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　４ 年　 ６ 月 ２８ 日現在

機関番号：３４４０６
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ１２９０５

研究課題名（和文）モバイル端末を用いたマルチモダル発声支援システムの研究

研究課題名（英文）Multi-Modal Speech Enhancement Using Mobile Device

研究代表者
松井　謙二（MATSUI, Kenji）

大阪工業大学・ロボティクス＆デザイン工学部・教授

研究者番号：３０６１３６８２

交付決定額（研究期間全体）：（直接経費）　　2,500,000円

研究成果の概要（和文）：本研究は,喉頭摘出者のように発声が困難なユーザーに対して、“既存のデバイスが使える”,“目立たない外観である”,“使いやすいインターフェースである”という発声支援装置開発を目標としている.具体的にはスマートフォンのような携帯型のデバイスを用いた口唇認識と音声合成による発声支援装置の開発を行っている.先ずPC上で36種類の口形素を単位としてVAEとCNNを用いた画像認識による読唇を行った。次に携帯端末に実装し使用感評価を行った。また認識性能向上のため深度画像を用いて主に子音の認識性能向上を目指した。現在までに20単語、特定話者で90%程度の性能を確認している。

研究成果の学術的意義や社会的意義
喉頭摘出者など病気や事故で発声が困難になった場合、電気式人工喉頭や食道発声等の代用音声を用いる．しかしこれらは使用時に目立つことや習得に時間がかかることが課題である．実際にユーザからは“既存のデバイスが使える”，“目立たない外観である”，“使いやすいインターフェースである”ことが望まれている．このことから機械読唇による発声支援が研究されている．本研究の特徴は口形素と変分オートエンコーダを用いて単語登録が極めて容易な機械読唇によるフレーズ認識方式であり、携帯端末への実装も行いその効果や課題を検証した．また、深度画像を用いて機械読唇での子音認識の精度向上を図っており、実証実験に向けて意義は大きい．

研究成果の概要（英文）：We have been developing a speech enhancement device for laryngectomees. Our approach is to use a lip-reading technology to be able to recognize Japanese words from lip images and generate speech outputs using mobile devices. The target words are translated into registered 36 viseme sequences, and converted into VAE (Variational Auto Encoder) feature parameters. Then the corresponding words are recognized using CNN-based model. PC-based prototype was tested, and observed more than 90% accuracy with 20 Japanese words and a well-trained single subject. Also, we developed a mobile device based prototype and conducted the preliminary recognition experiment with 26 words by a well-trained single subject, and 95% accuracy was obtained including the 1st through 6th candidates, which was almost equivalent to the PC-based system. To be able to improve consonant recognition, depth camera was introduced and obtained slightly better accuracy, however, more careful algorithm tuning is necessary.

研究分野：音声信号処理

キーワード：機械読唇　発声支援　変分オートエンコーダー　口形素　深度画像　携帯端末

１．研究開始当初の背景

喉頭摘出など多様な理由による発声障碍に対して，　過去５０年間，人工喉頭や拡声器などの発声支援器具が用いられているが，その技術進化は極めて乏しい．しかし，現代は長寿化，就労の高齢化など，健常者と同等に元気に働くユーザーが増加し，これらのユーザーに対して発声支援は一層重要である．一方，昨今の AI，センシング技術により革新的機器開発が可能になりつつある．

２．研究の目的

本研究による発声支援装置開発の目的は,（１）周囲の視線が気にならずストレスなく使える，（２）カメラ，スピーカー，ディスプレイなどの連携によるマルチモダル機能を活用，（３）広く普及している携帯端末を活用，などの見た目に自然，低コスト，軽量，かつ高度な信号処理技術を用いる革新的発声支援技術の実現を目指す．

３．研究の方法

真に役に立つ発声支援装置を開発するためには，ユーザー視点の開発プロセスは極めて重要である．本研究では，デザイン思考的アプローチ，およびイノベーション手法の一つである Foresight & Innovation を参考にしつつ，ユーザーである銀鈴会の方々のご意見を伺いながら研究開発を行った．

