九州大学学術情報リポジトリ
Kyushu University Institutional Repository

深層学習を用いたデータ駆動型調音・音声間変換に関する研究

田口，史朗

https://hdl.handle.net/2324/4475138

出版情報：Kyushu University, 2020, 博士（芸術工学）, 課程博士
バージョン：
権利関係：

深層学習を用いたデータ駆動型調音・音声間変換に関する研究

A Study on Data-Driven Conversion between Articulatory Movements and Speech by Using Deep Learning

田口 史朗
TAGUCHI Fumiaki

2021 年 3 月

目次

第 1 章　序論 1
1.1　音声の生成と調音・音声間変換 1
1.2　調音・音声間変換実現へのアプローチ 3
1.3　本論文の目的 5
1.4　本論文の構成 6

第 2 章　音声の分析と再合成 9
2.1　高品質ボコーダによる分析と再合成 9
2.2　線形予測分析 10
2.3　ボコーダフリーな分析と再合成 13
2.4　動的特徴量の導入 16
2.5　まとめ 18

第 3 章　深層学習 19
3.1　教師あり学習と勾配法 19
3.2　多層パーセプトロン 23
3.3　DNN を適切に学習するための工夫 29
3.4　深層学習による時系列モデリング 33
3.5　まとめ 37

第 4 章　日本語調音・音声パラレルデータの収集 39
4.1　発話文の選定 39
4.2　3D-EMA による調音情報の収集 46
4.3　口唇動画による調音情報の収集 51
4.4　まとめ 53

第 5 章　磁気センサによる調音–音声順変換の構築 55

5.1 調音情報からの音声合成法 56
5.2 実験 57
5.3 結果と考察 59
5.4 まとめ 63

第 6 章 Residual Time-Delay Neural Network による音声–調音逆マッピングの構築 65
6.1 Residual TDNN による音声–調音逆マッピング 66
6.2 実験 67
6.3 結果と考察 70
6.4 まとめ 76

第 7 章 口唇動画を用いた調音-音声変換の構築 77
7.1 畳み込み層を基底とした系列変換モデルを用いた口唇動画–音声変換 . . . 78
7.2 口唇動画–音声変換の複数話者モデルの検討 82
7.3 実験 82
7.4 結果と考察 84
7.5 おわりに 92

第 8 章 結論 93
8.1 総括 93
8.2 今後の展望 94

謝辞 99

付録 A 4.1.3 節で選択された音素バランス文 101

付録 B 4.1.4 節で選択された音素バランス文 117

参考文献 135

第 1 章

序論

1.1 音声の生成と調音・音声間変換

音声によるコミュニケーションでは，話し手が伝えたい言葉を発声し，空気の疎密波として伝播した音声が，聞き手の耳に届き，その音声に含まれる情報を聴覚系を通じて受け取ることで成立する．音声の生成に関する器官は調音器官 (Fig. 1.1)，声帯，呼気流の供給源となる肺を総称して音声器官と呼ばれる．音声を生成する際には，発話したい言葉をイメージし，それを実現するための筋司令の運動計画を運動前野にて行い，運動野からその筋司令が音声器官に送られ，音声器官が運動することで音声が生成されると考えられている．このとき，話し手が発した声は空気あるいは骨導によって，自身の聴覚器官に伝えられる．このフィードバックされた自身の音声を当初自身が想定した音声と比較しながら，音声器官の運動を逐次修正している．この一連の音声の生成，聴取の過程は Speech chain として知られている [1].

音声は声帯の自励振動などによって生じた音源波が声道を通過し，口唇から放射されることによって生成される．この音声生成の過程では，舌，口唇，下顎，軟口蓋といった調音器官の運動，すなわち調音運動が深く関与している．調音運動は声道の形状を制御し，その音響特性を決定するだけでなく，声道の狭めや閉鎖を形成することで声道を通過する呼気流を変化させ，摩擦音や破裂音といった子音音源の生成をおこなっている．

この音声生成の過程における調音器官の運動に関する情報を，何らかの方法によって取得し，生成された音声と調音情報の間の関係をモデリングする，調音・音声間変換の研究がこれまで行われてきた．調音情報から音声を推定する調音–音声順変換は，音声の生成過程の工学的模擬となる．また，音声の生成過程を逆向きにたどり，音声から調音情報を推定する音声–調音逆変換は，煩雑な調音観測を介さずに，調音運動の情報を活用できる．さらに，音声生成の聴覚フィードバックにおいて音声から調音運動の修正を行う際には，音声から調音に関する情報を得ることが必要になるはずであり，音声–調音逆変換はこの

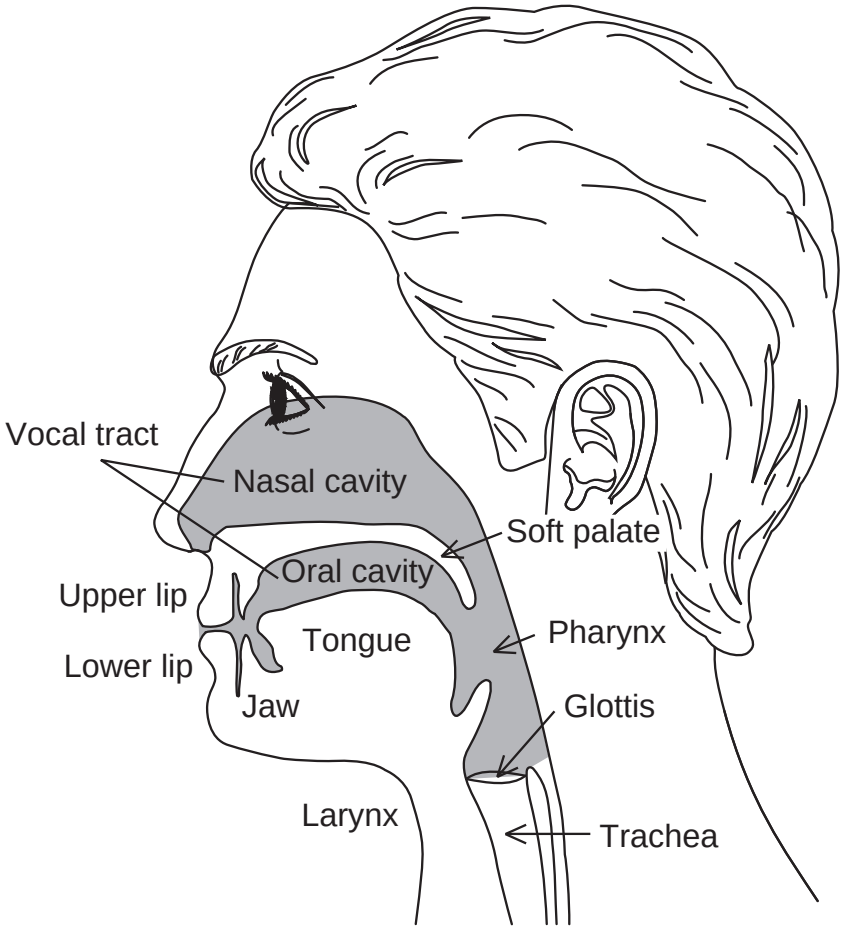


Fig. 1.1: 頭部正中断面と調音器官 [3].

脳の機能の工学的模擬であると言える．

事象の工学的模擬の実現はその事象の解明の一助となるのみならず，種々の応用が可能である．音声の生成過程の工学的模擬が実現されれば，がんなどの要因によって正常に発声が行えなくなった音声障害者に対して，欠落した機能を工学的模擬によって外的に補うことによって，音声コミュニケーションを維持する代用音声への応用が可能となる．また，音声から調音運動を得る脳の機能の工学的模擬が実現されれば，音声から調音方法へのフィードバックを外的に与えることで，非母語音素や聴覚障害者の発話訓練に活用できる．その他にも，声質変換処理，Silent speech interface (SSI) などへの工学的応用に関しても検討され，その有効性が確かめられている [2]．これは，調音情報は有声や無声といった発話時の音源の切り替えの影響を受けず，子音のような非定常音に関しても，声道の調音状態を安定に表現できるためであると考えられる．

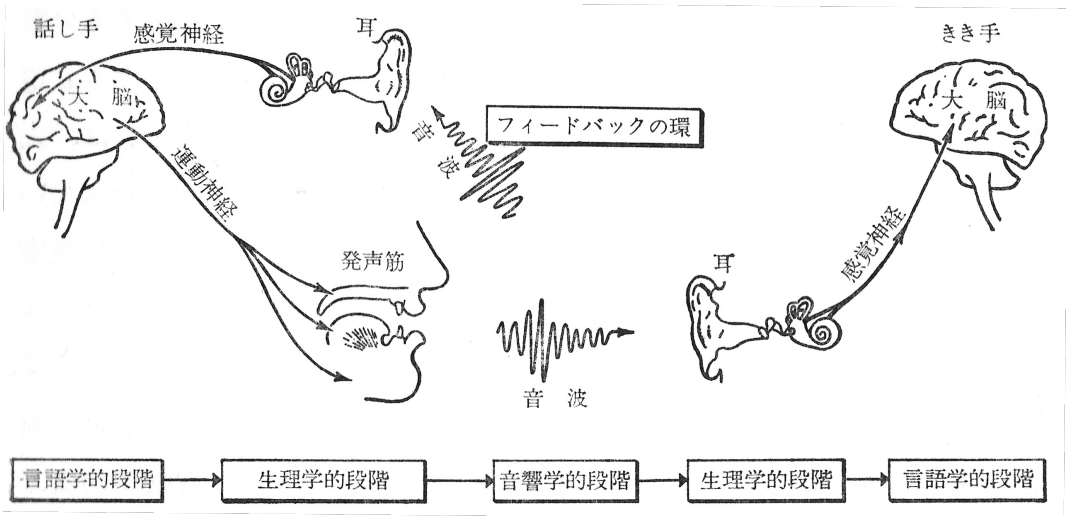


Fig. 1.2: Speech chain．文献 [1] の和訳書 (神山ら，1966) から引用．

1.2　調音・音声間変換実現へのアプローチ

これまでに検討されてきた調音・音声間変換は，音声生成において，声道は音響フィルタに対応すると考えられることから，調音情報から声道のフィルタ特性を得る調音–音響順マッピング，あるいは音声のスペクトル包絡から調音パラメータを推定する音響–調音逆マッピングであった．この調音・音響間マッピングは，物理モデルに基づく手法と，調音情報ならびに音声を同時に収録した調音・音声パラレルコーパスを利用したデータ駆動型アプローチに大別される．

物理モデルに基づく調音–音響順マッピングとしては，文献 [4] に代表されるように，声道を平面波が伝搬すると仮定し，円筒形の音響管の縦続接続として声道を近似しその音響特性を数値的に求めるという方法である．この方法は平面波近似の仮定から，波長が声道の横断長より十分に短い 3.5 kHz までの周波数帯域に関して有効であり，この帯域中に母音の特徴として重要な音声スペクトルのピーク (フォルマント) が概ね 4 つ含まれる．近年では，磁気共鳴画像法 (magnetic resonance imaging; MRI) の発展により，3 次元的に詳細な声道形状を捉えることが可能となったため，有限要素法や時間領域有限差分法などの数値シュミレーションによって，平面波伝搬に加えて高次のモードの発生を含むより広帯域の声道の音響特性を調べることが可能になってきている．周波数領域の音響解析手法は，声道が十分に励起された定常状態を仮定しており，通常の会話音声のように様々な子音と母音が連続して発音され，声道が過渡的かつ時変である状態の分析には適さない．

物理モデルに基づく音響–調音逆マッピングは，この発声の物理的プロセスを反映した

上で，音声の生成過程を逆向きにたどるアプローチを取る．例えば，文献 [5] では声道形状を表すパラメータの変化と声道のフォルマント周波数との関係を表す感度関数 [6] をもとに，声道形状パラメータを所望のフォルマント周波数を持つように決定するという手法が提案されている．この音響–調音逆マッピングは不良設定問題であり，ある音響特性を与えうる声道形状が多数存在する [7] ことが広く知られている．このような逆問題における解の非一意性に対して，例えば文献 [8] において，調音運動の滑らかさを規範として解を特定することが行われているように，声道および調音器官の物理的制約を基に声道パラメータの自由度を下げる必要がある．物理モデルに基づく逆マッピングは，データ駆動型アプローチとは異なり，調音情報の事前収録データを必要とせず，音声のスペクトル包絡が正確に抽出可能な範囲内で，調音情報を推定することが可能である．その一方で，仮定している物理モデルの限界から，高ピッチの音声や声道が音響的に十分に励振されない過渡的な状態に対して適用することは難しいという問題が挙げられる．

それに対して，コーパスを利用したデータ駆動型アプローチでは，あらかじめ用意された調音・音声パラレルコーパスを用いて，音声と調音運動の変換モデルを学習によって構築し，それに基づき音声から調音情報の推定を行う．この方法は，コーパスさえ用意できれば，物理モデルに基づく手法では困難な連続発話にも適用できる長所がある．データ駆動型アプローチの初期の検討としては，調音情報と音声特徴量の対を格納したコードブックを作成する手法 [9, 10] が提案されており，近年では種々の機械学習手法によって実現されている．

調音・音声パラレルコーパスを利用した方法では，発話時の調音情報をいかに正確に得るかが問題となる．これまで，X 線撮像，X 線マイクロビーム [11]，磁気センサ (Electromagnetic articulography; EMA)[12]，超音波エコー [13]，電気的パラトグラフ [14]，磁気共鳴画像 (Magnetic resonance imaging; MRI) [15] など，様々な手法によって調音運動を観測する方法が提案されている．

特に現在広く用いられているのは磁気センサである．これは，複数の小型マーカーコイルを調音器官に取り付けて，それらの位置情報をリアルタイムで高精度に測定することによって，コイルが固定された調音器官の運動を観測する手法である．磁気センサによって収録される調音情報は他の手法と比較して，時間分解能が高く，侵襲性が低く，座位や立位での収録になるため自然な姿勢での発話が可能である．また，得られたコイルの運動データは煩雑な後処理なしでそのまま調音情報として使用でき，動作音もないので雑音の少ない音声を同時に収録可能である．他方，マーカーコイルを舌の後部や軟口蓋に装着することは困難であり，また観測によって得られる情報は装着点の位置情報のみであることから，磁気センサデータから声道全体の形状を復元することは難しいという問題がある．

1.3　本論文の目的

本研究では，深層学習を用いてデータ駆動型調音・音声間変換を構築するために，次の 4 点について検討を行った．

1. 大規模日本語調音・音声パラレルデータの収集
2. 調音・音声間変換における音源情報の活用
3. 深層学習の導入による，調音・音声間変換の精度向上
4. 実応用のための新たな調音情報の検討

深層学習は Deep Neural Network を用いた機械学習である．Deep Neural Network は多数のパラメータを持つため，これらを最適化するためには多量のデータが必要となる．磁気センサによる調音・音声パラレルコーパスに関して，以前から用いられてきた英語のものは 1 話者 30 分以下の発話内容となっている．しかし近年では単一話者に対して測定された 1 時間超のコーパスが公開され，調音・音響間変換の研究に広く用いられてきた．一方，日本語に関しては，磁気センサによる公開された調音・音声パラレルコーパスは存在しない．そこで，本研究では，英語音声のコーパスを参考に，1 時間程度の発話内容を目指して，発話文リストを選定するとともに，実際に磁気センサによる調音・音声パラレルデータの収集を行った．

これまで，調音情報と声道のフィルタ特性の間の関係をマッピングする研究が行われてきたのは，音源と声道フィルタが独立して機能するという，音声生成のモデル化において広く用いられる考え方からであった．一方，実際の音声生成においては，その限りではない．例えば，声道は破裂音や摩擦音を含む一部の子音の音源生成に深く関与している．生理学的には，舌と声帯が位置する喉頭は舌骨を媒介としてつながっており，舌を動かすと喉頭が上下する．喉頭が上下すると声道の長さが変わり，声道の共鳴特性が変化すると同時に，声帯の振動周波数や音声の基本周波数が変化する現象が起こる．さらに，文献 [16] の MRI による声道観察によると，110Hz から 164Hz の通常発話の範囲内において基本周波数を変化させると，声道長だけではなく声道形状が変化することが報告されている．また，文献 [17] では，同じ軟口蓋破裂音に分類される無声音/k/と有声音/g/に関して，舌が口蓋に達する時刻が/k/のほうが早かったことが報告されているなど，声帯振動の有無および，基本周波数と調音様態の関係性を示す観察結果の報告もある．加えて，音声の音韻表現には声道の音響特性が関与しており，音素文脈とピッチの時間的パターンや有声/無声判定などの音源情報は言語的文脈を介して関連している．これらのことから，音声の音源に関する特徴量と調音動作との間には，暗黙的かつ間接的な関係があるのではないかと予想される．本研究では，調音・音響間マッピングを，調音情報と音声の音源に関する

情報の間の関係を考慮し，声道の音響的なフィルタ特性だけではなく，音源を考慮に入れた調音・音声間変換に拡張する．これによって，調音–音声順変換においては，フィルタ特性と音源特徴量の両方が調音情報から推定されるので，調音情報からの音声合成が可能となる．また，音声–調音逆変換においては，音源特徴量が入力に加わることでさらなる推定精度の向上が期待できる．

深層学習は近年様々な問題に対してその有効性が確かめられており，調音・音声間変換においても例外ではない．一方で，深層学習の進展の速度が非常に早く，これまで調音・音声間変換に用いられてきた手法は深層学習の初期の手法にとどまるというのが現状である．また，調音・音声間変換においては，同じコーパスを用いて，先行手法と性能を比較するという有効性検証のためのアプローチが取られることが少なかった．そこで本研究では，深層学習による最新の手法を導入し，調音・音声時系列の特性を陽に考慮した調音・音声間変換法を提案するとともに，共通のコーパスを用いて調音・音声間変換をそれぞれ構築し，先行手法との比較を行うことによって，提案法の有効性を検討した．

調音情報の収録には，多くの場合，大規模で特殊な装置が必要であるので，SSI 等への実応用には不向きである．そこで，本研究では調音情報としての口唇動画に着目する．口唇動画は，口唇周辺の領域をビデオカメラで収録するという非常に簡便な方法で収録が可能である．本研究では，磁気センサでの調音情報の収録と同様に，発話文リストを選定した上で 4.8 時間程度の日本語読み上げ音声を収録した．それを基に新たな口唇動画–音声順変換を提案し，より品質の高い音声を合成するための検討を行った．

1.4　本論文の構成

本論文は本章を含め全 8 章からなる．2 章および 3 章では，本研究で使用する音声の分析合成手法と深層学習について述べる．4 章では，データ駆動型調音・音声間変換を実現するための調音・音声パラレルコーパスの収集について述べる．既存の発話文セットに加えて，限られた文中に多様な音素文脈を含む発話文を新たに構築した上で，3 次元磁気センサシステムおよび口唇動画による調音情報の収集を行った．5 章では，磁気センサによる調音情報から音声を得る調音–音声順変換の検討を行ったことについて述べる．ここでは，新たに双方向再帰型ニューラルネットワークによって，声道のフィルタ特性だけではなく音源情報も同時に推定する調音–音声順変換を提案し，先行研究との比較を行った．6 章では，磁気センサによる調音情報を音声から得る音声–調音逆変換の検討を行ったことについて述べる．本研究では新たに Residual Time-Delay Neural Network による音声–調音逆変換法を提案し，複数の調音・音声パラレルコーパスから，話者依存逆変換モデルを複数構築した上で先行研究との比較を行うとともに，音源情報が推定精度に及ぼす影響を調査した．7 章では，口唇動画による調音情報から音声を得る調音–音声順変換の検討

を行ったことについて述べる．ここでは，畳み込み系列変換モデルを基とした調音–音声順変換を提案し，客観評価によって得られる音声の品質評価を行った．さらに複数話者の調音–音声順変換についても検討を行った．最語に，8 章で本論文の研究内容とその成果を総括する． また、今後の展望についても述べる．

第 2 章

音声の分析と再合成

音声を時間的な波形のまま取り扱うことはその複雑さから困難であり，音声を何らかのパラメータによって表現し，またそのパラメータから音声信号を合成する音声分析合成システムを介することが多い．ただし，近年の深層学習の発展によって，音声波形をそのまま入出力として扱えるようなネットワークが提案されていることには留意しておく必要がある．

本章では，音声の分析，符号化，再合成の手法について，音源フィルタ近似に基づく高品質ボコーダによる手法と，近年の深層学習での音声処理の発展とともに用いられるようになった音源フィルタ近似を用いないボコーダフリーの手法について説明する．また，分析，符号化された音声の特徴量時系列をより良くモデル化するための動的特徴量の導入に関する説明も行う．

2.1 高品質ボコーダによる分析と再合成

2.1.1 音源フィルタ近似

音源フィルタ近似 [18] によれば，音声の生成過程は次のような独立した線形時不変システムの合成として表すことができる．

$$X(\omega) = G_0(\omega)G_e(\omega)H(\omega)R(\omega) \tag{2.1}$$

ここで，$X(\omega)$ は音声，$G_0(\omega)$ は音源のピッチ周期に相当するインパルス列，$G_e(\omega)$ は音源の包絡，$H(\omega)$ は声道，$R(\omega)$ は口唇放射のスペクトル特性を表している．ここで，$G_e(\omega)$ がおよそ -12 db/oct. ，$R(\omega)$ が +6db/oct. の特性を持つことから考えると，音声のスペクトルのうち，スペクトルの包絡は主に声道のスペクトル特性 $H(\omega)$ が寄与し，スペクトルの微細構造は音源の特性 $G_0(\omega)$ が表している．

この近似をもとに，音声信号を分析し音源の特徴を表す量と声道特性 (フィルタ) を表す

量に符号化し，そこから音声信号を合成する一連のシステムのことをボコーダ (Vocoder, voice coder) と呼ぶ.

2.2 線形予測分析

線形予測分析 (Linear Prediction Analysis) は，現時刻の音声サンプルを過去のサンプルから線形的に表すことができると仮定し，その係数となる線形予測係数を求める分析手法である．現時刻の出力 $x[n]$ を p サンプル前までの $x[n]$ で表すと次のように書ける.

$$x[n] = -\sum_{i=1}^{p} a_i x[n-i] + e[n] \tag{2.2}$$

ここで，a_i が線形予測係数，$e[n]$ はこのシステムに対する入力，または (2.2) 式を次のように変形し，右辺第 2 項を予測値ととらえるならば，予測誤差と考えることができる.

$$e[n] = x[n] - (-\sum_{i=1}^{p} a_i x[n-i]) \tag{2.3}$$

この予測誤差ができるだけ小さくなるように線形予測係数 a_i は決定される．予測誤差 $e[n]$ の二乗和 E_p を求めると次のようになる.

$$E_p = \sum_{i=-\infty}^{\infty} e[n] = \sum_{i=-\infty}^{\infty} \{x[n] + a_1 x[n-1] + \cdots + a_p x[n-p]\}^2 \tag{2.4}$$

(2.4) 式から，E_p は a_i の二次式であるから，E_p を最小にする係数は次の条件を満たす.

$$\frac{\partial E_p}{\partial a_i} = 0 \quad (i = 1 \ldots p) \tag{2.5}$$

実際に (2.4) 式を微分すると条件式は次のようになる.

$$\sum_{i=-\infty}^{\infty} x[n-i]x[n] + \sum_{k=1}^{p} a_k \left\{ \sum_{n=-\infty}^{\infty} x[n-i]x[n-k] \right\} = 0 \tag{2.6}$$

ここで，$x[n]$ の自己相関関数 $\phi[\tau]$ は以下の通りである.

$$\phi[\tau] = \sum_{i=-\infty}^{\infty} x[n-i]x[n] + \sum_{k=1}^{p} a_k \left\{ \sum_{n=-\infty}^{\infty} x[n-i]x[n-k] \right\} = 0 \tag{2.7}$$

この $\phi[\tau]$ を用いて (2.7) 式を書き直すと次のようになる.

$$\phi[i] + \sum_{k=1}^{p} a_k \phi[k-i] = 0 \tag{2.8}$$

さらに，この式を連立方程式として書き直すと，

$$\begin{bmatrix} \phi[0] & \phi[1] & \phi[2] & \cdots & \phi[p-1] \\ \phi[1] & \phi[0] & \phi[1] & \cdots & \phi[p-2] \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \phi[p-1] & \phi[p-2] & \phi[p-3] & \cdots & \phi[0] \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} \phi[1] \\ \phi[2] \\ \vdots \\ \phi[p] \end{bmatrix} \tag{2.9}$$

となる．(2.9) 式から，線形予測係数 a_i は信号 $x[n]$ の自己相関関数 $\phi[\tau]$ から求めることができる．(2.2) 式は自己回帰モデルと同形である．つまり，信号 $x[n]$ は $e[n]$ をフィルタ係数 a_k を持つ無限インパルス応答 (IIR) フィルタに入力して得られる出力であると解釈できる．これを音源フィルタ近似と対応させると，$e[n]$ は音源，a_k は声道のフィルタ特性を表すと解釈できる．

線形予測分析は IIR フィルタを仮定したパラメトリック推定であるため，4 kHz 以上の帯域を含む零点があらわれるような音声に対しては有効ではない．よって，音声分析合成においてはノンパラメトリックなスペクトル包絡推定を含む後述の高品質ボコーダが用いられることが多い．

2.2.1　高品質ボコーダ

高品質ボコーダは STRAIGHT [19] に始まる．音声から基本周波数を推定し，さらに音声と推定した基本周波数からスペクトル包絡および各周波数成分における波形全体のパワーに対する非周期成分 (雑音成分) の比を表す非周期性指標を高精度に分析し，得られた基本周波数，スペクトル包絡，非周期性指標から音声を合成する手法である．本研究では World [20] (D4C edition [21]) を用いる．

基本周波数は声の高さに関係するパラメータであり，(2.1) 式の $G_0(\omega)$ のピッチ周期に対応する．音声には声帯振動によって音源が生成される有声音と，声道の狭めを空気が通過することによって生じる雑音を音源とする無声音がある．無声音のときにはこの基本周波数の値は存在しない．基本周波数の推定法としては，様々な倍音を含む音声の基本波成分を低域通過フィルタによって取り出し，この基本波のゼロ交差点から周期を求め，周波数を得るゼロ交差法がある．World の基本周波数分析法 Harvest[22] では複数の低域通過フィルタを用いて，ゼロ交差法を自動化するとともに，従来の時間フレームに独立なアプローチとは異なり，より良い基本周波数軌跡を求めるための種々の工夫が施されている．

スペクトル包絡はフィルタ特性に相当する．ただし，ここでは声道フィルタではなく，(2.1) 式の $G_e(\omega)H(\omega)R(\omega)$ に当たる，音源の特性と口唇の放射特性も含んだものとなっている．World のスペクトル包絡分析法 CheapTrich[23] は，音声をその基本周期の 3 倍のフレーム長をもつ Hanning 窓によって切り出し，周波数分析を行った上でパワースペクトルの平滑化と補償を行うことで，音声を切り出す時刻に対して不変で，基本周波数の

整数倍での値を正確に保持するスペクトル包絡を得ることができるという手法である．

有声音と無声音は 2 値で区別できるものでなく，有声音にも何らかの非周期成分が含まれている．それを表現するために導入されたのが非周期性指標である．非周期性指標は波形全体のパワーに対する非周期成分比を，各周波数，各時刻について，完全に非周期性な場合は 1，完全に周期性な場合は 0 で保持する，時刻と周波数ビンの 2 次元の行列で表される量である．World の非周期性指標分析法 D4C[21] は，基本周期の 4 倍のフレーム長をもつ Blackman 窓によって信号を切り出し，群遅延をもとにした，切り出す時刻に対して不変な特徴量から非周期性指標を計算している．

2.2.2 メルケプストラム

ある適当な位相特性 $\tilde{\omega} = \beta(\omega)$ をもつ因果的なオールパス関数を

$$\tilde{z}^{-1} = \Psi(z) \tag{2.10}$$

ただし，

$$\Psi(e^{j\omega}) = e^{-j\tilde{\omega}} \tag{2.11}$$

として，

$$\tilde{c}(m) = \oint_C \log \tilde{X}(\tilde{z}) \tilde{z}^{m-1} d\tilde{z} \tag{2.12}$$

$$\log \tilde{X}(\tilde{z}) = \sum_{m=-\infty}^{\infty} \tilde{c}(m) \tilde{z}^{-m} \tag{2.13}$$

で，メルケプストラム $\tilde{c}(m)$ を定義する．ここで $X(z) = \tilde{X}(\tilde{z})$ は安定な実係数 $z(n)$ の z 変換，C は単位円を含む $\log X(z)$ の収束領域内で，原点を左回りに 1 周する閉路とする．(2.12)，(2.13) 式は単位円上で，

$$\tilde{c}(m) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \tilde{X}(e^{j\tilde{\omega}^j}) e^{j\tilde{\omega}^m} d\tilde{\omega} \tag{2.14}$$

$$\log \tilde{X}(e^{j\tilde{\omega}}) = \sum_{m=-\infty}^{\infty} \tilde{c}(m) e^{-j\tilde{\omega}m} \tag{2.15}$$

ただし，

$$\tilde{X}(e^{j\tilde{\omega}}) = \tilde{X}(e^{j\beta(\omega)}) = X(e^{j\omega}) \tag{2.16}$$

と書けて，$\tilde{c}(m)$ は対数スペクトル $\log X(e^{j\omega})$ を非直線周波数軸 $\tilde{\omega} = \beta(\omega)$ に周波数変換したときのフーリエ変換となる．ここで，オールパス関数を

$$\tilde{z}^{-1} = \Psi(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \tag{2.17}$$

$$|\alpha| < 1$$

とすれば，$\tilde{z}^{-1} = e^{-j\tilde{\omega}}$ の位相特性は，

$$\tilde{\omega} = \beta(\omega) = \tan^{-1} \frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha} \tag{2.18}$$

で与えられる．この α が 0 の場合，メルケプストラムは単純なケプストラムと等価になる．また，α をサンプリング周波数に対して適切な値に設定することで非線形周波数軸がメル尺度やバーク尺度と一致する．メルケプストラムは低い周波数域では細かく，高い周波数域では荒い分解能を持つ人間の聴覚特性を反映したケプストラムということができ，通常のケプストラムの半分程度の次数で音声スペクトルを表現することが可能である．

2.2.3　高品質ボコーダを用いた音声特徴量の抽出，再合成

ここでは，パラメトリック音声合成によく用いられる，音声から特徴量を抽出する方法および再合成方法を説明する．まず音声に対して World を適用し，基本周波数，スペクトル包絡，非周期性指標を得る．次に，スペクトル包絡とメルケプストラムによって表現されるスペクトルが一致するように最適化を行うことで，メルケプストラムを計算する．

非周期性指標は周波数ビン × 時刻の 2 次元特徴量であるが，5 帯域に圧縮しても合成音声の品質がほぼ変化しないことから，5 帯域平均非周期性指標 [24, 25] が用いられることが多い．また，World においては，0 次の非周期性指標が 0.5 より小さい場合が有声音，大きい場合が無声音というように非周期性指標から有声/無声判定を得ることができる．

音声の基本周波数は無声音の場合値が存在しないこととなる．多くの場合は，無声音のときの基本周波数は 0 として表されるが，この基本周波数の断続的な変化は，推定には不利となるので，無声区間に線形補間等で適当な値を入れ，基本周波数とは別に有声/無声判定を推定するという方法 [26] が取られる．さらに，F0 の対数をとった連続対数 F0 を用いる事が多い．

Fig. 2.1 に World を用いて抽出された特徴量の 1 例を示す．

2.3　ボコーダフリーな分析と再合成

2.3.1　短時間フーリエ変換

短時間フーリエ変換 (short-time Fourier transform; STFT) は窓関数と呼ばれる特殊な信号を用いて，もとの信号から短時間の区間を切り出し，切り出した信号に対して離散フーリエ変換 (discrete fourier transform; DFT) を適用するという処理を，信号を切り出す区間を時間方向に少しずつ移動させながら行う演算である．窓関数の長さをフレーム長，窓を移動させるサンプル数をシフト長と呼ぶ．ここで，信号 $x[n]$ を STFT して得られた結果を $X(j,k)$ と表すこととする．X は複素行列であり，j が時刻，k が周波数ビン

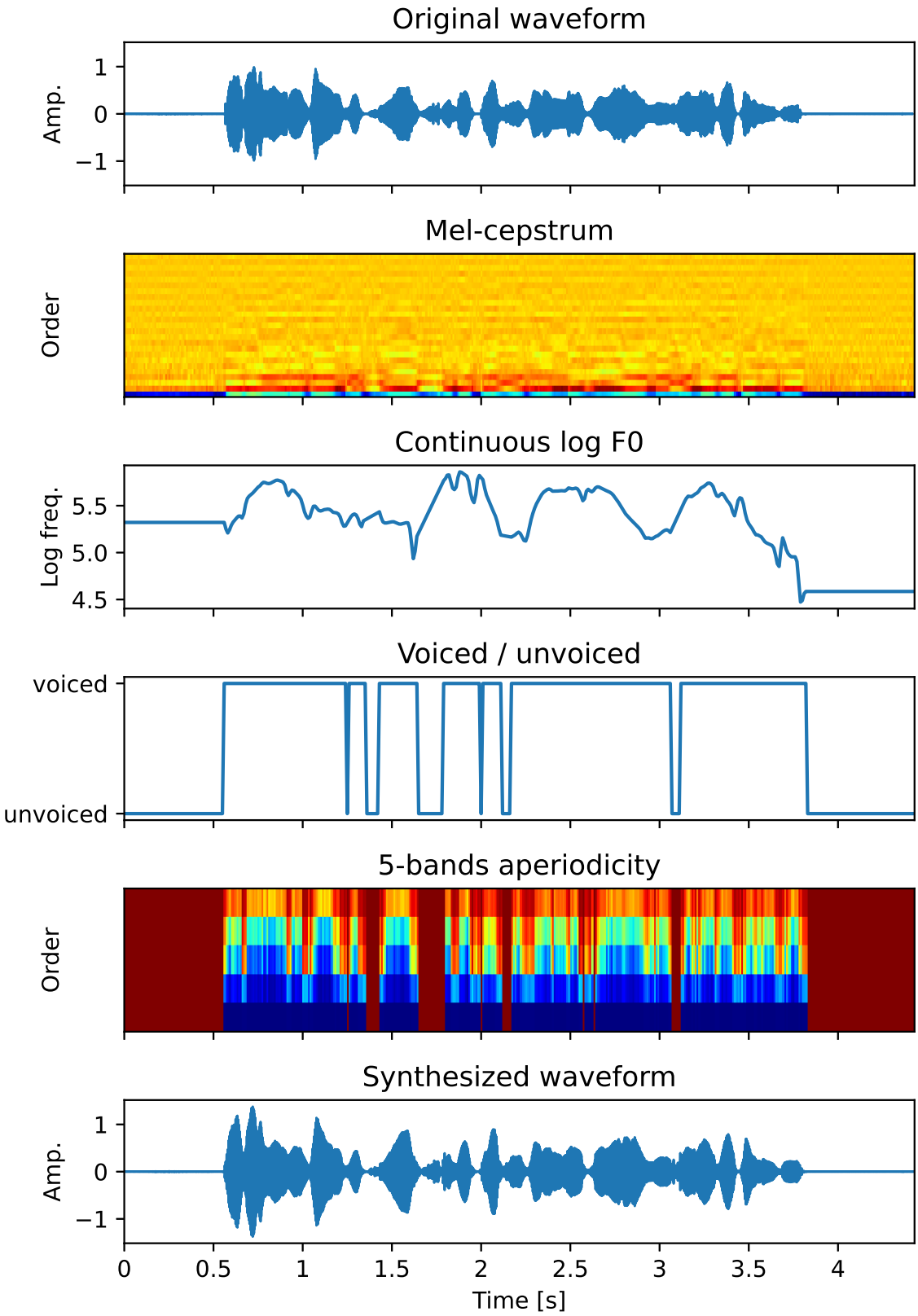


Fig. 2.1: World を用いて抽出された特徴量の例.

のインデックスにそれぞれ対応する．この X のことを複素スペクトログラムと呼び，周波数成分の時間変化を観察することができる．$|X|^2$ をパワースペクトログラムあるいは単にスペクトログラム，$|X|$ を振幅スペクトログラム，$\angle X$ を位相スペクトログラムと呼ぶ．

X から信号 $x[n]$ を得る逆 STFT を行う場合は上記と逆の操作，つまり，$X(j)$ に対して逆 DFT を行い，得られた短区間の波形を少しずつ移動させながら加算していけば良い．この演算をオーバーラップ加算と呼ぶ．ただし，完全な再合成を行うためには窓関数をオーバーラップ加算した結果が一定になる必要がある．これを満たす条件は多岐にわたるが，hanning 窓であればフレーム長がシフト長の 2 以上の整数倍であれば良い．以降特に記述がなければフレーム長としてシフト長 ×4 を用いる．

2.3.2　メルスペクトログラム

メル尺度 [27] は聴覚心理的実験から得られた人間の聴覚特性を考慮した音高の心理尺度で，1000 Hz の正弦波から知覚される音高を 1000 mel とする比例尺度である．実装上は次の式で表されることが多く，後述のメル周波数ケプストラム係数のような，人間の聴覚に関する特性を反映した符号化や情報圧縮に広く用いられている．

$$\mathrm{Mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.19}$$

メルスペクトログラムは振幅スペクトルにメルフィルタバンクを適用することで得られる．メルフィルタバンクは任意の周波数帯域で交互に重なり合う，メル尺度上で同じ幅の最大振幅 1 の三角フィルタの集合である．フィルタの数は任意である．メルフィルタバンクは (周波数ビン数)×(フィルタバンク数) の行列として表され，メルスペクトログラムはメルフィルタバンクと $|X(j)|$ の積を取ることによって得られる．更に対数をとった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任意の次元数まで圧縮することで，音声の特徴量として広く用いられる，メル周波数ケプストラム係数 (Mel Frequency Cepstral Coefficient; MFCC) が得られる．

2.3.3　ボコーダフリーな音声特徴量の抽出，再合成

音声に STFT とメルフィルタバンクを適用することで，メルスペクトログラムを得る．この時点で位相の情報が欠落しており，メルフィルタバンクも不可逆な変換である．メルスペクトログラムから振幅スペクトルへの復元は非負拘束最小二乗法 (NNLS)[28] によって得られる．また，位相は STFT と逆 STFT を振幅を付け替えながら反復する Griffin–Lim(GL) アルゴリズム [29] で復元することができる．しかしこうして得られた

音声はもとの波形を完全に復元することはできない．Fig. 2.2 に NNLS によってメル尺度から線形尺度に復元された振幅スペクトログラムを示す．概形は保たれているものの，主に高域成分において微細構造が復元できていないことがわかる．

メルスペクトログラムが音声を推定する際の特徴量として用いられるようになったのは，Deep Neural Network によって音声波形を得るニューラルボコーダが提案されたからである．代表的なものとして，Wavenet Vocoder[30] がある．欠落した情報の復元をニューラルボコーダによって行うことで，メルスペクトログラムから NNLS と GL 法を用いるよりも高品質な音声波形が得られる．ニューラルボコーダは高品質ボコーダによって符号化した特徴量にも有効であり，ソースフィルタ近似によって欠落した情報を復元した上で音声波形を合成することができる．

2.4 動的特徴量の導入

2.4.1 動的特徴量

特徴量の時間方向の動的変化を表す特徴量を動的特徴量と呼ぶ．特徴量の時間に関する導関数が用いられる場合が多い．より実際的には中心差分が用いられる．1 次の導関数を Δ 特徴量，2 次の導関数を $\Delta\Delta$ 特徴量と呼ぶ場合もある．また，もとの特徴量のことを動的特徴量に対して静的特徴量と呼ぶ．

入力する時系列に動的特徴量を加えることで，特徴量の時間方向の動的変化を直接的に考慮することが可能となる．また，動的特徴量を出力される特徴量の制約条件として用いることで，歪のないなめらかな出力が得られる．

2.4.2 最尤パラメータ生成

最尤パラメータ生成 (Maximum-likelihood parameter generation; MLPG) [31] は，静的特徴量と動的特徴量から，動的特徴を考慮した静的特徴量を生成する手法である．

モデルから出力された特徴量を $\boldsymbol{o}_t$ とする．$\boldsymbol{o}_t$ は次式のように静的特徴量 $\boldsymbol{c}_t$ とその動的特徴量 $\Delta\boldsymbol{c}_t$ から構成されているとする．

$$\boldsymbol{o}_t = [\boldsymbol{c}_t^{\mathrm{T}}, \Delta\boldsymbol{c}_t^{\mathrm{T}}]^{\mathrm{T}} \tag{2.20}$$

ここで，$\boldsymbol{o}_t$ と $\boldsymbol{c}_t$ の時系列をそれぞれ $\mathrm{O} = [\boldsymbol{o}_1^{\mathrm{T}}, \boldsymbol{o}_2^{\mathrm{T}}, \cdots, \boldsymbol{o}_T^{\mathrm{T}}]^{\mathrm{T}}$，$\mathrm{C} = [\boldsymbol{c}_1^{\mathrm{T}}, \boldsymbol{c}_2^{\mathrm{T}}, \cdots, \boldsymbol{c}_T^{\mathrm{T}}]^{\mathrm{T}}$ とすると，O は C から次のように計算できる．

$$\mathrm{O} = \mathrm{WC} \tag{2.21}$$

ここで，W は静的特徴量系列 C から静的・動的特徴量系列 O を計算するために使用され

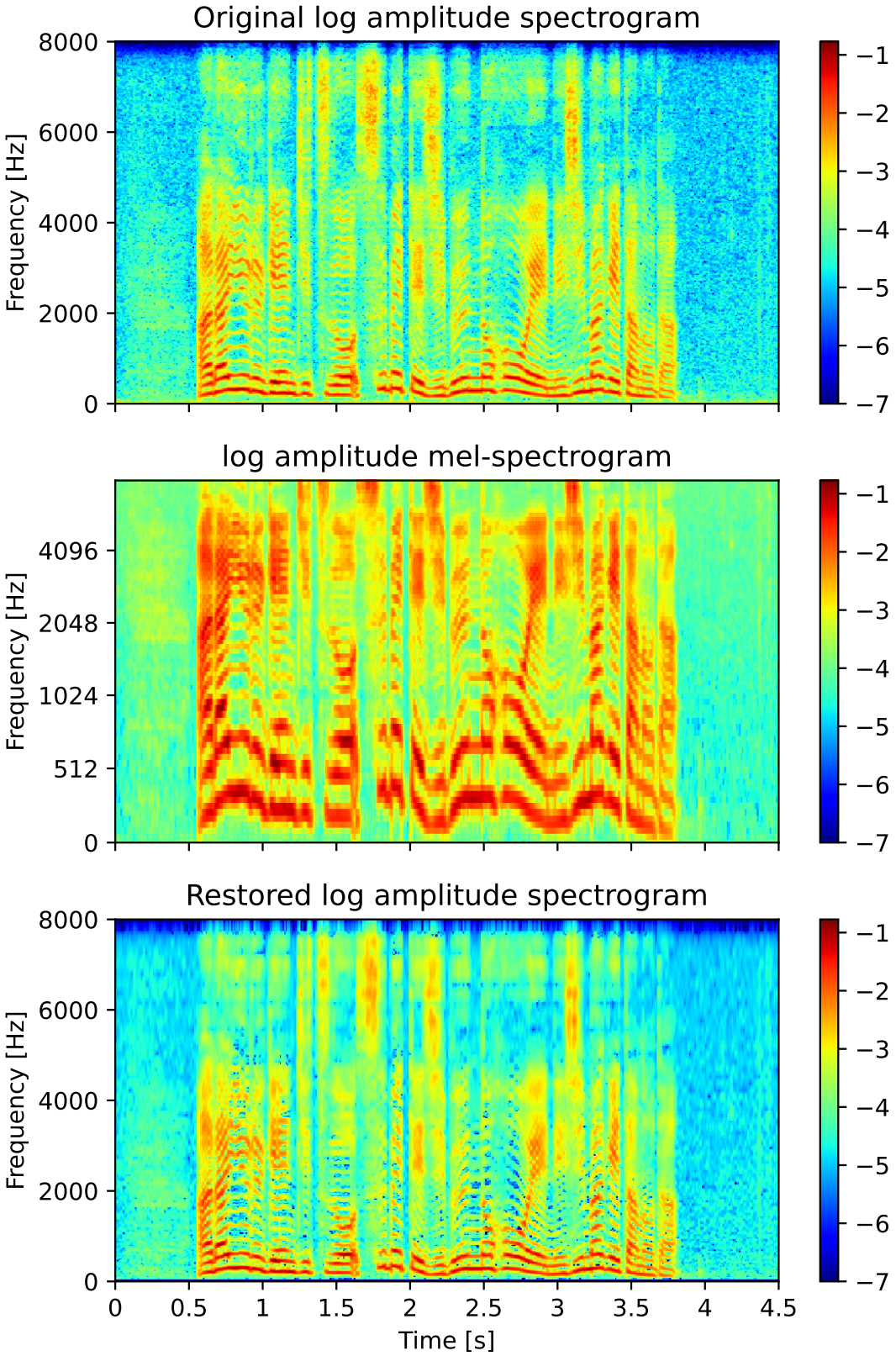


Fig. 2.2: メルスペクトログラムおよび NNLS によって復元された振幅スペクトログラムの例.

