様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月 ２４ 日現在

機関番号：５２３０１
研究種目：若手研究(B)
研究期間：2015～2018
課題番号：１５Ｋ２１０２４

研究課題名（和文）音声と映像との相互作用を利用した発話アニメーションの印象制御に関する研究

研究課題名（英文）Study of the impression change for speaking animation using interaction between voice and related animation

研究代表者
川本　真一（Kawamoto, Shinichi）

群馬工業高等専門学校・電子情報工学科・准教授

研究者番号：７０４１８５０７

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：本研究課題では，発話アニメーションにおける映像と音声をそれぞれ別の人物から由来する素材を用いて制作し，音声素材を入れ替えた発話アニメーションに対する主観的な年齢の印象評定を行うことで，見た目と声の相互作用が発話者に対する印象へ与える影響について調査した。その結果、静止顔画像と声の組み合わせでは声の主観年齢の影響を強く受けたことを示唆する結果が得られ、声に同期した発話アニメーションと声の組み合わせでは顔の主観年齢の影響が強くなったことを示唆する結果を得た。

研究成果の学術的意義や社会的意義
映画やアニメーションにおけるアフレコ、CGや合成音声などによって制作されたものなどのように、映像と音声が別々に作成され組み合わされることによって様々なコンテンツが制作されている。このようなコンテンツにおいて適切に印象を設計する上で、音声とその音声を発する人の顔や発話動画像の組み合わせによって印象が変わるのか、あるいは「音声の聞こえのみ」もしくは「顔の見た目のみ」の印象に支配されているのかについて把握することは、今後の視聴覚メディア設計を考える上で有益な知見と考える。

研究成果の概要（英文）：In this study, perceptual age of combination of voice and face was examined. As a result, the perceptual age of combination of voice and still facial image were seems to get slightly close to the perceptual age of voice only. On the other hand, the perceptual age of combination of voice and lip-synched face images were seems to get slightly close to the perceptual age of face only.

研究分野：音声情報処理

キーワード：リップシンク

様　式　C－１９，F－１９－１，Z－１９，CK－１９（共通）

１．研究開始当初の背景

音声と映像によるマルチモーダル(複数の伝達手段の組み合わせた)コミュニケーションは，人間が用いる最も基本的な情報伝達手段の一つであり，人間同士の対面対話のみならず，人間と機械とのインタラクションや，機械を介した人間同士のコミュニケーションなど，人間が関わるメディアにおいて重要な役割を果たしている．このような音声と映像との同時提示環境においては，音声が有する情報や映像が有する情報がそのまま伝達されるだけではなく，各モダリティの相互作用が生じる．例えば，発話アニメーションのような音声と CG キャラクタアニメーションの同期ずれは，アニメーション全体の自然性に大きく影響する．特にリップシンク（音声に同期した口形状アニメーション）は発話アニメーションの最も基本的な要素であり，リップシンクの同期ずれが自然性や了解度に影響することが知られている．しかし，リップシンクが非言語情報に与える影響について扱った研究はほとんど行われていない．アニメや映画の吹き替えなどにおいて映像に合わせて台詞音声を収録すること（アフレコ）が多用されており，このような場合においても映像と音声との相互作用が生じることが考えられる．本研究では特に，声や見た目，発話アニメーションから受ける話者の知覚年齢を中心に，リップシンクの影響について検討する．

２．研究の目的

本研究課題では，発話アニメーションにおける映像と音声をそれぞれ別の人物から由来する素材を用いて制作し，音声素材を入れ替えた発話アニメーションに対する印象評定を行うことで，見た目と声の相互作用が発話者に対する印象へ与える影響について検証する．

３．研究の方法

以下の手順で実験を積み重ねることで，検証を進める．また，関連する知見の蓄積も並行して進める．

(1) 顔単体および声単体の主観年齢の調査

顔画像単体，および音声単体について，視聴実験を行い，視聴者が感じた年齢（主観年齢）を付与する. この基礎データを基に, 顔と声を組み合わせたときの主観年齢について検討する. 顔と声の組み合わせとなるデータについては，顔単体の主観年齢データおよび，声単体の主観年齢データを参考に，顔の主観年齢と声の主観年齢の差が大きな組を抽出する．

(2) 静止画と音声の組み合わせによる主観年齢の調査

静止顔と声を組み合わせたものの視聴実験を行い，主観年齢を評定する．このとき，顔と声を組み合わせたときの主観年齢（「顔＋声」主観年齢）が，顔のみで評価した主観年齢（「顔」主観年齢）と声のみで評価した主観年齢（「声」主観年齢）のどちらに近いかを評価することで，見た目と声の相互作用の影響が確認されるかどうか，もし相互作用の影響が確認される場合はどちらのメディアが強く影響しているかを確認する．

(3) 静止画から作成したリップシンク動画像と音声の組み合わせによる主観年齢の調査

静止顔画像に対して音声に同期した唇の動き（リップシンク）を付与した顔動画像と声を組み合わせたものの視聴実験を行い，主観年齢を評定する．このとき，顔と声を組み合わせたときの主観年齢（「顔＋声」主観年齢）が，顔のみで評価した主観年齢（「顔」主観年齢）と声のみで評価した主観年齢（「声」主観年齢）のどちらに近いかを評価することで，どちらのメディアが強く影響しているかを確認する．先の静止顔画像と声の組み合わせによる主観年齢の結果と比較することにより，リップシンクの付与による影響が現れるかどうかについて検討する．