４．研究成果

（１）口唇画像認識方式の概要

発声支援装置の全体的なシステムのイメージの概略図を図１に示す．このシステムの特徴として，①既存のデバイス（スマートフォン）が利用可能，②システムを使っている姿が不自然ではない，③ スマートフォンの機能を利用した使いやすいインタフェースである，の 3 点が挙げられる．

（２）口唇領域画像の抽出

発話動画を撮影し，動画を 30 fps で画像に変換する．これらの顔画像を図２の流れで口唇領域画像の抽出を行った．まず人の画像から，HOG 特徴量を抽出して SVM を用いて顔を検出する．次に勾配ブースティング決定木（GBDT）を用いて顔から，各部位を検出し，そこから口唇領域画像を抽出する．最後に，抽出された口唇領域画像にヒストグラム正規化を行い，口唇のアスペクト比を維持したまま 64 × 64 pixel にリサイズする．この画像を入力画像とした．図３に無発声と各母音の前処理後の画像を示す．

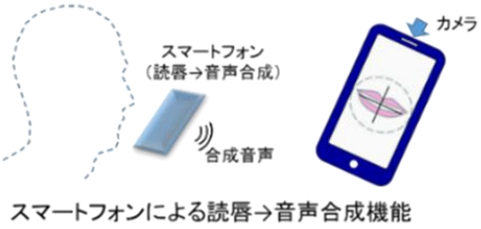


図１　システムのイメージ

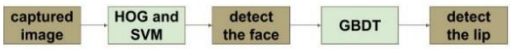


図２　口唇領域画像の抽出

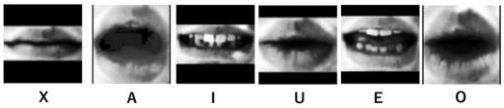


図３　前処理を行った口唇領域画像

（３）モデルの学習に使用するデータ

使用するデータは，表１にある各母音(A，I，U，E，O)と無発声(X)の口形列 36 種類を発声した動画を撮影し，前処理を行ったものである．本研究では使用者に最適化するため，使用者が 36 種類の口形列を撮影した発声動画を学習用データとした．二音節の場合は，記載されている１つ目の母音を発声し，録画を開始する．そして，2 つ目の母音を発声し続けながら録画を終了する．

表１　学習用口系列一覧

X	A	I	U	E	O
XA	AX	IX	UX	EX	OX
XI	AI	IA	UA	EA	OA
XU	AU	IU	UI	EI	OI
XE	AE	IE	UE	EU	OU
XO	AO	IO	UO	EO	OE

（４）変分オートエンコーダ[１]

口唇領域画像の特徴量抽出を行うためのモデルとして変分オートエンコーダ(VAE)を用いた． VAE は，少ない次元数で特徴量を抽出でき，入力データに対して潜在変数 Z を生成するパラメータに変換する．一般的に確率分布として正規分布が用いられるため，パラメータは平均と分散になる．図 4 に VAE のイメージ図を示す．VAE のエンコード（図５）からサンプリングされる特徴量は，原点を中心に連続的に変化するベクトルとなる．よって，入力データが連続的に変化するタスクや，データが少量な場合において，データを生成するタスクなどに有効である．

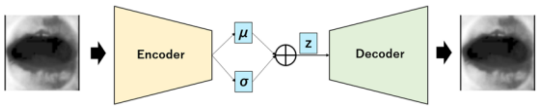


図４　VAE のイメージ図

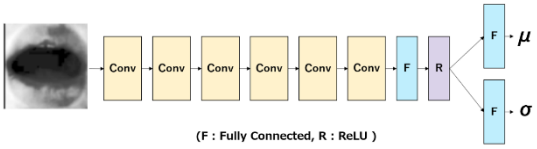
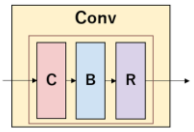


