様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 １０ 日現在

機関番号：１１３０１
研究種目：挑戦的萌芽研究
研究期間：2012～2014
課題番号：２４６５０１００

研究課題名（和文）視聴覚話者情報がもつ音韻・感性情報の分析とクロスモーダル推定・合成手法の模索

研究課題名（英文）Analysis and synthesis method of phonetic/emotional information in audio-visual speech information

研究代表者
鈴木　陽一（SUZUKI, Yo-iti）

東北大学・電気通信研究所・教授

研究者番号：２０１４３０３４

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：本研究の目的は，高品位ではない環境下でも高次感性情報の通信が可能な視聴覚音声コミュニケーションシステムの実現にある。ここでは，コミュニケーションにおいて最も重要な要素が，話者の発する音声の意味を聴衆が理解することであると考え，音韻情報の取得において話者映像のどの部分が寄与するのかについて分析を進めた。実験の結果，話者の音韻情報とその話者の静止画像がある環境下では，特に唇音に着目して，話者の口唇部分を何らかの画像処理により編集，操作して，音声と同期して組み合わせることで，本研究が目指す視聴覚音声に関するクロスモーダル推定・合成法が実現できることが示唆された。

研究成果の概要（英文）：Moving images of a talker's face carry much information for speech understanding. Interpretation of that information is known as lip-reading. For the development of advanced multi-modal communications systems, such information should be well considered. To aim at developing such systems, we have focused on the relationship between speech sound information and moving image of talker's face. In this study, we have been particularly examining which parts of moving image of talker's face contribute most to speech understanding. We performed audio-visual speech intelligibility tests and investigated the relationship between speech intelligibility and effects of the parts of moving image of talker's face. Results of the experiments indicated that the mouth area alone provides sufficient information for speech intelligibility. The results suggested that the cue of lip-reading around the mouth might be able to generate from speech sound information.

研究分野：情報学

キーワード：視聴覚音声知覚　マルチモーダルインタフェース　感性情報処理

１．研究開始当初の背景

音声コミュニケーションシステムの構築において，話者映像を含めた視聴覚音声情報の高感性・高品位提示技術は，極めて重要となっている。これまでの研究では，単に音声伝達に寄与する視聴覚情報の分析，および，その情報の伝達手法に着目された研究がなされてきた。近年は，国内では NTT，早稲田大，海外では Tilburg Univ.（オランダ）などの研究機関において，感情や情動に着目し，映像と音声の両者の含む感性情報と視聴者が知覚する印象との関連に関する研究が盛んに行われてきたものの，各感覚情報の持つ感性情報の組み合わせの効果を知覚実験により明らかにするという域を出ていない。実際に視聴覚情報の持つどのような要素がキーとなって視聴者の感性情報を引き起こすのか，もしくは，個々の要素がどのように組み合わされることにより，全体の印象知覚や音韻知覚が形成されるのかという点については，個々の要素の抽出・特定が困難なことも相まって，全く手つかずの状況である。しかし，システム構築や工学応用という点を考えると，このような要素還元的なアプローチでの研究を行うことは非常に重要である。

申請グループは，これまで視聴覚音声知覚に関する，知覚心理的・工学的研究を精力的に進め，音声知覚に寄与する視覚情報の特定を行ってきた。その実験の中で，口の動きの再現程度により口の動きの本物らしさの印象が変化することが明らかとなった。このような印象変化を定量的に明らかにすることは，視聴覚音声の持つ感性情報を操作する手がかりとなり得るものと思われる。

２．研究の目的

本研究の目的は，高品位ではない環境下でも高次感性情報の通信が可能な視聴覚音声コミュニケーションシステムの実現にある。まず視聴覚音声情報について，視覚と聴覚情報それぞれの特徴量が，感性情報を含めた音声の知覚に与える効果を要素還元的に定量評価する。更に，視覚と聴覚情報の操作により元情報の持つ感性情報をより忠実に，更には，より強調して伝達することが可能な視聴覚音声コンテンツ創成技術を創出する。最終的には，視聴覚いずれか一方のみしかないコンテンツから存在しない方の感覚情報を最適設計・作成しうる技術の創出を目指す。

