# 口腔動作を起点とした音声生成による代用発声技術の実現

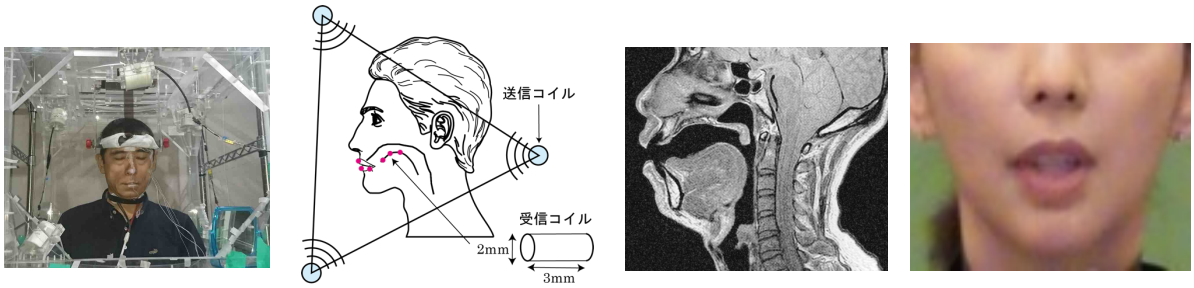
研究代表者　　鏑木時彦　　九州大学大学院芸術工学研究院・教授

## 1　はじめに

重度の喉頭疾患として、喉頭ガンの新規罹患数は毎年約 5,000 人に達する。このような器質的疾患によって喉頭を摘出した場合、その後の一生において声を出すことができなくなる。神経麻痺などの機能性疾患に罹患した場合にも、日常のコミュニケーションに大きな支障をきたし、生活意欲の低下など連鎖的にさまざまな問題を生じる。喉頭摘出者のための代用発声法としては、音源装置を使用する電気式人工喉頭や、食道の粘膜を声帯の代わりに振動させる食道発声などがある。しかしながら、電気式人工喉頭は、ピッチの抑揚のない機械的な発声になるだけでなく、使用者がもともと持っている個人的な声の質が失われてしまう。食道発声は、胃に空気を取り込んで吐き出すため、特に高齢者では習得が難しい。このような社会的問題の解決において、喉頭疾患では、口腔の音声器官（舌、唇、下顎、軟口蓋）は維持されることから、これらの音声器官の発話動作から音声を合成することにより、いわば「口パク」することで、音声による意図の伝達を行う方法が考えられる。超高齢化した社会状況に鑑みても、喉頭疾患によるコミュニケーションの喪失に対処し得る情報技術の創出は不可欠であり、本研究が目指す新しい代用発声技術の着想に至った。

研究代表者はこれまで、長期にわたって人の発話動作の観測に携わってきており、特に代表者が中心となって開発した 3 次元磁気センサ[1,2]は、音声器官に固定した小型コイルの位置を高精度に計測するもので、国内有数のシステムとなっている。磁気センサは、音声を同時に記録できるため、口腔動作と音声の同時測定によって、日本語としては最大規模の発話データベースを構築した[3]。舌、唇、下顎、軟口蓋の動作は、口腔の音響特性を調整し、音声のスペクトル包絡特性と密接な関係がある。そこで従来は、このデータベースを基として、口腔動作とスペクトル包絡特性の間の非線形な写像関係を、パーティクルフィルタや深層学習などによってモデル化する研究を行ってきた[4,5]。一方、自然で明瞭な合成音を得るには、言語の音韻性を表出する口腔の音響特性とともに、音声の音源情報（声の高さ、大きさ、有声・無声の区別など）を正確に推定し、合成音に反映させる必要がある。入力となる口腔の動作は、口腔の音響特性とは直接的な関わりがあり、良好な推定が期待できる。それに対して、声の高さや大きさは肺や声帯の機能によって定まるものであるため、口腔の動作とこれらの音源情報との間には、直接的な因果関係はほとんど存在しない。このように、本研究の本質的な困難さは、口腔の動作からいかに音声の音源情報を推定するかにある。