る係数を含む行列である [31]．そして，静的・動的特徴量系列 O から，動的特徴量を考慮した静的特徴量系列 $\hat{\mathrm{C}}$ は次のように計算できる．

$$\hat{\mathrm{C}} = \mathrm{RO} \tag{2.22}$$

$$\mathrm{R} = \left(\mathrm{W}^{\mathrm{T}}\Sigma^{-1}\mathrm{W}\right)^{-1}\mathrm{W}^{\mathrm{T}}\Sigma^{-1} \tag{2.23}$$

ここで，$\Sigma = \mathrm{diag}[\Sigma_1, \Sigma_2, \cdots, \Sigma_T]$ は静的・動的特徴量系列 O の共分散行列である．Σ_t は時刻 t における特徴量の共分散行列であり，Σ は学習データから計算可能である．

2.4.3　MGE 学習

MGE 学習 [32, 33] は MLPG によって生成した特徴量と静的特徴量の正解値の誤差を最小化する手法である．深層学習の枠組みで考えると，生成した特徴量と静的特徴量の正解値の誤差は従来よく用いられる平均自乗誤差等でよく，ネットワークから出力される静的・動的特徴量系列から静的特徴量を求めるための式 (2.22) の乗算が加わるだけであるため，誤差逆伝搬法の枠組みで勾配を求めることができる．式 (2.23) の R は密行列であることから，MGE 学習によって計算されたあるフレームの誤差は全フレームのネットワークの出力に影響する．そのため，たとえネットワークそのものが時系列を考慮していなくても，MGE 学習を導入することで時系列を考慮した学習が可能になる．

2.5　まとめ

ここでは，音声の分析，符号化，再合成の手法について，高品質ボコーダ World による手法と，メルスペクトログラムを用いたボコーダフリーの手法について説明を行い，分析，符号化された音声の特徴量時系列をより良く表現するための動的特徴量と最尤パラメータ生成，MGE 学習に関する説明を行った．

第 3 章

深層学習

深層学習は Deep Neural Network (DNN) が持つパラメータを損失が最小となるように勾配法によって最適化する手法である．最適化のことを学習と呼ぶ場合もある．また学習したネットワークによって値を求めることを推論と呼ぶ．本章では，多岐にわたる深層学習に関する話題のうち，本研究の前提となる教師あり学習と勾配法による最適化，多層パーセプトロンとそれへの誤差逆伝播法の適用，DNN を適切に学習するための種々の工夫，時系列を取り扱うためのネットワーク構造についての説明を行う．

3.1 教師あり学習と勾配法

3.1.1 教師あり学習

教師あり学習は，あるデータに対してそれに対応するラベルや正解値 (目標値) が存在するデータ対をもとに最適化を行う手法である．画像分類や音声認識，本研究の主題となる調音・音声間変換もこの教師あり学習のアプローチが取られる．画像分類では画像とその分類ラベル，音声認識であれば音声とテキスト，調音・音声間変換であれば調音情報と音声のデータ対を用いて入力から所望の出力が得られるようにモデルの最適化を行うことで，画像分類器や音素認識器，調音・音声間変換モデルが実現される．訓練データを学習させることで，学習データに含まれないデータに対しても推定が可能になる．これを汎化という．一方，学習データに対しては推定精度が良好なものの，学習外データの推定精度が悪い場合を過学習と呼ぶ．

ネットワークの学習の際に使用するデータは，学習 (training) データの他に，学習外検証 (validation) データ，評価 (test) データにわけられる．これらの 3 つは互いに独立であり，同じデータは含まれない．ネットワーク学習中に定期的に学習外検証データの推定精度を計算することで，学習の終了判定や学習に必要なハイパーパラメータの調節に用い

る．評価データは学習後のネットワークの最終的な推定精度を判定するために使用する．学習外検証データと学習データをそれぞれ用意しなければならないのは，評価データから学習の終了判定や学習に必要なハイパーパラメータの調節を行うと，ネットワークが評価データに対して過剰に適合してしまい，汎化性能を検証できなくなるためである．

3.1.2　損失

損失は推定の悪さを表すスカラー量であり，この損失を最小とするモデルを最適解とする．損失を計算するための関数を損失関数と呼ぶ．学習段階において入力から正解値を推定する回帰問題においては，出力と正解値との平均二乗誤差が用いられることが多い．

$$L = \frac{1}{n}\sum_{i=1}^{n}(y_i - t_i)^2 \tag{3.1}$$

ここで，y_i は推定値ベクトル $\boldsymbol{y}$ の i 番目の要素，t_i は正解値ベクトル $\boldsymbol{t}$ の i 番目の要素である．

また，ラベルを推定する分類問題においては交差エントロピーが用いられることが多い．

$$L = -\sum_{i=1}^{n} t_i \log(y_i) \tag{3.2}$$

ここで，y_i は推定値ベクトル $\boldsymbol{y}$ の i 番目の要素，t_i は正解ラベルベクトル $\boldsymbol{t}$ の i 番目の要素である．$\boldsymbol{y}$ と $\boldsymbol{t}$ は正かつ和が 1 である必要があるが，$\boldsymbol{y}$ に関しては後述の Softmax 関数を適用することでこれを実現している．また，$\boldsymbol{t}$ は当該要素のみ 1 でそれ以外は 0 である one-hot ベクトルであることが多い．

3.1.3　勾配法

深層学習では損失を最小化するための最適化手法として勾配法を用いる．DNN のパラメータを θ とおくと，DNN によって推定した値と正解値から計算した損失 L は θ の関数 $L(\theta)$ となる．このとき，学習率と呼ばれるハイパーパラメータを η とおくと，ステップ k におけるパラメータは損失関数の偏微分 $\partial L(\theta)/\partial\theta$ を用いて (3.3) 式のように更新される．

$$\theta_{k+1} = \theta_k - \eta\frac{\partial L(\theta_k)}{\partial\theta_k} \tag{3.3}$$

このときパラメータ θ の初期値 θ_0 は適切な乱数によって初期化されている点に注意が必要である．また，勾配法の概念図を Fig. 3.1 に示す．損失関数 $L(\theta)$ は θ がある値のときに最小値を取るという凸性を仮定すると，θ を勾配 $\partial L(\theta)/\partial\theta$ と逆向き (最急降下方向) に更新することで，$L(\theta)$ の最小値を与える θ に近づいていく．

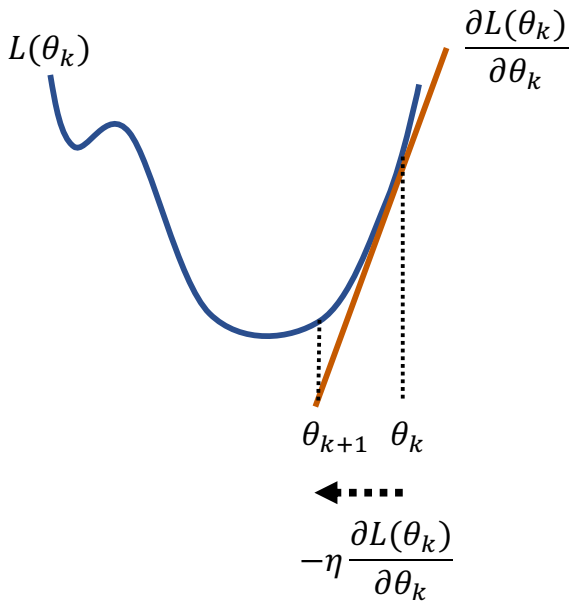


Fig. 3.1: 勾配法の概念図

深層学習においては，学習に使用する学習データから無作為に一部を抽出し勾配法を実行する確率的勾配降下法 (stochastic gradient descent; SGD) を用いる．この時，学習データから複数のデータを抽出して用いるミニバッチ学習が広く用いられている．すべての学習データから勾配法を用いるバッチ学習の場合はデータの規模によって計算時間が膨大になり，また，最急降下方向は局所的な性質であることから，勾配を正確に求めることよりもパラメータの更新回数を増やすことのほうが，早く良い解にたどり着ける可能性がある．一方，学習データから 1 つのデータを抽出して勾配法を行うオンライン学習は，勾配の分散が大きくなることで，解への収束が遅くなることが知られている．よって，バッチ学習とオンライン学習の折衷案として，ミニバッチ学習が用いられる．全学習データを 1 度学習させることを 1 epoch と呼ぶ．全学習データから取り出してくるデータの数 (ミニバッチ数) によって，同じ 1 epoch でもパラメータが更新される回数が異なるので注意が必要である．

学習率 η は，一度にパラメータをどの程度更新するかを決定するハイパーパラメータである．η が大きすぎると解が収束しない．逆に小さすぎると収束まで時間がかかり，局所解に陥る可能性もある．そのため，学習率 η を適切な値に調節することは極めて重要である．近年では，学習中に学習率を変更 (主に減少させる) ことも広く行われている．

(3.3) 式で示した勾配法では，全てのパラメータに対して共通の学習率を用いている．これには困難が伴うため，確率的勾配法の様々な拡張が考案されている．ここでは，本論文で用いる 2 つの改良法について述べる．一つは，パラメータの更新に慣性項を導入する

モーメンタム法であり，次式で表される．

$$\theta_{k+1} = \theta_k - \eta \frac{\partial L(\theta_k)}{\partial \theta_k} + \beta \Delta \theta_k \tag{3.4}$$

ここで，$\Delta\theta_k$ は前時刻のパラメータの更新量，β は慣性項を制御するハイパーパラメータである．慣性項を導入することで，パラメータの更新の際に起こる振動を抑制し，収束が早まることが期待できる．もう一つはパラメータごとに適応的に学習率を変化させる方法である．RMSprop[34] は勾配を直近の勾配の移動平均で正規化する．以下に RMSprop によるパラメータの更新式を示す．

$$m_k = \aleph m_{k-1} + (1 - \aleph) g_k \tag{3.5}$$

$$v_k = \aleph v_{k-1} + (1 - \aleph) g_k^2 \tag{3.6}$$

$$\Delta_k = \beth \Delta_{k-1} - \gimel \frac{g_k}{\sqrt{v_k - m_k^2 + \epsilon}} \tag{3.7}$$

$$\theta_k = \theta_{k-1} + \Delta_k \tag{3.8}$$

ここで，$g_k = \partial L(\theta_k)/\partial\theta_k$，$\aleph$，$\beth$，$\gimel$，$\epsilon$ はハイパーパラメータである．ハイパーパラメータに関しては，$\aleph = 0.95$，$\beth = 0.9$，$\gimel = 0.0001$，$\epsilon = 0.0001$ を用いることが推奨されている．RMSprop は勾配の自乗の累積に移動平均を用いることで，必要以上の過去の履歴を取り除き，学習率が小さくなりすぎてしまうことを防いでいる．

これらの改良を組み合わせたような Adam[35] も提案されており広く用いられている．Adam によるパラメータの更新式は以下の通りである．

$$m_k = \beta_1 \cdot m_{k-1} + (1 - \beta_1) \cdot g_k \tag{3.9}$$

$$v_k = \beta_2 \cdot v_{k-1} + (1 - \beta_2) \cdot g_k^2 \tag{3.10}$$

$$\hat{m}_k = \frac{m_k}{1 - \beta_1^t} \tag{3.11}$$

$$\hat{v}_k = \frac{v_k}{1 - \beta_2^t} \tag{3.12}$$

$$\theta_k = \theta_{k-1} - \frac{\alpha \cdot \hat{m}_k}{\sqrt{\hat{v}_k} + \epsilon} \tag{3.13}$$

ここで，$g_k = \partial L(\theta_k)/\partial\theta_k$，$\alpha$，$\beta_1$，$\beta_2$，$\epsilon$ はハイパーパラメータである．ハイパーパラメータに関しては，$\alpha = 0.01$，$\beta_1 = 0.9$，$\beta_2 = 0.999$，$\epsilon = 10^{-8}$ を用いることが推奨されている．RMSprop と Adam は 1 次と 2 次のモーメントを用いるという点で非常に似ているということがわかる．

ここで紹介した以外にも最適化器は数多く提案されているが，どれも 1 長 1 短があり，ハイパーパラメータの 1 つとしてタスクやネットワーク構造によって調整しているのが現状である．

3.2　多層パーセプトロン

多層パーセプトロン (Multilayer perceptron; MLP) は Feed-forward Neural Network (FFNN) とも呼ばれ，典型的には全結合層と活性化関数の積層で構築される．入力 $\boldsymbol{x}$ を受け取る全結合層のことを入力層，出力の直前の全結合層のことを出力層，それ以外を隠れ層と呼ぶことが多い．ここでは，全結合層と活性化関数，および MLP への勾配法の適用について述べる．

3.2.1　全結合層

全結合層 (Fully connected layer) は MLP の構成要素の一つであり，(3.14) 式で表される．

$$\boldsymbol{o} = \mathrm{W}\boldsymbol{x} + \boldsymbol{b} \tag{3.14}$$

ここで，$\boldsymbol{x}$ は全結合層への入力となる列ベクトル，$\boldsymbol{o}$ は全結合層の出力である列ベクトルであり，W は重み行列，$\boldsymbol{b}$ はバイアスベクトルである．ここで $\boldsymbol{x}$ の次元を N，$\boldsymbol{o}$ の次元を M とすると，W は M 行 N 列の行列，$\boldsymbol{b}$ は M 次元の列ベクトルとなる．W，$\boldsymbol{b}$ はパラメータであり，前述の勾配法によって最適化を行う．全結合層は Affin 変換とも呼ばれ，行列の積和で表される線形変換であり，入力に回転や伸縮，平行移動を施す．出力 $\boldsymbol{o}$ の次元数は任意であり，次元の拡大や圧縮を行うことも可能である．

全結合層は複数の入力 $\boldsymbol{x}_1$, ..., $\boldsymbol{x}_N$ に対して，次式のように 1 回の行列計算として処理することができる．

$$(\boldsymbol{o}_1, ..., \boldsymbol{o}_N) = \mathrm{W}(\boldsymbol{x}_1, ..., \boldsymbol{x}_N) + \boldsymbol{b} \tag{3.15}$$

これは，上記のミニバッチ学習と相性がよく，効率的な計算が可能となる．$\boldsymbol{b}$ について，厳密には入力するデータ数にあわせて $\boldsymbol{b}$ の列数を増やす必要があるが，深層学習の文脈においてはこの操作は表記されないことが多い．

3.2.2　活性化関数

前述の通り全結合層は線形変換である．(3.16) 式に示すように，2 つの全結合層の合成は 1 つの全結合層と等価であり，全結合層を複数組み合わせても線形変換しか表すことができない．

$$\mathrm{W}_2(\mathrm{W}_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2 = \mathrm{W}_2\mathrm{W}_1\boldsymbol{x} + \mathrm{W}_2\boldsymbol{b}_1 + \boldsymbol{b}_2 = \mathrm{W}'\boldsymbol{x} + \boldsymbol{b}' \tag{3.16}$$

よって，DNN で非線形性をモデリングするためには非線形な活性化関数が必要不可欠である．ニューラルネットワークが考案された当初はステップ関数が用いられていたが，後述の誤差逆伝搬法の登場に伴って sigmoid 関数 ((3.17) 式，Fig. 3.2) や同じ sigmoid 属であるハイパボリックタンジェント ((3.18) 式，Fig. 3.3)，ソフトサイン ((3.19) 式，Fig. 3.4) が用いられるようになった．近年では，ReLU [36]((3.20) 式，Fig. 3.5) や Leaky-ReLU [37]((3.21) 式，Fig. 3.6) といった ReLU 属と呼ばれる関数が広く用いられている．

$$sigmoid(x) = \frac{1}{1+e^{-x}} \tag{3.17}$$

$$tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{3.18}$$

$$softsign(x) = \frac{x}{1+|x|} \tag{3.19}$$

$$ReLU(x) = max(0, x) \tag{3.20}$$

$$Leaky-ReLU(x) = max(ax, x) \tag{3.21}$$

また，分類問題においては，DNN の出力ベクトル $\boldsymbol{h}$ に次の softmax 関数を適用することが多い．

$$y_i = \frac{e^{h_i}}{\sum_k e^{h_k}} \tag{3.22}$$

ここで，h_i は DNN の出力ベクトル $\boldsymbol{h}$ の i 番目の要素であり，y_i は softmax の出力ベクトル $\boldsymbol{y}$ の i 番目の要素である．softmax 関数の出力ベクトル $\boldsymbol{y}$ について，$\boldsymbol{y} > 0$ と $\sum_k y_k = 1$ を満たすことから，$\boldsymbol{y}$ は離散確率分布のように扱うことができ，交差エントロピーなどの情報理論の知見を適用できる．

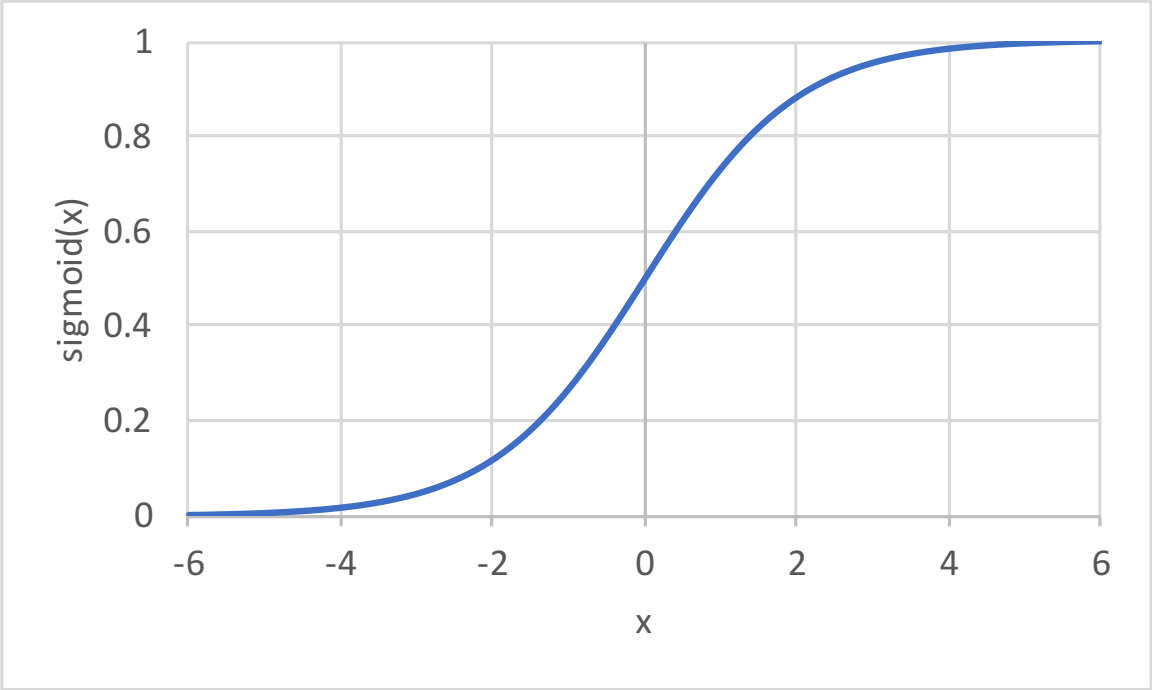


Fig. 3.2: sigmoid 関数

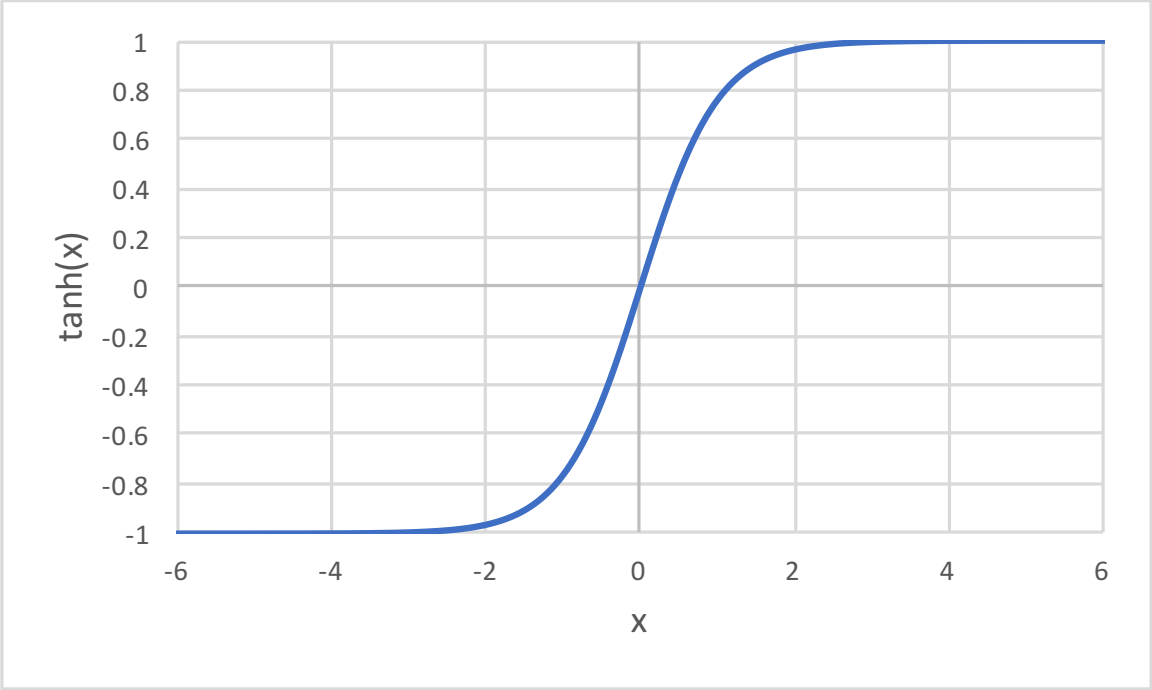


Fig. 3.3: tanh 関数

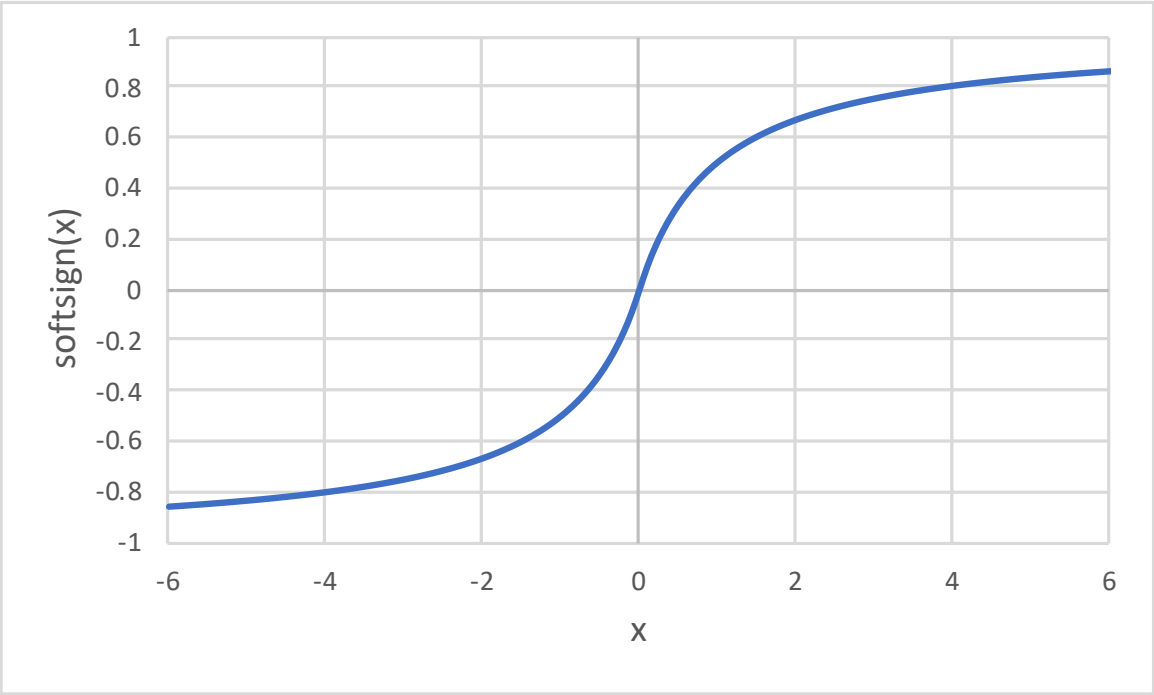


Fig. 3.4: softsign 関数

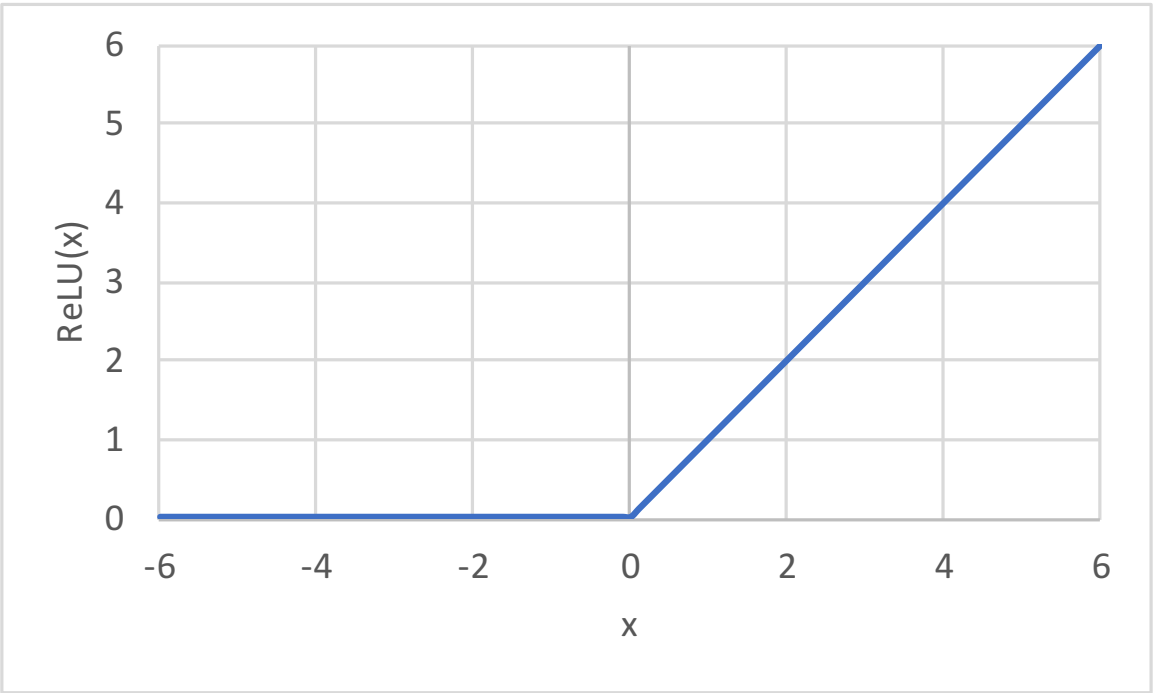


Fig. 3.5: ReLU 関数

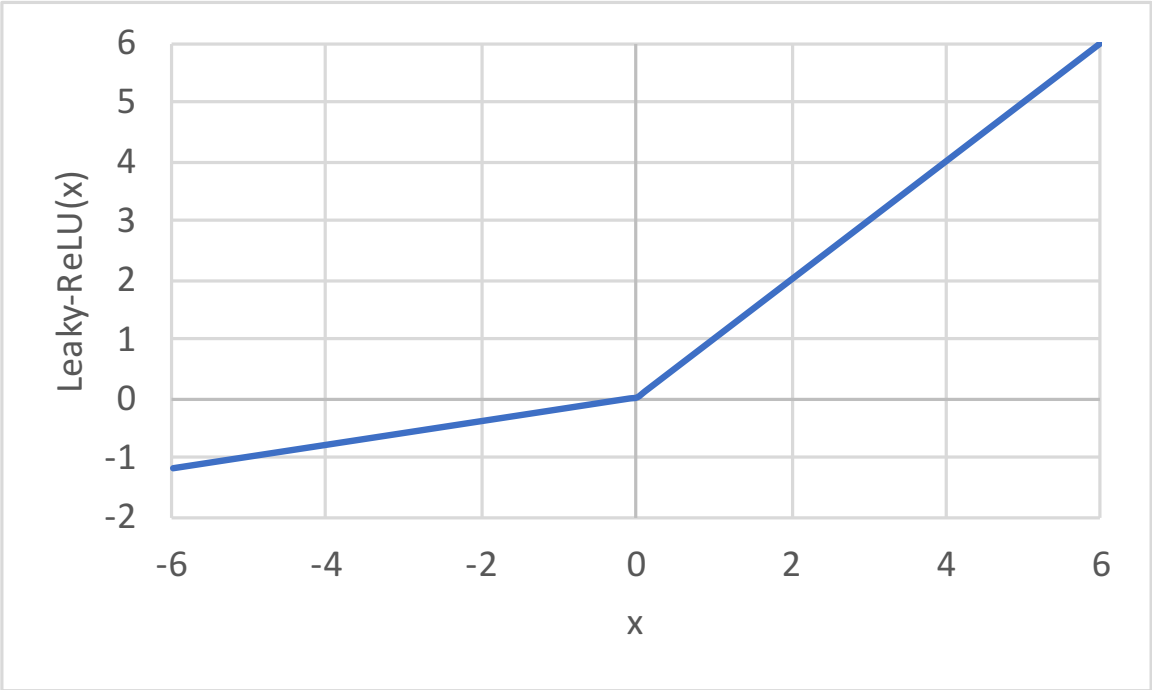


Fig. 3.6: Leaky-ReLU 関数

3.2.3　多層パーセプトロンへの勾配法の適用

MLP を勾配法を用いて最適化するために，MLP の持つ重みパラメータの勾配を求める方法について述べる．まず，k 層目の全結合層を $\mathrm{W}_k\boldsymbol{x}+\boldsymbol{b}_k$，$k$ 層目の活性化関数を $f_k(\cdot)$ と表し，以下のような 2 層のパーセプトロンの出力について，平均自乗誤差によって損失を計算する場合を考える．

$$\boldsymbol{h}_1 = \mathrm{W}_1\boldsymbol{x}+\boldsymbol{b}_1 \tag{3.23}$$
$$\boldsymbol{h}_2 = f_1(\boldsymbol{h}_1) \tag{3.24}$$
$$\boldsymbol{h}_3 = \mathrm{W}_2\boldsymbol{h}_2+\boldsymbol{b}_2 \tag{3.25}$$
$$\boldsymbol{y} = f_2(\boldsymbol{h}_3) \tag{3.26}$$
$$L = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{3.27}$$

誤差関数の平均自乗誤差の微分は以下のように表される．

$$\frac{\partial L}{\partial \boldsymbol{y}} = \frac{2}{n}(\boldsymbol{y}-\hat{\boldsymbol{y}}) \tag{3.28}$$

活性化関数 f_1, f_2 は入力ベクトルの要素それぞれに適用されるものだと仮定すると，損失 L に対する 2 層目の全結合層のパラメータ $\mathrm{W}_2, \boldsymbol{b}_2$ および入力 $\boldsymbol{h}_2$ の微分は次のように書ける．

$$\frac{\partial L}{\partial \boldsymbol{b}_2} = \frac{\partial L}{\partial \boldsymbol{y}} \odot \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{h}_3} \tag{3.29}$$
$$\frac{\partial L}{\partial \mathrm{W}_2} = \left(\frac{\partial L}{\partial \boldsymbol{y}} \odot \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{h}_3}\right)\boldsymbol{h}_2^{\mathrm{T}} \tag{3.30}$$
$$\frac{\partial L}{\partial \boldsymbol{h}_2} = \mathrm{W}_2^{\mathrm{T}}\left(\frac{\partial L}{\partial \boldsymbol{y}} \odot \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{h}_3}\right) \tag{3.31}$$

ここで，$\odot$ は要素ごとの積を表す，さらに，(3.31) を用いて損失 L に対する 1 層目の全結合層のパラメータ $\mathrm{W}_1, \boldsymbol{b}_1$ および入力 $\boldsymbol{h}_1$ の微分は次のように書ける．

$$\frac{\partial L}{\partial \boldsymbol{b}_1} = \frac{\partial L}{\partial \boldsymbol{h}_2} \odot \frac{\partial \boldsymbol{h}_2}{\partial \boldsymbol{h}_1} \tag{3.32}$$
$$\frac{\partial L}{\partial \mathrm{W}_1} = \left(\frac{\partial L}{\partial \boldsymbol{h}_2} \odot \frac{\partial \boldsymbol{h}_2}{\partial \boldsymbol{h}_1}\right)\boldsymbol{x}^{\mathrm{T}} \tag{3.33}$$
$$\frac{\partial L}{\partial \boldsymbol{h}_1} = \mathrm{W}_1^{\mathrm{T}}\left(\frac{\partial L}{\partial \boldsymbol{h}_2} \odot \frac{\partial \boldsymbol{h}_2}{\partial \boldsymbol{h}_1}\right) \tag{3.34}$$

このように，2 層のパーセプトロンの損失に対するパラメータの勾配をすべて求めることができる．またこの方法は，2 層以上のパーセプトロンにも同様に適用可能である．

このように連鎖律を用いることで，微分可能な関数の合成関数の微分は各関数の微分の積で表すことができる．これを Neural Network に適用することで，損失に対するパラメータの勾配を計算するのが誤差逆伝搬法である．誤差逆伝搬法を適用するために，ニューラルネットワークを構成する計算は全て，出力に対して入力およびパラメータで偏微分可能である必要がある．一方，この条件を満たすあらゆる計算を全てニューラルネットワークに導入することが可能である．活性化関数 ReLU ((3.20) 式)，Leakey-ReLU ((3.21) 式) は $x = 0$ の一点において微分不可能であるが，実際の計算上は $x = 0$ となることがないことから問題は生じない．実装上は右側微分あるいは左側微分のどちらかを $x = 0$ の微分として扱うことが多い．

深いネットワークの学習の際にこの偏微分の値が微小であった場合，所望の勾配が偏微分の積によって得られることから，勾配が 0 に近い値となりパラメータが更新されなくなる勾配消失が起こりうる．一方で，偏微分が大きな値であった場合，その積によって得られる勾配が非常に大きな値になることが考えられる．これを勾配爆発と呼ぶ．パラメータが大きく更新されることは過学習の原因の一つとなるほか，場合によっては，勾配が浮動小数点の範囲を超えてしまい，計算自体が不能になる場合がある．

3.3 DNN を適切に学習するための工夫

3.3.1 正規化

深層学習において取り扱うデータについて，正規化が広く行われている．特によく用いられるのは，データが 0 平均，単位分散になるようにデータの平均を引き，標準偏差で割る方法である．

また近年では，入出力のデータだけではなく，ネットワークの中間において中間層の出力である隠れ表現を正規化することが学習に有効であることが確かめられてきた．あるパラメータの偏微分が示す勾配は他のパラメータが変わらないという仮定のもとでの最急降下方向であり，勾配法では他のパラメータとの関係が考慮されない．特にニューラルネットにおいては，前層の出力から次層の出力を計算することから，パラメータの間に強い相関があることは自明である．よって，全てのパラメータを同時に更新する勾配法ではパラメータ間に依存関係があると特に学習率が大きい場合に損失が減少せず学習が安定しない．そこで，ネットワークの中間において隠れ表現ベクトルを正規化することで，このパラメータの依存性を低減する Batch Normalization [38] が考案された．ネットワークの計算途中である隠れ表現ベクトルのミニバッチ $B = \{\boldsymbol{h}_j\}_{j=1}^{|b|}$ について，

Batch Normalization は以下のように計算できる．

$$m_i = \frac{1}{|B|}\sum_{j=1}^{|B|} h_{i,j} \tag{3.35}$$

$$v_i = \frac{1}{|B|}\sum_{j=1}^{|B|} (h_{i,j} - m_i)^2 \tag{3.36}$$

$$\tilde{\boldsymbol{h}} = \frac{\boldsymbol{h} - \boldsymbol{m}}{\sqrt{\boldsymbol{v} + \epsilon}} \tag{3.37}$$

$$\boldsymbol{h}_{\mathrm{out}} = \boldsymbol{\psi} \odot \tilde{\boldsymbol{h}} + \boldsymbol{\beta} \tag{3.38}$$

ここで，$\boldsymbol{\psi} \odot \tilde{\boldsymbol{h}}$ は要素ごとの積を表す．(3.37) 式は要素ごとの除算である．ϵ は 0 除算を防ぐための微小量である．$\boldsymbol{m}$，$\boldsymbol{v}$ はそれぞれ隠れ表現ベクトルのミニバッチにおける平均と分散を表す．この $\boldsymbol{m}$ と $\boldsymbol{v}$ を用いて隠れ表現ベクトルを正規化する．その後，$\boldsymbol{h}$ と同じ次元のベクトル $\boldsymbol{\psi}$ と $\boldsymbol{\beta}$ で隠れ表現ベクトルの平均と分散の再調整を行っている．この $\boldsymbol{\psi}$ と $\boldsymbol{\beta}$ はその値によっては正規化をキャンセルできるパラメータであり，勾配法によって学習する．決定的な推論を行うため，学習時に学習データ全体の平均 $\boldsymbol{m}$ と分散 $\boldsymbol{v}$ をミニバッチから逐次的に計算しておき，推論時にはその値を用いる．Batch Normalization は，ミニバッチ内で平均と分散を計算することで，容易に誤差逆伝搬法の枠組みの中に導入することができる．Batch Normalization を導入する利点としては，ネットワークの中間においてデータを正規化することで学習率を大きく設定することができ，収束が早まる他に，ネットワークの初期値への依存性を軽減することや，正則化としても機能し過学習を抑制し汎化性能が向上する．Batch Normalization は，Feedforward Neural Network においては全結合層の直後に行われることが多い．これは非線形な活性化関数を適用したあとでは，隠れ表現ベクトルの統計量が大きく変化してしまうため，正規化の効果が低いためである．

その他の正規化法として，Layer Normalization [39] が考案されている．N 次元の隠れ表現ベクトル $\boldsymbol{h}$ について Layer Normalization は以下のように計算される．

$$m = \frac{1}{N}\sum_{i=1}^{N} h_i \tag{3.39}$$

$$v = \frac{1}{N}\sum_{i=1}^{N} (h_i - m)^2 \tag{3.40}$$

$$\tilde{\boldsymbol{h}} = \frac{\boldsymbol{h} - m}{\sqrt{v + \epsilon}} \tag{3.41}$$

$$\boldsymbol{h}_{\mathrm{out}} = \boldsymbol{\psi} \odot \tilde{\boldsymbol{h}} + \boldsymbol{\beta} \tag{3.42}$$

Batch Normalization との大きな違いとしては，Layer Normalization は平均 m と分散

v を隠れ表現ベクトルの要素に対して計算している点である．これによって，ミニバッチ数によらず Layer Normalization は同じふるまいをするということになり，大きなミニバッチサイズを設定できない状況 (特に後述の Recurrent Neural Network) に適した手法である．

入出力や隠れ表現ではなく，ネットワークの重みに正規化をかける Weight Normalization [40] も考案されている．DNN 内の重みテンソル W について，次のように再パラメータ化する．

$$\mathrm{W} = \frac{g}{||\mathbf{v}||}\mathbf{v} \tag{3.43}$$

ここで，$||\cdot||$ はユークリッドノルムである．この $\mathbf{v}$ と g をパラメータとして最適化する．提案された時点では $||\mathbf{v}||$ は入力に関する軸に関して独立に計算されることから，重み層の入力次元数と同じ大きさのベクトルであり，g はスカラー量であった．一方，$||\mathbf{v}||, g$ ともに，出力に関する軸に関して独立に計算し，重み層の出力次元数と同じ大きさのベクトルとする実装も広く用いられている．Weight Normalization も Layer Normalization と同様，ミニバッチ数によらず同じ振る舞いをするため，あらゆる問題に適用できる．また，一般に隠れ表現よりも重みのほうがサイズが小さいため，Batch Normalization や Layer Normalization よりも計算量が少ないという利点がある．

3.3.2 正則化

DNN は大量のパラメータを持ち，非常に自由度が高いため，汎化性能が向上しない過学習に陥りやすくなる．そこで，DNN の自由度を下げるような正則化を導入することで過学習を防ぐ事を考える．

L2 正則化は重回帰分析等にも用いられる手法で，深層学習の文脈では重み減衰 (weight decay) とも呼ばれる．パラメータ θ を最適化する際に損失に θ のユークリッドノルムの自乗に正則化の強さを調節するハイパーパラメータ λ をかけた $\lambda||\theta||^2$ を加えるという方法である．θ が大きいほど損失が大きくなることから，損失を最小化する過程で，パラメータ θ が過剰に増大することを防ぎ，過学習を低減する事ができる．

また，DNN 特有の手法として Dropout[41] が広く用いられている．これは隠れ表現をある割合 r で 0 にするという手法である．推論時は隠れ表現に r を乗算する．典型的には $r = 0.5$ を用いることが多いが，ハイパーパラメータとして調整する必要がある．推論時は隠れ表現に r を乗算する代わりに，学習時に隠れ表現に対して r を除算することで．推論時の乗算が不要になるという実装もある．Dropout が有効なのは，Dropout は DNN のなかに複数の小さな DNN を作り，その平均を出力するアンサンブル学習を擬似的に行っているからであるとされている．

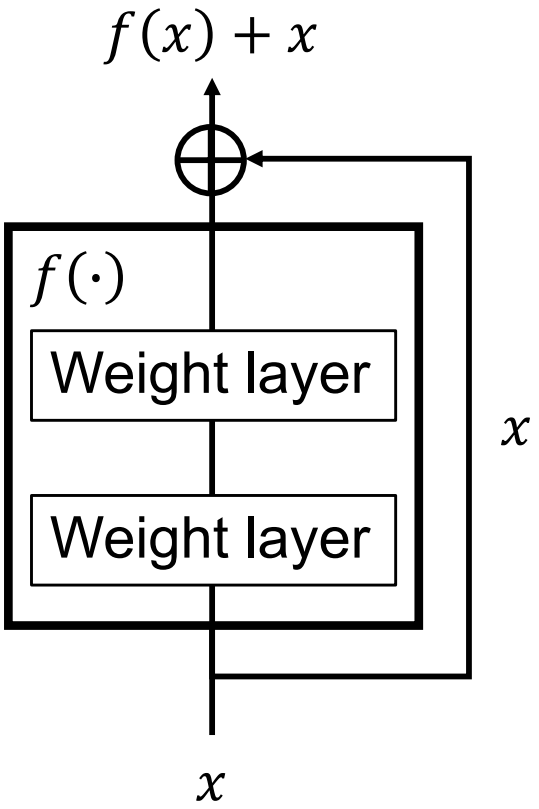


Fig. 3.7: Residual connection の概要図

他にも，Batch Normalization 等の正規化手法は正則化の役割を担っていることが知られている．

3.3.3 Residual connection

DNN の層数が多くなるほど勾配が適切に浅い層まで伝わらず，学習が困難となる．Residual Network[42] は画像識別の文脈で提案されたネットワーク構造であり，1000 層以上の DNN の学習を可能にした．特徴的なのは Fig. 3.7 のように複数層の重み層をスキップするようなショートカットを導入したことである．この Residual connection によって，出力層の勾配がショートカットをたどって浅い層にまで伝わるようになるため，層を深くしても学習が可能となる．

3.3.4 勾配の調整

前述の勾配消失，勾配爆発を防ぐ方法の一つとして勾配の値を調整するという方法がある．例えば，勾配がある閾値を超えた場合には，勾配を閾値に制限する，勾配のノルムを

計算しそれが閾値以下になるように勾配をスケーリングする，勾配にノイズを加えることでパラメータを摂動させ，最適化における鞍点を回避するなどのアプローチがある．

3.3.5　Incremental 法

深いネットワークを学習させる際の工夫の一つとして，Incremental 法 [43] と呼ばれる手法がある．これはまず，入力層と出力層からなるネットワークを学習させ，それが収束した後に，入力層に近い側の隠れ層を追加しさらにネットワークを学習させていくことを，隠れ層が所望の数になるまで繰り返すという手法である．隠れ層を追加する際には出力層は再初期化され，隠れ層追加後はこれまで学習してきた部分を含めてネットワーク全体を学習することに注意が必要になる．この手法は全ての層が出力層の直前の層としてトレーニングされるため，深いネットワークの隠れ層であっても有効に学習が可能である．

3.4　深層学習による時系列モデリング

3.4.1　Recurrent Neural Network

Recurrent Neural Network(RNN，再帰型ニューラルネットワーク) は，広義には何かしらの再帰構造を持つニューラルネットワークである．再帰構造によって前時刻の出力を次時刻の推論で用いるため，時系列の推定に適していると考えられる．最も単純な再帰構造としては，次式のように全結合層と活性化関数を適用した出力を次時刻の前層の出力と結合して全結合層の入力とする構造である．

$$\boldsymbol{h}_k = f\left(FC\left(\begin{bmatrix}\boldsymbol{x}_k \\ \boldsymbol{h}_{k-1}\end{bmatrix}\right)\right) \tag{3.44}$$

ここで，k は時刻を表す．この構造の積層によって RNN が構成される．再帰構造が表れても誤差逆伝搬法は適用可能であり，勾配法による最適化が可能である．

再帰構造を 2 つ組み合わせることで順方向と逆方向の時系列を同時に考慮する Bi-directional Recurrent Neural Network[44] が考案されている．ある時系列データ $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T)$ が与えられたとき，$X_{-1} = (\boldsymbol{x}_T, ..., \boldsymbol{x}_2, \boldsymbol{x}_1)$ と表すことにする．2 つの再帰構造をそれぞれ $\boldsymbol{R}_f$ と $\boldsymbol{R}_b$ と表すことにすると，Bi-directional RNN は以下のように書ける．

$$H_f = (\boldsymbol{h}_{f1}, \boldsymbol{h}_{f2}, ..., \boldsymbol{h}_{fT}) = \boldsymbol{R}_f(X) \tag{3.45}$$

$$H_{b-1} = (\boldsymbol{h}_{bT}, ..., \boldsymbol{h}_{b2}, \boldsymbol{h}_{b1}) = \boldsymbol{R}_b(X_{-1}) \tag{3.46}$$

$$H = \begin{bmatrix}H_f \\ H_b\end{bmatrix} = \left(\begin{bmatrix}\boldsymbol{h}_{f1} \\ \boldsymbol{h}_{b1}\end{bmatrix}, \begin{bmatrix}\boldsymbol{h}_{f2} \\ \boldsymbol{h}_{b2}\end{bmatrix}, ... \begin{bmatrix}\boldsymbol{h}_{fT} \\ \boldsymbol{h}_{bT}\end{bmatrix}\right) \tag{3.47}$$

時系列データ X を再帰構造 $\boldsymbol{R}_f$ に入力することで，順方向の時系列を考慮した出力 H_f を得る．さらに，逆方向の時系列データ X_{-1} を再帰構造 $\boldsymbol{R}_b$ に入力することで，逆方向の時系列を考慮した出力 H_{b-1} を得る．この 2 つの出力を時系列を揃えて結合することで，順方向と逆方向の時系列を同時に考慮した出力 H が得られる．Bi-directional RNN は未来の情報を利用することから，推論には全時刻の入力時系列データが必要になる．

RNN の再帰構造を時間方向に展開し，巨大な計算グラフとして解釈することができる．そこから RNN は時間方向に積層された深いネットワークであると考えることができる．深いネットワークで問題となるのが勾配消失あるいは勾配爆発である．これは偏微分の乗算の回数が増えるほど顕著に現れる．よって RNN においては，時間的に離れた場所で生じた誤差を伝搬することは困難になる．そのため，RNN は長期間の依存関係を考慮することが難しく直近の時系列の依存関係のみを学習してしまう傾向にある．この対策として，RNN にゲートを設けた構造が数多く提案されている．

Long Short-Term Memory(LSTM)[45] は，最も代表的なゲート付き再帰構造である．LSTM では入力ゲート $\boldsymbol{i}$，忘却ゲート $\boldsymbol{f}$，出力ゲート $\boldsymbol{o}$ と呼ばれる 3 つのゲートを使用する．また，長期の情報を保持するセルと呼ばれる隠れ表現ベクトルを使用する．LSTM は次のように計算される．

$$\begin{bmatrix} \overline{\boldsymbol{h}}_k \\ \boldsymbol{i}_k \\ \boldsymbol{f}_k \\ \boldsymbol{o}_k \end{bmatrix} = \begin{bmatrix} tanh \\ sigmoid \\ sigmoid \\ sigmoid \end{bmatrix} \left(\begin{bmatrix} W_{\overline{h}} \\ W_i \\ W_f \\ W_o \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_k \\ \boldsymbol{h}_{k-1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{b}_{\overline{h}} \\ \boldsymbol{b}_i \\ \boldsymbol{b}_f \\ \boldsymbol{b}_o \end{bmatrix} \right) \tag{3.48}$$

$$\boldsymbol{c}_k = \boldsymbol{i}_k \odot \overline{\boldsymbol{h}}_k + \boldsymbol{f}_k \odot \boldsymbol{c}_{k-1} \tag{3.49}$$

$$\boldsymbol{h}_k = \boldsymbol{o}_k \odot tanh(\boldsymbol{c}_k) \tag{3.50}$$

上式の $\overline{\boldsymbol{h}}_k$ は (3.44) 式における隠れ表現ベクトル $\boldsymbol{h}_k$ と同値である．$\overline{\boldsymbol{h}}_k$ を入力ゲート $\boldsymbol{i}_k$ で調節した値をセル $\boldsymbol{c}$ の更新分としている．さらに，忘却ゲート $\boldsymbol{f}_k$ で過去のセルの値を減少させている．言い換えれば，入力ゲートと忘却ゲートで短期と長期のバランスを調整してセルの値を更新している．最後に更新されたセルの値を出力ゲート $\boldsymbol{o}_k$ で調整した値が，LSTM の出力となる．

