計算資源が限られた音声合成システムに用いる
深層学習モデルの学習法に関する研究

2021 年 3 月

松永　悟行

1. 序論

音声は最も基本的，効率的な情報伝達手段のひとつである．音声には，個人性や情緒性の情報も含まれており，言語的情報を伝える以上の役割を担っている．現在では，様々な製品やサービスの音声インタフェースのひとつとして，人工的に音声を生成する技術である音声合成が利用されている．文字列から音声を合成する技術をテキスト音声合成または単に音声合成と呼ぶ．

音声合成技術はこれまで数多く研究されてきた．音声合成技術は録音編集方式と規則合成方式に大別できる．録音編集方式は最も簡単な合成方式であり，音声を，単語や文節ごとに収録し，記憶装置に蓄積して，これらの収録音声を適切に接続することで音声を合成する [1]．この方式では合成できる音声の内容は収録音声の語彙の組み合わせに限られるが，肉声のような自然性の高い音声を合成できる．また，この方式はシステムを簡単に構築できるため，合成する音声の内容が特定できる自動音声応答装置やカー・ナビゲーション・システムなどの製品やサービスでは多く利用されている．

規則合成方式は，音素の結合規則や、韻律の規則により生成した合成パラメータに基づいて音声を合成する．この方式は，音素レベルで合成を行うため，任意の文章に対して音声を合成できる．初期のころは，専門家が音声波形を分析し，専門知識によって合成規則を決めていた [2]．しかし，専門知識があったとしても，一貫性のある合理的な規則を策定することは難しい．また，専門知識による合成規則の表現力の限界や，音声波形を生成する際の励振信号の近似のため，合成音声は機械的な音声であった．

1990 年代ごろからは，規則合成方式のひとつであるコーパスベース方式の音声合成が研究されはじめた．この方式は，大規模な音声コーパスを構築し，専門知識による合成規則の代わりに統計的手法より合成パラメータを生成する [1]．コーパスベース方式は，波形接続方式と統計モデル方式に大別できる．波形接続方式は，言語解析部，音声特徴量予測部，波形生成部の 3 つのサブシステムで構成される．言語解析部は言語解析により文字列から言語特徴量を算出する．音声特徴量予測部は統計モデルで言語特徴量から音声特徴量を予測する．波形生成部は，言語特徴量と音声特徴量に従い音声コーパスから最適な音素波形を選択し，それらの音素波形を接続することで音声を合成する [3]．この方式は録音編集方式と同様に収録音声を直接利用するので，肉声に近い音声が合成できる．

統計モデル方式は、波形接続方式と同様に，言語解析部，音声特徴量予測部，波形生成部の 3 つのサブシステムで構成される．ただし，波形生成部は，ボコーダ（vocoder：voice coder）と呼ばれる音声分析変換合成システムによって，音声特徴量から音声波形を合成する [4] [5]．この方式は音声特徴量を編集することで波形接続方式よりも柔軟に合成音声を制御することができる．

音声合成に利用される統計モデルは，時系列のモデル化に適しており，効率的なモデルパラメータの学習アルゴリズムを必要とする．2000 年ごろには，音声認識で利用されていた

隠れマルコフモデル（HMM：Hidden Markov Model）が音声合成で利用された［6］．2006年には深層学習モデル（DNN：Deep Neural Network）の効率的な学習アルゴリズムが考案され［7], 2011 年ごろから DNN による音声認識が実用化されはじめ［8], 2013 年に DNN による音声合成が実現した［9]. HMM は各音素の音声特徴量を数個の状態でモデル化するのに対し，DNN は時間フレームごとに音声特徴量をモデル化する．このように，DNN は HMM よりも緻密に音声特徴量をモデル化できるため，DNN を利用することにより合成音声の品質を向上させることができる．

自然言語処理の分野でも DNN は利用されている．単語は記号であるため，単語を DNN で扱うには数値で表す必要がある．単語を効率よく数値ベクトルとして表現する方法として単語埋め込み法が考案された［10］．単語埋め込み法は，単語の共起関係から DNN を介して単語ごとに固有の数値ベクトルを与える．DNN 音声合成においても，言語解析部に単語埋め込み法を用いた学習法が提案されている［11] [12］．単語埋め込み法により，言語規則に基づいた言語特徴量を使わなくても，単語と音声特徴量の関係を直接学習できるようになった．ただし，日本語のように分かち書きされない言語においては，単語埋め込み法の前処理として文字列を単語ごとに分割する処理が必要となる．

ボコーダには励振信号をインパルスと白色雑音で近似する問題がある．この問題に対して DNN で音声波形を直接モデル化する方法が提案され，自然音声と遜色のない音声の合成が可能になった［13]. 文献［13]を発端として様々な音声波形のモデル化法が報告されている［14] [15] [16]．これらのような音声波形を生成する DNN は，ボコーダという言葉に因んでニューラル・ボコーダと呼ばれる．

音声特徴量予測部の DNN に加え，単語埋め込み法の DNN とニューラル・ボコーダを連結させることにより，単語列と音声波形の関係を直接モデル化できるようになった［11] [12］．このような学習方法を一貫学習または end-to-end 学習と呼ぶ．一貫学習により専門知識に基づいた合成規則や近似はほぼ必要なくなったが，学習外データに対する頑健性やユーザの制御性という点で課題が残っている［17].

音声合成技術の研究が進むにつれて，音声合成システムを利用した製品やサービスが登場してきた．音声合成システムが利用され始めたころは音声の明瞭度が主な要求であった．しかし，コンテンツ制作に利用されるようになると，単語の読み方の指定，話速や抑揚などの韻律の制御，合成音声の個人性も要求されるようになった．さらに，音声対話に利用されるようになると，感情，演技，発話意図の表現までも要求されるようになった．これらの要求に対応できるようするために，製品やサービスで利用される音声合成方式は，録音編集方式や波形接続方式から統計モデル方式へと変遷していった［17] [18]．合成方式の変遷に伴い合成処理は複雑化し，計算コストは増大した．一方で，音声合成システムの要件としては，応答が高速であること，あらゆる日本語文章の文字列に対して頑健性が高いこと，保守性が高いことが求められる．頑健性とは，学習外のデータに対しても破綻することなく合成パラメータを生成できることである．保守性とは，機能の変更や追加の容易さのことである．ま

た，音声合成システムは，組み込み機器などの演算装置や記憶装置の制約が大きいものから，画像処理や深層学習用の高性能な演算装置を搭載したものまで，様々な性能の計算機での動作も求められる．

音声合成システムへの要求の高度化に伴い，音声特徴量を柔軟に制御できる統計モデル方式の需要は高まっている．近年の統計モデル方式では DNN が利用されるため DNN の計算コストが問題となるが，演算装置の性能の向上によりその問題は解決されつつある．しかし，製品やサービスの要求，要件，仕様，制約により必ずしも高性能な演算装置を利用できるとは限らないため，DNN の計算コストの削減は必要である．

そこで本論文では，音声合成システムの保守性や制御性を考慮しつつ，計算資源が限られた計算機においても，頑健かつ高速に動作する音声合成システムを目指すために，音声合成システムの音声特徴量を予測する深層学習モデルの学習法を考案し，その有効性を評価した結果について述べる．

1 章は序論で，音声合成の背景と学位論文の範囲を述べる．2 章は DNN 音声合成システムの概要と音声コーパスについて述べる．3 章は，計算資源が限られた計算環境に適した音声特徴量予測部の構成について述べる．4 章は DNN で音声特徴量を頑健に予測するための言語特徴量の新たな正規化法について述べる．5 章は，3 章の音声特徴量予測部で利用する深層学習モデルが時系列を考慮して学習するための新たな損失関数による学習法について述べる．6 章は，3 章の音声特徴量予測部で利用する深層学習モデルが時系列を考慮して学習するための新たな敵対的ネットワークによる学習法について述べる．7 章は結論である．

2. DNN 音声合成システム

2.1. DNN 音声合成システムの構成

本論文では，音声合成システムの保守性や制御性を考慮して，一貫学習の音声合成システムの構成ではなく，図 2.1 に示す基本的な構成の DNN 音声合成システムを対象とする．音声合成システムの保守性は，システムの不具合の修正や，機能の変更や追加の容易さを表す．音声合成システムの制御性は，ユーザからの単語の読み方，アクセント型，話速，音高，抑揚，声質などの制御指令に対応できるかを表す．特に，一貫学習の音声合成システムでは，一部の変更がシステム全体に影響を与えるため，保守性は低い．

また，言語解析部については，ユーザからの単語の読み方やアクセント型の指定に対応できるように，単語埋め込み法ではなく，言語規則に基づいた言語解析法を用いる．単語埋め込み法では，音声コーパス内の単語と音声特徴量の関係を直接学習するため，単語の読み方やアクセント型は制御できない．

さらに，波形生成部については，計算コストを考慮して，ニューラル・ボコーダではなく，ボコーダを用いる．ニューラル・ボコーダは 1 サンプルごとに音声波形を予測するため，1 秒間に数万サンプルの予測を行わなければならない．このため，計算コストは非常に高く，未だにボコーダの計算コストの方が低い．

図 2.1 に示す DNN 音声合成システムでは，DNN は音声特徴量予測部でのみ利用される．一般的に，音声合成システムを利用する製品やサービスでは，ユーザが DNN を再学習する機能は提供されない．このため，DNN の学習時間は重要ではなく，合成時の音声特徴量の予測が高速，頑健，かつ高精度であればよい．そこで本論文では，簡素な構造の DNN を用いることや，音声特徴量予測部の処理を削減することで，合成処理の高速化を図り，簡素な構造の DNN でも音声特徴量を頑健で高精度に予測できる新しい学習法を提案する．

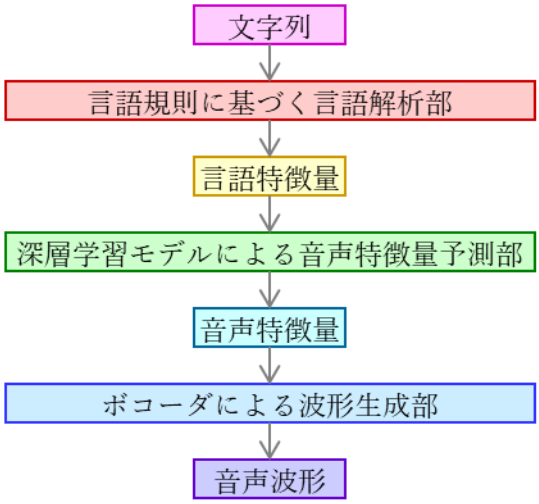

図 2.1　DNN 音声合成システムの構成

2.1.1.　言語規則に基づく言語解析部

言語規則に基づく言語解析法は，形態素解析で文字列を解析して，その結果から言語規則に基づいて言語特徴量を算出する．言語特徴量は呼気段落，アクセント句，モーラ，音素などの言語的な属性で構成される．形態素解析器は，文中の部分文字列と形態素辞書に登録されている形態素と照合することで，文章の文字列を構成する形態素に分割する [19] [20]．形態素は，表記文字，原形，読み，アクセント型，アクセント結合型，品詞，活用型の属性を持つ．アクセント句は，形態素のアクセント結合型からアクセント結合規則に基づき形態素を結合することによって得られる [21]．呼気段落は文中の息継ぎの箇所であるポーズからポーズまでの間に含まれるアクセント句群のことである．ポーズの位置は基本的に句読点の位置で決まる．モーラや音素は形態素の読みの属性から決まる．音素は破裂音や摩擦音などの調音様式や唇音や歯茎音などの調音部位の属性を持つ．調音方式や調音位置は音素ごとに固有に決まっている [1]．本論文では，形態素解析器は MeCab を使用し [19]，アクセント結合規則は [21]に従い，音素ラベルや音素の調音は HMM 音声合成システム（HTS：H Triple S＝HMM/DNN Speech Synthesis System）に従った [22]．

2.1.2.　ボコーダによる波形生成部

ボコーダは音声波形から音声特徴量を抽出する分析部と，音声特徴量から音声波形を生成する合成部で構成される．ボコーダの分析部により抽出される音声特徴量は，音高を表す基本周波数と，声色を表すスペクトル包絡と，有声音と無声音の混合比を表す非周期性指標である．ボコーダの合成部は以下のように音声波形を生成する．まず，有声音の励振信号を表す基本周波数に基づいた周期的なインパルスと，無声音の励振信号を表す白色雑音を生成する．次に，スペクトル包絡と非周期性指標から有声音の励振信号用のスペクトル包絡と無声音の励振信号用のスペクトル包絡をそれぞれ算出する．そして，これらのスペクトル包

絡を有声音および無声音それぞれの励振信号に畳み込み，これらの信号を加算することで音声波形を生成する．音声合成システムの波形生成部には，ボコーダの合成部だけを使用する．本論文で用いたボコーダは WORLD（D4C edition）である [4] [23].

2.2. 音声特徴量予測部の構成

図 2.1 に示す DNN 音声合成システムにおいて，音声を合成するために必要な音声特徴量は，継続長，基本周波数，スペクトル包絡，非周期性指標である．継続長は音素レベルの音声特徴量であり，基本周波数，スペクトル包絡，非周期性指標は時間フレームレベルの音声特徴量である．時間フレームレベルの音声特徴量の予測には時間フレームの情報が必要であり，時間フレーム情報を得るためには継続長が必要である．そのため，まず，音素レベルの言語特徴量から継続長を予測し，次に，継続長から求めた時間フレーム情報が付加された時間フレームレベルの言語特徴量から基本周波数，スペクトル包絡，非周期性指標を予測する．

図 2.2 に示す音声特徴量予測部は，音素レベルの音声特徴量である継続長の予測部と，基本周波数，スペクトル包絡，非周期性指標をまとめた時間フレームレベルの音声特徴量の予測部で構成される．図 2.3 に示す音声特徴量予測部は，音声特徴量ごとに個別の予測部を持つ．図 2.2 の構成の利点は，ひとつの DNN で基本周波数，スペクトル包絡，非周期性指標を予測できるため，計算量が少なくて済むことである．図 2.3 の構成の利点は，音声特徴量ごとに，モデルの更新や，不具合の修正ができることである．本論文では，図 2.3 に示す音声特徴量ごとの予測部を持つ音声特徴量予測部を用いる．

図 2.4 に示すように音声特徴量予測部の構成は学習時と予測時で異なる．学習時の音声特徴量予測部の構成は，言語特徴量の正規化部，DNN，音声特徴量の正規化部，モデルパラメータ更新部で構成される．予測時の音声特徴量予測部の構成は，言語特徴量の正規化部，学習済みの DNN，音声特徴量の逆正規化部，後処理部で構成される．

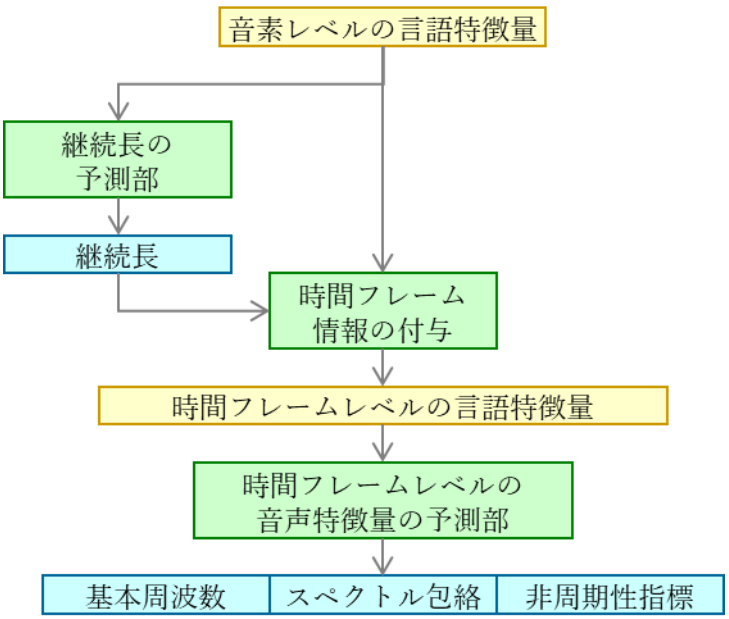

図 2.2　音声特徴量を一括で予測する音声特徴量予測部の構成

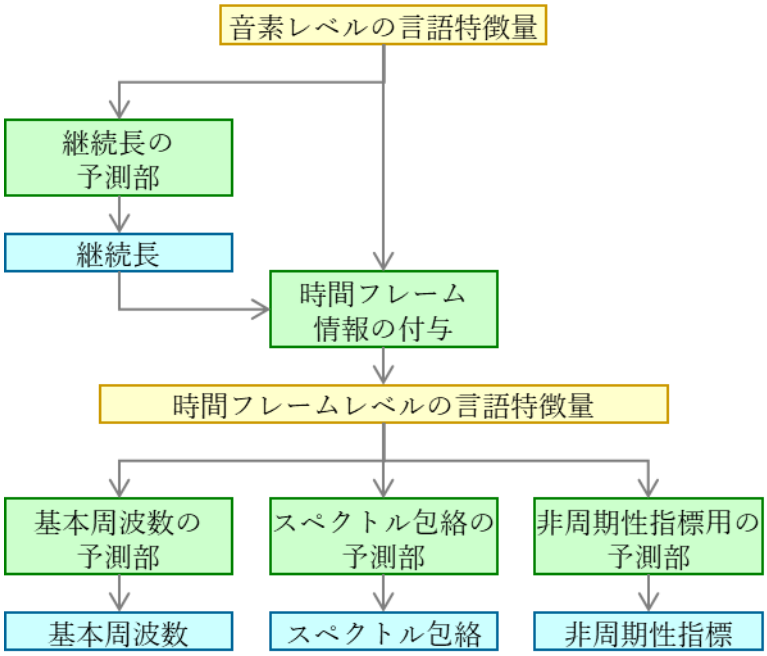

図 2.3　音声特徴量ごとの予測部を持つ音声特徴量予測部の構成

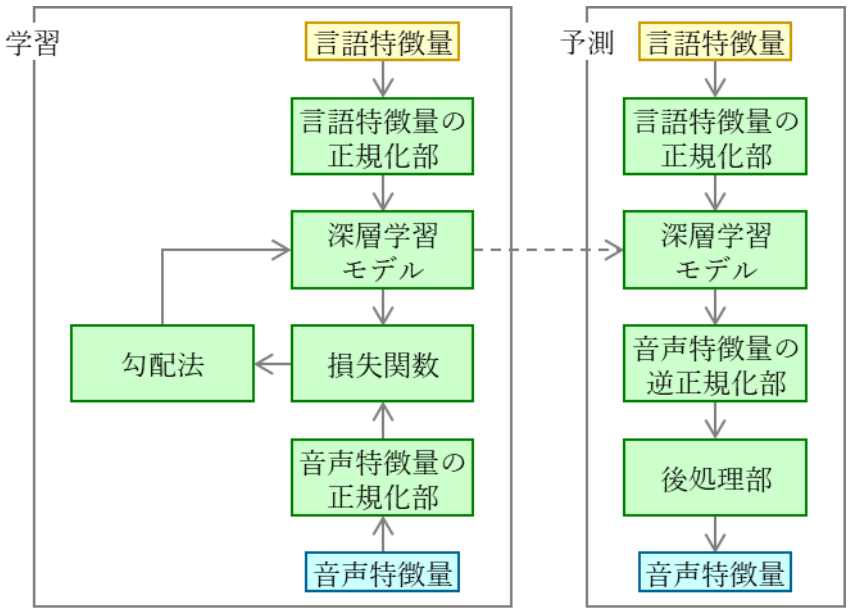

図 2.4　各音声特徴量の予測部の構成

2.2.1.　正規化部と逆正規化部

DNN は行列の積和で表現されるため，入力データの要素のうち大きな値をとる要素が支配的になる．また，DNN は出力データと教師データ間の誤差に基づいて学習されるため，教師データの要素のうち大きな値をとる要素の誤差が支配的になる．これらの問題を防ぐため，データの正規化が必要となる．一般的な正規化法には，Min-Max 正規化法と Mean-Var 正規化法がある [9]．Min-Max 正規化法は，最小値が 0，最大値が 1 となるようにデータのスケールを変化させる．Mean-Var 正規化法は，平均値が 0，標準偏差が 1 である標準正規分布に従うようにデータのスケールを変化させる．また，教師データを正規化して DNN を学習すると，DNN の出力データのスケールは，正規化後の教師データのスケールと同じになる．出力データのスケールをもとに戻すには，教師データに適用した正規化法の処理と逆の処理を出力データに適用する必要がある．

言語特徴量については、呼気段落の総数やアクセント句の総数など文章の構成によってこれらの属性値が大きく変化するため，正規化が必要である．一般的に，言語特徴量の正規化には Min-Max 正規化法が利用される．

音声特徴量については，モデル化の戦略や後処理に応じて、正規化の必要性を検討する．例えば，音声特徴量の次元間の関係性を保つ場合，正規化は適用されない．一方で，複数話者のデータから平均声のモデルを学習する場合，話者間の音声特徴量の差をなくすために，話者ごとに音声特徴量を正規化する．また，ユーザからの制御指令に応じて音声特徴量の平均値や標準偏差を変更する場合，あらかじめ音声特徴量を正規化しておくと処理の都合が良い．一般的に，音声特徴量の正規化には Mean-Var 正規化法が利用される．

2.2.2. 深層学習モデル

DNN は，複数の人工神経の結合で構成される．人工神経は，生体神経をモデル化したものである（図 2.5）．生体神経細胞においては，樹状突起のシナプスが神経伝達物質を受け取ると細胞体の電位が上昇し，その電位が一定の閾値電位を超えると細胞体から軸索へ活動電位が伝わる．人工神経においては，入力の加重和に閾値を加え，活性化関数を適用すると，出力が得られる．荷重は樹状突起のシナプスの結合強度を模擬し，閾値は細胞体の閾値電位を模擬し，活性化関数は細胞体の活動電位を模擬する．

複数の人工神経を並列に配置したものを層と呼び，DNN は層を積み重ねた構造を持っている（図 2.6）．図中の白丸はひとつの人工神経を表す．入力データを受ける層を入力層と呼び，出力データを渡す層を出力層と呼び，入力層と出力層の間にある層を隠れ層と呼ぶ．人工神経の荷重や閾値を DNN のモデルパラメータと呼び，ユニット数，層の数，活性化関数を DNN のハイパーパラメータと呼ぶ．ユニット数は 1 層あたりの人工神経の数である．DNN のモデルパラメータは，基本的に乱数で初期値が与えられ，その値は学習を繰り返すことによって更新される．DNN のハイパーパラメータは固定であり，学習の前に決定しておく．