この問題に対して、本研究では、文節や文章全体にわたる口腔動作の時系列性に着目する。このような長期の口腔動作（舌、唇、下顎、軟口蓋の動かし方）には、通常の発話時の言語的内容に対応した音韻系列の情報が、何らかの形で反映されるはずである。さらに、音韻系列の情報が与えられれば、単語のアクセントを表出する声の高さの変化（ピッチパタン）や有声・無声の区別など、各種の音源情報を推定できる可能性がある。そこで本研究では、人の発話動作の観測を行なって、発話動作と音声とが対になった調音・音声パラレルコーパスを作成し、得られるデータベースと深層学習とを基として、口腔動作から音源情報への写像関係を特定し、目的とする音声合成法の実現へと発展させる。




(1) 磁気センサ　　(2) リアルタイム MRI　　(3) 口唇動画

図 1. 発話動作と音声のパラレルコーパス

発話動作の観測手法としては、先に述べた磁気センサとともに、リアルタイム MRI（磁気共鳴画像）、口唇動画を検討する（図 1）。口腔の音声器官、特に舌や軟口蓋は外部から直接見ることができないため、測定



公益財団法人電気通信普及財団
研究調査助成報告書　第36号　2021年度

が難しく、その方法も限られている。磁気センサ[1,2,3]は、時間分解能が高い点が特徴である。得られる情報は、受信コイルを装着した観測点の位置情報となるため、情報のサイズとしては画像よりもコンパクトである一方、言語音の調音で重要となる舌の後部や軟口蓋に受信コイルを装着することは困難である。リアルタイム MRI は、声道全体の断層イメージを測定するもので、時間分解能は劣るが、磁気センサーでは困難な舌後部や軟口蓋の運動をも得ることができる。口唇動画は、ビデオカメラで容易に測定できる反面、唇の正面画像という最小限の運動情報しか得られず、時間分解能も劣る。このように、口腔動作の測定手段にはそれぞれ特徴があり、音源情報の推定や音声合成における有効性についても未知である。

## 2　複数の観測手法による調音・音声パラレルコーパスの作成

本章では、本研究において機械学習を用いて音声合成システムを構築するために必要となる、調音・音声パラレルコーパスの作成について述べる。磁気センサを用いた調音データの取得に関しては、発話時の調音運動を 3 次元的に観測するシステムの理論設計や、ハードウェア装置の構築を含めて、すでに数年来の実績があったため、本課題においては、リアルタイム MRI と口唇動画による調音・音声パラレルコーパスの作成を中心に研究を実施した。

### 2-1 リアルタイム MRI による調音・音声パラレルコーパスの作成

MRIの基本的な撮像原理としては、まず生体に対して静磁場とラジオ波を加える。このラジオ波の周波数は、原子を構成する原子核の歳差運動と同じにする。この時、原子核はエネルギーを吸収し、平衡状態から励起状態に変わる。ラジオ波を止めると、原子核は吸収したエネルギーを歳差運動と同じ周波数のラジオ波として放出しながら、平衡状態に戻る（核磁気共鳴現象）。このラジオ波を測定することで、生体の断層イメージを得ることができる。無侵襲で安全性が高く、横断像、冠状断像、矢状断像のみならず、任意の断面について断層像を得ることも可能である。通常、MRIを用いた生体観測は静止状態で行われることが多いが、MRIでは、特定のスライス面において、断層イメージを動画の形で連続的に撮像することができる。本研究では、このリアルタイムMRIを用いて、連続音声を発声中の声道の動的観測を行った。

リアルタイムMRIによる測定実験は、京都府にあるATR Promotions社の脳活動イメージングセンタが所有する3テスラMRI装置を有償で使用した。測定日は、2019年10月10日、2019年12月24日、2020年9月18日、2020年10月22日の4回、被験者はいずれも健常者の成人男性1名である。発声リストには、音韻バランスの考慮された文章リスト（ATR503文）を用いた。被験者は、MRI装置のガントリ内で仰向けの状態となり、制御室でのパソコン操作によって提示される文章を目視で確認しながら、1文ずつ発声する。撮像条件は、フレームレートが毎秒27.17フレーム、連続的に撮像できる時間は19秒であり、この時間内で発声可能な文章を順に記録していった。得られるデータはdicom形式のヘッダ付き画像データであり、19秒間の画像枚数は516枚となる。これらの画像データより、プログラミング言語のMATLABを用いて後処理でムービーファイルを作成した。

