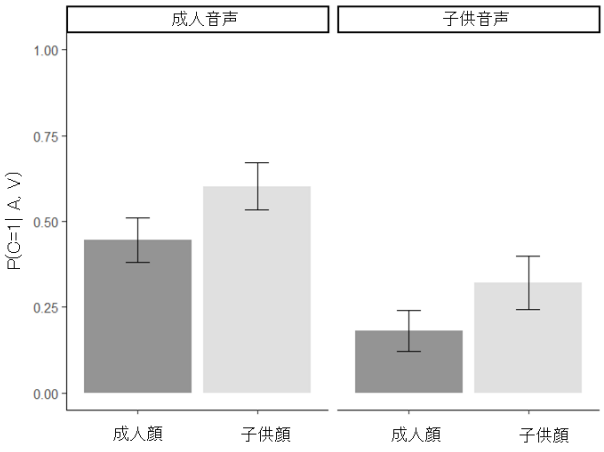
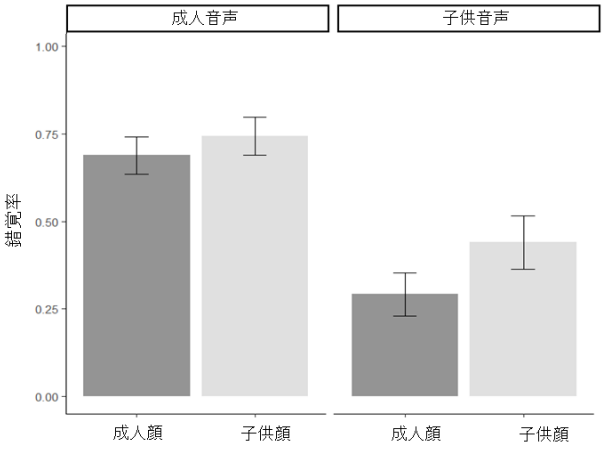


図 4. CIMS モデルから予測された錯覚率



視覚と聴覚の情報が同一の信号源とする因果推定の確率（P(C＝1)）を実際の錯覚率データから推定し，刺激条件別に図 3 に示した。話者属性の影響について，線形混合効果モデルを用いて解析を行った結果，話者の声の効果が有意であり，刺激に子供音声が含まれた場合は成人音声と比べて，因果推定の確率が有意に低いことが示された（β = −.71，95% CI [−1.16, -.27]，p = .002）。一方で，話者の顔の効果や顔と音声の交互作用は有意でなかった（顔: β = .42，95% CI [−.21, .52]，p = .062，交互作用: β = −.05，95% CI [−.68, .58]，p = .88）。

CIMS モデルに基づき，因果推定確率，統合表象，知覚ノイズを用いて，刺激条件ごとの錯覚率の予測値を算出した（図 4）。錯覚率の予測値を指標として，話者属性の効果を線形混合効果モデルにより検討した結果，話者の声の効果が有意であり，刺激に子供音声が含まれた場合は成人音声が含まれる場合と比べて，錯覚率が有意に低下することが示された（β = −1.11，95% CI [−1.50, -.72]，p < .001）。一方で，話者の顔の効果や顔と音声の交互作用は有意でなかった（顔: β = .16，95% CI [−.23, .54], p = .422, 交互作用: β = .26, 95% CI [−.29, .80], p = .351）。さらに，CIMS モデルによる錯覚率の予測値と実測値との相関（r = .86, p <.001）は，non-CIMS モデル（統合判断の確率を常に 1 と仮定）との相関よりも有意に高く（r = -.04, p =.72），因果推定構造の妥当性が支持された。

4. 考察

本研究では，CIMS モデルを用いて，視覚および聴覚の知覚ノイズ，統合表象（感覚の重みづけ），因果推定確率，錯覚予測率を指標とし，McGurk 効果における話者の顔および声の属性（成人・子供）の影響を検討した。その結果，因果推定確率および錯覚予測率において，話者の声の属性に有意な効果が認められ，子供の声が提示された条件において，視覚と聴覚の情報が同一の信号源に由来すると判断される確率が有意に低下し，それに伴い錯覚率も低下することが明らかとなった。