４．研究成果

当初予定していた研究項目に対する研究成果概要を(1)～(3)に，関連項目に関する研究成果概要を(4)に示す．

(1) 顔単体および声単体の主観年齢の調査

10 名の顔画像データおよび 19 名の文音声データに対して主観年齢を評定した. それぞれのデータに対し 13 名の実験協力者によって付与した結果, 顔画像は 26.0～33.2 歳, 音声は 19.6～53.5 歳の主観年齢の幅をデータであることを確認した. 顔画像に比べ，音声のばらつきが大きかったため, 以降の実験では 10 名の顔画像に対し, 5 名の音声データを組み合わせ

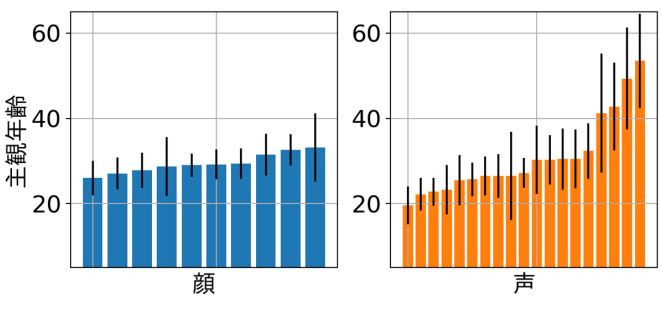

図１：顔単体・声単体の主観年齢評定値

ることで実験を進めることとした.

(2) 静止画と音声の組み合わせによる主観年齢の調査
　顔画像と音声を同時に 2 秒間提示し，その組み合わせから知覚される主観年齢を評定した．顔画像単体の主観年齢と声単体の主観年齢との間には差があり，静止顔画像と声を組み合わせた主観年齢がその間に入る場合（内挿）と，入らない場合（外挿）とが確認された．また，いずれの場合においても静止顔画像と声を組み合わせた主観年齢が，顔画像単体の主観年齢と声単体の主観年齢のどちらに近いかを比較した結果，声単体の主観年齢に近くなることが確認された．どちらか一方の単一刺激に対する主観評価年齢に完全に一致していないことから，静止画像と声を組み合わせるだけでも,「指定の顔の人物が，同時に示された声で話している」ことを想起することで，見た目と声の相互作用が影響し，結果として主観年齢にも反映されたことを示唆する結果と考える.

(3) 静止画から作成したリップシンク動画像と音声の組み合わせによる主観年齢の調査
　顔画像の静止画から作成したリップシンク動画と音声を同時に 2 秒間提示し，その組み合わせから知覚される主観年齢を評定した．リップシンク動画像と声を組み合わせた主観年齢が，顔画像単体の主観年齢と声単体の主観年齢のどちらに近いかを比較した結果，静止顔画像と声の組み合わせの時と比べて顔画像単体の主観年齢に若干近くなることが確認された．静止顔画像と声の組み合わせの時は，声単体の主観年齢に近くなったことから，リップシンクを加えることにより，顔と声を組み合わせたときの主観年齢の知覚に影響を与えたことを示している．これは,リップシンクの付加が非言語情報に影響を与えることを示唆する結果と考える.なお，この結果についてはリップシンクの自然性などにも影響する可能性が考えられるため，詳細な分析が今後必要と考えられる.

(4) 関連研究の展開
　本研究に関連する知見の蓄積のため，いくつかの研究テーマについても並行して展開を進めた.
・音声刺激の作成において，複数の話者の音声から中間的な音声を作成するための音声モーフィング技術について，時間周波数空間上で 2 話者の音声間で対応する特徴点を自動付与する技術を提案し，音声モーフィングによる音声刺激の系統的な作成を実現した.
・モーフィング音声によって 2 話者間の中間的な声を作成し，その主観年齢を付与することで主観年齢評価のためのデータ拡充を実現した．なお，特徴量的に中間的な音声であるからといって，必ずしもモーフィング音声の主観年齢が混合元の音声に対する主観年齢の算術平均にはならないことも確認した.
・2 話者の顔画像に対する中間的な顔画像（平均顔）を作成し，その主観年齢についても評価したところ，音声モーフィングと同様に，平均顔の主観年齢が必ずしも混合元の顔画像に対する主観年齢の算術平均にならないことを確認した.

５．主な発表論文等

〔雑誌論文〕（計　０　件）

〔学会発表〕（計　７　件）
① 及川隼平，川本真一，“顔と声を同時提示した際の主観年齢分析，” 情報処理学会第８１回全国大会，4V-01，査読無，2019 年 3 月.
② Shumpei Oikawa, Shinichi Kawamoto, “Analysis of Impression Change by Combination of Faces and Voices,” 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE 2018)，査読有，2018 年 10 月.
③ 滝澤 照太，川本 真一,“音声モーフィングのための基準点自動付与手法，” 平成 29 年度北陸地区学生による研究発表会，査読無，2018 年 3 月.
④ Shota Takizawa, Shinichi Kawamoto, “Automatic Reference Point Assignment Technique for Voice Morphing,” 2017 IEEE 6th Global Conference on Consumer Electronics (GCCE 2017)，査読有，2017 年 10 月.
⑤ 野村 竜暉，川本 真一，“顔と声の主観的関連性の分析，” 電子情報通信学会 2017 年総合大会，査読無，2017 年 3 月.
⑥ 園田 浩之介，川本 真一，赤木 正人，“仮説検証による特定話者音声の音素アライメント，” 平成 27 年度北陸地区学生による研究発表会，査読無，2016 年 3 月.
⑦ 尾島 康浩，川本 真一，“頭部位置を固定しない視聴覚提示実験システムの開発，” 平成 27 年度北陸地区学生による研究発表会，査読無，2016 年 3 月.

〔図書〕（計　０　件）

〔産業財産権〕

○出願状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年：
国内外の別：

○取得状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
ホームページ等

６．研究組織

(1)研究分担者
なし

(2)研究協力者
なし