３．研究の方法

音声コミュニケーションに寄与の大きい感性パラメータを，生理指標を含め，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。特に，音韻理解・感性知覚に着目し，音韻については読唇効果，感性知覚については情動情報を中心として，視覚，聴覚情報のそれぞれが持つ特徴量同士の相互関係に着目するとともに，元々持っている音韻情報や感性情報を忠実に，もしくは，強調して表現するためのパラメータの操作手法を検討する。

その結果に基づいて，視聴覚音声情報，もしくは，視覚，聴覚それぞれの単一感覚情報に対する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，視聴覚高感性音声コンテンツの作成を試みる。

平成 24 年度は，本研究期間を通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を抽出することを目的とする。収録音声の品質を一定にし，かつ，様々な感性情報に応じて発話を変えることができ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 25 年度以降は，平成 24 年度に収録した刺激素材を用いて実際に心理実験を行い，映像，音声の各刺激の持つ物理パラメータがどのように作用して刺激全体の音韻情報，並びに，感性情報を形作っているかを分析する。

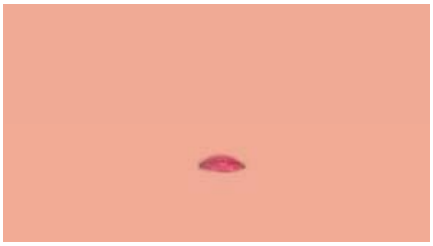
これらの心理実験結果から，感性パラメータとして有効な物理パラメータを抽出し，話者映像，並びに，音声刺激からの感性パラメータの高精度抽出，高効率提示アルゴリズムを構築する。

４．研究成果

コミュニケーションにおいて最も重要な要素は，話者の発する音声の意味を聴衆が理解することである。そこで本研究を行うにあたり，まず，音韻情報の取得において話者映像のどの部分が寄与するのかについて分析を進めた。この知見が得られれば，話者の顔画像と，発声された音声を用いて，その音韻の聞き取りに寄与の大きい顔画像の部分を，その音韻に適した形で加工することで，本研究の目指す視聴覚話者情報のクロスモーダル推定・合成法が実現できると考えられる。

実験に使用した素材を図 1 に示す。元動画像（a）に対し，音韻情報の取得に重要と思われる口周辺の映像に着目し，口唇のみを抽出した動画像（b）を作成し，それと合わせ，唇の厚さに関係なく人間は話者の音韻情報を取得することができることに着目し，口唇の内側のみを抽出した動画像（c）を作成した。（b）については，動画像に含まれる色情報から輝度値を算出し，口唇とそれ以外の閾値となる値を算出して領域を抽出した後に，境界が滑らかとなるように時間・空間において境界の平滑化を行った。（c）については，（b）で閾値から算出した境界線を基準に，少なくとも唇の内側の輪郭線が全て表示されるように内側に一定画素分境界面をシフトさせ，（b）と同様に境界面が滑らかになるような処理を行った。発声された音声は，日本語の 100 音節を組み合わせた無意味三連音節で，音韻列を工夫することで全ての母音と子音の組が現れるようにした。これは，特に母音から子音への渡りの部分に口唇情報の寄与が大きいと考えたからである。

(a) 元動画像 (Original)



(b) 口唇のみ抽出した画像 (Lips only)

(c) 口唇内側のみ抽出した画像 (Inner lips)

図 1：実験に使用した画像

聴取者 8 名の了解度試験結果を図 2 に示す。図から明らかのように，話者映像なしの条件（Audio only）に比べ，他の 3 条件で了解度が有意に高くなっている。また，話者映像なしの条件よりも了解度が高かった 3 条件については，統計的な有意差は認められなかった。この結果は，話者映像の口唇，それも，唇の内側の情報さえ音声と合わせて正しく提示できれば，話者の発する音声の聞き取りが充分に行えることを示唆している。

さらに詳しく調べるため，音韻ごとに今回提示した話者映像の寄与が異なるかを個々の音素に分けて話者映像の寄与を分析した。代表的な例として，Inner lips 条件に関して，他の条件に比べて聞き取りに有意差のあった音韻をまとめた結果を表 1 に示す。分析の結果，話者映像なし条件に比べ，/n/，/h/，/m/，/w/，/z/，/d/，/b/，/p/の各音素で映像を付加することにより了解度が向上するという結果が得られ，この傾向は今回使用した映像刺激のいずれにおいてもおおむね観測された。このことから，先に示した口唇の内側のみ音声と合わせて提示すれば，話者の発する音声の聞き取りが充分に行えることが，音韻レベルで示された。今回の示された音韻はいずれも唇に特徴的な動きがある唇音であり，口唇の内側さえ示されれば，唇の開閉の情報を聴取者は得ることができるため，これらの音韻で差が得られたものと推察される。

