ダイバー同士の水中会話を支援するための深層学習を用いた音声認識手法

土田, 修平 ; Worachat, ARUNOTHAIKRIT ; Haomin, MAO ; 大西, 鮎美 ; 寺田, 努 ; 塚本, 昌彦 ; 滝口, 哲也

(Citation)
マルチメディア, 分散協調とモバイルシンポジウム2021論文集:317-324

(Issue Date)
2021-07

(Resource Type)
conference paper

(Version)
Version of Record

ダイバー同士の水中会話を支援するための深層学習を用いた音声認識手法

土田 修平[1]　Worachat Arunothaikrit[1]　Haomin Mao[1]　大西 鮎美[1]　寺田 努[1]　塚本 昌彦[1]
滝口 哲也[2]

概要：ダイバーが水中でコミュニケーションをとることは，安全面や娯楽面などの観点から重要である．水中における即時性・柔軟性に優れたコミュニケーション方法としては，聴覚を介した音声コミュニケーションが挙げられる．しかし．ダイバーは口にレギュレータを装着しているため，口を正しく動かすことができず，正しい発音で話すことができない．そこで本研究では，水中での円滑な音声コミュニケーションの実現を目指し，レギュレータを口で咥えた状態での不明瞭な音声から音素を認識するシステムを提案する．ダイバー用レギュレータに防水マイクを取り付け，水中での音声を録音し，深層学習を用いて音素の推定を試みた．また，発音の際の口の変化に着目し，ダイバー用レギュレータの上面と左面に圧力センサを設置し，推定精度の向上を試みた．その結果，音声のみを用いたデータの場合の方が推定精度が高かったが，特定の音素においては音声と圧力センサ値の混合データを用いた場合の方が推定精度が優れていることがわかった．

1. はじめに

ダイバーが他のダイバーと水中でコミュニケーションをとることは，安全面や娯楽面などの観点から重要である．水中でのコミュニケーションは，手話などの視覚を介したコミュニケーションやトランシーバなどの聴覚を介したコミュニケーション（音声コミュニケーション）がある．しかし視覚を介したコミュニケーションは，水の抵抗や視界の制限により伝えられる情報が限られる．一方，音声コミュニケーションは伝達できる情報が多いものの，レギュレータを口で咥えているため口が大きく変形してしまい，口を正しく動かすことができず正しい発音で話すことができない．音声通信には，マイクを内蔵したフルフェイスマスクを使用する方法もある．しかし，フルフェイスマスクは高価であり，マスクの面が破損すると，水がマスクで守られている口腔・鼻腔内に浸入し，空気の供給が妨げられてダイバーが危険にさらされてしまう恐れがある．そのため，従来のレギュレータを装着したまま音声コミュニケーションを行うことが望ましい．ここで，レギュレータを口で咥えた状態での不明瞭な音声を自動音声認識技術（Automatic Speech Recognition: ASR）で認識できれば，認識した音声を元に合成音声を出力することで水中における円滑な音声コミュニケーションにつながると考えた．

[1] 神戸大学大学院工学研究科
Graduate School of Engineering, Kobe University
[2] 神戸大学 都市安全研究センター
Research Center for Urban Safety and Security, Kobe University

本研究では，水中での円滑な音声コミュニケーションの実現を目指し，レギュレータを口で咥えた状態での不明瞭な音声から音素を認識するシステムを提案する．まず水中での発話音声のデータセットを作成するために，ダイビング用レギュレータ内に防水マイクを取り付け，水を溜めた浴槽内で音声を録音する．また音素の認識精度を向上させるために，発話時の口の形状の変化に着目し，レギュレータに圧力センサを取り付け，発話時の圧力値の変化を音声の録音と同時に記録したデータセットを作成した．

音素の認識は，MFCC とスペクトログラムをそれぞれ特徴量として，CNN-BRNN を用いた深層学習モデルを使用した．具体的には，CNN-BRNN は Bidirectional Recurrent Neural Networks (BRNN) および Convolutional Neural Networks (CNN)，Time Delay Neural Network (TDNN) を組み合わせた学習モデルである．本研究では音声のみのデータと音声と圧力センサ値の混合データの 2 種類のデータを用いて提案システムのモデルの性能を評価した．ASR の標準的な性能指標である Word Error Rate (WER) による評価を行う．

2. 関連研究

発音，自動音声認識技術（ASR），オーディオ・ビジュアル音声認識，サイレントスピーチインタフェース（Silent Speech Interface: SSI）に関する研究について述べる．