音声については、MRI測定時に被験者の音声を制御室で録音してムービーファイルに挿入する方法を用いたが、MRI装置から非常に大きな騒音が発生する問題があり、この雑音を音声データから除去することが今後の検討課題として残された。なお、現在のところ、スペクトル減算法などの信号処理に基づく従来法と比べて、機械学習の一種である変分オートエンコーダや非負値行列因子分解を用いて音声信号と雑音信号のそれぞれをモデル化し、両者を分離する方法がより性能が高いという予備的な知見が得られており、今後、客観評価や主観評価を通して有効性を検証する予定である。

リアルタイムMRIによる調音運動測定については、実際の実験を通してわかったことがいろいろとあった。測定を行った脳活動イメージングセンタでも、毎秒27.17フレームでのMRI動画収録については十分な経験がないとのことであり、試行錯誤しながらの測定となった。従来のリアルタイムMRIは、フレームレートが今回の半分程度であり、蓄積される画像データの枚数や、MRI装置本体から制御用の計算機に転送される画像データの量、転送時間もさほど問題とはならなかったが、今回はそれらが非常にシビアとなり、測定実験が進むにつれてMRI画像構築のためのデータ処理とデータ転送のために収録ができない待ち時間が増えていった。その結果、測定実験の途中からは、19秒間の記録のあとに数分の待ち時間ができてしまう状態になってしまった。この長い待ち時間の発生によって、収録できた文章の数は当初の予定を大幅に下回ってしまい、4回の測定実験で1名の被験者に対してようやくATR503文を収録することができた。複数話者のデータベースを作成するには、本課題の終了後も引き続き測定実験を継続していく必要がある。

他方、今回の実験によって、毎秒27.17フレームでのMRI動画収録の有効性も明らかとなった。これまではフレームレートが毎秒14フレーム程度しかなかったため、発声中の舌の素早い動きを十分な時間分解能でとらえることができず、画像がぼけてしまうなどの問題があった。しかし、今回の測定では非常にクリアな動画データを得ることができ、高速な動画収録の有効性を十分に確認することができた。

### 2-2 口唇動画による調音・音声パラレルコーパスの作成

次に、口唇動画による調音・音声パラレルコーパスの作成について述べる。このデータ収集は基本的に簡便であり、被験者に指定の文章リストを発声してもらい、その際の口唇付近の動画をビデオ撮影する。本課題では、動画撮影に必要な機材（ムービーカメラ、背景用スクリーン、照明用ライトなど）をそろえた後に、3 名の被験者についてデータを収集した。1 名はプロの男性ナレーター（有償で雇用）、1 名はプロの女性ナレーター（有償で雇用）、もう 1 名は本課題の研究代表者である。動画の撮影は座位で行い、発声する文章をパソコンから提示した。被験者は提示された文章を 1 文ずつ発声し、ムービーカメラで口唇を含む被験者の顔全体を撮像した。ムービーカメラとは別に、高性能マイクロホンにより音声をリアルタイムでパソコンに収録した。

収録は九州大学大橋キャンパスの無響室で行った。使用機材については、口唇動画はビデオカメラ（Panasonic 社 HC-W580M）を用いて 60fps で撮影した。音声信号データの収録には、ブリューアンドケア製 Type 4191 外部偏極型自由音場マイクロホン（測定範囲 3.15 Hz～40 kHz）、Type2669-L プリアンプ、NEXUS Tpye2690-A-0S2 マイクコンディショニングアンプを使用した。オーディオインターフェイス（Zoom 社 UAC-2 か Focusrite 社 scarlett 6i6）を使用してサンプリング周波数 48 kHz で PC に取り込んだ。頭部の固定は特に行わず、発話中はできるだけ頭部を動かさないように教示した。マイクロフォンと話者の距離は 50 cm、カメラと話者の距離は 170 cm とし、正面の画角から話者の顔全体を録画した。ビデオと音声は別々に収録しており、収録後にビデオカメラで収録された音声とマイクロフォンで収録された音声の間の相互相関を計算することによって同期をとった。発話内容としては、ATR 音素バランス文、JSUT の basic5000 のうち前半の 2500 文に加えて、独自に音素バランス文を作成し、合計で 3887 文からなる発声リストを収録に用いた。各被験者の実質的な発声時間は、5 時間弱になる。以下では、独自に作成した音素バランス文について説明する。