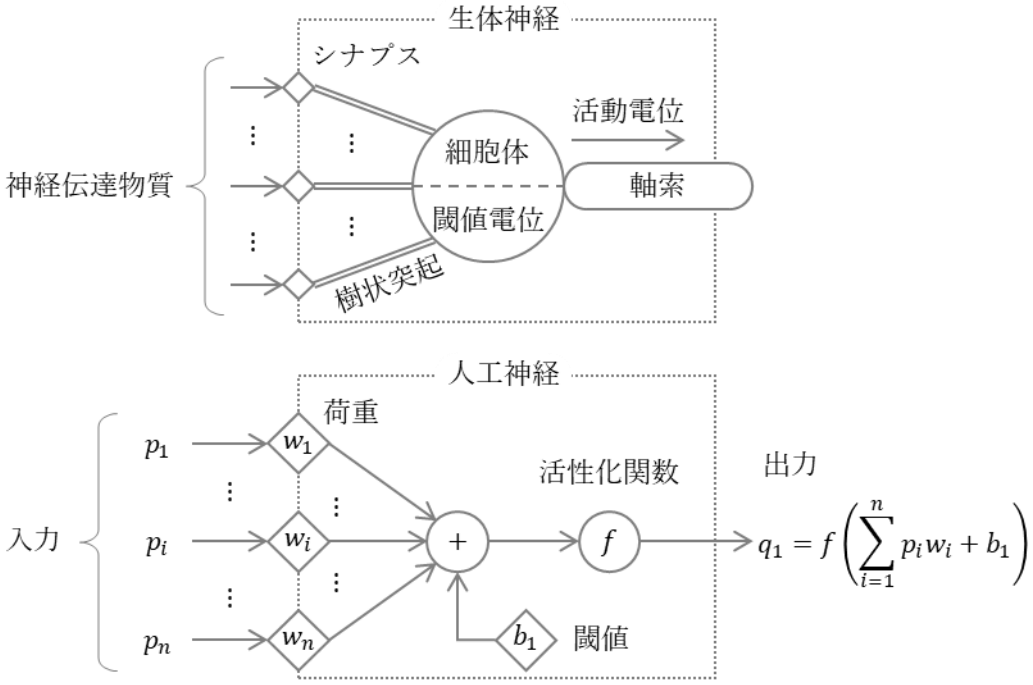

図 2.5　生体神経と人工神経の構造

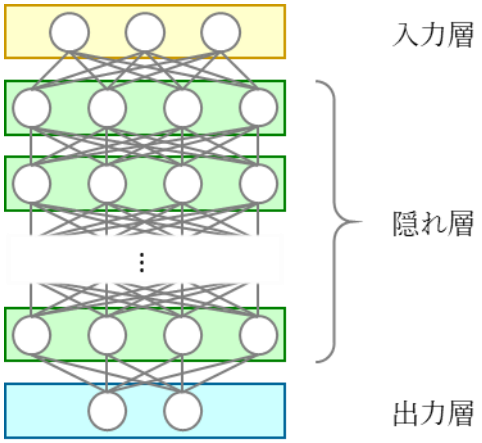

図 2.6　深層学習モデルの構造

2.2.2.1. 順伝搬型の深層学習モデル：全結合層

順伝搬型の深層学習モデル（FFNN：Feed-Forward Neural Network）は基本的な深層学習モデルである．FFNN は全結合層のみで構成される（図 2.7）．$\boldsymbol{p}$はN_i次元の入力ベクトル，$\boldsymbol{q}$はN_o次元の出力ベクトル，$\boldsymbol{W}$は$N_i \times N_o$の荷重行列，$\boldsymbol{b}$はN_o次元の閾値ベクトル，fは活性化関数である．荷重行列と閾値ベクトルは全結合層のモデルパラメータである．

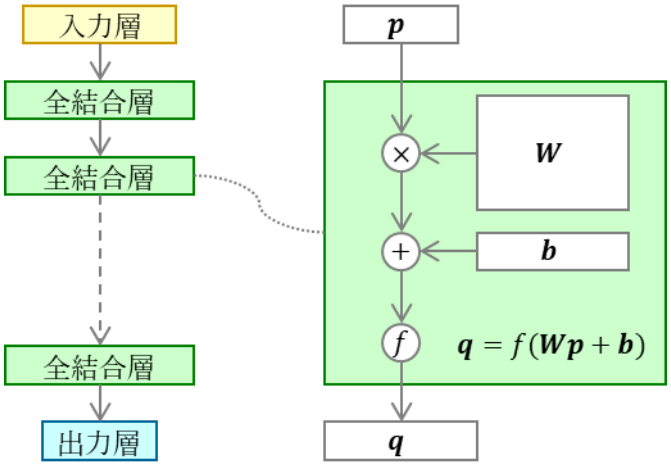

図 2.7　順伝搬型の深層学習モデルの構造

2.2.2.2. 再帰型の深層学習モデル：再帰層

再帰型の深層学習モデル（RNN：Recurrent Neural Network）は時系列をモデル化するのに適した深層学習モデルである．RNN は一つ以上の再帰構造を持った層により構成される．再帰層は最も基本的な再帰構造を持つ，RNN を構成する層のひとつである（図 2.8）．$\boldsymbol{p}_t$は時間フレームtのN_i次元の入力ベクトル，$\boldsymbol{q}_{t-1}$は時間フレーム$t-1$のN_o次元の出力ベクトル，$\boldsymbol{q}_t$は時間フレームtのN_o次元の出力ベクトル，$\boldsymbol{W}$は$N_i \times N_o$の荷重行列，$\boldsymbol{R}$は$N_o \times N_o$の

再帰荷重行列，$\boldsymbol{b}$はN_o次元の閾値ベクトル，×は内積，＋は要素ごとの和，fは活性化関数である．荷重行列，再帰荷重行列，閾値ベクトルは再帰層のモデルパラメータである．

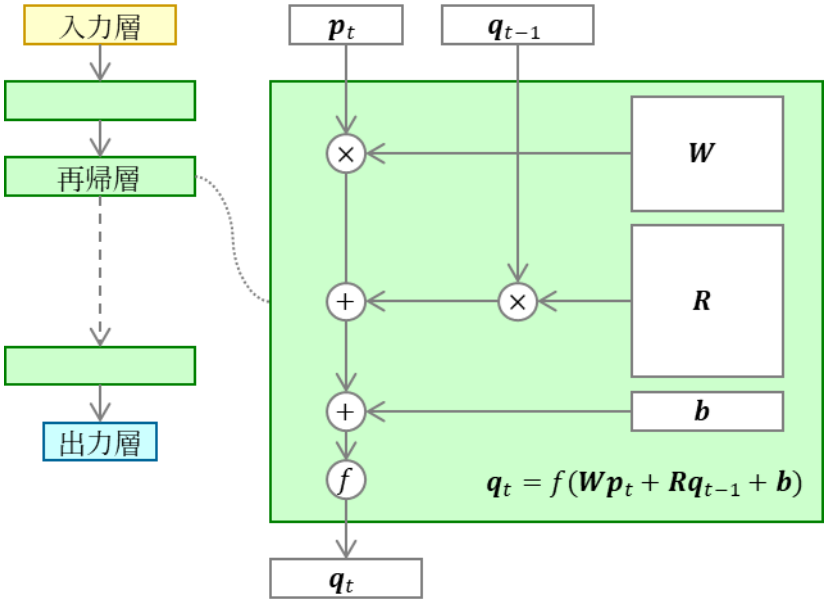

図 2.8　再帰層の構造

2.2.2.3. 再帰型の深層学習モデル：長短期記憶層

高性能な再帰層として長短期記憶層（LSTM 層：Long Short -term Memory 層）がある．LSTM 層は，4 つの再帰層と記憶セルで構成される（図 2.9）．$\boldsymbol{p}_t$は時間フレームtのN_i次元の入力ベクトル，$\boldsymbol{q}_{t-1}$は時間フレーム$t-1$のN_o次元の出力ベクトル，$\boldsymbol{q}_t$は時間フレームtのN_o次元の出力ベクトル，$\boldsymbol{q}_t^{(\mathrm{r})}$は時間フレーム$t$の$N_o$次元の一時ベクトル，$\boldsymbol{q}_t^{(\mathrm{i})}$は時間フレーム$t$の$N_o$次元の入力ゲートの出力ベクトル，$\boldsymbol{q}_t^{(\mathrm{f})}$は時間フレーム$t$の$N_o$次元の忘却ゲートの出力ベクトル，$\boldsymbol{q}_t^{(\mathrm{o})}$は時間フレーム$t$の$N_o$次元の出力ゲートの出力ベクトル，$\boldsymbol{r}_{t-1}$は記憶セルが保持する時間フレーム$t-1$の$N_o$次元のベクトル，∘は要素ごとの積，＋は要素ごとの和，fは活性化関数である．各ゲートの活性化関数はシグモイド関数であり，これらのゲートの出力ベクトルは 0 から 1 までの値をとる．これにより，各ゲートはそれぞれ入力，記憶セル，出力の情報の取捨選択や流量を調整する．4 つの再帰層の荷重行列，再帰荷重行列，閾値ベクトルは LSTM 層のモデルパラメータである．

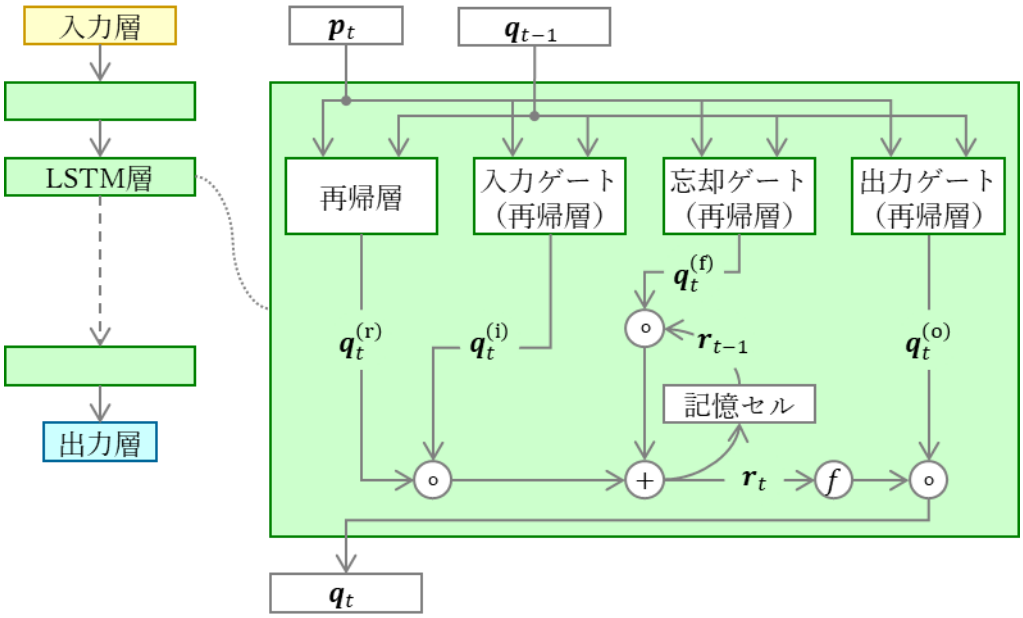

図 2.9　長短期記憶層の構造

2.2.3.　損失関数と勾配法

DNN のモデルパラメータは，教師データと予測データ間の誤差に基づいて更新される．損失関数は教師データと予測データの誤差を算出するものであり，勾配法は損失関数が算出した誤差に基づいて DNN のモデルパラメータを更新するものである．

DNN の学習においては，学習データセットを複数のバッチに分割して，バッチごとにモデルパラメータを更新する．このような学習の仕方をミニバッチ学習と呼ぶ [24]．ひとつのバッチに含まれるデータの数をバッチサイズと呼ぶ．言語特徴量と音声特徴量の関係を学習する音声特徴量予測部の DNN では，バッチサイズは時間フレーム数で表現され，固定の時間フレーム数や，1 文ごとの言語特徴量と音声特徴量の時間フレーム数に設定される．すべてのバッチを学習するサイクルをエポックと呼び，このサイクルの繰り返し回数をエポック数と呼ぶ．エポック数は予測データと教師データ間の誤差が集束するように設定される．

2.2.4.　後処理部

統計モデル方式の音声合成では，統計モデルの性能を補うために，統計モデルによって予測された音声特徴量に後処理が適用されることがある．HMM 音声合成のときから利用されてきた基本的な後処理として，尤度最大化に基づくパラメータ生成法と [6]，ケプストラム強調がある [25]．

2.2.4.1.　尤度最大化に基づくパラメータ生成法

尤度最大化に基づくパラメータ生成法（MLPG：Maximum Likelihood Parameter

Generation）は HMM 音声合成における区間定常の問題を解決するために考案されたものである．DNN 音声合成においては，FFNN が時間フレームごとに独立して音声特徴量をモデル化する問題を解決するために MLPG が利用される．MLPG は音声特徴量の動的特徴量の持つ正規分布が与えられたとき，音声特徴量の動的特徴量の尤度が最大になるような音声特徴量を求める．音声特徴量の動的特徴量の対数尤度を次式で定義する．

$$\log P\left(\boldsymbol{W}\boldsymbol{\psi} \mid \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{U})\right) \tag{2.1}$$

ここで，$\boldsymbol{\psi}$は MLPG が生成する音声特徴量ベクトル系列，$\boldsymbol{W}$は動的特徴量を求めるための係数行列，$\boldsymbol{\mu}$は音声特徴量の動的特徴量の平均ベクトル系列，$\boldsymbol{U}$は音声特徴量の動的特徴量の共分散行列，$\mathcal{N}$は$\boldsymbol{\mu}$と$\boldsymbol{U}$を持つ正規分布，$\log P$は$\mathcal{N}$が与えられたときの$\boldsymbol{W}\boldsymbol{\psi}$の対数尤度である．$\log P$を最大にする$\boldsymbol{\psi}$は，$P$の$\boldsymbol{\psi}$についての導関数が 0 のときの式から導き出される．

$$\begin{aligned}
&\underset{\boldsymbol{\psi}}{\operatorname{argmax}} \log P\left(\boldsymbol{W}\boldsymbol{\psi} \mid \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{U})\right) \\
&\frac{\partial \log P\left(\boldsymbol{W}\boldsymbol{\psi} \mid \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{U})\right)}{\partial \boldsymbol{\psi}} = \boldsymbol{0}
\end{aligned} \tag{2.2}$$

$$\begin{aligned}
\boldsymbol{\psi} &= \mathrm{MLPG}(\boldsymbol{\mu}, \boldsymbol{U}^{-1}, \boldsymbol{W}) \\
&= (\boldsymbol{W}^{\top}\boldsymbol{U}^{-1}\boldsymbol{W})^{-1}(\boldsymbol{W}^{\top}\boldsymbol{U}^{-1}\boldsymbol{\mu})
\end{aligned} \tag{2.3}$$

$$\begin{aligned}
\boldsymbol{\mu} &= [\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_t, \cdots, \boldsymbol{\mu}_T]^{\top} \\
\boldsymbol{\mu}_t &= \left[\boldsymbol{\mu}_t^{(0)}, \boldsymbol{\mu}_t^{(1)}, \boldsymbol{\mu}_t^{(2)}\right] \\
\boldsymbol{\mu}_t^{(n)} &= \left[\mu_t^{(n,1)}, \cdots, \mu_t^{(n,d)}, \cdots, \mu_t^{(n,D)}\right] \quad (n = 0, 1, 2)
\end{aligned} \tag{2.4}$$

$$\begin{aligned}
\boldsymbol{U}^{-1} &= \mathrm{diag}\,[\boldsymbol{U}_1^{-1}, \cdots, \boldsymbol{U}_t^{-1}, \cdots, \boldsymbol{U}_T^{-1}] \\
\boldsymbol{U}_t &= \begin{bmatrix} \boldsymbol{U}_t^{(0,0)} & \boldsymbol{U}_t^{(0,1)} & \boldsymbol{U}_t^{(0,2)} \\ \boldsymbol{U}_t^{(1,0)} & \boldsymbol{U}_t^{(1,1)} & \boldsymbol{U}_t^{(1,2)} \\ \boldsymbol{U}_t^{(2,0)} & \boldsymbol{U}_t^{(2,1)} & \boldsymbol{U}_t^{(2,2)} \end{bmatrix}_{(3D \times 3D)}
\end{aligned} \tag{2.5}$$

$$\begin{aligned}
\boldsymbol{W} &= [\boldsymbol{W}_1, \cdots, \boldsymbol{W}_t, \cdots, \boldsymbol{W}_T]^{\top} \\
\boldsymbol{W}_t &= \left[\boldsymbol{W}_t^{(0)}, \boldsymbol{W}_t^{(1)}, \boldsymbol{W}_t^{(2)}\right] \\
\boldsymbol{W}_t^{(n)} &= \left[\underset{1\mathrm{st}}{\boldsymbol{0}_{(D\times D)}}, \cdots, w_{\tau}^{(n)} \underset{(t+\tau)\,\mathrm{th}}{\boldsymbol{I}_{(D\times D)}}, \cdots, \underset{T\,\mathrm{th}}{\boldsymbol{0}_{(D\times D)}}\right]^{\top} \quad \begin{cases} (n = 0, 1, 2) \\ (\tau = -1, 0, 1) \end{cases}
\end{aligned}$$

$$\begin{aligned}
w_{\tau}^{(0)} &= \begin{cases} 0 & (\tau = -1) \\ 1 & (\tau = 0) \\ 0 & (\tau = 1) \end{cases} \\
w_{\tau}^{(1)} &= \begin{cases} -0.5 & (\tau = -1) \\ 0 & (\tau = 0) \\ 0.5 & (\tau = 1) \end{cases} \\
w_{\tau}^{(2)} &= \begin{cases} 1 & (\tau = -1) \\ -2 & (\tau = 0) \\ 1 & (\tau = 1) \end{cases}
\end{aligned} \tag{2.6}$$

ここで，$\mu_t^{(n,d)}$は時間フレームtにおける次元dの音声特徴量のn次の動的特徴量，$\boldsymbol{U}_t^{(n_1,n_2)}$は時間フレーム$t$における音声特徴量の$n_1$次の動的特徴量と$n_2$次の動的特徴量の$D \times D$の共分散行列，$\boldsymbol{0}_{(D\times D)}$は$D \times D$の零行列，$\boldsymbol{I}_{(D\times D)}$は$D \times D$の単位行列，$w_{\tau}^{(n)}$は相対時間フレーム$\tau$の$n$

次の動的特徴量を求める係数である．ただし，計算量を削減するため，$\boldsymbol{U}_t$は対角行列として，次元ごとに独立して式（2.3）を計算することが多い［6］．本論文でも，$\boldsymbol{U}_t$は対角行列とする．

HMM 音声合成においては，$\boldsymbol{\mu}_t$や$\boldsymbol{U}_t$は HMM の各状態のモデルパラメータである．DNN 音声合成においては，$\boldsymbol{\mu}$は DNN で予測され，$\boldsymbol{U}_t$は時間フレームに依らず一定とし，次式で計算される学習データセット全体の音声特徴量の分散とする．

$$
\begin{aligned}
&\boldsymbol{U}^{-1} = \mathrm{diag}\left[\boldsymbol{u}^{-1}, \cdots, \boldsymbol{u}^{-1}\right] \\
&\boldsymbol{u} = \mathrm{diag}\left[\boldsymbol{u}^{(0)}, \boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}\right] \\
&\boldsymbol{u}^{(n)} = \left[u^{(n,1)}, u^{(n,d)}, u^{(n,D)}\right] \quad (n = 0,\ 1,\ 2) \\
&u^{(n,d)} = \underset{t,\mathbb{U}}{\mathrm{var}}\left(\sum_{\tau=-1}^{1} y_t^{(d)} w_\tau^{(n)}\right) \quad (\tau = -1,\ 0,\ 1) \\
&\boldsymbol{y} \in \mathbb{U} \\
&\boldsymbol{y} = [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_t^\top, \cdots, \boldsymbol{y}_T^\top]^\top \\
&\boldsymbol{y}_t = \left[y_t^{(1)}, \cdots, y_t^{(d)}, \cdots, y_t^{(D)}\right]
\end{aligned}
\tag{2.7}
$$

ここで，varは分散を算出する関数，$\mathbb{U}$は学習データセット，$\boldsymbol{y}$は$\mathbb{U}$に含まれる音声特徴量ベクトル系列，$\boldsymbol{y}_t$は時間フレームtにおける音声特徴量ベクトル，$y_t^{(d)}$は時間フレームtにおける次元dの音声特徴量，$u^{(n,d)}$は$\mathbb{U}$全体の$y_t^{(d)}$から算出した次元dの音声特徴量のn次の動的特徴量の分散である．

2.2.4.2. ケプストラム強調

統計モデルは音声特徴量を統計的にモデル化する．モデル化の際に平均などの統計処理が音声特徴量に施されるため，統計モデルで予測される音声特徴量は平滑化されている．また，MLPG は隣接する時間フレーム間の音声特徴量が連続的に変化するように音声特徴量の動的特徴量に基づいて音声特徴量を平滑化する．音声特徴量の中でもスペクトル包絡は合成音声の音色を制御する特徴量である．スペクトル包絡の適度な平滑化は，隣接する時間フレーム間のスペクトル包絡の不連続を緩和し，合成音声の音色が急に変化することを防ぐ．一方で，スペクトル包絡の過剰な平滑化は，スペクトル包絡の起伏を緩やかにして平坦なスペクトル包絡に近づける．起伏が緩やかなスペクトル包絡を励振信号に畳み込んでも，励振信号のスペクトル形状は十分に変化しないため，励振信号のブザーのような音色が合成音声に現れるようになる．ケプストラム強調はこの問題を解決する．

2.3.3 で述べるが，スペクトル包絡の表現法のひとつにメルケプストラムがある．ケプストラム強調はメルケプストラムの係数を定数倍することで，スペクトル包絡のフォルマントを強調し，起伏のあるスペクトル包絡にする．ケプストラム強調は次式で定義される．

$$
\begin{aligned}
\boldsymbol{c}_t &= \left[c_t^{(1)}, \cdots, c_t^{(d)}, \cdots, c_t^{(D)}\right] \\
\tilde{\boldsymbol{c}}_t &= \left[\tilde{c}_t^{(1)}, \cdots, \tilde{c}_t^{(d)}, \cdots, \tilde{c}_t^{(D)}\right] \\
\tilde{c}_t^{(d)} &= \begin{cases} c_t^{(1)} - (\beta - 1)\sum_{d_1=3}^{D} (-\alpha)^{(d_1-1)} c_t^{(d_1)} & (d = 1) \\ c_t^{(2)} & (d = 2) \\ \beta c_t^{(d)} & (3 \le d \le D) \end{cases} \\
\hat{\boldsymbol{c}}_t &= \left[\tilde{c}_t^{(1)} + r_t, \tilde{c}_t^{(2)}, \tilde{c}_t^{(3)}, \cdots, \tilde{c}_t^{(D)}\right] \\
r_t &= \frac{1}{2}\log\frac{\rho_t}{\tilde{\rho}_t}
\end{aligned}
\tag{2.8}
$$

ここで，$\boldsymbol{c}_t$は時間フレームtにおけるメルケプストラム，$\tilde{\boldsymbol{c}}_t$は時間フレームtにおける強調メルケプストラム，$\hat{\boldsymbol{c}}_t$は時間フレームtにおける補正した強調メルケプストラム，ρ_tは$\boldsymbol{c}_t$から得られる最小位相インパルス応答の振幅の 2 乗の和で与えられるエネルギー，$\tilde{\rho}_t$は$\boldsymbol{b}_t$から得られる最小位相インパルス応答の振幅の 2 乗の和で与えられるエネルギー，r_tはρ_tと$\tilde{\rho}_t$から算出される補正係数，αはメルケプストラムの周波数伸長パラメータ，βは強調係数である．ただし，$c_t^{(1)}$はケプストラムの 0 次項を表す．βの値は経験則に基づいて決定され，文献 [25] では 1.5，文献 [22]では 1.4 となっている．本論文では，βを 1.4 とした．

2.3. 音声コーパス

本論文では，数人の男性話者と数人の女性話者の音声コーパスの中から，1 名の女性話者の音声コーパスを使用した．この女性話者は日本語を母語とするプロのアナウンサーである．この音声コーパスは朗読音声，感情音声など様々なスタイルの音声を含むため，本論文では，朗読調の音声のみを使用した．この音声コーパスを朗読調の音声に限ると，文章数は 3000 で，音声の合計時間は約 5 時間となる．900 文（約 2 時間）は一文章あたりの平均モーラ数が約 56 の長文セットであり，残り 2100 文（約 3 時間）は ATR503 文を含む音素バランス文セットである [26].