2.1 発音

本節では，発音の仕組みを確認する．発音は口の動きだ

けでなく喉，鼻の動きによっても音声が異なる．そのため，それぞれの場所の動きでどのような音が出るか理解しておく必要がある．発音は大きく分けて母音と子音の 2 種類がある．母音は，「a」「e」「i」「o」「u」などの声道をあまり妨げずに発音される言語音である [9]．子音は，声道の完全または部分的な妨げによって発声される言語音である．子音は，調音点（Point of articulation）および調音法（Manner of articulation）で，両唇音（Bilabial sounds），唇歯音（Labiodental sounds），歯音（Dental sounds），歯茎音（Alveolar sounds），硬口蓋音（Patatal sounds），軟口蓋音（Velar sounds），声門音（Glottal sounds）の 7 つのタイプに分けることができる．調音点とは，子音の生成に関わる音声器官を指す．調音法とは，肺から口への空気の流れを意味し，破裂音（Plosives），閉鎖音（Stops），摩擦音（Fricative），破擦音（Affricates），鼻音（Nasals），液体（Liquids），滑舌（Glides）の 6 つのタイプがある．

2.2 自動音声認識

ASR に関する研究の概要は，Deng ら [19] が述べている．ASR は，音声データをテキストデータに変換するプロセスであり，Speech-to-Text とも呼ばれる．音声認識には長い歴史があり，現在では，深層学習の進歩を活用した研究が盛んである．

Alex ら [10] は，エンド・ツー・エンドの認識のために，CTC 損失関数を用いて，双方向リカレントニューラルネットワーク（LSTM）を構築している．このモデルは，Wall Street のコーパスの文字レベルの転写において，0.07 の WER を達成している．また Alex らは，LSTM RNNs アーキテクチャと CTC 損失関数を用いて，シーケンスデータの各音素ラベルを予測し，0.18 の WER を達成している．シーケンスデータは，TIMIT コーパスの音素を抽出したものである．

本研究では，深層学習モデルを活用した音素の認識を試みる．

2.3 オーディオ・ビジュアル音声認識

ASR の性能を向上させるために，読唇術のような視覚データを活用した研究がある．野田ら [12] は，口元の画像と音素ラベルで学習した畳み込みニューラルネットワーク（CNN）を用いて，視覚的音声認識（VSR）のための視覚的特徴抽出を行っている．また彼らの提案するシステムでは，隠れマルコフモデルを用いて孤立した単語の音素ラベルを認識している．実験の結果，6 人の話者の 40 音素に対する平均音素認識率は 58%で，VSR は子音よりも母音の認識に優れていることがわかった．すべての母音の認識率は 60-100%，子音の認識率は 20-80%程度であった．

本研究においても口の動きに着目し，センサを活用した ASR の向上に試みる．

2.4 サイレントスピーチインターフェース（SSI）

サイレントスピーチインターフェースとは，公共の場における周囲の騒音問題を解決するために，声を出すことを求められていない人間の発話をセンサで検出する方法である．SSI は電子的な読唇術の一種といえる [20]．舌の動き，口の動き，唇の動きの検出には，超音波センサや光学カメラ等が用いられる．暦本 [16] は，加速度センサと角速度センサを顎の下に装着し，顎の動きと舌の動きを測定した．それらセンサから 12 次元のデータを取得し，深層学習を用いて学習させた．その結果，35 種類の音声コマンド・フレーズを 94%の認識率で認識することに成功した．また Fukumoto [17] は，日常生活での使用を目的とした SSI システム「SilentVoice」を提案している．提案されたシステムは，マイクからの気流の方向をキャッチしつつ，音声を取り込むことができる．実験では，話者に口を閉じて話してもらい，口の横に標準的なマイクを置いて音声を収集した．その結果，85 の限られた命令文において，WER は話者従属条件で 0.02，話者独立条件で 0.07 となった．

このように口の周りに装着したセンサやマイクを用いることで発話音声を推定できるが，水中環境では適切な推定方法とはいえない．水中では，ダイバーに触れる水の流れによりセンサの測定値が不安定となる可能性がある．そこで我々は，防水マイクと圧力センサをダイビング用レギュレータに搭載することで，水中での発話の音素の認識を行う．

3. 音声データによる音声認識

唇は話し手の声を認識するための音声形成に重要な役割を果たしているが，ダイビング用レギュレータを装着した状態で話すと唇が自由に動かなくなるため，発音がはっきりしない場合がある．国際音声記号（IPA）によると，特にダイビング用レギュレータの影響を受けやすい子音は，両唇で発音する「両唇音」（p，b，m）と，下唇と上前歯で発音する「唇歯音」（f，v）の 2 種類と考えられる．これはレギュレータを装着していると，空気を遮断するための口を閉じることができないためである．