図５　VAE のエンコーダのモデル

VAE のデコーダは図６のような畳み込みを多段にしたモデルであり，潜在変数 Z を入力とし入力画像を復元する．エンコーダの図５との違いは，逆畳み込みを適用しており，サイズをアップサンプリングする層となっている点である．図７にデコーダのモデルを示す．



(R : ReLU , C : Convolution(transpose), B : Batch Normalization)

図６　デコーダモデルに使用する畳込層

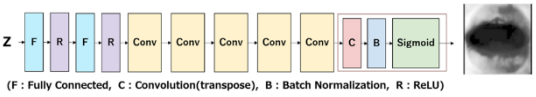


図７　VAE のエンコーダのモデル

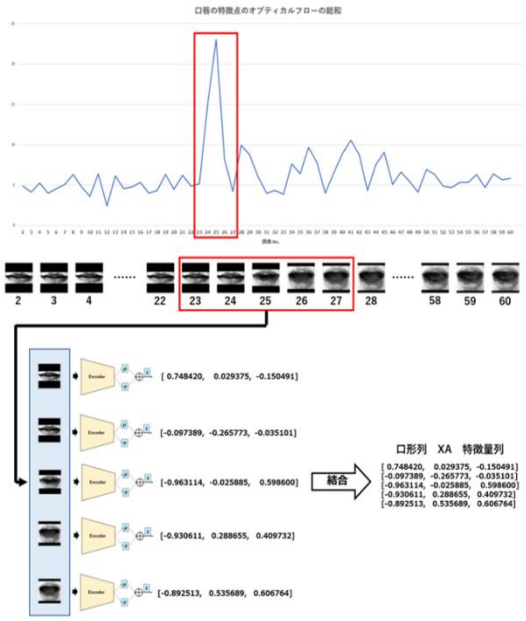


図８　２音節の場合の特徴量データの作成

（５）単語認識のための系列データの生成

まず，口形列が単音節（短母音）の場合，発声動画の画像化を行い，各画像を VAE のエンコーダに入力し，得られた特徴量の平均値を対応する特徴量とした．口形列が２音節の場合，（２）の口唇の特徴点を用いてオプティカルフローの総和を算出し，変化量が最も大きかった時点から前後 2 枚ずつ合計 5 枚を変化途中の画像列とした．この 5 枚の各画像を VAE のエンコーダに入力し，口形列に対応した特徴量の系列データとした．上記に記述した処理を表 1 の口形列すべてに行い，口形列に対応する特徴量の系列データを作成した．図８に２音節の場合の特徴量データの作成方法のイメージ図を示す．

（６）学習用データ作成

発話したい単語を発話リストに登録し，登録された単語を口形列に変換した．単語を口形列に返還するにあたり，ひらがなに対応する口形列を，宮崎らが提案した口形順序コード[31]を基に作成した．表２に基本となる 44 音の口形列を示す．

本研究での口形列は，「ありがとう」の場合に「X, XA, A, AI, I, IA, A, AU, UO, O, OX, X」となる．

（７）単語認識モデル

作成した学習用データを同じ時系列の長さとなるようにパディングを行い，モデルの学習を行った．入力データとしてパディングを行ったデータ(本研究では，長さ 238 要素が 3 のデータ)を使用し，発話リストに登録した単語のうち確率が高いものを選択する．通常，時系列データを対象とする場合は再帰型ニューラルネットワーク(RNN)を用いるが，本研究では単語認識モデルに畳み込みニューラルネットワーク(CNN)を用いた．さらに，CNN を用いることで認識に必要な時間を短縮することができる．図９に単語認識モデルを示す．