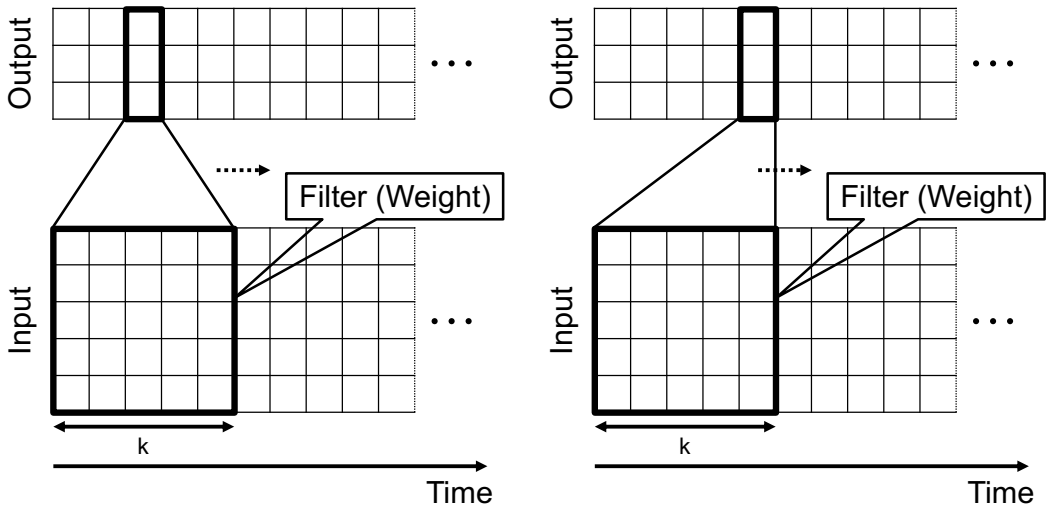
その他にも Gated recurrent unit (GRU)[46] と呼ばれる構造も提案されており，次のように計算される．

$$\begin{bmatrix} \boldsymbol{r}_k \\ \boldsymbol{z}_k \end{bmatrix} = sigmoid \left(\begin{bmatrix} W_r \\ W_z \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_k \\ \boldsymbol{h}_{k-1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{b}_r \\ \boldsymbol{b}_z \end{bmatrix} \right) \tag{3.51}$$

$$\tilde{\boldsymbol{h}}_k = tanh \left(W \begin{bmatrix} \boldsymbol{x}_k \\ \boldsymbol{r}_k \odot \boldsymbol{h}_{k-1} \end{bmatrix} + \boldsymbol{b} \right) \tag{3.52}$$

$$\boldsymbol{h}_k = (1 - \boldsymbol{z}_k) \odot \tilde{\boldsymbol{h}}_k + \boldsymbol{z}_k \odot \boldsymbol{h}_{k-1} \tag{3.53}$$

GRU は LSTM とは異なり，セルを用いず，隠れ状態 $\boldsymbol{h}_k$ のみに過去の情報を集約してい



(a) 非因果的　　(b) 因果的

Fig. 3.8: Time-Delay Neural Network の重み層

る．また，GRU は LSTM より 1 つ少ない 2 つのゲートを用いて，忘却と状態の更新を操作している．再設定ゲート $\boldsymbol{r}_k$ によって，1 時刻前の隠れ状態ベクトルを減衰させてから，更新分 $\tilde{\boldsymbol{h}}_k$ を計算し，更新ゲートを用いて，1 時刻前の隠れ状態ベクトルからの更新率を調整する．

LSTM と GRU はタスクによってどちらが優れているかが変わる．ただし，同じユニット数の場合は GRU のほうがパラメータ数と計算量が少ない利点がある．

3.4.2　Time-Delay Neural Network

Time-Delay Neural Network (TDNN) [47] は Fig. 3.8 で示すような，固定した時間幅を持つ時不変なフィルタで入力時系列を処理する重み層をもつニューラルネットワークである．全結合層との違いは，出力において時系列構造が保たれるという点である．TDNN の重みはこのフィルタであり，これを誤差逆伝播法の枠組みで最適化を行っていく．深層学習の文脈で TDNN は時間方向の 1 次元畳み込みニューラルネットワークと言い換えることができ，TDNN を用いた系列処理の有効性が確かめられている．これは，TDNN は RNN のような再帰構造を持たないため，時系列の並列計算が可能であり，なおかつ時間方向の誤差逆伝播による勾配消失の影響を受けないためであると考えられる．

バッチ処理的なタスクであれば，当該フレームの前後の情報を用いる非因果的な畳み込みを用いることができる．一方，リアルタイム処理的なタスクであれば，当該フレームの

過去の情報のみを用いる因果的な畳み込みを用いる必要がある．

3.4.3 Scaled dot-product self attention

Scaled dot-product self attention はニューラル機械翻訳で注目を集めた Transformer [48] というネットワークに使われていた構造である．

時刻と次元の 2 次元で表される時系列行列 S がある．この S に異なる全結合層を適用し，Key 行列 K と Value 行列 V を得る．また，S のある時刻のベクトルに対し，全結合層を適用し，query ベクトル $\boldsymbol{q}$ を得る．ここで，K と $\boldsymbol{q}$ の次元数は同じである必要がある．この Key 行列 K と Value 行列 V は辞書型の構造であり，そこから query ベクトル $\boldsymbol{q}$ を用いて，当該時系列からどの時刻を注意すればよいかを決定するのが Attention 構造であり，ここでは，K，V，$\boldsymbol{q}$ がすべて同じ時系列行列由来であるので，Self attention と呼ばれる．$\boldsymbol{q}$ と K の積から logit ベクトル $\boldsymbol{l}$ を得る．

$$\boldsymbol{l} = \frac{\boldsymbol{q}\mathrm{K}^T}{\sqrt{d}} \tag{3.54}$$

ここで，d は K と $\boldsymbol{q}$ の次元数である．$\boldsymbol{l}$ は $\boldsymbol{q}$ と K の各時刻での要素ベクトルの内積であり 2 者の類似度を表している．この logit ベクトル $\boldsymbol{l}$ に Softmax 関数を適用することで，重みベクトル $\boldsymbol{w}$ を得る．この重みベクトル $\boldsymbol{w}$ と Value 行列 V の積によって，Scaled dot-product self attention の出力ベクトル $\boldsymbol{o}$ を得る．

$$\boldsymbol{o} = \boldsymbol{w}\mathrm{V} \tag{3.55}$$

この $\boldsymbol{o}$ は Value 行列 V を重みベクトル $\boldsymbol{w}$ によって時刻に関しての重み付け和をとったものである．このようにして，query ベクトルに対して注意すべき時刻に重み付けされた出力 $\boldsymbol{o}$ が得られた．

この Scaled dot-product self attention は TDNN のように，時間方向の誤差逆伝播による勾配消失の影響を受けない．さらに，TDNN とは異なり，大局的な文脈を捉えることができる．また，K，V，$\boldsymbol{q}$ を分割することで，複数の Attention を計算する Multi-head attention も提案されている．また，Attention 機構は別の時刻に同じベクトルがある場合にはそれを区別できない．これに対しては位置エンコーディングと呼ばれる，時刻と次元から決まる行列を隠れ表現に加えることによって時刻の情報を付加することで対処している．

Transformer は Fig. 3.9 のような構造を基本単位として構築される．Multi-Head Attention と 2 層の全結合層が交互に繰り返し積層されており，それぞれの出力に対して正則化のために Dropout が適用されている．また，それぞれに Residual connection が導入されており，その後に Layer normalization を導入することで，隠れ表現の正規化が

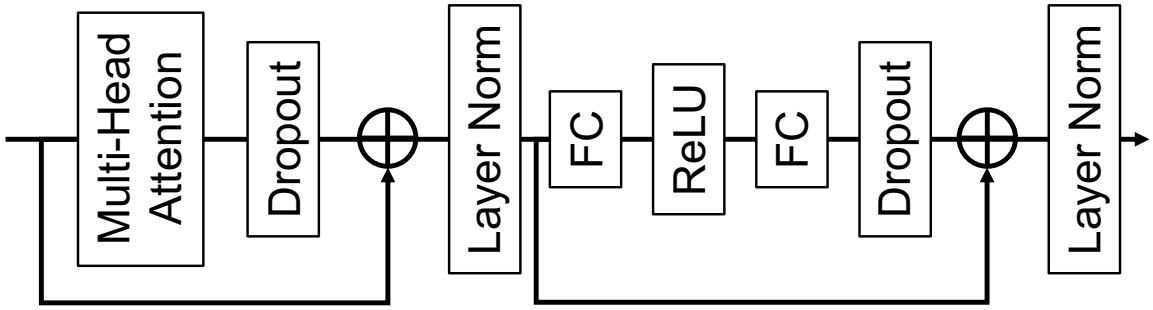


Fig. 3.9: Transformer の基本構造

行われている．

3.4.4 時空間畳み込み

時空間畳み込み [49] は，空間 (画像の縦と横) と時刻の 3 次元で表される動画データに対して適用される畳み込み層であり，使われるフィルタも同様に時空間の 3 次元で，動画データの時空間の局在性をもとに特徴量を抽出する方法である．動作識別の文脈で提案され，その有効性が確かめられている．畳み込みニューラルネットワークは主に畳み込み層，活性化関数の他に pooling が用いられる．pooling を導入し，隠れ表現の局所特徴を集約することで，微小な変化に対して頑健になることが期待される．よく用いられるものとしては時空間の局所から最大値のみを上層に伝える max-pooling と呼ばれるものがある．また，局所ではなく次元全体を対象とした global pooling と呼ばれる演算も用いられる．例えば，次元，空間，時刻の 4 次元で表されるテンソルに対して，空間に対して平均値を取る global average pooling (GAP) を適用することで，次元と時刻の 2 次元行列へ縮約できる．

3.5 まとめ

ここでは，本研究の前提となる深層学習に関する話題のうち，教師あり学習アルゴリズムと勾配法による最適化，多層パーセプトロンとその誤差逆伝播法の適用，DNN を適切に学習するための種々の工夫，時系列を取り扱うためのネットワーク構造である，RNN，TDNN，Scaled dot-product self attention，時空間畳み込みの説明を行った．

第 4 章

日本語調音・音声パラレルデータの収集

本章では，調音・音声間変換の基礎的データとなる，日本語調音・音声パラレルデータの収集を行ったことについて述べる．まず，調音データの収録を効率的に行うために限られた文章の中に多様な音素文脈が登場するように文選択を行う方法について説明する．

これまで，調音情報の収録方法としては 3 次元磁気センサシステム (3D-EMA) およびビデオカメラによる口唇動画の 2 つの手法を用いて収集を行っており，ここでは 3D-EMA による調音情報の測定原理を紹介し，これまでに収集したデータの一例を紹介する．

4.1 発話文の選定

ヒトは舌や唇などの調音器官の位置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う．これらの言語音のうち，言語体系の中で同じ音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，連続的に調音を行うことで所望の音素列を生成しているが，調音器官の物理的な制約により，同じ音素であっても前後の音素の影響によって調音運動の様態が変化する調音結合が生じる．また，音声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この音素文脈を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． ダイフォンやトライフォンの例を Table 4.1 に示す．

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．日本語において，音声研究に広く用いられているコーパスとして，ATR デジタル音声データベース セット B [50] がある．このコーパスに用いられている文章リストは，新聞，雑誌，小説，手紙，教科書等の文献から無作為に抽出した約 1 万文をもとに，

Table 4.1: 音素表現の例.

単語	あらゆる
音素列	/a/, /r/, /a/, /y/, /u/, /r/, /u/
ダイフォン	a, a/r, r/a, a/y, y/u, u/r, r/u, u,
トライフォン	a+r, a-r+a, r-a+y, a-y+u, y-u+r, u-r+u, r-u,

ダイフォン 402 種類とトライフォン 223 種類がバランスよく出現するように作成された，30 分弱で読み上げ可能な 503 個の音素バランス文である．この文章セットは「ATR 音素バランス文」とも呼ばれており，研究用途での日本語の音声収録における事実上の標準となっている．その一方，近年の計算機の発展に伴って，機械学習に用いられるデータの量は増加の一途を辿っており，音声においても例外ではない．日本語音声コーパスとしては，JSUT[51] という 10 時間の音声を含むコーパスが公開されており，この中の basic5000 と呼ばれる文章セットは常用漢字の音読み・訓読みを全てカバーするように選定されており，日本語音声に登場するであろう音素文脈がほぼ網羅されている．

調音情報の収集という観点から考えると，収録には特殊な装置が必要で，場合によっては自然な発話が阻害されるなど，発話者が特殊な環境に置かれることが多く，長時間のデータを収録することは困難であることから，調音情報の収集に用いる発話文セットは限られた文章の中に多様な音素文脈が出現する音素バランス文を用いることが望ましいと考えられる．一方，近年広く用いられる英語の EMA コーパスである MNGU0[52] が 1 時間程度であることから，これに合わせて，日本語 EMA データを収集する際にも 1 話者につき 1 時間を目安に収録を行うのが望ましい．この場合 ATR 音素バランス文だけでは足りず，追加の文章リストを用意する必要がある．そこで本研究では，ATR 音素バランス文に加えて用いるための，できるだけ多様なトライフォンが登場する，ATR 音素バランス文と同程度のモーラ数を持つ音素バランス文を新たに構築した．

4.1.1 文章候補の選定と音素付与

音素バランス文の文候補としては，日本語版 Wikipedia の本文データを用いた．これは文章量が多く，二次配布が可能であり，文章がある程度整っているという利点を有するからである．音素バランス文の選択に当たっては，まず，文章への読みの付与を行い，その読みを音素列へと変換する必要がある．最初に，形態素解析ソフト MeCab [53] と新語に対応するための MeCab 用の辞書である Neologd [54] を用いて文章に読みを付与した．続いて音声認識ソフトウェア Julius [55] の音素変換表に従い読みから音素列への変換を

行い，この音素列をトライフォンへ変換することで文章にトライフォン列を付与した．文章と得られたトライフォン列の 1 例を Table 4.2 に示す．

Table 4.2: 文章から自動付与されたトライフォン列の例.

文章	少ない施肥量が過剰施肥を防ぐ。
読み	すくないせひりょうがかじょーせひをふせぐ。
音素列	s u k u n a i s e h i ry o u g a k a j o: s e h i o f u s e g u
トライフォン列	s+u s-u+k u-k+u k-u+n u-n+a n-a+i a-i+s i-s+e s-e+h e-h+i h-i+r i-ry+o y-o+u o-u+g u-g+a g-a+k a-k+a k-a+j a-j+o j-o:+s o-s+e s-e+h e-h+i h-i+o i-o+f o-f+u f-u+s u-s+e s-e+g e-g+u g-u

4.1.2　Minoux の改良貪欲法 [56] を用いた文選択 [57]

文集合 U と許容コスト B が与えられた時，U の部分集合 S の効用が最大になるように S を選択する問題を考える．このとき，$J(S)$ を文部分集合 S の効用を評価する集合関数とし，文部分集合選択問題を次のように定式化する．

$$S^* = \arg\max_{S \subseteq U} J(S) \ \text{ subject to } \ \sum_{s \in S} c(s) \leq B \tag{4.1}$$

ただし，$c(s)$ は文 s のコストである．例えば全ての文に対して単位コスト $c(s) = 1$ を用いる場合，最大で B 個の文を選択できるという制約条件になる．このような制約のことを「サイズ制約」と呼ぶ．ただし，この制約においては限られた文数でより多くのトライフォンを網羅するとなると，長い文が優先的に選択されることは容易に予測できる．そこで本研究では，コスト $c(s)$ として文に含まれるモーラ数を用いる．つまり長い文ほどコストが高いということになる．このような制約のことを「ナップザック制約」と呼ぶ．

上記の効用を示す関数 J を目的関数と呼ぶことにし，その設定方法を示す．まず，P を全音素の集合とすると，全音素の種類数は $|P|$ で表される．π_i を i 番目の音素の所望の確率，$\pi = (\pi_1, ..., \pi_{|P|})$ を所望の音素分布，$f_i(S)$ を文の集合 S における i 番目の音素の頻度とする．この時，文部分集合 S の効用を次式により定義する．

$$J(S) = \sum_{s \in S}^{|P|} \pi_i \log f_i(S) \tag{4.2}$$

この目的関数は，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くなる性質を持っている．本研究においては一様分布，すなわち全トラ

イフォンが等価となるように π を設定した．さらに，この目的関数は劣モジュラ性を満たす．

任意の $S \subseteq T \subseteq U$ と $s \in U \backslash T$ に対して，$J(\cdot)$ が以下の性質を持つとき，その集合関数は劣モジュラであるという．

$$J(S \cup \{s\}) - J(S) \geq J(T \cup \{s\}) - J(T) \tag{4.3}$$

文部分集合選択の文脈でいうと，「ある文 s を小さな文部分集合 S に加える場合と，同じ文を大きな文部分集合 T に加える場合とでは，前者の方が効用の増分が大きい」という意味である．

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ことが知られているため，近似アルゴリズムを用いる必要がある．$J(\cdot)$ 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．ナップザック制約に対しては厳密解の評価値に対して最悪でも約 32% の評価値を持つ解が得られることが理論的に保証される．Algorithm 1 に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選択するアルゴリズムを示す．しかしながら，このアルゴリズムでは，Algorithm 1 の 3 行目で行われる，追加したときの効用が最大となる文 s^* を求める演算量が $O(n!)$ となり膨大な時間がかかってしまう．そこで貪欲法に Minoux の改良 [56] 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回数を大幅に削減することができる．

Algorithm 1 ナップザック制約による文部分集合選択アルゴリズム

Require: U, B

1: $S \Leftarrow \phi$
2: **while** $\sum_{s \in S} c(s) \leq B$ **do**
3: $\quad s^* \Leftarrow \arg\max_{s \in U \backslash S} \frac{J(S \cup \{s\}) - J(S)}{c(s)}$
4: $\quad S \Leftarrow S \cup \{s^*\}$
5: **end while**

Ensure: S

Minoux の改良を説明するために priority queue Q というデータ構造を導入する．Q は文 s とその文を文部分集合 S に加えたときの効用の増分 $\alpha = (J(S \cup \{s\}) - J(S))/c(s)$ のペア (s, α) を保持する構造である．また，それぞれの α は計算された直後は"fresh"であるとし，文部分集合 S が更新されたときに α は"fresh"ではなくなる．Q に関して次の 3 つの演算を定義しておく．

1. 新たな (s, α) を Q に挿入する．

$$\mathrm{INSERT}(Q, (s, \alpha)) \tag{4.4}$$

2. 最大の α を持つペア (s, α) を Q から取り出す．

$$(s, \alpha) \Leftarrow \mathrm{POP}(Q) \tag{4.5}$$

3. Q の中の最大の α を求める．

$$\mathrm{MAX}(Q) \tag{4.6}$$

この priority queue Q 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 Minoux の改良貪欲法によって選択するアルゴリズムを Algorithm 2 に示す．Algorithm 2 においては α が 12 行目の条件を満たせば，その文 s を文部分集合 S に追加することができ，Algorithm 1 と比べて評価回数が減少することが期待できる．

Algorithm 2 Minoux の改良を導入したナップザック制約による文部分集合選択アルゴリズム

Require: U, B

1: $S \Leftarrow \phi$
2: Initialize Q
3: **for** $s \in U$ **do**
4: $\quad \alpha \Leftarrow \frac{J(S \cup \{s\}) - J(S)}{c(s)}$
5: $\quad \mathrm{INSERT}(Q, (s, \alpha))$
6: **end for**
7: **while** $\sum_{s \in S} c(s) \leq B$ **do**
8: $\quad (s, \alpha) \Leftarrow \mathrm{POP}(Q)$
9: $\quad$ **if** α is not "fresh" **then**
10: $\quad\quad \alpha \Leftarrow \frac{J(S \cup \{s\}) - J(S)}{c(s)}$
11: $\quad$ **end if**
12: $\quad$ **if** (α in line 8 is "fresh") OR ($\alpha \geq \mathrm{MAX}(Q)$) **then**
13: $\quad\quad S \Leftarrow S \cup \{s\}$
14: $\quad$ **else**
15: $\quad\quad \mathrm{INSERT}(Q, (s, \alpha))$
16: $\quad$ **end if**
17: **end while**

Ensure: S

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音素バランス文を完成させた．

ATR 音素バランス文の全 503 文は 50 文のサブセット (1 つだけ 53 文) からなっており，このサブセットそれぞれで音素バランスが保たれている．それを参考に，本研究では選択した文部分集合をモノフォンのバランスを保った上でサブセットに分割を行った．手順としては，まず文部分集合の中で最も出現回数が少ない音素が含まれる文を 1 文ずつサブセットに配置する．つまり，文部分集合の中で最も出現回数が少ない音素が含まれる文数がそのままサブセット数となる．そして，残った文部分集合の中で最も出現回数が少ない音素を最も多く持つ文を取り出し，その音素が最も少ないサブセットに追加するということを繰り返すことでサブセットを作成した．

4.1.3　音素バランス文の作成と評価

新たに作成する音素バランス文は，後述する EMA 装置の制約から文章の長さができるだけ一定の短文の集合であることが望ましい．そこで 4.1.1 節で選定した文章群のうち 7〜12 モーラの長さの文書を文候補として，文選択を行った．得られた文部分集合を付録 A に掲載している．このバランス文に関する情報を Table 4.3 に示す．ATR 音素バランス文に得られた音素バランス文を加えることで，総モーラ数が 1.74 倍，2 回以上登場したダイフォン種類数が 1.42 倍，同じくトライフォン種類数が 1.51 倍となり，より多様な音素文脈の収集が可能になると考えられる．

Table 4.3: 付録 A の音素バランス文に関する情報．ダイフォン種類数とトライフォン種類数のカッコ内の数字は 2 回以上登場した種類数を表す．

	付録 A	ATR 音素バランス文	合算
文章数	1298	503	1801
総モーラ数	12641	16970	29611
各文のモーラ数	7〜12	10〜73	7〜73
サブセット数	3	10	13
サブセットごとの文章数	419〜441	50, 53	50〜419
ダイフォン種類数	627 (573)	503 (441)	675 (628)
トライフォン種類数	5333 (2222)	2834 (2142)	5378 (3248)

4.1.4　音素バランス文の再設計と評価

4.1.3 節で構築した音素バランス文を用いて予備的に磁気センサによる調音データの収集を行ったところ，2 つの問題が生じた．一つは，1 文の長さが想像以上に短く，文章としての体裁をなしていないものも多く含まれること．もう一つが，外国語の固有名詞のカタカナ表記が数多く選択されており，その中に日本語には含まれない音素文脈が多く含まれていて，調音器官にコイルを貼り付けた状態では発声が困難になるという点である．そこで，4.1.1 節の文章群のうち，以下の条件を満たすもののみを文候補として，文選択を行再度行った．

1. 文章のモーラ数が 18〜29.
2. 「・」を含まない.
3. 文末が句点である.
4. 1 文字以上ひらがなが含まれる.
5. 形態素解析の結果，固有名詞が含まれない.
6. 形態素解析の結果，文頭が助詞でない.

また，ここで，構築した音素バランス文を ATR 音素バランス文と併用することを前提として，ATR 音素バランス文に含まれないトライフォンを優先的に選択する様に目的関数を次のように変更した．

$$J(S) = \sum_{s \in S}^{|P|} \log(f_i(S) + f_i(S_{\mathrm{ATR}}) + 1) \tag{4.7}$$

ここで，S_{ATR} は ATR 音素バランス文を表す．このように目的関数を変更することで，ATR 音素バランス文にすでに多く含まれている音素が追加された場合の効用の増分が小さくなるので，ATR 音素バランス文に含まれないトライフォンを優先的に選択するようになることが期待される．

得られた文部分集合を付録 B に掲載している．このバランス文に関する情報を Table 4.4 に示す．ATR 音素バランス文に得られた音素バランス文を加えることで，総モーラ数が 2.09 倍，2 回以上登場したダイフォン種類数が 1.37 倍，同じくトライフォン種類数が 1.70 倍となり，より多様な音素文脈の収集が可能になると考えられる．4.1.3 節と比較すると ATR503 文と合算した際の 2 回以上登場したトライフォン種類数が向上しており，目的関数の変更による効果が見られる．

Table 4.4: 付録 B の音素バランス文に関する情報．ダイフォン種類数とトライフォン種類数のカッコ内の数字は 2 回以上登場した種類数を表す．

	付録 B	ATR 音素バランス文	合算
文章数	884	503	1387
総モーラ数	18474	16970	35444
各文のモーラ数	18〜29	10〜73	10〜73
サブセット数	8	10	18
サブセットごとの文章数	107〜120	50, 53	50〜120
ダイフォン種類数	636 (586)	503 (441)	641 (605)
トライフォン種類数	4626 (3173)	2834 (2142)	4846 (3643)

4.2　3D-EMA による調音情報の収集

4.2.1　先行研究

これまで EMA コーパスとしては，男性 1 名，女性 1 名が TIMIT 文 (460 文) [58] を読み上げた際の調音運動を 2 次元の磁気センサ (EMA) システム (Carstens 社 AG200) を用いて測定した，MOCHA-TIMIT コーパス [59] が広く用いられてきた．この英語音声コーパスは，調音データに大局的な時間変動が見られ，これが音響–調音逆推定に悪影響を及ぼすことが報告されている [60]．2 次元測定システムでは，マーカーコイルが話者の頭部正中面上に正確に位置していることが必要であり，長時間の測定によって頭部が測定平面からずれた場合には，このような測定誤差を生じる原因となる．この問題は，コイル位置の 3 次元測定 [61] を用いることで解決可能である．例えば，EMA データから頭部の動きの影響を除くために，調音とは関係のない鼻や上顎にも受信コイルを貼り付け，これらが固定点となるように他のコイルの座標位置を補正する操作が行われる．この処理によって，3 次元計測では，2 次元計測では不可能であった正中断面からの位置ずれの補正が可能となった．

3 次元測定による調音コーパスは，1 名のイギリス人英語話者による 1354 文の読み上げに関する MNGU0 [52]，7 名の健常英語話者及び 8 名の構音障害者 (小児性脳性麻痺 7 名，ALS1 名) の調音運動からなる，計 23 時間を含む TORGO[?]，USC-TIMIT[63] や EMA-IEEE[62] 等が良く知られているが，現状で最も利用されているのは MNGU0 である．また，現在までのところ，日本語の調音コーパスは公開されたものが存在しない．日

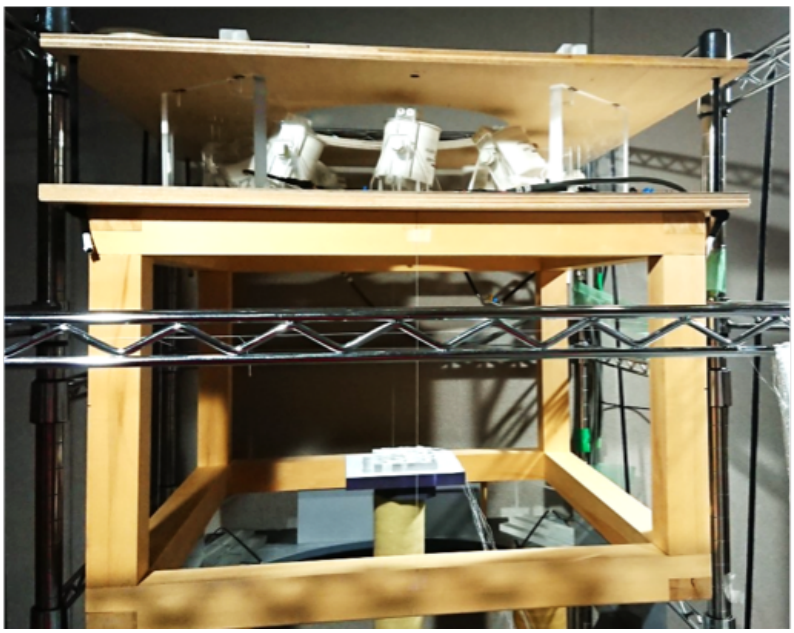


Fig. 4.1: 研究室作成の EMA 装置．左の写真が送信コイルが取り付けられた筐体，右の写真が受信コイルを話者に取り付けた様子である．

本語の調音運動コーパスを公開することができれば，日本語音声に固有の特徴を含む調音運動のモデル化など，日本語の音声研究分野での活用が期待できる．そこで，ここでは，収集データの公開も視野に入れ，MNGU0 コーパスと同程度の発話量を持つ大規模日本語調音運動コーパスの構築を目指し，調音・音声パラレルデータを収集したことについて述べる．

4.2.2　3 次元磁気センサによる調音運動の観測原理

本研究で用いる 3 次元磁気センサシステムは，8 個の送信コイルにより生成した磁界中に位置マーカとしての受信コイルをおくことにより，受信コイルに誘導起電力が生じ，その信号強度から受信コイルの 3 次元位置を推定するシステムである．この受信コイルを舌，唇，下顎等の調音器官に貼付すれば，発話中の調音器官の運動がコイルを貼付された観測点の運動という形で測定できる．Fig. 4.1 に研究室で作成した EMA 装置を示す．

磁気センサは，大きな雑音の発生源がないため，音声と調音運動の同時収録が可能となっている．また，磁気センサによって収録される調音情報は，他の手法と比較して時間分解能が高く，得られた運動データを煩雑な後処理なしでそのまま調音情報として使用できるメリットがある．他方，調音器官にマーカーコイルを取り付けた状態で発話を行うことは話者にとって大きな負担となることから，長時間の使用は難しい点には注意が必要である．

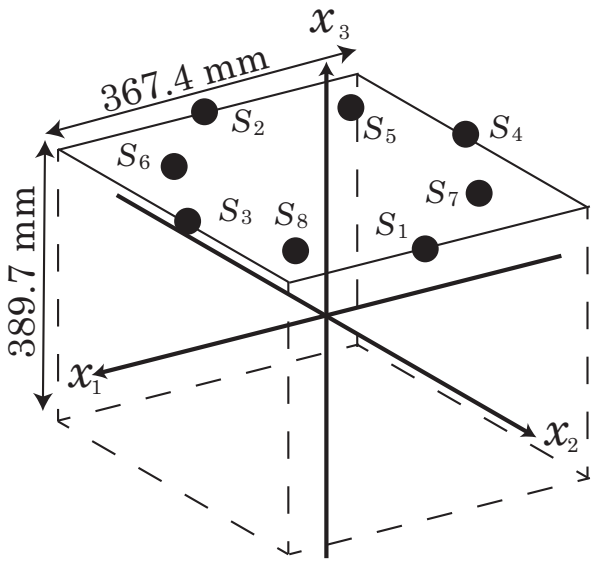


Fig. 4.2: 8 個の送信コイルを持つ 3 次元磁気センサシステムの送信コイルの配置

3 次元磁気センサシステムを用いた位置推定法の概要は次の通りである．Fig. 4.2 に示す位置に配置された 8 個の送信コイル $(S1, S2, \ldots, S8)$ は互いに異なる周波数の交流磁界を生成する．生成された交流磁界中に置かれた受信コイルには受信信号が誘導される．この時の受信コイルの位置及び傾きの状態 $\boldsymbol{s} = (x_1, x_2, x_3, \theta_1, \theta_2)$ を，Fig. 4.3 のように定める．本システムでは，送信コイルを磁気双極子とみなし，送信コイルの生成する磁界をクーロン則に基づく磁界モデルにより表現する．本磁界モデルを用いることにより，任意の位置 $\boldsymbol{x} = (x_1, x_2, x_3)$ における磁界 $\mathbf{y}(\boldsymbol{x})$ が求められる．状態 $\boldsymbol{s}$ における受信信号予測値はこのとき，下記の式で表現される．

$$\hat{z}(\boldsymbol{s}) = g\mathbf{y}(\boldsymbol{x}) \cdot \mathbf{e}(\theta_1, \theta_2) \tag{4.8}$$

ここで，g は受信コイルの校正時に決定されるゲインパラメータであり，$\mathbf{e}(\theta_1, \theta_2)$ は受信コイルの向きを表す単位ベクトルである．受信コイルの状態 $\boldsymbol{s}$ は，測定された受信信号 z_l $(l = 1, 2, \ldots, 8)$ とその予測値 $\hat{z}_l$ の間の自乗誤差 E が最小となるように定められる．

$$E = \sum_{l=1}^{8} (z_l - \hat{z}_l(\boldsymbol{s})) \tag{4.9}$$

この問題は非線形最小自乗問題となるため，本研究では繰り返し最適化手法の一種である Gauss-Newton 法を用いて最適解を求めることで，受信コイルの状態 $\boldsymbol{s}$ を推定する．

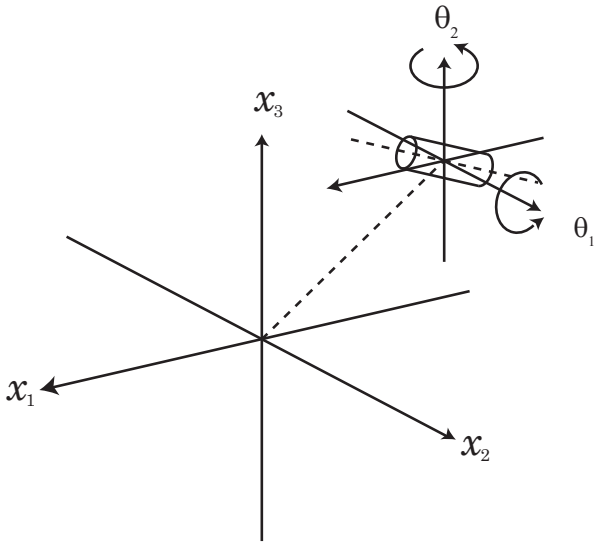


Fig. 4.3: 受信コイルの状態

4.2.3　測定条件

システムのハードウェア構成と受信コイルの取り付け位置を Fig. 4.4 に示す．本システムは，8 個の送信コイルを搭載した 655 × 655mm の木製の筐体と受信コイル，合計 8 チャネル分の送信アンプ，12 チャンネルの受信アンプ，サンプリングレート 100 kHz で 8 チャンネル同時 DA 可能な DA 変換器 (DASmin- E2000 model 1608-100k-DA，コーメックス電子)，サンプリングレート 50 kHz で 16 チャンネル同時 AD 可能な AD 変換器 (DASmin-E2000 model 1616-100k-AD，コーメックス電子) および Linux PC で構成されている．PC は送信コイルに関する正弦波駆動信号の生成と受信コイルからの信号の処理のために使用される．送信コイルは図 4.2 に示すように半径 18.37 cm の円周上に等間隔に配置する．送信コイルの生成する交流磁界の周波数はそれぞれ，12375 Hz，8500 Hz，13375 Hz，11375 Hz，9375 Hz，10375 Hz，11875 Hz，10875 Hz とした．

また，本システムでは，3D-EMA による調音運動，マイクを用いて収集される音声信号に加え，発話中の声帯の接触の程度を電気喉頭計 (Electroglottography; EGG) により測定した．3D-EMA による測定点は，上唇 (UL)，下唇 (LL)，下顎 (J)，舌上の 3 点 (T1, T2, T3) とした．また，発話中の頭部の動きを補正するための参照点として，鼻背 (ND)，鼻尖 (NA)，上顎 (UI) の 3 点の位置を測定した，なお，全ての測定点は頭部のほぼ正中面上にとった．発話中の各受信コイルの受信信号と音声信号並びに EGG 信号を AD 変換器により 50 kHz で標本化し PC に取り込んだ．

各受信コイルの受信信号から調音運動データを得るために，下記の処理を行った．ま

ず，受信信号を送信コイルの送信周波数に対応した周波数成分に分けた．その際，窓長 8 ms (400 サンプル) の矩形窓を使用し，シフト幅を 4 ms(200 サンプル) とすることにより，サンプリングレート 250 Hz で各送信コイルに対応する受信信号を取得した．次に，その受信信号をもとに受信コイルの状態を推定することで受信コイルの位置情報を得た．最後に，得られた受信コイルの位置情報に対し．頭部参照点の位置情報を用いて発話中の頭部の動きを補正し，遮断周波数 12.5 Hz のローパスフィルタにより平滑化処理を行った．得られた位置情報データは 3 次元位置情報となっており，x 座標値は頭部の左右方向の変位，y 座標値は前後方向の変位，z 座標値は上下方向の変位をそれぞれ示している．また，各軸の正の方向を x 軸方向では発話者から見て左に，y 座標方向では後方に，z 軸方向では上方にとった．

音声信号データの収録には, Brüel & Kjær 社製 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)，Type 2669-L プリアンプ，NEXUS Tpye 2690-A-0S2 マイクコンディショニングアンプを使用した，その際，マイクスタンドを用いて，発話者の口唇とマイク間の距離を 20〜30 cm に設定した．得られた音声信号を 50 Hz のハイパスフィルタを通すことにより，低域のノイズ及び DC 成分を除去した．EGG 信号データの収録には，Glottal enterprises 社製の EG-2 を使用した．得られた EGG 信号にも 50 Hz のハイパスフィルタを適用した．

4.2.4 収録データの例

3D-EMA システムを用いて，男性話者 1 名の ATR 音素バランス文と 4.1.3 節で構築した音素バランス文の読み上げデータの収録を行った．総発話時間は 67 分となり MNGU0 コーパスとほぼ同等となった．Fig. 4.5 に「あらゆる現実をすべて自分のほうへねじ曲げたのだ」と発話した際の収録データを示す．図の横軸は時間，上から順に音声波形，EGG 波形，上唇 (UL)，下唇 (LL)，下顎 (J)，舌 (T1, T2, T3) の測定点のデータを表示している．縦の点線は各音素セグメントの区切りを示している．どの測定点においても，x 方向 (正中面に対して垂直方向) の変位が他の方向と比べて少ないことがわかる．これは，通常発話においてにおいてこの方向に調音器官を運動させることが少ないためである．よって，3D-EMA を用いて調音運動を収録する場合も正中面上の運動にのみ着目することが多い．

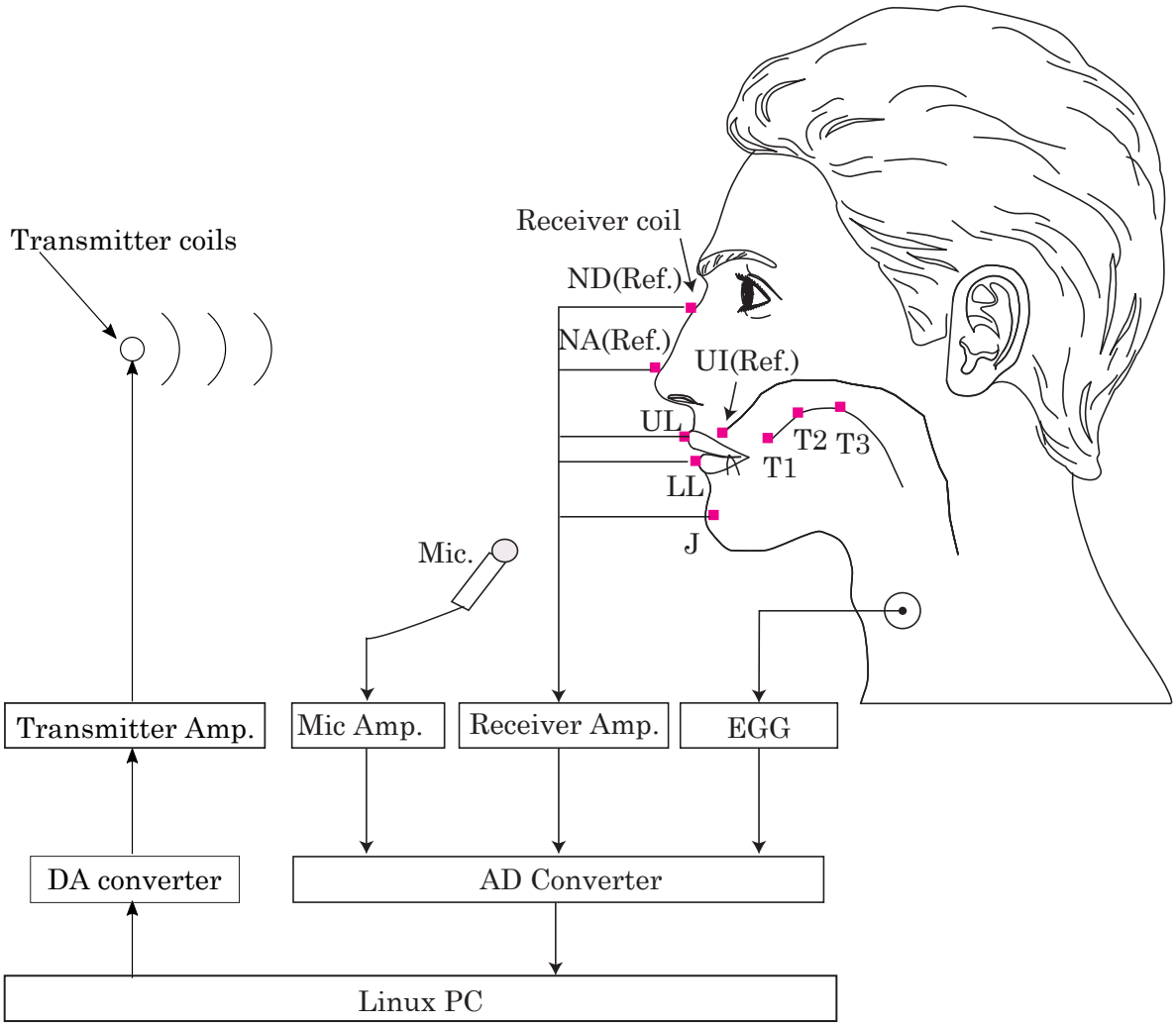


Fig. 4.4: 調音・音声パラレルデータ収集システムの構成並びに 3D-EMA の測定点

4.3　口唇動画による調音情報の収集

4.3.1　先行研究

口唇動画はビデオカメラで口唇の動きを録画することで収録できるため，最も実応用しやすい調音情報であると考えている．公開されている口唇動画コーパスとしては，単語の発話や，語彙が制限されているなど評価のために強い制約下のもとに設計されていたり，音声の品質が非常に悪いことが多い．例えば，広く用いられている口唇動画コーパスである GRID[64] は 34 名の英語話者それぞれが 1 時間程度の発話を行っているが，発話文は 51 単語からランダムに 6 単語を組み合わせて構築されており，語彙が制限されたコーパスである．また，TCD-TIMIT[65] コーパスは 59 名のボランティアと 3 名のプロのリッ

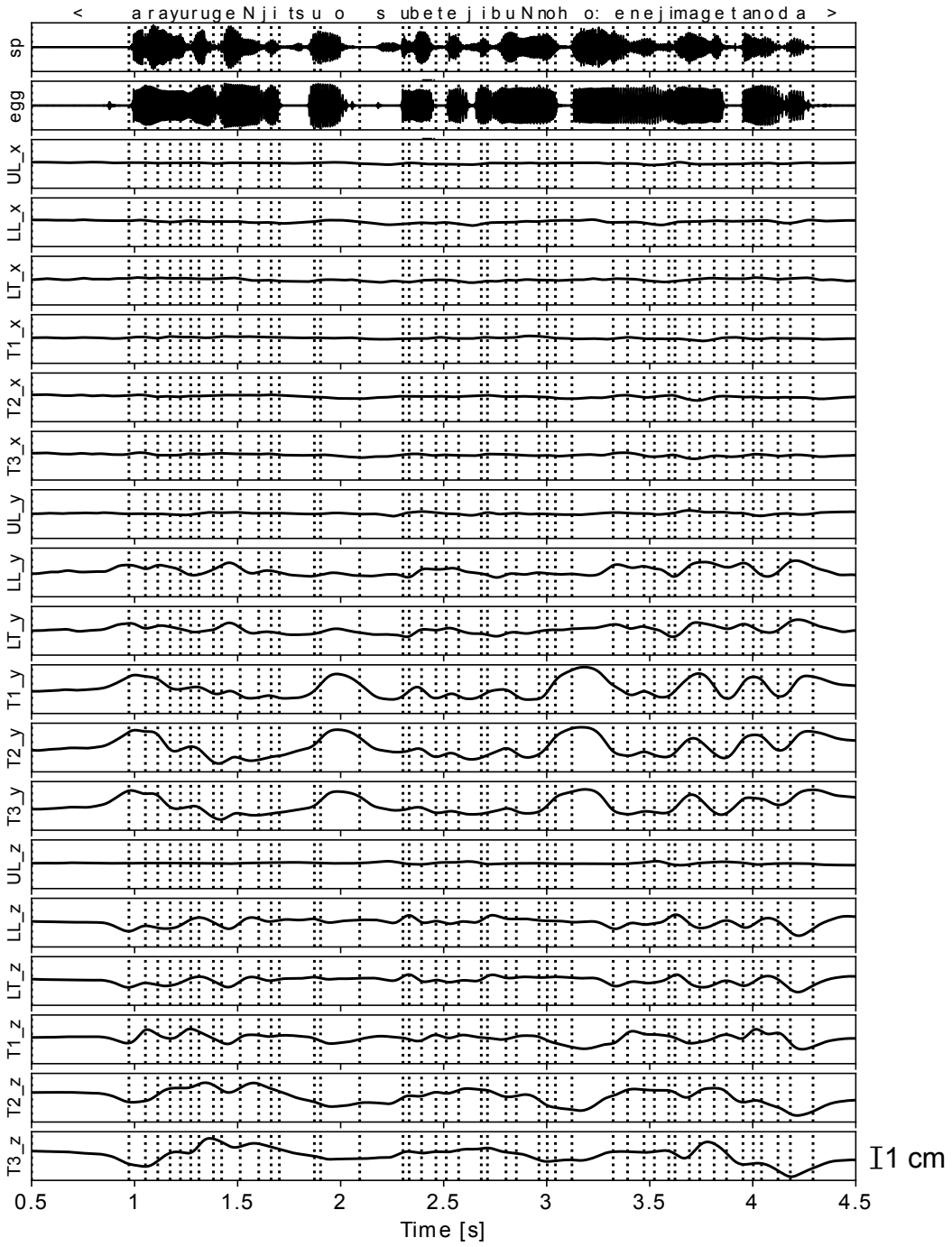


Fig. 4.5: 「あらゆる現実をすべて自分のほうへねじ曲げたのだ」と発話した際の収録データ

プスピーカが全 6913 文の TIMIT 文の一部をそれぞれ読み上げているが，収録されている波形の振幅のレベルが低く十分な SN が確保できていない．一方で，Youtube の動画をもとにした Lip2Wav コーパス [66] が公開されており，1 話者につき約 20 時間と非常に大規模で，頭部の固定等はとくにされておらず，非常に実環境に近い状態となっている．

そこで，この 2 者の間を埋めるような，数時間程度の発話時間で頭部の動きが少なく，かつ，音声の品質も高いような口唇動画コーパスが今後の研究のために重要であると考えている．よって，本研究ではこの条件を満たすような日本語口唇動画データの収集も行った．

4.3.2　測定条件

発話内容としては，ATR 音素バランス文，4.1.4 節で選択した音素バランス文，JSUT の basic5000 のうち前半の 2500 文の計 3887 文であった．収録は九州大学大橋キャンパス 3 号館 1 階の無響室で行った．マイクロフォン，マイクアンプ，EGG は 4.3.2 節と同じものを用い，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．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．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 2 者の同期を行った．話者はプロのナレータ男女 1 名ずつの計 2 名であった．

4.3.3　収録データの例

総発話時間は 1 話者につき約 4.8 時間となった．Fig. 4.6 に収録した口唇動画の 1 フレームを示している．十分な大きさで口唇を捉えることができている．また， Fig. 4.7 に口唇動画と同時に収録した「あらゆる現実をすべて自分のほうへねじ曲げたのだ」を発話した際の音声波形を示す． Fig. 4.7 を見ると良好な SN 比のもとで音声が適切な振幅レベルで記録されていることがわかる．

4.4　まとめ

本章では，調音情報の収録を行うために ATR 音素バランス文と同時に用いる新たな音素バランス文を構築した．この音素バランス文を用いることで，ATR503 文単体で用いるよりも多様な音素文脈を収録することが可能となった，さらに，この音素バランス文を用

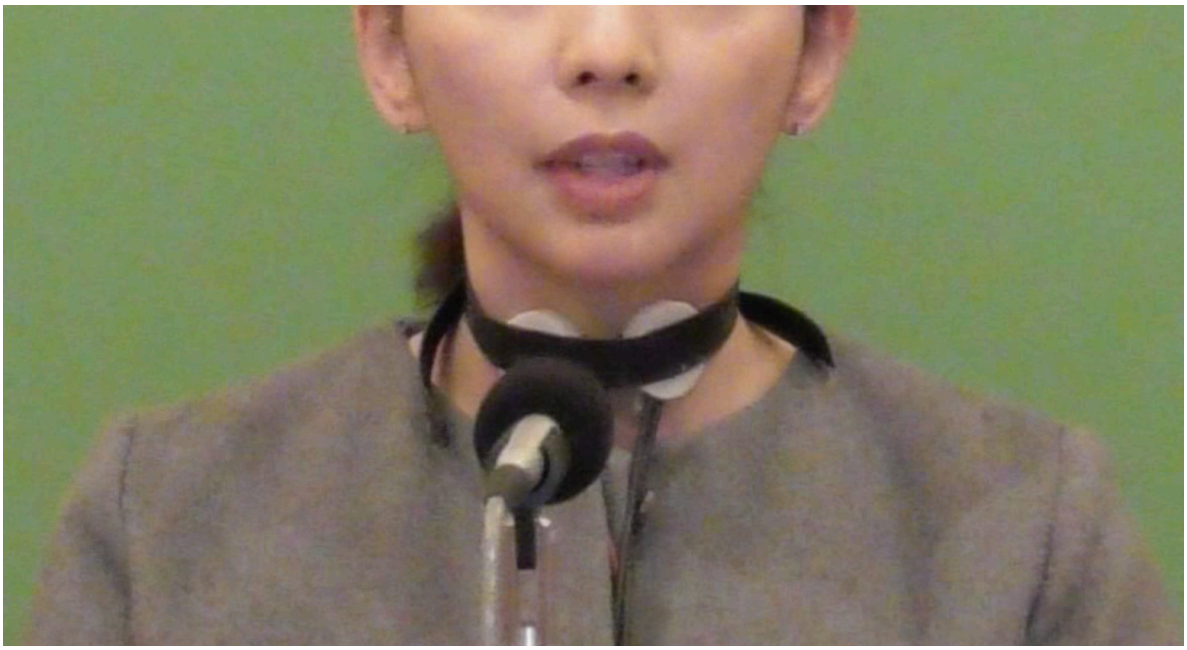


Fig. 4.6: 収録した口唇動画の例.

