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　 ５ 日現在

機関番号：１５３０１
研究種目：基盤研究(C)（一般）
研究期間：2018～2021
課題番号：１８Ｋ１１３７６

研究課題名（和文）ディープニューラルネットワークによる舌亜全摘出者の音韻明瞭性改善の研究

研究課題名（英文）A Study on Algorithms to Improve Intelligibility of Glossectomy Patients' Speech Using Deep Neural Networks

研究代表者
阿部　匡伸（Abe, Masanobu）

岡山大学・ヘルスシステム統合科学学域・教授

研究者番号：７０５９５４７０

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：舌癌の治療で舌の一部を切除された患者（舌亜全摘出者）は，カ行，サ行，タ行の音韻を明瞭に発声できなくなり日常生活に支障を来す．本研究では音声信号処理によって，舌亜全摘出者が発声した音声を明瞭で聞き易い音声に変換する方式を検討した．提案方式はリアルタイム処理を実現するために差分スペクトル補正に基づく声質変換方式をベースとし，健常者音声と舌亜全摘出者音声との差分スペクトルをDeep Neural Networks（DNN）によって推定する．明瞭性の一層の向上を目指し，知識蒸留アプローチによる音韻情報の利用と，補助的な口唇形状の利用を提案し，評価実験により変換精度の改善を示した．

研究成果の学術的意義や社会的意義
音声はコミュニケーションの手段としてばかりでなく，人間としての尊厳を保ち豊かな生活を送るうえで重要な役割を果たしている．舌，顎，唇（以下，調音器官）の癌治療のために調音器官を切除して明瞭な音声を発声できなくなることは日常生活に測り知れない損失をもたらす．本研究では，癌治療によって舌を切除したために，音声を明瞭に発声できなくなった患者を対象に，患者が健常であった頃の音声を取り戻すための技術を提案し，その有効性を示した．2017年の国立がん研究センターの推計によれば，口腔・咽頭癌の患者数は約22,800人（癌患者の約２％を占める）であり，これらの患者が声を取り戻せる可能性を示した．

研究成果の概要（英文）：In this study, we investigate voice conversion algorithms to improve intelligibility of speech uttered by a patient who has articulation disorders because of wide glossectomy and/or segmental mandibulectomy. To achieve real time processing, voice conversion directly modifies waveform using spectrum differential between a healthy speaker and a glossectomy speaker. The spectrum differential is estimated by Deep Neural Networks(DNN). To improve the performance, we proposed to use lip shapes as auxiliary inputs and to introduce knowledge distillation approach to make best use of phoneme labels as auxiliary inputs. Experimental results showed that both approaches work well, and phoneme labels with knowledge distillation has better performance than the usage of lip shapes.

研究分野：音声情報処理

キーワード：声質変換　音声合成　舌亜全摘出者　DNN　知識蒸留

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

音声はコミュニケーションの手段としてばかりでなく，人間としての尊厳を保ち豊かな生活を送るうえで重要な役割を果たしている．舌，顎，唇（以下，調音器官）の癌治療のために調音器官を切除して明瞭な音声を発声できなくなることは日常生活に測り知れない損失をもたらす．現状ではリハビリテーションによる訓練のみが声を取り戻す唯一の手段であるが，調音器官の切除量の程度によってはその手段すらままならないことが多い．2017 年の国立がん研究センターの推計によれば，口腔・咽頭癌の患者数は約 22,800 人（癌患者の約２％を占める）であり，多くの患者が声を取り戻すことを望んでいる．そこで本研究では，癌治療によって舌を切除したために，音声を明瞭に発声できなくなった患者を対象に，患者が健常であった頃の音声を取り戻すための技術を確立する．

２．研究の目的

舌亜全摘出者の音韻明瞭性を音声信号処理によって改善する方式を確立する．これにより，従来の器具装着方式に比べて音韻明瞭性を向上させる．これと同時に，食事など器具を装着できない状況も含めて，患者が音声コミュニケーションを広く享受することに資する．