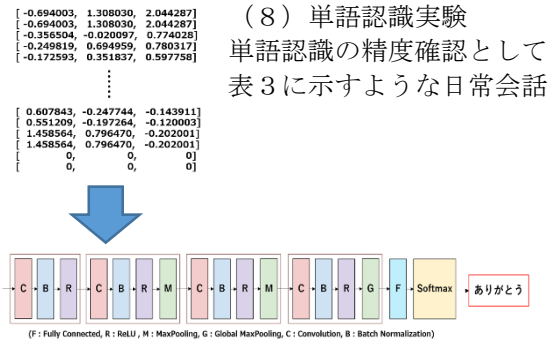


図９　CNN を用いた単語認識モデル

（８）単語認識実験

単語認識の精度確認として 2 種類の発話リストでの単語認識実験を行った．表３に示すような日常会話での使用頻度の高い単語を 15 種類選択した．

表２　学習用データ　口系列一覧

あ	か	さ	た	な	は	ま	や	ら	わ
A	A	IA	IA	IA	A	XA	IA	IA	UA
い	き	し	ち	に	ひ	み		り	
I	I	I	I	I	I	XI		I	
う	く	す	つ	ぬ	ふ	む	ゆ	る	
U	U	U	U	U	U	XU	U	U	
え	け	せ	て	ね	へ	め		れ	
E	E	IE	IE	IE	E	XE		IE	
お	こ	そ	と	の	ほ	も	よ	ろ	
O	O	UO	UO	UO	O	XO	UO	UO	

表３　登録単語

ありがとう	いいえ	おはよう	おめでとう	おやすみ
ごめんなさい	こんにちは	こんばんは	さようなら	すみません
どういたしまして	はい	はじめまして	またね	もしもし

表４　15 単語　認識結果

	正解文	認識結果	
		第一候補	第二候補
1	ありがとう	さようなら	ありがとう
2	いいえ	いいえ	はじめまして
3	おはよう	おはよう	こんばんは
4	おめでとう	おめでとう	おはよう
5	おやすみ	またね	こんばんは
6	ごめんなさい	ごめんなさい	すみません
7	こんにちは	こんにちは	もしもし
8	こんばんは	こんばんは	こんにちは
9	さようなら	さようなら	おやすみ
10	すみません	すみません	いいえ
11	どういたしまして	おめでとう	またね
12	はい	はい	はじめまして
13	はじめまして	はじめまして	すみません
14	またね	またね	すみません
15	もしもし	こんばんは	はい

表５　登録 20 単語

ありがとう	いいえ	おはよう	おめでとう	ごめんなさい
すみません	どういたしまして	はい	もしもし	あか
あお	きいろ	みどり	しろ	ぜろ
いち	に	さん	よん	ご

表６　20 単語　認識結果

	正解文	認識結果		
		第一候補	第二候補	第三候補
1	ありがとう	ぜろ	ありがとう	おめでとう
2	いいえ	いいえ	すみません	ぜろ
3	おはよう	おはよう	あお	よん
4	おめでとう	おめでとう	ぜろ	ありがとう
5	ごめんなさい	ごめんなさい	おめでとう	すみません
6	すみません	すみません	いいえ	さん
7	どういたしまして	おめでとう	ごめんなさい	どういたしまして
8	はい	はい	に	さん
9	もしもし	さん	よん	に
10	あか	さん	すみません	あか
11	あお	あか	さん	あお
12	きいろ	いち	に	きいろ
13	みどり	に	いち	みどり
14	しろ	しろ	に	さん
15	ぜろ	ぜろ	さん	いいえ
16	いち	いち	に	みどり
17	に	に	いち	みどり
18	さん	さん	ぜろ	いいえ
19	よん	ご	よん	あお
20	ご	ご	よん	あか