3.1 システム構成

提案システムは，TS モノラルマイク（MM-SPAMP6HM，サンワサプライ），IC レコーダ（DR-22WLVER2-J，TASCAM），標準的なダイビング用レギュレータ（METALSUBSECOND-STAGE-THUNDER，METAL-SUB），PC で構成される．図 1 に，水中音声の記録用デバイスを示す．標準的なダイビング用レギュレータに TSmono マイクを搭載しており，図 1 の赤丸の位置にマイクを固定した．また図 2 にデータの収集経路を示す．PC は，6GB のグラフィックメモリ（GeForce RTX 2060）を搭載しており， TensorFlow を用いて深層学習モデルを構築・学習

図 1　水中音声記録用デバイス: ダイビング用レギュレータの内部に防水の TS モノラルマイクが固定されている．

図 2　音声データの収集経路．

に用いた．

3.2　実験

3.2.1　環境

本節では，水中環境における音声データの収集方法について述べる．音声データの収集は，図 3 に示すように，日本語を母国語とする著者が浴槽の中で提案デバイスを装着し，ATR 音素バランス 503 文 [18] を読み上げることで，音声データを収録した．読み上げた例文を表 1 にデータの収集経路を示す．ATR 音素バランス 503 文は，503 の日本語文からなる日本語読み上げ音声コーパスであり，音声の総時間は 1 時間程度である．音声データは Wave 形式のファイルに保存した．データの最初と最後に無音の部分がない，話者の音声のみを含む音声データを使用し，dBFS

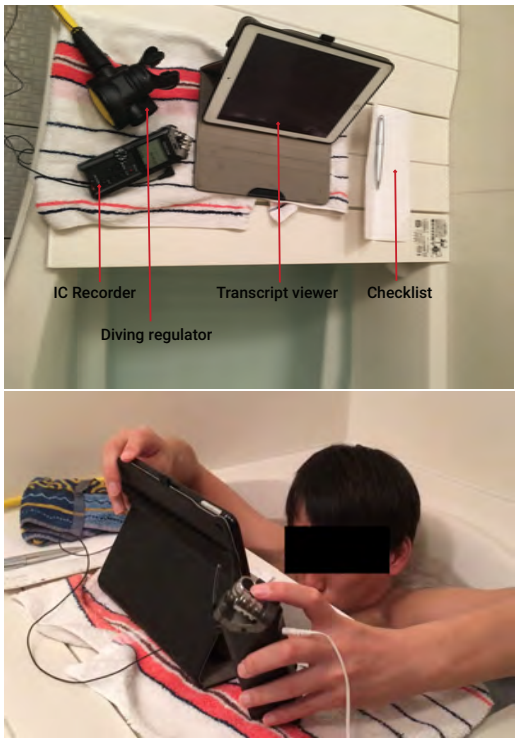

図 3　データ収集環境: バスタブに湯を張り，話者がレギュレータを咥えながら潜って ATR 音素バランス 503 文を読み上げた．

表 1　記録に用いたテキスト例

Type	Transcript
Kanji	あらゆる 現実を
	すべて自分の
	ほうへ ねじ曲げたのだ
Hiragana	あらゆる げんじつを
	すべてじぶんの
	ほうへ ねじまげたのだ
Pheneme	a r a y u r u g e N j i ts u o
	s u b e t e j i b u N n o
	h o: e n e j i m a g e t a n o d a

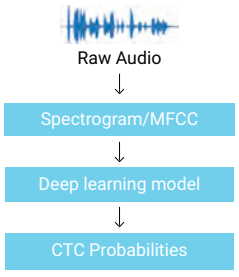

図 4　音声データのみを用いた音声認識手法

(decibels relative to full scale) が-40 以上のノイズを除去した．書き起こした音声は A-J 形式のファイルにし，各ラベルをスペースで区切って音素ラベルとして保存した．

3.2.2　モデル

音声データを用いた音声認識手法を図 4 に示す．まず，生の音声データを特徴量（スペクトログラム [15] または

表 2　特徴量の入力次元数

Feature Representation	Input Dimension
MFCCs	13
Spectrogram	161

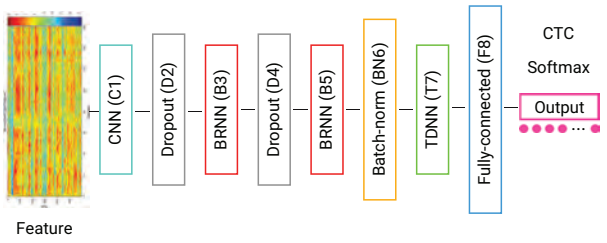

図 5　CNN-BRNN を用いた深層学習モデル: C，D，B，BN，T，F はそれぞれ，畳み込み層，ドロップアウト層，Bidirectional 層，Batch normalization 層，Time distribution 層，Fully-connected 層を表す．