連続音声の特徴である調音結合の問題について考えると、人は舌や唇などの調音器官の位置や形を調整することによって、様々な言語音を作り出して発声を行っている。これらの言語音のうち、言語体系の中で同じ音色として知覚されるものが、音素として分類される。日本語の場合には、ヘボン式ローマ字による表音表記が音素とよく対応している。発話の際には、連続的に調音器官の運動を行うことで、発話意図に沿った形で所望の音素列を生成しているが、質量を持った調音器官の物理的な制約によって、同じ音素であっても前後の音素の調音的な影響や、発話速度、強調などの要因によって、調音運動の様態が変化する調音結合と呼ばれる現象が生じる。調音結合はあくまで調音運動のレベルで生じる現象だが、同じ音素でもこのように調音運動が変化することによって、声道の音響特性や、ひいては音声の特性にもこの調音結合の影響があらわれる。そのため、音声合成や音声認識においては、特定の音素の前後に現れる音素文脈を陽に考慮に入れた、ダイフォンやトライフォンと呼ばれる音素表現を音声単位として用いることが多い。したがって、効率的な音声データの収集には、限られた文章の中に多様な音素文脈が出現することが望ましい。このような考えのもとに作成された発声リストが、音素バランス文である。

本課題では、Minoux の改良貪欲法による文選択を基として、音素バランスを考慮しながら日本語版 Wikipedia の本文データから文章サンプルを選択、収集した。音素バランス文の文候補として Wikipedia を用いたのは、文章量が多く、二次配布が可能であり、文章がある程度整っているからである。文選択の手順としては、まず、個々の文章へ読みの付与を行い、その読みを音素記号列へ変換する。この音素記号列は、文中の音素の並びに従ったトライフォンの列となる。次に、文選択のための目的関数として、文部分集合が多くの音素を含むほど、また音素分布が所望のものに近いほど評価値が高い値をとるように設定する。本研究では、一様分布、すなわち全トライフォンが等価となることを仮定した。さらに、目的関数は劣モジュラ性を持つようにした。これは、文部分集合選択の文脈では、「ある文を小さな文部分集合に加える場合と、同じ文を大きな文部分集合に加える場合とでは、前者の方が効用の増分が大きい」という意味になる。

効用を最大化する文部分集合を厳密に探索する組み合わせ最適化問題は解くのが困難であるため、近似アルゴリズムを用いる必要がある。評価関数が非負かつ単調な劣モジュラ関数のとき、サイズ制約下において、劣モジュラ関数を最大化する問題に対して貪欲法は準最適であることが知られている。そこで貪欲法に

Minoux の改良を導入することで、目的関数の劣モジュラ性を活かして、関数評価の探索回数を削減する。Minoux の改良では priority queue というデータ構造を導入し、ナップザック制約において多様なトライフォンが出現するように、文部分集合を貪欲法によって選択することができる。文選択が終了した後に、音素バランス文に用いるのに不適切な表現や、機械的に付与した音素列が誤っている文を人手で省き、再度効用が大きくなるように文章を追加することを繰り返して、最終的な音素バランス文のセットを完成させた。ATR503 文と本課題で作成した文章セットの比較を、以下の表 1 に示す。

表 1. ATR503 文と本課題で作成した文章セットの比較

| | 文章数 | 総モーラ数 | 各文のモーラ数 | ダイフォンの種類の数 | トライフォンの種類の数 |
|---|---|---|---|---|---|
| ATR503 文 | 503 | 16970 | 10～73 | 503 | 2834 |
| 本課題 | 884 | 18474 | 18～29 | 636 | 4626 |

## 3　磁気センサデータによる調音・音声間変換の構築

磁気センサとは、舌、唇、下顎に発声を阻害しないような小型の受信コイルを装着して、外部の送信コイルから生成される交流磁界を用いることによって、発声中のそれらの位置を連続的に毎秒数 100 フレームのレートでリアルタイム計測するものである。本研究では、我々の研究室で開発した 3 次元測定装置[1,2]により、被験者の舌面に 3 個、上下の唇にそれぞれ 1 個、下顎に 1 個の受信コイルを装着して、それらの調音運動を音声とともに収録した調音・音声パラレルコーパスを作成している。コーパスのサイズは、成人男性話者の約 1 時間分の発声データである。