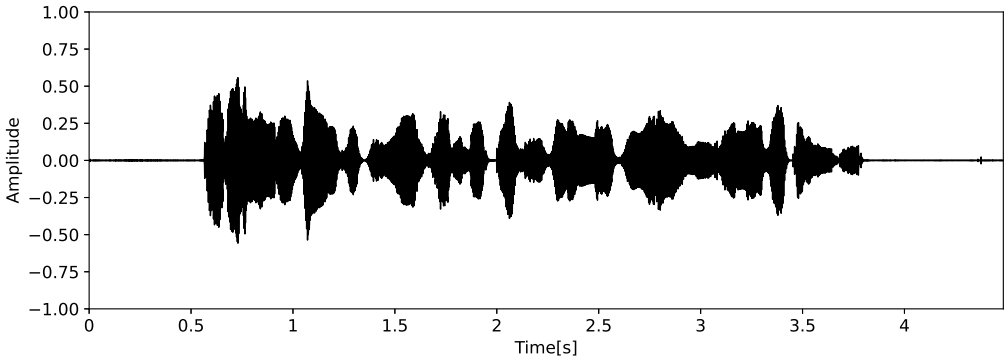


Fig. 4.7: 「あらゆる現実をすべて自分のほうへねじ曲げたのだ」と発話した際の音声波形.

いて，日本語の調音・音声パラレルデータの収集を行ったところ，3D-EMA に関しては男性 1 名で 67 分のデータ，口唇動画は男女 1 名ずつ各 4.8 時間のデータを得ることができた.

第 5 章

磁気センサによる調音–音声順変換の構築

EMA データから声道全体の形状を復元することは困難であるので，音声と EMA データを同時に記録した調音・音声パラレルコーパスを用いて，調音情報から音声特徴量を推定するデータ駆動型のアプローチがこれまで取られてきた．調音情報と音声特徴量の対を格納したコードブックをもとに，入力となる調音情報と最も近い調音情報をもつデータ対をコードブックから検索することで，声道のスペクトル包絡を得る手法 [9] が提案されている．しかし，このコードブックに基づく方法は，コードブックのサイズが大きくなるとデータ対を選択するために必要な計算量が増加する問題点がある．その他にも，混合ガウス分布 (Gaussian mixture model; GMM)[67]，混合密度ネットワーク (Mixture density network; MDN)[60]，feed-forward neural network [68], DNN[69, 70], LSTM[71] に基づく手法が提案されている．

これまでは，メルケプストラムなどの音声のスペクトル包絡を表すパラメータが音響特徴量として用いられることが多かった．これは，声道が音声の生成においてフィルターの役割を果たし，そのスペクトル特性を決定するためである．一方で，声道は破裂音や摩擦音を含む一部の子音の音源生成に深く関与しており，また，音声の音韻表現には声道の音響特性が関与し，さらに，音素文脈とピッチの時間的パターンや有声/無声判定などの音源情報は関連がある．これらのことから，音声の音源に関する特徴量と調音動作との間には，暗黙的かつ間接的な関係があるのではないかと予想される．音源情報の推定は検討されつつある [68, 71] ものの，これらの研究で使用された EMA コーパスは発話時間が 30 分以下と比較的小さいものであった．

本章では，磁気センサ (EMA) で得られた調音情報から，スペクトル包絡を表す特徴量だけでなく，音源に関するパラメータも同時に推定する調音–音声順変換を構築した結果について述べる．この変換モデルは，1 時間以上のデータが収録されている MNGU0 コー

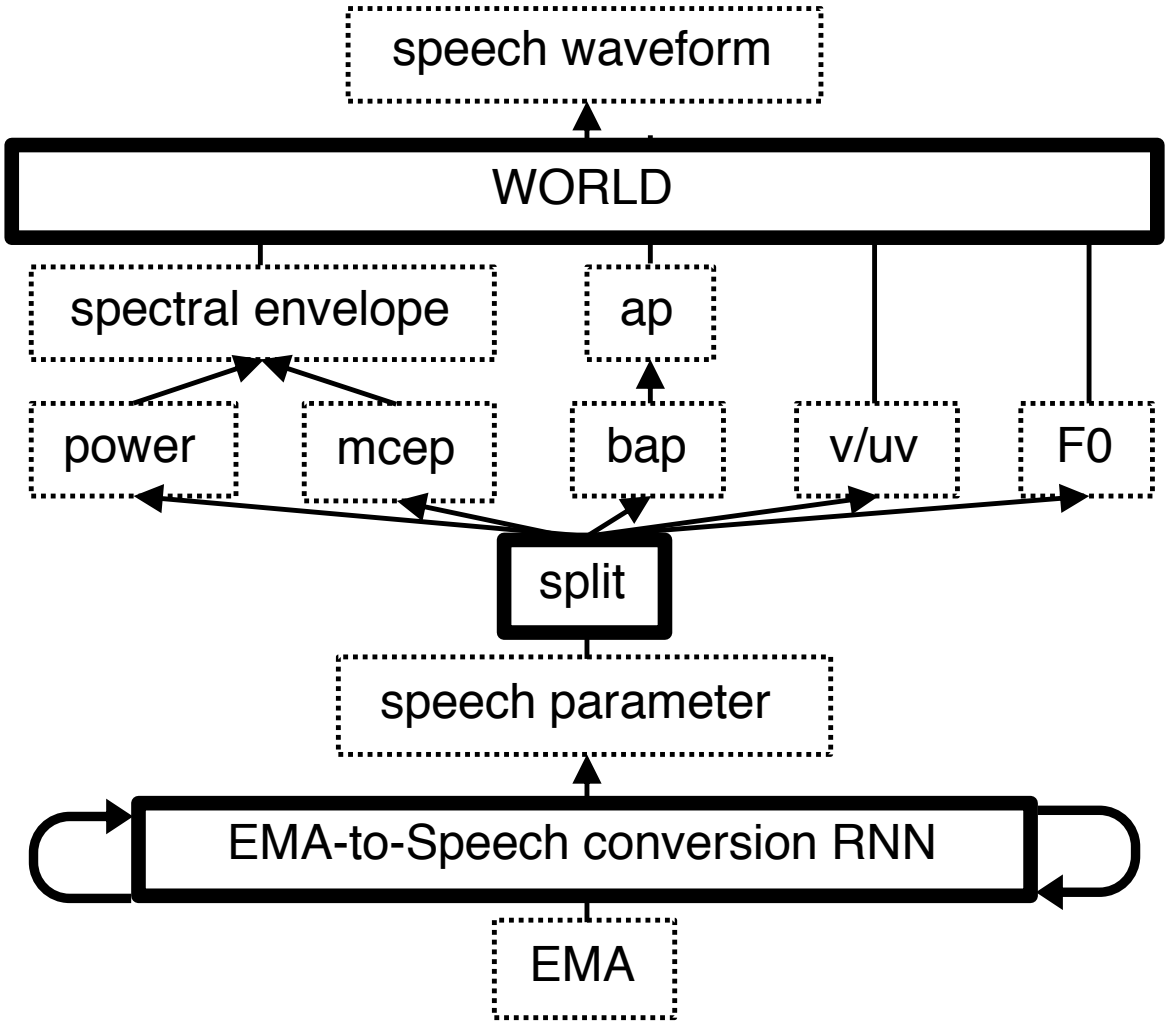
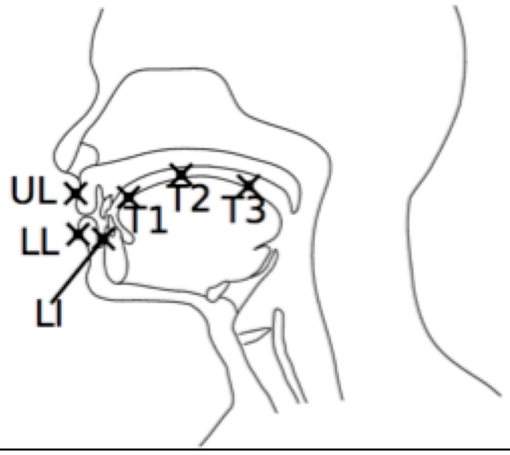


Fig. 5.1: 提案した調音-音声変換システムの概要図．ここで，”mcep”はメルケプストラム，”ap”は非周期性指標，”bap”は 5 帯域平均非周期性指標，”v/uv”は有声/無声判定を意味する．

パス [52] を用いて，双方向リカレントニューラルネットワーク (Bi-RNN) を学習させることによって構築した．また，提案法の有効性を検討するために，推定した音声特徴量の誤差に関する客観評価と，調音情報から合成された音声の自然性と了解性を調べる主観評価実験を行った．

5.1　調音情報からの音声合成法

Fig. 5.1 は，観測された調音運動から音声を合成する提案法を示す概要図である．図の下方にある”EMA-to-speech conversion RNN”は，入力された EMA データの順方向と逆方向の時系列を同時に考慮できる双方向 LSTM を組み込んだ構造である．この RNN は，EMA データの時系列が入力されると，音声パラメータの静的特徴量と動的特徴量の



label	location	label	location
UL	Upper lip	T1	Tongue tip
LL	Lower lip	T2	Tongue body
LI	Lower incisor (Jaw)	T3	Tongue dorsum

Fig. 5.2: 受信コイルの取り付け位置 (文献 [52] をもとに作成)

結合ベクトルを出力する．この結合ベクトルの各要素はおのおの独立であるという仮定をおいた上で，最尤パラメータ生成 (MLPG) アルゴリズム [31] を用いて静的・動的特徴量系列から滑らかな時間的軌跡を持つ音声特徴量系列を生成する．これらの結合された音声特徴量を後処理し，音声合成器として WORLD を駆動し，システム全体の出力として音声信号が得られる．パラメータ推定を繰り返して他のパラメータを推定するカスケード法 [72] が提案されているが，本研究では全ての音声特徴量を単一のネットワークで推定した．RNN は，音声特徴量を連結したパラメータベクトルに対する平均二乗誤差が最小となるように訓練される．本手法では，変換モデルを単一のネットワークで構築するため，訓練の手間やハイパーパラメータの複雑なチューニングを大幅に削減することができる．

5.2　実験

5.2.1　調音・音声パラレルコーパスの前処理

提案法においては，音声と EMA による調音運動の情報を同時に記録した調音・音声パラレルコーパスから得られる，調音・音声データ対を用いて，Fig. 5.1 に示すシステムの “ EMA-to-speech conversion RNN ” を学習した．評価には MNGU0 コーパス [52] を用いた．このコーパスは，男性話者 1 名の英語音声に関して 3 次元 EMA システムであるドイツ Carstens 社の AG500 によって収録された．サンプリング周波数 16 kHz の音声

データと，サンプリング周波数 200 Hz の EMA データが含まれている．コーパスにおいて学習 / 検証 / 評価用のサブセットが指定されており，その指定のとおりに分割したところ，発話時間はそれぞれ 60 分，3 分，3 分となった．EMA データには受信信号からの位置推定に失敗した NaN データが含まれていたため，NaN データの前後の値からスプライン補間により内挿することでこれらを取り除いた．受信コイルはすべて頭部の正中断面上に配置されている．上唇，下唇，下歯，舌 3 点の計 6 点 (Fig. 5.2) 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 次元のデータとした．

音声特徴量は WORLD の分析機能を用いて計算した，0〜40 次のメルケプストラム，対数連続 F0，有声/無声判定，5 帯域平均非周期性指標である．メルケプストラムについて，カットオフ周波数 50 Hz の FIR ローパスフィルタを forward-backward filtering を適用することで trajectory smoothing [73] を行った．ここで，音声の分析シフト長を EMA データのサンプリングレート (5 ms) と揃えることで，EMA データと音声特徴量が時間的に 1 対 1 に対応する．各フレームごとに決定された静的特徴量に加えて，調音，音声ともに 1 次導関数 (Δ 特徴量) を加えた静的・動的特徴量系列として時系列を陽に考慮した．なお Δ 特徴量は 2 次中心差分により近似的に求めた．コーパスに含まれる音素ラベルを用いて，発話の前後の無音区間は取り除いた．EMA 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．

5.2.2　RNN の学習

RNN の学習は，調音・音声データ対を用いて，RNN に調音データを入力し得られた出力と音声特徴量の間の損失を最小化することで行った．RNN は 3 層の全結合層と 2 層の双方向 LSTM，出力の全結合層からなる．全結合層 3 層は活性化関数が sigmoid 関数で layer normalization が適用されておりユニット数は 128 である，また，双方向 LSTM のユニット数は 256 である．

まず，音声の静的・動的特徴量の平均二乗誤差が最小となるように，incremental 法を用いて学習を行った．続いて，Minimum generation error (MGE) 学習を行った． この手法では，静的特徴量と動的特徴量の両方から音声パラメータの静的特徴量を生成し，生成された静的特徴量に関して平均二乗誤差を最小化するようにネットワークの学習を行った．学習の各段階で，最適化器としては Grave's RMSprop を用い，5.0 の gradient clipping を適用した．訓練を繰り返しながら，1 epoch ごとに検証データの損失を計算し，それが最小となった時点のパラメータを評価に用いた．

Table 5.1: 各手法おける音声特徴量の推定誤差.

	mcep [dB]	F0 [Hz]	vuv [%]	bap	power
GMM	5.176	13.74	14.09	0.1517	0.5139
MLP	4.843	13.66	**10.50**	0.1327	0.4698
RNN	**4.801**	**12.59**	10.55	**0.1263**	**0.4085**

5.2.3 先行手法との比較

提案法に加えて，GMM による最尤推定 [67] ならびに多層パーセプトロン (MLP) を用いた変換モデルを構築し，相互の比較を行った.

GMM による最尤推定 [67] は，入力となる時系列 $\boldsymbol{X}$ と出力となる時系列 $\boldsymbol{y}$ の静的・動的特徴量 $\boldsymbol{Y}$ の結合ベクトル $\boldsymbol{Z}$ に関して，GMM を最適化し得られたパラメータセット $\lambda^{(\boldsymbol{Z})}$ を用いて，与えられた入力時系列 $\boldsymbol{X}$ から出力時系列の推定値 $\hat{\boldsymbol{y}}$ は次のように得られる.

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P(\boldsymbol{Y}|\boldsymbol{X}, \lambda^{(\boldsymbol{Z})}) \tag{5.1}$$

この $\hat{\boldsymbol{y}}$ は EM アルゴリズムによって得られるが, ここでは suboptimum mixture sequence による近似 [74] を導入し解析的に $\hat{\boldsymbol{y}}$ を求めた．調音データは当該フレームとその前後 5 フレームの計 11 フレームを結合したベクトルを，主成分分析によって累積寄与率が 80% となるように次元削減を行ったものを用いた．出力となる音声特徴量は提案法と同じく静的特徴量と Δ 特徴量の結合ベクトルである．GMM の要素数は 256 とした.

提案法のほかに MLP を用いた変換モデルに関しても検討を行った．MLP の構造は隠れ層 5 層，ユニット数 512 であり，活性化関数として Leaky ReLU を用いた．また，隠れ層には Layer Normalization と 20% の Dropout を適用した．その他の学習手順は RNN と同じである.

5.3 結果と考察

5.3.1 客観評価の結果

本手法の推定精度を先行研究と比較するために，音声特徴量パラメータの推定誤差を Table 5.1 に示す．ここで，1 次から 40 次までのメルケプストラム (mcep) の誤差をメルケプストラム歪 [dB] で求めた．また，5 帯域平均非周期性指標 (bap)，0 次のメルケプス

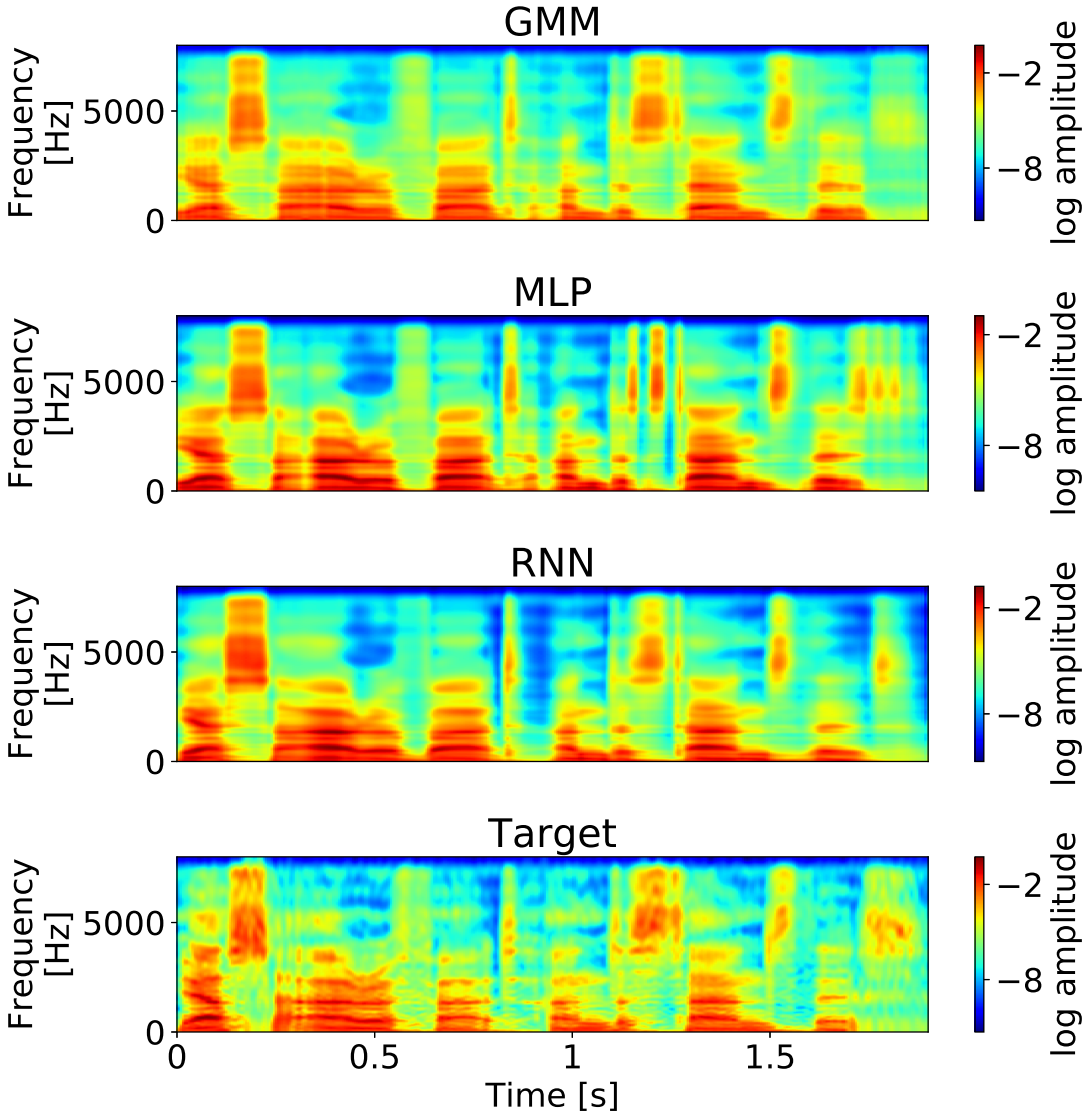


Fig. 5.3: MNGU0 のテストセットの 1 文”Yasser Arafat understands this”におけるスペクトル包絡の推定値と正解値.

トラム (power)，基本周波数 (F0) は平均 2 乗誤差根 (RMSE)，有声/無声判定 (v/uv) については誤り率 [%] によって評価を行った．GMM と提案法である RNN を比較するとメルケプストラム歪が 5.176 から 4.801 dB，基本周波数が 13.74 から 12.59 Hz，有声/無声判定が 14.09% から 10.55%，5 帯域平均非周期性指標が 0.1517 から 0.1267，パワが 0.5139 から 0.4085 とすべての特徴量の誤差を改善していることがわかる．一方，MLP と RNN を比較すると多くの特徴量で誤差が改善しているものの，有声/無声判定のみ 10.50% から 10.55% と微小に悪化していることがわかった．

MNGU0 のテストセットに含まれる 1 文 ”Yasser Arafat understands this.”でのスペクトル包絡を Fig. 5.3，基本周波数を Fig. 5.4 に示す．スペクトル包絡に関して GMM と深層学習による手法を比較すると，深層学習による手法の方が値の分散が大きく，フォ

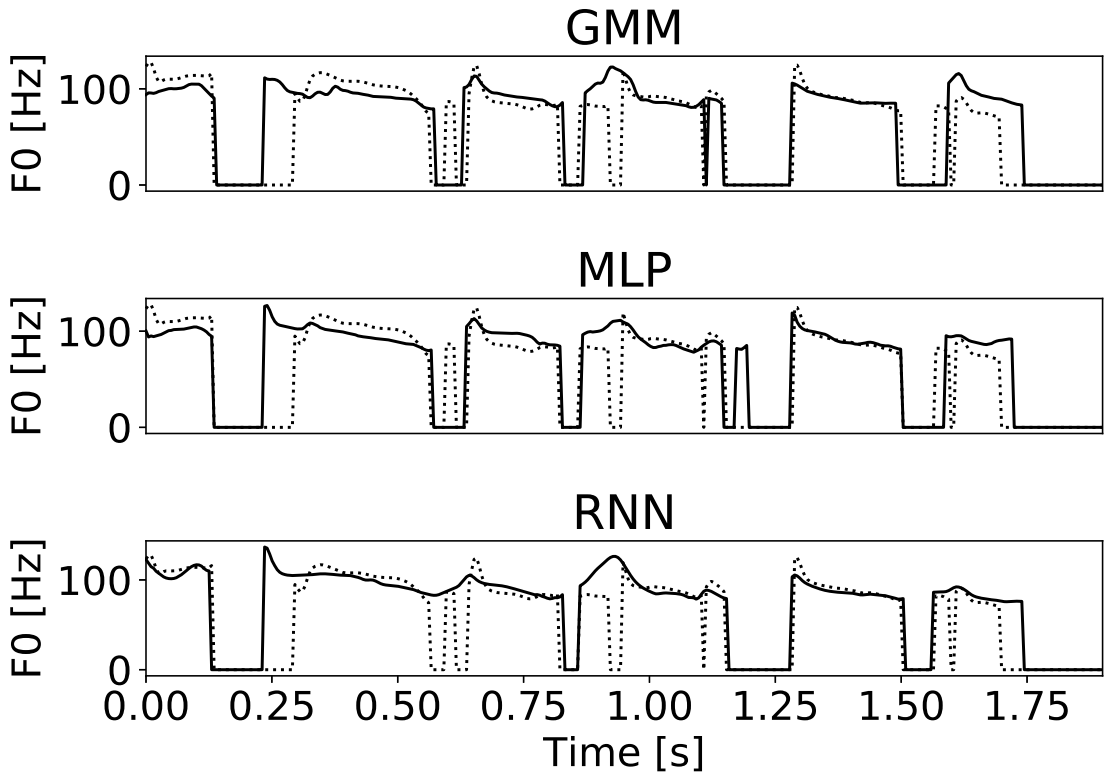


Fig. 5.4: MNGU0 のテストセットの 1 文"Yasser Arafat understands this"における F0 の推定値と正解値. 実線が各手法による推定値，破線が正解値を表す.

ルマント等もよく推定できていることがわかる．さらに，MLP と提案法を比較すると，RNN の方がより正確にスペクトル包絡を推定できている．基本周波数に関してはどの手法もよく推定できており，従来あまり行われたこなかった，有声音と無声音の調音的な差異や調音時系列に内包される言語的な特徴文脈をもとに基本周波数を推定することは，調音–音声合成において有用である．提案法と他の手法での基本周波数を比較すると，無声部分を有声に誤判定する事が多く見られ，これが有声/無声判定の誤り率にも反映されていると考えられる.

5.3.2 主観評価

次に，合成された音声の主観的評価を行うために，Amazon Mechanical Turk(mturk) を介して EMA データから生成された音声サンプルの自然性を 5 段階評価する MOS テストおよび発話内容の書き取りテストを行い，GMM，MLP，RNN，World による分析再合成音 (モデルの学習における目標値)，原音声の 5 条件についての比較を行った.

実験参加者

自然性に関する MOS テストに 50 名，発話内容の書き取りテストに 25 名が実験に参加した．すべての実験参加者はイギリスまたはアメリカから mturk を利用しており，実験後のアンケートで母国語が英語であると回答した．実験参加者には，(1) イヤホン，ヘッドホンまたはモニタースピーカの使用を強く推奨すること，(2) 音声サンプルは英語であり，文法的に正しいこと，(3) ノイズや歪みにより音声サンプルの品質が低下したことの 3 点について実験前に教示を行った．また，実験参加者は必要に応じて音声を聞き返すことが可能であった．

刺激

自然性の評価に用いたのは MNGU0 のテストセット全 65 文である．各実験参加者には，文章の重複なく，できるだけ均等に各手法の音声が含まれる 13 サンプルについて評価を行ってもらった．よって，全 325 個のサンプルそれぞれについて 10 名の話者が評価したことになる．13 個のサンプルはランダムな順序で 2 順提示され，1 順目を練習試行，2 順目を本試行として 2 順目の回答のみを結果の集計に用いた．

書き取りテストに用いたのは MNGU0 のテストセット全 65 文のうち固有名詞を含まない 25 文であり，このうち文章が最長と最短の 2 文章を練習試行に，残りの 23 文章を本試行に用いた．文章の重複なくそれぞれの手法による音声サンプルが 5 つずつの計 25 文章をランダム順で実験参加者に提示した．よって，それぞれのサンプルは，5 名の話者によって書き起こされたことになる．

結果

Table 5.2 に各手法における平均オピニオン評点を示す．GMM，MLP，RNN の三者に関して，Kruskal-Wallis 検定で有意差は出なかったものの RNN の MOS 値が 3.115 と最も高く，提案法を用いることでより自然な音声を合成できる可能性があることが示唆された．

Table 5.3 に各手法における単語誤り率 (Word error rate; WER) を示す．単語誤り率は次のように表される．

$$WER = \frac{S + D + I}{N}. \tag{5.2}$$

ここで，N は参照文の単語数，S は置き換わった単語数，D は削除された単語数，I は挿入された単語数を表す．WER に関しても GMM，MLP，RNN の三者に関して，Kruskal-Wallis 検定で有意差は出なかったが，深層学習を用いた手法は GMM の 24.554% よりも了解度の高い音声を合成できる傾向があり，その中でも MLP が WER で 17.071% とな

Table 5.2: 各手法の平均オピニオン評点と標準誤差.

GMM	2.769 ± 0.115
MLP	2.885 ± 0.111
RNN	$\mathbf{3.115} \pm 0.102$
分析再合成音	3.985 ± 0.082
原音声	4.246 ± 0.073

Table 5.3: 各手法の単語誤り率 [%] と標準誤差.

GMM	24.554 ± 2.939
MLP	$\mathbf{17.071} \pm 1.930$
RNN	19.186 ± 2.320
分析再合成音	2.733 ± 0.683
原音声	3.039 ± 0.864

り，最も単語誤り率が低い傾向にあるということがわかった．RNN と MLP を客観手法で比較すると，有声/無声判定のみ RNN に比べて MLP が精度が高いことから，本来有声であった部分が無声化，あるいはその逆の現象が起こったことで，了解性が低下したことが考えられる．

5.4　まとめ

本章では，調音器官の運動軌跡に基づいて，スペクトル包絡と音源情報を含む音声の特徴量を推定する新たな手法を検討した．その結果，本章で提案した手法は調音情報から音声を合成することが可能であることを明らかにした．MNGU0 コーパスから得られた調音・音声パラレルデータを用いて，調音情報から音声への変換を実現するための RNN を学習した．実験の結果，提案手法は従来の手法よりも優れた音声特徴量の推定が可能であることが示された．また，主観評価の結果，提案法は従来法に比べてより優れた自然性と了解性を持つ可能性が示唆された．

第 6 章

Residual Time-Delay Neural Network による音声–調音逆マッピングの構築

これまで，EMA によって取得された調音・音声パラレルコーパスを使用して，音声からマーカーコイル装着点の位置情報を調音情報として推定する逆マッピングモデルがさまざまに検討されてきている．逆推定における非一意性の解消法としては，舌や口唇などの調音器官の物理的な制約から，EMA によって測定される調音運動が連続でなめらかな軌道となることが，先験的な知見として利用できる．すなわち，逆マッピングモデルにおいては，入力される音声の特徴量の大局的な時間構造や文脈を陽に考慮し，いかに連続的な調音運動を再現するかが重要となる．初期には，調音情報と音響特徴量が対となったコードブックをパラレルコーパスから作成し，変換時には入力された音響特徴量の近傍の対データをコードブックから検索して，対応する調音情報を得る手法 [10] が検討された．さらには，隠れマルコフモデルを用いて音声の生成過程をモデル化し，事後確率最大化法を用いて調音情報を逆推定する手法 [75] や，混合ガウス分布 (Gaussian mixture model; GMM) を用いて変換モデルを構築する手法 [67] などが提案されてきた．近年では，Mixture density network [60]，Deep belief network [76]，再帰型ニューラルネットワーク (Recurrent Neural Networks; RNN) [77] など，深層学習を用いた手法も様々に検討されている．

調音観測手法として EMA を用いた場合の問題点としては，同一の話者に対しても，収録時期が異なると完全に同一の位置にマーカーコイルを装着することが難しいことが挙げられる．さらに，調音器官，特に舌の形状や大きさは話者によって異なるため，逆マッピングにおいてこれらの要因を一般化して扱うことは困難であることから，話者依存の逆マッピングが広く研究されている．また，調音器官にマーカーコイルを取り付けた状態で

発話を行うことは話者にとって大きな負担となることから，大規模なデータを得ることは音声のみの収録に比べてはるかに難しく，同一話者かつ同一収録の限りあるデータから安定なマッピングモデルを構築することが必要となる．他方，従来の逆マッピングモデルでは，音声の特徴量としてメルケプストラムなどのスペクトルの包絡特性を表す特徴量が用いられてきた．これは，音声生成の音源–フィルタ近似から考えれば，声道は音響フィルタの役割を果たすため妥当と言える．一方，一部の子音において，声道は摩擦性や破裂性の音源生成にも深く関わっていることや，声道の音響特性は音声の音韻性を担っており，さらにピッチパターンや有声無声などの音源情報は音韻情報と無関係ではないこと，および，先行研究 ([16] 等) から示唆される，喉頭と調音器官の生理的な相互作用を考えると，調音情報と音声の音源に関わるパラメータの間にも何らかの関係性があることが考えられる．この関係を利用できれば，音声からさらに精度良く調音情報を推定できることが期待される．

そこで本章では，長期の時系列を考慮しつつも，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，Residual Time-Delay Neural Network (Residual TDNN) を用いた音声–調音逆マッピングを提案するとともに，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．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，層数を大き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．また，音声特徴量に関しては，情報量削減，音源・フィルタ分離の有無による調音情報の推定精度の比較を行うことで，音源情報が音声–調音逆マッピングに及ぼす影響を検討するとともに音声–調音逆マッピングに最適な音声特徴量を検討する．手法および特徴量の評価には 3 次元 EMA を用いて収集された複数のコーパスを用いて，性別，言語，発話量が多様な 10 名の話者に関して話者依存マッピングモデルをそれぞれ構築し，その推定誤差を比較した．

6.1　Residual TDNN による音声–調音逆マッピング

深層学習に基づく音声–調音逆マッピングは，調整可能なパラメータを多量に持つ巨大な合成関数であるニューラルネットワークに音声の特徴量を入力し，得られる出力と入力に対応した調音データの間の平均自乗誤差を最小化するように，ニューラルネットワークが持つパラメータを誤差逆伝播法によって最適化することで構築される．

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ことが知られている．そこで本研究では，Residual Network のショートカットを持つ構造

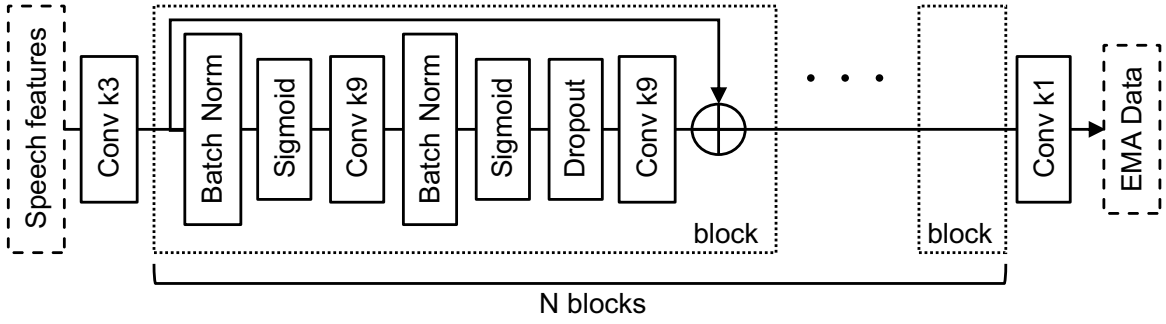


Fig. 6.1: Residual TDNN の概要図．”Conv kn”はフィルタ幅 n の畳み込み層を表す．

を TDNN に導入した Residual TDNN を提案した．Residual TDNN の概略を Fig. 6.1 に示す．Residual TDNN は，長期の時系列を考慮できる深い TDNN に対し，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．本研究で用いるのは pre-activation 型の Residual Network[78] を参考にしたものである．入出力の畳み込み層を除いた中間層は residal block を基本単位として構成される．block の中には 2 層の sigmoid 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ようなショートカットを持つ．また，residal block 内では正則化のために Dropout を加えている．

6.2 実験

6.2.1 音声・調音データ対

EMA コーパスとしては，5 章でも用いた MNGU0[52] に加えて，EMA-IEEE[62] とともに，4.3.2 節で収集した日本語のコーパスを用い，英語男性話者 5 名，英語女性話者 4 名，日本語男性話者 1 名の計 10 名に関してそれぞれ話者依存の音声–調音逆マッピングを構築し評価を行った．

MNGU0 については 5.2.1 節で述べたとおりである．EMA-IEEE[62] はカナダ NDI 社の WAVE によって収録された，男性話者 4 名，女性話者 4 名がそれぞれ IEEE 推奨発話文セット [79] を読み上げたコーパスである．サンプリング周波数 44.1 kHz の音声データと，サンプリング周波数 100 Hz の EMA データが含まれている．1 つの文章に関して，複数の話速で 1，2 回読み上げを行っているが，そのうち中速のものの最後の発話のみを取り出し，1 文章に対して 1 発話のデータベースとして用いた．IEEE 推奨発話文セットのうちのブロック 11 を検証用に，ブロック 12 を評価用に用いて残りを学習データとしたところ，発話時間はそれぞれ学習用 21 分，検証用 2 分，評価用 2 分となった．

また，日本語のコーパスについて，音声のサンプリング周波数は 50 kHz，EMA データのサンプリング周波数 は 250 Hz であった．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．

すべてのコーパスに対して，音声は 16 kHz に，EMA データは 100 Hz にリサンプリングを行った．さらに EMA-IEEE コーパスを参考に，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 EMA 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た．上唇，下唇，下歯，舌 3 点の計 6 点 (Fig. 5.2) 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 次元のデータとした．

音声特徴量については，音声の分析シフト長を EMA データのサンプリングレート (10 ms) と揃えることで，EMA データと音声特徴量が時間的に 1 対 1 に対応する．本研究では，高品質ボコーダによって音源・フィルタ分離を行うことによる調音データの推定精度の変化を含め，音声–調音逆マッピングに最適な特徴量についても検討した．用いた特徴量としては，(1) 321 次の対数振幅スペクトログラム，(2) 振幅スペクトログラムに 40 次のメルフィルタバンクを適用して得られた対数メルスペクトログラム，(3) 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 Mel-frequency cepstral coefficients (mfcc)，(4) REAPER[80] と CheapTrick[23] を用いて得られたスペクトル包絡にメルフィルタバンクを適用した 40 次の対数メルスペクトログラム，(5) 更にこれを離散コサイン変換によって次元削減した 24 次の mfcc の計 5 種類である．データベースに含まれる音素ラベルを用いて，発話の前後の無音区間は取り除いた．EMA 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．

6.2.2　各手法の設定

提案法に加えて，5.2.3 節で説明した GMM による最尤推定 [67] と，深層学習に基づく方法として MLP，RNN による逆マッピング [77] を構築し，相互の比較を行った．モデル構造やハイパーパラメータはそれぞれの手法について音声特徴量として mfcc を用いた MNGU0 データベースの検証用セットを用いて調節し，得られた最適値を他の全ての特徴量と話者に使用した．

GMM による最尤推定では，音声特徴量は当該フレームとその前後 5 フレームの計 11 フレームを結合したベクトルを，主成分分析によって寄与率 80% への次元削減と白色化を行ったものを用いた．調音データは静的特徴量と Δ 特徴量の結合ベクトルである．GMM の要素数は $\{1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$ から探索したところ，512

が最適であった．

深層学習に基づく方法として，提案法のほかに MLP と RNN [77] を用いた．共通する設定として，最適化器として Adam[35] を用い，学習イテレーション数は 50000 とし，Adam の 学習率 (alpha パラメータ) を学習の 50%，75% で 0.2 倍にした．勾配ノルム閾値は 3.0 とし，1.0e-6 の重み減衰を適用した．損失は平均自乗誤差である．学習時には各発話に関して固定長を設定し，発話のフレーム数が固定長より長ければイテレーションごとに固定長をランダムで切り出し，短ければ終端の値でパディングを行うことで入力フレーム数を固定長に揃え学習を行った．このとき，パディングしたフレームは損失の計算から除外した．その他の実験設定は Table 6.1 に示すとおりである．ミニバッチサイズは {4, 8, 16, 32, 64, 128, 256}, Adam の初期学習率は {0.01, 0.005, 0.001, 0.0005, 0.0001}, dropout 率は {0%, 20%, 50%, 70%}，固定時間フレーム長は {50, 100, 150, 200, 250, 300}，ネットワークの活性化関数は {sigmoid, tanh, ReLU, Leaky ReLU(LReLU)} から各手法ごとに最適値を探索した．さらに提案法に関しては，畳み込みチャンネル数は {32, 64, 128, 256, 512}，畳み込みカーネルサイズは {3, 5, 7, 9, 11, 13, 15, 17, 19, 21}, residual block 数は {10, 15, 20, 25, 30, 35, 40} から探索を行った．

MLP は 中間層に batch normalization を適用し，音声特徴量は当該フレームとその前後 5 フレームの計 11 フレームを結合したベクトルを用いた．隠れ層数は {1, 2, 3, 4, 5, 6, 7, 8}，ユニット数は {128, 256, 512, 1024, 2048, 4096} から探索を行った．RNN は全結合層 (Fully connected layer; FC) と双方向 Long short-term memory (LSTM) の積層から成る．全結合層には batch normalization を適用し，音声特徴量は当該フレームのみを入力した．全結合層数および LSTM 層数は {1, 2, 3, 4}，ユニット数は {16, 32, 64, 128, 256, 512, 1024} から探索を行った．

6.2.3 ローパスフィルタによる後処理

EMA によって収録された調音データは連続的でなめらかな軌道を持つことから，推定した調音データをローパスフィルタ (LPF) で後処理し，推定値にあらわれる軌道の不連続性を低減することで推定精度が向上することが知られている [67]．本研究では，逆マッピングの出力に対してタップ数 5 でカットオフ周波数 15 Hz の FIR ローパスフィルタを FIR フィルタの時間遅れが生じないように非因果的に適用した．また，MLP, RNN, 提案法に関しては順伝搬に LPF を組み込み，LPF を適用したあとの出力と正解値との間の損失が最小化されるように学習を行った [81]．

Table 6.1: 深層学習に基づく手法におけるハイパーパラメータ，ネットワーク構造とネットワークの総パラメータ数.

	MLP	RNN	Proposed
Mini-batch		32	
Learning rate	0.005	0.001	0.005
Dropout [%]		50	
Time frames	100	100	200
Network units	2048	256	128
Layers	4	2 FC + 2 LSTM	30 blocks
Kernel size	-	-	9
Activation	LReLU	LReLU	tanh
Parameters	17,362,956	2,709,516	8,873,484

Table 6.2: 音声特徴量 mfcc 時の各手法の推定誤差 (RMSE)[mm]

	MNGU0	Japanese	IEEE-M01	IEEE-M02	IEEE-M03
GMM	1.050	1.343	2.695	2.187	2.054
MLP	1.000	1.261	2.560	1.983	1.953
RNN	0.8639	0.9591	2.706	1.977	1.828
proposed	**0.7510**	**0.8903**	**2.360**	**1.727**	**1.628**
	IEEE-M04	IEEE-F01	IEEE-F02	IEEE-F03	IEEE-F04
GMM	1.753	2.215	2.147	2.172	2.117
MLP	1.559	1.937	2.074	1.994	1.852
RNN	1.551	1.870	1.879	2.000	1.737
proposed	**1.260**	**1.619**	**1.716**	**1.799**	**1.443**

6.3　結果と考察

6.3.1　提案法の有効性に関する検討

構築した音声–調音逆マッピングを用いて評価用データの推定を行い，推定された EMA データと正解値の EMA データとの間の二乗平均平方根誤差 (Root Mean Square Error;

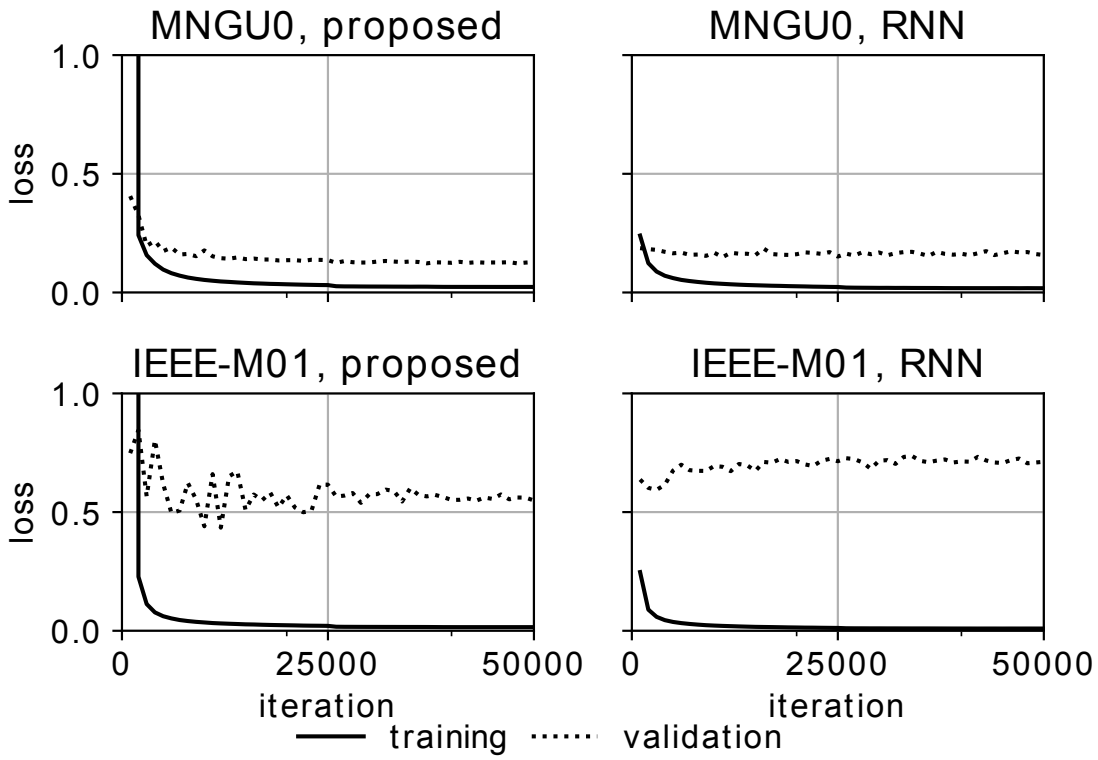


Fig. 6.2: 特徴量 mfcc における MNGU0 と IEEE-M01 の学習曲線.