３．研究の方法

従来の音声分析や声質変換の技術は，人間が発声した「正常な」音声のモデリングが主たる目的であった．本研究では，このモデリングを舌亜全摘出者の音声という「異常な」音声へ応用する．具体的には，舌の切除によって失われた情報を復元するために，声質変換のアプローチを導入する．発想のポイントは，調音器官の連続的な動きによって音声が生成されていることは，「正常音声」でも「異常音声」でも同じであると考える点にある．つまり，舌亜全摘出者と健常者とが発声した音声データを用いて，復元すべき音声スペクトルを推定する．大雑把にいえば次のようなことである．/t/や/r/という音素は舌が無いと発音できないが，健常者の発声した/tobira/と患者が発声した/tobira/とを突き合わせてみると，/t/や/r/の前後のスペクトルの変化から「患者の音声に/t/や/r/がありそうなスペクトルパタン」が抽出できる．このように前後音韻のコンテキストを制約として用いることで，復元すべき特徴量を推定する．

４．研究成果

（１）舌亜全摘出者のシミュレーション

アルゴリズムの検討には，舌亜全摘出者患者の音声データが欠かせない．一方，舌亜全摘出者患者に研究のためのデータを発声させることはかなりの負担になる．また，アルゴリズム開発においては，同じ患者の舌摘出前後の音声であることが望ましい．これは，調音器官の大きさや動かし方が話者毎に異なるためであり，舌亜全摘出者患者の音声と別の健常者音声を比較しても舌摘出の影響だけを明確にできないためである．以上のことから，舌摘出をシミュレートするために図１に示すような舌抑制器具を作成した．このプレートを下歯にはめ込むことによって，舌がプレートより上に移動することを抑制する．これにより，舌を用いた摩擦音や破裂音を発声できなくなり，舌亜全摘出者の音声がシミュレートできる．以下では．この音声を疑似舌亜全摘出者音声と呼ぶ．本器具を５人の健常者ごとに作成し，この器具の装着の有無により，同じ話者の健常音声と疑似舌亜全摘出音声を収録した．各話者 503 文である．スペクトログラムの観察によって，疑似舌亜全摘出音声が本物の舌亜全摘出者音声と酷似していることを確認した．

図１　舌抑制器具

（２）リアルタイム動作

舌亜全摘出者を支援するアプリケーションを提供するためにはリアルタイム動作が必須となる．差分スペクトル補正における MLSA フィルタリング処理はボコーダ方式と比較して計算コストが安価であり，リアルタイム処理が実現できる．さらに，以下の特徴がある．

・差分スペクトル補正では F0 の分析が不要であり，F0 抽出ミスの音質劣化がない．
・舌摘出者は声帯には問題が無いため，入力音声の声帯音源に含まれる特徴を利用できる．
・舌亜全摘出者は口腔内に狭めを作り乱気流を発生させることが難しく，舌を必要とする音素については励振力が弱くなる傾向がある．従って，ボコーダ方式では音声分析の段階で音源情報の抽出が難しいが，差分スペクトル補正では音源情報の抽出が不要である．

差分メルケプストラムを用いた変換処理の概要を図２に示す．差分メルスペクトラムの MLSA フィルタを音声波形で直接駆動しているが，これは数式的には推定したメルスペクトラムの MLSA フィルタを残差信号を用いてを駆動することと等価である．この方式は差分スペクトル補正に基づく声質変換と呼ばれる．なお，入力話者と目標話者が同一人物であるならば，入力音声の F0 は目標音声の F0 と等しいといえるので問題は無い．逆に同一人物でない場合は入力話者

F0を目標話者のF0に合わせる処理が必要となる．評価実験の結果，差分スペクトル補正は従来のボコーダ方式に比べて，話者性の変換精度が高いこと，高い自然性を持つことが明らかとなった．また，変換処理がリアルタイム以下の時間で可能であることを示した．

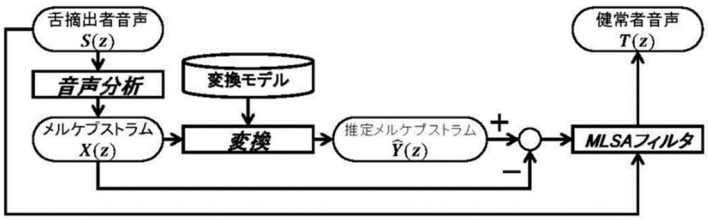


図２ 差分メルケプストラムを用いた変換処理

（３）音韻明瞭度の改善方式の検討

①音響特徴量の時間変化の導入と補助情報としての音素ラベル導入

音韻明瞭度改善の基本方式はDeep Neural Networks（DNN）を用いた声質変換法である．入力を疑似舌亜全摘出者の音声とし，同一話者が発声した健常音声を出力するようにDNNで差分スペクトルを推定する．