この朗読調の音声コーパスを以下のように分けて，以降の章の実験で用いた．音素バランス文セットから 2000 文を選択して学習データセット$\mathbb{U}_{2000}$とし，残り 100 文を標準的な評価データセット$\mathbb{U}_{\mathrm{s}}$とした．また，長文セットから 100 文を選択して例外的な評価データセット$\mathbb{U}_{\mathrm{e}}$とした．さらに，学習データセット$\mathbb{U}_{2000}$から 100 文，200 文，300 文，400 文，500 文，1000 文を選択して，それぞれ$\mathbb{U}_{100}$，$\mathbb{U}_{200}$，$\mathbb{U}_{300}$，$\mathbb{U}_{400}$，$\mathbb{U}_{500}$，$\mathbb{U}_{1000}$とした．さらに，100 文の学習データセット$\mathbb{U}_{100}$を学習内の評価データセット$\mathbb{U}_{\mathrm{c}}$とした．

この女性話者の音声コーパスを使用した理由は，他の音声コーパスよりも朗読調の音声が多いこと，ボコーダの分析部による音声波形の分析が安定していること，ボコーダの合成部による合成音声の品質が安定していること，音声コーパスを小規模にしても波形接続方式の音声合成システムによる合成音声の品質が高いからである．

2.3.1. 収録音声

雑音基準値（NC 値：Noise-Criterion 値 [27]）が NC-15 の防音設備のあるスタジオで音声を収録した．マイクは C414（AKG），マイクアンプは SB-2024（ADgear），オーディオインターフェースは Pro Tools HDX（Avid Technology）を使用した．サンプリング周波数は 48 kHz，量子化精度は 16 bit とした．サンプリング周波数を 48 kHz とした理由は，48 kHz のサンプリング周波数で収録した音声の品質の方が 16 kHz のサンプリング周波数で収録した音声の品質よりも良いからである．また，音声を扱う製品やサービスの多くがサンプリング周波数を 44.1 kHz や 48 kHz に設定しているため，音声合成システムはこれらのサンプリング周波数に対応している必要があるからである．

2.3.2. 言語特徴量

音声コーパスの原稿を 2.1.1 の言語規則に基づく言語解析部を用いて言語特徴量を算出した．時間フレーム情報は継続長から算出される．言語特徴量は呼気段落，アクセント句，モーラ，音素，時間フレームの階層構造を持ち，音声特徴量と時間的な対応関係を持つ（図 2.10）．

言語特徴量の属性の一覧を表 2.1 に示す．表中の属性名の「：」は区切り文字であり，区切り文字で属性名を分割したとき，左側にある要素は上位の概念を表す．例えば，「fall:org」は「fall:org:prv」，「fall:org:cur」，「fall:org:nxt」をまとめて表す．「fall:org」と「fall:mod」の違いは 0 型アクセントの表現方法である．「fall:org」では，0 型アクセントは 0 で表現され，「fall:mod」では，0 型アクセントはそのアクセント句のモーラの総数で表現される [28]．

言語特徴量は実数型の属性と列挙型の属性を持つ．言語特徴量を DNN の学習や予測に用いるには数値ベクトルでの表現が必要である．実数型の属性は属性値をそのまま使用するが，列挙型の属性は局所表現のベクトルに変換する．局所表現のベクトルは，ひとつの要素が 1 で，残りの要素が 0 であるベクトルのことである．例えば，3 つの項目がある列挙型の属性の局所表現のベクトルは，3 次元のベクトルで，各項目はそれぞれ$[1,0,0]$，$[0,1,0]$，$[0,0,1]$と表現される．これにより，言語特徴量の数値ベクトルは，521 次元のベクトルとなる．

2.3.3. 音声特徴量

2.1.2 のボコーダの分析部により音声波形から基本周波数，スペクトル包絡，非周期性指標を抽出した（図 2.10）．分析フレーム周期は 5 ms，離散フーリエ変換長は 2048 とした．基本周波数は声帯振動の時刻検出を用いた方法と基本波抽出法によって抽出される [4]．スペクトル包絡はピッチ同期分析法とケプストラム法によって抽出される [4]．非周期性指標は群遅延に基づくパラメータから推定する方法によって抽出される [23]．継続長は HMM で予測した音素境界から算出した．ただし，熟練したラベラーがほとんどの音素境界を音声の聴取とスペクトログラムを目視で確認して手動で修正した．

DNN でモデル化するために，基本周波数とスペクトル包絡に以下の前処理を適用した．基本周波数は無音区間および無声区間を補間して対数基本周波数に変換した．無音区間および無声区間を補間する理由は，基本周波数が抽出できない無音区間や無声区間における基本周波数は 0 Hz であり、基本周波数を統計的に扱う場合の外れ値となるからである．また，基本周波数を対数化するのは，音高の知覚が対数スケールに従うからである．

スペクトル包絡はリフタリングによる重み付けをされたケプストラムを離散フーリエ変換することで得られる．ケプストラムは対数パワースペクトルを離散フーリエ逆変換することで得られる．対数パワースペクトルは音声波形を離散フーリエ変換することで得られる．従って，離散フーリエ変換の対称性を考慮すると，スペクトル包絡の次元数は離散フーリエ変換の長さの半分に 1 を足した値であり，数百次元や数千次元となる．メルケプストラムはスペクトル包絡を小さい次元数で効率的に表現する．メルケプストラムは，ケプストラム領域において，対数パワースペクトル包絡の周波数スケールを線形スケールから人間の聴覚特性を考慮した周波数スケールへ変換することで得られる [29]．スペクトル包絡からメルケプストラムへの変換は次式で定義される．

$$
\begin{aligned}
\boldsymbol{y}_t &= \left[y_t^{(1)}, \cdots, y_t^{(d_1)}, \cdots, y_t^{(D_1)}, y_t^{(D_1-1)}, \cdots, y_t^{(2)}\right] \\
\boldsymbol{Y}_t &= \Re\left(\mathfrak{F}^{-1}(\log(\boldsymbol{y}_t \circ \boldsymbol{y}_t))\right) \\
\boldsymbol{C}_t &= \left[C_t^{(1)}, \cdots, C_t^{(d_1)}, \cdots, C_t^{(D_1)}\right] \\
C_t^{(d_1)} &= \begin{cases} \dfrac{Y_t^{(d_1)}}{2} & (d_1 = 1, D_1) \\ Y_t^{(d_1)} & (2 \le d_1 \le D_1 - 1) \end{cases} \\
\boldsymbol{c}_t &= \left[c_t^{(1)}, \cdots, c_t^{(d_2)}, \cdots, c_t^{(D_2)}\right] \\
c_t^{(d_2)} &= c_t^{(1, d_2)} \\
c_t^{(d_1, d_2)} &= \begin{cases} C_t^{(d_1)} + \alpha c_t^{(d_1-1, 1)} & (d_2 = 1) \\ (1-\alpha^2) c_t^{(d_1-1, 1)} + \alpha c_t^{(d_1-1, 2)} & (d_2 = 2) \\ c_t^{(d_1-1, d_2-1)} + \alpha\left(c_t^{(d_1-1, d_2)} - c_t^{(d_1, d_2-1)}\right) & \begin{cases}(d_2 = 3, 4, \cdots, D_2) \\ (d_1 = D_1, \cdots, 2, 1)\end{cases} \end{cases} \\
\alpha &= \frac{\alpha_1 - \alpha_2}{1 - \alpha_1 \alpha_2} \\
\boldsymbol{c}_t &\equiv \mathrm{freqt}(\boldsymbol{C}_t \mid D_1, \alpha_1, D_2, \alpha_2)
\end{aligned}
\tag{2.9}
$$

ここで，$\boldsymbol{y}_t$は時間フレームtにおけるスペクトル包絡，$\mathfrak{F}^{-1}$は離散フーリエ逆変換，$\Re$は実部を抽出する関数，$\circ$は要素ごとの積，$\boldsymbol{C}_t$は時間フレームtにおけるケプストラム，$\boldsymbol{c}_t$は時間フレームtにおけるメルケプストラム，D_1はスペクトル包絡の次元数，D_2はメルケプストラムの次元数，α_1はケプストラムの周波数伸長パラメータ，α_2はメルケプストラムの周波数伸長パラメータである．「freqt」は周波数変換関数であり，音声信号処理ツールキット（SPTK：Speech Signal Processing Toolkit）に定義されている [30]．ケプストラムの周波数スケールは線形であるため，α_1は 0 である．また，「freqt」により，メルケプストラムの周波数スケールを線形に変換することでケプストラムに変換することもできる．

本論文では，メルケプストラムの周波数伸長パラメータを 0.55 として，1025 次元のスペクトル包絡を 60 次元のメルケプストラムに変換した．20，30，40，50，60 次元のメルケプストラムから再合成した 5 つの音声と，スペクトル包絡から再合成した音声を比較した．その結果，60 次元のメルケプストラムから再合成した音声と，スペクトル包絡から再合成した音声とでは，音質に差がなかったことから，メルケプストラムの次元数を 60 とした．また，周波数伸長パラメータの値はサンプリング周波数に対応する値である．

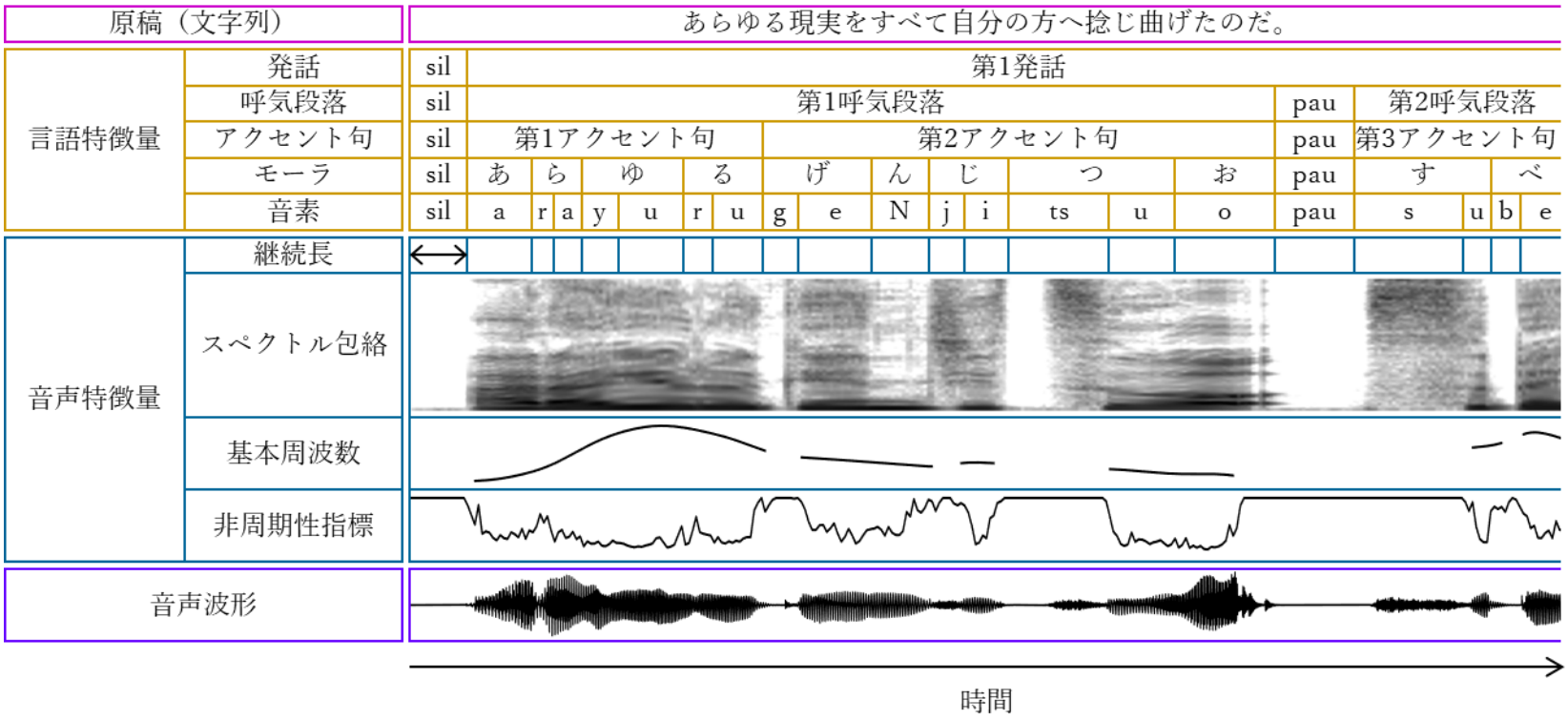

図 2.10 音声コーパスのデータ構造

言語特徴量の「sil」は無音であり，「pau」は息継ぎのポーズを表す．スペクトル包絡の縦軸は周波数であり，色の濃淡はスペクトル強度を示す．基本周波数の縦軸は周波数である．非周期性指標の縦軸は割合である．音声波形の縦軸は振幅である．

表 2.1　言語特徴量の属性の一覧

インデックス	属性名	型	説明	所属階層
1	n_bre:utt	実数型	当該発話における呼気段落の総数	発話
2	n_acc:utt	実数型	当該発話におけるアクセント句の総数	
3	n_mora:utt	実数型	当該発話におけるモーラの総数	
4	b_bre:utt:fwd	実数型	当該発話における呼気段落の呼気段落レベルでの昇順位置	
5	b_bre:utt:bwd	実数型	当該発話における呼気段落の呼気段落レベルでの降順位置	
6	a_bre:utt:fwd	実数型	当該発話における呼気段落のアクセント句レベルでの昇順位置	
7	a_bre:utt:bwd	実数型	当該発話における呼気段落のアクセント句レベルでの降順位置	
8	m_bre:utt:fwd	実数型	当該発話における呼気段落のモーラレベルでの昇順位置	
9	m_bre:utt:bwd	実数型	当該発話における呼気段落のモーラレベルでの降順位置	
10	a_acc:utt:fwd	実数型	当該発話におけるアクセント句のアクセント句レベルでの昇順位置	
11	a_acc:utt:bwd	実数型	当該発話におけるアクセント句のアクセント句レベルでの降順位置	
12	m_acc:utt:fwd	実数型	当該発話におけるアクセント句のモーラレベルでの昇順位置	
13	m_acc:utt:bwd	実数型	当該発話におけるアクセント句のモーラレベルでの降順位置	
14	m_mora:utt:fwd	実数型	当該発話におけるモーラのモーラレベルでの昇順位置	
15	m_mora:utt:bwd	実数型	当該発話におけるモーラのモーラレベルでの降順位置	
16	n_acc:bre:prv	実数型	当該呼気段落の前の呼気段落におけるアクセント句の総数	呼気段落
17	n_acc:bre:cur	実数型	当該呼気段落におけるアクセント句の総数	
18	n_acc:bre:nxt	実数型	当該呼気段落の次の呼気段落におけるアクセント句の総数	
19	n_mora:bre:prv	実数型	当該呼気段落の前の呼気段落におけるモーラの総数	
20	n_mora:bre:cur	実数型	当該呼気段落におけるモーラの総数	

表 2.1　言語特徴量の属性の一覧

21	n_mora:bre:nxt	実数型	当該呼気段落の次の呼気段落におけるモーラの総数	
22	a_acc:bre:fwd	実数型	当該呼気段落におけるアクセント句のアクセント句レベルでの昇順位置	
23	a_acc:bre:bwd	実数型	当該呼気段落におけるアクセント句のアクセント句レベルでの降順位置	
24	m_acc:bre:fwd	実数型	当該呼気段落におけるアクセント句のモーラレベルでの昇順位置	
25	m_acc:bre:bwd	実数型	当該呼気段落におけるアクセント句のモーラレベルでの降順位置	
26	m_mora:bre:fwd	実数型	当該呼気段落におけるモーラのモーラレベルでの昇順位置	
27	m_mora:bre:bwd	実数型	当該呼気段落におけるモーラのモーラレベルでの降順位置	
28	n_mora:acc:prv	実数型	当該アクセント句の前のアクセント句におけるモーラの総数	アクセント句
29	n_mora:acc:cur	実数型	当該アクセント句におけるモーラの総数	
30	n_mora:acc:nxt	実数型	当該アクセント句の前のアクセント句におけるモーラの総数	
31	m_mora:acc:fwd	実数型	当該アクセント句におけるモーラのモーラレベルでの昇順位置	
32	m_mora:acc:bwd	実数型	当該アクセント句におけるモーラのモーラレベルでの降順位置	
33	fall:org:prv	実数型	当該アクセント句の前のアクセント句のアクセント下降位置	
34	fall:org:cur	実数型	当該アクセント句のアクセント下降位置	
35	fall:org:nxt	実数型	当該アクセント句の次のアクセント句のアクセント下降位置	
36	fall:mod:prv	実数型	当該アクセント句の前のアクセント句の修正したアクセント下降位置	
37	fall:mod:cur	実数型	当該アクセント句の修正したアクセント下降位置	
38	fall:mod:nxt	実数型	当該アクセント句の次のアクセント句の修正したアクセント下降位置	
39	rise:prv	実数型	当該アクセント句の前のアクセント句のアクセント上昇位置	
40	rise:cur	実数型	当該アクセント句のアクセント上昇位置	
41	rise:nxt	実数型	当該アクセント句の次のアクセント句のアクセント上昇位置	

表 2.1　言語特徴量の属性の一覧

42	dur:utt	実数型	当該発話の時間フレームの総数	発話
43	t:utt:fwd	実数型	当該発話における時間フレームの時間フレームレベルでの昇順位置	
44	t:utt:bwd	実数型	当該発話における時間フレームの時間フレームレベルでの降順位置	
45	dur:bre	実数型	当該呼気段落の時間フレームの総数	呼気段落
46	t:bre:fwd	実数型	当該呼気段落における時間フレームの時間フレームレベルでの昇順位置	
47	t:bre:bwd	実数型	当該呼気段落における時間フレームの時間フレームレベルでの降順位置	
48	dur:acc	実数型	当該アクセント句の時間フレームの総数	アクセント句
49	t:acc:fwd	実数型	当該アクセント句における時間フレームの時間フレームレベルでの昇順位置	
50	t:acc:bwd	実数型	当該アクセント句における時間フレームの時間フレームレベルでの降順位置	
51	dur:mora	実数型	当該モーラの時間フレームの総数	モーラ
52	t:mora:fwd	実数型	当該モーラにおける時間フレームの時間フレームレベルでの昇順位置	
53	t:mora:bwd	実数型	当該モーラにおける時間フレームの時間フレームレベルでの降順位置	
54	dur:ph	実数型	当該音素の時間フレームの総数	音素
55	t:ph:fwd	実数型	当該音素における時間フレームの時間フレームレベルでの昇順位置	
56	t:ph:bwd	実数型	当該音素における時間フレームの時間フレームレベルでの降順位置	
57-59	pau_id:prv	列挙型	当該呼気段落とその前の呼気段落の間のポーズの種類	呼気段落
60-62	pau_id:nxt	列挙型	当該呼気段落とその次の呼気段落の間のポーズの種類	
63-70	eos_id:prv	列挙型	当該アクセント句の前のアクセント句の文末表現	アクセント句
71-78	eos_id:cur	列挙型	当該アクセント句の文末表現	
79-86	eos_id:nxt	列挙型	当該アクセント句の次のアクセント句の文末表現	
87-138	ph_id:prv2	列挙型	当該音素の 2 つ前の音素の名前	音素

表 2.1　言語特徴量の属性の一覧

139-190	ph_id:prv	列挙型	当該音素の前の音素の名前	
191-242	ph_id:cur	列挙型	当該音素の名前	
243-294	ph_id:nxt	列挙型	当該音素の次の音素の名前	
295-346	ph_id:nxt2	列挙型	当該音素の 2 つ次の音素の名前	
347-381	ph_art:prv2	列挙型	当該音素の 2 つ前の音素の調音	
382-416	ph_art:prv	列挙型	当該音素の前の音素の調音	
417-451	ph_art:cur	列挙型	当該音素の調音	
452-486	ph_art:nxt	列挙型	当該音素の次の音素の調音	
487-521	ph_art:nxt2	列挙型	当該音素の 2 つ次の音素の調音	

3. 合成処理を高速化するための音声特徴量予測部の構成

3.1. はじめに

計算資源が限られた計算環境においては，演算装置と記憶装置の制約が大きい．DNN の処理のほとんどは行列の積算であるため，低性能な演算装置では，DNN の処理時間が問題になり，小容量の記憶装置では，DNN のモデルサイズが問題となる．DNN の処理時間とモデルサイズは，DNN の種類，層数，ユニット数によって決まる．また，後処理についても，低性能な演算装置では，後処理の処理時間が問題になり，小容量の記憶装置では，後処理に必要なメモリサイズが問題になる．

そこで，本章では，予測時における音声特徴量予測部の DNN と後処理の処理時間を明らかにするとともに，DNN のモデルサイズと各処理に必要な記憶領域について述べ，計算資源が限られた計算環境を想定した音声特徴量予測部の構成を決める．

3.2. 音声特徴量予測部の構成

3 つの音声特徴量予測部の構成について述べる．1 つめは FFNN を用いた基本的な音声特徴量予測部の構成で，2 つめは RNN を用いた基本的な音声特徴量予測部の構成で，3 つめは計算資源が限られた音声特徴量予測部の構成である．

3.2.1. FFNN を用いた基本的な音声特徴量予測部の構成

FFNN を用いた基本的な音声特徴量予測部の構成を図 3.1 に示す [9]．継続長は，音素レベルの言語特徴量から言語特徴量の正規化部，FFNN，音声特徴量の逆正規化部を介すことで生成される．対数基本周波数は，時間フレームレベルの言語特徴量から言語特徴量の正規化部，FFNN，音声特徴量の逆正規化部，MLPG を介すことで生成される．メルケプストラムは，時間フレームレベルの言語特徴量から言語特徴量の正規化部，FFNN，音声特徴量の逆正規化部，MLPG，ケプストラム強調を介すことで生成される．非周期性指標は，時間フレームレベルの言語特徴量から言語特徴量の正規化部，FFNN を介すことで生成される．非周期性指標の値は 0 から 1 までの範囲にあり，正規化の必要がないため，音声特徴量の逆正規化部の処理を省略した．