Table 6.3: 特徴量 mfcc における MNGU0 使用時の Residual Block 数を変化させた場合の RMSE.

blocks	5	10	20	30	40
RMSE[mm]	0.8151	0.7820	0.7799	0.7510	0.7507

RMSE) を計算することで評価を行った． Table 6.2 に音声特徴量として mfcc を用いたときの RMSE を各手法，各データベースごとに示す．ここで，話者の Japanese は我々が収集した男性話者， IEEE-M01〜04 は EMA-IEEE コーパスの男性話者，IEEE-F01〜04 は女性話者を表す．提案法は話者にかかわらず他の手法より低い RMSE 値を示し，EMA データを良好に推定できていることがわかる．GMM を基準に RMSE の変化率

$$\frac{\text{その手法での RMSE} - \text{GMM の RMSE}}{\text{GMM の RMSE}} \times 100 \quad [\%] \tag{6.1}$$

の話者平均を求めると，提案法が GMM に対して 24% 減少，RNN が 13% 減少，MLP が 7.8% 減少となり，GMM よりも深層学習に基づく手法のほうがより少ない誤差で調音

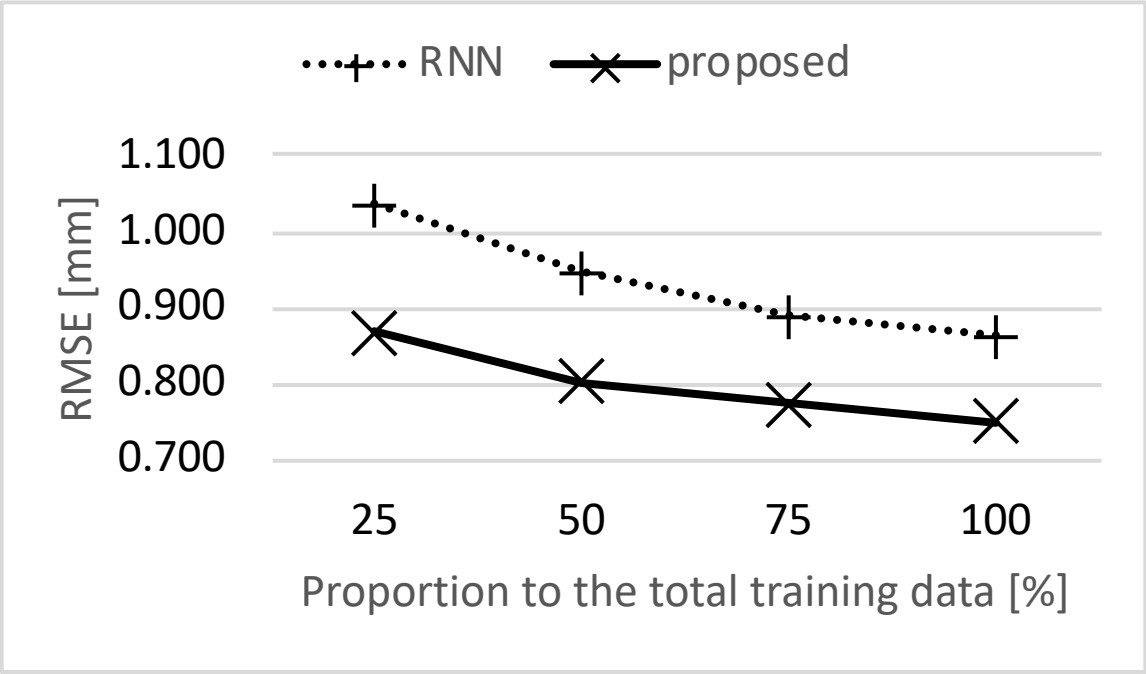


Fig. 6.3: 特徴量 mfcc における MNGU0 の学習データの使用率を変化させたときの推定精度の変化

データを推定できるということがわかった．この結果は，音声–調音逆マッピングには主成分分析と GMM の組み合わせではカバーすることのできない複雑な非線形性に対処することが必要になることを示唆している．

提案法に次いで RMSE が減少する傾向にあった RNN に関して，話者 IEEE-M01 と IEEE-F03 に関しては他の手法よりも RMSE が増加した．Fig. 6.2 に音声特徴量として mfcc を用いたときの提案法と RNN について，学習 (training) データと学習外検証 (validation) データとして MNGU0 と IEEE-M01 を用いたときの学習曲線を示す．MNGU0 においてはいずれの手法も学習データと検証データともに損失が減少し収束が得られている一方で，IEEE-M01 に関しては，RNN は学習データに対する損失が減少する一方で，検証データの損失が微小に増大する過学習の状態にあったことがわかる．一方，提案法では，RNN よりも学習するパラメータ数が 3 倍以上も多いにもかかわらず，学習前半には検証用データに対する損失が増減しているものの，学習後半では損失は減少し収束しており過学習は見られなかった．これは，MNGU0(学習データが 60 分) でハイパーパラメータの調整を行い，それをそのまま EMA-IEEE(学習データが 20 分程度) に適用したため，学習データの減少によりネットワークを十分に学習できなかったことが原因であると考えられる．提案法の学習データの量に対する頑健性を調べるため，

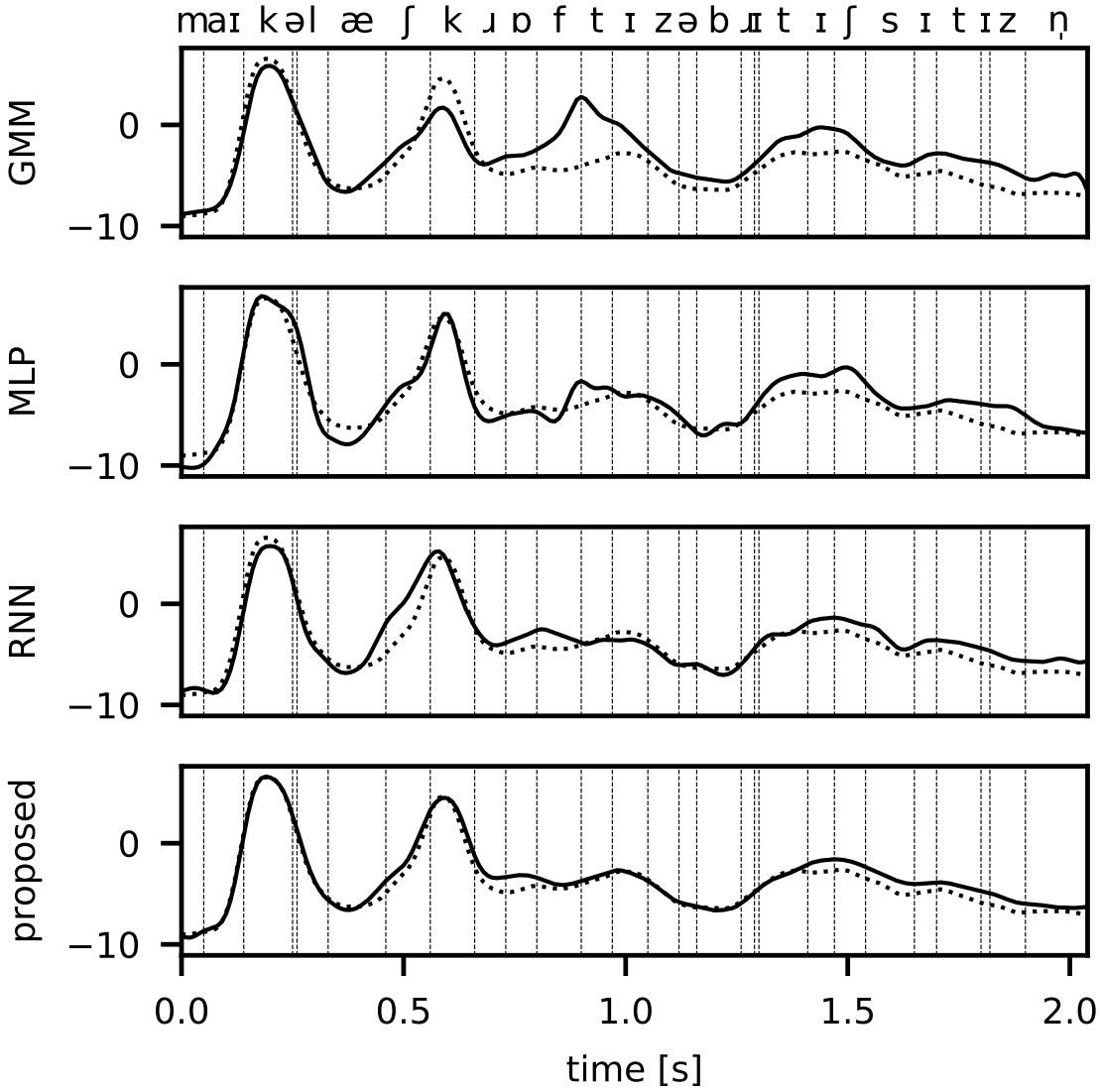


Fig. 6.4: MNGU0 のテストセットの 1 文 ”Michael Ashcroft is a British citizen.”において推定された MNGU0 の後舌の鉛直方向の軌道 [mm]．実線が推定値，破線は正解値を表す．

音声特徴量として mfcc を用い MNGU0 データベースに含まれる学習データの使用率を 25% から 100% まで 4 段階に変化させた．提案法と RNN について RMSE を求めたところ Fig. 6.3 のようになり，学習データ量によらず RNN よりも提案法のほうが優れた精度で推定可能であり，学習データが少なくなるほど RNN と提案法の間で性能差が広がることがわかった．

また，提案法に関してネットワークのブロック数を変化させた場合の RMSE を Ta-

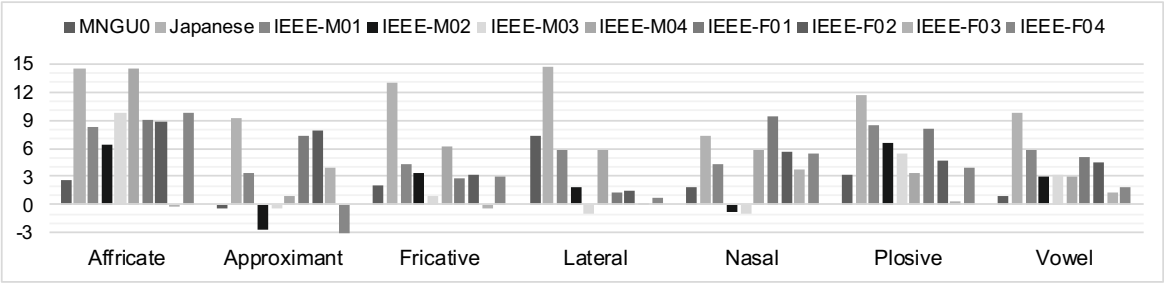


Fig. 6.5: mfcc(env) に対する mfcc の調音方法ごとの RMSE の改善率 [%].

ble 6.3 に示す．最適となった 30 ブロックよりもブロックを減らすと推定精度が悪化したのに対して，10 ブロック (畳み込み層 20 層) を余分に追加した 40 ブロックの場合は 30 ブロックのときと比べて大きな変化は見られなかった．このことから提案法では，ブロック数が最適値よりも多く設定されている場合に関して頑健に推定が可能であることが示唆された．

このように，提案法は話者および学習データの量に対して頑健な傾向が見られたことから，ハイパーパラメータを学習データごとに調節する必要性が低く，限られたデータ量のなかで推定精度の良い逆マッピングを構築可能であることを示唆している．

Fig. 6.4 に音声特徴量として mfcc を用い MNGU0 から構築された逆マッピングによって推定された，英語文章”Michael Ashcroft is a British citizen”発話時の舌後部に取り付けられた受信コイルの鉛直方向の軌道を示す．どの手法でも LPF による後処理によりなめらかな軌道が得られていることがわかる．ただし，MLP には他の手法に比べて細かな変動が見られる．これは，LPF に加えて，GMM では最尤推定の中で，RNN と提案法ではネットワーク構造によって出力の文脈を考慮する一方，MLP では LPF 以外に出力の時系列の滑らかさを反映していないことが原因である．また，最も滑らかな軌道が得られているのが提案法であり，これは，TDNN の畳み込み演算と，Residual Network のアンサンブル的な振る舞い [82] が連続的で滑らかな EMA データの表現に適しているからであると考えられる．

6.3.2　音声–調音逆変換に最適な特徴量の検討

Table 6.4 に音声特徴量として mfcc を用いたときの RMSE を各手法，各データベースについて示す．ここで，mspec は対数メルスペクトログラム，spec は対数振幅スペクトログラム，mfcc(env) および mspec(env) は音声信号から抽出したスペクトル包絡を元に特徴量を算出したことを表す．IEEE-M01 をのぞいて，mfcc を用いたときに最も RMSE

Table 6.4: 提案法における各音声特徴量を用いた際の推定誤差 (RMSE)[mm]

	MNGU0	Japanese	IEEE-M01	IEEE-M02	IEEE-M03
mfcc	**0.7510**	**0.8903**	2.360	**1.727**	**1.628**
mspec	0.7991	0.9338	2.466	1.750	1.678
spec	0.8400	0.9308	**2.205**	1.769	1.791
mfcc(env)	0.7658	0.9677	2.579	1.802	1.656
mspec(env)	0.8161	0.9777	2.556	1.795	1.707
	IEEE-M04	IEEE-F01	IEEE-F02	IEEE-F03	IEEE-F04
mfcc	**1.260**	**1.619**	**1.716**	**1.799**	**1.443**
mspec	1.332	1.663	1.779	1.835	1.505
spec	1.448	1.812	1.913	1.944	1.648
mfcc(env)	1.324	1.706	1.807	1.813	1.493
mspec(env)	1.357	1.740	1.831	1.846	1.502

が小さくなった事がわかる．mfcc を基準に RMSE の変化率

$$\frac{\text{各音声特徴量での RMSE} - \text{mfcc での RMSE}}{\text{mfcc での RMSE}} \times 100 \quad [\%] \tag{6.2}$$

の話者平均を求めると, mspec については 3.9%, spec 8.3%, mfcc(env) 4.6%, mspec(env) で 6.4% の増加が生じた．まず, mfcc, mspec, spec の三者に着目すると, この順で RMSE が増加している．これは特徴量の次元の多さと対応している．特徴量の次元の増加は，学習するパラメータの増加を意味する．一般に，より多くのパラメータを十分に最適化するためには，より多くの学習データを必要とする．次元が少なくなるほど RMSE が改善されたことは，学習データの量が限られた音声–調音逆マッピングの構築において，従来広く用いられてきたメルフィルタバンクおよび離散コサイン変換による特徴量の次元削減が依然として有効であることを意味する．

また，次元数が同一で音源・フィルタ分離を用いたか否かが異なる mfcc と mfcc(env), あるいは mspec と mspec(env) の比較において，いずれも音源・フィルタ分離を行わないほうが RMSE が小さくなった．これには 2 つの理由が考えられる．1 つは音源の情報が音声–調音逆マッピングにおいて有用な特徴であったということである．EMA データの各次元ごとに音声特徴量として mfcc(env) を用いたときを基準として，mfcc を用いたときの各調音方法ごとの RMSE の改善率

$$-\frac{\text{mfcc での RMSE} - \text{mfcc(env) での RMSE}}{\text{mfcc(env) での RMSE}} \times 100 \quad [\%] \tag{6.3}$$

を計算し，EMA データの次元で平均したものを Fig. 6.5 に示す．改善率が正なら，音源情報を用いる (音源・フィルタ分離を行わない) ことで RMSE が改善されていることを意味する．多くの話者，調音方法に関して，音源情報を用いることで RMSE が改善されていることがわかる．話者平均の改善率を見ると，破擦音，破裂音，鼻音，摩擦音の順で改善率が大きい．これらは，声道の極端な狭めや閉鎖をともなう子音であり，音源情報はこれらの子音の推定において大きな手がかりとなった．

もう一つは，スペクトル包絡の分析手法の雑音に対する頑健性の問題である．EMA の収録の際には，送信コイルが生成する磁界の中で音声を収録することになる．音声の収録に用いるマイクロホンがこの磁界の影響を受け，収録音声に定常的な雑音として現れる．振幅スペクトルや mfcc であれば，特徴量の標準化で定常雑音の影響はある程度抑制されていると考えられるが，スペクトル包絡を用いた特徴量においては，F0 推定法や包絡抽出法が雑音の影響を受けてスペクトル包絡が適切に抽出されなければ，それが推定精度の悪化につながると考えられる．

6.4　まとめ

本章では，Residual TDNN を用いた音声–調音逆マッピングを提案し，3 次元 EMA を用いて収集された複数のコーパスを用いて，英語男性話者 5 名，英語女性話者 4 名，日本語男性話者 1 名の計 10 名に関してそれぞれ話者依存の音声–調音逆マッピングを構築し評価を行った．推定誤差を RMSE によって比較したところ，評価に用いたすべての話者において提案法が優れた精度で EMA データの推定が可能であった．また，提案法において複数の種類の音声特徴量について推定精度の比較を行ったところ，次元削減は逆マッピング問題において有効であり，音源に関する情報は，声道の閉鎖や極端な狭めを伴うような調音時の推定精度を改善し推定精度の向上に貢献することがわかった．

第 7 章

口唇動画を用いた調音-音声変換の構築

口唇動画は他の調音観測手段と異なり，普及した端末で容易に収録可能であることから，実応用に近い調音情報として，口唇動画からのテキスト認識や，音声合成の研究も行われている．文献 [83] では口唇動画と超音波エコーによる舌動画から画像特徴量を抽出し，画像特徴量から HMM によって音素を推定し，推定した音素と画像特徴量から波形接続 HMM によって音声を合成する手法が提案されている．この手法では，実際に了解性のある音声は合成できず，口唇動画と舌動画では舌動画の寄与がほとんどであることが報告された．文献 [84] では口唇動画から離散コサイン変換によって抽出した画像特徴量を用いて，LSTM によって STRAIGHT による音声特徴量の推定が行われた．文献 [85] では二次元の畳み込みニューラルネットワークを用いて，口唇動画とそのオプティカルフローから振幅スペクトルが推定された．また，読唇 (Lip-to-text) において時空間畳み込みの有効性が確認 [86] されてからは，時空間畳み込みを用いたネットワークを用いた手法も数多く提案されている [87, 88, 89, 66].

しかしながら，音声合成の研究に利用できるようなクリーンな音声が収録された口唇動画コーパスは数が少なく，現在広く用いられている GRID コーパス [64] は非常に語彙が制限されたコーパスであり，このコーパスでシステムを評価しても，実際に実応用にシステムを拡張できるかは不明である．一方で，Youtube の動画をもとにした Lip2Wav コーパス [66] が公開されている．1 話者につき約 20 時間と非常に大規模であり，頭部の固定等はとくにされていないため非常に実環境に近い口唇動画データであるが，音声の品質も様々であり，参照信号が必要な各種客観評価指標への適用可能性について疑問が残る．

本章では，畳み込み層を基底とした系列変換モデルを用いた口唇動画–音声変換法を提案するとともに，4.3 節で収集した，比較的長時間の大語彙連続発話で，良好な SN 比を持つ日本語の口唇動画・音声パラレルデータを用いて，提案法の客観評価を行った．さら

に複数話者モデルについての検討も行った.

7.1 畳み込み層を基底とした系列変換モデルを用いた口唇動画–音声変換

提案法の概略を Fig. 7.1 に示す．提案法は，口唇動画から特徴を抽出する Encoder と，出力となるメルスペクトログラムを自己回帰推定する Decoder からなる.

Encoder への入力は RGB の口唇動画である．まず，動画処理に有効な時空間 Residual network 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 Residual network の出力は空間方向に平均化し時間 × 次元の 2 次元に次元削減を行い，Encoder の出力となる特徴量マップとした.

Decoder は因果的畳み込み層の積層により，Encoder の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 Fig. 7.2 に示した Gated linear unit(GLU) block [90] である．この構造は LSTM など近年のゲート付き再帰構造の影響を受けており，因果的畳込みの出力が 2 つに別れ，片方がそのまま出力へ，もう片方が Sigmoid 関数を適用することでゲートとして機能している．ゲートを適用した直後に Residual connection が導入されており，勾配消失が起きづらくなっている．Residual connection の合流の後に，Encoder の出力である feature map がフィルタ幅 1 の畳み込みによって次元数を調整されて加えられる．これによって再帰推定の中で，口唇動画の情報が反映される．前時刻のメルスペクトログラムを処理する prenet は，2 層の活性化関数 ReLU の全結合層であり，50% の Dropout が適用されている．この prenet 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.

post-net[91] 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，Batch Normalization と 50% の Dropout 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，post-net に入力する前の Decoder 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 post-net 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の和である.

また，口唇動画からより良いメルスペクトログラムを得るために次の 4 つについて検討を行った.

1 つ目は，動的特徴を考慮した損失である．メルスペクトログラムは時系列なので，損失においてその動的特徴の誤差を考慮することで，より品質の高い音声が得られることが

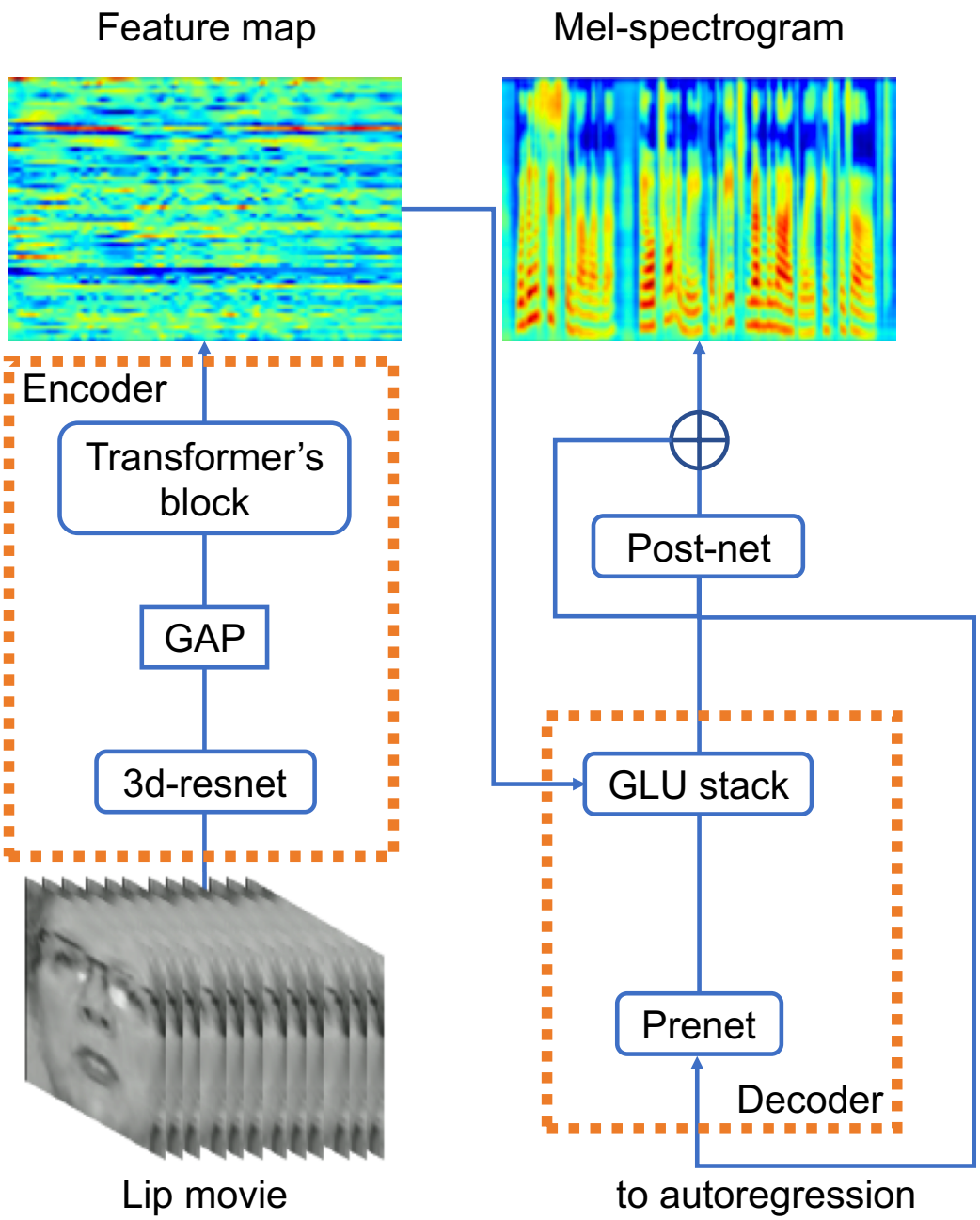


Fig. 7.1: 提案法の概要図.

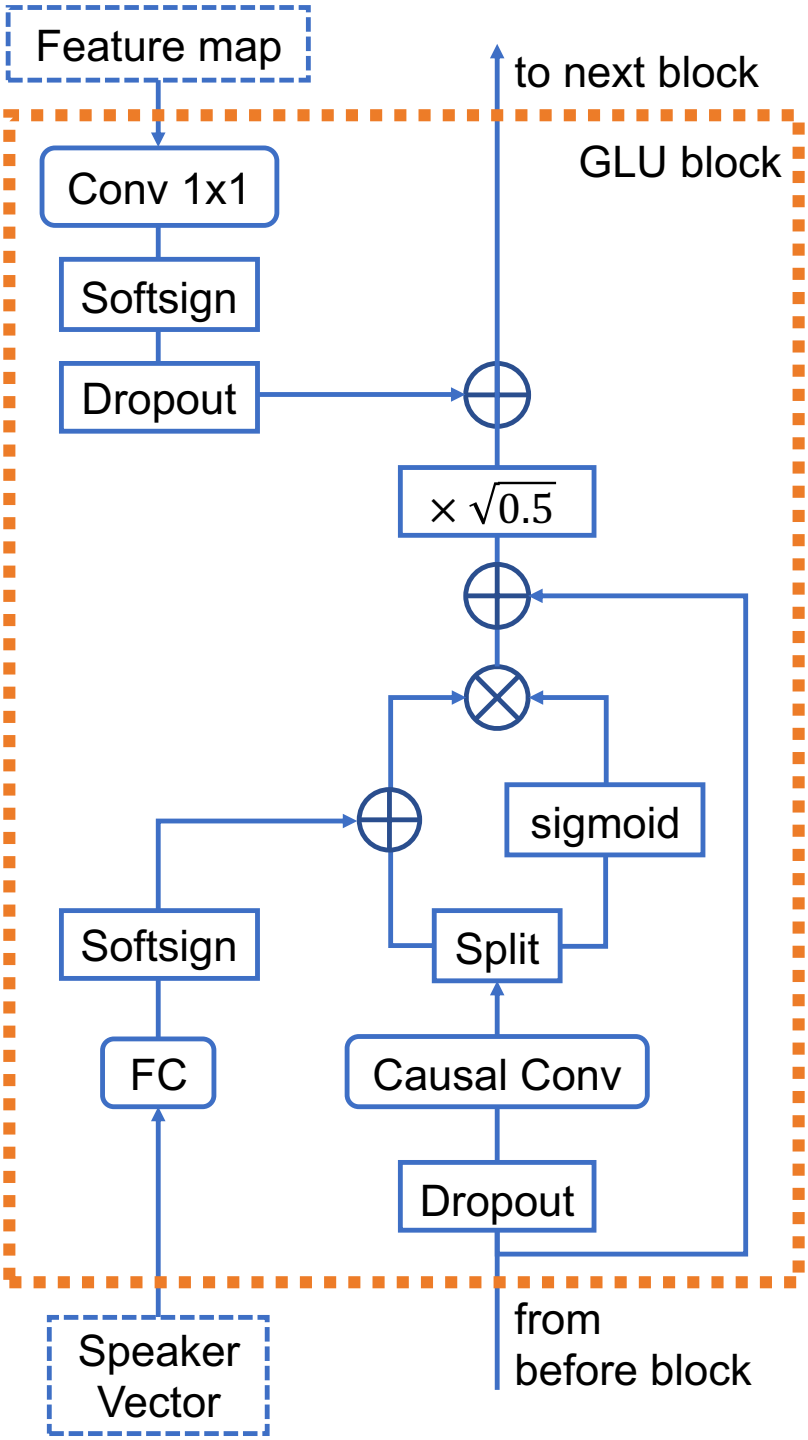


Fig. 7.2: Gated linear unit block の構造.

期待できる．そこで本研究では，ネットワークの学習時に出力されたメルスペクトログラムからその動的特徴量を計算しその誤差を最小化することを検討した．MGE 学習と異なるのは，ネットワークの出力は静的特徴量のみである点である．具体的には次の項を損失に加える．

$$E\left(\left(\frac{Wy - Wt}{\sigma}\right)\right) \tag{7.1}$$

ここで，t はメルスペクトログラム系列，y はネットワークの出力系列，W は静的特徴量系列を静的・動的特徴量系列に変換する行列，σ は学習データ全体の Wt の各次元ごとの標準偏差である．σ は学習時に逐次的に計算を行っていく．(7.1) 式は言い換えれば，正規化された，メルスペクトログラムの静的・動的特徴量系列の平均自乗誤差である．

2 つ目は，Transformer's block(Fig. 3.9) の導入である．畳み込みニューラルネットワークによって考慮できる時系列は，畳み込みのカーネルサイズによって制限される一方で，音声には発話全体にかかる長期の特徴がある．そこで，Encoder の出力である特徴マップに対して Scaled dot-product self attention を含む Transformer's block を導入することで，発話全体の文脈を考慮可能になることが期待できる．Transformer's block に入力する前に位置エンコーディングを加える必要があることに注意が必要である．

3 つ目は，音源情報に関するマルチタスク学習の導入である．調音情報には音源情報 (基本周波数，有声/無声，音声の強さ) が陽に含まれていない．そのため，調音–音声マッピングにおいては，音源情報は言語的な文脈，あるいは調音器官 (主に舌) と声帯がある喉頭との生理的な非独立性をもとに推定されると考えられる．その一方で，口唇動画は舌の動きを捉えることができないため，音源情報の推定が他種の調音情報よりも難しくなることが予測される．そこで，Decoder 出力を分岐し，音声の対数基本周波数とパワを推定するネットワークを同時に学習するマルチタスク学習を導入することで，音源情報の推定精度を向上させることを検討した．

最後に，Scheduled Sampling for Transformers[92] の導入である．本手法では畳み込みニューラルネットワークを用いていることから，学習時の自己回帰の入力にメルスペクトログラムの 1 時刻前の正解値を入力する teacher-forcing によって高速な学習が可能になる一方で，推論時には誤差の累積によって出力が大きく劣化する可能性がある．Scheduled Sampling for Transformers は一度 teacher-forcing で順伝搬を行い，得られた出力を前時刻の正解値のメルスペクトログラムと混合し，自己回帰の入力を劣化させることで，誤差の累積の緩和を図る手法である．Scheduled Sampling for Transformers の導入によって，ネットワークが誤差の累積に対して頑健になるか検討を行った．

7.2　口唇動画–音声変換の複数話者モデルの検討

本研究では，1 つのネットワークを 2 人の話者のデータで学習を行う複数話者モデルについても検討した．複数話者口唇動画–音声変換モデルの概要を Fig. 7.3 に示す．口唇動画から Encoder によって抽出される特徴量マップは言語的な特徴を表している．Decoder は Encoder が出力した言語的特徴と話者表現をもとにメルスペクトログラムを自己回帰推定するというのが提案する方法である．話者表現はベクトルとして得られ，Fig. 7.2 に示すように GLU block の中で隠れ表現に加算される．口唇動画における顔の個人性と音声の個人性を対応付けるアプローチも考えられるが，この方法は学習データに含まれない話者に対しての拡張性が不明であり，提案法を今後データベースに含まれない話者に拡張することを検討しているため不採用とした．一方，音声から話者の個人性を抽出し，それをもとに未知話者についての音声を合成することは可能 [93] であるため，話者情報を別に与えることとした．ただし，本研究では，2 名の話者を同じネットワークで学習するにとどめた．よって合成できるのも学習に用いた 2 名の話者のみである．話者表現はニューラル自然言語処理における単語の表現方法と同様に，各話者に対応する表現ベクトルをネットワークの学習の中で獲得する手法を用いた．

提案法では，口唇動画に含まれる個人性はメルスペクトログラムの推定において有用な情報ではない．そこで，本研究では，ドメイン敵対学習 [94] によって，Encoder において話者性を抽出しないような拘束条件を導入することを検討した．Encoder の時空間 resnet の出力を分岐させ，それに 2 層の全結合層を適用することで口唇動画から話者を認識するマルチタスク学習を行った．ここで，時空間 resnet と 2 層の全結合層の間に勾配反転層を追加する．勾配反転層は文字通り勾配の正負を入れ替える演算を行う．勾配反転層の導入によって，口唇動画から話者を認識するマルチタスク学習において，時空間 resnet はその損失を最大化するように学習が行われる．つまり，時空間 resnet の出力は話者性に関して有用な情報を持たないような出力をするようになる．また，同様の手法を Decoder の prenet の出力に対しても適用した．この手法を導入することによって，推定されるメルスペクトログラムにどのような影響があるか検討を行った．

7.3　実験

7.3.1　音声・調音データ対

口唇動画コーパスとしては 4.3 節で収集した男女 1 名ずつ，1 話者につき約 4.8 時間の日本語音声のデータを用いた．読み上げ文のうち ATR 音素バランス分の J セットを評価

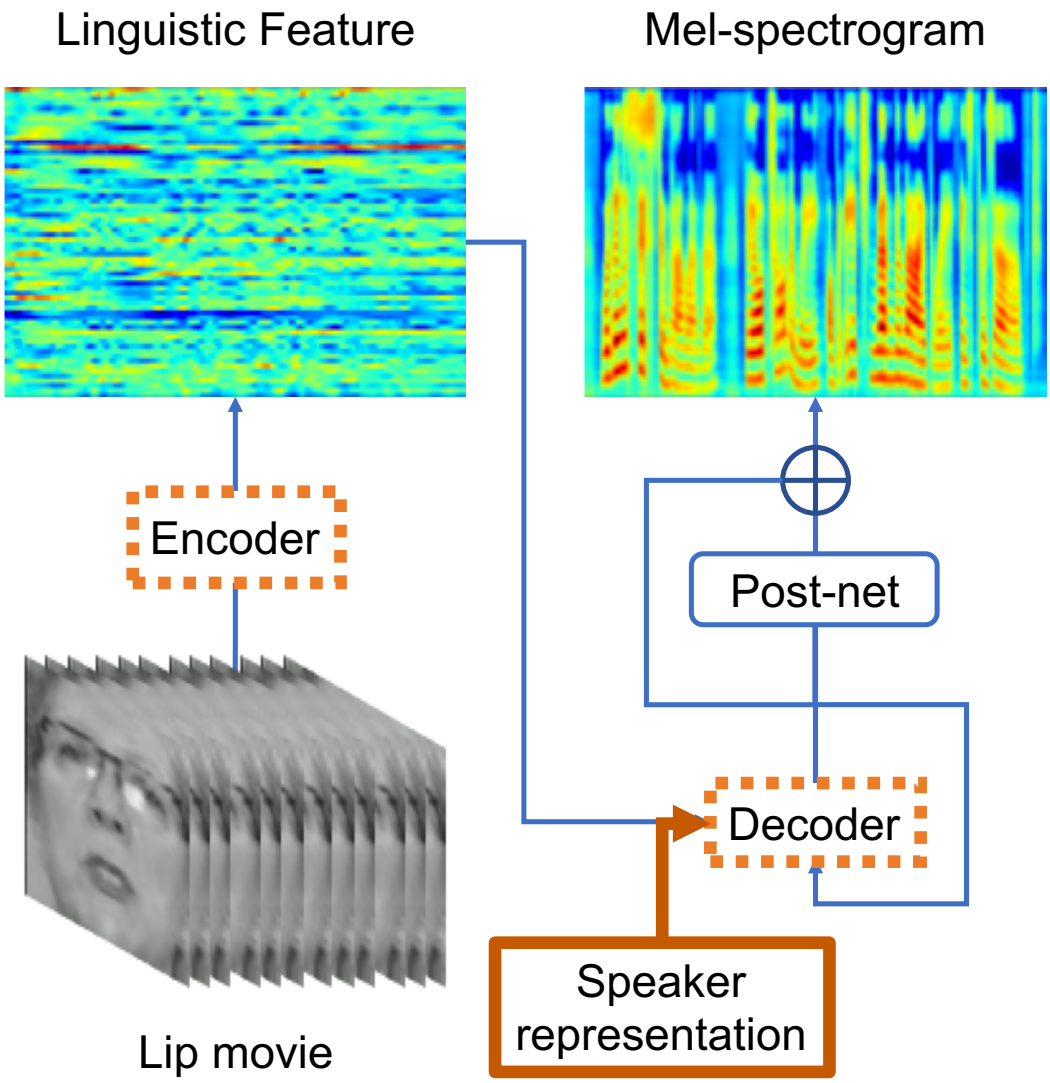


Fig. 7.3: 複数話者口唇動画–音声変換の概要図.

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 分であった．

動画は 60fps で収録したが，フレームを間引いて 50fps にした．S^3FD [95] によって動画の各フレームから顔領域を矩形で検出した．各フレーム顔領域の中心を中心座標とし，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．さらに口唇動画は (48×48) の大きさにリサイズした．

音声は 16kHz 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，70Hz から 8000Hz の帯域に対して適用した．ここで，音声の分析シフト長を 10ms とすることで，口唇動画 1 フレームに対し，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 Decoder は 1 ループで 2 フレームのメルスペクトログラムを推定するように設定した．出力のメルスペクトログラムからは，NNLS 法と GL 法によって音声を得た．

7.3.2　ネットワーク構造

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．時空間 resnet の出力に Global avarege pooling を適用し空間方向の軸を削減した上で，位置エンコーディングを加算し，transformer's block に入力する．ユニット数 128 で Multi-head Attention のヘッド数は 2 とした．正則化として 10% の dropout を適用した．

Decoder はユニット数 256 の prenet と 6 層の GLU block からなる．GLU block はチャンネル数は 256 で因果的畳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 dropout を適用した．また，post-net のチャンネル数は 512 とした．

7.3.3　学習条件

Encoder 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 Batch Normalization を適用することで正規化の代わりとした．teacher-forcing 時の入力となる前時刻のメルスペクトログラムは時間，周波数方向にマスクをかけることで劣化させ，汎化を促進させた．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 α パラメータを半減させた．学習時はメルスペクトログラムで 300 フレーム (3 秒) の固定長で入力し，これより発話が長い場合はランダムに切り出し，短い場合は padding を行い padding 部分は損失の計算から除外した．

7.4　結果と考察

評価は明瞭性に関する客観指標である STOI[96]，ESTOI[97]，自然性に関する客観指標である PESQ[98, 99]，音声のパワの RMSE(POWER [dB])，SWIPE[100] を用いて計算した対数 F0 の RMSE(LF0)，有声/無声判定の誤り率 (VUV [%]) を用いて行った．

さらに，Julius dictation-kit の DNN-HMM 音響モデルを用いて求めた音素列をもとにした音素誤り率 (Phoneme error rate; PER) についても検討を行った．PER は 5.3.2 節で説明した WER と同様の計算方法を音素列に適用して計算する．ただし本手法によ

る音素列の推定には，言語モデルの知識が加味されている点に注意が必要である．参照系列は原音声から Julius を用いて得られた音素列とした．

7.4.1 改善法の効果

まず，女性話者 1 名のデータを用いて，7.1 節で説明した，より良いメルスペクトログラムを得るための 4 つの手法の効果を確認した．すべての手法を導入していない場合を Baseline とし，それに動的特徴量を考慮した損失 (DL)，Transformer's block (TB)，音源情報に関するマルチタスク学習 (MT)，Scheduled Sampling for Transformers(SS) をそれぞれ，あるいは組み合わせて導入した際の客観評価指標を求めた．

Table 7.1 に各条件での評価指標の値を示す．マルチタスク学習以外の手法に関して，音声のパワの RMSE を除き，手法の導入によって客観評価指標が改善している．最も改善が大きかったのが，動的特徴量を考慮した損失，Transformer's block，Scheduled Sampling for Transformers を同時に適用した場合であり，PESQ が 1.254，STOI が次点で 0.6632，ESTOI が 0.5711，対数 F0 の RMSE(LF0) が 0.4951，有声/無声判定の誤り率が 13.89%，PER が 13.02% となった．Fig 7.4 にテストセットの 1 文「小さな鰻屋に，熱気のようなものがみなぎる.」について口唇動画から推定した音声を再分析したスペクトログラムの例を示す．中段の動的特徴量を考慮した損失，Transformer's block，Scheduled Sampling for Transformers を同時に適用した条件と下段の正解値を比較してみると，推定したメルスペクトログラムはメルフィルタバンクを適用した影響で高域の微細構造は表現できていないものの，概形はよく推定できている．一方，上段の Baseline では 0 から 0.5 秒の区間にもともと存在しなかった信号成分が推定されている．これは発話の予備動作のブレスから誤って音声を推定してしまったものと考えられる．この問題は改善法を導入することで改善されており，口唇動画とパワの間の関係は発話全体での長期の文脈を考慮することでモデリングが可能になる．

しかしながら，この予備動作の取り違えの問題があるにも関わらず，対数パワについての誤差は Baseline が 2 番目に良く 8.650 という結果になった．これに関しては，コーパス中の音声信号の正規化方法が適切ではなかったことが原因と考えられる．現状は発話ごとに波形のピークを正規化しているが，発声の強さが変われば調音にも影響が生じるので，この方法が適切ではなく，学習，評価の両方に悪影響を及ぼしている可能性がある．

7.4.2 ネットワーク構造による効果

続いて，提案したネットワークの性能の評価を行う．男女それぞれのデータに対して，それぞれ調音–音声変換を構築し比較を行った．比較対象としては，Lip2AudSpec[88] の

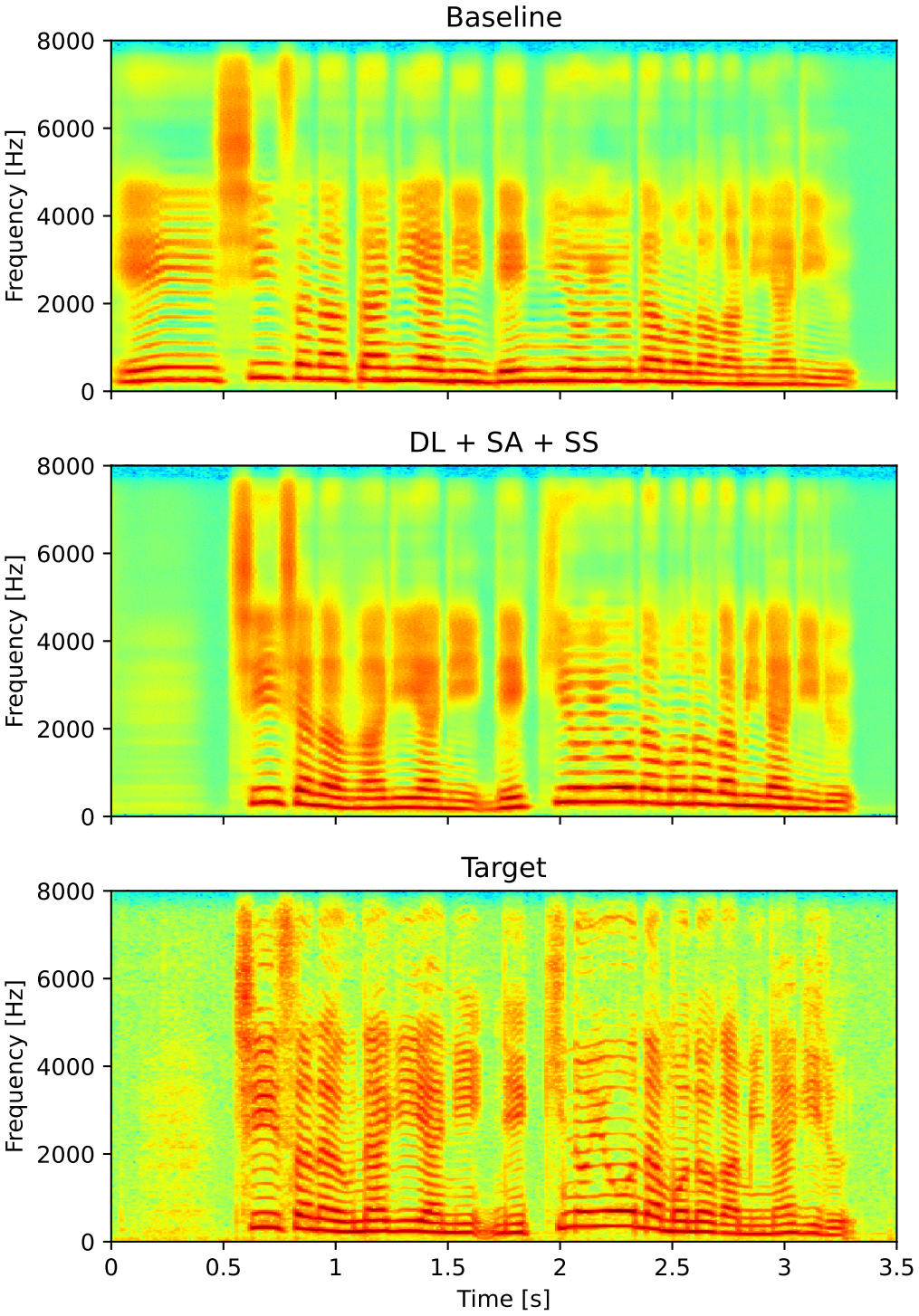


Fig. 7.4: 「小さなうなぎ屋に熱気のようなものがみなぎる」を発話した際のスペクトログラム.