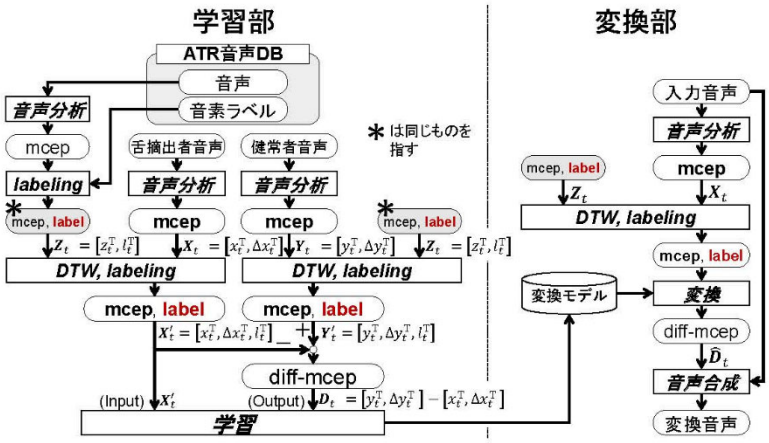
舌亜全摘出者の音声は調音器官の一部が欠落しているため，複数の音韻が明確に分離されず，あたかも１つの音韻に聞こえてしまう現象（多対一対応）が起こる．そこで，音響特徴量の時間変化情報を用いて音韻の分離を向上させる目的で，Bidirectional Long Short-Term Memory based Recurrent Neural Network (BLSTM-RNN)を適用する方式を提案した．また，多対一対応の問題が解決された場合の上限値を明確にするため，補助情報として音韻ラベルが与えられると仮定し，その場合の変換性能を明らかにした．

提案方式の学習部(図３(a))では，大きく以下の４つの処理をおこなう．

- ATR DB の音響特徴量に対して音素ラベルを付与する
- 舌亜全摘出者の音響特徴量に対してDTW により音素ラベルを付与する
- 舌亜全摘出者と健常者間の差分音響特徴量を生成する
- 変換モデル(BLSTM-RNN)を学習する．

提案方式の変換部(図３(b))では，大きく以下の３つの処理をおこなう．

- 舌亜全摘出者の音響特徴量に対してDTW により音素ラベルを付与する
- 学習済み変換モデル(BLSTM-RNN)により差分特徴量を推定する
- 変換音声を合成する



(a) (b)

図３ 音素補助情報利用型BLSTM-RNN変換処理

変換モデル(BLSTM-RNN)のネットワーク構成を図４に，学習パラメータを表１に示す．

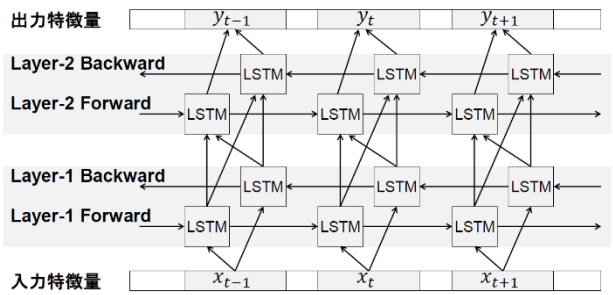


図４ BLSTM-RNNのネットワーク構成

表１ BLSTM構成及び学習条件

BLSTMの条件	
ネットワーク構成	
[I, L-128, PReLU, BLSTM-128, BLSTM-128, L-128, O-50]	
I: 入力層 (Input layer)	PLAなし: I-50，PLAあり: I-95
O: 出力層 (Output layer)	
L: 線形層 (Linear layer)	
PReLU [20]: 活性化関数	
(付随する数字はユニット数を表す)	
損失関数	平均二乗誤差
最適化手法	Adam ($\alpha = 0.001$, $\beta1 = 0.9$, $\beta2 = 0.999$)
ミニバッチサイズ	2 sentences