これまで研究室では、調音データから再帰型のリカレントニューラルネットワークを用い、音声合成に必要な音源パラメータと声道の音響特性を表すスペクトルパラメータを推定して、World ボコーダを用いて音声を合成するシステムを開発してきた[5]。そこで、本課題においては、この研究を発展させる形で、調音運動と音声の相互変換を可能にするために、時間遅延ニューラルネットワークに残差構造を取り入れることで改良型のネットワークを構築し、これによって音声から調音運動を予測する逆変換の検討をおこなった[6]。本章では、この逆推定法に関しての検討結果について述べる。

### 3-1 Residual TDNN による推定法

連続的に発声された音声の長期の時系列性を考慮しつつ、同一話者かつ同一収録条件の限られた発話量の調音・音声パラレルコーパスから、高精度な変換則を構築するための手法として、ここではResidual Time-Delay Neural Network (Residual TDNN) を用いた音声・調音逆マッピングを提案し、さらに音声の音源情報を考慮することによる音声・調音逆マッピングへの効果を検討する。本システムの構成を図2に示す。Residual TDNNは、深く積層されたTDNNによって長期の時系列を考慮しつつも、層数を深くした際に起きる勾配消失問題を回避するように設計されたネットワークであり、少量の学習用データにおいても効果的に逆マッピングを構築できることが期待される。

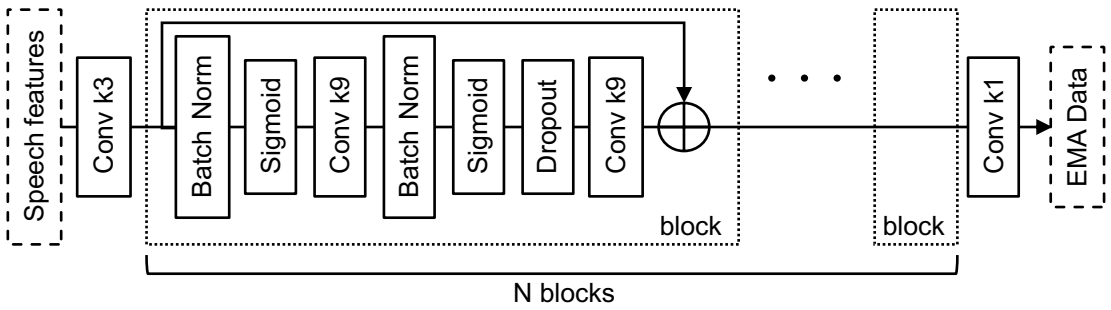



図 2. Residual TDNN の構成図。Conv は畳み込み層であり、数字はフィルタ幅を表す。

TDNNの重み層を積層することで、より長期の時系列を考慮することができる一方、ニューラルネットワークは層を深くすると勾配消失問題によって最適化が困難になる。そこで本研究では、Residual Networkのショートカットを持つ構造をTDNNに導入した、Residual TDNNを検討した。Residual TDNNの概略を以下の図2に示す。Residual TDNNは、長期の時系列を考慮できる深いTDNNに対して、限られたデータ量の調音・音声パラレルコーパスでも、勾配消失の影響を避けつつ逆マッピングの構築が可能となることが期待される。本研究では、pre-activation型のResidual Networkを参考にした。入出力の畳み込み層を除いた中間層は、residual blockを基本単位として構成される。blockの中には2層のシグモイド関数を活性化関数に持つ畳み込み層と、それをスキップするショートカットを持つ。また、residual block内では正則化のためにドロップアウトを加えている。

### 3-2 推定精度の検証

研究室で磁気センサを用いて収集した日本語男性話者1名の調音・音声パラレルコーパスを用いて、ネットワークの学習と評価をおこなった。音声データのサンプリング周波数は16 kHz、磁気センサデータのサンプリング周波数は100 Hzであった。エイリアシングの影響や、調音とは無関係と考えられる高域のノイズ成分を低減するために、磁気センサデータにカットオフ周波数20 Hz、タップ数17のFIRローパスフィルタを適用した。受信コイルはすべて頭部の正中断面上に配置されており、上唇、下唇、下歯、舌3点の計6点の観測点について、3次元直交座標系で表される位置情報のうち、正中断面に対応する2次元の座標位置を用い計12次元のデータを用いた。音声特徴量の抽出については、音声の分析シフト長を磁気センサデータのサンプリングレートとそろえることで、磁気センサデータと音声特徴量が時間的に1対1に対応する。磁気センサデータ、音声特徴量ともに、各次元について0平均単位分散への標準化を行った。