Table 7.1: 各改善法を導入した際の客観評価指標

	PESQ	STOI	ESTOI	
Baseline	1.145	0.6060	0.4673	
+ DL	1.162	0.6314	0.5134	
+ TB	1.180	0.6331	0.5101	
+ MT	1.125	0.5870	0.4338	
+ SS	1.188	0.6266	0.5109	
+ DL + TB	1.241	**0.6639**	0.5663	
+ DL + TB + SS	**1.254**	0.6632	**0.5711**	
	POWER	LF0	VUV	PER
Baseline	8.650	0.6209	0.2339	0.2851
+ DL	10.023	0.5508	0.1876	0.2067
+ TB	**6.920**	0.5581	0.1688	0.1945
+ MT	10.789	0.6489	0.2647	0.3272
+ SS	9.017	0.6077	0.1799	0.1967
+ DL + TB	8.752	0.5105	0.1400	0.1570
+ DL + TB + SS	8.923	**0.4951**	**0.1389**	**0.1302**

ネットワーク構造を用いる．Lip2AudSpec は時空間 CNN と LSTM，MLP の積層から構成されている．文献の方法から，時空間 CNN の出力後に GAP を適用し，口唇動画からメルスペクトログラムの時間解像度にアップサンプリングする部分に最近傍補間法を用いるように変更を加えている．また，ネットワークのみを参考にし，その他の手続きは提案法と同じである．提案法に関しては前節の動的特徴量を考慮した損失，Transformer's block，Scheduled Sampling for Transformers を同時に適用した条件を用いる．どちらのネットワークも学習イテレーション数を 15 万とした．

Table 7.2 にネットワーク構造による客観指標の変化を示す．女性話者に関してはパワを除いたすべての指標に関して提案法が上回っている．一方，男性話者に関してはその限りではないが，有声/無声の誤り率が約 7%，PER が約 10% と大きく改善している．よって男性話者に関しては提案法のほうがより音素文脈を再現した音声を合成することが可能であり，提案したネットワーク構造は口唇動画–音声変換に適していると言うことができる．

Table 7.2: ネットワーク構造による客観指標の変化

(a) 女性話者

	PESQ	STOI	ESTOI	
Lip2AudSpec	1.188	0.6129	0.4973	
Proposed	**1.257**	**0.6330**	**0.5415**	
	POWER	LF0	VUV	PER
Lip2AudSpec	**6.528**	0.5701	0.1707	0.1457
Proposed	8.286	**0.5104**	**0.1445**	**0.1084**

(b) 男性話者

	PESQ	STOI	ESTOI	
Lip2AudSpec	**1.585**	**0.6454**	0.4629	
Proposed	1.535	0.6400	**0.4883**	
	POWER	LF0	VUV	PER
Lip2AudSpec	**12.28**	**0.1373**	0.1958	0.1725
Proposed	15.54	0.1404	**0.1212**	**0.0759**

7.4.3 複数話者口唇動画–音声変換

最後に，男女 2 名の話者のデータを同じネットワークで学習する複数話者口唇動画–音声変換を検討した．話者表現ベクトルは 256 次元の単位ベクトルで表されるとして学習を行った．ネットワーク構造は前節までと同様で，動的特徴量を考慮した損失，Transformer's block を同時に適用した条件を用いる．学習イテレーション数を 20 万とした．ドメイン敵対学習に関しては勾配スケールパラメータ λ に関して，$\lambda = 0.1$ と $\lambda = 0$(ドメイン敵対学習を適用しない) の 2 条件に関して検討を行った．

まず，話者それぞれを別のモデルで学習する場合と，同じネットワークで学習する場合の客観評価指標の変化を比較する．Table 7.3 に各条件での客観評価指標を示す．ここで，Speaker Dependent はそれぞれの話者に関して別々に学習した結果であり，Table 7.2 の提案法の客観指標を男性話者と女性話者で平均をとったものである．よって，この条件のみ学習イテレーション数が 15 万回となっている．複数話者モデルにすることで，客観

評価指標は低下していることがわかる．特に PER を見ると Speaker Dependent 条件の 9.22% から，6〜7% ほど低下しており，音声の了解性が劣化してしまっていることがわかる．これらの条件間では学習イテレーション数が異なるため，複数話者モデルが学習回数の増分の 5 万イテレーションの間に過学習が起こってしまっている可能性が考えられるが，$\lambda = 0.1$ のモデルの学習曲線 (Fig. 7.5) を見ると，実際には過学習は生じていないことがわかる．よって要因として考えられるのは，ネットワークの持つパラメータが 2 名の話者を一般化するには少なかったことが挙げられる．学習曲線を見ても，検証用データよりも学習データに対する損失が大きいので，ネットワークを大きくしてパラメータを増やしモデルの自由度を上げることはより良い結果につながると考えられる．ただし，この学習曲線は teacher-forcing 適用下のものであり，自己回帰による誤差の累積の影響は考慮されていない点で注意が必要である．

ドメイン敵対学習の有無では，適用していない $\lambda = 0$ の場合と比べて，$\lambda = 0.1$ の場合の方がパワの RMSE を除いて微小に改善した．すなわち，PESQ が 1.319 から 1.338, STOI が 0.5880 から 0.6021, ESTOI が 0.4520 から 0.4601, 対数基本周波数の RMSE が 0.3472 から 0.3418，有声/無声判定の誤り率が 16.32% から 14.95%，PER が 16.58% から 15.68% となり，ドメイン敵対学習を適用し，口唇動画から話者性を取り除いたことで，合成音声の品質が微小に改善された．

提案法は，口唇動画から音韻情報を抽出し，それとは別に話者情報を入力することによって口唇動画から音声を合成する．よって，口唇動画と話者情報の話者が異なっていた場合でも音声を得ることができる．Fig. 7.6 に女性話者の話者情報と男性話者の「小さなうなぎ屋に熱気のようなものがみなぎる」と発話した際の口唇動画を入力して得られた音声を再分析したスペクトログラムを示す．合成された音声は，音韻性については入力した口唇動画のものとなっている．よって，Fig. 7.6 中段の男性話者のスペクトログラムと概形は一致している．一方で，話者性は入力された女性話者のものとなっている．基本周波数 (スペクトログラムの横縞間の幅) を見ると，合成した音声は，Fig. 7.6 中段の男性話者のものよりも，Fig. 7.6 上段の女性話者に近くなっていることがわかる．

入力する口唇動画と話者情報が一致している場合を SAME 条件，そうでないものを CROSS 条件として PER を求めたものを Table 7.4 に示す．前述の通り，SAME 条件ではドメイン敵対学習で PER が 16.58% から 15.68% に改善された一方，CROSS 条件では 23.57% から 25.41% に増加している事がわかる．ドメイン敵対損失の導入によって，モデル内の話者依存性が低減されれば，CROSS 条件においても PER は改善されるはずである．CROSS 条件の合成音声を確認すると，発話リズムなどの大域的な特徴はよく再現されているものの，音韻性が一部異なって聞こえることが多い．現状，口唇動画の時系列から 1 つの話者ラベルを推定しているので，時系列内に含まれる発話の癖などの個人性がドメイン敵対学習によって抽出できなくなっている事が考えられる．よって，ドメイン

Table 7.3: 複数話者口唇動画–音声変換の客観評価指標

	PESQ	STOI	ESTOI	
Speaker Dependent	**1.396**	**0.6365**	**0.5149**	
$\lambda = 0$	1.319	0.5880	0.4520	
$\lambda = 0.1$	1.338	0.6021	0.4601	
	POWER	LF0	VUV	PER
Speaker Dependent	**11.91**	**0.3254**	**0.1329**	**0.0922**
$\lambda = 0$	13.65	0.3472	0.1632	0.1658
$\lambda = 0.1$	14.01	0.3418	0.1495	0.1568

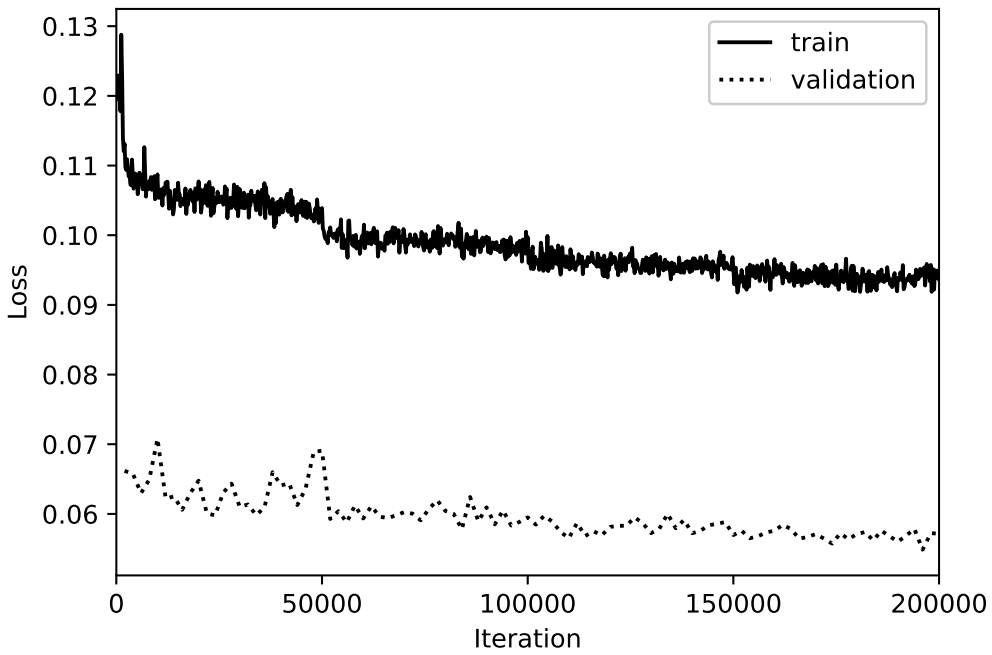


Fig. 7.5: 複数話者モデル ($\lambda = 0.1$) の学習曲線

敵対学習を時間フレームに関して独立に適用することで，振る舞いの改善が期待できる．また，口唇動画を話者の目尻間距離で正規化を行う，口唇の特徴点を抽出しその変位を調音情報として用いるなど，話者に対して頑健に音素特徴を抽出するための口唇動画の前処理の工夫についても今一度検討する必要がある．

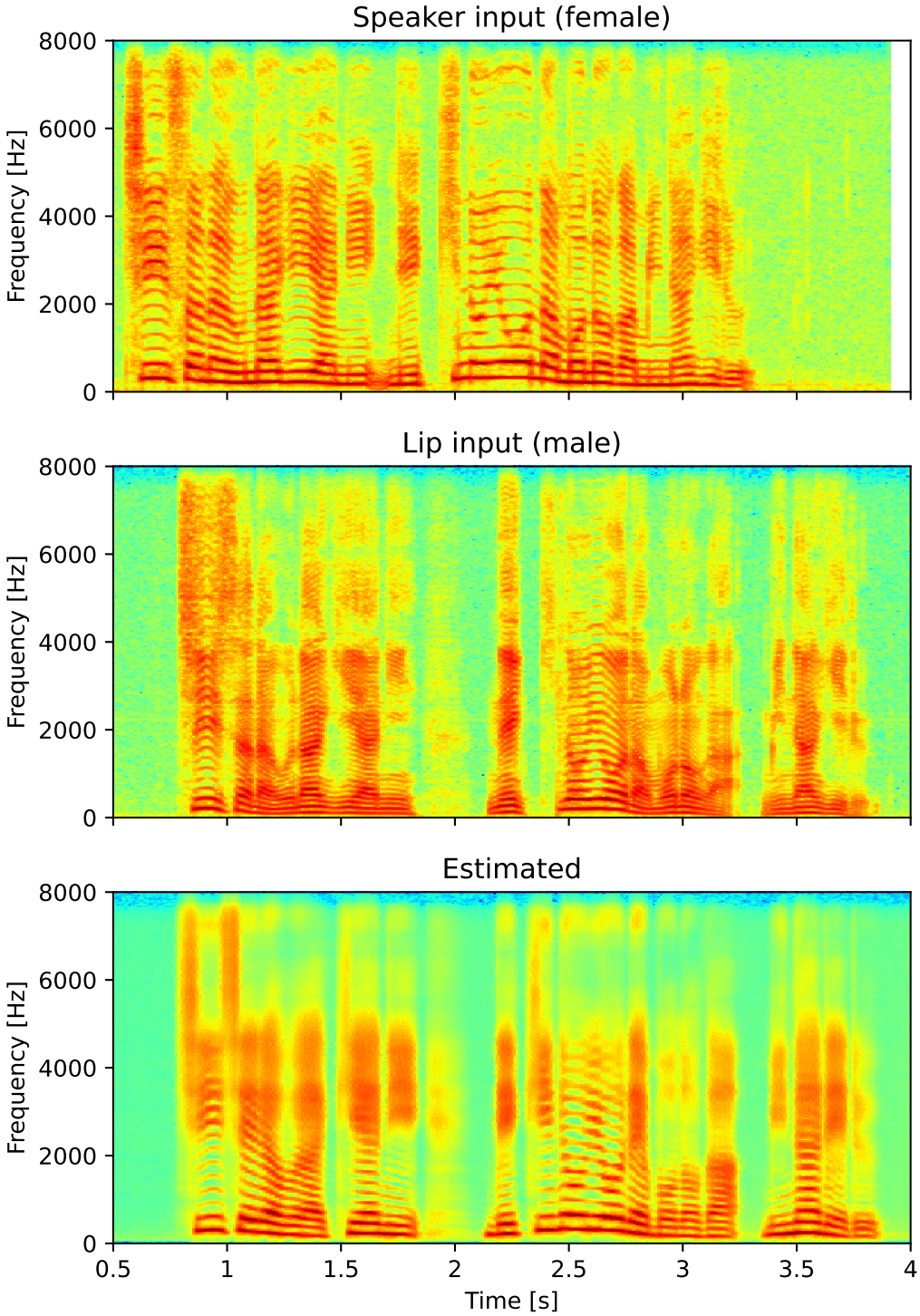


Fig. 7.6: 男性話者の「小さなうなぎ屋に熱気のようなものがみなぎる」と発話した際の口唇動画と女性話者の話者情報を入力して合成された音声のスペクトログラム.

Table 7.4: 口唇動画と話者情報の一致不一致による PER の違い

	SAME	CROSS
$\lambda = 0$	0.1658	**0.2357**
$\lambda = 0.1$	**0.1568**	0.2541

7.5 おわりに

本研究では，畳み込み層を基底とした系列変換モデルを用いた口唇動画–音声変換法を提案し，男女 2 名の日本語話者のデータ用いて，客観評価指標により評価を行った．まず，より良いメルスペクトログラム時系列を得るための改善法に関して検討を行った．多くの改善手法は客観指標に良い効果をもたらしたものの，マルチタスク学習では予想された音源情報の推定精度の向上が達成されなかったため，新たな音源情報に関する制約を検討する必要がある．また，ネットワーク構造に関して先行研究と提案法を比較した結果，特に PER に関して 4% から 10% の改善が見られ，提案法がより了解性の高い音声を合成できることを示した．最後に 1 つのネットワークを複数話者のデータで学習する複数話者口唇動画–音声変換について予備的な検討を行ったところ，話者性や発話リズムなど大局的な特徴はよく表現できているものの，音韻性の再現性という点に関しては話者依存モデルよりも大きく劣化したため，さらなる検討が必要である．

第 8 章

結論

8.1 総括

本研究では，調音器官の運動パタンに関する情報である調音情報と音声を同時に記録したデータを基に，調音情報と音声の間の関係をモデリングするデータ駆動型調音・音声間変換に関して，大規模日本語調音・音声パラレルデータの収集，調音・音声間変換における音源情報の活用，深層学習の導入による調音・音声間変換の精度向上，実応用のための新たな調音情報の検討を行った．

4 章では，データ駆動型調音・音声間変換を実現するための日本語調音・音声パラレルコーパスの収集をおこなった．既存の ATR 音素バランス文に加えて，Wikipedia 日本語版を文候補として，所望の総モーラ数の中で，できるだけ多様なトライフォンが登場するよう Minoux の改良貪欲法によって文選択を行った．得られた音素バランス文と ATR 音素バランス文を同時に用いることで，2 回以上登場したトライフォンの種類数が ATR 音素バランス文の 1.7 倍となり，この音素バランス文を用いることで，より多様な音素文脈の収録が可能となった．この音素バランス文をもとに，3D-EMA と口唇動画の収録を行った．3D-EMA では，日本語男性話者 1 名の約 1 時間のデータ，口唇動画では日本語話者男女 1 名ずつの各 4.8 時間のデータとなり，日本語としては唯一で，英語の公開データベースと比較しても，単一話者の発話時間は比較的長期となるデータを収集することができた．

5 章では，磁気センサによる調音情報から音声を得る調音–音声順変換の検討を行った．ここでは，双方向再帰型ニューラルネットワークによって，声道のフィルタ特性だけではなく音源情報も推定する調音–音声変換を新たに提案した．先行研究である GMM および MLP と同一のコーパスを用いて比較を行ったところ，客観評価では提案法は多くの音声特徴量の推定誤差を改善することが示された．さらに，主観評価においては提案法の自然性に関する MOS 評点が 3.115 となり，先行研究と比べてより自然性の高い音声を得られ

る可能性が示唆された．

6 章では，磁気センサによる調音情報を音声から得る音声–調音逆変換の検討を行った．本研究では新たに Residual Time-Delay Neural Network による音声–調音逆変換法を提案し，公開されているコーパスと 4 章で収集した計 10 名の 3D-EMA データを評価に用いた．各話者に関して話者依存逆変換モデルをそれぞれ構築した上で先行研究との比較を行ったところ，すべての話者に関して提案法が最も調音軌道の推定精度が高いという結果が得られた．また，提案法はモデルの層数と学習データ量に対して頑健であることがわかった．さらに，提案法をもとに，音声–調音逆変換に最適な特徴量を検討したところ，メル周波数ケプストラム係数 (mfcc) が最も推定精度が良いこと，音源フィルタ分離は音声–調音逆変換に有効ではないこと，メルフィルタバンクと離散コサイン変換による特徴量の次元削減は限られたデータ量を扱う音声–調音逆変換には有効であることが示された．

7 章では，口唇動画による調音情報から音声を得る調音–音声順変換の検討を行った．本研究では新たに畳み込み系列変換モデルを基とした調音–音声変換を提案し，4 章で収集した男女 2 名の日本語話者のデータ用いて，客観評価指標により評価を行った．まず，より良いメルスペクトログラム時系列を得るための改善法に関して検討を行ったところ，動的特徴を考慮した損失，Self-Attention の導入，Scheduled sampling による入力特徴量の劣化が有効であった一方，音源情報に関するマルチタスク学習は結果に悪影響を及ぼすことがわかった．改善した手法をすべて盛り込むことで，合成音声の PER が 28.51% から 13.02% に改善された．また，ネットワーク構造に関して先行研究である Lip2Audspec と提案法を比較した結果，特に PER に関して 4% から 10% の改善が見られ，提案したネットワーク構造がより了解性の高い音声を合成できることが示された．最後に 1 つのネットワークを複数話者のデータで学習する複数話者口唇動画–音声変換について予備的な検討を行ったところ，話者性や発話リズムなど大局的な特徴はよく表現できているものの，音韻性の再現性という点に関して PER が 9.22% から 15.68% に低下し，話者依存モデルよりも大きく劣化したため，さらなる検討が必要である．

8.2　今後の展望

磁気センサデータの収録および話者に対する一般化表現

磁気センサデータは話者と収録状況に依存する．つまり，話者が異なったり，同じ話者でも収録時にコイルを装着しなおせば，それらのデータを一般化することは難しくなる．

文献 [101] では，測定された受信コイルの軌道から Tract Variables と呼ばれる値を計算することによって，複数話者の音声–調音逆変換を実現させているが，Tract Variables は同一話者内で，コイルの取り付け位置が異なる場合の一般化が達成できていない．

理想的には，簡便な調音器官の測定を基に得られた個人性と一般化された調音表現から，調音軌道を表す特徴表現が得られるのが望ましい．それへのアプローチとして，話者と収録条件に非依存な一般化調音表現空間の構築が考えられる．一般化調音表現空間は生の調音軌道から収録バイアスと話者バイアスの影響を除去することによって，一般化調音表現空間に転写し，一般化調音表現空間上の時系列から，収録バイアスと話者バイアスを加味することで調音軌道を復元するというアプローチである．この空間が実現されれば，複数話者調音・音声間変換を容易に実現することが可能となり，調音・音声パラレルコーパスのデータ量の上限から解放される．

他人の音声を自身の調音に変換する音声–調音逆変換

ある音声を聞いてそれを声真似することを考えると，言語情報に限らず，非言語情報やパラ言語情報に関しても，その目的音声に近づけるように発話を行うはずである．その過程には，言語的な再現以外に他人の音声を自身の調音運動によって再現しうるかの検討が行われているはずであり，これは，他人の音声を自身の調音に変換する音声–調音逆変換であるといえる．この機能の再現は音声の生成，知覚を考える上で有益であると考えられる．

他人の音声と自身の調音のパラレルデータは存在し得ないので，本研究で用いた教師あり学習によるアプローチは使用することができず，評価方法も単純な推定誤差を使用することはできない．深層学習においては文献 [102] のように非パラレルな時系列の変換が可能であるので，その手法を参考にすることで，変換自体は実現しうる．一方で，評価方法に関してはよく検討する必要がある．

調音–音声順変換に適したニューラルボコーダ

近年，Deep Neural Network によって音声波形を得るニューラルボコーダの研究が進んでいる．代表的なものとして，Wavenet Vocoder[30] があり，これを導入したテキスト音声合成法によって合成された英語音声は，肉声と聞き分けができないほどの自然性を持つ [91]．ここで用いられた WaveNet Vocoder は 24.6 時間の単一話者のデータで学習されており，高品質な音声を得るためには大量の学習データが必要になる．

調音・音声パラレルコーパスは，調音収録の特殊さからデータ量が限られる．よって，ニューラルボコーダを調音–音声順変換に適用する場合には，調音・音声パラレルコーパスとは別に，同一話者の音声だけのデータを大量に収集しニューラルボコーダを学習するか，学習データに含まれない未知の話者および未知の背景雑音に対して頑健であるか，あるいは少量のデータを用いて適応可能なニューラルボコーダを検討する必要がある．前者

のアプローチは，既存の公開調音コーパスで実現するのは難しい．後者の方法を実現できれば，調音–音声順変換だけではなく，声質変換等の様々な問題に対して音声の品質向上に貢献できる．

既存のニューラルボコーダを大量の話者を含む音声データを用いて学習することで，未知話者に対しても有効なニューラルボコーダを構築するのは現状でも可能であるが，学習データに含まれる話者と含まれない話者の間では得られる音声の品質に差が生じてしまう．

また，Wavenet Vocoder は，波形を 1 サンプルごとに自己回帰によって生成する．よって，合成速度が非常に遅くなってしまう．一方，MelGAN [103] のように波形の時系列をバッチ生成するアプローチも数多く提案されているが，自己回帰型の手法と比べて品質が低下する問題点がある．

さらに，ニューラルボコーダに入力する音声特徴量は，多くの場合正解値よりも劣化しているので，この入力の劣化に対しても頑健になるように学習法を工夫する必要がある．

調音–音声順変換のアプリケーション化

調音–音声順変換は代用音声や SSI への応用が期待されるが，現状のレベルで十分な了解性をもつ音声が得られているので，実応用のためのアプリケーションとしての開発を進めていく必要がある．調音情報として 7 章で検討した口唇動画を選択すれば，普及した端末で容易に収録できることから，ソフトウェア的な開発のみでアプリケーションが実現できる．

代用音声として応用する場合は，低遅延性が重要となってくる．7 章で提案したモデルを代用音声として用いる場合，まず，因果性を満たすように改良すべきであろう．Decoder の部分ではこの要件は満たされているので，Encoder の口唇動画から特徴を抽出する部分に関して，未来の情報を使用しないように変更する必要がある．未来の情報が使用できなくなる事によって，生成される音声の品質がどのように変化するかは検討が必要である．また，Encoder に用いられる時空間畳み込みの演算量が多いため，代替となる手法が必要となるかもしれない．また，現状の方法では 5 時間程度の口唇動画–音声データを事前に収録しておく必要がある．これは代用音声としては現実的ではないため，モデルの未知話者への zero-shot 適用を検討する必要がある．

本研究では，読み上げ音声を収録した調音・音声パラレルコーパスから，調音–音声順変換を構築した．その一方で，代用音声は音声コミュニケーションの維持のためのものであることから，会話音声など，読み上げ音声より多様な自由発話の再現が求められる．自由発話は調音の観点からも，読み上げ音声より多様になることが予測されるので，読み上げ音声を用いて構築した調音–音声順変換が自由発話によるコミュニケーションにそのま

ま導入できるかは検討が必要である．

代用音声に入力される口唇動画は本研究で用いたものとは異なり，実際の音声の生成を伴わない．この場合，聴覚フィードバックが機能しないため，調音運動の様態が異なるものになる可能性がある．実際の音声の生成を伴わない調音運動に対して，調音–音声順変換が有効に働くかを検証するためには，調音収録の段階において，そのような条件下でデータを収集する必要がある．また，低遅延な調音–音声順変換が実現された際には，合成された音声がフィードバックされることによって調音運動にどのような影響が現れるかについては，興味深い研究対象になりうる．

謝辞

本研究を進めるにあたり多くのご指導，ご助言を頂いた，九州大学大学院芸術工学研究院の鏑木時彦教授に心から感謝の意を表します.

副査として本論文をご精読いただき，大変有意義なご指摘やコメントを賜りました，長崎大学の松永昭一教授と九州大学大学院芸術工学研究院の吉永幸靖准教授に心より感謝申し上げます.

インターンシップや共同研究でお世話になった，NTT コミュニケーション科学基礎研究所の持田岳美氏，廣谷定男氏，上江洲安史氏，また，共同研究でお世話になった，東京理科大の桂田浩一准教授に深く感謝申し上げます.

研究についての議論や実験の補助などを通じて本研究に協力いただいた，九州大学大学院芸術工学研究院の若宮幸平助教ならびに，研究室の先輩，同輩，後輩に心から感謝申し上げます.

最後に，本研究を遂行するにあたって，お世話になったすべての方々にこの場を借りて厚く御礼申し上げます.

付録 A

4.1.3 節で選択された音素バランス文

A-1　スーザフォーンとも。
A-2　眉唾と言える。
A-3　ハルピュイアの姉。
A-4　茶目っ気を見せた。
A-5　エッセイ等あり。
A-6　チューダー朝とも。
A-7　東胡へ贈った。
A-8　ラージプートなど。
A-9　ホソトビウオなど。
A-10　通風管とも。
A-11　キャンペーンガール。
A-12　ガーンジーである。
A-13　異父兄は何苗。
A-14　首都はポドゴリツァ。
A-15　硫黄臭がする。
A-16　本阿弥光悦。
A-17　官位は正四位。
A-18　モーツァルトの妻。
A-19　杖で打ち据えた。
A-20　スィパーヒーだった。
A-21　コンペを開いた。
A-22　メーン州生まれ。
A-23　それじゃ普通だぞ。
A-24　夫はヒュプノス。
A-25　子は建部政世。
A-26　甘えちゃいけない。
A-27　カンピ・ビゼンツィオ。
A-28　アーラ・ディ・ストゥーラ。
A-29　江戸定府だった。
A-30　詭弁を弄した。
A-31　ポピュラーな種族。
A-32　県都はペシュコピ。
A-33　全米オープン。
A-34　下馬評は割れた。
A-35　パーカッショニスト。
A-36　グディーズであった。
A-37　ロンツォ＝キエーニス。
A-38　非折りたたみ式。
A-39　プロレスレフェリー。
A-40　ギャラリーオーナー。
A-41　首都はフェアアーラ。
A-42　新しいジーグ。
A-43　ベリャーエフは言う。
A-44　古墳跡もある。
A-45　ボアーラ・ピザーニ。
A-46　ポッジョ・ブストーネ。
A-47　百里四方ある。
A-48　ペスカッセーロリ。
A-49　家へ連れ帰る。
A-50　ハノーファー王妃。
A-51　産業技術史。
A-52　トニー・オーバンら。
A-53　偶数月刊。
A-54　俗に言う亜人。
A-55　庶民院議長。
A-56　鰻屋へ向かう。
A-57　レジャーシートなど。
A-58　ソゾポルとなった。
A-59　バフィーのウォッチャー。
A-60　ビオトープがある。
A-61　トゥチャ族出身。
A-62　朝比奈安兵衛。
A-63　属性は炎。
A-64　ダプネーを生んだ。
A-65　オウェキノ湖がある。
A-66　バッシャ・サルダーニャ。
A-67　違法で無効だ。
A-68　獣医学博士。
A-69　愛称ギューティー。
A-70　伊勢湾へ注ぐ。
A-71　ティーアハウプテン。
A-72　ベビーキャリーとも。
A-73　防具を調合。
A-74　ペチョルスクである。

A-75　神功皇后。
A-76　描線は細め。
A-77　ニーチェを専攻。
A-78　ボフスラーウなど。
A-79　およびマネージャー。
A-80　盗癖があった。
A-81　オープニングより。
A-82　ヘリザウ出身。
A-83　どうするシンディー。
A-84　デンマーク王女。
A-85　ため池百選。
A-86　ファンガヌイ生まれ。
A-87　解脱を得させた。
A-88　マーフィー市がある。
A-89　セビージャ・ラ・ヌエバ。
A-90　ブルーアイランド。
A-91　一塁手だった。
A-92　趣味は旅行、絵。
A-93　法蔵部所伝。
A-94　ペースを握った。
A-95　トリガーを引いた。
A-96　ウェストバリー駅。
A-97　ブーテースがいる。
A-98　シャンパーニュ伯妃。
A-99　最も分厚い。
A-100　名前はダグニー。
A-101　夫婦デュオである。
A-102　同技師補である。
A-103　リパーゼ陽性。
A-104　首府はワガドゥグー。
A-105　ディープフライする。
A-106　台中州知事。
A-107　ルーシェの親友。
A-108　ブルッヘの生まれ。
A-109　エウィンへと帰る。
A-110　園城寺長吏。
A-111　醤油醸造家。
A-112　路線案内色。
A-113　もう一人のジョセフ。
A-114　上空を制圧。
A-115　上り列車発車。
A-116　ズウォティの補助単位。
A-117　州都はトゥクピータ。
A-118　空路を利用する。
A-119　ニルチッイと出会う。
A-120　カーブも投げ分ける。
A-121　重労働はない。
A-122　レオポルト・アウアー。
A-123　エンスヘデーがある。
A-124　変異がおきやすい。
A-125　産声があがった。
A-126　妻はペルセポネー。
A-127　半発酵茶とも。
A-128　大いに潤った。
A-129　ヘ音記号となる。
A-130　魚の目治療薬。
A-131　それゆえ増えやすい。
A-132　うまく合えば増える。
A-133　バッサーノ・ロマーノ。
A-134　ニャーンも戦死した。
A-135　羽織る上着である。
A-136　良いプレーができる。
A-137　まれにローズウッド。
A-138　シオーフォクにて死去。
A-139　心血を注いだ。
A-140　ティーコゼーを使う。
A-141　ライナーノーツ付き。
A-142　老上単于の子。
A-143　病院へ直行。
A-144　代々木上原など。
A-145　ウェルウィッチアがある。
A-146　もうよせばいいのに。
A-147　絵本ナビゲーター。
A-148　ザポリージャ出身。
A-149　キウーザ・ディ・ページオ。
A-150　ブレッツォ・ディ・ベーデロ。
A-151　ボジージオ・パリーニ。
A-152　ヌーブール出身。
A-153　フォークを武器とした。
A-154　ボクのティーチャーです。
A-155　イーゾラ・デル・リーリ。
A-156　カステッラフィウーメ。
A-157　ジョイオーザ・マレーア。
A-158　剛毛は持たない。
A-159　南方に二社ある。
A-160　ヤーヤー一揆とも。
A-161　パーシパエーがいる。
A-162　テーマはエコロジー。
A-163　そうマッギンは言う。
A-164　周悦と号した。
A-165　アイフェ、イーフェとも。
A-166　子供を湯屋へやる。
A-167　雌蕊は多数ある。
A-168　ジョシュアが住んでいた。
A-169　現テヘラン市長。
A-170　主な町はデュズジェ。
A-171　スペインへ留学。
A-172　オーディションを通過。
A-173　河南按察使となる。
A-174　パエトゥーサと姉妹。
A-175　紺綬褒章授与。
A-176　伊藤景経とも。
A-177　荘重なアダージョ。
A-178　成長をアピール。
A-179　考察を進めた。
A-180　ティーシポネーの父。
A-181　エピメラーゼである。
A-182　テナーホーン奏者。
A-183　フルーヒウ連隊。

A-184　アンチテーゼがある。
A-185　ウドンメンチェイ州。
A-186　キーウンギャルと呼ぶ。
A-187　公安委員長。
A-188　轟音が響いた。
A-189　アウト・オブ・バウンズ。
A-190　エアバルーンアート。
A-191　見知らぬ部屋に居た。
A-192　モンペリエの生まれ。
A-193　ライプツィヒに居住。
A-194　ディンウィディ郡である。
A-195　グッズのみ販売。
A-196　ペプチダーゼである。
A-197　王の赦免を得た。
A-198　トゥッツィングにて死去。
A-199　防犯アナリスト。
A-200　木造平屋建て。
A-201　蓮華座上に坐す。
A-202　ハイデッガーがいる。
A-203　外郎売である。
A-204　ドローン評論家。
A-205　イーストヘスがある。
A-206　ローマを封鎖した。
A-207　普通は匂わない。
A-208　シンカーを操る。
A-209　体力を消耗。
A-210　イーピゲネイアとなる。
A-211　新潟県阿賀野市。
A-212　テーテュースと兄弟。
A-213　ターンアウトスイッチ。
A-214　口説く場面もあった。
A-215　住居不能となった。
A-216　略してイーペーとも。
A-217　かりいぬ座と呼ばれた。
A-218　荘王にこう言った。
A-219　自称はアマーズィーグ。
A-220　アシードへ譲渡した。
A-221　さらにこれをあおった。
A-222　冬はスキーもできる。
A-223　ジャギュアと表記される。
A-224　タイポグラファーである。
A-225　抑え捕手を務めた。
A-226　カンピョーネ・ディターリア。
A-227　手術適応はない。
A-228　英語でいうファースト。
A-229　越後広瀬駅など。
A-230　ブロンソンっぽくなる。
A-231　ギュイエンヌと呼ばれる。
A-232　ベビーパウダーがある。
A-233　ネツァワルコヨトルとも。
A-234　母親は家を出た。
A-235　正一位を追贈。
A-236　軍官はツーピース。
A-237　県都はナーシリーヤ。
A-238　諸王の王を示す。
A-239　ヘアメイクアーティスト。
A-240　コーンウォールに移る。
A-241　ブッティリエーラ・アルタ。
A-242　ヤンキー風の生徒。
A-243　チュン・ウー・チェンの渾名。
A-244　ダフィは変死していた。
A-245　ゼーロ・ブオン・ペルシコ。
A-246　ベルツォ・インフェリオーレ。
A-247　省都はバクリエウ市。
A-248　プエニャーゴ・スル・ガルダ。
A-249　チペワ語とも呼ばれる。
A-250　チュウオウアメリカハブ。
A-251　ニッツァ・モンフェッラート。
A-252　フォンタネート・ダゴーニャ。
A-253　グラーツへ赴いた。
A-254　輪になって歌う唄。
A-255　オルチャーノ・ディ・ペーザロ。
A-256　サミュエル・ウォードである。
A-257　ジェシー・ジェイムズである。
A-258　これを十如是という。
A-259　創造を司る。
A-260　バーベキュー場もある。
A-261　縦横差し替え式。
A-262　実家は呉服問屋。
A-263　人口は微増中。
A-264　ハーブを摘んで食べた。
A-265　ターチャーン川である。
A-266　酢味噌和えなどにする。
A-267　県都はギュミュシュハーネ。
A-268　ヒンドゥー教の寺院。
A-269　ラムペティエーと姉妹。
A-270　ポウォネチュカなどである。

A-271　ポーランドのジェンビツェ。
A-272　オオアメーバ属とも。
A-273　エレベーターメーカー。
A-274　豆腐茶屋と呼ばれた。
A-275　中条詮秀の子。
A-276　溝口善兵衛など。
A-277　花椒は欠かせない。
A-278　これを蝦夷梅雨という。
A-279　王妃はアティであった。
A-280　那智勝浦町長。
A-281　ベンチャーキャピタリスト。
A-282　ツァツォンを戦死させた。
A-283　テチーターに残留。
A-284　バークシャーで育った。
A-285　県都はバジェドゥパール。
A-286　フェノール臭を放つ。
A-287　ギー・ドゥボールらがいる。
A-288　ブーイーの町がある。
A-289　ウースチー州の都市。
A-290　白紺のツートーン。
A-291　家が貧しいボンビ。
A-292　武茂氏を鎮圧する。
A-293　漢字学習アプリ。
A-294　エニイ・ギブン・サンデー。
A-295　キャンプインを迎えた。
A-296　ダウンフォースを稼ぐ。
A-297　首府はアベンゴウロー。
A-298　カボチャで見いだされた。
A-299　マレー諸島を含む。
A-300　雄蘂より前に出る。
A-301　イーメイアスで終わる。
A-302　西はケープフィア川。
A-303　趣味はネットサーフィン。
A-304　ジャールナー県の都市。
A-305　ロープウェーの終点。
A-306　首都はクィズィルである。
A-307　スイス バーゼル生まれ。
A-308　スキーとスノーボード。
A-309　慈善病院入院。
A-310　実在説も根強い。
A-311　それをラーチャウォンと呼ぶ。
A-312　真っ直ぐ飛ぶだけだった。
A-313　ベージュも若干淡い。
A-314　ミャンマーへ譲渡された。
A-315　指輪を盗んでしまう。
A-316　ベーコンエッグバーガー。
A-317　アヒ・アマリージョを入れる。
A-318　アフタヌーン・ティーと呼ぶ。
A-319　臨命終時の略語。
A-320　母親のヘザーは画家。
A-321　中心駅は有家駅。
A-322　ウフィツィの名がつけられた。
A-323　脚部不安を発症。
A-324　応用案も出ている。
A-325　家へと帰っていった。
A-326　ベーケーシュチャバ出身。
A-327　被収容者処遇法。
A-328　自称ファッションフリーク。
A-329　地域委員会がある。
A-330　ギースナー盆地となる。
A-331　ニックネームはズーパー。
A-332　ローズ大尉を銃撃。
A-333　ゲチョなどを結んでいる。
A-334　アジピン酸が得られる。
A-335　モンティチェッリ・ブルザーティ。
A-336　ザリエーシェとも呼ばれた。
A-337　サウディアの航空事故。
A-338　いぼが一対ずつある。
A-339　視点はクオータービュー。
A-340　ポッツァーリオ・エドゥニーティ。
A-341　ウェーブヘアーをしている。
A-342　モンタージュ風スロット。
A-343　捕虫網等へ落とす。
A-344　岐阜県揖斐町生まれ。
A-345　ゴルトヘーフェ駅がある。
A-346　兄はフン・フンアフプー。
A-347　衆十万を集めた。
A-348　正四位を贈位される。
A-349　表記はスタウファーとも。
A-350　オルレアン朝とも呼ぶ。
A-351　フォチャ地方の主都である。

A-352　鞭毛をもって泳ぐ。
A-353　エピファネスを自称した。
A-354　アロイジアとジュゼッパに。
A-355　バントゥー語族の部族。
A-356　綱の上へあがってく。
A-357　ナーディル・シャーの異母弟。
A-358　ベオグラード市ゼムン区。
A-359　テレビ放映もされた。
A-360　動く速球と言える。
A-361　古名はマゼープィンツィ。
A-362　コミューンを横断する。
A-363　坂東舜一が居る。
A-364　彩雲寮へ入寮。
A-365　エーベレーベン出身。
A-366　非常通報器がある。
A-367　元大阪府副知事。
A-368　中心地はアジアーゴ。
A-369　ネザーウッド地区にある。
A-370　ディーワーンなどがあった。
A-371　ギターアンプも製造。
A-372　シャーチョプカとも呼ばれる。
A-373　そこへプシシェが登場。
A-374　スカダーは失脚する。
A-375　カパアウに設置された。
A-376　ピレネー山脈の山。
A-377　フーヤー付きの宦官。
A-378　マーハーは顧客に説く。
A-379　アチェ王国のパンリマ。
A-380　笑顔でキャンディをくれる。
A-381　無限宇宙は矛盾する。
A-382　フォノカートリッジともいう。
A-383　ウィッティヒ反応 を挙げる。
A-384　ホメオパシーの愛好者。
A-385　エグゼクティブ・プロデューサー。
A-386　カンザス州ウィチタ生まれ。
A-387　議員運営委員理事。
A-388　ジャンドゥイヤに因るとされる。
A-389　かれらの船は奪われた。
A-390　操る怪獣は不明。
A-391　慶永の補佐に務めた。
A-392　それを全数調査する。
A-393　若ハゲがチャームポイント。
A-394　八王子市へやって来た。
A-395　エネルゴンウェポンが付属。
A-396　ペーパージャムとも呼ばれる。
A-397　ア・コルーニャ空港がある。
A-398　安全寺坂へ通ずる。
A-399　指が露出した手袋。
A-400　欧州ツアーも成功。
A-401　いずれもフィーチャーフォン用。
A-402　技の種類も多種多様。
A-403　ログハウス風平屋建て。
A-404　多くのサケが遡上する。
A-405　旧制宇部中学校。
A-406　社是という言い方もある。
A-407　医者が処方箋を渡す。
A-408　スウィーニィ自身が付けた。
A-409　ルイーゼ・ビューヒナーがいる。
A-410　オーディションへ応募できる。
A-411　モーター雑誌記者である。
A-412　怪獣や宇宙人たち。
A-413　十八期中央委員。
A-414　布衣以下で御目見以上。
A-415　スーフィー王朝であった。
A-416　ロバツェは、ボツワナの都市。
A-417　ヨトゥンヘイムに攻め込んだ。
A-418　ガウワーは難を逃れた。
A-419　母親はビーガンのシェフ。
A-420　サナーイーはその両目だ。
A-421　アドベンチャーレースである。
A-422　シェウチェーンコ区に属する。
A-423　シーウィードが助けに来る。
A-424　御船神社と同座する。
A-425　サッカークウェート代表。
A-426　スピロペンタジエンがある。
A-427　個人所有物も多い。
A-428　コワイ妖怪がうじゃうじゃ。
A-429　再び封書を受け取る。

A-430　ネバダはあえて座礁した。
A-431　好物はビーフジャーキー。
A-432　ビッルング家とも呼ばれる。
A-433　見知らぬ場所へ流れ着く。
A-434　青梅中央署へ向かう。
A-435　ガッルーラ語とも呼ばれる。
A-436　娘エウアドネーの父。
A-437　小麦アレルギーを含む。
A-438　ジェネラル・サントス出身。
B-1　偶像になった。
B-2　社員配置駅。
B-3　テルアビブ生まれ。
B-4　ケファともいわれる。
B-5　凝着 を 防ぐ。
B-6　尊皇攘夷派。
B-7　ナースィルの息子。
B-8　調はヘ長調。
B-9　塗装はグリーン。
B-10　階級は中佐。
B-11　鋳造ピストン。
B-12　総じて小規模。
B-13　従一位男爵。
B-14　黒や赤もある。
B-15　蝦夷地のオオカミ。
B-16　乎加神社の社地。
B-17　オールラウンダー。
B-18　ウールのギャバジン。
B-19　チャースラフ生まれ。
B-20　ローラーブレード。
B-21　苗字は大島。
B-22　斯波雄蔵とも。
B-23　そう、言い切った。
B-24　フリーウェアである。
B-25　膳所藩が管理。
B-26　背骨を骨折。
B-27　邪悪な女よ。
B-28　ンバケ県がある。
B-29　ウェーダーともいう。
B-30　タヒチ島生まれ。
B-31　伊予大洲城主。
B-32　しゃべるのも下手だ。
B-33　メディアアーティスト。
B-34　メアーナ・ディ・スーザ。
B-35　モンテーウ・ダ・ポー。
B-36　デマンド運行。
B-37　ウェヌスが有名。
B-38　父親も病死。
B-39　シェイプアップした。
B-40　善悪を分かつ。
B-41　省都はトゥイホア。
B-42　エレメカの一種。
B-43　アポフィシスはない。
B-44　ノセシェチカがある。
B-45　般若姫は娘。
B-46　インドや中国。
B-47　ホールン伯爵。
B-48　トッレ・サン・ジョルジョ。
B-49　冬場は短い。
B-50　ノチェーラ・ウンブラ。
B-51　パラッツォ・ピニャーノ。
B-52　ピッツィゲットーネ。
B-53　女王がモチーフ。
B-54　ローアバッハ郡。
B-55　ポテンツァ・ピチェーナ。
B-56　ムーロ・レッチェーゼ。
B-57　パーチェ・デル・メーラ。
B-58　ディリーパ王の子。
B-59　寒くて動けず。
B-60　ヒーヌムンと呼ぶ。
B-61　誉めてあげようよ。
B-62　褒美を与えた。
B-63　ジャーンシャーという。
B-64　モンペザ伯爵。
B-65　異名はバズーカ。
B-66　ギズボーン生まれ。
B-67　細谷十太夫。
B-68　素足を表現。
B-69　メチャうれしいです。
B-70　知恩院末寺。
B-71　カピラワットゥとも。
B-72　州都はオショッボ。
B-73　マーマウスだった。
B-74　スボティツァ出身。
B-75　とりあえず勝利。
B-76　ゼネバ機構とも。
B-77　ゲジを思わせる。
B-78　ゴアは主張した。
B-79　オンエアーされた。
B-80　ツィター演奏家。
B-81　李百薬の父。
B-82　チャット機能あり。
B-83　モーミー在住。
B-84　コンピュータビジョン。
B-85　諱は光秀。
B-86　スージーを探す。
B-87　社会民主主義。
B-88　場所は現小渕。
B-89　チェアスキー選手。
B-90　幸運を祈る。
B-91　居合道範士。
B-92　シッスィなどである。
B-93　シャウトキーを押す。
B-94　デン・ハーグ生まれ。
B-95　幼虫越冬。
B-96　マフィオが糾弾。
B-97　ウィニペグの生まれ。
B-98　へし折ってしまう。
B-99　バーレーンの村。
B-100　ザースフェーである。
B-101　ファシディウム菌科。
B-102　会議通訳者。

B-103　オーフス在住。
B-104　フィジーのゴルファー。
B-105　フリースタイラー。
B-106　エウフェミアである。
B-107　ジンバブエがある。
B-108　映画著述業。
B-109　武装パーツ類。
B-110　目や皮膚に悪い。
B-111　ティピーに被せた。
B-112　ご当地ムービー。
B-113　姉ヤーラに会う。
B-114　カフェを訪れる。
B-115　トゥラジャープルとも。
B-116　ローツァンパのこと。
B-117　アイヌ語ではピヤパ。
B-118　データベース機能。
B-119　蝦夷に滅ぼされる。
B-120　ポホヨラへ向かった。
B-121　トリアージュとも言う。
B-122　アティテュードなんだよ。
B-123　旧山田町域。
B-124　刻みネギを添える。
B-125　ミャオーまたはクワー。
B-126　立ち位置は中央。
B-127　風防、風除け。
B-128　母はデーイピュレー。
B-129　従五位勲四等。
B-130　愛称はロザミィ。
B-131　美しいゲームよ。
B-132　トゥオーロ山がある。
B-133　フォールスルーはない。
B-134　落ち着いた雰囲気。
B-135　タイガーシャークなど。
B-136　セーブ機能は無い。
B-137　ラグジュアリーホテル。
B-138　ロンドンへ亡命。
B-139　まずビアッジをパス。
B-140　今思えば無謀。
B-141　州都はイーラーム。
B-142　キーロフスキー地区。
B-143　公務員住宅。
B-144　デュシェス・ド・ブルゴーニュ。
B-145　パハール語群とも。
B-146　ゴッホを所蔵する。
B-147　モーゼは否定する。
B-148　サービサーともいう。
B-149　セリーヌはほほ笑む。
B-150　ガードフォワードとも。
B-151　疼痛を伴う。
B-152　リエパーヤ出身。
B-153　首都ファドゥーツ生まれ。
B-154　アッツァーノ・デーチモ。
B-155　フィナーレ・エミーリア。
B-156　バンクーバー生まれ。
B-157　サパテアドを踊る。
B-158　モンジュッフィ・メリーア。
B-159　ベーブ・ルースだった。
B-160　解雇を覆す。
B-161　レンズシャッター式。
B-162　張歩は恥じ入った。
B-163　大型アスリート。
B-164　主部はスケルツォ風。
B-165　ウィキペディアにはない。
B-166　公家植松家の祖。
B-167　従一位右大臣。
B-168　音を増幅する。
B-169　舌動脈の枝。
B-170　小出英亮室。
B-171　ター・ハーを参照。
B-172　オデッサで働く。
B-173　ヘンリー・ウッドなど。
B-174　賀名生へ移された。
B-175　バザー収入など。
B-176　高難易度モード。
B-177　模式種はタヌキモ。
B-178　藤波伊兵衛室。
B-179　助言者を意味する。
B-180　パッチムを用いる。
B-181　ペルー中部の都市。
B-182　スイーパーをこなす。
B-183　エフスターフィイ級。
B-184　古名はスィドルィー。
B-185　女流絵師であった。
B-186　司法省へ出仕。
B-187　やや複雑になる。
B-188　ジアゼパムであった。
B-189　コメディエンヌである。
B-190　うめえなあーと思う。
B-191　ノウハウを教える。
B-192　チェシャ猫の飼い主。
B-193　次郎がプロポーズ。
B-194　刑務所さえあった。
B-195　安土駅から徒歩。
B-196　ベネゼエラに戻る。
B-197　テレボーウリャである。
B-198　ディフューザーを装備。
B-199　ドーバーやイースト。
B-200　バージョンアップする。
B-201　マギステル・ミリトゥム。
B-202　茶室を寄贈した。
B-203　ウェザビーは喜ぶ。
B-204　ネパールの王朝。
B-205　タウィタウィ州の島。
B-206　カーシャーン出身。