②知識蒸留アプローチの導入

知識蒸留とは，学習済みのモデルを教師モデルとし，教師モデルが獲得した知識を利用して生徒モデルを学習する枠組みである．大規模データを用いて汎用性の高い教師モデルを学習して，小規模データしかない特定タスクの生徒モデルを学習したり，大規模なDNN構造で教師モデルを学習して，高速に処理するための小規模な生徒モデルを学習したりする．この考え方を適用し，①の方式により音声データと音素ラベルとを用いて教師モデルを学習し，音声データだけで生

徒モデルを学習する方式を提案した．これは，実際の場面では，音素ラベルが得られないためである．

図５に学習済み教師モデルを用いて生徒モデルを学習する方法を示す．学習部では，教師モデルの中間層(以下,ML-KD 層(MiddleLayer for Knowledge Distillation と呼ぶ)の出力と生徒モデルの中間層の誤差を最小化することで，知識蒸留をおこなう．教師モデルは①によって学習する．生徒モデルの学習では，音素ラベルを用いず舌摘出者の音響特徴量のみを入力特徴量とし，教師モデルの ML-KD 層の出力を学習に用いる．

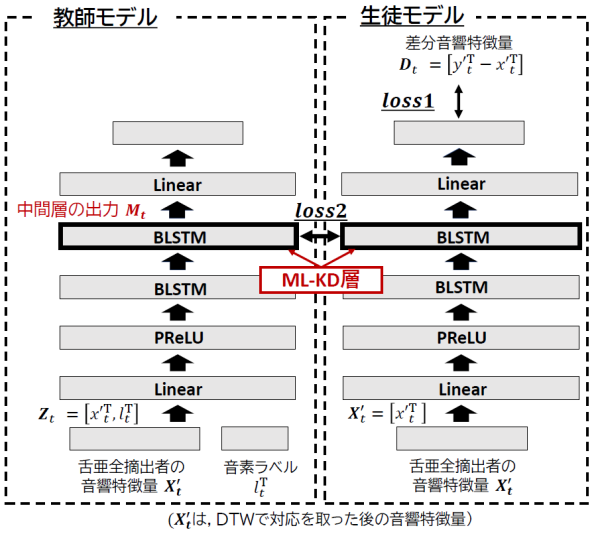


図５ 知識蒸留学習モデルの学習法

③補助情報としての口唇形状導入

多対一対応の問題を解決するためには，舌亜全摘出者が意図した音素を推定するために，音響情報以外の情報を補助的に利用することが考えられる．難聴者がリップリーディングできることから唇の動きは音素特定に有効であると考えられる．そこで口唇形状を補助情報として取り入れる方式を提案した．口唇特徴量は CNN-Autoencoder のボトルネック特徴量である．この特徴量を音響特徴量と同様にフレーム単位で抽出し，①の音響特徴量と並行して学習する．

CNN-Autoencoder の概要図を図６に示す． マイクロソフト Kinect（赤外線レーザーと赤外線カメラを用いた計測方式）で収録した２次元顔座標データ群（700 次元）を口唇特徴量として用いる．Kinect で収録した顔座標には，顔の向きの正規化と音声との時間軸調整をおこなっておく．まず，口唇特徴量を入力として，入力特徴量と復元された特徴量の誤差を最小化するように CNN-Autoencoder を学習する．そして，学習済み CNN-Autoencoder の Encoder 部を利用して，口唇特徴量からボトルネック特徴量を抽出する．ボトルネック特徴量は，口唇特徴量よりも小さい次元で元の特徴量に復元することが可能である．ボトルネック特徴量を変換モデルの学習のための追加の入力特徴量として利用する．