ネットワークのハイパーパラメータであるが、音声特徴量の固定フレーム数は200、残差ネットワークのブロック数は30、畳み込みチャネル数は128、カーネルサイズは9、活性化関数はtanh関数、パラメータの総数は約887万個であった。最適化器にはAdamを使用し、学習率は0.005、学習回数は50000回とした。発声リストとして用いたATR音素バランス文のうち、Iセットを検証用に、Jセットを評価用に用いて、残りを学習データとしたところ、発話時間は学習用61分、検証用3分、評価用2分半となった。音声特徴量として対数メルスペクトログラムを離散コサイン変換によって次元削減した24次のメル周波数ケプストラム係数を用いたところ、磁気センサデータの推定精度は平均で約0.89 mmとなった。ネットワークに入力する音声特徴量としては、これ以外にも、対数振幅スペクトログラム、対数メルスペクトログラム、音声のスペクトル包絡から求めた対数メルスペクトログラムなどを用い、推定精度を比較したが、結果的にメル周波数ケプストラム係数が最も良い推定結果となることがわかった。

## 4　口唇動画からの音声合成システムの構築

最後に、本章では、口唇動画からの音声合成システムについて述べる[7,8]。磁気センサやリアルタイムMRIでは、舌、唇、下顎といった、調音器官の全体的な運動をとらえることができるのに対して、口唇動画で得られる調音情報は、基本的に口唇周辺の動きだけである。特に、口腔の形状を決定する上で最も主要な器官は舌であり、その重要な調音情報を陽に得ることができないことから、口唇動画からどの程度の品質を持った音声を合成できるかは、非常にチャレンジグな問題であると言える。他方、口唇動画はビデオカメラのみで撮影することができ、磁気センサやリアルタイム MRI のような特殊な観測装置を必要としない。したがって、代用発声技術の実現を目指す上で、最も実現性の高い方法であると言える。

### 4-1 畳み込み層を基底とする系列変換モデルによる音声合成システム

今回は、動画の処理に適すると考えられる畳み込み型のネットワークを基として、口唇動画から音声を生成する方法を検討した。音声合成システムの全体的な概要を図 3 に示す。提案法は、口唇動画から特徴量を抽出するエンコーダと、特徴量からメルスペクトログラムを自己回帰推定するデコーダからなる。推定されたメルスペクトログラムより、位相を復元した後に音声を合成する。より詳細には、エンコーダでは、動画処理に有効な時空間残差ネットワークを用いて、時空間両方向の関係を同時に考慮し、動画から特徴を抽出する。時空間残差ネットワークの出力は、空間方向に平均化して時間×次元の 2 次元に次元削減を行い、エンコーダの出力となる特徴量マップを得る。デコーダは、因果的畳み込み層の積層により、エンコーダ出力と前時刻のメルスペクトログラムから次時刻のメルスペクトログラムを推定する。主要な構造は、Gated

linear unit (GLU) ブロックである。この構造は、因果的畳み込みの出力が 2 つに別れており、片方はそのまま出力へ接続し、もう一方はシグモイド関数を適用することでゲートとして機能する。