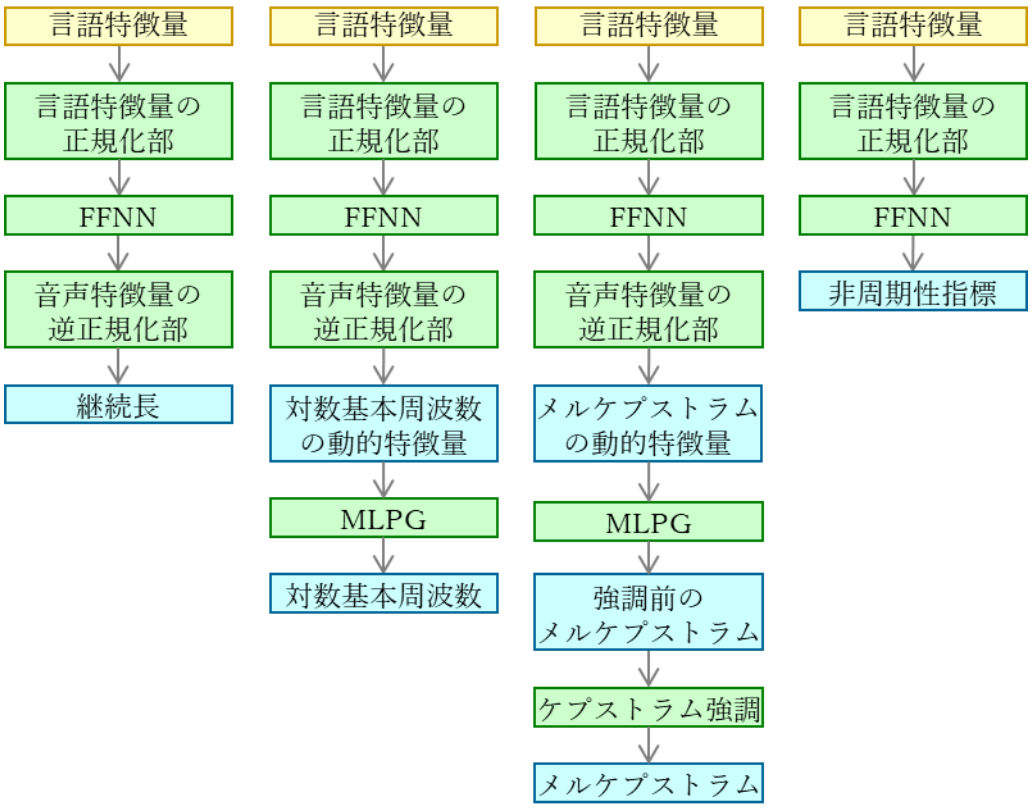

図 3.1　FFNN を用いた基本的な音声特徴量予測部の構成

3.2.2.　RNN を用いた基本的な音声特徴量予測部の構成

RNN を用いた基本的な音声特徴量予測部の構成を図 3.2 に示す [31]. 継続長は，音素レベルの言語特徴量から言語特徴量の正規化部，RNN，音声特徴量の逆正規化部を介すことで生成される．対数基本周波数は，時間フレームレベルの言語特徴量から言語特徴量の正規化部，RNN，音声特徴量の逆正規化部を介すことで生成される．メルケプストラムは，時間フレームレベルの言語特徴量から言語特徴量の正規化部，RNN，音声特徴量の逆正規化部，ケプストラム強調を介すことで生成される．非周期性指標は，時間フレームレベルの言語特徴量から言語特徴量の正規化部，RNN を介すことで生成される．非周期性指標の値は 0 から 1 までの範囲にあり，正規化の必要がないため，音声特徴量の逆正規化部の処理を省略した．隣接する時間フレーム間の音声特徴量の関係を RNN が学習するため，MLPG は必要ない．

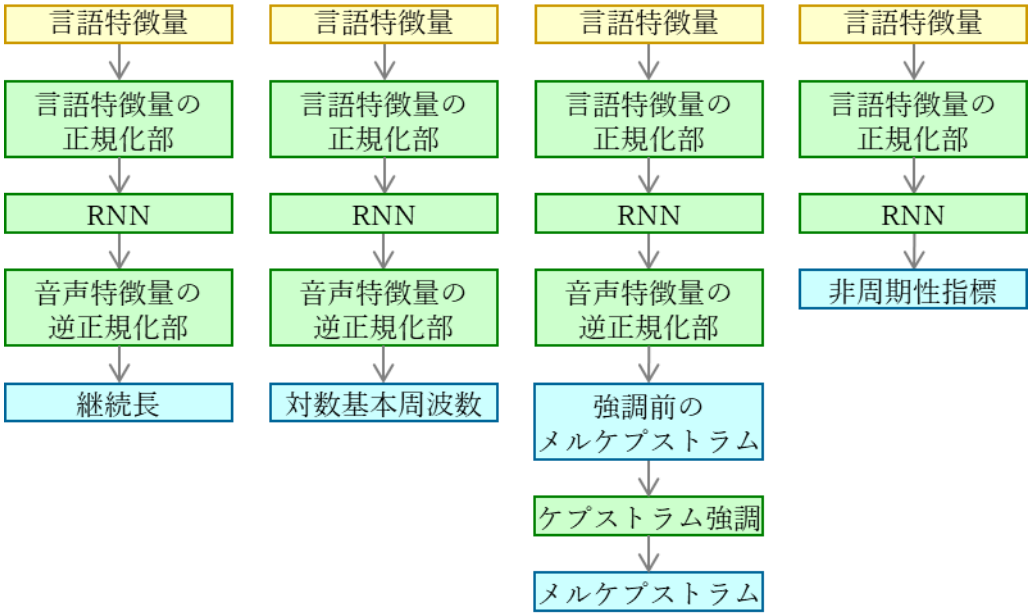

図 3.2　RNN を用いた基本的な音声特徴量予測部の構成

3.2.3.　計算資源が限られた音声特徴量予測部の構成

計算資源が限られた音声特徴量予測部の構成を図 3.3 に示す．継続長は，音素レベルの言語特徴量から言語特徴量の正規化部，FFNN，音声特徴量の逆正規化部を介すことで生成される．対数基本周波数は，時間フレームレベルの言語特徴量から言語特徴量の正規化部，FFNN，音声特徴量の逆正規化部を介すことで生成される．メルケプストラムは，時間フレームレベルの言語特徴量から言語特徴量の正規化部，FFNN，音声特徴量の逆正規化部，を介すことで生成される．非周期性指標は，時間フレームレベルの言語特徴量から言語特徴量の正規化部，FFNN を介すことで生成される．非周期性指標の値は 0 から 1 までの範囲にあり，正規化の必要がないため，音声特徴量の逆正規化部の処理を省略した．計算量を削減するために，MLPG とケプストラム強調を排除した．2.2.2 節で述べたように，FFNN は時間フレームごとに独立して音声特徴量をモデル化するため，時系列のモデル化には適していない．この構成で，RNN で生成した音声特徴量や，後処理を適用した音声特徴量に匹敵する音声特徴量を予測するには，FFNN の学習法を工夫する必要がある．その学習法は 5 章と 6 章で述べる．

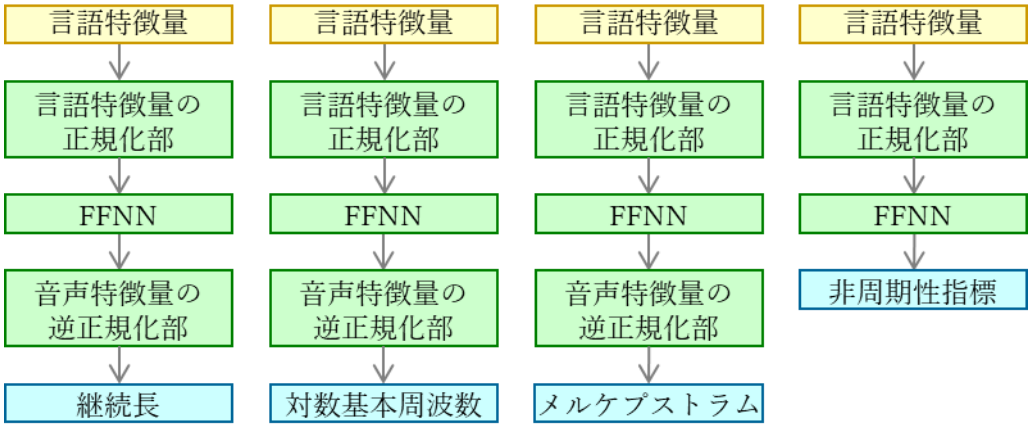

図 3.3　計算資源が限られた音声特徴量予測部の構成

3.3.　実験方法

3.3.1.　DNN の構成

3.2.1，3.2.2，3.2.3 節で述べた音声特徴量予測部で用いる DNN の構成を表 3.1 に示す．表中のDは音声特徴量の次元数である．FFNN-3.2.1 は，ユニット数が 512 で，活性化関数が正規化線形関数（ReLU 関数：Rectified Linear Unit 関数 [32]）の 4 層の全結合層と，ユニット数が$3D$またはDで，活性化関数を持たない全結合層で構成される．RNN-3.2.2 は，ユニット数が 320 で，活性化関数が双曲線正接関数（tanh 関数：hyperbolic tangent 関数）の LSTM 層と，ユニット数がDで，活性化関数を持たない再帰層で構成される．FFNN-3.2.3 は，ユニット数が 512 で，活性化関数が ReLU 関数の 4 層の全結合層と，ユニット数がDで，活性化関数を持たない全結合層で構成される．

後処理に MLPG を用いる対数基本周波数とメルケプストラムについては，0 次から 2 次までの動的特徴を考慮するため，FFNN-3.2.1 の 5 層目のユニット数を 3 倍にした．DNN のモデルサイズは，DNN のモデルパラメータを単精度浮動小数点として算出したものである．DNN のモデルパラメータ数は，荷重行列と閾値ベクトルの要素数の合計である．音声特徴量ごとに次元数は異なるが，いずれの音声特徴量の DNN のモデルサイズも，概ね表中の値となる．計算資源が限られた計算環境を考慮して，モデルサイズが数 MB になるように層数やユニット数を設定した．

表 3.1　音声特徴量予測部で用いる DNN の構成

識別名	FFNN-3.2.1	RNN-3.2.2	FFNN-3.2.3
1 層目	全結合層 512 units / ReLU	LSTM 層 320 units / tanh	全結合層 512 units / ReLU
2 層目	全結合層 512 units / ReLU	再帰層 D units / —	全結合層 512 units / ReLU
3 層目	全結合層 512 units / ReLU		全結合層 512 units / ReLU
4 層目	全結合層 512 units / ReLU		全結合層 512 units / ReLU
5 層目	全結合層 $\begin{cases} 3D \text{ units } (\text{MLPG あり}) / - \\ D \text{ units } (\text{MLPG なし}) / - \end{cases}$		全結合層 D units / —
モデルサイズ	約 4 MB	約 4 MB	約 4 MB

3.3.2. 計算機の構成と実装方法

DNN や MLPG の処理は，音声特徴量の次元数による違いを除き，どの音声特徴量についても同じであるため，音声特徴量の中で最も次元数が大きいメルケプストラムについて処理時間を計測した. 計測に用いた計算機の中央演算装置（CPU：Central Processing Unit），画像処理用演算装置（GPU：Graphics Processing Unit），主記憶装置容量（メモリサイズ）を表 3.2 に示す．各処理は Python で実装した．GPU を搭載しない計算機における音声特徴量予測部では，DNN の実装には DNN ライブラリ（TensorFlow をバックエンドとする Keras），MLPG の実装には疎行列ライブラリ（SciPy の sparse），ケプストラム強調の実装には数値演算ライブラリ（NumPy）を使用した．GPU を搭載する計算機における音声特徴量予測部では，DNN，MLPG，ケプストラム強調の実装には Keras を使用した．TensorFlow, NumPy, SciPy は C 言語によって実装されている．GPU が利用可能な場合，TensorFlow は GPU 向けの汎用並列コンピューティング・アーキテクチャ（CUDA：Compute Unified Device Architecture）を介して GPU により処理を行う．処理時間の計測は，時間長が 5000 ms の言語特徴量を用いた．この時間長は，合成音声の時間長と同じであり，時間フレーム周期（5 ms）と時間フレーム数（1000 フレーム）の積算により計算される．処理時間は 100 回の計測値の平均とした．

表 3.2　処理時間の計測に用いた計算機の構成

識別名	CPU	GPU	メモリサイズ
i5-3317U	Intel i5-3317U 2 cores / 4 threads 1.7 GHz / 269 GFLOPS	—	4 GB
i3-5005U	Intel i3-5005U 2 cores / 4 threads 2.0 GHz / 326 GFLOPS	—	4 GB
i7-6700K	Intel i7-6700K 4 cores / 8 threads 4.0 GHz / 442 GFLOPS	—	32 GB
i9-9900K	Intel i9-9900K 8 cores / 16 threads 3.6 GHz / 461 GFLOPS	—	32 GB
GTX-1080	Intel i7-6700K 4 cores / 8 threads 4.0 GHz / 442 GFLOPS	NVIDIA GTX-1080 2560 cores / 1607 MHz 8 GB / 8.9 TFLOPS	32 GB
RTX-2080	Intel i9-9900K 8 cores / 16 threads 3.6 GHz / 461 GFLOPS	NVIDIA RTX-2080 2944 cores / 1515 MHz 8 GB / 10.1 TFLOPS	32 GB

3.4.　実験結果

DNN，MLPG，ケプストラム強調の処理時間を表 3.3 に示す．合計の処理時間は，合成音声の時間長である 5000 ms よりも短かった. GPU を搭載しない計算機においては, FFNN の処理時間は RNN の処理時間の 1/4 倍以下であった. GPU を搭載する計算機においては，FFNN の処理時間は RNN の処理時間の 1/100 倍以下であった．これは，FFNN は並列処理ができるのに対し, RNN は再帰構造のため並列処理ができないためである．FFNN-3.2.1 の処理時間については，どの計算機においても MLPG の処理時間が占める割合が大きかった．当然ではあるが，FFNN-3.2.3 は，MLPG やケプストラム強調の処理がなく，DNN が再帰構造を持たないため，合計の処理時間は最も短かった.

CPU の性能と処理時間については，コア数やスレッド数が多く，動作周波数が高い CPU ほど処理時間は短くなった．CPU の性能と処理時間は単純な比例関係ではないが，CPU の性能が向上すれば，処理時間は短くなるという相関関係があった．また，GPU を搭載した計算機における RNN-3.2.2 の DNN の処理時間については，GTX-1080 の処理時間の方が RTX-2080 の処理時間よりも短かった．これは RNN の構造上，並列処理ができないため，動作周波数が高い GTX-1080 の方が有利となる．一方で，MLPG やケプストラム強調は並

列処理できるため，コア数の多い RTX-2080 の方が有利となる．

表 3.3　計算機と音声特徴量予測部の構成について処理時間（ミリ秒）

識別名	音声特徴量 予測部の構成	合計	DNN	MLPG	ケプストラム 強調
i5-3317U	3.2.1	3669	122	2961	586
	3.2.2	1080	494	—	586
	3.2.3	122	122	—	—
i3-5005U	3.2.1	3753	53	3205	495
	3.2.2	898	403	—	495
	3.2.3	53	53	—	—
i7-6700K	3.2.1	1521	23	1303	195
	3.2.2	317	122	—	195
	3.2.3	23	23	—	—
i9-9900K	3.2.1	1287	23	1123	141
	3.2.2	242	101	—	141
	3.2.3	23	23	—	—
GTX-1080	3.2.1	395	1	370	24
	3.2.2	189	165	—	24
	3.2.3	1	1	—	—
RTX-2080	3.2.1	351	1	328	22
	3.2.2	206	184	—	22
	3.2.3	1	1	—	—

3.5.　考察

6 種類の計算機において，3 種類の音声特徴量予測部の処理時間を計測した．計測した処理時間は，音声特徴量予測部の処理時間だけであり，言語処理部や波形生成部の処理時間は考慮されていない．そのため，音声合成システム全体の処理時間を考慮すると，i5-3317U や i3-5005U における FFNN-3.2.1 の構成の処理時間は，合成音声の時間長を超える可能性がある．合成処理をしながら生成された音声波形を再生デバイスに書き込むような逐次処理では，音声合成システム全体の処理時間が合成音声の時間長よりも長くなると，再生される音声が途切れてしまう．また，音声波形をファイルとして出力し，そのファイルを利用するようなバッチ処理では，合成処理が完了するまでファイルは利用できないため，処理時間はできる限り短い方が好ましい．

MLPG の処理時間は，DNN やケプストラム強調の処理時間よりも数十倍以上長かった．

これは MLPG の実装が最適でないためである．C 言語での実装や，再帰型の MLPG を利用すれば，MLPG の処理時間は短くできる [33].

DNN のモデルサイズは，ひとつの音声特徴量に対して約 4 MB であり，4 つの音声特徴量の DNN のモデルサイズの合計は 20 MB 以下である．波形接続方式の音声合成システムが音声波形を生成するために利用する音声コーパスのサイズは数百 MB から数 GB である．近年の補助記憶装置の容量を考慮しても，表 3.1 の DNN のモデルサイズは小さいといえる．

合成処理に必要な記憶領域は，各処理の実装方法によって異なる．そのため，ここでは，単精度浮動小数点で換算したときの各処理に最低限必要な記憶領域について述べる．DNN の処理では, DNN のモデルパラメータ, 言語特徴量, 音声特徴量の記憶領域が必要である．DNN のモデルパラメータの記憶領域は DNN のモデルサイズと同じである．言語特徴量と音声特徴量の記憶領域は，時間フレーム数に依る．時間フレーム数が 1000 の場合，言語特徴量の記憶領域は約 2 MB，動的特徴量を考慮した音声特徴量の記憶領域は約 1 MB，動的特徴量を考慮しない音声特徴量の記憶領域は数百 KB となる．

MLPG の処理では，式（2.3）の$\boldsymbol{W}$や$\boldsymbol{U}^{-1}$に関連する記憶領域が必要であり，その記憶領域は時間フレーム数に依る. 時間フレーム数が 1000 の場合, $\boldsymbol{U}^{-1}$が対角行列であることや, $\boldsymbol{W}^{\top}\boldsymbol{U}^{-1}\boldsymbol{W}$が対称性のある帯行列であることを考慮すると，MLPG の処理に必要な記憶領域は数 MB となる．

ケプストラム強調の処理では, 補正係数の計算に離散フーリエ変換を使うため, 離散フーリエ変換の長さ分の記憶領域が必要となる．ケプストラム強調は時間フレームごとに処理を行うため，1 フレーム分の処理に必要な記憶領域を考慮する．離散フーリエ変換の長さが 2048 の場合，記憶領域は約 8 KB となる．

以上より，モデルサイズはどの DNN の構成も同じであるが，計算資源が限られた計算環境においては，処理時間や処理に必要な記憶領域を考慮すると，FFNN-3.2.3 の構成は他の構成よりも優れている．

3.6. まとめ

音声特徴量予測部の DNN と後処理の処理時間を明らかにするとともに，DNN のモデルサイズと各処理に必要な記憶領域について述べた．その結果，後処理のない FFNN による構成が計算資源の限られた計算環境に適していることを明らかにした．

4. 頑健な音声特徴量の予測を可能にする言語特徴量の正規化法

4.1. はじめに

本章では, 音声特徴量予測部に用いられる DNN の学習外データに対する頑健性について述べる. あらゆる日本語文章の文字列から音声を合成する場合, 入力された文字列の文章構造が, 学習データセットの原稿の文章構造と異なることがある. これは学習データセットの言語特徴量が網羅する値の範囲外の値が言語特徴量に含まれることを意味する. 本章では、このような学習データセットが網羅しない値を外れ値と呼ぶ. DNN は入力ベクトにモデルパラメータを乗算し、それに活性化関数を適用することにより出力ベクトルを生成するため, DNN に外れ値が入力されると, その影響が出力ベクトルに現れる. また, DNN のモデルパラメータは, 学習データセットを基準にして学習されるため, 外れ値には対応できず, 不安定な出力ベクトルを生成してしまう. 一般的な外れ値の対策としては, 学習データを増やして, 多くの言語特徴量のパターンを学習する. しかし, この方法はすべての入力パターンを網羅できないため, 完璧な外れ値の対策にはならない.

本研究では, 学習データと構造が異なる文章に対しても頑健な音声特徴量の予測を可能にするために, いかなる文章に対しても外れ値が発生しない新しい言語特徴量の正規化法を提案する. 文章構造と密接に関係する基本周波数について, 提案法と従来法を比較することで, 提案法の有効性を示す. また, 言語特徴量の各属性と基本周波数の関連性を明らかにすることで, 正規化法が異なっていても, 言語特徴量の各属性と基本周波数の関連性が保たれることを確認する.

4.2. 言語特徴量の正規化法

本章では 4 つの言語特徴量の正規化法について述べる. 1 つめは従来法で, 学習データセットの言語特徴量の最小値と最大値で言語特徴量を正規化する. 2 つめは広範囲版の従来法で, 外れ値を想定した言語特徴量の最小値と最大値で言語特徴量を正規化する. 3 つめはクリッピング版の従来法で, 学習データセットの言語特徴量の最小値から最大値までの範囲に値を制限して言語特徴量を正規化する. 4 つめは提案法で, 2 つの言語特徴量の属性値の比を取ることで正規化する.

4.2.1. 従来法：Min-Max 正規化法

Min-Max 正規化法は, 学習データセットから求めた言語特徴量の最小値と最大値により言語特徴量を正規化する. 一文の言語特徴量を次式で定義する.

$$\begin{aligned} \boldsymbol{x} &= [\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_T^\top]^\top \\ \boldsymbol{x}_t &= \left[x_t^{(1)}, \cdots, x_t^{(k)}, \cdots, x_t^{(K)}\right] \end{aligned} \tag{4.1}$$

ここで, $\boldsymbol{x}$は言語特徴量ベクトル系列, $\boldsymbol{x}_t$は時間フレームtの言語特徴量ベクトル, tは時間フレームインデックス, Tは時間フレーム数, kは言語特徴量ベクトルの次元インデックス,

Kは言語特徴量ベクトルの次元数である．kは表 2.1 の 1 列目を指す．k次の言語特徴量の最小値と最大値は次式となる．

$$\begin{aligned} l^{(k)} &= \min_{t,\mathbb{T}} x_t^{(k)} \\ u^{(k)} &= \max_{t,\mathbb{T}} x_t^{(k)} \\ \mathbb{T} &\in \{\mathbb{U}_{100}, \mathbb{U}_{200}, \mathbb{U}_{300}, \mathbb{U}_{400}, \mathbb{U}_{500}, \mathbb{U}_{1000}, \mathbb{U}_{2000}\} \\ \boldsymbol{x} &\in \mathbb{T} \end{aligned} \tag{4.2}$$

ここで，$l^{(k)}$はk次の言語特徴量の最小値，$u^{(k)}$はk次の言語特徴量の最大値，$\mathbb{T}$は学習データセットである．$l^{(k)}$と$u^{(k)}$により正規化したk次の言語特徴量は次式となる．

$$\frac{x_t^{(k)} - l^{(k)}}{u^{(k)} - l^{(k)}} \tag{4.3}$$

この正規化法では，正規化された外れ値は外れ値のままであるため，この正規化法は外れ値の問題を抱えたままである．

4.2.2. 広範囲版の従来法：広範囲版の Min-Max 正規化法

この正規化法は本質的には Min-Max 正規化法と同じである．ただし，この正規化法は，学習データセットから算出した最小値以下の最小値と，学習データセットから算出した最大値以上の最大値を使用する．この正規化法の最小値と最大値は次式で定義される．

$$\begin{aligned} \hat{l}^{(k)} &= \min_{t,\mathbb{U}} x_t^{(k)} \quad \left(\hat{l}^{(k)} \le l^{(k)}\right) \\ \hat{u}^{(k)} &= \max_{t,\mathbb{U}} x_t^{(k)} \quad \left(\hat{u}^{(k)} \ge u^{(k)}\right) \\ \mathbb{U} &= \mathbb{T} \cup \mathbb{U}_{\mathrm{e}} \end{aligned} \tag{4.4}$$

ここで，$\hat{l}^{(k)}$は外れ値を考慮したk次の言語特徴量の最小値，$\hat{u}^{(k)}$は外れ値を考慮したk次の言語特徴量の最大値である．$\hat{l}^{(k)}$と$\hat{u}^{(k)}$により正規化したk次元の言語特徴量は次式となる．

$$\frac{x_t^{(k)} - \hat{l}^{(k)}}{\hat{u}^{(k)} - \hat{l}^{(k)}} \tag{4.5}$$

この正規化法と 4.2.1 の正規化法における言語特徴量ベクトルの外れ値について述べる．時間フレームtのk次の言語特徴量ベクトルの外れ値を次式で定義する．

$$\hat{x}_t^{(k)} = x_t^{(k)} + e^{(k)} \tag{4.6}$$

ここで，$e^{(k)}$は$l^{(k)}$から$u^{(k)}$までの範囲を超える差分とする．$\hat{x}_t^{(k)}$を$l^{(k)}$と$u^{(k)}$および$\hat{l}^{(k)}$と$\hat{u}^{(k)}$で正規化したときの差分の関係は次式となる．

$$\frac{e^{(k)}}{u^{(k)} - l^{(k)}} \ge \frac{e^{(k)}}{\hat{u}^{(k)} - \hat{l}^{(k)}} \tag{4.7}$$