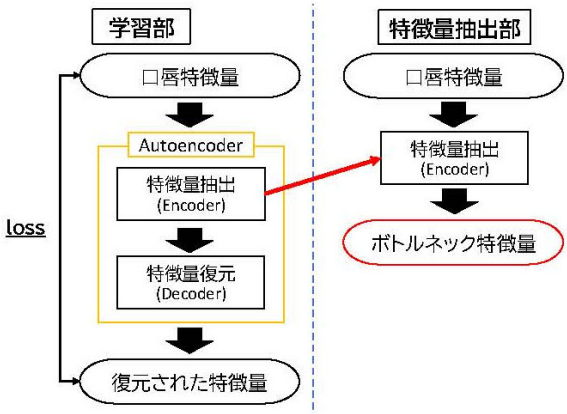


図６ CNN-Autoencoder の概要

表２ CNN-Autoencoder のネットワーク構成

Encoder 部				Decoder 部			
Layer type	cannel	ksize	stride	Layer type	cannel	ksize	stride
L_1 *convolution*	128	(1,2,2)	(1,2,2)	L_9 *convolution*	1	(1,1,2)	(1,1,2)
L_2 *convolution*	64	(1,1,2)	(1,1,2)	L_{10} *convolution*	2	(1,1,2)	(1,1,2)
L_3 *convolution*	32	(1,1,3)	(1,1,1)	L_{11} *convolution*	4	(1,1,2)	(1,1,2)
L_4 *convolution*	16	(1,1,2)	(1,1,1)	L_{12} *convolution*	8	(1,1,2)	(1,1,2)
L_5 *convolution*	8	(1,1,2)	(1,1,2)	L_{13} *convolution*	16	(1,1,2)	(1,1,1)
L_6 *convolution*	4	(1,1,2)	(1,1,2)	L_{14} *convolution*	32	(1,1,3)	(1,1,1)
L_7 *convolution*	2	(1,1,2)	(1,1,2)	L_{15} *convolution*	64	(1,1,2)	(1,1,2)
L_8 *convolution*	1	(1,1,2)	(1,1,2)	L_{16} *convolution*	128	(1,2,2)	(1,2,2)

（４）評価実験

①音響特徴量の時間変化の導入と音素ラベル導入の効果

健常者男性(Male1) が発声した音声と Male1 が発声した疑似舌摘出者音声 SPM1 (Simulated Patient Male1)を用いて評価した．DNN モデル構築には 400 文の音声データを学習に 50 文の音声データを検証に用いた．また，評価実験はこれらとは別の 50 文の音声データを用いた． これらを用いて，下記のモデルを構築した．

・DNN-ボコーダ型：DNN-VVC
・DNN-差分スペクトル型：DNN-DVC
・DNN-音韻ラベル付き-ボコーダ型：DNN-PLA-VVC
・DNN-音韻ラベル付き-差分スペクトル型：DNN-PLA-DVC
・BLSTM-ボコーダ型：BLSTM-VVC
・BLSTM-差分スペクトル型：BLSTM-DVC
・BLSTM-音韻ラベル付き-ボコーダ型：BLSTM-PLA-VVC
・BLSTM-韻ラベル付き-差分スペクトル型：BLSTM-PLA-DVC

疑似舌摘出者の音韻明瞭度改善をメルケプストラム歪みにより評価した．音素ごとに算出したメルケプストラム歪みを図７に示す．ボコーダ方式(VVC) と差分スペクトル補正方式(DVC)を比較すると．全体的に DVC の方が良い．これは，DVC では舌亜全摘出者の声帯音源をそのまま用いて変換するため，ボコーダ方式では無視されてしまうような励振力の弱い音素も復元できたためと考えられる．PLA (音素ラベル付与)は，特に調音に舌を必要とする摩擦音(/s/, /z/, /sh/) 等の音素について大幅にスペクトル誤差を低減できている．これは，音素ラベルを付与することで多対一対応の問題が解消できたためと考えられる．PLA の効果は BLSTM のモデル構造で顕著である．時間依存の強い音素ラベルに対して BLSTM が長期の時間依存情報を考慮できるためと考えられる．特に破裂音の音素については提案方式の有効性が顕著であるが，これは BLSTM により継続時間長の短い破裂音の音素を捉えることが容易になったためと考えられる．