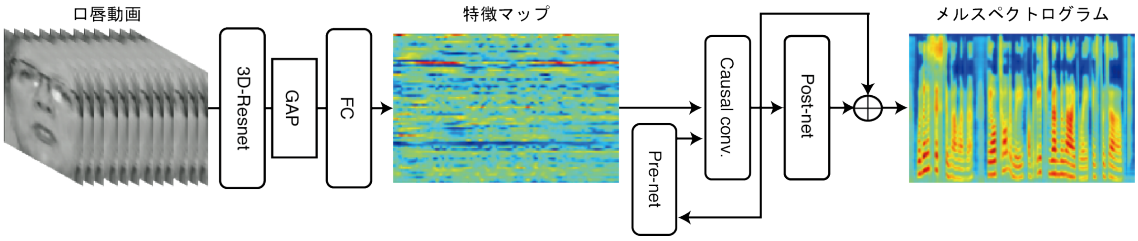



図 3. 口唇動画からの音声合成システムの概要

さらに、残差構造も取り入れることで勾配消失問題に対応し、合成音の高品質化のためにテキスト音声合成において広く用いられているプリネットやポストネットの導入も試みた。前時刻のメルスペクトログラムを処理するプリネットは、2 層の活性化関数 ReLU の全結合層であり、50%のドロップアウトが適用されている。このプリネットは近年の系列変換モデルによるテキスト音声合成において広く用いられており、モデルの汎化に重要な役割を果たす。ポストネットは Residual コネクションを持つ 5 層の非因果的畳み込みニューラルネットワークである。活性化関数は tanh で、バッチ正規化と 50%のドロップアウトが適用されている。自己回帰によって、メルスペクトログラムに生じる誤差の累積の影響を低減する役割を持つ。提案したネットワークを学習するための損失は、ポストネットに入力する前のデコーダーの出力と、正解となるメルスペクトログラムとの平均自乗誤差と、ポストネットの出力と正解となるメルスペクトログラムの平均自乗誤差の和である。

### 3-2 音声合成システムの評価

口唇動画コーパスとしては、2-2節で収集したうちの男女各1名、それぞれ約4.8時間のデータを用いた。読み上げ文のうち、ATR音素バランス分のJセットを評価データとし、残りのデータの95%を学習データ、5%を学習外検証データとした。評価データは約3分の長さであった。口唇動画は60fpsで収録したが、フレームを間引いて50fpsにし、動画の各フレームから顔領域を矩形で抽出した。各フレームで顔領域の中央を中心座標とし、動画の中での最大の矩形を切り出す範囲として、頭部の移動に追従するように口唇動画の切り出しを行った。一方、口唇動画に対応した音声データは、16 kHzにダウンサンプリングを行った。メルフィルタバンクは80次で、70 Hzから8000 Hzの帯域に対して適用した。ここで、音声の分析シフト長を10 msとすることで、口唇動画の各フレームに対して、メルスペクトログラムが2フレームで対応する状態となる。これに合わせて、合成システムのデコーダーは1ループで2フレーム分のメルスペクトログラムを出力するように設定した。

時空間resnetは、いわゆるボトルネック型のresidual blockの積層である。カーネルサイズは3、チャンネル数は128で、5ブロックを積層した。デコーダーはユニット数256のプリネットと6層のGLU blockからなる。GLU blockは、チャンネル数は256、因果的畳み込みのカーネルサイズは5である。GLU block内の畳み込み層には、すべてWeight Normalizationを適用している。正則化として、10%のドロップアウトを適用した。ポストネットのチャンネル数は512とした。エンコーダーの入力は、口唇動画と画素ごとに計算したΔ特徴量、ΔΔ特徴量である。口唇動画には、データ拡張として、明度、彩度、コントラストの摂動、回転、平行移動を学習時にランダムに適用した。口唇動画にはバッチ正規化を適用した。バッチサイズは16で10万イテレーション学習を行い、学習の25%、50%、75%の時点でAdamのアルファパラメータを半減させた。

種々の客観評価指標を用いて合成音を評価したところ、明瞭性に関する客観指標であるSTOIは0.6、ESTOIは0.5、自然性に関する客観指標であるPESQは1.4、音声パワのRMSEは12dB、対数基本周波数のRMSEは0.3、有声/無声判定の誤り率は0.13%という結果が得られた。合成音の了解性や自然性に関する主観評価実験を用いたシステムの評価は今後の課題であるが、本課題で収集した口唇動画データと、提案したネットワーク構造を用いることにより、口唇の調音情報のみを用いた場合でも、十分に了解性のある音声を合成できることがわかった。

## 5　まとめ

本課題においては、喉頭摘出などの理由により発声機能を失うことによる音声コミュニケーションの機能喪失に対処するため、喉頭疾患では口腔の調音器官（舌、唇、下顎、軟口蓋）は維持されることに着目し、これらの調音器官の発話動作（調音運動）から音声を生成する、代用発声技術の開発をおこなった。調音運動は音声の音源情報との直接的な関係がほとんどないため、目的とする音声合成では機械学習の援用によって、調音運動の長期的なパタンより音源情報の再現も含めて音声を合成する必要がある。従って、システム構築のためには、調音運動と音声を同時に記録した調音・音声パラレルコーパスが必要である。