MFCC [21]）に変換する．各特徴量の入力次元数を表 2 に示す．図 5 に CNN-BRNN を用いた深層学習モデルの構成を示す．双方向リカレント・ニューラルネットワーク（BRNN） [8]，畳み込みニューラルネットワーク（CNN），時間遅延ニューラルネットワーク（TDNN）を含んでいる．第 1 の畳み込み層（C1）は，入力層から MFCC またはスペクトログラムの特徴量を受け取る．双方向層（B3 および B5）は，前のドロップアウト層（D2 および D4）からのドロップアウト・レスポンスを入力として受け取り，2 つの隠れ層でデータを処理する．Batch normalization 層（BN6）は，モデルを高速化し，ミニバッチごとに入力をより安定させるために使用される．Time distribution 層（T7）には，2 つの畳み込み 2D 層が含まれる．Fully-connected 層 (F8) は，活性化関数としてソフトマックス関数を使用し，T7 からの出力を受け取り，41 音素の確率を出力する．CTC 損失関数は，現在のニューラルネットワークにおいて，時系列問題を解決する特性を持つ．そこで，正しい音素に割り当てられる確率を最大にするために，CTC 損失関数を使用した．また，確率ベクトルを音素に変換する文字マッピングを用いた．

3.3　結果

収集した音声データをうち，全音声データの 80%を「トレーニングデータ」，全音声データの 10%を「バリデーションデータ」，全音声データの 10%を「テストデータ」とし，モデルの性能評価を行った．WER は，音声認識システムと機械翻訳システムの標準的な性能指標であり，トランスクリプトの精度を比較するために使用される．WER は以下のように計算され，

$$WER = \frac{S + D + I}{N}$$

表 3　各特徴量における深層学習モデルの WER

Models	Data	Features	WER
CNN-BRNN	Audio-only	MFCCs	0.25
CNN-BRNN	Audio-only	Spectrogram	0.24

表 4　スペクトログラムと MFCC を用いた深層学習モデルの音素の予測のエラー数．テストセットの全音素数は 2158 個．緑色の背景のセルの方がエラー数が少ない．最下段は表示されている音素以外の全ての音素を指す．

		Phon-emes	The number of error prediction	
			Spectrogram	MFCCs
Vowel sounds		a	128	142
		e	52	56
		i	42	78
		o	82	112
		u	84	62
Consonant sounds	Bilabial sounds	b	22	14
		p	6	8
		m	50	40
		w	36	24
	Labiodental sounds	f	4	2
	Others	d	26	26
		h	48	52
		t	28	40
		n	28	22
		N	36	42
		r	40	32
		y	26	30
		...	186	188

S は置換単語数，I は挿入単語数，D は削除単語数，N は正解単語数を示す．

音声のみのデータに対して，MFCC とスペクトルグラムの各特徴量における深層学習モデルの WER を表 3 に示す．CNN-BRNN を用いた場合，スペクトログラムの結果が良く，WER は 0.24 であった．CNN-BRNN のエラー予測数を表 4 に示す．スペクトログラムを用いた場合の性能は，母音では MFCC の性能よりも優れているが，子音の Bilabial 音と Labiodental 音では劣っている．その他の子音では，特定の音素に対してスペクトログラムと MFCC がそれぞれ高い精度を示している．この 2 つの特徴量の差は小さく，どちらの特徴量を利用すべきかは明らかにならなかったが，水中での ASR に用いる特徴量として活用できる可能性がある．また母音の方が母音のエラー数が多くなっているが，日本語では母音が含まれる頻度が高いことが原因と考えられる．