正規化の範囲を拡大することで，DNN へ入力される外れ値の差分が小さくなるため，出力ベクトルへの影響を小さくできる．ただし，この正規化法と 3.2.1 の正規化法の違いは、正規化後の言語特徴量ベクトルの値のスケールだけであり，この正規化法は外れ値の問題を根本的に解決しない．

4.2.3. クリッピング版の従来法：クリッピング版の Min-Max 正規化法

この正規化法は本質的には Min-Max 正規化法と同じである．ただし，この正規化法は，言語特徴量の値を学習外データセットから算出した最小値と最大値の範囲に制限する．これにより，正規化後の言語特徴量ベクトルの値は必ず 0 から 1 までの範囲に収まる．この正規化法により正規化したk次の言語特徴量は次式となる．

$$\begin{cases} 0 & \left(x_t^{(k)} < l^{(k)}\right) \\ \dfrac{x_t^{(k)} - l^{(k)}}{u^{(k)} - l^{(k)}} & \left(l^{(k)} \leq x_t^{(k)} \leq u^{(k)}\right) \\ 1 & \left(x_t^{(k)} > u^{(k)}\right) \end{cases} \tag{4.8}$$

言語特徴量ベクトルの値を最小値から最大値までの範囲に限ることにより、DNN は安全に学習データセットから獲得したモデルパラメータに基づいて出力ベクトルを予測できる。しかし，この正規化法は外れ値を最小値から最大値までの範囲へ丸め込むだけであり，外れ値の問題を根本的に解決しない．

4.2.4. 提案法：2 つの言語特徴量の属性値の比を取る正規化法

この正規化法は，学習データセットから算出する最小値と最大値に依存せず，一文の言語特徴量ベクトル系列内の値のみを用いて正規化をする．この正規化法は次式に従う．

$$\frac{x_t^{(k_1)}}{x_t^{(k_2)}} \quad (k_1 \neq k_2) \tag{4.9}$$

k_1とk_2の具体的な組み合わせを表 4.1 に示す．図 2.10 のように，言語特徴量は，発話，呼気段落，アクセント句，モーラ，音素の階層構造をしている．言語特徴量の実数型の属性は，子要素の総数や，下位の階層レベルで数えたときの位置を表すものばかりである．この正規化法は，言語特徴量の階層構造に着目して，属性値を相対的な値で表現する．例えば，表 2.1 の「fall:org:cur」を「n_mora:acc:cur」で除することで，「fall:org:cur」は当該アクセント句における相対的なアクセント下降位置として表現できる．また，表 2.1 の「n_mora:acc:cur」を「n_mora:utt」で除することで，「n_mora:acc:cur」は発話全体のモーラの総数に対する当該アクセント句のモーラの総数の割合として表現できる．このように，この正規化法は言語特徴量の階層構造に基づくため，いかなる文章であっても正規化後の言語特徴量ベクトルの値は必ず 0 から 1 までの範囲に収まる．

表 4.1　提案する正規化法を適用したときの言語特徴量の属性の一覧

提案する正規化は実数型の属性のみが対象であり，列挙型の属性は局所表現のベクトルをそのまま利用する．k_1とk_2は表 2.1 のインデックスを指す．

インデックス	属性名	k_1	k_1が指す属性名	k_2	k_2が指す属性名	所属階層
1	n_bre_acc:utt	1	n_bre:utt	2	n_acc:utt	発話
2	n_bre_mora:utt	1	n_bre:utt	3	n_mora:utt	
3	n_acc_mora:utt	2	n_acc:utt	3	n_mora:utt	
4	b_bre:utt:fwd	4	b_bre:utt:fwd	1	n_bre:utt	
5	b_bre:utt:bwd	5	b_bre:utt:bwd	1	n_bre:utt	
6	a_bre:utt:fwd	6	a_bre:utt:fwd	2	n_acc:utt	
7	a_bre:utt:bwd	7	a_bre:utt:bwd	2	n_acc:utt	
8	m_bre:utt:fwd	8	m_bre:utt:fwd	3	n_mora:utt	
9	m_bre:utt:bwd	9	m_bre:utt:bwd	3	n_mora:utt	
10	a_acc:utt:fwd	10	a_acc:utt:fwd	2	n_acc:utt	
11	a_acc:utt:bwd	11	a_acc:utt:bwd	2	n_acc:utt	
12	m_acc:utt:fwd	12	m_acc:utt:fwd	3	n_mora:utt	
13	m_acc:utt:bwd	13	m_acc:utt:bwd	3	n_mora:utt	
14	m_mora:utt:fwd	14	m_mora:utt:fwd	3	n_mora:utt	
15	m_mora:utt:bwd	15	m_mora:utt:bwd	3	n_mora:utt	
16	n_acc:bre:prv	16	n_acc:bre:prv	2	n_acc:utt	呼気段落
17	n_acc:bre:cur	17	n_acc:bre:cur	2	n_acc:utt	
18	n_acc:bre:nxt	18	n_acc:bre:nxt	2	n_acc:utt	
19	n_mora:bre:prv	19	n_mora:bre:prv	3	n_mora:utt	

表 4.1　提案する正規化法を適用したときの言語特徴量の属性の一覧

20	n_mora:bre:cur	20	n_mora:bre:cur	3	n_mora:utt	
21	n_mora:bre:nxt	21	n_mora:bre:nxt	3	n_mora:utt	
22	a_acc:bre:fwd	22	a_acc:bre:fwd	17	n_acc:bre:cur	
23	a_acc:bre:bwd	23	a_acc:bre:bwd	17	n_acc:bre:cur	
24	m_acc:bre:fwd	24	m_acc:bre:fwd	20	n_mora:bre:cur	
25	m_acc:bre:bwd	25	m_acc:bre:bwd	20	n_mora:bre:cur	
26	m_mora:bre:fwd	26	m_mora:bre:fwd	20	n_mora:bre:cur	
27	m_mora:bre:bwd	27	m_mora:bre:bwd	20	n_mora:bre:cur	
28	n_mora:acc:prv	28	n_mora:acc:prv	3	n_mora:utt	
29	n_mora:acc:cur	29	n_mora:acc:cur	3	n_mora:utt	
30	n_mora:acc:nxt	30	n_mora:acc:nxt	3	n_mora:utt	
31	m_mora:acc:fwd	31	m_mora:acc:fwd	29	n_mora:acc:cur	
32	m_mora:acc:bwd	32	m_mora:acc:bwd	29	n_mora:acc:cur	
33	fall:org:prv	33	fall:org:prv	28	n_mora:acc:prv	
34	fall:org:cur	34	fall:org:cur	29	n_mora:acc:cur	
35	fall:org:nxt	35	fall:org:nxt	30	n_mora:acc:nxt	
36	fall:mod:prv	36	fall:mod:prv	28	n_mora:acc:prv	
37	fall:mod:cur	37	fall:mod:cur	29	n_mora:acc:cur	
38	fall:mod:nxt	38	fall:mod:nxt	30	n_mora:acc:nxt	
39	rise:prv	39	rise:prv	28	n_mora:acc:prv	
40	rise:cur	40	rise:cur	29	n_mora:acc:cur	

表 4.1　提案する正規化法を適用したときの言語特徴量の属性の一覧

41	rise:nxt	41	rise:nxt	30	n_mora:acc:nxt	
42	t:utt:fwd	43	t:utt:fwd	42	dur:utt	発話
43	t:utt:bwd	44	t:utt:bwd	42	dur:utt	
44	dur:bre:utt	45	dur:bre	42	dur:utt	呼気段落
45	t:bre:fwd	46	t:bre:fwd	45	dur:bre	
46	t:bre:bwd	47	t:bre:bwd	45	dur:bre	
47	dur:acc:utt	48	dur:acc	42	dur:utt	アクセント句
48	dur:acc:bre	48	dur:acc	45	dur:bre	
49	t:acc:fwd	49	t:acc:fwd	48	dur:acc	
50	t:acc:bwd	50	t:acc:bwd	48	dur:acc	
51	dur:mora:utt	51	dur:mora	42	dur:utt	モーラ
52	dur:mora:bre	51	dur:mora	45	dur:bre	
53	dur:mora:acc	51	dur:mora	48	dur:acc	
54	t:mora:fwd	52	t:mora:fwd	51	dur:mora	
55	t:mora:bwd	53	t:mora:bwd	51	dur:mora	
56	dur:ph:utt	54	dur:ph	42	dur:utt	音素
57	dur:ph:bre	54	dur:ph	45	dur:bre	
58	dur:ph:acc	54	dur:ph	48	dur:acc	
59	dur:ph:mora	54	dur:ph	51	dur:mora	
60	t:ph:fwd	55	t:ph:fwd	54	dur:ph	
61	t:ph:bwd	56	t:ph:bwd	54	dur:ph	

表 4.1　提案する正規化法を適用したときの言語特徴量の属性の一覧

62-64	pau_id:prv	—	—	—	—	呼気段落
65-67	pau_id:nxt	—	—	—	—	
68-75	eos_id:prv	—	—	—	—	アクセント句
76-83	eos_id:cur	—	—	—	—	
84-91	eos_id:nxt	—	—	—	—	
92-143	ph_id:prv2	—	—	—	—	音素
144-195	ph_id:prv	—	—	—	—	
196-247	ph_id:cur	—	—	—	—	
248-299	ph_id:nxt	—	—	—	—	
300-351	ph_id:nxt2	—	—	—	—	
352-386	ph_art:prv2	—	—	—	—	
387-421	ph_art:prv	—	—	—	—	
422-456	ph_art:cur	—	—	—	—	
457-491	ph_art:nxt	—	—	—	—	
492-526	ph_art:nxt2	—	—	—	—	

4.3. 音声特徴量予測部の構成

対数基本周波数の予測部の構成は 3.2.1 であり，DNN の構成は表 3.1 の FFNN-3.2.1 であり，図 2.4 の学習時の構成で FFNN-3.2.1 を学習した．FFNN は正規化された言語特徴量と正規化された対数基本周波数の動的特徴量の関係を学習する．損失関数は正規化された対数基本周波数の動的特徴量の平均二乗誤差を計算する．対数基本周波数の次元数Dは 1 である．勾配法は適応モーメント推定法（Adam 法：Adaptive Moment Estimation 法 [34]）であり，Adam 法のパラメータについては，学習率を 0.001，β_1を 0.9，β_2を 0.999，微小量を10^{-7}，学習率減衰を 0.0 とした．エポック数は 20 とし，バッチサイズは 1 文ごとの対数基本周波数の時間フレーム数とした．

4.4. 学習データセットと評価データセット

本章で使用する学習データセットと評価データセットについて述べる．学習データ量と対数基本周波数の予測誤差の関係を調べるために，$\mathbb{U}_{100}$，$\mathbb{U}_{200}$，$\mathbb{U}_{300}$，$\mathbb{U}_{400}$，$\mathbb{U}_{500}$，$\mathbb{U}_{1000}$，$\mathbb{U}_{2000}$の 7 つの学習データセットを使用した．また，外れ値に対する対数基本周波数の予測誤差を調べるために，$\mathbb{U}_{\mathrm{e}}$を評価データセットに使用した．各学習データセットの言語特徴量に対する$\mathbb{U}_{\mathrm{e}}$の言語特徴量が外れ値を含む割合を表 4.2 に示す．表中の最小値と最大値は，各学習データセットから算出した各属性の最小値と最大値で正規化されたものである．また，割合の「—」は 0%，最小値と最大値の「—」は学習データセットから算出した各属性の最小値と最大値の範囲以内であることを表している．表中の外れ値の割合は，1 文の言語特徴量の時間フレーム数あたりの外れ値を含む時間フレーム数であるため，発話関連の属性が外れ値を持つと，外れ値の割合は必然的に高くなる．$\mathbb{U}_{\mathrm{e}}$は長文セットから作成したため，当該発話におけるモーラの総数など，発話関連の属性が外れ値を含む可能性は高い．一方で，呼気段落，アクセント句，モーラ，音素関連の属性は，ほとんど外れ値を含んでいないこととから，今回使用した学習データセットは，$\mathbb{U}_{100}$であっても，これらの言語特徴量の属性を広く網羅するものであるといえる．

表 4.2　評価データセット$\mathbb{U}_e$の言語特徴量が外れ値を含む時間フレームの割合と，言語特徴量の各属性の正規化された最小値と最大値

学習データセット	$\mathbb{U}_{100}$			$\mathbb{U}_{200}$		
属性名	割合	最小値	最大値	割合	最小値	最大値
n_bre:utt	92.3	—	3.80	92.3	—	3.17
n_acc:utt	94.0	—	3.36	92.3	—	3.17
n_mora:utt	94.1	-0.02	3.18	94.0	—	2.18
b_bre:utt	82.6	—	3.18	92.8	—	2.75
a_bre:utt	93.0	—	3.67	71.1	—	3.18
m_bre:utt	94.0	—	3.48	92.0	—	2.83
a_acc:utt	92.5	—	3.36	68.6	—	2.18
m_acc:utt	94.0	—	3.34	90.6	—	2.77
m_mora:utt	95.7	—	3.10	89.9	—	2.73
dur:utt	94.0	—	4.34	94.0	—	3.27
t:utt	94.0	—	4.15	88.4	—	3.20
n_acc:bre	4.1	—	1.20	4.1	—	1.20
n_mora:bre	1.4	—	1.28	1.4	—	1.28
a_acc:bre	0.9	—	1.20	0.9	—	1.20
m_acc:bre	0.7	—	1.39	0.7	—	1.39
m_mora:bre	0.4	—	1.24	0.4	—	1.24
dur:bre	2.0	—	1.39	2.0	—	1.38
t:bre	0.6	—	1.37	0.6	—	1.37
n_mora:acc	1.0	-0.10	1.20	0.3	—	1.18
m_mora:acc	0.1	—	1.18	—	—	—
fall:org	1.3	—	1.62	0.3	—	1.12
fall:mod	1.3	—	1.62	0.3	—	1.12
rise	—	—	—	—	—	—
dur:acc	1.8	-0.11	1.40	0.1	-0.03	—
t:acc	0.3	—	1.34	—	—	—
dur:mora	0.1	-0.03	1.03	0.1	-0.03	1.03
t:mora	0.1	—	1.03	0.1	—	1.03
dur:ph	0.1	—	1.44	0.1	—	1.28
t:ph	0.1	—	1.43	0.1	—	1.28

表 4.2　評価データセット$\mathbb{U}_e$の言語特徴量が外れ値を含む時間フレームの割合と，言語特徴量の各属性の正規化された最小値と最大値

学習データセット	$\mathbb{U}_{300}$			$\mathbb{U}_{400}$		
属性名	割合	最小値	最大値	割合	最小値	最大値
n_bre:utt	92.3	—	3.17	92.3	—	3.17
n_acc:utt	92.3	—	2.18	92.3	—	2.18
n_mora:utt	94.0	—	2.75	94.0	—	2.75
b_bre:utt	92.8	—	3.18	92.8	—	3.18
a_bre:utt	71.1	—	2.18	71.1	—	2.18
m_bre:utt	92.0	—	2.83	92.0	—	2.83
a_acc:utt	68.6	—	2.18	68.6	—	2.18
m_acc:utt	90.6	—	2.77	90.6	—	2.77
m_mora:utt	89.9	—	2.73	89.9	—	2.73
dur:utt	94.0	—	3.27	94.0	—	3.27
t:utt	88.4	—	3.20	88.4	—	3.20
n_acc:bre	4.1	—	1.20	4.1	—	1.20
n_mora:bre	1.4	—	1.28	1.4	—	1.28
a_acc:bre	0.9	—	1.20	0.9	—	1.20
m_acc:bre	0.7	—	1.39	0.7	—	1.39
m_mora:bre	0.4	—	1.24	0.4	—	1.24
dur:bre	2.0	—	1.38	2.0	—	1.38
t:bre	0.6	—	1.37	0.6	—	1.37
n_mora:acc	0.3	—	1.18	0.3	—	1.18
m_mora:acc	—	—	—	—	—	—
fall:org	0.3	—	1.12	0.2	—	1.12
fall:mod	—	—	—	—	—	—
rise	—	—	—	—	—	—
dur:acc	0.1	-0.03	—	0.1	-0.03	—
t:acc	—	—	—	—	—	—
dur:mora	0.1	-0.03	1.02	—	—	—
t:mora	0.1	—	1.02	—	—	—
dur:ph	0.1	—	1.28	0.1	—	1.26
t:ph	0.1	—	1.28	0.1	—	1.25

表 4.2　評価データセット$\mathbb{U}_e$の言語特徴量が外れ値を含む時間フレームの割合と，言語特徴量の各属性の正規化された最小値と最大値

学習データセット	$\mathbb{U}_{500}$			$\mathbb{U}_{1000}$		
属性名	割合	最小値	最大値	割合	最小値	最大値
n_bre:utt	92.3	—	3.17	89.0	—	2.71
n_acc:utt	92.3	—	2.18	92.3	—	2.18
n_mora:utt	94.0	—	2.73	94.0	—	2.73
b_bre:utt	92.8	—	3.18	92.8	—	3.18
a_bre:utt	71.1	—	2.18	71.1	—	2.18
m_bre:utt	92.0	—	2.83	92.0	—	2.83
a_acc:utt	68.6	—	2.18	68.6	—	2.18
m_acc:utt	90.6	—	2.77	90.6	—	2.77
m_mora:utt	89.9	—	2.73	89.9	—	2.73
dur:utt	94.0	—	3.25	94.0	—	3.25
t:utt	88.4	—	3.20	88.4	—	3.20
n_acc:bre	4.1	—	1.20	0.7	—	1.17
n_mora:bre	1.4	—	1.28	1.4	—	1.28
a_acc:bre	0.9	—	1.20	—	—	—
m_acc:bre	0.7	—	1.39	0.5	—	1.24
m_mora:bre	0.4	—	1.24	0.4	—	1.24
dur:bre	2.0	—	1.38	2.0	—	1.33
t:bre	0.6	—	1.37	0.5	—	1.31
n_mora:acc	0.3	—	1.18	—	—	—
m_mora:acc	—	—	—	—	—	—
fall:org	0.2	—	1.12	—	—	—
fall:mod	—	—	—	—	—	—
rise	—	—	—	—	—	—
dur:acc	0.1	-0.03	—	0.1	-0.01	—
t:acc	—	—	—	—	—	—
dur:mora	—	—	—	—	—	—
t:mora	—	—	—	—	—	—
dur:ph	0.1	—	1.26	0.1	—	1.20
t:ph	0.1	—	1.25	0.1	—	1.20

表 4.2　評価データセット$\mathbb{U}_e$の言語特徴量が外れ値を含む時間フレームの割合と，言語特徴量の各属性の正規化された最小値と最大値

学習データセット	$\mathbb{U}_{2000}$		
属性名	割合	最小値	最大値
n_bre:utt	75.5	—	2.38
n_acc:utt	92.3	—	2.18
n_mora:utt	94.0	—	2.73
b_bre:utt	92.8	—	3.18
a_bre:utt	71.1	—	2.18
m_bre:utt	92.0	—	2.83
a_acc:utt	68.6	—	2.18
m_acc:utt	90.6	—	2.77
m_mora:utt	89.9	—	2.73
dur:utt	94.0	—	3.23
t:utt	88.4	—	3.20
n_acc:bre	0.7	—	1.17
n_mora:bre	1.4	—	1.16
a_acc:bre	—	—	—
m_acc:bre	0.3	—	1.10
m_mora:bre	0.2	—	1.09
dur:bre	2.0	—	1.26
t:bre	0.4	—	1.25
n_mora:acc	—	—	—
m_mora:acc	—	—	—
fall:org	—	—	—
fall:mod	—	—	—
rise	—	—	—
dur:acc	—	—	—
t:acc	—	—	—
dur:mora	—	—	—
t:mora	—	—	—
dur:ph	0.1	—	1.13
t:ph	0.1	—	1.13

4.5. 各正規化法と基本周波数の予測精度

4.5.1. 聴取実験方法

各正規化法を適用した言語特徴量ベクトル系列から予測した対数基本周波数系列を比較するために，合成音声の聴取実験により主観評価した．提案法の有効性を確認するため，表 4.3 の組み合わせで，合成音声の韻律についての聴取実験を行った．聴取実験の手順を図 4.1 に示す．参加者に音声 A と音声 B を順に聴かせ，音声 A に対する音声 B の韻律についての評価を表 4.4 のカテゴリから選択させた．ただし，評点は参加者には開示していない．各群の合成音声をセッションごとにランダムで音声 A と音声 B に割り当てた．評点は次式に従って集計した．

$$\mathbb{V} = \left\{ v_i \;\middle|\; v_i = \left(\sum_{s=1}^{S} (v_i)_s^{(\mathrm{G}_1)} \right) \div \left(\sum_{s=1}^{S} (v_i)_s^{(\mathrm{G}_1)} + \sum_{s=1}^{S} (v_i)_s^{(\mathrm{G}_2)} \right) \right\} \tag{4.10}$$

ここで，v_iは参加者iの群 1 の音声に対する評点，$(v_i)_s^{(G_1)}$はs回目のセッションにおける参加者iの群 1 の音声の評点，$(v_i)_s^{(G_2)}$はs回目のセッションにおける参加者iの群 2 の音声の評点，Sはセッション数，$\mathbb{V}$はv_iの集合である．

学習データセットは$\mathbb{U}_{100}$，評価データセットは$\mathbb{U}_{\mathrm{e}}$のうち外れ値を含む 96 文を使用した（2.3 を参照）．実験に用いた音声はすべて 2.1.2 のボコーダの合成部で生成した．対数基本周波数を予測するときは，原音声の継続長から算出した時間フレーム情報を付与した言語特徴量を使用した．予測された対数基本周波数と，その対数基本周波数に対応する原音声のスペクトル包絡と非周期性指標を用いて音声波形を生成した．$\mathbb{U}_{\mathrm{e}}$は長文セットから作成したため，合成音声は長くなる．長い合成音声を評価することは難しいため，96 個の合成音声を呼気段落ごとに分割して，232 個の短い合成音声を作成した．

参加者は合成音声の韻律や音質の違いに敏感な 10 名である．合成音声の韻律を評価するため，合成音声のアクセントや抑揚に注目して評価するように指示をした．また，各聴取実験のセッション数は 232 であるため，参加者を適宜休憩させた．

表 4.3　聴取実験における正規化法の比較の組み合わせ

識別名	群 1	群 2
聴取実験 1	提案法	従来法
聴取実験 2	提案法	広範囲版の従来法
聴取実験 3	提案法	クリッピング版の従来法

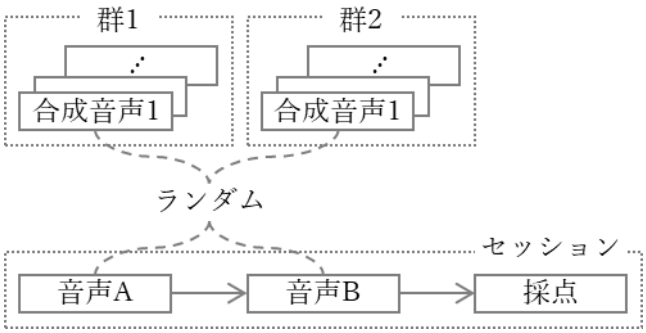

図 4.1　聴取実験の手順

表 4.4　評価カテゴリと評点

カテゴリ	音声 A の評点	音声 B の評点
音声 B は音声 A よりも明らかに良い	3	0
音声 B は音声 A よりも良い	2	0
音声 B は音声 A と同じくらい良いまたは悪い	1	1
音声 B は音声 A よりも悪い	0	2
音声 B は音声 A よりも明らかに良い	0	3