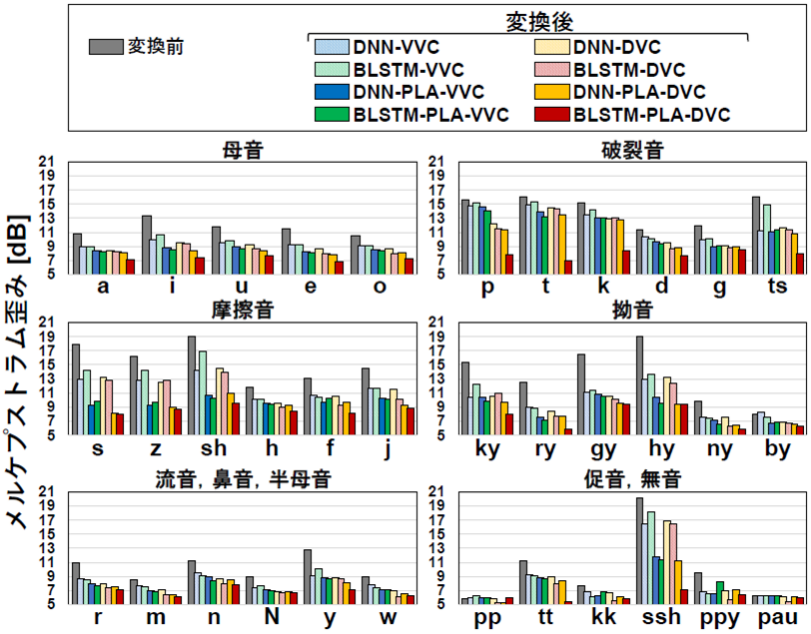


図７各モデルの音素ごとのメルケプストラム歪み

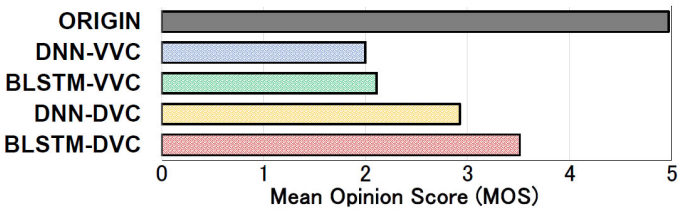


図８ 変換音声の自然性の MOS 評価

変換音声の自然性を MOS により評価した．結果を図８ に示す．ORIGIN とは，SPM1 の元音声である．音声の音韻明瞭度は評価対象としないため PLA (音素ラベル付与) に関する変換モデルについては評価から除外した．結果より，BLSTM-DVC が最も高い自然性を実現しており，ボコーダ型よりも差分スペクトル補正型方が有効であると言える．また，BLSTM が良い理由は，長期の順方向・逆方向の時間依存情報を考慮して変換しており，frame-by-frame の変換方式におけるフレーム境界での歪みの発生を回避できたためと考えられる．

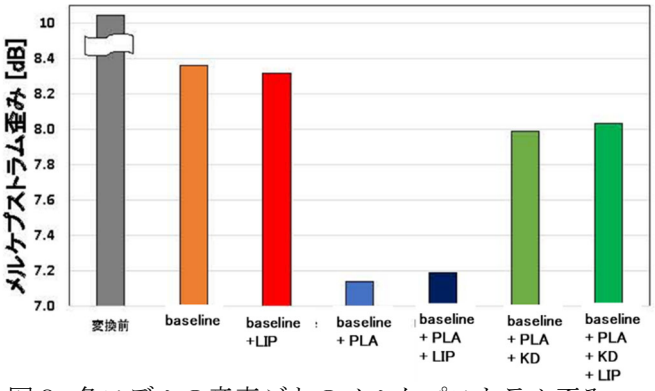


図９ 各モデルの音素ごとのメルケプストラム歪み

②知識蒸留と口唇形状導入の効果

BLSTM-差分スペクトル型 (BLSTM-DVC) をベース (baseline) として，知識蒸留 (KD)，口唇形状の利用 (LIP)，音素ラベル付与 (PLA) を組合せた声質変換モデルを構築して評価した．変換音声のメルケプストラム歪みを図９に示す．図９より，PLA の効果が最も大きいことが分かる．また，baseline に口唇情報を加えるよりも，知識蒸留の処理を行った方が効果が大きいことが分かる．PLA に口唇情報を加えると性能が低減した．これは，音素ラベルが補助情報として十分であるため，口唇情報を加えることで，かえって入力情報と出力情報の対応が曖昧になったことが原因だと考えられる．