4.　音声と圧力センサの混合データを用いた音声認識

ダイビング用レギュレータに圧力センサを取り付けたデバイスから収集する音声と圧力センサ値の混合データを用いた音声認識手法を提案する．唇の動きによってダイビン

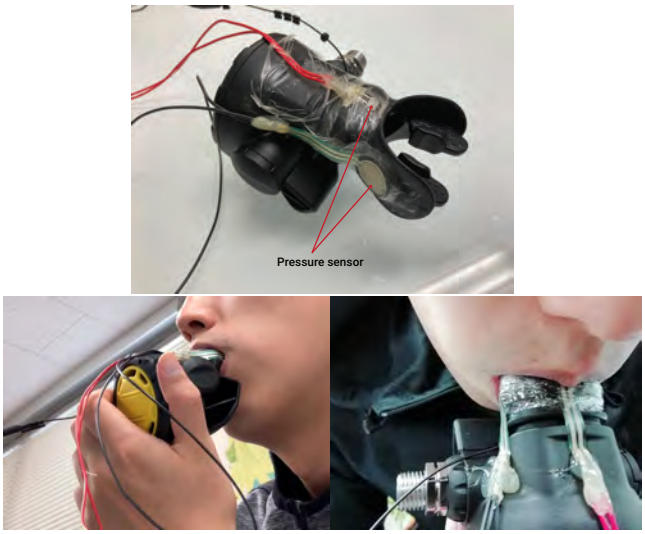

図 6　防水マイクと圧力センサを搭載した提案デバイス

グ用レギュレータにかかる圧力を圧力センサの値から読み取れる．発声に要する口の形が推測できるため，音声認識の精度の向上が期待できる．

4.1　システム構成

唇の動きを撮影することで読唇術を活用する音声認識手法があるが，水中ではダイバーがダイビング用レギュレータを咥えているため，唇の動きをカメラで撮影することは難しい．一方，唇からの力をセンサで測定すれば，唇の動きを推測するための情報を得られる．そこで，マイクロコントローラ（Arduino Nano）に接続された圧力センサ（FSR402：Interlink Electronics）を用いて，唇の動きによって発生するレギュレータにかかる圧力を計測する．

ただし，圧力センサである FSR402 は防水仕様ではない．センサの背面にある黒いパッド部分で力を測定しているため，そのパッドに水が流れ込むとセンサが破損して測定できなくなる．そのため防水性の圧力センサで代用を試みたが FSR402 に比べて感度が低く，唇の動きから発生する圧力は小さいため，圧力をうまく計測できなかった．そこで FSR402 の両面に医療用の防水粘着テープで FSR402 を覆い，FSR402 のパッド部分に水が入らないようにした．

ダイビング用レギュレータの上面と左面に圧力センサを取り付けた．提案デバイスを図 6 に示す．予備実験手として上下左右の 4 面すべてに圧力センサを配置し計測したが，図 7 に示すように，向かい合った 2 つのセンサが同じような値を示すことがわかった．そのため，上面と左面の反対側にあるセンサは取り除いた．

音声データと圧力センサから得られるデータの時刻同期を行うために，図 8 に示すようなインタフェースを Python フレームワークである Django を用いて開発した．また，Arduino のデータを扱うために PyFirmata という Python のライブラリを使用し，GUI の構築には Javascript を活用

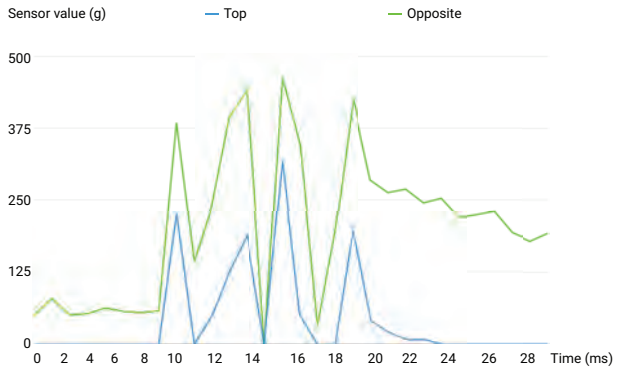

図 7　圧力センサを反対側に配置した場合のセンサ値の例

図 8　音声や圧力センサのデータを時刻同期して記録するためのインタフェース

している．インタフェースでは，Record ボタンをクリックすると，音声データと圧力センサデータを同時に記録できる．録音中は，インタフェースの右下にセンサから取得されるデータの波形が表示される，リアルタイムでセンサの取得エラーを確認できる．記録が完了すると，被験者は「Stop Record」ボタンをクリックするよう求められる．インタフェースの左下に直近の記録ファイル名が表示され，自動的に次のテキストが表示される．

音声と圧力センサ値の混合データを収集するシステムの構成を図 9 に示す．センサデータはタイムスタンプ，上面の圧力センサの値，左面の圧力センサの値の 3 つの列で CSV ファイル形式で保存される．音声データは WAV 形式で保存される．音声データと圧力センサのデータはタイムスタンプを用いて同期され，同期していないデータは削除される．

4.2　実験

4.2.1　実験環境

日本語を母国語とする著者が室内で提案デバイスを装着し，ATR 音素バランス 503 文 [18] を読み上げた．その際，開発したインターフェースを用いて音声データと圧力センサのデータを記録した．