1 単語につき 50 個のデータとなるようにデータの水増しを行い, 750 個のデータセットで単語認識モデルの学習を行った．発話動画を画像に変換し，口唇領域画像を作成した．各口唇領域画像に対して VAE のエンコーダによる特徴量抽出を行い，特徴量の時系列データを作成した．このデータに対してデータの長さの統一化を図るためパディング処理を行い，学習済みの単語認識モデル入力することで単語の認識を行った．
表 4 に単語の認識結果を示す．色付きになっている単語は，認識結果が正解文と一致していることを示す．表 4 より，最も可能性が高いと推定される第一候補では 15 単語中 11 単語を正しく認識することができた．次に， LipNet を日本語に適用した研究で使用された単語セットで認識精度の確認を行った．[２]
表 5 に 20 種類の発話リストに登録する単語を示す．また，表 6 に認識結果を示す．
第一候補では 20 単語中 12 単語を認識することができた．また第二候補と第三候補の結果を含めた場合では，20 単語中 19 単語を認識することができた．これは単語の認識率を算出すると，第一候補のみでは 60 %，第三候補までの含めた場合では 95%であった．LipNet での単語の認識率は 36%であったことから，提案手法での結果は比較的に良好であるといえる．[３]
（９）携帯用発声支援システム
次に携帯用発声支援システムの開発を行った．開発には AndroidStudio 上で言語は Python Kotlin を用いて行った．また PC 上のプログラムをスマートフォンでも実行するために Chaquopy を用いて実行した． この場合 Python ライブラリの 88 %をスマートフォン上でも実行できる．これにより PC 上のプログラムをスマートフォンでもほぼ実行することができる．カメラアプリは Google が Android で導入しているカメラアプリ Camera2API を用いた．
本システムは発声支援装置として使用者が円滑なコミュニケーションを取れることを考慮し，単語の認識精度や遅延といった課題点に注力した．まず認識精度を考慮したシステムを検討した．すなわち認識結果の候補を第 1 候補だけではなく第 6 候補まで出力し，選択できるようにした．これにより発話したい単語の誤認識補正が容易になる．
（１０）インタフェースの評価実験
以上の操作方法を健常者 10 名に説明し使用感調査を行った．(COVID19 対策のため健常者による実験とした．) 既存の発声支援装置である電気式人工喉頭を使用してもらい発声支援装置についての理解を深めてもらいながら開発した携帯用発声支援システムを体感してもらった．実験結果より発声支援システムは使用にストレスを感じない． 使用している姿に抵抗を感じない ．インタフェースが使いやすいなどの評価が得られた． また 候補ボタンに関しても使いやすいという意見があった．
（１１）携帯型単語認識評価実験
第 1 候補での認識精度は 40%であり PC より低い結果となった．その原因としてスマートフォンの取得した画像の枚数差がある．PC とスマートフォンの同じ速度で画像を取得した枚数 は PC の 1/3 程度であることから認識精度が低下した．そこで PC と同じ 特徴量を取得した結果，第一候補では 20%程度認識率が低下するが，第二候補以下では PC 版とほぼ同等となった．
（１２）深度情報を用いた読唇方式の検討
口唇の機械読唇方式を開発し，スマーとフォンに実装して評価を行った．課題として認識性能があり，次に深度情報による子音認識を検討した．2014 年の押尾らによる「口唇の深度画像を用

いたマルチモーダル音声認識」では Microsoft Kinect を使用して深度画像とカラー画像による認識を行なっている．この研究によると複数の話者データを用いるタスクに対して有意義な性能向上が見られた．音素単位での認識率は口を尖らせる形を取る音素や口腔を舌で塞ぐ動きをする音素に対して特に有効であることが確認されているためこれらの項目に注意して認識実験を行うものとした．本研究では子音での単音節認識を行い深度画像利用の基本的特性を把握することとした．このために変更した口形列を表７に示す．

表７　単音節認識に使用した基本口系列

あ	か	さ	た	な	は	ま	や	ら	わ
A	KA	SA	TA	NA	HA	MA	YA	RA	WA
い	き	し	ち	に	ひ	み		り	
I	KI	SHI	CHI	NI	HI	MI		RI	
う	く	す	つ	ぬ	ふ	む	ゆ	る	
U	KU	SU	TSU	NU	HU	MU	YU	RU	
え	け	せ	て	ね	へ	め		れ	
E	KE	SE	TE	NE	HE	ME		RE	
お	こ	そ	と	の	ほ	も	よ	ろ	
O	KO	SO	TO	NO	HO	MO	YO	RO	

