科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　２９　日現在

機関番号：　１１３０１
研究種目：　基盤研究(C)
研究期間：　2011～2013
課題番号：　２３５００２５２

研究課題名（和文）高度感性情報の抽出・提示を実現する視聴覚音声コミュニケーションシステムの構築

研究課題名（英文）Development of high-definition audio-visual speech communication systems based on the knowledge of /kansei/ information processing

研究代表者
坂本　修一（SAKAMOTO, Shuichi）

東北大学・電気通信研究所・准教授

研究者番号：６０３３２５２４

交付決定額（研究期間全体）：（直接経費）　　4,000,000円、（間接経費）　　1,200,000円

研究成果の概要（和文）：　本研究は，視聴覚音声通信において，音韻・感性情報の伝送に必要となる物理パラメータの抽出，定量化を進め，音韻情報と感性情報をより高精度，高品位に伝えることのできる高感性コミュニケーションシステムの構築を目指すものである。
　これまでに，話者映像と音声の印象の一致度，調和度といった感性情報が音韻知覚に影響を及ぼすこと，さらに，その音韻情報の伝達において唇の内側の動きが極めて大きく寄与することが明らかとなった。これらの結果から，口唇付近の映像を中心に，話者の印象が正しく伝わるように映像を提示することで，話者の伝えたい情報を正しく伝える視聴覚音声コミュニケーションシステムが実現されると考えている。

研究成果の概要（英文）：It is important to transmit talker's mind including emotion for developing advanced communication system. Because not only speech signal itself but also moving image of talkers face includes rich information to understand what talker wants to tell, these information should be transmit accurately to the listeners. However, the relationship between the physical parameter extracted from audio-visual speech signal and /kansei/ information is unclear.
The final goal of this study is to develop high-definition audio-visual speech communication systems. To do this, the effect of moving image of talker's face to speech understanding was investigated. The results of psychophysical experiments indicate that around the lips area is crucial to understand what talker says. Moreover, /kansei/ impression from audio-visual speech signal is also important for speech understanding.

研究分野：　総合領域

科研費の分科・細目：　情報学・感性情報学・ソフトコンピューティング

キーワード：　感性情報学　情報システム　音声認知　視聴覚コミュニケーション

１．研究開始当初の背景

近年のマルチメディア技術の進歩により，高品位な映像・音声情報がネットワークをとおして容易に通信可能となってきている。これに伴い，単に映像，音声情報の伝送だけでなく，話者の細かい表情の変化や，感情の表出による話声の変化といった，感性情報の一端をも視聴者へ伝送する可能性が広がりつつある。これまでの研究により，単に口形情報だけでなく，話者映像と音声からうける感性情報も，単に視聴者の感情に作用するだけでなく，音韻知覚に影響を及ぼす（坂本ら，2003）ことが示されている。したがって，本来あるべき感性情報を正しく伝送することが，音声情報通信という観点からも重要である。

しかし現状では，どのように視聴覚情報を処理すれば感性情報が正しく伝送出来るかといった知見は非常に少ない。更に，話者の映像，声質，話速，視聴覚情報の同期といった個々のパラメータ，もしくは，これら複数のパラメータの相互作用が，視聴者が知覚する音韻情報や，視聴者が受ける感性情報にどのように作用し，また，その両情報がどのように結びつくのかということは，ほとんど解明されていないといえる。

２．研究の目的

本研究の最終的な目的は，視聴覚音声通信において特に感性情報の伝送に重要な感性パラメータを抽出し，抽出された感性パラメータを精度高く伝送することで，話者の感性情報をも視聴者へ再現・提示可能な高感性コミュニケーションシステムを構築することにある。

そのために，話者映像の有無や，話速といった話声の性質，話者映像と話声の同期，話者の口の動きといったパラメータのうち，音韻知覚及び視聴者が知覚する感性情報に及ぼす影響を定量評価する。

次に，得られた知見に基づいて，実際の話者映像や音声情報から，音韻知覚及び視聴者の受ける感性情報に影響を与える要因のみを精度高く抽出し，明瞭，かつ，感性高く音声の提示が可能となるように，主観評価実験をとおして，抽出したパラメータを最適化する。