調音運動の観測手法としては、磁気センサ、リアルタイム MRI、口唇動画の 3 種類を用いたが、特に本報告においては、リアルタイム MRI と口唇動画による調音・音声パラレルコーパスの構築について詳細に述べた。このうち、リアルタイム MRI については、本課題の研究期間において男性話者 1 名の 20 分程度のデータ収集しか達成できず、音声合成への応用は今後の課題となってしまった。一方、口唇動画については、3名の話者についてそれぞれ5時間弱のコーパスを構築することができた。この大規模なデータ収集のため、音素バランスに考慮した発声用の文章セットを独自に作成した。

次に、磁気センサで計測した調音運動データを用いた調音・音声間変換について述べた。ここでは、調音運動と音声の相互変換を可能にするために、時間遅延ニューラルネットワークに残差構造を取り入れることで改良型のネットワークを構築し、これによって音声から調音運動を予測する逆変換の検討をおこなった。実験の結果、1 mm 以下の精度で調音運動の推定が可能であることが示された。この結果は、今後、発話訓練やサイレント音声インターフェースなど、広範な応用が可能になると考えられる[9,10]。

最後に、代用発声技術の実現にもっとも適する、口唇動画から音声を合成する方法について検討し、本課題で構築した調音・音声パラレルコーパスを用いた実験結果を示した。口唇動画でも十分に了解可能な音声の生成が可能であることがわかったが、口唇動画には主要な器官である舌の調音情報が反映されないため、合成音の了解性や自然性は、必ずしも十分であるとは言えない。今後は、本課題で収集することができたリアルタイム MRI による調音データを利用し、これを口唇動画と組み合わせることで、エンコーダの出力である特徴量マップの改良をおこない、合成音の品質向上が可能かを検討する予定である。

## 【参考文献】


[1] Tokihiko Kaburagi, Kohei Wakamiya, and Masaaki Honda, Three-dimensional electromagnetic articulography: A measurement principle, J. Acoust. Soc. Am., vol. 118, pp. 428-443 (2005).

[2] Hidetsugu Uchida, Kohei Wakamiya, and Tokihiko Kaburagi, Improvement of measurement accuracy for the three-dimensional electromagnetic articulograph by optimizing the alignment of the transmitter coils, Acoust. Sci. & Tech., vol. 37, pp. 106-114 (2016).

[3] 若宮幸平, 田口史朗, 渡辺莉子, 桂田浩一, 牧野武彦, 鏑木時彦, 大規模日本語調音・音声パラレルデータの収集, 信学技報, SP2019-2 (2019).

[4] 鏑木時彦, 誉田雅彰, 津村尚志, 音素ラベル付き調音・音響対コードブックの検索に基づく調音運動からの音声合成法の検討, 日本音響学会誌, vol. 54, pp. 207-214 (1997).

[5] Fumiaki Taguchi and Tokihiko Kaburagi, Articulatory-to-speech conversion using bi-directional long short-term memory, Proc. Interspeech2018, pp. 2499-2503 (2018).

[6] 田口史朗, 鏑木時彦, Residual Time-Delay Neural Network による音声-調音逆マッピングの検討, 日本音響学会誌, vol.77, pp. 103-111 (2021).

[7] 田口史朗, 鏑木時彦, Full CNN 構造を用いた口唇動画-音声変換, 日本音響学会春季研究発表会, 1-P-18 (2019).

[8] 田口史朗, 鏑木時彦, 口唇動画を用いた調音-音声変換の大語彙連続発話への適用, 日本音響学会秋季講論集, 2-2-4 (2020).

[9] Korin Richmond, 山岸順一, Zhen-Hua Ling, 調音運動の機械学習に基づく応用, 日本音響学会誌, vol. 71, pp. 539-545 (2015).

[10] 藪謙一郎, 障害者が必要としている音声技術, 日本音響学会誌, vol. 74, pp. 136-143 (2018).

## 〈発　表　資　料〉

| 題　　名 | 掲載誌・学会名等 | 発表年月 |
|---|---|---|
| 口唇動画を用いた調音–音声変換の大語彙連続発話への適用 | 日本音響学会秋季講論集 | 2020 年 9 月 |
| Residual Time-Delay Neural Network による音声-調音逆マッピングの検討 | 日本音響学会誌 | 2021 年 2 月 |