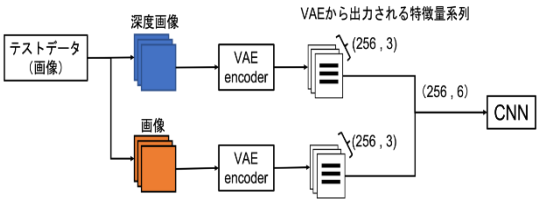


図１０　深度画像を加えた認識手順

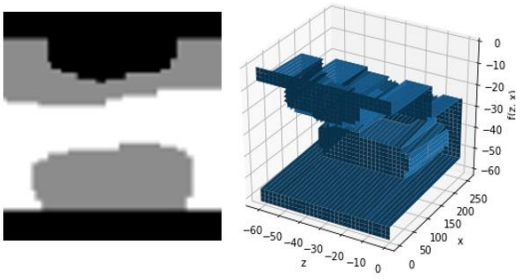


図１１　取得した深度画像「あ」の口形

表９　単音節の認識率

	認識結果(%)				
ファイル名	第１候補	第２候補	第３候補	第４候補	第５候補
test3-1(RGB)	6.25	18.8	21.9	25.0	31.2
test3-2(Depth)	6.25	15.6	24.0	34.4	34.4
test3-3(RGB)	13.5	13.5	26.0	33.3	43.8
test3-4(Depth)	6.25	31.3	44.8	50.0	56.3
test7-1(RGB&Depth)	6.25	13.5	15.6	18.8	29.2
test7-2(RGB&Depth)	0.00	6.25	20.8	31.3	43.8

本研究では通常の画像と深度画像をそれぞれ VAE ベクトルとして保存し，擬似的な口唇特徴量系列を作成する際に 256 要素６次元の入力データを作り CNN のモデル学習を行う． 認識の手順を示したものを図１０に示す．学習済みの VAE エンコーダを用いて得られた特徴量系列データを結合・パディングなどの処理を行なったのち 256 要素，3 次元の 2 つのデータを 256 要素，6 次元のデータに統合し学習済み CNN の入力とすることで発話単語の認識を行う．

（１３）深度画像を用いた認識実験

深度カメラ RealSenseD435 を用いて情報の取得と精度について検証実験を行なった．撮影した深度画像と 3D グラフを図１１に示す．

舌，歯，上唇などが確認できる．RGB 画像と深度画像を別々に分けて使用して学習，認識を行った．表９に第５候補までの認識結果を示す．

（１４）考　察

実験結果より，深度画像を用いた従来の lip2word による学習は単音節の識別に有効な場合が見られたがデータによってその精度は不安定であり，深度情報の組込み方に関する検討が必要である．例として閉唇状態や文字と文字の間の遷移状態を口形列生成の際に挿入することで精度向上が期待できる．また，現在の認識手法は CV のみの単音節認識なので VCV 単位のアルゴリズムを追加することで認識性能の向上につながると考える．

今後の課題に関して以下にまとめる．

(1) 携帯端末を考慮した深度情報の取得手法の検討および特徴量の組込み方に関する検討

(2)認識単位を VCV に変更して認識精度が CV に比較して向上することの確認

(3) VAE，CNN のフィルタを変更した際の認識精度の確認

(4) 既存の携帯型デバイスへの実装

以上の課題に取り組み，代用音声を必要とする喉頭摘出者などのユーザーに対し発声支援システムの実用化を目指す．

＜引用文献＞

［１］ Kingma, Diederik P., and Max Welling. ″Auto-encoding variational bayes.″ arXiv preprint arXiv:1312.6114 ,2013

［２］ Daiki Ito, Tetsuya Takiguchi, Yasuo Ariki, Lip image to speech conversion using LipNet, Acoustic Society of Japan articles, March 2018