以上の結果から，実際に高感性コミュニケーションシステムの構築を念頭に，感性情報を効率的に伝送する方法を提案する。

３．研究の方法

3 年間の研究期間をとおして，本研究では，音声コミュニケーションに影響の大きい感性パラメータを，主観的，客観的な測定により明らかにするとともに，それら感性パラメータの相互の影響を多面的に解析する。

その結果に基づいて，感性パラメータの高精度抽出手法，および，それらの効率伝送手法を検討し，話者の感性情報をも視聴者へ再現・提示可能な高感性コミュニケーションシステムを構築する。

初年度の平成 23 年度には，研究期間全体をとおして使用する刺激素材を収録し，その刺激素材自身の持つ様々な物理パラメータの抽出を目指す。ここで得られる物理パラメータは，その刺激自身の持つ感性情報との関連を分析するために必要となるものである。発話訓練経験のある話者に協力を依頼し，刺激を収録する。

平成 24 年度には，前年度収録した視聴覚音声刺激群を実際に視聴者に提示して感性評価を行い，刺激の持つ物理パラメータと，得られた感性情報との関連を分析する。

最終年度である平成 25 年度は，前年度までの分析に基づき，感性パラメータとして有効な物理パラメータを抽出し，感性情報も有効に提示でき，かつ，音韻の聴き取りも良好となりうる条件における，感性パラメータの取りうる範囲を検証する。以上の結果から，高感性コミュニケーションシステムの構築を念頭に，感性情報を効率的に伝送する方法を提案する。

４．研究成果

(1) 話者映像の感性情報と音韻情報の関連

我々は会話において，話者から様々な感性情報を取得しており，その取得過程には，話者の表情などの視覚情報や，音声の韻律などの聴覚情報が重要な役割を果たしている。その一方で，話者からの視聴覚情報が音韻知覚自体に影響を与えることは明らかである。しかし，感性情報と音韻知覚との相互関係については未解明の部分が多い。

そこで，唇音の音声と非唇音の話者映像の同時提示により，これらとは別の音韻が聞こえる McGurk 効果を指標として，複数の話者の映像と音声の合成刺激を用いて McGurk 効果を測定し，映像と音声の調和度，明瞭性といった感性情報が音韻知覚に与える影響を検討した。

実験時に使用した刺激は，映像/ke/，音声/pe/の視聴覚不一致刺激 1 種と，映像音声ともに/ke/,/pe/の視聴覚一致刺激 2 種である．McGurk 効果が見られた場合，被験者は視聴覚不一致刺激提示時に/te/と知覚することとなる。女性話者 5 名に/pe/,/ke/と発音させ，その時の発話映像，音声を刺激用素材として上記 3 種の刺激を作成した。18～22 才の大学生 9 名（いずれも健聴者で矯正視力を含め両眼で 1.0 以上の視力を持つ）に，作成した刺激を提示し，①画面に映った人が何と話したか，②提示映像からの印象と，提示音声からの印象の調和度，③提示された音声の明瞭性の 3 種を回答するように教示した。なお,②,③は 5 段階で評定尺度させた。

図 1 に実験結果の一例を示す。この図は，視聴覚不一致刺激での聴取結果から得られた，McGurk 効果の生起確率と映像と音声の調和度の関係を示したものである。図 1 を見る

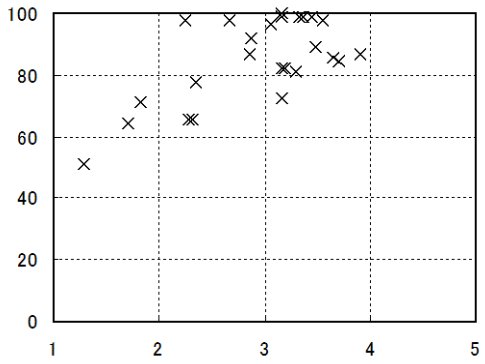


図 1 映像と音声の調和度と McGurk 効果の生起確率

と，映像と音声の調和度が高くなるにつれて，McGurk 効果の生起確率が上昇していることが分かる．これは，話者映像と音声のそれぞれから受ける印象が調和している場合，調和していない場合に比べ，被験者は映像と音声をより融合したものとして知覚したためと考えられる。同様の傾向は，McGurk 効果の生起確率と映像と音声の明瞭性との関連においても見られた。