図１０に変換音声のスペクトログラムを示す．PLA や KD では，破裂音の閉鎖区間が再現されているのが分かる．

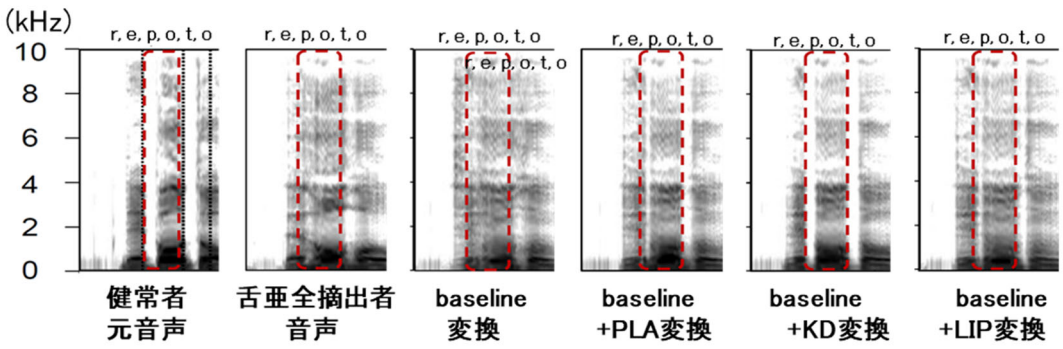


図１０ 変換音声のスペクトログラム

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計8件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 Hiroki Murakami, Sunao Hara, Masanobu Abe
２．発表標題 DNN-based Voice Conversion with Auxiliary Phonemic Information to Improve Intelligibility of Glossectomy Patients' Speech
３．学会等名 APSIPA Annual Summit and Conference 2019,（国際学会）
４．発表年 2019年

１．発表者名 荻野聖也，原直，阿部匡伸
２．発表標題 舌亜全摘出者の音韻明瞭度改善のための推定音素事後確率を用いた声質変換の検討
３．学会等名 電子情報通信学会総合大会
４．発表年 2020年

１．発表者名 荻野聖也，村上博紀，原直，阿部匡伸
２．発表標題 音声と口唇形状を用いた声質変換による舌亜全摘出者の音韻明瞭度改善の検討
３．学会等名 電子情報通信学会技術研究報告
４．発表年 2018年

１．発表者名 村上博紀，原直，阿部匡伸
２．発表標題 声質変換による舌亜全摘出者の音韻明瞭度改善のための補助情報の検討
３．学会等名 日本音響学会2018年秋季研究発表会
４．発表年 2018年

１．発表者名
村上博紀，原直，阿部匡伸

２．発表標題
舌亜全摘出者の音韻明瞭度改善のための Bidirectional LSTM-RNN に基づく音素補助情報を用いた声質変換方式の検討

３．学会等名
日本音響学会2019年春季研究発表会

４．発表年
2019年

１．発表者名
荻野聖也，村上博紀，原直，阿部匡伸

２．発表標題
声質変換による舌亜全摘出者の音韻明瞭度改善のための音素補助情報の推定方式の検討

３．学会等名
日本音響学会2019年春季研究発表会

４．発表年
2019年

１．発表者名
高島和嗣，原直，阿部匡伸

２．発表標題
音素情報を知識蒸留する舌亜全摘出者の音韻明瞭度改善法

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

１．発表者名
高島和嗣，原直，阿部匡伸

２．発表標題
口唇特徴量を利用した知識蒸留による舌亜全摘出者の音韻明瞭度改善法の検討

３．学会等名
電子情報通信学会技術研究報告

４．発表年
2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

ディープニューラルネットワークによる 舌亜全摘出者の音韻明瞭性改善の研究 http://site-330980-922-1588.mystrikingly.com/

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	原　直 (Hara Sunao) (50402467)	岡山大学・ヘルスシステム統合科学研究科・助教 (15301)	
研究分担者	皆木　省吾 (Minagi Shogo) (80190693)	岡山大学・医歯薬学総合研究科・教授 (15301)	

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