4.5.2. 聴取実験結果

各聴取実験で得られた評点$\mathbb{V}$を図 4.2 に示す．棒グラフは$\mathbb{V}$の平均値，エラーバーは 95%信頼区間である．評点が 0 に近づくほど，群 1 の合成音声は群 2 の合成音声に比べて韻律の品質が悪いことを示し，評点が 1 に近づくほど，群 1 の合成音声は群 2 の合成音声に比べて韻律の品質が良いことを示し，平均評点が 0.5 であれば，群 1 の合成音声は群 2 の合成音声と比べて韻律の品質が同じであることを示す．「平均評点は 0.5 である」という仮説を検定するために，各聴取実験の$\mathbb{V}$の平均値について両側検定の t 検定を行った．その結果，提案法と従来法を比較したときの$\mathbb{V}$の平均値は 0.5 よりも有意に大きく（$t(9) = 52.28, p < 0.001$），提案法と広範囲版の従来法を比較したときの$\mathbb{V}$の平均値は 0.5 よりも有意に大きく（$t(9) = 14.91, p < 0.001$），提案法とクリッピング版の従来法を比較したときの$\mathbb{V}$の平均値は 0.5 よりも有意に大きかった（$t(9) = 11.23, p < 0.001$）．以上より，提案法が従来法，広範囲版の従来法，クリッピング版の従来法よりも優れていることが示された．

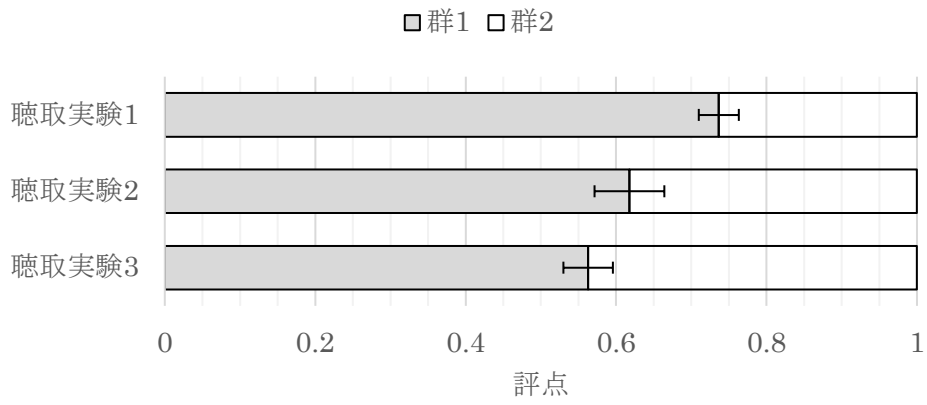

図 4.2　聴取実験の結果

4.5.3. 予測誤差の算出方法

聴取実験の結果を裏付けるために対数基本周波数の予測誤差を計算する．ただし，対数基本周波数の予測誤差と合成音声の品質との因果関係は絶対的なものではないため，対数基本周波数の予測誤差は聴取実験の結果を補足するために用いる．対数基本周波数の予測誤差を次式で定義する．

$$
\begin{aligned}
&\mathbb{T} \in \{\mathbb{U}_{100}, \mathbb{U}_{200}, \mathbb{U}_{300}, \mathbb{U}_{400}, \mathbb{U}_{500}, \mathbb{U}_{1000}, \mathbb{U}_{2000}\} \\
&\mathbb{P} \in \{\mathbb{U}_{\mathrm{c}}, \mathbb{U}_{\mathrm{e}}\} \\
&\boldsymbol{y} \in \mathbb{P} \\
&\boldsymbol{y} = [y_1, \cdots, y_t, \cdots, y_T]^{\top} \\
&\hat{\boldsymbol{y}} = [\hat{y}_1, \cdots, \hat{y}_t, \cdots, \hat{y}_T]^{\top} \\
&\mathbb{E}_{(\mathbb{T},\mathbb{P})} = \left\{ \varepsilon \;\middle|\; \varepsilon = \frac{1}{T} \sum_{t=1}^{T} |y_t - \hat{y}_t| \right\}
\end{aligned}
\tag{4.11}
$$

ここで，$\mathbb{T}$は学習データセット，$\mathbb{P}$は評価データセット，$\boldsymbol{y}$は$\mathbb{P}$に含まれる原音声の対数基本周波数系列，y_tは時間フレームtにおける$\boldsymbol{y}$の要素，$\hat{\boldsymbol{y}}$は$\boldsymbol{y}$に対応する$\mathbb{T}$で学習した FFNN で予測した対数基本周波数系列，$\hat{y}_t$は時間フレームtにおける$\hat{\boldsymbol{y}}$の要素，Tは$\boldsymbol{y}$の時間フレーム数，εは$\boldsymbol{y}$と$\hat{\boldsymbol{y}}$の時間フレームごとの絶対誤差の平均，$\mathbb{E}_{(\mathbb{T},\mathbb{P})}$は$\mathbb{T}$と$\mathbb{P}$についての$\varepsilon$の集合である．

4.5.4. 予測誤差の結果

各正規化法の$\mathbb{E}_{(\mathbb{T},\mathbb{P})}$を図 4.3 と図 4.4 に示す．箱の上端は$\mathbb{E}_{(\mathbb{T},\mathbb{P})}$の第 3 四分位数，箱の中線は$\mathbb{E}_{(\mathbb{T},\mathbb{P})}$の第 2 四分位数（中央値），箱の下端は$\mathbb{E}_{(\mathbb{T},\mathbb{P})}$の第 1 四分位数，上側のひげは$\mathbb{E}_{(\mathbb{T},\mathbb{P})}$の最大値，下側のひげは$\mathbb{E}_{(\mathbb{T},\mathbb{P})}$の最小値である．

図 4.3 については，学習データセットに関わらず，正規化法ごとで$\mathbb{E}_{(\mathbb{T},\mathbb{U}_{\mathrm{c}})}$の中央値に差はほぼなかった．各正規化法の$\mathbb{E}_{(\mathbb{T},\mathbb{U}_{\mathrm{c}})}$の中央値の差が最大のものは，従来法の$\mathbb{E}_{(\mathbb{U}_{100},\mathbb{U}_{\mathrm{c}})}$の中央値と提案法の$\mathbb{E}_{(\mathbb{U}_{100},\mathbb{U}_{\mathrm{c}})}$の中央値の差であり，約 0.006 であった．この差は話者の平均基本周波数の約 270 Hz に対して約 2 Hz の差であり，音声言語においては小さな差である．また，

いずれの正規化法も，学習データ量が増加するにつれて，$\mathbb{E}_{(\mathbb{T},\mathbb{U}_c)}$の中央値は増加した．その増加量は約 0.01 から 0.02 であり，上記と同じ理由で，音声言語においては小さな差である．

一方，図 4.4 については，学習データ量が小さいほど，正規化法ごとのの中央値の差は大きかった．各正規化法の中央値の差が最大のものは，従来法の$\mathbb{E}_{(\mathbb{U}_{100},\mathbb{U}_e)}$の中央値と提案法の$\mathbb{E}_{(\mathbb{U}_{100},\mathbb{U}_e)}$の中央値の差であり，約 0.05 であった．この差は話者の平均基本周波数の約 270 Hz に対して約 13 Hz の差であり，約 1/12 oct.の変化に相当する．この差だけでは従来法による対数基本周波数と提案法による対数基本周波数のどこに違いがあるのかは判断できない．しかし，聴取実験の結果を考慮すると，この差は韻律の知覚に影響を与えるものであるといえる．

各正規化法の$\mathbb{E}_{(\mathbb{T},\mathbb{U}_e)}$の平均値を Tukey-Kramer 法により比較した結果を表 4.5 に示す．$\mathbb{U}_{400}$と$\mathbb{U}_{2000}$の場合を除いて，提案法の$\mathbb{E}_{(\mathbb{T},\mathbb{U}_e)}$の平均値は従来法の$\mathbb{E}_{(\mathbb{T},\mathbb{U}_e)}$の平均値や広範囲版の従来法の$\mathbb{E}_{(\mathbb{T},\mathbb{U}_e)}$の平均値よりも有意に小さく，クリッピング版の従来法の$\mathbb{E}_{(\mathbb{T},\mathbb{U}_e)}$の平均値は従来法の$\mathbb{E}_{(\mathbb{T},\mathbb{U}_e)}$の平均値や広範囲版の従来法の$\mathbb{E}_{(\mathbb{T},\mathbb{U}_e)}$の平均値よりも有意に小さかった．このため，提案法やクリッピング版の従来法を用いることで，対数基本周波数の予測誤差を小さくすることができるといえる．

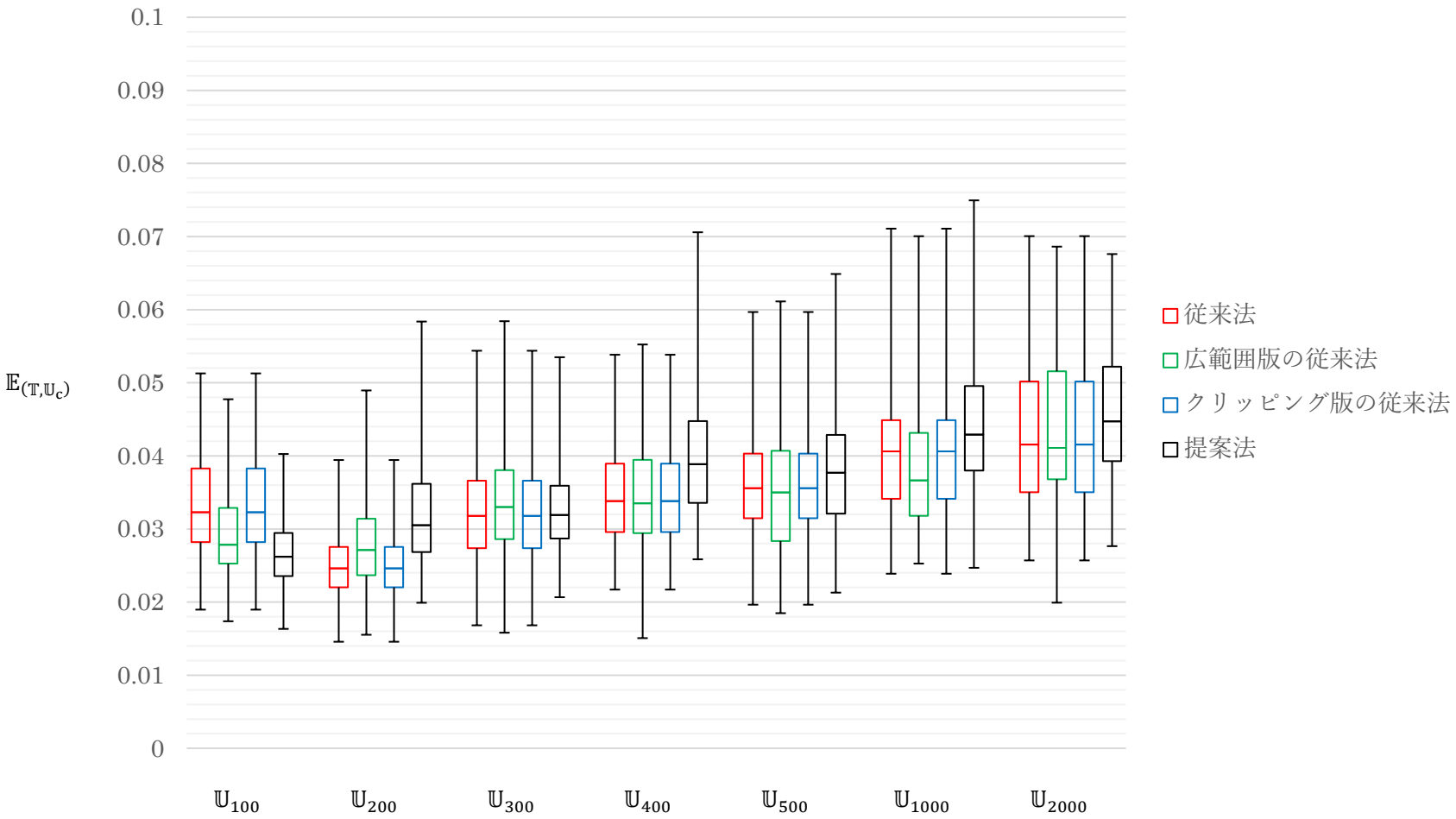

図 4.3　各正規化法における学習データ量と学習内データについての対数基本周波数の予測誤差の関係

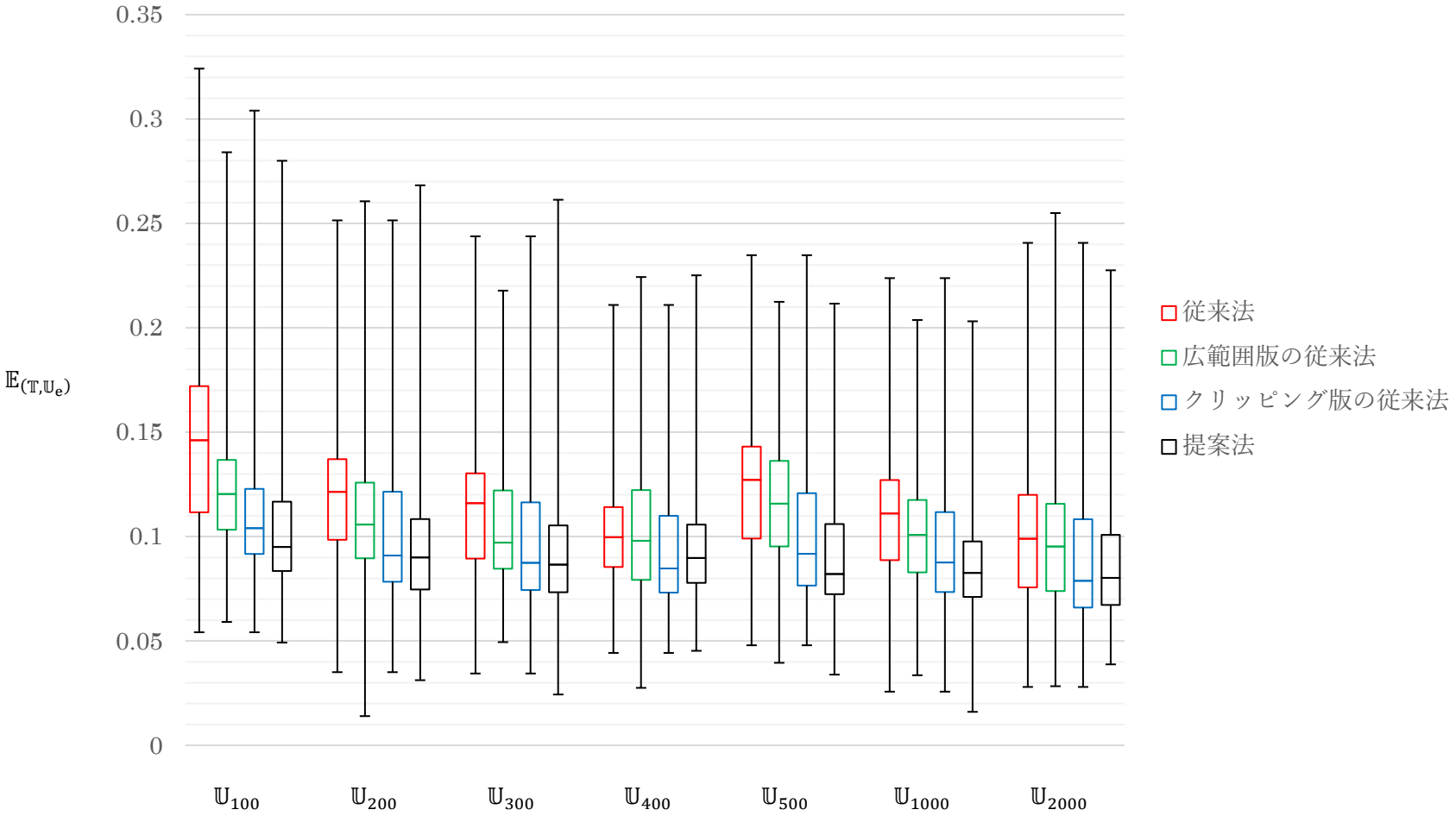

図 4.4　各正規化法における学習データ量と外れ値を含む学習外データについての対数基本周波数の予測誤差の関係

表 4.5　Tukey-Kramer 法による各正規化法の$\mathbb{E}_{(\mathbb{T},\mathbb{U}_e)}$の平均値の比較結果

表中の数値はスチューデント化された範囲分布の q 値と p 値である．群数は 4，自由度は 396，信頼区間は 95%である．「Conv」は従来法，「Extra」は広範囲版の従来法，「Clip」はクリッピング版の従来法，「Prop」は提案法を表す．

		$\mathbb{U}_{100}$	$\mathbb{U}_{200}$	$\mathbb{U}_{300}$	$\mathbb{U}_{400}$	$\mathbb{U}_{500}$	$\mathbb{U}_{1000}$	$\mathbb{U}_{2000}$
Conv – Extra	q 値	4.21	1.39	2.13	0.29	2.15	3.18	0.03
	p 値	0.016	0.732	0.435	0.900	0.429	0.112	0.900
Conv – Clip	q 値	6.98	4.13	3.90	2.64	5.85	4.40	2.91
	p 値	0.001	0.019	0.031	0.245	0.001	0.011	0.170
Conv – Prop	q 値	8.57	4.72	4.92	1.76	8.53	6.63	3.22
	p 値	0.001	0.005	0.003	0.587	0.001	0.001	0.106
Extra – Clip	q 値	2.78	2.74	1.77	2.93	3.71	1.23	2.88
	p 値	0.204	0.215	0.585	0.164	0.045	0.799	0.177
Extra – Prop	q 値	4.36	3.32	2.79	2.05	6.38	3.45	3.19
	p 値	0.012	0.089	0.201	0.468	0.001	0.072	0.111
Clip – Prop	q 値	1.59	0.58	1.02	0.88	2.68	2.23	0.31
	p 値	0.656	0.900	0.880	0.900	0.233	0.396	0.900

4.6.　言語特徴量の各属性と対数基本周波数の関連性

4.6.1.　実験方法

言語特徴量のほとんどの属性値は文章構造により一意に決まるが，実数型の属性のうちアクセントの下降位置であるアクセント型はユーザが自由に設定できる．このため，アクセント型の制御が正規化法の違いにより損なわれてはならない．従来法は，言語特徴量の各属性値のスケールを変化させるだけなので，言語特徴量の各属性は対数基本周波数と直接結びつく．一方で，提案法は，言語特徴量の 2 つの属性値の比をとるため，言語特徴量の各属性は対数基本周波数と直接結びつかない．従来法と提案法で，言語特徴量の各属性と対数基本周波数の関連性に変化があるかを確認する．

局所的に解釈可能なモデルによらない説明法（LIME 法：Local Interpretable Model-agnostic Explanations 法 [35]）により言語特徴量の各属性と対数基本周波数の関連性を確認した．画像識別において，LIME 法は，学習済みのモデルに虫食い画像を入力したときの虫食い領域と識別精度の関係を得ることで，画像のどの部分が識別に重要かを明らかにする．これを応用し，すべての時間フレームにおいて任意の属性の値が 0 の正規化された言語特徴量を学習済みの FFNN に入力したときの対数基本周波数の予測精度を算出することで、言語特徴量の各属性と対数基本周波数の関連性を明らかにする．FFNN は言語特徴量の各属性と対数基本周波数の関係を学習するため，言語特徴量の属性と対数基本周波数の関連性が強いほど，LIME 法による対数基本周波数の予測誤差は大きくなる．

4.6.2. 実験結果

LIME 法による対数基本周波数の予測誤差を図 4.5 と図 4.6 に示す. 右端の「reference」の箱ひげ図は LIME 法を適用しない場合の対数基本周波数の予測誤差である．LIME 法による対数基本周波数の予測誤差が大きいほど，下端の言語特徴量の属性名と対数基本周波数の関連性が強いことを表す．どちらの正規化法も対数基本周波数と言語的に関連する属性が上位を占めた. 学習時に明示的な規則を定義しなくても，モデルパラメータが対数基本周波数と対数基本周波数に関連性の高い言語特徴量の属性を関連付けるように学習されることが確認できた．また，アクセントの下降位置については,「fall:org」よりも「fall:mod」の方が対数基本周波数との関連性が強かった．この結果は，文献 [28]と一致した.

また，従来法と提案法の対数基本周波数との関連性が高い属性の上位 6 つは，どちらも「fall」,「fall:mod」,「fall:mod:cur」,「ph_art」,「m_mora:acc:fwd」,「rise」でり，いずれも共通していた．これらの属性の LIME 法による予測誤差の平均値と「reference」の予測誤差の平均値を両側検定の t 検定した．その結果，従来法も提案法も，上位 6 つの属性の LIME 法による予測誤差の平均値は「reference」の予測誤差の平均値よりも有意に大きかった（表 4.6）．以上より，上位 6 つの属性は対数基本周波数と関連性が高いといえる.

さらに，従来法と提案法で，上位 6 つの属性の LIME による予測誤差の平均値を両側検定のt検定した. その結果,「fall」,「fall:mod」,「fall:mod:cur」,「ph_art」,「m_mora:acc:fwd」の 5 つの属性については，提案法の LIME 法による予測誤差の平均値は，従来法の LIME 法による予測誤差の平均値よりも有意に大きかった（表 4.7）．以上より，提案法で正規化された言語特徴量は従来法で正規化された言語特徴量よりも対数基本周波数との関連性が高く，対数基本周波数をモデル化するのに適しているといえる.