これらの結果から，感性高く視聴覚音声情報を提示するためには，話者映像と音声の印象の調和や明瞭性が重要であることが示唆された。ここでは映像と音声から得られる全体的な印象をターゲットにして実験を行っているが，印象を決める要因には，映像と音声のずれや，大きさといった物理パラメータが含まれる。したがって，これらの物理パラメータを操作することで，本研究が目指す高感性コミュニケーションシステムの実現が可能となると考えている。

(2)音素の聴取における話者映像の寄与部位

音声コミュニケーションにおいて，話者がなんといっているのかを聴取者に正確に伝えることは，実現すべき最重要課題である。騒音環境下や聴覚障害者など，音声情報の取得が困難な場合には話者映像は極めて大きな役割を果たす。しかし，話者映像のどの部位が音声聴取に寄与するのかを要素還元的に明らかにした例はなく，音声コミュニケーションシステム実現において，話者映像のどの部位をどの程度精度高く伝送するのかについての知見はほとんどない。そこで，視覚情報が音声同定課題に与える影響について，顔のどの部分が音声知覚に影響を与えているかを，口唇周辺の映像に着目して実験的に明らかにした。

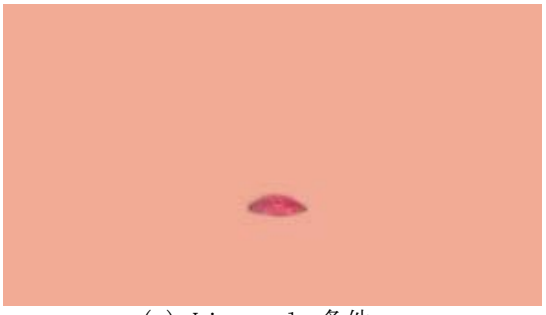
実験に先立って，どの音素が視覚的影響を強く受けるかについて網羅的に調べるため，全ての母音と全ての子音の組をバランスよく配置した無意味 3 連音節リストを構築した。このリストは，日本語 67 音節からなる無意味 3 連音節 155 個からなり，全ての母音と子音の組み合わせが含まれるように決定されている。この音声を発話経験のある女性話者に発話させ，音声と映像を収録して元の素材とした。その後，画像処理技術を用いて話者映像を部位ごとに加工し，音声と組み合わせることで，実験刺激を作成した。

元の素材と，音声のみの条件に加え，図 2 に示す Lips only 条件と Masked Lips 条件の実験刺激を用いて無意味 3 連音節了解度試験を実施した。視聴者は 16 名の大学生および大学院生であり，いずれも日本語母語話者であった。

実験結果を図 3 に示す。図 3 を見るとわかるように，音声のみの条件（Audio only）に比べ，他の映像付加条件全てにおいて，了解度が上昇している。これは，話者の口唇付近を隠した Masked Lips 条件についても同様である。一方で，元の話者映像を用いた条件（Original）と Lips only 条件とは同程度の了解度が得られており，Masked Lips 条件はそれよりも若干了解度が低くなっている。

これらの結果を考えると，少なくとも口唇付近の話者映像情報さえ正しく提示することができれば，話者の伝えたい音韻情報を正しく伝えることができるといえる。さらに，この実験に引き続いて行った実験結果から，口唇付近でも特に口唇内部の情報さえ提示されていれば，元々の話者映像と遜色なく音韻情報の伝達が可能であるといった結果が得られている。

以上得られた実験結果を踏まえることで，高感性視聴覚コミュニケーションシステムの実現が見えてくる。すなわち，と，口唇付近の映像を中心に，話者の印象が正しく伝わるように映像を提示することで，話者の伝え