4.2.2　モデル

本章で利用するモデルは，3.2.2 節をベースにしている．

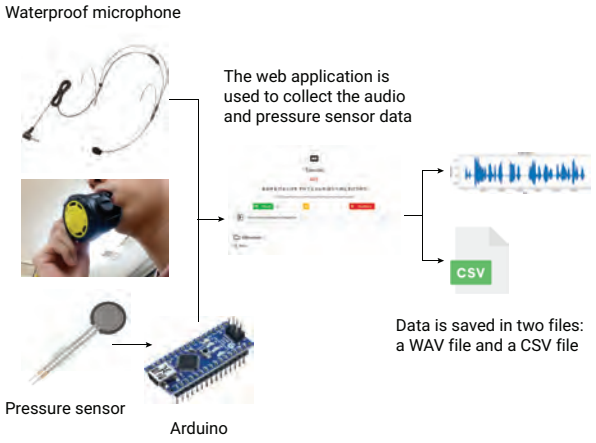

図 9　音声と圧力センサ値の混合データを収集するシステムの構成

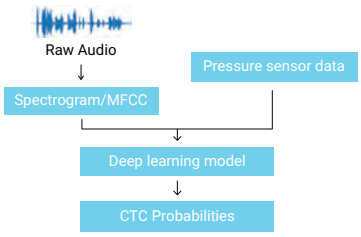

図 10　音声と圧力センサ値の混合データを用いた音声認識手法

スペクトログラムと MFCC の特徴量の間に精度の差はあまりないため，使用する特徴量はスペクトログラムのみとした．3.2.2 節との違いは，音声スペクトログラムの特徴量に圧力センサの特徴量を加えた点である．3.2.2 節では，スペクトログラムの特徴量の次元は 161 次元であったが，オーディオ・ビジュアル音声認識における映像データの特徴量と音声データの特徴量を組み合わせる方法と同様に，圧力センサから得られるデータの特徴量（2 次元）と音声データの特徴量（161 次元）を組み合わせ，163 次元の特徴量ベクトルとし，正規化を適用した．

音声と圧力センサ値の混合データに対する音声認識手法を図 10 に示す．図 5 に示した CNN-BRNN を用いた深層学習モデルを使用し，CTC 損失関数を用いた．モデルの出力は確率ベクトルで，出力の際に音素に変換した．

4.3　結果

音声のみのデータ，音声と圧力センサ値の混合データをそれぞれ活用した深層学習モデルの WER を表 5 に示す．また，モデルによる予測の一部を図 11 に示す．音声のみのデータを用いた場合よりも，音声と圧力センサ値の混合データを用いた場合の方が WER が高かった．追加した特徴量が精度を向上させなかった原因としては，スペクトログラムの特徴量次元が多く，圧力センサの特徴をうまく学習できなかった可能性が挙げられる．一方，表 6 に示すように，音声と圧力センサ値の混合データを用いたモデルの予測は，子音の一部，特に Bilabial 音と Labiodental 音で比較的良好な性能を示した．3 章で述べたように両唇や下唇と上前歯を用いて発音するためにレギュレータに圧力がかかるために精度が向上したと考えられ，これらについては期待した性能向上が確認できた．

表 5　音声のみのデータ，音声と圧力センサ値の混合データにおける WER

Models	Data	Features	WER
CNN-BRNN	Audio-only	Spectrogram	0.24
CNN-BRNN	Audio-pressure-sensor-mixed	Spectrogram	0.27

	Transcription
The result in using audio-pressure-sensor-mixed data	t e ch o: o n i N h a sh e r i: g a k i sh i t a k a s a z c k u e n o y i sh o w a y i a N N o m u m o n o N n o k o k o r o o f u k a t k u y u r i u g o k a s u k t o q k u m i a i w a y u o: g o a t o a m a r a d e ts u z u i t a
The result in using audio-only data	j i t e ch o: n i N y h a sh i h r i: g a k i sh i t o a k a z a k q c k u e n o i sh o w a i N y o m u m o n o n o k o k o r o o f u k a q t k u y u r i y u g o k a s u t o q k u m i a i o w a y u: o g o a t o a m a n o d e ts u z u i t a

図 11　音声のみのデータと音声と圧力センサ値の混合データをそれぞれ用いた深層学習モデルの予測の一部．緑色は正しい音素，赤色はモデルが間違えて予測した音素，無色はモデルが正しく予測した音素を示す．