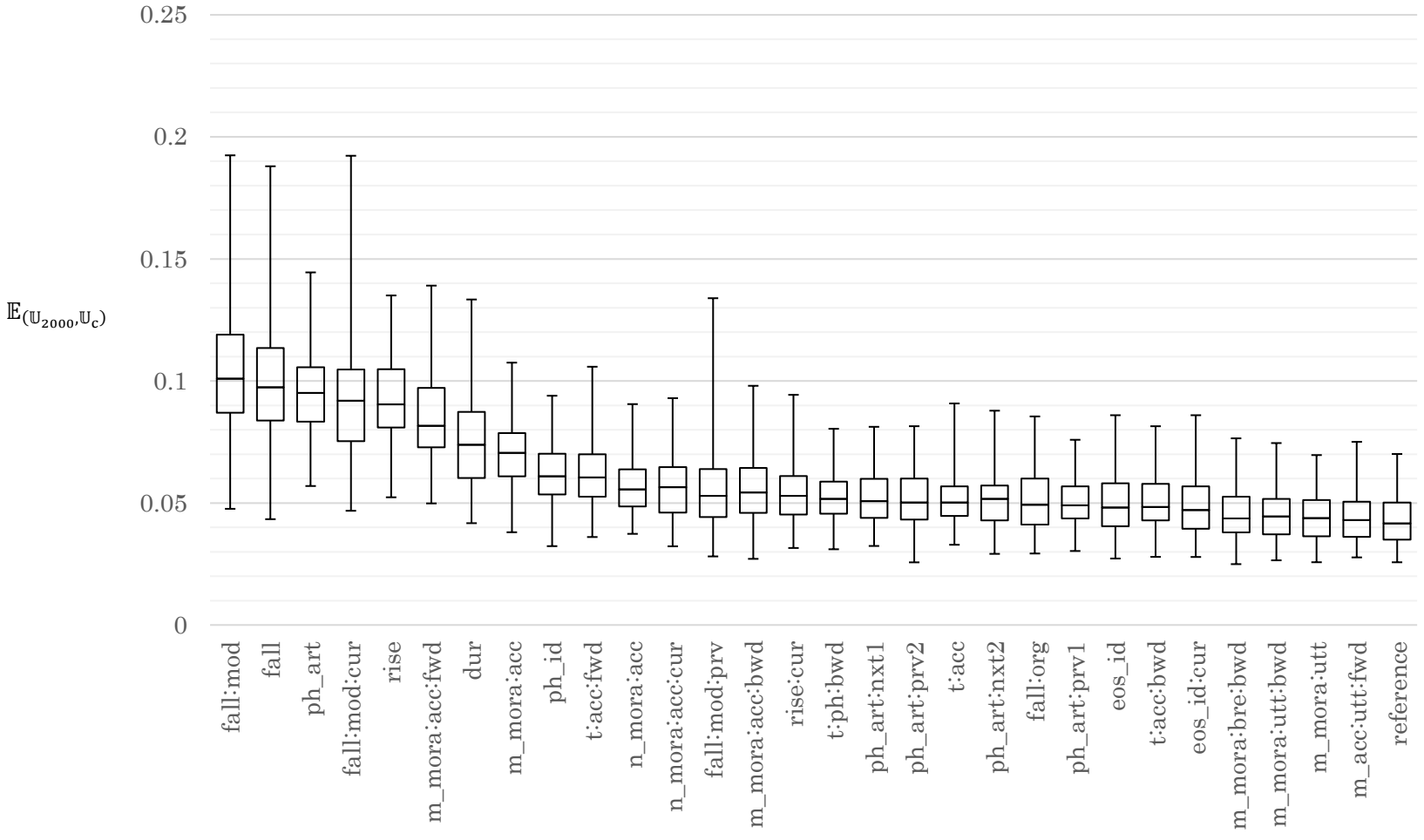

図 4.5　従来法の LIME 法による対数基本周波数の予測誤差

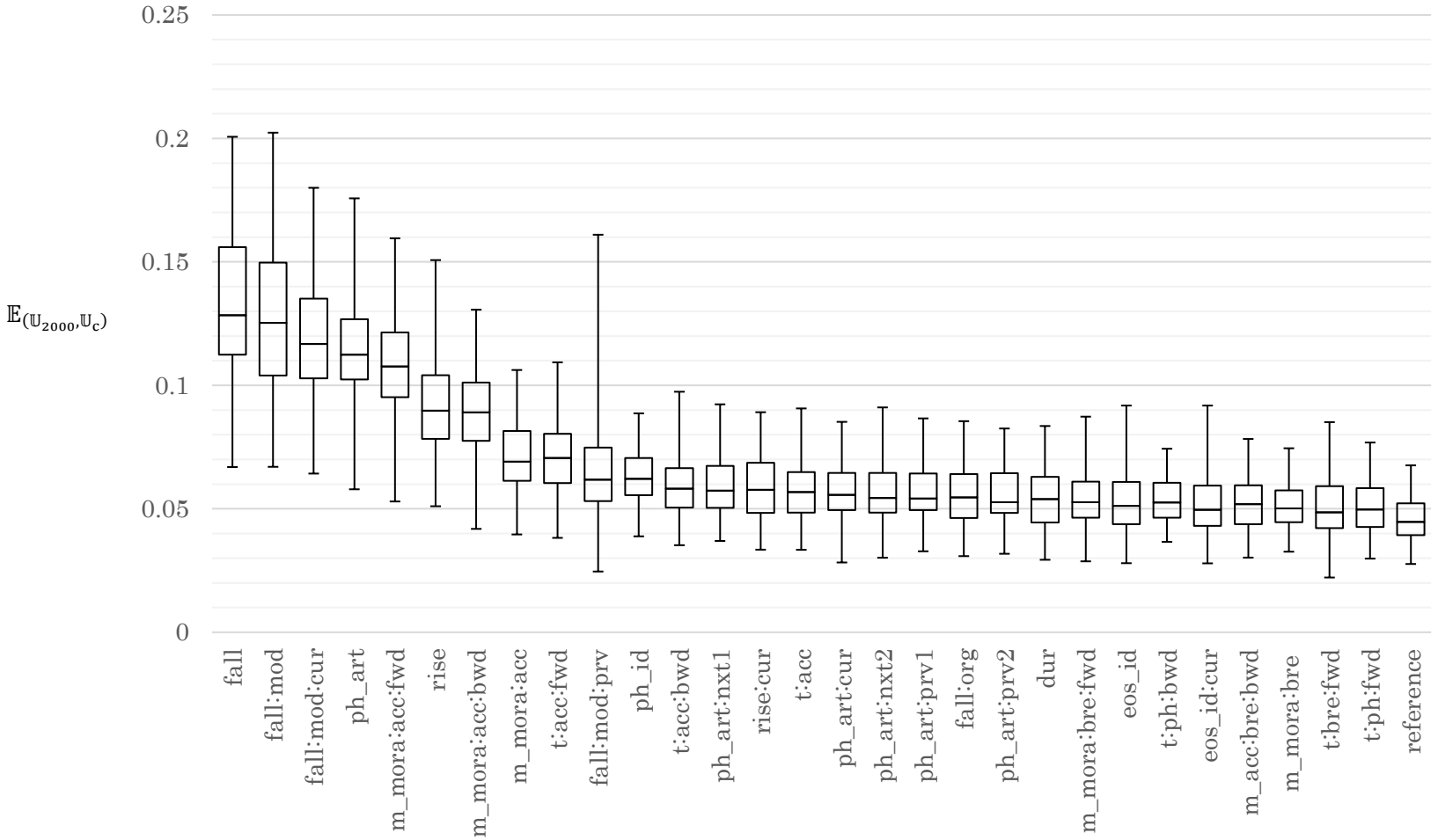

図 4.6　提案法の LIME 法による対数基本周波数の予測誤差

表 4.6　「reference」の予測誤差の平均値と各属性の LIME 法による予測誤差の平均値の両側検定の t 検定による比較

属性名	従来法			提案法		
	自由度	t 値	p 値	自由度	t 値	p 値
fall	127.5	21.02	$p < 0.001$	117.3	28.81	$p < 0.001$
fall:mod	125.1	21.24	$p < 0.001$	116.1	26.04	$p < 0.001$
fall:mod:cur	129.0	18.97	$p < 0.001$	126.4	27.71	$p < 0.001$
ph_art	150.7	24.88	$p < 0.001$	135.1	29.64	$p < 0.001$
m_mora:acc:fwd	156.8	20.76	$p < 0.001$	143.9	29.75	$p < 0.001$
rise	154.0	24.53	$p < 0.001$	143.2	21.42	$p < 0.001$

表 4.7　従来法と提案法の各属性の LIME 法による予測誤差の平均値の両側検定の t 検定による比較

属性名	自由度	t 値	p 値
fall	189.6	9.13	$p < 0.001$
fall:mod	190.2	6.56	$p < 0.001$
fall:mod:cur	194.0	7.34	$p < 0.001$
ph_art	190.4	7.11	$p < 0.001$
m_mora:acc:fwd	192.5	9.70	$p < 0.001$
rise	193.1	0.46	0.648

4.7. 考察

2 つの言語特徴量の属性値の比を取る正規化法を提案し，Min-Max 正規化法に基づく 3 つの従来法を比較対象として，聴取実験と予測誤差により評価した．従来法は最も予測誤差が大きく，聴取実験においても韻律が訛る問題が頻発した．これは，従来法は言語特徴量の属性値のスケールを変換するだけであり，学習した言語特徴量と文章構造が異なる文字列から作成された言語特徴量に含まれる外れ値には対応できないためである．

広範囲版の従来法は，外れ値を考慮して最小値と最大値の範囲を広くしたものである．聴取実験や予測誤差から判断すると，正規化の範囲を広くすることには一定の効果があるといえる．しかし，この正規化法でも，外れ値の問題は根本的には解決されないため，従来法と同様に韻律が訛る問題が残った．

クリッピング版の従来法は，外れ値を学習データセットから算出した最小値と最大値に丸め込むことで，DNN に外れ値を入力しないようにする．言語特徴量の属性値をクリッピングすることにより，言語特徴量の属性値が制限される問題は本実験では確認できなかった．これは，表 4.2 と図 4.5 から判断すると，対数基本周波数と関連性が高い言語特徴量の属性のほとんどが学習データセットから算出される最小値から最大値の範囲内にあるた

めである．このため，従来法や広範囲版の従来法よりも予測誤差が小さかった．また，表 4.2 と図 4.5 から判断すると，対数基本周波数と関連性が低い発話階層の言語特徴量の属性を使用せず，学習データ量を増やせば，対数基本周波数の予測についてはクリッピング版の従来法でも問題ないといえる．しかし，クリッピング版の従来法は，外れ値を最小値や最大値に丸め込むため，最小値や最大値に対応する対数基本周波数しか予測されない．このため，クリッピング版の従来法は外れ値の問題を根本的に解決していない．

提案法は，言語特徴量の階層構造に着目し，2 つの言語特徴量の属性値の比を取る．このため，いかなる文字列から抽出された言語特徴量であっても，正規化後の値は必ず 0 から 1 となる．このため，対数基本周波数の予測誤差も聴取実験の評点も他の正規化法よりも優れていた．このような結果となった要因は，提案法により正規化された言語特徴量が外れ値を含まなくなったことに加え，特にアクセント型を表すアクセントの下降位置の言語特徴量の値が基本周波数パターンを学習するのに適したものになったことにあると考える．例えば，4 モーラ 2 型のアクセント句と 6 モーラ 3 型のアクセント句では，基本周波数パターンはアクセント句の中央付近で下降する．2 つのアクセント句のアクセント下降位置の属性値は異なるが，基本周波数パターンは類似している．Min-Max 正規化法に基づく従来法では，属性値ごとに基本周波数パターンを学習するため，多くの言語特徴量と基本周波数パターンの組み合わせが必要である．これに対して，提案法では，どちらのアクセント句もアクセントの下降位置の属性値は 0.5 となり，アクセント句の中央付近で下降するという基本周波数パターンを表現するのに適した値となる．これにより，少量の学習データでも言語特徴量と基本周波数パターンを効率的に学習できる．

4.8. まとめ

学習データと構造が異なる文章に対しても頑健な音声特徴量の予測を可能にするために，2 つの言語特徴量の属性値の比を取る正規化法を提案した．文章構造と密接に関係する基本周波数について，合成音声の韻律を評価する聴取実験と対数基本周波数の予測誤差により提案法と従来法を比較した．その結果，提案法は少ない学習データ量においても頑健な対数基本周波数の予測を可能にした．また，提案法は言語特徴量の各属性と対数基本周波数の関連性を保ったまま，外れ値が発生しないように言語特徴量を正規化できることがわかった．

5. 時系列の複数の属性を考慮した損失関数による FFNN の学習法

5.1. はじめに

3 章では，後処理を用いない、FFNN のみで構成された音声特徴量予測部が最も高速であることを明らかにした．しかし，FFNN は RNN のように再帰構造を持たないため，一般的な学習法では隣接する時間フレーム間の音声特徴量の関係を考慮したモデルパラメータを獲得できない．また，後処理で MLPG を利用しないため，音声特徴量の動的特徴量を考慮して音声特徴量を生成できない．そこで，本章では，音声特徴量予測部に用いる FFNN が音声特徴量の時間構造を考慮したモデルパラメータを獲得できるようにするため，時系列の複数の属性を考慮した損失関数による学習法を提案する．音声合成の韻律と音質に関する音声特徴量である対数基本周波数とメルケプストラムを対象として，提案法と従来法を比較することで，提案法の有効性を示す．

5.2. 従来の損失関数

基本的な音声特徴量予測部の構成である 3.2.1 と 3.2.2 の学習で用いる 3 つの損失関数について述べる 1 つめは 3.2.2 についての損失関数で，音声特徴量の平均二乗誤差を計算する．2 つめは 3.2.1 についての損失関数で，音声特徴量の動的特徴量の平均二乗誤差を計算する．3 つめは 3.2.1 についての損失関数で，音声特徴量の動的特徴量から MLPG を介して生成した音声特徴量の平均二乗誤差を計算する．

5.2.1. 音声特徴量の平均二乗誤差

この損失関数は最も基本的なものであり，音声特徴量の平均二乗誤差（MSE：Mean Squared Error）を計算する．損失関数に入力される教師データの音声特徴量を次式で定義する．

$$\begin{aligned} \boldsymbol{y} &= [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_t^\top, \cdots, \boldsymbol{y}_T^\top]^\top \\ \boldsymbol{y}_t &= \left[y_t^{(1)}, \cdots, y_t^{(d)}, \cdots, y_t^{(D)}\right] \end{aligned} \tag{5.1}$$

ここで，$\boldsymbol{y}$は教師データとしての音声特徴量ベクトル系列，$\boldsymbol{y}_t$は教師データとしての時間フレームtにおける音声特徴量ベクトル，$y_t^{(d)}$は教師データとしての時間フレームtにおける次元dの音声特徴量，tは時間フレームインデックス，Tは時間フレーム数，dは次元インデックス，Dは次元数である．また，$\boldsymbol{y}$に対応する DNN で予測された音声特徴量であり，損失関数に入力される予測データの音声特徴量を次式で定義する．

$$\begin{aligned} \hat{\boldsymbol{y}} &= [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_t^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top \\ \hat{\boldsymbol{y}}_t &= \left[\hat{y}_t^{(1)}, \cdots, \hat{y}_t^{(d)}, \cdots, \hat{y}_t^{(D)}\right] \end{aligned} \tag{5.2}$$

ここで，$\hat{\boldsymbol{y}}$は予測データとしての音声特徴量ベクトル系列，$\hat{\boldsymbol{y}}_t$は予測データとしての時間フレームtにおける音声特徴量ベクトル，$\hat{y}_t^{(d)}$は予測データとしての時間フレームtにおける次元dの音声特徴量である．$\boldsymbol{y}$と$\hat{\boldsymbol{y}}$の平均二乗誤差は次式となる．

$$
\begin{aligned}
(e_{\mathrm{MSE}})_t^{(d)} &= \left(y_t^{(d)} - \hat{y}_t^{(d)}\right)^2 \\
\mathcal{L}_{\mathrm{MSE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= \frac{1}{TD}\sum_{t=1}^{T}\sum_{d=1}^{D}(e_{\mathrm{MSE}})_t^{(d)}
\end{aligned}
\tag{5.3}
$$

ここで，$(e_{\mathrm{MSE}})_t^{(d)}$は$\hat{y}_t^{(d)}$の$y_t^{(d)}$に対する二乗誤差である．この損失関数が算出した誤差に基づいて，勾配法が DNN のモデルパラメータを更新すると，$\hat{y}_t^{(d)}$に関連する DNN のモデルパラメータは，$(e_{\mathrm{MSE}})_t^{(d)}$のみに基づいて学習される．このため，この損失関数だけでは，DNN のモデルパラメータは$y_t^{(d)}$と$y_{t+1}^{(d)}$の関係性も，$y_t^{(d)}$と$y_t^{(d+1)}$の関係性も捉えることはできず，$y_t^{(d)}$を独立してモデル化してしまう．ただし，この損失関数と RNN の組み合わせるにおいては，RNN の再帰構造により，DNN のモデルパラメータは$\boldsymbol{y}$の時間構造を暗黙的に学習することができる．

5.2.2. 音声特徴量の動的特徴量の平均二乗誤差

この損失関数は 5.2.1 の損失関数と本質的に同じであり，音声特徴量の動的特徴量の平均二乗誤差を計算する．損失関数に入力される教師データの音声特徴量の動的特徴量を次式で定義する．

$$
\begin{aligned}
\boldsymbol{\mu} &= [\boldsymbol{\mu}_1^\top, \cdots, \boldsymbol{\mu}_t^\top, \cdots, \boldsymbol{\mu}_T^\top]^\top \\
\boldsymbol{\mu}_t &= \left[\boldsymbol{\mu}_t^{(0)}, \boldsymbol{\mu}_t^{(1)}, \boldsymbol{\mu}_t^{(2)}\right] \\
\boldsymbol{\mu}_t^{(n)} &= \left[\mu_t^{(n,1)}, \cdots, \mu_t^{(n,d)}, \cdots, \mu_t^{(n,D)}\right] \quad (n = 0, 1, 2)
\end{aligned}
\tag{5.4}
$$

ここで，$\boldsymbol{\mu}$は教師データとしての音声特徴量の動的特徴量ベクトル系列，$\boldsymbol{\mu}_t$は教師データとしての時間フレームtにおける音声特徴量の動的特徴量ベクトル，$\boldsymbol{\mu}_t^{(n)}$は教師データとしての時間フレームtにおける音声特徴量のn次の動的特徴量ベクトル，$\mu_t^{(n,d)}$は教師データとしての時間フレームtにおける次元dの音声特徴量のn次の動的特徴量である．また，$\boldsymbol{\mu}$に対応する DNN で予測された音声特徴量の動的特徴量であり，損失関数に入力される予測データの音声特徴量の動的特徴量を次式で定義する．

$$
\begin{aligned}
\hat{\boldsymbol{\mu}} &= [\hat{\boldsymbol{\mu}}_1^\top, \cdots, \hat{\boldsymbol{\mu}}_t^\top, \cdots, \hat{\boldsymbol{\mu}}_T^\top]^\top \\
\hat{\boldsymbol{\mu}}_t &= \left[\hat{\boldsymbol{\mu}}_t^{(0)}, \hat{\boldsymbol{\mu}}_t^{(1)}, \hat{\boldsymbol{\mu}}_t^{(2)}\right] \\
\hat{\boldsymbol{\mu}}_t^{(n)} &= \left[\hat{\mu}_t^{(n,1)}, \cdots, \hat{\mu}_t^{(n,d)}, \cdots, \hat{\mu}_t^{(n,D)}\right] \quad (n = 0, 1, 2)
\end{aligned}
\tag{5.5}
$$

ここで，$\hat{\boldsymbol{\mu}}$は予測データとしての音声特徴量の動的特徴量ベクトル系列，$\hat{\boldsymbol{\mu}}_t$は予測データとしての時間フレームtにおける音声特徴量の動的特徴量ベクトル，$\hat{\boldsymbol{\mu}}_t^{(n)}$は予測データとしての時間フレーム$t$における音声特徴量の$n$次の動的特徴量ベクトル，$\hat{\mu}_t^{(n,d)}$は予測データとしての時間フレーム$t$における次元$d$の音声特徴量の$n$次の動的特徴量である．$\boldsymbol{\mu}$と$\hat{\boldsymbol{\mu}}$の平均二乗誤差は次式となる．

$$
\begin{aligned}
(e_{\mathrm{MSE}})_t^{(n,d)} &= \left(\mu_t^{(n,d)} - \hat{\mu}_t^{(n,d)}\right)^2 \\
\mathcal{L}_{\mathrm{MSE}}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) &= \frac{1}{3TD}\sum_{t=1}^{T}\sum_{d=1}^{D}\sum_{n=0}^{2}(e_{\mathrm{MSE}})_t^{(n,d)}
\end{aligned}
\tag{5.6}
$$

ここで，$(e_{\mathrm{MSE}})_t^{(n,d)}$は$\hat{\mu}_t^{(n,d)}$の$\mu_t^{(n,d)}$に対する二乗誤差である．この損失関数が算出した誤差に基づいて，勾配法が DNN のモデルパラメータを更新すると，$\hat{\mu}_t^{(n,d)}$に関連する DNN のモデルパラメータは，$(e_{\mathrm{MSE}})_t^{(n,d)}$のみに基づいて学習される．このため，この損失関数だけでは，DNN のモデルパラメータは，$\mu_t^{(n,d)}$と$\mu_{t+1}^{(n,d)}$の関係性も，$\mu_t^{(n,d)}$と$\mu_t^{(n+1,d)}$の関係性も捉えることはできず，$\mu_t^{(n,d)}$を独立してモデル化する．しかし，予測時には，DNN が予測した$\hat{\mu}_t^{(n,d)}$に MLPG を適用するため，$\hat{\mu}_t^{(n,d)}$に基づいた音声特徴量が生成される．

5.2.3. 最小生成誤差法

この損失関数は DNN で予測された音声特徴量の動的特徴量から MLPG を介して生成した音声特徴量と教師データの音声特徴量の平均二乗誤差を計算することにより，$\mu_t^{(n,d)}$が独立してモデル化される 5.2.2 の損失関数の問題を解決する [36]．この学習法を最小生成誤差法（MGE 法：Minimum Generation Error 法）と呼ぶ．音声特徴量の動的特徴量から MLPG を介して生成した音声特徴量を次式で定義する．

$$
\begin{aligned}
\hat{\boldsymbol{\psi}} &= \mathrm{MLPG}(\hat{\boldsymbol{\mu}}, \boldsymbol{U}^{-1}, \boldsymbol{W}) \\
&= \left[\hat{\boldsymbol{\psi}}_1^{\top}, \cdots, \hat{\boldsymbol{\psi}}_t^{\top}, \cdots, \hat{\boldsymbol{\psi}}_T^{\top}\right]^{\top} \\
\hat{\boldsymbol{\psi}}_t &= \left[\hat{\psi}_t^{(1)}, \cdots, \hat{\psi}_t^{(d)}, \cdots, \hat{\psi}_t^{(D)}\right]
\end{aligned}
\tag{5.7}
$$

ここで，$\hat{\boldsymbol{\mu}}$は式（5.5）の予測データとしての音声特徴量の動的特徴量ベクトル系列，$\boldsymbol{U}^{-1}$は式（2.7）の音声特徴量の動的特徴量の分散の逆数の対角行列，$\boldsymbol{W}$は式（2.6）の動的特徴量を算出するための係数行列，$\hat{\boldsymbol{\psi}}$は予測データとしての MLPG で生成した音声特徴量ベクトル系列，$\hat{\boldsymbol{\psi}}_t$は予測データとしての時間フレーム$t$における音声特徴量ベクトル，$\hat{\psi}_t^{(d)}$は予測データとしての時間フレーム$t$における次元$d$の音声特徴量である．$\boldsymbol{y}$と$\hat{\boldsymbol{\psi}}$の平均二乗誤差は次式となる．

$$
\begin{aligned}
(e_{\mathrm{MGE}})_t^{(d)} &= \left(y_t^{(d)} - \hat{\psi}_t^{(d)}\right)^2 \\
\mathcal{L}_{\mathrm{MGE}}(\boldsymbol{y}, \hat{\boldsymbol{\psi}}) &= \frac{1}{TD}\sum_{t=1}^{T}\sum_{d=1}^{D}(e_{\mathrm{MGE}})_t^{(d)}
\end{aligned}
\tag{5.8}
$$

ここで，$(e_{\mathrm{MGE}})_t^{(d)}$は$\hat{\psi}_t^{(d)}$の$y_t^{(d)}$に対する二乗誤差である．時間フレーム$t$の周辺の複数の時間フレームを$t+\tau$で表す．ここで，$\tau = \{\cdots, -1, 0, 1, \cdots\}$であり，$\tau$の有効範囲は$\boldsymbol{U}^{-1}$に依る．$\hat{\psi}_t^{(d)}$は$\hat{\mu}_t^{(n,d)}$だけでなく，$\hat{\mu}_{t+\tau}^{(n,d)}$も考慮されて生成される．このため，$(e_{\mathrm{MGE}})_t^{(d)}$は$\hat{\mu}_{t+\tau}^{(n,d)}$に関連する DNN のモデルパラメータの学習に寄与する．このようにすることで，隣接する時間フレーム間の音声特徴量の動的特徴量の関係を学習できる．ただし，この学習法でも，予測時には MLPG が必要である．