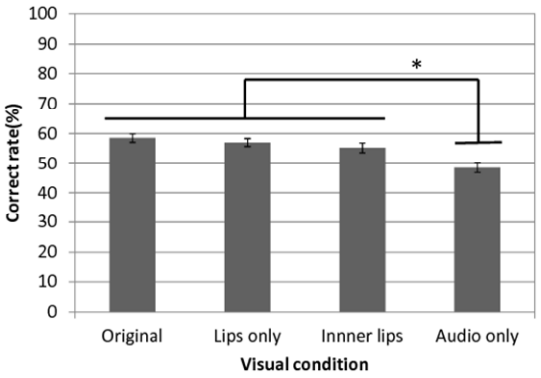


図 2：各映像条件における了解度試験結果

表 1：Inner lips 条件での個々の音韻の聞き取りの様相

+:Inner lips条件が他の条件より有意に正答率が高い箇所
−:Inner lips条件が他の条件より有意に正答率が低い箇所

		a	i	u	e	o	Ave			a	i	u	e	o	Ave
n	Original	-						z	Original	-					-
	Lips only								Lips only	-					-
	Audio						+		Audio	-					-
h	Original							d	Original						
	Lips only								Lips only						
	Audio	+					+		Audio			+			+
m	Original							b	Original						
	Lips only								Lips only						
	Audio	+	+				+		Audio	+					+
w	Original					-		p	Original						
	Lips only				+				Lips only						
	Audio			+		+	+		Audio	+	+			+	+

以上，これまでの研究結果から，話者の音韻情報とその話者の静止画像がある環境下では，特に唇音に着目して，話者の口唇部分を何らかの画像処理により編集，操作して，音声と同期して組み合わせることで，本研究が目指す視聴覚音声に関するクロスモーダル推定・合成法が実現できることが示唆された。この技術は，例えば高齢者や難聴者など，音声の聞き取りが困難な聴衆に対して，映像情報を付加することにより聞き取りの向上を図るというような福祉応用の観点でも発展が期待できる。

今後は，実際にどのような動きが音韻の聞き取りに貢献するのかと言ったさらなる詳細な分析が必要であるほか，今回充分な検討を行うことができなかった感性情報の伝達といった観点でも研究が必要であると考えている。

５．主な発表論文等

〔雑誌論文〕（計 3 件）

[1] Shuichi Sakamoto, Gen Hasegawa, Toru Abe, Tomoko Ohtani, Yôiti Suzuki and Tetsuaki 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) (査読無)

[2] 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)

[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視覚情報の寄与の分析,” 日本音響学会聴覚研究会資料, H-2013-102, 595-600 (2013) (査読無)

〔学会発表〕(計 5 件)

[1] Shuichi Sakamoto, Gen Hasegawa, Toru Abe, Tomoko Ohtani, Yôiti Suzuki and Tetsuaki Kawase, “The contribution of the detailed parts around talker's mouth for speech intelligibility (invited lecture),” The 21st International Congress on Sound and Vibration (ICSV21), 2014 年 7 月 13～17 日, Beijing, China

[2] Shuichi Sakamoto, Gen Hasegawa, Toru Abe, Tomoko Ohtani, Yôiti Suzuki and Tetsuaki Kawase, “Contribution of detailed parts around talker's mouth for audio-visual speech perception,” 167th Meeting of the Acoustical Society of America, 2014 年 5 月 5～9 日, 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 (東京都千代田区)

[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 (宮城県仙台市)

[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 (兵庫県神戸市)

６．研究組織

(1)研究代表者

鈴木　陽一 (SUZUKI, Yôiti)
東北大学・電気通信研究所・教授
研究者番号：20143034

(2)研究分担者

川瀬　哲明 (KAWASE, Tetsuaki)
東北大学・大学院医工学研究科・教授
研究者番号：50169728

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