5.　考察

5.1　圧力センサ

ダイビング用レギュレータの曲面に圧力センサを設置すると，測定中に圧力センサの値が不安定になり得る．特に，ダイビング用レギュレータの側面は柔らかい素材であるため，口の形状が変化すると，圧力センサの値がより不安定になる．そのため，同じ発音でも側面を大きく動かすような発音では圧力センサの値が安定せず，結果としてモデルの学習がうまくいかなくなると考えられる．

また防水性の観点において制限が存在する．圧力センサは水の影響を受けやすいため，医療用の防水粘着テープを使用して圧力センサを覆うことで唾液などの水による影響を取り除いた．しかし，水中でのデータ取得では小さな隙間にも水が入ってくるために，デバイス全体を防水仕様にする必要がある．また，発音によっては口とレギュレータの間に水が侵入し，圧力センサの値に影響を及ぼす可能性がある．実際に水中でデータを取得する際には，デバイスの構造を再設計する必要がある．

2 つの圧力センサのみでは，深層学習モデルの性能を音声のみのデータを活用した場合よりも向上させるには不十

表 6　音声のみのデータ，音声と圧力センサ値の混合データを用いた場合の音素の予測のエラー数．スペクトログラムと MFCC を用いた深層学習モデルの予測のエラー数．テストセットの全音素数は 2158 個である．緑色の背景のセルの方がエラー数が少ない．最下段は表示されている音素以外の全ての音素を指す．

		Phonemes	The number of error prediction	
			Audio-only	Audio-pressure-sensor-mixed
Vowel sounds		a	128	106
		e	52	56
		i	42	44
		o	82	128
		u	84	86
Consonant sounds	Bilabial sounds	b	22	14
		p	6	4
		m	50	38
		w	36	28
	Labiodental sounds	f	4	2
	Others	d	26	36
		h	48	46
		t	28	32
		n	28	28
		N	36	38
		r	40	48
		y	26	22
		...	186	218

分であると考える．1 つの圧力センサでは 1 次元の特徴量しか得られないため，せん断応力といった横方向の力も計測できるセンサを利用し，より多くの次元の特徴量でモデルを学習する必要があると考える．

5.2　音声認識

表 4 と表 6 によると，提案モデルでは母音の認識のエラーが多い．ダイビング用レギュレータを装着している際は，舌を自由に動かすことができず，母音をはっきりと発音することができないために．誤認識を増やしているのではないかと考えられる．

音声のみのデータと，音声と圧力センサ値の混合データの WER の差は小さくほぼ一致しているが，音素ごとの予測結果は異なる．表 5 によると，音声のみのデータの WER は，音声と圧力センサ値の混合データの WER よりも 0.03 低い．しかし表 6 によると，音声と圧力センサ値の混合データでは，唇と下唇の接触で発生する両唇音，上前歯の接触で発生する唇歯音の予測に優れていることがわかる．圧力センサをダイビング用レギュレータの上部と左側に設置したことで，圧力センサの位置が Bilabial 音と Labiodental 音の発声に使用される場所の圧力の変化を捉えられたため，結果が良くなったと考えられる．以上より，子音はそれぞれ発声で使用する場所が異なるため，他の子音の認識率を上げるにはその子音に対応する場所にセンサを設置する必要があるといえる．

6.　まとめ

本論文では，水中での音声コミュニケーションのための音声認識システムを提案した．音声のみのデータと，音声と圧力センサ値の混合データの 2 種類のデータを用いて，CNN-BRNN を活用した深層学習モデルを学習させた．まず音声のみのデータでは，特徴抽出にはスペクトログラムと MFCC という 2 つの特徴量を用いて評価した．評価の結果，スペクトログラムと MFCC を用いたそれぞれの精度の差は比較的小さかった．次に音声と圧力センサ値の混合データを用いた場合では，スペクトログラムと圧力センサのデータを連結させた特徴量をモデルに入力し，評価した．その結果，音声のみのデータで学習したモデルは，音声と圧力センサ値の混合データで学習したモデルよりもわずかながら性能が良かったが，特定の音素では後者のモデルの精度が高かった．

今後の課題としては，より多くの話者から音声と圧力センサ値の両方のデータを取得し，深層学習モデルの汎化性能を確認する．また，今回室内でのデータ取得であったが，水中でのデータ収集も実施する．唇の動きを 3D で追跡するなど，より多彩な機能を実現できるダイビング用レギュレータの開発を目指す．

謝辞

本研究の一部は，JST CREST(JPMJCR18A3) の支援によるものである．ここに記して謝意を表す．

参考文献

[1] B. Woodward and H. Sari: Breathing Noise Elimination in Through-water Speech Communication between Divers, *Journal of the Acoustical Society of America*, Vol. 121, pp. 156–160 (Jan. 2007).