(a) Lips only 条件

(b) Masked Lips 条件

図 2 実験で用いた話者映像刺激の例

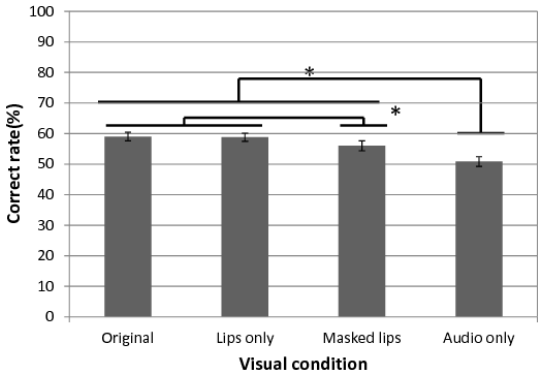


図 3 無意味 3 連音節了解度試験結果

たい情報を正しく伝える視聴覚音声コミュニケーションシステムが実現されると考えている。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 5 件）

[1] S. Sakamoto, G. Hasegawa, T. Abe, T. Ohtani, Y. Suzuki and K. Kawase, “The contribution of the detailed parts around talker’s mouth for speech intelligibility,” Proc. The 21st International Congress on Sound and Vibration (ICSV21), 7-page manuscript, 2014（掲載確定，査読有）

[2] 長谷川玄, 坂本修一, 阿部亨, 大谷智子, 鈴木陽一, 川瀬哲明 “口唇以外の話者映像情報が無意味 3 連音節を用いた音声明瞭度に与える影響,” 日本音響学会講演論文集, 2-P5-21, 641-642, 2014（査読無）

[3] 長谷川玄, 坂本修一, 阿部亨, 大谷智子, 鈴木陽一, 川瀬哲明, “無意味 3 連音節を用いた音素別明瞭度における話者映像の寄与の分析,” 電子情報通信学会技術研究報告, HIP2013-60, 1-6, 2013（査読無）

[4] 長谷川玄, 坂本修一, 阿部亨, 大谷智子, 鈴木陽一, 川瀬哲明, “無意味 3 連音節を用いた音素別明瞭度における視覚情報の寄与の分析,” 日本音響学会聴覚研究資料, H-2013-102, 595-600, 2013（査読無）

[5] S. Sakamoto, H. Mishima and Y. Suzuki, “Effect of consonance between features and voice impression on the McGurk effect,” Interdisciplinary Information Sciences Journal, 18, 83-85, 2012（査読有）

〔学会発表〕（計 5 件）

[1] S. Sakamoto, G. Hasegawa, T. Abe, T. Ohtani, Y. Suzuki and K. Kawase, “The contribution of the detailed parts around talker’s mouth for speech intelligibility,” Proc. The 21st International Congress on Sound and Vibration (ICSV21), 2014.7.13-17, Beijing, China（発表確定，招待講演）

[2] S. Sakamoto, G. Hasegawa, T. Abe, T. Ohtani, Y. Suzuki and K. Kawase, “Contribution of detailed parts around talker’s mouth for audio-visual speech perception,” 167th Meeting of the Acoustical Society of America, 2014.5.5-9, Providence, USA

[3] 長谷川玄, 坂本修一, 阿部亨, 大谷智子, 鈴木陽一, 川瀬哲明, “口唇以外の話者映像情報が無意味 3 連音節を用いた音声明瞭度に与える影響,” 日本音響学会 2014 年春季研究発表会, 2014 年 3 月 10～12 日, 日本大学

[4] 長谷川玄, 坂本修一, 阿部亨, 大谷智子, 鈴木陽一, 川瀬哲明, “無意味 3 連音節を用いた音素別明瞭度における話者映像の寄与の分析,” 電子情報通信学会ヒューマン情報処理(HIP)研究会, 2013 年 11 月 19～20 日, 東北大学

[5] 長谷川玄, 坂本修一, 阿部亨, 大谷智子, 鈴木陽一, 川瀬哲明, “無意味 3 連音節を用いた音素別明瞭度における視覚情報の寄与の分析,” 日本音響学会聴覚研究会, 2013 年 10 月 10～11 日, 神戸セミナーハウス

６．研究組織
(1)研究代表者
坂本　修一（SAKAMOTO Shuichi）
東北大学・電気通信研究所・准教授
研究者番号：60332524