B-207　パーンドゥをもうけた。
B-208　同フェスで演奏。
B-209　アイゼンバッハ川。
B-210　ンクドゥとも呼ばれる。
B-211　ボヘミアを旅した。
B-212　ほぞ穴に差し込む。
B-213　ウェイイに寝返った。
B-214　オプス・デイの信者。
B-215　数式エディタ機能。
B-216　メフォディとも表記する。
B-217　欧州議会議員。
B-218　カーアワーと呼ばれる。
B-219　すべて愛媛県道。
B-220　フォアシュピールともいう。
B-221　官位相当は無い。
B-222　ウェブ経由でアクセス。
B-223　県都はビャウィストック。
B-224　江戸幕府大目付。
B-225　キンポウゲを捕える。
B-226　おばあちゃんっ子である。
B-227　第二オーダーである。
B-228　一昼夜放置する。
B-229　爆破処理を行う。
B-230　一言で言えばピュア。
B-231　ウプウアウトとされた。
B-232　ヤハウェを冒涜した。
B-233　シチュエーションを募集。
B-234　俗に言うギャグ漫画。
B-235　空軍は青である。
B-236　レシピは以下の通り。
B-237　非常に稀有といえる。
B-238　いつも言い合っていた。
B-239　ダル・セーニョを用いる。
B-240　ボルゲット・ロディジャーノ。
B-241　サン・パオロ・ディ・イェージ。
B-242　パールサファイアブルー。
B-243　コメッツァノ＝チッツァーゴ。
B-244　空戦用フレーム。
B-245　宇和海に浮かぶ島。
B-246　斬れ味ゲージを持つ。
B-247　ドビュッシー研究家。
B-248　フェザー級チャンピオン。
B-249　小坪隧道のこと。
B-250　ハードウェアエンジニア。
B-251　演説を聞いていた。
B-252　アーサー王の王妃。
B-253　彼女目当てに入部。
B-254　英語読みでパンサー。
B-255　セーフティカーが入る。
B-256　ウィジェットとも呼ばれる。
B-257　フォートワース出身。
B-258　首都はニーシャープール。
B-259　虚言癖を伴う。
B-260　プレーオフ絶望に。
B-261　母フェクラは看護婦。
B-262　ギーラーン語とも呼ぶ。
B-263　ティガも翻弄される。
B-264　サイバーアーキテクト。
B-265　ヌールッディーンの兄。
B-266　山号は象頭山。
B-267　カルディツァ県のひとつ。
B-268　娘ペーノーの父。
B-269　女性のボーカルデュオ。
B-270　テーアを演じている。
B-271　アビトゥーアに合格。
B-272　イェヌーファは喜ぶ。
B-273　ペリメーデーを生んだ。
B-274　クーデターを起こした。
B-275　王城はスウス城。
B-276　ヤズィーディーなどがいた。
B-277　テネシー州メンフィス。
B-278　学名ディモルフォセカ 。
B-279　膝蹴りが得意技。
B-280　ジョアンを厚遇した。
B-281　ボタンウミウサギなど。
B-282　代打でメジャーデビュー。
B-283　別称に、ねね姫。
B-284　まんのう町全域。
B-285　シャパリュ等と称する。
B-286　別名メーサーヘリ。
B-287　ブルーオーバーもある。
B-288　フィブリンへ変化する。
B-289　シドニーへ引っ越した。
B-290　モリーゼ州出身。
B-291　スリーツアーズバギー。
B-292　モアヘッドに敗れた。
B-293　エレン・アーサーである。

B-294　冷えた麦茶を出した。
B-295　ウォーゲームデザイナー。
B-296　キウイフルーツによる。
B-297　アーキャロも同意した。
B-298　主要都市はオプース。
B-299　ウッジャイン県の都市。
B-300　蕾はほぼ球形。
B-301　首都はツァイツであった。
B-302　イアーソーンを生んだ。
B-303　今後も増える予定。
B-304　オーバーダビングした。
B-305　スタッフも同じだし。
B-306　彼に犬を譲った。
B-307　劉子羽は、軍略家。
B-308　水呑百姓がいた。
B-309　暴風雨に遭い座礁。
B-310　ノルウェー領ブーベ島。
B-311　チャアダイ、チャプタイとも。
B-312　軽率でおっちょこちょい。
B-313　モルフェウス、モルペウス。
B-314　政府批判を封じた。
B-315　その保護を得ようとする。
B-316　自叙伝を執筆中。
B-317　トゥアハ・デ・ダナーンの王。
B-318　モーシェは考えていた。
B-319　ウァッロは逃亡した。
B-320　異次元へ退却する。
B-321　ジャコモ・プッチーニである。
B-322　防爆スーツともいう。
B-323　クザーゴまたはクサーゴ。
B-324　ガーナのサッカー選手。
B-325　ツァッディークとほぼ同義。
B-326　ビッグウルフが活躍。
B-327　ファウンデーションの一種。
B-328　県都はディーワーニーヤ。
B-329　同じ場合は省略。
B-330　マンゴーアレルギー持ち。
B-331　カンピリオーネ＝フェニーレ。
B-332　自由の気風が強い。
B-333　サン・フェリーチェ・チルチェーオ。
B-334　撰者は二条為氏。
B-335　ラッピ県南部の都市。
B-336　エミー賞を受賞する。
B-337　アストゥディージョで生まれた。
B-338　艦長はチェホフ大佐。
B-339　下記に車種別で示す。
B-340　モンテルーポ・アルベーゼ。
B-341　モンテマーレ・ディ・クーネオ。
B-342　数万人を数える。
B-343　両耳にはめて眠る。
B-344　ジール王国の王子。
B-345　トリガーフィッシュに因む。
B-346　主要母語はズールー語。
B-347　リチウムを含む雲母。
B-348　スタートも首位をキープ。
B-349　強行輸送に従事。
B-350　藤井家へ養子に行く。
B-351　完璧なプロポーション。
B-352　ベルギービールの一種。
B-353　ジェームズ・ポークであった。
B-354　書風は典麗高雅。
B-355　白毫寺に布陣した。
B-356　何も彼も同様ぞ。
B-357　出走が危ぶまれた。
B-358　ガローチャ郡に属する。
B-359　あるいはケーウクスの子。
B-360　ゴッドファーザーでもある。
B-361　メキシコへ越境する。
B-362　陸軍中尉従七位。
B-363　ガルーでは逆転する。
B-364　イプティーメーをもうけた。
B-365　右肘痛が再発。
B-366　リトムニェジツェ出身。
B-367　プレーから遠ざかった。
B-368　冬眠シェルターがある。
B-369　数百年を要する。
B-370　ジョニーの死の謎を追う。
B-371　ゴツェ・デルチェフに生まれる。
B-372　ターペー郡と接する。
B-373　アゴーギクを造語した。
B-374　天女ツェフボバと出会う。
B-375　母親はミャンマー人。
B-376　マイブームはスペア

リブ。
B-377　ウェッブは捕まえられた。
B-378　トルコではアヒーという。
B-379　ガーズィープル県の都市。
B-380　ラーイセーン県の都市。
B-381　チヌーク・ジャーゴンである。
B-382　元羽は太尉を兼ねた。
B-383　フランスのパリで育つ。
B-384　前方へ跳躍する。
B-385　チロシンキナーゼをする。
B-386　問い合わせをしたという。
B-387　およびアキテーヌ王妃。
B-388　アヌーププル県の都市。
B-389　デーモニーケーを生んだ。
B-390　モチベーションスピーカー。
B-391　マギーとホーピーである。
B-392　ティーンホーフェンにて没。
B-393　エネルギー準位の一種。
B-394　よく祈って決断せよ。
B-395　コメディータッチで描いた。
B-396　交通委員長である。
B-397　オーバーオールのジーンズ。
B-398　伊勢茶栽培地のひとつ。
B-399　顔を縫う負傷を負った。
B-400　こちらも釣る事ができる。
B-401　歌詞はアポリネールによる。
B-402　遊びや塾に忙しい。
B-403　ユニバーシアード出場。
B-404　北条流兵法の祖。
B-405　クウォドニツァ川 に跨る。
B-406　サルデーニャ語ではヌーゴロ。
B-407　上条上杉家当主。
B-408　ビーズ・ニーズと記述する。
B-409　セッティモ・トリネーゼ育ち。
B-410　大手門を造営した。
B-411　オヒョウの木をアッニと呼ぶ。
B-412　サン・ラッファエーレ・チメーナ。
B-413　アエギュプトゥスでは途絶えた。
B-414　そしてウォータールー駅へ。
B-415　メヌアは、ウラルトゥの王。
B-416　マーティン医師が死亡する。
B-417　鉱山や玉泉もある。
B-418　謎としか言いようがない。
B-419　エーマティオーンと兄弟。
B-420　ヌァードゥワーが挙げられる。
B-421　ローマのゲシュタポ長官。
B-422　ロジェ王は反応しない。
B-423　とても球離れが速い。
B-424　冒険ファンタジー漫画。
B-425　ウォーターフォードに大きい。
B-426　テイエムオペラオーが出た。
B-427　優しい普通の少年。
B-428　ヤクゥとドクゥより受領。
B-429　高島藩へ訴えた。
B-430　老朽化と言い続けた。
B-431　姉ジェリーがそれぞれいる。
B-432　死者十数人を出した。
B-433　花鳥諷詠を追求。
B-434　投法はオーバースロー。
B-435　サルペードーンに討たれた。
B-436　イタリア王ピピンの庶子。
B-437　キリスィマスィ島に位置する。
B-438　カフェはベーグルが名物。
B-439　主たるポジションはプロップ。
B-440　ドイツのセム語研究者。
B-441　メルコスールの立法府。
C-1　ショーンを封印。
C-2　右投右打。
C-3　そのドーチェとなる。
C-4　ゾフィーらを救う。
C-5　ユッケジャンスープ。
C-6　無痛性である。
C-7　種数は少ない。
C-8　カスティーリャ女王。
C-9　ガバチョは撃てない。
C-10　パピーミルと言う。

C-11 奥羽の豪族。
C-12 母はジョゼフィーヌ。
C-13 ベビーターンする。
C-14 恵慶法師とも。
C-15 テーベの守護神。
C-16 以後が羽左衛門。
C-17 ディフェンシブハーフ。
C-18 アーチャーは死んだ。
C-19 子に織田信家。
C-20 空襲で被災。
C-21 芽生えを意味する。
C-22 パデルノ・ドゥニャーノ。
C-23 メカアニメーター。
C-24 得手不得手がある。
C-25 ベウラ＝カルデッツァ。
C-26 駅からも遠い。
C-27 林羅山の著。
C-28 ポルト・チェザーレオ。
C-29 プレーリードッグ。
C-30 フィギュア集めほか。
C-31 稲永へ向かう。
C-32 ショツァハ川がある。
C-33 マジェスティック級。
C-34 また事件っスか。
C-35 宗祖は空海。
C-36 そう言い残して。
C-37 宇佐神宮禰宜。
C-38 県都はメルスィン。
C-39 おじいちゃんと呼ぶ。
C-40 イウヌラーがいる。
C-41 オフタースハイム。
C-42 ユーモア溢れる。
C-43 ニッポンイイズナ。
C-44 幡豆小笠原氏。
C-45 ボーイズグループ。
C-46 ボイスパフォーマー。
C-47 県都はマアーン。
C-48 初主演映画。
C-49 老母を頼った。
C-50 草の根運動。
C-51 スウェーデン王。
C-52 出番も激減。
C-53 香椎宮宮司。
C-54 バーンサーイ郡。
C-55 太秦在住。
C-56 ツォリコンにて没。
C-57 ロビン・ウィリアムズ。
C-58 停留所はない。
C-59 バウツェンの生まれ。
C-60 ウルドゥー語詩人。
C-61 アンカーパーソン。
C-62 たろうの妹。
C-63 ダーツプレイヤー。
C-64 坊城を号す。
C-65 ハーモニー参加。
C-66 フィエーゾレ生まれ。
C-67 ミュンヒノンがある。
C-68 ロチェスター生まれ。
C-69 首都はアチャルプル。
C-70 空調メーカー。
C-71 妄想エッセイ。
C-72 州都はガローウェ。
C-73 愛称はベアー。
C-74 合意が進んだ。
C-75 アルバート・マーチ。
C-76 犬よりも猫派。
C-77 駐ペルー公使。
C-78 引退を望む。
C-79 増上寺法主。
C-80 旧姓はウォジャス。
C-81 カリフォルニアポピー。
C-82 ピエスモンテと言う。
C-83 リビュアを飛行した。
C-84 オアフ島へ移住。
C-85 イレーヌに送った。
C-86 フォーフォーズともいう。
C-87 やがて寿命となる。
C-88 情緒豊かである。
C-89 文字多重放送。
C-90 ウンウンエンニウム。
C-91 バハーイー教など。
C-92 同校を中退。
C-93 対義語はチョベリグ。
C-94 血を奪おうとする。
C-95 フードゥーと呼ばれる。
C-96 ピシウム病がある。
C-97 関東由緒あり。
C-98 ボディタイプはクーペ。
C-99 決着をつけよう。
C-100 右衛門兵衛尉。
C-101 野木宮に潜んだ。
C-102 ピューロノエーである。
C-103 ジュエリーアーティスト。
C-104 レブンウスユキソウ。
C-105 考烈王の庶子。
C-106 ピチャフィンウェであった。
C-107 ホップを使用せず。
C-108 フェーザー砲 装備。
C-109 愛称はファビーニョ。
C-110 ボッビオ・ペッリーチェ。
C-111 カーゾラ・ディ・ナーポリ。
C-112 チェレゾーレ・レアーレ。
C-113 ボッファローラ・ダッダ。
C-114 トレッツォ・スッラッダ。
C-115 店舗数も増加。
C-116 愛称はシーミュウ。
C-117 グラディスカ・ディゾンツォ。
C-118 むしろツィターである。
C-119 モントゥ・ベッカリー

ア。
C-120　声はメゾソプラノ。
C-121　サーラ・ビエッレーゼ。
C-122　トゥルグ・ジウ出身。
C-123　ファッジェート・ラーリオ。
C-124　チェッレート・グイーディ。
C-125　あるいは如意宝珠。
C-126　現在閉鎖中。
C-127　メーミィが登場。
C-128　フラーフェで生まれた。
C-129　ωは周波数。
C-130　クロアチアの女優。
C-131　作風を一変。
C-132　ソフィーティアの夫。
C-133　ブラスエデュケーター。
C-134　言わばパロディなんだ。
C-135　ンド族などが住む。
C-136　ウシ、ブタで多い。
C-137　ウティカへ移された。
C-138　テオノエーを生んだ。
C-139　僧侶は応じない。
C-140　千種有能室。
C-141　ペニーはそれを拒否。
C-142　情報メディア法。
C-143　夜明け近づいたぞ。
C-144　いい明日をつくる。
C-145　州都はジーザーン。
C-146　ノースリーブ仕様。
C-147　ポモージェ県生まれ。
C-148　賞品はトロフィー。
C-149　諸部に分配した。
C-150　東道を慰撫した。
C-151　デズフール出身。
C-152　子に毛利包詮。
C-153　こずえは走り出す。
C-154　エベネザー・プレイス。
C-155　文芸評論へ。
C-156　ヘーニオケーの父。
C-157　収税所があった。
C-158　弓削法皇社とも。
C-159　おへぎが作られる。
C-160　トシェビーチ出身。
C-161　西へ枝を伸ばせ。
C-162　オイルシェールとなる。
C-163　アイウォ地区出身。
C-164　憂鬱症だった。
C-165　トィシャツキーである。
C-166　イェファーに生まれた。
C-167　巨大なヘビである。
C-168　パッケージ版向け。
C-169　エセー等創作。
C-170　私のヒーローよ。
C-171　ナディヤー県の都市。
C-172　フェイシャルマッサージ。
C-173　自筆譜蒐集家。
C-174　ジュネーブに滞在。
C-175　パナシェシュティ出身。
C-176　未来へ歩みだす。
C-177　シェアウィズアクトなど。
C-178　胃病を患った。
C-179　イーデー山に捨てた。
C-180　上下線で共通。
C-181　同宇都宮藩主。
C-182　レギュラーコーヒー用。
C-183　アフターケアの義務化。
C-184　怏々たるメンバー。
C-185　エアインテークとなる。
C-186　伊勢路を防衛した。
C-187　未だ邦訳はない。
C-188　デセチュオ島が浮かぶ。
C-189　ゼツェッシオンともいう。
C-190　ラーオダマースの父。
C-191　アルバーノ・ラツィアーレ。
C-192　ニュージャージー州知事。
C-193　ツァーバーフェルトである。
C-194　クウェート国の首都。
C-195　思考パズルの一種。
C-196　ピアノやチェロも学ぶ。
C-197　南米ツアーにいく。
C-198　ハプニングが発生。
C-199　海を見て呟いた。
C-200　別称、ビアジョッキ。
C-201　キーボードプレーヤー。
C-202　ウォーウィーとも呼ばれる。
C-203　諸事情で頓挫した。
C-204　安いギャラで奮闘。
C-205　新ウルガータである。
C-206　エリザベス女王杯。
C-207　自然治癒力はない。
C-208　仮病説という意味。
C-209　ドゥジーノ・サン・ミケーレ。
C-210　フラボーザ・ソプラーナ。
C-211　ログローニョへ戻った。
C-212　スケッジャ・エ・パシェルーポ。
C-213　アルツァーテ・ブリアンツァ。
C-214　ムアーウィヤの盟友。
C-215　続行不能となる。

C-216 閲覧可能である。
C-217 グウォグフに建てられた。
C-218 ランチャー対応型。
C-219 無事ウィーンへ戻った。
C-220 オーウェンディルを殺す。
C-221 ブゼート・パリッツォーロ。
C-222 パチュリ、パチュリーとも。
C-223 アルビノのエゾヒグマ。
C-224 任侠肌の姐御。
C-225 弱火でゆっくり煮る。
C-226 ウジェーヌ・プベルである。
C-227 女房はとっかえる。
C-228 シフィドニツァで始まる。
C-229 ホイールアーチはない。
C-230 右絵図面参照。
C-231 ビュフェやブッフェともいう。
C-232 違う言い訳をする。
C-233 レンダーファームと呼ぶ。
C-234 もう遊びは終わりだ。
C-235 旧姓はベーリェー。
C-236 剛直な枝を出す。
C-237 イーオーとゼウスの子。
C-238 その後プロデューサーへ。
C-239 万太郎を応援。
C-240 受け入れを躊躇する。
C-241 ニューオリンズが舞台。
C-242 可愛いシール集め。
C-243 移動時はトップビュー。
C-244 サーモグラフィーである。
C-245 ムハンマド・アリー・シャー。
C-246 ダービーシャー出身。
C-247 その時怪しい目が。
C-248 池村で自殺した。
C-249 アンジュ湾沿いにある。
C-250 旧朝来町地域。
C-251 伊達騒動で著名。
C-252 両腕からのビーム。
C-253 ジョフリーが王となる。
C-254 朝日山妙見寺。
C-255 エアシャワーを行う。
C-256 堀内氏善の子。
C-257 オーセージ郡である。
C-258 カードをプレイできる。
C-259 ヒューマンインタフェース。
C-260 ミルウォーキー出身。
C-261 ウォルフェーハ駅がある。
C-262 イージーは拒絶した。
C-263 シャー・ジャハーンの后妃。
C-264 そういう文化がある。
C-265 シローヒー県の都市。
C-266 神田神保町へ。
C-267 歯が丈夫なビーバー。
C-268 ハラーハーを注解。
C-269 どう対応すべきか。
C-270 昭王を葬った。
C-271 片足を上げている。
C-272 メシェデに通じている。
C-273 焚書坑儒を起こした。
C-274 ドゥルセ・デ・レチェなどがある。
C-275 フィレンツェにおいてである。
C-276 単純硫黄温泉。
C-277 レピドゥスを敗死させた。
C-278 全てやり方が違う。
C-279 願望を成就させる。
C-280 楚王負芻を捕らえる。
C-281 レジャースポーツが盛ん。
C-282 ピャーは使われていない。
C-283 母はニザーム・バーイー。
C-284 ウチは作りは無茶苦茶。
C-285 プーアル茶のティーバッグ。
C-286 レンジファインダーカメラ。
C-287 プロ初ゴールを挙げた。
C-288 のち、兼イタリア王。
C-289 ルディ・ドゥチュケは負傷した。
C-290 天禄元年卒去。
C-291 シーズン半ばに復帰。
C-292 アーマー進化が可能。
C-293 テューポーンともいわれる。
C-294 ワールドツアーを実施。
C-295 ラゲーン基地副司令。
C-296 タイムマシンを強奪。
C-297 ユーロエアポートがある。
C-298 強いフェーン風となる。
C-299 建王に封ぜられた。
C-300 ドイツ風ミートローフ。
C-301 その経緯は謎である。

C-302　オリビエは覚えていた。
C-303　ドゥアルテなどに対応。
C-304　デール・カーネギーの著書。
C-305　ランペドゥーザ・エ・リノーザ。
C-306　普通ビーンと呼ばれる。
C-307　ファミリー向けスキー場。
C-308　日本の馬術選手。
C-309　チェコ語ではコルナと呼ぶ。
C-310　アッバディーア・チェッレート。
C-311　水分を貯蔵できる。
C-312　ティーガーは応戦した。
C-313　イメージカラーは、青。
C-314　パクティヤー州出身。
C-315　神馬舎へ延焼した。
C-316　ジェモーナ・デル・フリウーリ。
C-317　ツィアーノ・ピアチェンティーノ。
C-318　アーティファクトが生じる。
C-319　ジャパンオープン優勝。
C-320　エジュデルハーと呼ばれる。
C-321　仏法も信奉する。
C-322　ドゥズィッパ川などがある。
C-323　右投げアンダースロー。
C-324　中央軍委秘書長。
C-325　それに準じて繰り下げ。
C-326　王座は空位となった。
C-327　ニーダーシェーネンフェルト。
C-328　ノーマルテープ専用。
C-329　十分謹んでおれ。
C-330　濃厚な風味がある。
C-331　ツェレ包囲中に急死。
C-332　オルーミーイェ出身。
C-333　チェコのアートフォトグラファ。
C-334　すべてがオーダーメイド。
C-335　ハーカー夫妻と出会う。
C-336　打順は五番を打った。
C-337　フォミーンらが挙げられる。
C-338　エントリーフィーは無料。
C-339　鍋パーティーをしている。
C-340　竿は延竿が多い。
C-341　ピンダアースとよばれる。
C-342　よって坐禅石という。
C-343　シービーを含んでいる。
C-344　日本ツアー最終日。
C-345　バターやチーズを食べた。
C-346　メッヒャーニッヒの生まれ。
C-347　頂上駅で下車する。
C-348　ジャズヒップホップダンス。
C-349　ファスティングアドバイザー。
C-350　ウォキーガンへ伸びていた。
C-351　農業上 役に立つ。
C-352　ウェウェテナンゴ県である。
C-353　コルナーレ・エ・バスティーダ。
C-354　そこからは導き出せる。
C-355　山城国と為すべし。
C-356　諏訪頼重の娘など。
C-357　規模は本社より小さい。
C-358　ハープーンを搭載する。
C-359　地方変異も数多い。
C-360　小倉百人一首から。
C-361　一般客は乗車不可。
C-362　居酒屋など数店ある。
C-363　母はアーレーテー王妃。
C-364　流通量は多くない。
C-365　高性能パワークラス。
C-366　セウタのモンテアチョである。
C-367　ウェハースチョコレートをいう。
C-368　徐々に街は寂れていく。
C-369　その父ズーミングゾーン。
C-370　ステージ数で表記する。
C-371　カール・パーマーを加える。
C-372　ざる法という余地がある。
C-373　土俵際へ一直線。
C-374　フォイヤーターレン出身。
C-375　ストシェムホビーで生まれた。
C-376　ミッドウェー島を経由。
C-377　犬種はジャーマン・シェパード。
C-378　エンルートへ引き継

がれる。
C-379　ギリシャ語のヒュペル が語源。
C-380　マチュピチュまで続いている。
C-381　ペルシア語ではヘジャブとも。
C-382　合わせてパジチョゴリと呼ぶ。
C-383　それがチュシ・ガンドゥクである。
C-384　未収録エピソードあり。
C-385　ルピーはセーシェルの通貨。
C-386　ミュセドーラスも捕虜になる。
C-387　異形にツェツィーリエがある。
C-388　ルッツェも重傷となった。
C-389　ベヘール・デ・ラ・フロンテーラ。
C-390　先発ローテーション入り。
C-391　ミャオ族などが多数住む。
C-392　異形にエチェベリアがある。
C-393　大宮のほうがやや上。
C-394　中心都市はホツィムスク。
C-395　厚底ブーツを発表。
C-396　アーデンアーケードである。
C-397　映像はスチール実写。
C-398　ンガプーはこれは拒否した。
C-399　ホラーアドベンチャーゲーム。
C-400　スーフィズムのホージャである。
C-401　アマチュアチームへ移った。
C-402　ラウール・プーニョに献呈。
C-403　小字として広田がある。
C-404　県都はティジ・ウズーである。
C-405　ベドゥム駅が置かれている。
C-406　摂津八部郡で生まれる。
C-407　阿蘇中流域に多い。
C-408　県都はスーク・アフラース。
C-409　当初石津王を名乗る。
C-410　フォックスキャッチャーを去った。
C-411　アルジャンタルへあふれ出た。
C-412　首都はダール、マーンドゥー。
C-413　ゾウパーゲ科も使われる。
C-414　マーダーインクの殺し屋。
C-415　干渉イオンを除去する。
C-416　ペラ州イポーで育った。
C-417　別名ブラッザグエノン。
C-418　テーファーロートが属した。
C-419　ショーパブ勤務を始める。

付録 B

4.1.4 節で選択された音素バランス文

A-1 少ない施肥量が過剰施肥を防ぐ。
A-2 ファイアーはバーナー攻撃をしてくる。
A-3 そのせいで長女がへそを曲げてしまう。
A-4 地酒や地ビールに呼応した名称。
A-5 ピタリとシャフト折れの不具合は止んだ。
A-6 節水を謳う商品も数多い。
A-7 早速レコード作りへ動き出した。
A-8 刻印付き円銀を流通させた。
A-9 道端や荒れ地などに生える雑草。
A-10 なぜ少女雑誌を出版しないのか。
A-11 邪悪な美しい女王とも言われる。
A-12 こちらはほぼ白青のツートンカラー。
A-13 ボギー車はエアブレーキを常用した。
A-14 共通通貨ユーロを誕生させた。
A-15 もうペニーロイヤルなんぞどうでもいい。
A-16 列車に乗り遅れたりで右往左往。
A-17 父は小作農兼ぞうり売りだった。
A-18 牛そり、犬ぞり、牛馬も同様。
A-19 あるいは罪人が憐れみを乞う歌。
A-20 エアバッグの意匠一部変更など。
A-21 かまぼこ兵舎が多数造営された。
A-22 議員って野郎どもはそういうもんだ。
A-23 ほぼキープコンセプトでのモデルチェンジ。
A-24 アウターループを通過し森の中へ。
A-25 機銃座も数箇所据えつけられていた。
A-26 やめちゃうなんて本当に残念だわ。
A-27 彼は両腕を前方へ突き出した。
A-28 場所中初めて星を五分に戻した。
A-29 解体後は骨や所持品を焼却。
A-30 大筋のストーリーは共通である。
A-31 長考派で重厚沈着な碁風。
A-32 型は一切メソッドに依存しない。
A-33 漢音や呉音、音読みを参照。
A-34 ボルシェビキを同地で虐殺している。
A-35 キャッチフレーズは、世界のホビーハウス。
A-36 デュアルエアバッグが標準装備された。
A-37 ナンバーワンアルバムが多いアーティスト。
A-38 同王座挑戦を改めてアピール。
A-39 合唱のオーディションと間違えて応募。

A-40 ほぼパーフェクトな演技で総合首位に。
A-41 細胞死や臓器不全をも起こしうる。
A-42 樹皮は淡い灰茶色ではげ落ちやすい。
A-43 ダブルルーフの屋根で前面は切妻。
A-44 ドゥーマンは母語として日本語を覚えた。
A-45 シャンプー、シェービングクリーム等である。
A-46 望遠スコープとサーモスコープがある。
A-47 フィールディングもよいオールマイティな投手。
A-48 ドキュメンタリーアドベンチャー映画である。
A-49 出生直後よりチアノーゼを呈する。
A-50 そのままアポイントを放置してしまった。
A-51 声調符号は主母音の上につける。
A-52 カメや野鳥やアヒルの餌となっている。
A-53 兵糧は尽き、野草で餓えをしのいだ。
A-54 個人での受賞はメンバー別ページへ。
A-55 ビュッフェで飲食物を買うことができる。
A-56 労役や出費の軽減を具申した。
A-57 ボディーペインティングでの実績も多い。
A-58 学生女流アマチュアの頂点に立つ。
A-59 例としてパパイヤ賞やグレープ賞など。
A-60 黄色いウサギをポスター用紙に描いた。
A-61 空襲情報伝達網に加わった。
A-62 鶯の鳴き声はフルートで示される。
A-63 浮世絵および絵入版本の研究者。
A-64 テレパシーで相手との意思疎通ができる。
A-65 アニメコーナーとダジャレコーナーを放送。
A-66 女子はセーラー服にシアンブルーのリボン。
A-67 筆力粘り強く、落ち着きある書風。
A-68 ポリエチレン、ポリプロピレンなどが主流。
A-69 味付けはこんぶと醤油で薄味にする。
A-70 王国へ旅に出るアドベンチャー作品。
A-71 スフィンクスコースはファミリー、初心者向け。
A-72 王妃は王子と謀って王を幽閉した。
A-73 メジャー系プロレスファンの憎悪を浴びていた。
A-74 スピーディーさやスマートさをイメージしている。
A-75 彼を祀って所願成就を願う信仰。
A-76 非合法な仕事を請け負うプロフェッショナル。
A-77 選者中で唯一ネガティブな評価をした。
A-78 押しも押されもせぬエースの座に君臨する。
A-79 メジャーからマイナーアーティストまで幅広い。
A-80 あれよあれよと勝ち抜いてファイナルに進出。
A-81 本路線初の運賃値上げを実施する。
A-82 ヘッダはあて先とメッセージタイプを含む。
A-83 アルバムリリース前にショートツアーを実施。
A-84 夜、寝る前に部屋でビュンビュンとバットを振る。
A-85 アトリエやギャラリーなどを誘致するプロジェクト。
A-86 勉強はせず芸者遊びに明け暮れ中退。
A-87 アナウンサー室から社長室秘書部へ異動。
A-88 ケーブルを売るメーカーがロイヤリティを支払う。
A-89 ルポルタージュを中心に著書を多数執筆。
A-90 ニャーニャー人形を合体素材にすればできる。
A-91 エレベーターと上下エスカレーターを設置する。
A-92 そこに世情不安や天変地異は含まれない。
A-93 いわゆるゴージャスでファンシーなサウンドが持ち味。
A-94 雌しべが長くほぼま

っすぐに伸びるのが特徴。
A-95 ボディ中心の攻撃で中盤以降圧倒。
A-96 スーパーファミコンへとプラットフォームを移された。
A-97 ほんと久しぶりにすっごい悔しかったんですよ。
A-98 ディレクター兼任アナウンサーとしても活動。
A-99 踏み出した足の膝をやや落としてウェートを乗せる。
A-100 再びジュニアタッグ王座への挑戦をアピール。
A-101 全身にファスナーが付いたオーバーボディを着用。
A-102 鉱物性油は重油や軽油、灯油である。
A-103 カーネーションのメンバーをサポートに迎えてツアー。
A-104 ウサギがぴょんぴょん跳ねる擬態語のぴょんと合わせたから。
A-105 腎盂炎、蛋白尿、神経痛を病んでいた。
A-106 レザーシートなどがパッケージでオプションとして選べる。
A-107 ルビー山脈およびシェルクリーク山脈の中にある。
A-108 事務所に言われオーディションを渋々受ける日々が続く。
A-109 かつての校風はすっかり消え失せ大きく変化した。
A-110 有象無象が舞い踊る空前絶後の武闘大会。
A-111 アゲハチョウ科アゲハチョウ属に分類されるチョウの一種。
A-112 このペア以降、コンビネーションジャンプを飛ぶペアが増えた。
A-113 優れたフィジカルを活かしてディフェンダーでのプレーも可能。
A-114 モカ、カプチーノ、エスプレッソ、シフォンについては省略。
A-115 マンゴーなどの新しい風味のバラエティーも増えつつある。
A-116 フィギュア、キーホルダー、文房具などのイラストをデザイン。
A-117 ドルフィンとケープハート、双方劣らぬしたたかさだった。
A-118 部署異動して実務にも急遽務め、アナウンサーに復帰。
A-119 尾を使った泳ぎ方はウナギやウミヘビなどと共通する。
A-120 右舷船首部および船尾中央にランプウェイを装備する。
B-1 金融コンツェルンもコンツェルンである。
B-2 ガッツあふれた守備でファンを魅了した。
B-3 所持しているアークを全て奪われる。
B-4 失ったシェアを取り戻しつつあった。
B-5 ペダルは左がしゃがみ、右がジャンプ。
B-6 穏やかな風味のウイスキーが多い。
B-7 その後双方が交互に二手ずつ打つ。
B-8 売れない兄弟デュオがカフェを始める。
B-9 エイ、カレイ、イセエビを水揚げする。
B-10 家の前を暴力団がうろついた。
B-11 のべぼう系アイテムのみ売値がつく。
B-12 草庵風茶室の源流とされる。
B-13 いぶし銀プレーヤー賞を受賞した。
B-14 バックアップメディア所属のディレクター。
B-15 しばしばうわべだけの即興を含む。
B-16 熱心なホビーユーザーの支持を得た。
B-17 レギュラーチームの人数が減ったため。
B-18 切り刻まれたあのポシェットを見つけた。
B-19 抜き足差し足でチーズを盗み出す。
B-20 寝技柔術の競技部門を作る。
B-21 ジャポニズムブームの一翼を担った。
B-22 いわゆる古文復興運動である。
B-23 末は大関横綱に必ずなる。
B-24 ジャンルは黄金体験アドベンチャー。
B-25 鉄道グッズを扱う雑誌

である。
B-26 違うエネルギー状態へ遷移する。
B-27 プライベートジェットをチャーターし帰京。
B-28 再びオッズメーカーの椅子に就いた。
B-29 瑞穂残務処理事務所は撤去された。
B-30 来客があれば時を選ばずウィスキー。
B-31 不穏な気分を一層増長させる。
B-32 産卵直後の卵嚢は薄青色。
B-33 南北で別々のウォンを発行した。
B-34 どういうわけか人より寿命が短い。
B-35 港湾エリアまでビーバーを輸送した。
B-36 耳は立ち耳、尾はふさふさした垂れ尾。
B-37 そしてワイヤー入りのフープを用いない。
B-38 アプリケーションにビデオ通話を導入。
B-39 神社が氏子などに頒布する授与品。
B-40 本尊は観世音菩薩座像である。
B-41 登録ポジションはディフェンダー、フォワード。
B-42 湯茶の入る茶碗の下に敷く受け皿。
B-43 ディナーパーティーステークスと名付けられた。
B-44 廊下や壁は緩やかに波打っている。
B-45 色鉛筆、ペン、シャープペンシルなど。
B-46 根強いファンの要望を受けて復活。
B-47 それぞれ北西方面へ徒歩数分。
B-48 ジャンプ部長兼ヘッドコーチに就任。
B-49 傍ら中外医事新報社に入社。
B-50 カフェ、ホテルなどが順次オープン予定。
B-51 老翁を責め上げるだけで十分らしい。
B-52 義援金保全法、義援金保護法。
B-53 常に大声を張り上げチームを鼓舞した。
B-54 雷雨は特に州東部と北部で多い。
B-55 ジェフィメンコ方程式として与えられる。
B-56 ウイスキー蒸留所の閉鎖が相次いだ。
B-57 厳密に言えばワークシェアリングではない。
B-58 ペンネームを逆から読むと本名になる。
B-59 サラダフォークやバーベキューフォークなどがある。
B-60 電流分布図をプラズマに適応した。
B-61 区間運転および直通運転あり。
B-62 上がパーソナリティ、下がコメンテーター。
B-63 ローラースケートを履いて行うホッケー。
B-64 シンセサイザーやキーボードを演奏する。
B-65 エネルギー量もおおよそ平均値である。
B-66 花粉うんぬんについては記述されていない。
B-67 インタビュー集やルポルタージュ等を執筆。
B-68 各郡は立法府委員を選挙で選ぶ。
B-69 はせべや応援ガールなどは登場しない。
B-70 雑でちぐはぐな印象を増幅している。
B-71 ピットレーンやパドックなどの諸設備である。
B-72 個人ユニットあやめ十八番を旗揚げ。
B-73 頭に湯気が立つほど喜びのぼせ上がる。
B-74 オーバーホールのためフィラデルフィアへ向かった。
B-75 デフォルメは大袈裟にギャグタッチで描かれる。
B-76 現場で通報者に会い事象を把握せよ。
B-77 エージェントに所属しアーティストビザを取得。
B-78 あのうすぎたねえ家を借りるからそう思え。
B-79 一部アイテムはグラブジャンプで防御可能。
B-80 昆虫類は飛翔して捕らえ樹上で食べる。
B-81 数兆円規模の政府収益を生んでいる。
B-82 ぴょんぴょん飛び跳ねながら女子の方へ向かうもの。
B-83 おいしいレシピだけでなく、まずいレシピもある。

B-84 レース終盤までデッドヒートを繰り広げた。
B-85 他のプレイヤーとスコア数で競い合うモード。
B-86 ファンタジー色の強い集団ヒーローアニメ。
B-87 窒素、ヘリウム雰囲気中で安定である。
B-88 合格者には中規模のトロフィーが授与された。
B-89 めしべに他家受粉が起きなければ散ってしまう。
B-90 そのままスコアは動かず前半を折り返す。
B-91 アニメキャラクターのトレーディングフィギュアのシリーズ。
B-92 グラフィックデザイナーを経て絵本をかきはじめる。
B-93 同月下旬に救援要員へ再転向。
B-94 ルーキーイヤーに次ぐ成績でシーズンを終えた。
B-95 無病開運、来世は必ず仏果を得べし。
B-96 ホイールと車体をはじいて壁沿いに停止した。
B-97 ポジションは右ウイング、右インサイドフォワード。
B-98 そのためアンケートの意義を問う住民も多い。
B-99 脱皮直後の節足動物は無防備である。
B-100 カードはすべてディーラーがオープンするのみである。
B-101 通常の賭けポーカーよりも勝負がつきやすい。
B-102 写像も座標を用いて表現することができる。
B-103 無人インフォメーションカウンターも設置されている。
B-104 思わぬハプニングで宇宙船の機能がフリーズ。
B-105 バーツへ銃を突き付け、そのまま右目を撃ち抜く。
B-106 ビビンパ、ピビンバ、ピビンパなどとも表記される。
B-107 これに携わる者たちをファシリティマネジャーという。
B-108 このほかベビーベッドやベビーチェアも設置している。
B-109 いくえちゃんが次女、めばえちゃんが末っ子とされている。
B-110 プレミアムグッズが当たる抽選応募券を封入。
B-111 ティーンエージャーをターゲットとしたカルチャー番組である。
B-112 スパゲティ、ピザのローテーションとインタビューで語っていた。
B-113 陸上渦巻ポンプ、水中ポンプの専門メーカー。
B-114 メールはバディの赤ちゃんの誕生パーティへの招待であった。
C-1 プッシュ型でメッセージが着信する。
C-2 とぼけた味わいのコメディ系がメイン。
C-3 受注量や収益状況が悪化。
C-4 同時に青いゲージが左へ増える。
C-5 特技はワープロの早打ち、卓球。
C-6 与党はブリッジ法案は取り下げた。
C-7 手足や鼻、唇は紫色。
C-8 夫ポールに離婚を言い出していた。
C-9 その後に臥牛館道場へ移った。
C-10 知恵を揮ってその状況を打破する。
C-11 冷たい北風がピューピュー吹いてくる。
C-12 金箔などに比べ貼付が困難。
C-13 通称ホームヘルパーまたはヘルパー。
C-14 命知らずのパフォーマンスアーティスト。
C-15 ギブスは完全にレゲエにシフトした。
C-16 子供が遊ぶ巨大なオブジェがあった。
C-17 頭部は金色がかったオリーブ色。
C-18 通常ビュッフェと飲み物が供される。
C-19 ホラーのリーディングカンパニーとなった。
C-20 かつてムーアやヒースに覆われていた。

C-21 衰えを知らぬ歌声を披露した。
C-22 まずソプラノに主導される冒頭部。
C-23 握力の低下や疼痛を生ずる。
C-24 カーブやチェンジアップも操っていた。
C-25 バスウッド展望塔の付け根である。
C-26 一方、皮膚病変は癌化しない。
C-27 生傷、青あざ、骨折は普通。
C-28 世界ウォーキングデーを主催している。
C-29 墓地の中を通って旧道へ下りる。
C-30 衣装に新たな布を縫い付けて演技。
C-31 バリアフリートイレや切符売場がある。
C-32 頭を打ったうえ膝から出血した。
C-33 麻雀を打つというモダンボーイだった。
C-34 療養のために硫黄温泉を目指す。
C-35 風味を増すためピーナッツが添えられる。
C-36 人々と牛を不運や病気にさせる。
C-37 見える像は上下左右逆像である。
C-38 オープン当初の愛称はあんあんむら。
C-39 キャンペーン目的のグッズ等も多い。
C-40 ルーペを通して本を読むという趣向。
C-41 エーカージャであり、輪廻しないとされた。
C-42 ミュージアムショップやカフェテリアも充実。
C-43 後頭部は鮮やかな赤茶色である。
C-44 フォアドリブルに比べスピードは出にくい。
C-45 ロボを全員集めて競技を行う。
C-46 頭や首にも若干羽毛が残る。
C-47 ゾーン北部にはウエスタンエリアがある。
C-48 長編コンペティション部門にノミネート。
C-49 下馬評を覆し決勝へ進出。
C-50 右へ真一文字作法で切腹した。
C-51 シェパードはエッジフィールドにおいて死去した。
C-52 テレタイプインターフェースを増設可能。
C-53 猛烈に馬を農場へ疾走させた。
C-54 パワーユーザー向けという位置づけであった。
C-55 ローンパイン、ビッグパイン、ビショップなど。
C-56 大型種は鳥類や哺乳類も食べる。
C-57 列車上空をヘリコプターが追尾する。
C-58 以後もシードペアを連破して勝ち進んだ。
C-59 エメラルドボア、グリーンボアとも呼ばれる。
C-60 貯蔵タンクや工場の壁も破損した。
C-61 ポールはそんなプレッシャーとは無縁だった。
C-62 ジェスチャームービーと粘土ムービーからなる。
C-63 ホームランを何本打てるか競うモード。
C-64 きょうは心ゆくまでプレーをエンジョイした。
C-65 薄いベージュのジャケットにスカート、パンツ。
C-66 ボア科ツリーボア属に分類されるヘビ。
C-67 ウイルス粒子がエンベロープに複数個。
C-68 バスルームにはシャワーブース、テレビも完備。
C-69 その後デザインアクセサリーをプロデュースした。
C-70 ディメンション表はファクト表と結合される。
C-71 腕試し編オーダーは実力が問われる。
C-72 イセエビの水揚げ量日本一を誇る。
C-73 衛士長、衛士副長及び衛士が置かれる。
C-74 マップ上でエフェクトアニメが挿入される。
C-75 グラフィックデザイナー、ディレクターを担当。
C-76 スポンサーより支援を受け運営している。
C-77 レギュラーパーソナ

リティは前シリーズと同じ。
C-78 雄大な眺望にあざやかな色を添える。
C-79 輸送車両やジープを次々と撃破した。
C-80 地獄で亡者を責めさいなむ獄卒のひとつ。
C-81 東西を往復するロープウェーなどもあった。
C-82 その頂点に達するとスケルツォ主部へ戻る。
C-83 数多くのコンビをタッグ王者に導いた。
C-84 総社員数十数人のガチャガチャメーカー。
C-85 画像はほとんどモノクロームかつファジーである。
C-86 ソファーが突然動きだし、子供は驚く。
C-87 フェンダーアーチが盛り上がった精悍な風貌。
C-88 趣味はヨガ、スノーボード、スキー、絵本集め。
C-89 空母保有への命脈はかろうじて保たれた。
C-90 またサブディスプレーはタッチ操作に対応する。
C-91 公式ジャンル名は妄想科学アドベンチャー。
C-92 弾道フェーズから滑空フェーズへと移行する。
C-93 素材からピッツァを作りあげていく課程を追う。
C-94 開幕から絶不調でマイナー落ちを経験。
C-95 そしてアヘンの徴税請負で利益を目指した。
C-96 そしてケチャップやマヨネーズ、生肉をぬりつけた。
C-97 デコーダはファームウェアやソフトウェアで実装される。
C-98 突如マフティーを名乗る集団にハイジャックされる。
C-99 同町長名義の借入申込書を偽造。
C-100 バスおよびフェリーの交通網も運営している。
C-101 美少女フィギュア的なタイプの外装パーツを持つ。
C-102 対処方法も牛乳アレルギーと同じである。
C-103 多種多様なものをほぼオーダーメードで製造する。
C-104 足をハーネスの中に収納してファスナーで閉じる。
C-105 ゼラチンフィルターやアセテートフィルターがこれに当たる。
C-106 突如トロンボーン、チューバのファンファーレが鳴り響く。
C-107 スプーンの背の上で小さなピッチャーの外で注ぐ。
C-108 一度もワーストに入ることなくフィナーレへ進出。
C-109 カーソルキーまたはテンキーで王子を左右に動かす。
C-110 とりあえず汗を流してやろうと、子供を湯屋へやる。
C-111 続編というよりもバージョンアップ版と言う方が近い。
D-1 敢えてプロサッカーへの道を選んだ。
D-2 イオウ成分を含まないため無臭。
D-3 同僚たちがおりなす呑兵衛コメディ。
D-4 ゲシュタポ本部での尋問に応じた。
D-5 イエローゾーン、ブルーコロニーのボス。
D-6 党中央文書保管局に勤務。
D-7 生地商社や素材メーカーを訪ねた。
D-8 ジャージーの空港の一覧を示す。
D-9 極端に図々しい一面も持つ。
D-10 隠れた温泉として情緒あふれる。
D-11 リレー方式でゴールを目指すモード。
D-12 リテールプロモーションアワードを受賞。
D-13 またはカモフラージュセラピーと呼ばれる。
D-14 腎盂膀胱吻合術を行う。
D-15 岩を隔てた穴へ生き埋めにされた。
D-16 序盤から打ち合ってペースを掌握。
D-17 花粉は風で運ぶ風媒花である。
D-18 この犬はプードルの先祖にもなった。