[2] Sequence Modeling with CTC, https://distill.pub/2017/ctc/ (Accessed in 2021/05/10).

[3] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio: Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, *Proc. of the 27th International Conference on Neural Information Processing Systems (NIPS 2014)*, Vol. abs/1412.3555 (Dec. 2014).

[4] W. Gevaert, G. Tsenov, and V. Mladenov: Neural Networks used for Speech Recognition, *Journal of Automatic Control 20 (JAC)*, Vol. 20, pp. 1–7 (Jan. 2010).

[5] T. Afouras, J. Son Chung, A. Senior, O. Vinyals, and A. Zisserman: Deep Audio-Visual Speech Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, Vol. 2, pp. 99–112 (Dec. 2018).

[6] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio: On the Properties of Neural Machine Translation: Encoder-Decoder Approaches, *Proc. of the Eighth*

Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8), pp. 103–111 (Oct. 2014).

[7] L. Bottou: Stochastic Learning, *Advanced Lectures on Machine Learning*, Vol. 3176, pp. 146–168 (Feb. 2003).

[8] M. Schuster and K. Paliwal: Bidirectional Recurrent Neural Networks, *IEEE Transactions on Signal Processing*, Vol. 45, No. 11, pp. 2673–2681 (Nov. 1997).

[9] P. Ladefoged and L. Maddieson: The Sounds of the World's Languages, *Oxford: Blackwell*, ISBN 978-0-631-19815-4 (Feb. 1996).

[10] A. Graves and N. Jaitly: Towards End-to-End Speech Recognition with Recurrent Neural Networks, *Proc. of the 31st International Conference on Machine Learning (PMLR 32)*, pp. 1764–1772 (Jan. 2014).

[11] A. Graves, A. Mohamed, and G. Hinton: Speech Recognition with Deep Recurrent Neural Networks, *Proc. of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, pp. 6645–6649 (Mar. 2013).

[12] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata: Lipreading using Convolutional Neural Network, *Proc. of 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)*, pp. 1149–1153 (Jan. 2014).

[13] M. Wand, J. Koutnik, and J. Schmidhuber: Lipreading with Long Short-Term Memory, *Proc. of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, pp. 6115–6119 (Jan. 2016).

[14] B. Lapatki, D. F. Stegeman, and M. J. Zwarts: Selective Contractions of Individual Facial Muscle Subcomponents Monitored and Trained with High-density Surface EMG, *The facial palsies. Complementary approaches*, pp. 89–109 (Jan. 2005).

[15] D. Toledano, D. Ramos, J. G. Dominguez, and J. G. Rodriguez: Speech Analysis, *Encyclopedia of Biometrics*, pp. 1284–1289 (Sep. 2009).

[16] 暦本 純一, 西村 悠: Derma: 皮膚運動計測によるサイレントスピーチインタラクション, 情報処理学会 インタラクション *2020* 論文集, pp. 11–20 (Mar. 2020).

[17] M. Fukumoto: SilentVoice: Unnoticeable Voice Input by Ingressive Speech, *Proc. of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST 2018)*, pp. 237–246 (Oct. 2018).

[18] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano: ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis, *Journal of Speech Communication*, Vol. 9, No. 4, pp. 357–363 (Aug. 1990).

[19] L. Deng and X. Li: Machine Learning Paradigms for Speech Recognition: An Overview, *IEEE Transactions on Audio, Speech, and Language Processing (ITASL 2013)*, Vol. 21, No. 5, pp. 1060–1089 (May. 2013).

[20] B. Denby, T. Schultz, K. Honda, J. M. Gilbert, and J. S. Brumberg: Silent Speech Interfaces, *Speech Communication 52*, Vol. 52, No. 4, pp. 270–287 (Apr. 2010).

[21] M. A. Hossan, S. Memon, and M. A. Gregory: A Novel Approach for MFCC Feature Extraction, *Proc. of the 4th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pp. 1–5 (Dec. 2010).

[22] Y. Jia, X. Chen, J. Yu, L. Wang, Y. Xu, S. Liu, and Y. Wang: Speaker Recognition based on Characteristic Spectrograms and an Improved Self-organizing Feature Map Neural Network, *Complex and Intelligent Systems*, pp. 1–9 (June 2020).

[23] D. Klakow and J. Peters: Testing the Correlation of Word Error Rate and Perplexity, *Speech Communication*, Vol. 38, pp. 19–28 (Sep. 2020).