本研究では，年齢による親近性の効果（e.g., Kuefner et al., 2007）を想定し，話者の顔と声の属性を操作した。親近性仮説では，親近性の高い成人話者（顔・声）の条件において視聴覚統合判断が促進されると予測しており，本研究の結果は親近性仮説を部分的に支持するものといえる。

一方，知覚ノイズについては話者属性による有意な差は認められず，効果量も小さかった。本研究の明瞭性仮説では，音声の明瞭度の違いが錯覚率に影響を与えることを想定していたが，成人および子供の話者による音声刺激における知覚精度に差は見られなかった。

この結果から，本研究では刺激の明瞭度が統合判断に寄与していた可能性は低いと考えられる。また，統合表象（感覚の重みづけ）に関しても，話者属性による有意な効果は認められなかった。全体としては視覚情報への引き寄せ傾向が見られたが，これは音声知覚における視覚の影響を反映したものであると考えられる（McGurk & MacDonald, 1976）。

さらに，顔と声の話者属性の一致性に関しては，因果推定確率や錯覚予測率において，顔と声の交互作用は有意ではなく，顔と声が一致することによる統合の促進効果（e.g., Green et al., 1991）は確認されなかった。また，顔と声の一致性判断課題においても，全ての条件で高い正答率が得られ，刺激による違いも示されなかった。したがって，顔と声の一致は観察者にとって顕在的に判断可能であったにもかかわらず，視聴覚統合判断に対して大きな影響を及ぼさなかったことが示唆される。

以上の結果より，明瞭性仮説および一致性仮説は支持されず，親近性仮説のみが部分的に支持された。視聴覚統合の判断は，単なる感覚的な一致や知覚精度によるものではなく，観察者が経験や学習を通じて形成する話者属性に関する信頼性評価といった，より高次の因果的意味づけに基づいてなされる可能性が示唆される。

文献

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. Perception & Psychophysics, 50(6), 524–536. https://doi.org/10.3758/BF03207536

Kuefner, D., Macchi Cassia, V., Picozzi, M., & Bricolo, E. (2008). Do all kids look alike? Evidence for an other-age effect in adults. Journal of Experimental Psychology: Human Perception and Performance, 34(4), 811–817. https://doi.org/10.1037/0096-1523.34.4.811

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal Inference in Multisensory Perception. PLoS ONE, 2(9), e943. https://doi.org/10.1371/journal.pone.0000943

Magnotti, J. F., & Beauchamp, M. S. (2017). A Causal Inference Model Explains Perception of the McGurk Effect and Other Incongruent Audiovisual Speech. PLOS Computational Biology, 13(2), e1005229. https://doi.org/10.1371/journal.pcbi.1005229

Magnotti, J. F., Ma, W. J., & Beauchamp, M. S. (2013). Causal inference of asynchronous audiovisual speech. Frontiers in Psychology, 4. https://doi.org/10.3389/fpsyg.2013.00798

Mcgurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. Nature, 264(5588), 746–748. https://doi.org/10.1038/264746a0

Rosenblum, L. D., & Yakel, D. A. (2001). The McGurk effect from single and mixed speaker stimuli. Acoustics Research Letters Online, 2(2), 67–72. https://doi.org/10.1121/1.1366356

Shams, L., & Kim, R. (2010). Crossmodal influences on visual perception. Physics of Life Reviews, 7(3), 269–284. https://doi.org/10.1016/j.plrev.2010.04.006

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. The Journal of the Acoustical Society of America, 26(2), 212–215. https://doi.org/10.1121/1.1907309

Ujiie, Y., & Takahashi, K. (2022). Own-race faces promote integrated audiovisual speech information. Quarterly Journal of Experimental Psychology, 75(5), 924–935. https://doi.org/10.1177/17470218211044480

Ujiie, Y., & Wakabayashi, A. (2022). Intact lip-reading but weaker McGurk effect in individuals with high autistic traits. *International Journal of Developmental Disabilities*, *68*(1), 47–55. https://doi.org/10.1080/20473869.2019.1699350

Walker, S., Bruce, V., & O’Malley, C. (1995). Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. Perception & Psychophysics, 57(8), 1124‑1133. https://doi.org/10.3758/BF03208369