D-19 ウィンにカーブとチェンジアップを教えた。
D-20 メビウスの成長を促そうとする。
D-21 塗装は暗いスチールブルー一色。
D-22 痛みを分かち合う癒しの場を主宰。
D-23 また俗にいう自惚れに相当する。
D-24 トレーラーに乗せられ集合地点へ。
D-25 ポジションはピッチャーでサウスポーだった。
D-26 パラソル、パーゴラ、シェードなどがある。
D-27 決勝でもワンツーフィニッシュを果たした。
D-28 手術と現役の続行へ踏み切った。
D-29 ピッチャー返しを狙うなどが挙げられる。
D-30 チェロの女王は玉座でウトウトしていた。
D-31 作風全体としてもパロディは多め。
D-32 ビューティーケア用品を扱うブランド。
D-33 主にジャズ調の曲をレコーディングした。
D-34 新オープニングよりイブニング復活。
D-35 一方はもう一方より小さかった。
D-36 王立舞踏アカデミーを創立した。
D-37 大幅な賃上げは難しい状況。
D-38 そして大きく落ちるチェンジアップが武器。
D-39 チークウッドの所蔵美術品となった。
D-40 番組テーマカラーやテロップはブルー。
D-41 苦戦の末祝家荘を攻め滅ぼした。
D-42 文字色が青や赤ならばチャンスアップ。
D-43 農村風景の中にぽつんとある駅。
D-44 犠牲者へのレクイエムというべきアダージョ。
D-45 以後、徐々にバリエーションを増やしつつある。
D-46 うなり声を上げたという言い伝えがある。
D-47 一番多いポピュラーな名前を選んだ。
D-48 完璧な守備法を持つプレイヤーもいる。
D-49 統合売上税も逆進税である。
D-50 ひどいペナルティが発生する場合もある。
D-51 同時に受注できるオーダー数が増える。
D-52 戦略パターン暗記型のゲームである。
D-53 映像は木馬座製作の影絵だった。
D-54 シェパードの死後に、マーガレットが寄付した。
D-55 せめてその権威失墜を防ごうとした。
D-56 暴動での犠牲者と遺族へ謝罪した。
D-57 是々非々で行うべきだと主張している。
D-58 コンピューターセキュリティ上の手法の一つ。
D-59 彼らに脅えてさらにウィスキーを煽った。
D-60 数百年以上邪神につかえている巫女。
D-61 壮絶な暴れっぷりで馬場を蹂躙した。
D-62 最年少リーダー録音記録を樹立。
D-63 まぁ、いいや、って思ったら、もう終わりよ。
D-64 カード数を大幅に増量した続編。
D-65 略綬は通例ボタンホールに佩用する。
D-66 普通六つ以上のバッジを所有していた。
D-67 特技はピアノ、ビーズアクセサリー製作。
D-68 右シェークハンドの両面裏ソフトラバー。
D-69 如意棒を手に往年の演武を披露した。
D-70 顎鬚や口髭、眉毛もふさふさしてる。
D-71 以後数年アレンジやツアー等で活動。
D-72 水平あるいはやや斜め上向きに開く。
D-73 女優を目指し演技経験ゼロで上京。
D-74 オフェンスのキャッチ時にディフェンスが起こすファウル。
D-75 デート終了後、スタジオでファイナルジャッジ。
D-76 おおよそ吸収ピークエネルギーに等しい。
D-77 議案は議長オフィスを通じて提出される。
D-78 メジャーメーカーはへ

ッドデザインはコピーしない。

D-79 時折ジェネレーションギャップに悩むこともある。

D-80 調性が不安定で短調ゆえに陰鬱。

D-81 ピヨピヨ鳴いたり翼を羽ばたかせたりもする。

D-82 守衛番長補を守衛班長に改めた。

D-83 突破力を生かしたオーバーラップが持ち味。

D-84 主にゼリーやバウムクーヘンを製造している。

D-85 これも往時のウッドフレームを彷彿とさせる。

D-86 各種の周辺ハードウェアもエミュレートしている。

D-87 人々が集い、笑顔でケーキを食べ、踊る。

D-88 レギュラーアナウンサーが以下のメニューをプロデュース。

D-89 保有台数と店舗網は日本一の規模。

D-90 ダービーは浜辺の家で見ながら笑みを浮かべる。

D-91 突如地球に現れた謎の巨大宇宙船。

D-92 荘厳と神秘、力強さと天衣無縫さ。

D-93 ナチュラルファウンデーションを経て、モチーフを設立。

D-94 最早往年の出来映えは見られず不評に終わる。

D-95 チームをリーグ上位に導く手腕を披露する。

D-96 現在でいうティッシュペーパーのようにも使用した。

D-97 二人は怪しげなヒッピーコミュニティに遭遇する。

D-98 サブシェルの如何なる変化もメインシェルへ与えない。

D-99 ウールのカーペットやヘッドライニングが奢られた。

D-100 そのポエジーを推測し、受け継がなければならない。

D-101 集合インターホンによる異常通報が対応。

D-102 ウェディングフラワー、フラワーギフトも取り揃えている。

D-103 一覧表示はフォローユーザーへの追加時間順。

D-104 リウマチ、痛風や皮膚病に大きな効果がある。

D-105 タピオカパールやゼリー入りのものも販売されている。

D-106 チェッカー間際にピットアウトし、チェッカーを受けている。

D-107 ほぞ穴に比してほぞを大きめに作って嵌め合わせる。

D-108 合法的なコピーをもつユーザーは、通知は受けない。

D-109 実際にオンエアは数秒で後姿だけの出演。

D-110 ギフトコーナーや婦人雑貨などを扱う売り場を備える。

E-1 カフェインレス、カフェインフリーとも呼ぶ。

E-2 コーディネーターがプレゼンテーションする。

E-3 ええ旅プランを求めてリサーチ旅。

E-4 序文、緒言、序論などを意味する。

E-5 アドベンチャー形式の任務に挑む。

E-6 恋愛エッセイを数多く執筆。

E-7 醤油醸造業と漁業で発展。

E-8 民芸茶屋やパブなどを経営した。

E-9 ファームウェアアップデートは失敗する。

E-10 このバージョンはバッファに格納される。

E-11 サーバが応答パケットにコピーする。

E-12 往々にして強風に遭遇する。

E-13 冬は幼虫や蛹で越冬する。

E-14 いずれもペーパープランに終わっている。

E-15 論調は中道右派、もしくは右派。

E-16 地上では鈍い女子社員を装う。

E-17 餅つきを行う足踏み式の臼。

E-18 ブリッジした状態でフォールを奪う。

E-19 カールを打ち破って王座を奪った。
E-20 独奏チェロとファゴットがそれに続く。
E-21 ティー以外なら名前はなんでもいいわ。
E-22 その途端、自分が極楽へスーッ。
E-23 一座の座長として地方を巡業。
E-24 舟運が交通の動脈であった。
E-25 夢への大きな第一歩を踏み出す。
E-26 シンフォニーコンサートも行っている。
E-27 ゼーゼー言うような大きな声で鳴く。
E-28 はじめに北方軍閥を迎え撃った。
E-29 ルーズリーフファイルに入れて保管される。
E-30 長距離用宇宙船のノウハウを得る。
E-31 ソフトウェアのバージョンアップで対処した。
E-32 バター、卵およびベーキングパウダー。
E-33 自我やアイデンティティの超越だとされる。
E-34 ワールド柔道ツアーには含まれない。
E-35 乙級リーグへの参加を可能にする。
E-36 架空の文字アースによって表記される。
E-37 そのまま順位を維持しシーズンを終えた。
E-38 グレーのスーツにちょびひげなしで登場。
E-39 射撃は射手足下のペダルで行う。
E-40 武器は塩酸を発射するレーザー銃。
E-41 ひざを負傷しシーズン絶望となった。
E-42 アーティチョークの利権を部下に奪われた。
E-43 ウィスキーをベースとするレシピが有名。
E-44 冬は男女ともブレザーを着用する。
E-45 彼が唯一得た政府ポストとなった。
E-46 趣味は切手集めとテディベアのコレクション。
E-47 キッカーなどパークアイテムもいくつかある。
E-48 ブティックやジュエリーショップなどが入居する。
E-49 その主要ポストは野党議員に譲った。
E-50 アマリロ地区にてプロレスラーとしてデビュー。
E-51 ディーゼルエンジン車がターボつきに統一。
E-52 それぞれ強奪した金品を浪費した。
E-53 幼虫は触れるとピシッという音を出す。
E-54 コケの生えた身体をシェイプアップさせた。
E-55 空軍少尉兼レーダー技師を務めた。
E-56 異常部位へのエネルギー供給を増やす。
E-57 ボヘミアン達を一網打尽に捕らえる。
E-58 母親は主婦で父親は美容師だった。
E-59 以前家出をした実家へ一時居候。
E-60 みづほもかつては防除班に所属していた。
E-61 プレイヤーが教えずとも豊富な語彙を持つ。
E-62 酷い冷え性でクーラーも苦手だとのこと。
E-63 メンバー共通のモチーフは地上動物。
E-64 ウオータージェットを搭載するタイプもある。
E-65 いつもはちょっぴり、おっちょこちょいな女の子。
E-66 種無しウエハースや普通のパンを用いる。
E-67 睦月は轟沈寸前の窮地に陥る。
E-68 流通在庫をもって売り切り終売となる。
E-69 バラの色は薄いピンクから薄い赤の内。
E-70 趣味は旅行、ぬいぐるみ集め、犬と散歩。
E-71 ヨットクラブ、保護地や保存地も数か所ある。
E-72 いや、そういう脆さがあるうちはダメでしょうね。
E-73 精舎建立と仏像安置を発願する。
E-74 正門辺りはキャンパスで一番古い場所。
E-75 情報や知識の順序付け戦略でもない。
E-76 所有する暗号キーの特許を譲り受けた。
E-77 脳出血の発作による心

不全で死亡。
E-78 初めは本茶と非茶を比べ当てる遊びだった。
E-79 瞼を黒く塗ってピエロ風の格好をする。
E-80 勲一等に叙せられ旭日大綬章を受勲。
E-81 または解除ゲーム数で概ね判別出来る。
E-82 プラズマスパークのエネルギーコアを強奪する。
E-83 風を自在に操って冒険を進めてゆく。
E-84 デュアルエアバッグとヒーター付きシートを備えた。
E-85 補助魔法を使いこなすオールマイティーな兵士。
E-86 小型のアクチュエータやモーターへの応用もある。
E-87 後衛パーサーは遮蔽板を前方にばらまく。
E-88 初々しいピュアな純愛模様も描かれている。
E-89 マグネシウムイオン、カリウムイオンなどを保持する。
E-90 その後ファーストウエーブを経て、現事務所に所属。
E-91 犬グッズショップ、夏場は犬用プールを備える。
E-92 アイディア、アイデンティティを重視するアプローチである。
E-93 イベントを通知するオブジェクト側のインターフェース。
E-94 ティーファを連れて魔物討伐の冒険へと向かった。
E-95 オアシスはオーディエンスを惹きつける力を持っていた。
E-96 全日本フェンシング選手権で、エペ個人優勝。
E-97 レシーバー右後部の大型レバーへ変更された。
E-98 レンズシャッターはレンズ付近に位置するシャッターである。
E-99 復興チャリティイベントのナビゲーター等も務めている。
E-100 サブルーチン、プロシージャ、コルーチンにおけるネスティング。
E-101 それぞれの製品を釜焙茶、籠焙茶といった。
E-102 数百フィートから数千フィート程度しか伝達しない。
E-103 アンチテーゼが存在しなければアウフヘーベンは起きない。
E-104 ゲストというよりはウィークリーパーソナリティ同然であった。
E-105 その他はストップウォッチ使用で秒単位の消費は切り捨て。
E-106 スペシャルプログラムやミュージックビデオスペシャルをオンエアする。
E-107 合流バッファ、リソースバッファという概念を導入した。
E-108 もう一方はハート型の錠前を持ち再び中央へ。
F-1 主なユニフィケーションは数種類ある。
F-2 特技は腕相撲、趣味はマッサージ。
F-3 フォースアウトのことは封殺ともいう。
F-4 可変幅フォントのサポートなどがある。
F-5 同日より注文受付を開始。
F-6 小さな子どもがチョロチョロ入ってきた。
F-7 頂点をノード、枝をアークと呼ぶ。
F-8 ストーリー重視の長編ファンタジー。
F-9 商品および所持金を没収した。
F-10 ボスゲームはアーケードモードと同じ。
F-11 無茶をする姉を後ろからフォローする。
F-12 趣味はサッカー、料理、フィギュア集め。
F-13 テーマを設けて、メッセージを募集。
F-14 のち家族と共にダービーへ移った。
F-15 古風で素朴な作風を打ち立てた。
F-16 単位波長あたりのエネルギーである。
F-17 前方へ延伸し頬骨と接する。
F-18 好きな食べ物はあんみつやポテトチップ。
F-19 ビーチバレーなど遊びを行うレジャー。
F-20 オブザーバーシートの

空間も得られた。
F-21 ファゴットのカデンツァが浮かび上がってくる。
F-22 レスリング選手、元アマチュアボクサー。
F-23 ベーキングパウダーを多めに入れて焼く。
F-24 ギスギスした雰囲気で翌日のロケへ。
F-25 運休や大幅な遅延が相次いだ。
F-26 思わぬ形でプロ初セーブを挙げた。
F-27 美しい黄色い花を持つポピュラー種。
F-28 東西に長いフィヨルド状の湖。
F-29 オプションのパズルは島のあちこちにある。
F-30 ついにゲシュタポが彼の家を訪れた。
F-31 テニスウエアにパステルカラーを導入。
F-32 多重通信、多重伝送とも言う。
F-33 大学ではワンダーフォーゲル部に所属。
F-34 平日のラッシュアワーのみ運行する。
F-35 チャリティージーンズのデザインを行った。
F-36 広い範囲で大雨や暴風となった。
F-37 同センター副所長等を経て教授へ。
F-38 ラジオは大雨注意報をながしていた。
F-39 村の規模よりも大きい住居地区は無い。
F-40 ジャズとオーディオに関する著書も多数ある。
F-41 それぞれパッケージエアコンが設置された。
F-42 将軍家への不忠を詫びて切腹する。
F-43 名前はペンネームで本名は未公表。
F-44 ティーンエイジャーと友人がダンスを踊る。
F-45 専門分野は獣医疾病予防学。
F-46 微笑む赤い牛の顔がトレードマーク。
F-47 ややスケルツォ的な雰囲気を漂わせる。
F-48 一帯はエビやタチウオなどの好漁場。
F-49 信号が脈打つことをパルス状という。
F-50 プロシージャ呼び出しのオーバヘッドをなくす。
F-51 カジュアルウエアを販売するチェーンストア。
F-52 自家用宇宙艇を盗んで宇宙へ出る。
F-53 編集パッドとウインドウを生成する。
F-54 エアジェット織機のシェアは世界トップである。
F-55 撮影にシネマ風エフェクトを用いている。
F-56 徐々に他作品のパロディやアドリブが増加。
F-57 右派であるティーパーティーとは相性が悪い。
F-58 他の土地や公へ仕える自由もあった。
F-59 ジャファーの望み通り洞窟に入っていく。
F-60 冬服は紺のセーターと襟なしブレザー。
F-61 ツンツンした茶髪で顎鬚を生やしている。
F-62 投げ上げた布を空中で縫う荒業を持つ。
F-63 鉄砲に秀で、威風流砲術を開く。
F-64 カーブを織り交ぜて勝負する速球派投手。
F-65 モデルエディター、フォントエディターも存在する。
F-66 そのため接頭符号を語頭符号とも呼ぶ。
F-67 手順は、まず小麦粉やオートミールをふるう。
F-68 マンスズとペア競技での競技復帰を目指す。
F-69 目覚しい戦果を上げ復活をアピールした。
F-70 まっすぐな畦道にして、すじ植えをしてみた。
F-71 そこへ近いメーターほど通信量が増える。
F-72 どんどん客は奪われ売上は落ち続ける。
F-73 翼は鮮やかなブルーで胸は濃いイエロー。
F-74 複合部分発作や部分発作もみられる。
F-75 発音数やエフェクト数も向上している。
F-76 胸ポケットやネームホルダーにかけて使用する。
F-77 やわらかいキャンディーを砂糖でコーティングしている。

F-78 普及モデルは外付けアダプターで対応した。
F-79 長靴を履いてアイスホッケーを行うスポーツ。
F-80 ピアノコンチェルトをピアノの伴奏付きで演奏。
F-81 わざと違う意味で伝えたというエピソードもある。
F-82 それからハーブ茶やコーヒーを飲んで儀式は終わる。
F-83 斎宮は入内して梅壺に入り女御となった。
F-84 受像器で両者を合成しカラー画像を作る。
F-85 財産を投げ打ち病院へ復讐を企てる。
F-86 ユニバーシアードサッカー競技メンバーに選出。
F-87 プロデューサーからアーティストデビューを誘われていた。
F-88 後者はサディズム、マゾヒズムとの関係が強い。
F-89 ポテトチップスやポップコーンの製造を行う。
F-90 恵比寿校オープンと同時に本社も恵比寿へ移転。
F-91 シュート技がキーパー技を上回ればゴールとなる。
F-92 パンツ、パジャマ、スカーフ、毛布などに用いられる。
F-93 シチュエーションホラーの女王と評されるホラー作家。
F-94 防風樹、防潮樹、街路樹として利用される。
F-95 ポップスなど幅広いジャンルやシーンで活動中。
F-96 スパゲッティ屋やパスタハウスと呼ばれるレストランもある。
F-97 健気な戦争未亡人を演じ、注目を浴びる。
F-98 慈悲ある庇護者だったが強欲な略奪者でもあった。
F-99 そして店特製の大粒の甘酸っぱいチェリージャム。
F-100 ギザギザのついた鋭い刃をもち、幅広でぶ厚い。
F-101 インドア派と見せかけてメンバー随一のアウトドア派。
F-102 ペパーミントグリーンのタクシー車両を特徴とする。
F-103 近年の東洋っぽいアプローチがいい具合に光る。
F-104 ほうじ茶をベースに烏龍茶と紅茶をブレンドしたお茶。
F-105 コーヒーチェリーは茶同様、カフェインを含有している。
F-106 ネイティブアメリカンがインディアンジュエリーを作り始める。
F-107 魚粉製造時に水と分離して出た油を魚油という。
G-1 天国を上へ上へと登りつめる。
G-2 フィギュアやおもちゃ集めを趣味としている。
G-3 新世代のパーティーロックプロジェクト。
G-4 山小屋の蓬ヒュッテが存在する。
G-5 自身もベストスパイカー賞を受賞。
G-6 その後負傷兵をキールへ輸送した。
G-7 躊躇無く魔法を使うことが多い。
G-8 ハロウィン以後に初放送されている。
G-9 ザブーンを封印してから半年後。
G-10 サードエナジーが再び暴走した。
G-11 ああいう明るい歌はダメなんですよ。
G-12 およそ検閲を議論する余地はない。
G-13 培養液を注げば準備完了。
G-14 ラテン語文法家、聖書注釈者。
G-15 フェミニスト闘争運動の代表。
G-16 操縦はガンツァー側が担当する。
G-17 ラグビー応援団を自認している。
G-18 脳中枢へ情報が伝えられる。
G-19 少年時はディフェンダーとしてもプレー。
G-20 神社は海中へ没したといわれる。
G-21 重いギアでグイグイ漕

ぐことができる。
G-22 日常をウクレレに載せて唄うデュオ。
G-23 イアン曰く、お人よしばかりの家。
G-24 サッカージャーナリストとして著書多数。
G-25 リーグはオールスターモードに対応。
G-26 装着されるジュピター専用の装備。
G-27 バリケード封鎖に関与し除籍処分。
G-28 バーンウェルの現在の領域になった。
G-29 各種装備やエネルギーが補充される。
G-30 軍部とファシズムを痛烈に批判した。
G-31 また、ハードウェアエンベロープも持たない。
G-32 胆汁中や尿中に排出される。
G-33 ブリッジや脚を絡む動きを封じる。
G-34 芸能エージェントが彼女を見出した。
G-35 直通便で往復が可能となった。
G-36 スピードよりもヘビーなサウンドへ変化。
G-37 砂糖、水飴をよく混ぜ合わせこねる。
G-38 低いいぼ状隆起を備えてざらつく。
G-39 つまりポップミュージックのルーツですよね。
G-40 メフィストフェレスの回想的モノローグ。
G-41 地下空洞は普通土砂で埋め戻された。
G-42 違反した場合逮捕、訴追もありうる。
G-43 祝言も挙げさせず女中扱いにする。
G-44 彼女は岸辺で白いヤギに遭遇した。
G-45 基本的なディスクオペレーティングシステム。
G-46 キャンディチップやチョコクランチを付けて食べる。
G-47 夜空を飛ぶヘリコプター内でプロポーズ。
G-48 ノルディック複合、アルペンスキー選手。
G-49 彼は常にああいうプレーが披露できる。
G-50 運営部長から業務部長へ更迭。
G-51 これを取ればプレイヤーはパワーアップする。
G-52 フォトロケーションやスペシャルブースが登場。
G-53 閲覧数はウィルスのように増え続ける。
G-54 フェリーによるピストン輸送は幕を閉じた。
G-55 カメラは上空へとズームアウトしていく。
G-56 ダーウィニズムが崩壊しつつあると述べた。
G-57 犬と遊ぶこと、絵を描くこと、エクレア。
G-58 思わぬプライバシーが筒抜けになりやすい。
G-59 彼女候補はすべて女マネージャーである。
G-60 のちオーバーホールされ運用に復帰した。
G-61 エッジの乱入により王座防衛となる。
G-62 はっぴーはっぴーフライデーの中で発表。
G-63 レギュラーだと暴れる同じパターンを披露。
G-64 電磁波遮蔽パウダーを浴びて崩壊する。
G-65 例えば防壁での見張りの任務があった。
G-66 十分打ち合わせをしてゲームを開始する。
G-67 ヘアフィールド病院などキャリアを重ねていく。
G-68 ゾーンのアルファベット符号付けがなされている。
G-69 すなわちブルジョアジーとプロレタリアートである。
G-70 上部にエアジェット、下部にメインジェットがつく。
G-71 ウィッシュマンはヌーディストのジャンルに見切りをつけた。
G-72 大きなクエスチョンマークがボディ裏に描かれた。
G-73 オーディオバッファ、単一のリスナーで処理される。
G-74 なお、チェスはセミプロとも言うべき腕前だった。
G-75 マッハ主義に立脚しエネルギー論を展開。
G-76 茶エキス、及び粉末酒を製造する企業。
G-77 プレイヤー馬や順位も

各店舗同一である。
G-78 言いたいことを強調して大げさに言う手法。
G-79 ディフェンダーをこなすユーティリティープレーヤーだった。
G-80 圧倒的なパワーを秘める巨漢のスラッガー。
G-81 アニメプラザではコラボレーションカフェが催された。
G-82 この停留所で下車した方が当駅へは早い。
G-83 タワーの前方にはエアーエースの銅像がある。
G-84 野球、ラグビー、ホッケーなどスポーツに熱中。
G-85 縦横方向にうねうねと亀裂が走っている。
G-86 フォワードなどでもプレーできるユーティリティプレーヤー。
G-87 旅行用鞄やアタッシェケースは施錠ができる。
G-88 アフィン変換はちょうど直線を直線に写す。
G-89 アスファルト付着防止剤噴霧装置を具備する。
G-90 エネルギッシュなリズム、メロディとハーモニーであった。
G-91 アヘン輸出は政府の税収増に用いられた。
G-92 インディアン部族の同意を得ずに決議したのである。
G-93 プレーヤーオプションを破棄しフリーエージェントとなった。
G-94 肋骨を折られながらも前へ前へと攻め続けた。
G-95 救急車で病院行きになるのはあなたのほうだぞ。
G-96 主砲射撃方位盤及び通信装置は破損。
G-97 ファイアウォーキングを実行するためのソフトウェアツール。
G-98 ジュニアチームもハーフタイムなどにパフォーマンスしている。
G-99 木像よりは絵像、絵像よりは名号といふなり。
G-100 数字はダウンロード数、年月日は達成した日。
G-101 しばしばフェスティバルまたはキャンプファイアハープと呼ばれる。
G-102 パパはおじいさんの家での食事の準備で大忙し。
G-103 ハンドル操作はジャイロ、バーチャルパッドのどちらでも可能。
G-104 棒付きキャンディーにシャーベットを乗せて口へ運んでもよい。
G-105 形状からスプーンペン、たまペン、さじペンとも呼ばれる。
G-106 ポップアップ液晶パネルなどの機構は省略されている。
G-107 警官はウィーナーモービルを止めて運転手を拘束した。
H-1 ハードウェアを持つセキュリティ機能のこと。
H-2 賃金や手当の格差を是正せよ。
H-3 オートジャイロの開発ノウハウを得た。
H-4 最後は概ねハッピーエンドである。
H-5 運賃を払えば乗車可能である。
H-6 ビッグホーク、クイーンホークに化身。
H-7 バッグにレザーパッチワークを加えた。
H-8 ウォータークローバーという呼称がある。
H-9 ブルドッグはこっぴどくやられてしまう。
H-10 お百も無駄骨を折ったと怒り出す。
H-11 地方通貨として流通もしていた。
H-12 悲運の名将はユニフォームを脱いだ。
H-13 主要構造材として著名である。
H-14 ベビーシャンプーなどを発売している。
H-15 本尊は鬼門封じ不動明王。
H-16 凄まじく熱いんで評判なんだよ。
H-17 イデオロギーそれ自体は不可避である。
H-18 プライベートでは多趣味なアウトドア派。
H-19 バナジウムジルコニウム青を参照。
H-20 屈強なボディガード付きだったという。

H-21 一部の特急うずしおが停車する。
H-22 妖艶なウェットボイスなどが特徴。
H-23 オープン時の弁士オーディション出身。
H-24 父は銀行員、母は主婦だった。
H-25 リスナーの友情エピソード募集中。
H-26 レーサーを夢見る爽やかなお兄さん。
H-27 傲慢で自惚れ屋だという者もいる。
H-28 残ったファイバーを特異ファイバーと言う。
H-29 バーディーに自分をもう一度殺させた。
H-30 ポンプジャック、ビームポンプとも呼ばれる。
H-31 その後プロテニスツアーを回り始めた。
H-32 マイナーキーのスローなブルースナンバー。
H-33 そのままピアノのカデンツァに突入する。
H-34 全株主に書面を追加送付した。
H-35 現像所や事務所、ステージを全焼。
H-36 モータースポーツへのサポートを縮小。
H-37 大急ぎで絵を下流に移したという。
H-38 エックスにパワーアップパーツを授ける。
H-39 異性の好みはお姉さんっぽい女性。
H-40 今夜でディスクジョッキーを中退します。
H-41 プレッツェル部分はビスケットと呼ばれる。
H-42 いまだポピュラーなパーティードラッグである。
H-43 シール、模擬パスケース等が付属する。
H-44 水軍遺跡や言い伝えも多数ある。
H-45 輜重や衆人が多数略奪された。
H-46 河川の土砂運搬量は非常に多い。
H-47 オリンピック初のアーチェリー競技である。
H-48 スーパーエージェントチームを出動させた。
H-49 航空用語で操縦不能を意味する。
H-50 初演時はポレフがカデンツァを作曲した。
H-51 停車場所で手を上げ合図して乗車する。
H-52 トップチェッカーを受けチャンピオンに輝く。
H-53 自転車にまつわるエッセーや著作がある。
H-54 羽を動かさずに空中浮遊ができる。
H-55 黄熱病によりフィラデルフィアで死去した。
H-56 同地のサマースクールで英語を習得。
H-57 社長が大当りを狙うリーチアクション。
H-58 安全なショルダーチャージは合法である。
H-59 より大きい寿命のスケールで保持される。
H-60 音楽ディレクター、映像プロデューサー。
H-61 絵筆で自由に絵の具をのせて絵を描いた。
H-62 空軍アカデミー作戦学部を卒業。
H-63 そちらのページも是非、覗いてみてください。
H-64 ウォーターハザードと並ぶハザードに属する。
H-65 民生方面へ業務に情熱を注ぐ。
H-66 秘伝奥義を伝授され下山を許された。
H-67 サイドポンツーンは大幅に縮小された。
H-68 アンチドーピング運動に携わっていく。
H-69 店舗前でオープニングセレモニーを実施。
H-70 フォーム一家は新しいアパートに引っ越す。
H-71 アップウィンドで離陸後安定上昇する。
H-72 ユウダ科ミズベヘビ属に分類されるヘビ。
H-73 南方へ分岐する歩道沿いのアーケード。
H-74 タワー上部中央に風穴が開けられた。
H-75 レフェリーは元のペナルティーへとプレーを戻す。
H-76 日本語ではチュッチュッといった擬音に相当。
H-77 シリアスとギャグを織り交ぜた作風が特徴。
H-78 カウンターテーブルを備えたビュッフェを設置した。
H-79 ブーツはマーチより長く、ハッピーより短い。

H-80　運輸保安、事故調査等を所掌事務とする。

H-81　彼女のエッセンスは永遠に奪わてしまうわ。

H-82　その他収納や別売りのアクセサリーも豊富。

H-83　フィクションを交えつつ激動的に描いている。

H-84　ウェストベルトやボディハーネスにあぶみをぶらさげる。

H-85　コメディまで幅広いジャンルの作品へ出演。

H-86　この他、編著や共著、ノンフィクションなどがある。

H-87　あべを除いたメンバーにて前身バンドが始動。

H-88　スープ付きのピラフ、たこ焼きやパフェなどを扱う。

H-89　ファンファーレ風の移行句をへて再現部へ移る。

H-90　セーフのジェスチャーをフェアグラウンドへ向けて行う。

H-91　写実的な風景画を多数添えた通俗地誌。

H-92　手及び腕でボールを投げ送る守備行為を指す。

H-93　秩序と調和を意味する宇宙をイメージしている。

H-94　真空と亜鉛粉末を利用して改善した。

H-95　ボギー車はエアブレーキとエアホイッスルを使用した。

H-96　ハンバーガーを主力としたファストフードチェーンである。

H-97　少林寺拳法部、オーディオビジュアル部も強豪。

H-98　増補、新増補、新増補追補から成っている。

H-99　王者の左ジャブを浴び続け、右目が腫れ上がる。

H-100　神秘的なムードとエキゾチックなメロディーが特徴。

H-101　絶頂ラッシュ中は毎ゲーム上乗せが行われる。

H-102　ファッションフォト、ファッショングラフィックがベースになっている。

H-103　腹から巨大なサナダムシがうじゃうじゃと這い出している。

H-104　連邦ハイウェイ、州立ハイウェイ、地方道路である。

H-105　過去の作風を風刺した自虐的なエピソードさえある。

H-106　唾液や血清沈着物、壊死した骨を歯から除去した。

H-107　卵黄を温めて溶いた同量のエッグクリームを注ぐ。

参考文献

[1] P.B. Denes and E.N. Pinson, The Speech Chain: The Physics and Biology of Spoken Language, W.H. Freeman, 1993.

[2] K. Richmond, Z. Ling, and J. Yamagishi, "The use of articulatory movement data in speech synthesis applications: An overview — Application of articulatory movements using machine learning algorithms —," Acoustical Science and Technology, vol.36, no.6, pp.467–477, 2015. https://www.jstage.jst.go.jp/article/ast/36/6/36_E156001/_article

[3] 鏑木時彦, "調音運動に基づく音声の合成法に関する研究," PhD thesis, 九州芸術工科大学, 1998. https://ci.nii.ac.jp/naid/500000155445

[4] M.M. Sondhi and J. Schroeter, "A Hybrid Time-Frequency Domain Articulatory Speech Synthesizer," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.35, no.7, pp.955–967, 1987.

[5] B.H. Story, "Technique for " tuning " vocal tract area functions based on acoustic sensitivity functions," The Journal of the Acoustical Society of America, vol.119, no.2, pp.715–718, 2006.

[6] G. Fant and S. Pauli, "Spatial characteristics of vocal tract resonance modes," Proceedings of the Speech Communication Seminar, pp.121–132, Stockholm, 1975.

[7] B.S. Atal, J.J. Chang, M.V. Mathews, and J.W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," Journal of the Acoustical Society of America, vol.63, no.5, pp.1535–1555, 1978.

[8] J. Schroeter and M.M. Sondhi, "Speech coding based on physiological models of speech production," Advances in Speech Signal Processing, eds. by F. Sadaoki and M.M. Sondhi, pp.231–268, Marcel Dekker, New York, 1992.

[9] T. Kaburagi and M. Honda, "Determination of the vocal tract spectrum from

the articulatory movements based on the search of an articulatory-acoustic database," International Conference on Spoken Language Processing, p.0425, 1998. http://www.isca-speech.org/archive/icslp_1998/i98_0425.html

[10] J. Hogden, A. Lofqvist, V. Gracco, I. Zlokarnik, P. Rubin, and E. Saltzman, "Accurate recovery of articulator positions from acoustics: New conclusions based on human data," The Journal of the Acoustical Society of America, vol.100, no.3, pp.1819–1834, 1996. https://doi.org/10.1121/1.416001

[11] S. Kiritani, K. Itoh, and O. Fujimura, "Tongue pellet tracking by a computer controlled X-ray microbeam system," The Journal of the Acoustical Society of America, vol.57, pp.15–16, 1975. https://ci.nii.ac.jp/naid/20000958404/

[12] P.W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader, and B. Conrad, "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," Brain and Language, vol.31, no.1, pp.26–35, 1987.

[13] M. Stone, "A Guide to Analysing Tongue Motion from Ultrasound Images," PhD thesis, University of Maryland, 2004. https://pdfs.semanticscholar.org/6764/7f247993d3f2065de114d3917f0e7f927661.pdf

[14] 藤村靖，桐谷滋，柴田貞雄，"電気的パラトグラフによる調音運動の記録，" 音響学会講演論文集，1967.

[15] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," The Journal of the Acoustical Society of America, vol.115, no.4, pp.1771–1776, 2004.

[16] T. Kitamura, P. Mokhtari, and H. Takemoto, "Changes of vocal tract shape and area function by f0 shift," Proceedings of The International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA), pp.85–88, 2005.

[17] R.O. Tachibana, T. Kitamura, and M. Fujimoto, "Differences in articulatory movement between voiced and voiceless stop consonants," Acoustical Science and Technology, vol.33, no.6, pp.391–393, 2012.

[18] G. Fant, Acoustic Theory of Speech Production, De Gruyter Mouton, Berlin, Boston, 1971.

[19] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, vol.27, no.3, pp.187–207, 1999.

[20] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," IEICE Transactions on Information and Systems, vol.E99D, no.7, pp.1877–1884, 2016. https://www.jstage.jst.go.jp/article/transinf/E99.D/7/E99.D_2015EDP7457/_pdf/-char/en

[21] M. Morise, "D4C, A band-aperiodicity estimator for high-quality speech synthesis," Speech Communication, vol.84, pp.57–65, 2016. http://dx.doi.org/10.1016/j.specom.2016.09.001

[22] M. Morise, "Harvest: A high-performance fundamental frequency estimator from speech signals," Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp.2321–2325, 2017.

[23] M. Morise, "CheapTrick, a spectral envelope estimator for high-quality speech synthesis," Speech Communication, vol.67, pp.1–7, 2015. http://dx.doi.org/10.1016/j.specom.2014.09.003

[24] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," Proceedings of International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA), pp.59–64, 2001.

[25] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing, INTERSPEECH 2006 - ICSLP, vol.5, pp.2266–2269, 2006.

[26] K. Yu and S. Young, "Continuous F0 Modeling for HMM Based Statistical Parametric Speech Synthesis," IEEE Transactions on Audio, Speech and Language Processing, vol.19, no.5, pp.1071–1079, 2011.

[27] S.S. Stevens, J. Volkmann, and E.B. Newman, "A Scale for the Measurement of the Psychological Magnitude Pitch," Journal of the Acoustical Society of America, vol.8, no.3, pp.185–190, 1937.

[28] R.F. Ling, C.L. Lawson, and R.J. Hanson, Solving Least Squares Problems., vol.72, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1977.

[29] D.W. Griffin and J.S. Lim, "Signal Estimation from Modified Short-Time Fourier Transform," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.32, no.2, pp.236–243, apr 1984.

[30] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp.1118–1122, 2017.

[31] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol.3, pp.1315–1318, 2000.

[32] Y.J. Wu and R.H. Wang, "Minimum generation error training for HMM-based speech synthesis," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol.1, pp.89–92, 2006.

[33] Z. Wu and S. King, "Improving trajectory modelling for dnn-based speech synthesis by using stacked bottleneck features and minimum generation error training," IEEE/ACM Transactions on Audio Speech and Language Processing, vol.24, no.7, pp.1255–1265, 2016.

[34] A. Graves, "Generating Sequences With Recurrent Neural Networks," arXiv preprint, pp.1–43, 2013. `http://arxiv.org/abs/1308.0850`

[35] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Proceedings of International Conference on Learning Representations (ICLR), vol.94, pp.172–179, 2015.

[36] V. Nair and G.E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," Proceedings of International Conference on Machine Learning (ICML), vol.33, pp.384–387, 2010.

[37] A.L. Maas, A.Y. Hannun, and A.Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," Processing of ICML Workshop on Deep Learning for Audio, Speech and Language, vol.28, pp.1–6, 2013.

[38] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," Proceedings of International Conference on Machine Learning (ICML), vol.1, pp.448–456, 2015. `http://jmlr.org/proceedings/papers/v37/ioffe15.pdf`

[39] J.L. Ba, J.R. Kiros, and G.E. Hinton, "Layer normalization," arXiv preprint, pp.1–14, 2016. `http://arxiv.org/abs/1607.06450`

[40] T. Salimans and D.P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," Proceedings of Advances in Neural Information Processing Systems, pp.901–909, 2016.

[41] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol.15, no.1, pp.1929–1958, jan 2014. http://dl.acm.org/citation.cfm?id=2627435.2670313

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp.770–778, jun 2016. http://arxiv.org/abs/1512.03385

[43] M. Hermans and B. Schrauwen, "Training and Analysing Deep Recurrent Neural Networks," Proceedings of Advances in Neural Information Processing Systems, eds. by C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, pp.190–198, Curran Associates, Inc., 2013. http://papers.nips.cc/paper/5166-training-and-analysing-deep-recurrent-neural-networks.pdf

[44] M. Schuster and K.K. Paliwal, "Bidirectional recurrent neural networks," IEEE Transactions on Signal Processing, vol.45, no.11, pp.2673–2681, nov 1997.

[45] S. Hochreiter and J. Urgen Schmidhuber, "Long short-term memory," Neural Computation, vol.9, no.8, pp.1735–1780, 1997.

[46] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.1724–1734, 2014.

[47] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K.J. Lang, "Phoneme Recognition Using Time-Delay Neural Networks," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.37, no.3, pp.328–339, mar 1989.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All you Need," Proceedings of Advances in Neural Information Processing Systems, eds. by I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, vol.8, pp.5998–6008, Curran Associates, Inc., 2017. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[49] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional neural networks for human action recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.35, no.1, pp.221–231, 2013.

[50] Y. Sagisaka, K. Takeda, M. Abel, S. Katagiri, T. Umeda, and H. Kuwabara, "A

Large-Scale Japanese Speech Database," International Conference on Spoken Language Processing, pp.1089–1092, 1990.

[51] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis," arXiv preprint, pp.1–4, 2017. http://arxiv.org/abs/1711.00354

[52] K. Richmond, P. Hoole, and S. King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp.1505–1508, 2011.

[53] T. Kudo, "MeCab: Yet another part-of-speech and morphological analyzer," 2005. http://chasen.org/taku/software/mecab/

[54] T. Sato, "Neologism dictionary based on the language resources on the Web for Mecab," 2015. https://github.com/neologd/mecab-ipadic-neologd

[55] A. Lee, T. Kawahara, and K. Shikano, "Julius-An open source real-Time large vocabulary recognition engine," Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH), pp.1691–1694, 2001.

[56] M. Minoux, "Accelerated greedy algorithms for maximizing submodular set functions," Optimization Techniques, ed. by J. Stoer, pp.234–243, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[57] Y. Shinohara, "劣モジュラ最適化を用いた文部分集合選択によるコーパス構築法," 情報処理学会研究報告，第 2014 巻，pp.1–5，2014.

[58] J.S. Garofolo, "TIMIT: Acoustic-phonetic Continuous Speech Corpus," 1993. http://perso.limsi.fr/lamel/TIMIT_NISTIR4930.pdf

[59] A. Wrench, "The MOCHA-TIMIT articulatory database," 1999. http://www.cstr.ed.ac.uk/artic/mocha.html

[60] K. Richmond, "Estimating articulatory parameters from the acoustic speech signal," PhD thesis, pp.1–214, 2002.

[61] T. Kaburagi, K. Wakamiya, and M. Honda, "Three-dimensional electromagnetic articulography: A measurement principle," The Journal of the Acoustical Society of America, vol.118, pp.428–443, 2005.

[62] M. Tiede, C.Y. Espy-Wilson, D. Goldenberg, V. Mitra, H. Nam, and G. Sivaraman, "Quantifying kinematic aspects of reduction in a contrasting rate production task," The Journal of the Acoustical Society of America, vol.141, no.5, pp.3580–3580, 2017.

[63] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee,

K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein, D. Byrd, E. Bresch, P. Ghosh, A. Katsamanis, and M. Proctor, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," The Journal of the Acoustical Society of America, vol.136, no.3, pp.1307–1311, 2014. http://dx.doi.org/10.1121/1.4890284

[64] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," The Journal of the Acoustical Society of America, vol.120, no.5, pp.2421–2424, 2006. http://asa.scitation.org/doi/10.1121/1.2229005

[65] N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," IEEE Transactions on Multimedia, vol.17, no.5, pp.603–615, may 2015.

[66] K.R. Prajwal, R. Mukhopadhyay, V.P. Namboodiri, and C.V. Jawahar, "Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis," Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR), pp.13793–13802, 2020.

[67] T. Toda, A.W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," Speech Communication, vol.50, no.3, pp.215–227, 2008.

[68] C.T. Kello and D.C. Plaut, "A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters," The Journal of the Acoustical Society of America, vol.116, no.4, pp.2354–2364, 2004. http://asa.scitation.org/doi/10.1121/1.1715112

[69] F. Bocquelet, T. Hueber, L. Girin, P. Badin, B. Yvert, G. France, U.M.R. Cnrs, and I.N.P. Ujf, "Robust articulatory speech synthesis using deep neural networks for BCI applications," Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol.152, pp.2288–2292, 2014.

[70] F. Bocquelet, T. Hueber, L. Girin, C. Savariaux, and B. Yvert, "Real-time control of a dnn-based articulatory synthesizer for silent speech conversion: A pilot study," Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp.2405–2409, 2015.

[71] Z.C. Liu, Z.H. Ling, and L.R. Dai, "Articulatory-to-acoustic conversion with cascaded prediction of spectral and excitation features using neural networks," Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp.1502–1506, 2016.

[72] Z.C. Liu, Z.H. Ling, and L.R. Dai, "Articulatory-to-acoustic conversion using BLSTM-RNNs with augmented input representation," Speech Communication, vol.99, pp.161–172, 2018.

[73] S. Takamichi, K. Kobayashi, K. Tanaka, T. Toda, and S. Nakamura, "The NAIST text-to-speech system for the Blizzard Challenge 2015," Proceedings of Blizzard Challenge workshop, pp.351–356, 2015.

[74] T. Toda, A.W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," IEEE Transactions on Audio, Speech, and Language Processing, vol.15, no.8, pp.2222–2235, 2007.

[75] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an hmm-based speech production model," IEEE Transactions on Speech and Audio Processing, vol.12, no.2, pp.175–185, 2004.

[76] B. Uria, I. Murray, S. Renals, and K. Richmond, "Deep architectures for articulatory inversion," Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp.866–869, 2012.

[77] P. Liu, Q. Yu, Z. Wu, S. Kang, H. Meng, and L. Cai, "A deep recurrent approach for acoustic-to-articulatory inversion," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4450–4454, 2015.

[78] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," Proceeding of European Conference on Computer Vision (ECCV), pp.630–645, 2016.

[79] IEEE, "IEEE Recommended Practice for Speech Quality Measurements," IEEE Transactions on Audio and Electroacoustics, vol.17, no.3, pp.225–246, jun 1969.

[80] D. Talkin, "REAPER: Robust Epoch And Pitch EstimatoR [Online]," 2015. `https://github.com/google/REAPER`

[81] M. Parrot, J. Millet, and E. Dunbar, "Independent and automatic evaluation of acoustic-to-articulatory inversion models," arXiv preprint, pp.1–5, 2019. `http://arxiv.org/abs/1911.06573`

[82] A. Veit, M. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," Proceedings of Advances in Neural Information Processing Systems, pp.550–558, 2016.

[83] T. Hueber, E.L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," Speech Communication, vol.52, no.4, pp.288–300,

2010.

[84] T.L. Cornu and B. Milner, "Generating intelligible audio speech from visual speech," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.25, no.9, pp.1751–1761, sep 2017. https://ieeexplore.ieee.org/abstract/document/7949073

[85] A. Ephrat, T. Halperin, and S. Peleg, "Improved speech reconstruction from silent video," Proceedings of IEEE International Conference on Computer Vision Workshops (ICCVW), vol.2018-Janua, pp.455–462, 2018.

[86] Y.M. Assael, B. Shillingford, S. Whiteson, and N.D. Freitas, "LipNet: End-to-end sentence-level lipreading," arXiv preprint, pp.1–13, 2016. http://arxiv.org/abs/1611.01599

[87] A. Ephrat and S. Peleg, "Vid2Speech: speech reconstruction from silent video," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5095–5099, 2017. http://arxiv.org/abs/1701.00495

[88] H. Akbari, H. Arora, L. Cao, and N. Mesgarani, "Lip2AudSpec: Speech reconstruction from silent lip movements video," arXiv preprint, pp.1–9, 2017. http://arxiv.org/abs/1710.09798

[89] Y. Kumar, R. Jain, K.M. Salik, R.R. Shah, Y. Yin, and R. Zimmermann, "Lipper: Synthesizing thy speech using multi-view lipreading," 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, pp.2588–2595, 2019.

[90] W. Ping, K. Peng, A. Gibiansky, S.O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep Voice 3: Scaling text-to-speech with convolutional sequence learning," Proceedings of International Conference on Learning Representations (ICLR), pp.1–16, 2018. http://arxiv.org/abs/1710.07654 https://openreview.net/forum?id=HJtEm4p6Z

[91] J. Shen, R. Pang, R.J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R.A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4779–4783, 2018.

[92] T. Mihaylova and A.F.T. Martins, "Scheduled Sampling for Transformers," Pro-

ceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pp.351–356, Association for Computational Linguistics, Florence, Italy, jul 2019. https://www.aclweb.org/anthology/P19-2049

[93] Y. Jia, Y. Zhang, R.J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I.L. Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," Proceedings of Advances in Neural Information Processing Systems, pp.4480–4490, 2018.

[94] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-Adversarial Training of Neural Networks," Journal of Machine Learning Research, vol.17, no.59, pp.1–35, 2016. http://jmlr.org/papers/v17/15-239.html

[95] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S.Z. Li, "S3FD: Single Shot Scale-Invariant Face Detector," Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp.192–201, 2017.

[96] C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4214–4217, IEEE, 2010.

[97] J. Jensen and C.H. Taal, "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers," IEEE/ACM Transactions on Audio Speech and Language Processing, vol.24, no.11, pp.2009–2022, 2016.

[98] ITU-T, "Rec. P.861, Objective quality measurement of telephone-band (300-3400 Hz) speech codecs," 1996.

[99] ITU-T, "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," 2007.

[100] A.G Camacho, "SWIPE: A Sawtooth Waveform Inspired Pitch Estimator For Speech And Music," Journal of Chemical Information and Modeling, vol.53, no.9, pp.1689–1699, 2007.

[101] G. Sivaraman, V. Mitra, H. Nam, M. Tiede, and C. Espy-Wilson, "Unsupervised speaker adaptation for speaker independent acoustic to articulatory speech inversion," The Journal of the Acoustical Society of America, vol.146, no.1, pp.316–329, 2019.

[102] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: non-parallel many-to-many Voice Conversion Using Star Generative Adversarial Networks,"

Proceedings of IEEE Spoken Language Technology Workshop (SLT), pp.266–273, dec 2018.

[103] K. Kumar, R. Kumar, T. deBoissiere, L. Gestin, W.Z. Teoh, J. Sotelo, A. de-Brebisson, Y. Bengio, and A. Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," Proceedings of Advances in Neural Information Processing Systems, vol.32, pp.14910–14921, 2019.