［３］ 浅見，石川，渡辺,”機械学習による日本語話者の自動読唇の基礎検討”，第 33 回画像符号化シンポジウム・第 23 回映像メディア処理シンポジウム（PCSJ/IMPS2018），P-3-08，Nov. 2018

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　2件／うちオープンアクセス　0件）

１．著者名 Fumiaki Eguchi, Kenji Matsui, Yoshihisa Nakatoh, Yumiko O. Kato, Alberto Rivas, Juan Manuel Corchado	４．巻 1
２．論文標題 Development of Mobile Device-Based Speech Enhancement System Using Lip-Reading	５．発行年 2021年
３．雑誌名 Distributed Computing and Artificial Intelligence	６．最初と最後の頁 210 - 220
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-86261-9	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Tomonori Nakahara, Kohei Fukuyama, Mitsuru Hamada, Kenji Matsui, Yoshihisa Nakatoh, Yumiko O. Kato, Alberto Rivas, Juan Manuel Corchado	４．巻 1237
２．論文標題 Mobile Device-based Speech Enhancement System Using Lip-reading	５．発行年 2020年
３．雑誌名 Advances in Intelligent Systems and Computing	６．最初と最後の頁 159 - 167
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-53036-5	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

〔学会発表〕　計8件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 Kenji Matsui, Kohei Fukuyama, Yoshihisa Nakatoh, Yumiko O. Kato
２．発表標題 Speech Enhancement System Using Lip-reading
３．学会等名 2nd IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET 2020)（国際学会）
４．発表年 2020年

１．発表者名 中原智典，福山晃平，松井謙二，中藤良久，加藤弓子
２．発表標題 発声支援のための口形素列によるフレーズ認識方式の検討
３．学会等名 日本音響学会2020年秋季研究発表会
４．発表年 2020年

１．発表者名
Yuto Kinoshita, Rin Hirakawa, Hideaki Kawano, Kenichi Nakashi, Yoshihisa Nakatoh

２．発表標題
Speech Enhancement System Using SVM for Train Announcement

３．学会等名
The 39th IEEE International Conference on Consumer Electronics（IEEE ICCE 2021）（国際学会）

４．発表年
2021年

１．発表者名
Tomonori Nakahara, Kohei Fukuyama, Mitsuru Hamada, Kenji Matsui, Yoshihisa Nakatoh, Yumiko O. Kato, Alberto Rivas, Juan Manuel Corchado

２．発表標題
Mobile Device-based Speech Enhancement System Using Lip-reading

３．学会等名
17th International Conference on Distributed Computing and Artificial Intelligence（国際学会）

４．発表年
2020年

１．発表者名
福山晃平，濵田三 弦，松井謙二，中藤良久，加藤弓子

２．発表標題
携帯機器と口唇情報利用による発声支援方式の検討

３．学会等名
日本音響学会2019年秋季研究発表会

４．発表年
2019年

１．発表者名
濵田三 弦，福山晃平，松井謙二，中藤良久，加藤弓子

２．発表標題
携帯機器 を用いた口唇情報利用 発声支援デバイスの開発

３．学会等名
日本音響学会2020年春季研究発表会

４．発表年
2020年

１．発表者名 福山晃平，松井謙二，中藤良久，加藤弓子
２．発表標題 発声支援のための読唇手法の検討
３．学会等名 日本音響学会2020年春季研究発表会
４．発表年 2020年

１．発表者名 江口文耀、松井謙二、中藤良久、加藤弓子
２．発表標題 携帯端末を用いた口唇認識による発話支援の検討
３．学会等名 日本音響学会2021年秋季研究発表会
４．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	中藤　良久 (Nakatoh Yoshihisa) (10599955)	九州工業大学・大学院工学研究院・教授 (17104)	
研究分担者	加藤　弓子 (Kato O. Yumiko) (10600463)	聖マリアンナ医科大学・医学部・研究員 (32713)	
研究分担者	水町　光徳 (Mizumachi Mitsunori) (90380740)	九州工業大学・大学院工学研究院・准教授 (17104)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関