5.3. 提案する損失関数

従来の損失関数は，教師データの音声特徴量と予測データの音声特徴量の平均二乗誤差のみに基づいて DNN のモデルパラメータを学習する．ひとつの誤差基準では教師データの音声特徴量の構造を捉えられないため，RNN や MLPG が必要となる．一方で，3.2.3 に示す FFNN のみの構成では，RNN も MLPG を利用できないため，FFNN 自体が教師データの音声特徴量の構造を捉える必要がある．そこで，提案する損失関数では，複数の誤差基準により，教師データの音声特徴量の構造を多角的に捉えるようにする．提案する損失関数を時系列の複数の属性の損失関数（MATS 損失関数：Multiple Attributes of Temporal Sequence 損失関数）と命名した．MATS 損失関数は，直結型（DC：Direct Coupling）の損失関数, 時間領域（TD：Time Domain）の損失関数, 次元領域（DD：Dimensional Domain）の損失関数，局所内分散（LV：Local Variance）の損失関数 [37]，局所内共分散（LC：Local Covariance）の損失関数，系列内分散（GV：Global Variance）の損失関数 [38]，系列内共分散（GC：Global Covariance）の損失関数で算出される誤差の重み付き和で定義される．

$$\mathcal{L}_{\mathrm{MATS}}(\boldsymbol{y},\hat{\boldsymbol{y}}) = \sum_{l} \omega_l \mathcal{L}_l(\boldsymbol{y},\hat{\boldsymbol{y}}) \qquad (5.9)$$
$$l = \mathrm{DC},\ \mathrm{TD},\ \mathrm{DD},\ \mathrm{LV},\ \mathrm{LC},\ \mathrm{GV},\ \mathrm{GC}$$

ここで，lは損失関数の識別子，ω_lは識別子の損失関数の重み，$\mathcal{L}_l$は識別子の損失関数である．

5.3.1. 直結型の損失関数

DC 損失関数は式（5.3）と同じであり，DNN に音声特徴量の概形を教える．また，DC 損失関数は，TD 損失関数の限定版であり，音声特徴量の 0 次の動的特徴量の平均二乗誤差のみを考慮する．音声特徴量の 1 次以上の動的特徴量を考慮する場合は，TD 損失関数を使用する．MATS 損失関数においては，DC 損失関数で音声特徴量の 0 次の動的特徴量の平均二乗誤差を計算し，TD 損失関数で音声特徴量の 1 次以上の動的特徴量の平均二乗誤差を計算するような使用法は基本的に禁止である．音声特徴量の 1 次以上の動的特徴量を考慮するか，しないかを明確にするために DC 損失関数と TD 損失関数を別々に定義した．MATS 損失関数においては，DC 損失関数か TD 損失関数が音声特徴量を学習する上で基礎の損失関数となる．そのため，まず，DC 損失関数か TD 損失関数をどちらかを決めてから，他の損失関数を組み合わせる．

5.3.2. 時間領域の損失関数

TD 損失関数は，隣接する時間フレーム間の音声特徴量の関係を表す特徴量である TD 特徴量の誤差を計算することによって，DNN に隣接するフレーム間の音声特徴量の関係を教える．TD 損失関数で教師データとしての音声特徴量から算出される TD 特徴量を次式で定

義する．

$$
\begin{aligned}
\boldsymbol{y}_{\mathrm{TD}} &= [(\boldsymbol{y}_{\mathrm{TD}})_1^{\top}, \cdots, (\boldsymbol{y}_{\mathrm{TD}})_t^{\top}, \cdots, (\boldsymbol{y}_{\mathrm{TD}})_T^{\top}]^{\top} \\
(\boldsymbol{y}_{\mathrm{TD}})_t &= \left[(\boldsymbol{y}_{\mathrm{TD}})_t^{(1)}, \cdots, (\boldsymbol{y}_{\mathrm{TD}})_t^{(n)}, \cdots, (\boldsymbol{y}_{\mathrm{TD}})_t^{(N)}\right] \\
(\boldsymbol{y}_{\mathrm{TD}})_t^{(n)} &= \left[(y_{\mathrm{TD}})_t^{(n,1)}, \cdots, (y_{\mathrm{TD}})_t^{(n,d)}, \cdots, (y_{\mathrm{TD}})_t^{(n,D)}\right] \\
(y_{\mathrm{TD}})_t^{(n,d)} &= \sum_{\tau=L_{\mathrm{TD}}}^{R_{\mathrm{TD}}} y_t^{(d)} (w_{\mathrm{TD}})_\tau^{(n)}
\end{aligned}
\tag{5.10}
$$

ここで，$\boldsymbol{y}_{\mathrm{TD}}$は教師データとしての TD 特徴量ベクトル系列，$(\boldsymbol{y}_{\mathrm{TD}})_t$は教師データとしての時間フレーム$t$における TD 特徴量ベクトル，$(\boldsymbol{y}_{\mathrm{TD}})_t^{(n)}$は教師データとしての時間フレーム$t$における$n$次の TD 特徴量ベクトル，$(y_{\mathrm{TD}})_t^{(n,d)}$は教師データとしての次元$d$の音声特徴量についての時間フレーム$t$における$n$次の TD 特徴量，$N$は TD 特徴量の次元数，$(w_{\mathrm{TD}})_\tau^{(n)}$は相対時間フレーム$\tau$における$n$次の TD 特徴量を求める係数，$L_{\mathrm{TD}}$は後方参照時間フレーム数，$R_{\mathrm{TD}}$は前方参照時間フレーム数である．$L_{\mathrm{TD}}$は 0 以下の値であり，$R_{\mathrm{TD}}$は 0 以上の値である．また，$\boldsymbol{y}_{\mathrm{TD}}$と同様に，TD 損失関数で予測データとしての音声特徴量から算出される TD 特徴量を次式で定義する．

$$
\begin{aligned}
\hat{\boldsymbol{y}}_{\mathrm{TD}} &= [(\hat{\boldsymbol{y}}_{\mathrm{TD}})_1^{\top}, \cdots, (\hat{\boldsymbol{y}}_{\mathrm{TD}})_t^{\top}, \cdots, (\hat{\boldsymbol{y}}_{\mathrm{TD}})_T^{\top}]^{\top} \\
(\hat{\boldsymbol{y}}_{\mathrm{TD}})_t &= \left[(\hat{\boldsymbol{y}}_{\mathrm{TD}})_t^{(1)}, \cdots, (\hat{\boldsymbol{y}}_{\mathrm{TD}})_t^{(n)}, \cdots, (\hat{\boldsymbol{y}}_{\mathrm{TD}})_t^{(N)}\right] \\
(\hat{\boldsymbol{y}}_{\mathrm{TD}})_t^{(n)} &= \left[(\hat{y}_{\mathrm{TD}})_t^{(n,1)}, \cdots, (\hat{y}_{\mathrm{TD}})_t^{(n,d)}, \cdots, (\hat{y}_{\mathrm{TD}})_t^{(n,D)}\right] \\
(\hat{y}_{\mathrm{TD}})_t^{(n,d)} &= \sum_{\tau=L_{\mathrm{TD}}}^{R_{\mathrm{TD}}} \hat{y}_{t+\tau}^{(d)} (w_{\mathrm{TD}})_\tau^{(n)}
\end{aligned}
\tag{5.11}
$$

ここで，$\hat{\boldsymbol{y}}_{\mathrm{TD}}$は教師データとしての TD 特徴量ベクトル系列，$(\hat{\boldsymbol{y}}_{\mathrm{TD}})_t$は教師データとしての時間フレーム$t$における TD 特徴量ベクトル，$(\hat{\boldsymbol{y}}_{\mathrm{TD}})_t^{(n)}$は教師データとしての時間フレーム$t$における$n$次の TD 特徴量ベクトル，$(\hat{y}_{\mathrm{TD}})_t^{(n,d)}$は教師データとしての次元$d$の音声特徴量についての時間フレーム$t$における$n$次の TD 特徴量である．TD 損失関数は$\boldsymbol{y}_{\mathrm{TD}}$と$\hat{\boldsymbol{y}}_{\mathrm{TD}}$の平均二乗誤差で定義される．

$$
\begin{aligned}
(e_{\mathrm{TD}})_t^{(n,d)} &= \left((y_{\mathrm{TD}})_t^{(n,d)} - (\hat{y}_{\mathrm{TD}})_t^{(n,d)}\right)^2 \\
\mathcal{L}_{\mathrm{TD}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= \frac{1}{TDN} \sum_{t=1}^{T} \sum_{d=1}^{D} \sum_{n=1}^{N} (e_{\mathrm{TD}})_t^{(n,d)}
\end{aligned}
\tag{5.12}
$$

ここで，$(e_{\mathrm{TD}})_t^{(n,d)}$は$(\hat{y}_{\mathrm{TD}})_t^{(n,d)}$の$(y_{\mathrm{TD}})_t^{(n,d)}$に対する二乗誤差である．$(\hat{y}_{\mathrm{TD}})_t^{(n,d)}$は時間フレーム$t + L_{\mathrm{TD}}$から$t + R_{\mathrm{TD}}$までの$\hat{y}_{t+\tau}^{(d)}$から算出されるため，$(e_{\mathrm{TD}})_t^{(n,d)}$は$\hat{y}_{t+\tau}^{(d)}$に関連する DNN のモデルパラメータの学習に寄与する．このようにすることで，隣接する時間フレーム間の音声特徴量の関係を学習できる．

$(w_{\mathrm{TD}})_\tau^{(n)}$については，対象とする音声特徴量ごとに適した変換式や，経験則による知見に基づいて，各時間フレームの音声特徴量を関係づけることが好ましい．MLPG で用いる動

的特徴量と同様に，$L_{\mathrm{TD}}=-1$，$R_{\mathrm{TD}}=1$として，$(w_{\mathrm{TD}})_{\tau}^{(n)}$を式（2.6）と同じ値にしても良いが，本章では，RNN の再帰構造を模擬するように，L_{TD}，R_{TD}，$(w_{\mathrm{TD}})_{\tau}^{(n)}$を以下のように設定した．

$$\begin{aligned}
&L_{\mathrm{TD}}=-1\\
&R_{\mathrm{TD}}=0\\
&(w_{\mathrm{TD}})_{\tau}^{(1)}=\begin{cases}0 & (\tau=-1)\\ w_1 & (\tau=0)\end{cases}\\
&(w_{\mathrm{TD}})_{\tau}^{(2)}=\begin{cases}-w_2 & (\tau=-1)\\ w_2 & (\tau=0)\end{cases}
\end{aligned} \tag{5.13}$$

これらの値において，$(e_{\mathrm{TD}})_t^{(n,d)}$が 0 と仮定した場合，$\hat{y}_t^{(d)}$について式を整理すると次式の漸化式となる．

$$\hat{y}_t^{(d)}=y_t^{(d)}-\frac{w_1}{w_1+w_2}y_{t-1}^{(d)}+\frac{w_1}{w_1+w_2}\hat{y}_{t-1}^{(d)} \tag{5.14}$$

TD 損失関数は，この式を考慮するため，RNN と同様に再帰的な学習を可能にする．さらに，w_1の値やw_2の値を調整することで，再帰の強さを制御することができる．例えば，w_2の値をw_1の値よりも大きくすることで，$\hat{y}_t^{(d)}$が$\hat{y}_{t-1}^{(d)}$と$y_t^{(d)}-y_t^{(d)}$から学習されるようにできる．特に，日本語のアクセント知覚は基本周波数の相対的な変化に深く関係しているため，w_2の値をw_1の値よりも大きく設定することで，このような知見に基づいた学習を可能にする．

5.3.3. 次元領域の損失関数

DD 損失関数は，メルケプストラムのような多次元の音声特徴量に対して利用する損失関数であり，隣接する次元間の音声特徴量の関係を表す特徴量である DD 特徴量の誤差を計算することによって，DNN に隣接する次元間の音声特徴量の関係を教える．DD 損失関数で教師データとしての音声特徴量から算出される DD 特徴量を次式で定義する．

$$\begin{aligned}
\boldsymbol{y}_{\mathrm{DD}} &= [(\boldsymbol{y}_{\mathrm{DD}})_1^{\top},\cdots,(\boldsymbol{y}_{\mathrm{DD}})_t^{\top},\cdots,(\boldsymbol{y}_{\mathrm{DD}})_T^{\top}]^{\top}\\
(\boldsymbol{y}_{\mathrm{DD}})_t &= \left[(y_{\mathrm{DD}})_t^{(1)},\cdots,(y_{\mathrm{DD}})_t^{(m)},\cdots,(y_{\mathrm{DD}})_t^{(M)}\right]\\
(y_{\mathrm{DD}})_t^{(m)} &= \sum_{d=1}^{D} y_t^{(d)}(w_{\mathrm{DD}})_d^{(m)}
\end{aligned} \tag{5.15}$$

ここで，$\boldsymbol{y}_{\mathrm{DD}}$は教師データとしての DD 特徴量ベクトル系列，$(\boldsymbol{y}_{\mathrm{DD}})_t$は教師データとしての時間フレーム$t$における DD 特徴量ベクトル，$(y_{\mathrm{DD}})_t^{(m)}$は教師データとしての時間フレーム$t$における$m$次の DD 特徴量，$M$は DD 特徴量の次元数，$(w_{\mathrm{DD}})_d^{(m)}$は次元$d$の音声特徴量についての$m$次の DD 特徴量を求める係数である．また，$\boldsymbol{y}_{\mathrm{DD}}$と同様に，DD 損失関数で予測データとしての音声特徴量から算出される DD 特徴量を次式で定義する．

$$
\begin{aligned}
\hat{\boldsymbol{y}}_{\mathrm{DD}} &= [(\hat{\boldsymbol{y}}_{\mathrm{DD}})_1^{\top}, \cdots, (\hat{\boldsymbol{y}}_{\mathrm{DD}})_t^{\top}, \cdots, (\hat{\boldsymbol{y}}_{\mathrm{DD}})_T^{\top}]^{\top} \\
(\hat{\boldsymbol{y}}_{\mathrm{DD}})_t &= \left[(\hat{y}_{\mathrm{DD}})_t^{(1)}, \cdots, (\hat{y}_{\mathrm{DD}})_t^{(m)}, \cdots, (\hat{y}_{\mathrm{DD}})_t^{(M)}\right] \\
(\hat{y}_{\mathrm{DD}})_t^{(m)} &= \sum_{d=1}^{D} \hat{y}_t^{(d)} (w_{\mathrm{DD}})_d^{(m)}
\end{aligned}
\tag{5.16}
$$

ここで，$\hat{\boldsymbol{y}}_{\mathrm{DD}}$は予測データとしての音声特徴量の DD 特徴量ベクトル系列，$(\hat{\boldsymbol{y}}_{\mathrm{DD}})_t$は予測データとしての時間フレーム$t$における DD 特徴量ベクトル，$(\hat{y}_{\mathrm{DD}})_t^{(m)}$は予測データとしての時間フレーム$t$における$m$次の DD 特徴量である．DD 損失関数は$\boldsymbol{y}_{\mathrm{DD}}$と$\hat{\boldsymbol{y}}_{\mathrm{DD}}$の平均二乗誤差で定義される．

$$
\begin{aligned}
(e_{\mathrm{DD}})_t^{(m)} &= \left((y_{\mathrm{DD}})_t^{(m)} - (\hat{y}_{\mathrm{DD}})_t^{(m)}\right)^2 \\
\mathcal{L}_{\mathrm{DD}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= \frac{1}{TM} \sum_{t=1}^{T} \sum_{m=1}^{M} (e_{\mathrm{DD}})_t^{(m)}
\end{aligned}
\tag{5.17}
$$

ここで，$(e_{\mathrm{DD}})_t^{(m)}$は$(\hat{y}_{\mathrm{DD}})_t^{(m)}$の$(y_{\mathrm{DD}})_t^{(m)}$に対する二乗誤差である．$(\hat{y}_{\mathrm{DD}})_t^{(m)}$は次元1から$D$までの$\hat{y}_t^{(d)}$から算出されるため，$(e_{\mathrm{DD}})_t^{(m)}$は次元1から$D$までの$\hat{y}_t^{(d)}$に関連する DNN のモデルパラメータの学習に寄与する．このようにすることで，隣接する次元間の音声特徴量の関係を学習できる．

$(w_{\mathrm{DD}})_d^{(m)}$については，対象とする音声特徴量ごとに適した変換式や，経験則による知見に基づいて，各次元の音声特徴量を関係づけることが好ましい．メルケプストラムを対象とする場合は，メルケプストラムの各次元の係数を関連付けるため，式（2.9 の周波数変換関数「freqt」に従うように$(w_{\mathrm{DD}})_d^{(m)}$の値を設定する．

5.3.4. 局所内分散の損失関数

LV 損失関数は，音声特徴量の短区間における分散の誤差を計算することによって，DNN に短区間における音声特徴量の振幅の大きさや，時間変動の程度を教える．LV 損失関数で教師データとしての音声特徴量から算出される局所内分散を次式で定義する．

$$
\begin{aligned}
\boldsymbol{y}_{\mathrm{LV}} &= [(\boldsymbol{y}_{\mathrm{LV}})_1^{\top}, \cdots, (\boldsymbol{y}_{\mathrm{LV}})_t^{\top}, \cdots, (\boldsymbol{y}_{\mathrm{LV}})_T^{\top}]^{\top} \\
(\boldsymbol{y}_{\mathrm{LV}})_t &= \left[(y_{\mathrm{LV}})_t^{(1)}, \cdots, (y_{\mathrm{LV}})_t^{(d)}, \cdots, (y_{\mathrm{LV}})_t^{(D)}\right] \\
(y_{\mathrm{LV}})_t^{(d)} &= \frac{1}{-L_{\mathrm{LV}} + R_{\mathrm{LV}} + 1} \sum_{\tau=L_{\mathrm{LV}}}^{R_{\mathrm{LV}}} \left(y_{t+\tau}^{(d)} - \bar{y}_t^{(d)}\right)^2 \\
\bar{y}_t^{(d)} &= \frac{1}{-L_{\mathrm{LV}} + R_{\mathrm{LV}} + 1} \sum_{\tau=L_{\mathrm{LV}}}^{R_{\mathrm{LV}}} y_{t+\tau}^{(d)}
\end{aligned}
\tag{5.18}
$$

ここで，$\boldsymbol{y}_{\mathrm{LV}}$は教師データとしての音声特徴量の局所内分散ベクトル系列，$(\boldsymbol{y}_{\mathrm{LV}})_t$は教師データとしての音声特徴量の時間フレーム$t$における局所内分散ベクトル，$(y_{\mathrm{LV}})_t^{(d)}$は教師データとしての次元$d$の音声特徴量の時間フレーム$t$における局所内分散，$\bar{y}_t^{(d)}$は教師データとしての次元$d$の音声特徴量の時間フレーム$t$における局所内平均，$L_{\mathrm{LV}}$は後方参照時間フレーム

数，R_{LV}は前方参照時間フレーム数である．また，$\boldsymbol{y}_{\mathrm{LV}}$と同様に，LV 損失関数で予測データとしての音声特徴量から算出される局所内分散を次式で定義する．

$$
\begin{aligned}
\hat{\boldsymbol{y}}_{\mathrm{LV}} &= [(\hat{\boldsymbol{y}}_{\mathrm{LV}})_1^{\top}, \cdots, (\hat{\boldsymbol{y}}_{\mathrm{LV}})_t^{\top}, \cdots, (\hat{\boldsymbol{y}}_{\mathrm{LV}})_T^{\top}]^{\top} \\
(\hat{\boldsymbol{y}}_{\mathrm{LV}})_t &= \left[(\hat{y}_{\mathrm{LV}})_t^{(1)}, \cdots, (\hat{y}_{\mathrm{LV}})_t^{(d)}, \cdots, (\hat{y}_{\mathrm{LV}})_t^{(D)}\right] \\
(\hat{y}_{\mathrm{LV}})_t^{(d)} &= \frac{1}{-L_{\mathrm{LV}} + R_{\mathrm{LV}} + 1} \sum_{\tau=L_{\mathrm{LV}}}^{R_{\mathrm{LV}}} \left(\hat{y}_{t+\tau}^{(d)} - \bar{\hat{y}}_t^{(d)}\right)^2 \\
\bar{\hat{y}}_t^{(d)} &= \frac{1}{-L_{\mathrm{LV}} + R_{\mathrm{LV}} + 1} \sum_{\tau=L_{\mathrm{LV}}}^{R_{\mathrm{LV}}} \hat{y}_{t+\tau}^{(d)}
\end{aligned}
\tag{5.19}
$$

ここで，$\hat{\boldsymbol{y}}_{\mathrm{LV}}$は教師データとしての音声特徴量の局所内分散ベクトル系列，$(\hat{\boldsymbol{y}}_{\mathrm{LV}})_t$は教師データとしての音声特徴量の時間フレーム$t$における局所内分散ベクトル，$(\hat{y}_{\mathrm{LV}})_t^{(d)}$は教師データとしての次元$d$の音声特徴量の時間フレーム$t$における局所内分散，$\bar{\hat{y}}_t^{(d)}$は教師データとしての次元$d$の音声特徴量の時間フレーム$t$における局所内平均である．LV 損失関数は$\boldsymbol{y}_{\mathrm{LV}}$と$\hat{\boldsymbol{y}}_{\mathrm{LV}}$の平均絶対誤差で定義される．

$$
\begin{aligned}
(e_{\mathrm{LV}})_t^{(d)} &= \left|(y_{\mathrm{LV}})_t^{(d)} - (\hat{y}_{\mathrm{LV}})_t^{(d)}\right| \\
\mathcal{L}_{\mathrm{LV}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= \frac{1}{TD} \sum_{t=1}^{T} \sum_{d=1}^{D} (e_{\mathrm{LV}})_t^{(d)}
\end{aligned}
\tag{5.20}
$$

ここで，$(e_{\mathrm{LV}})_t^{(d)}$は$(\hat{y}_{\mathrm{LV}})_t^{(d)}$の$(y_{\mathrm{LV}})_t^{(d)}$に対する絶対誤差である．$(\hat{y}_{\mathrm{LV}})_t^{(d)}$は$\hat{y}_{t+\tau}^{(d)}\,(L_{\mathrm{LV}} \leq \tau \leq R_{\mathrm{LV}})$から算出されるため，$(e_{\mathrm{LV}})_t^{(d)}$は$\hat{y}_{t+\tau}^{(d)}\,(L_{\mathrm{LV}} \leq \tau \leq R_{\mathrm{LV}})$に関連する DNN のモデルパラメータの学習に寄与する．このようにすることで，短区間$[t + L_{\mathrm{LV}}, t + R_{\mathrm{LV}}]$における音声特徴量の分散を学習できる．

5.3.5. 局所内共分散の損失関数

LC 損失関数は，メルケプストラムのような多次元の音声特徴量に対して利用する損失関数であり，音声特徴量の短区間における共分散の誤差を計算することで，DNN に短区間における音声特徴量の相関関係を教える．LC 損失関数で教師データとしての音声特徴量から算出される局所内共分散を次式で定義する．

$$
\begin{aligned}
\boldsymbol{y}_{\mathrm{LC}} &= [(\boldsymbol{y}_{\mathrm{LC}})_1^{\top}, \cdots, (\boldsymbol{y}_{\mathrm{LC}})_t^{\top}, \cdots, (\boldsymbol{y}_{\mathrm{LC}})_T^{\top}]^{\top} \\
(\boldsymbol{y}_{\mathrm{LC}})_t &= \left[(\boldsymbol{y}_{\mathrm{LC}})_t^{(1)}, \cdots, (\boldsymbol{y}_{\mathrm{LC}})_t^{(d_1)}, \cdots, (\boldsymbol{y}_{\mathrm{LC}})_t^{(D)}\right] \\
(\boldsymbol{y}_{\mathrm{LC}})_t^{(d_1)} &= \left[(y_{\mathrm{LC}})_t^{(d_1,1)}, \cdots, (y_{\mathrm{LC}})_t^{(d_1,d_2)}, \cdots, (y_{\mathrm{LC}})_t^{(d,D)}\right] \\
(y_{\mathrm{LC}})_t^{(d_1,d_2)} &= \frac{1}{-L_{\mathrm{LV}} + R_{\mathrm{LV}} + 1} \sum_{\tau=L_{\mathrm{LC}}}^{R_{\mathrm{LC}}} \left(y_{t+\tau}^{(d_1)} - \bar{y}_t^{(d_1)}\right)\left(y_{t+\tau}^{(d_2)} - \bar{y}_t^{(d_2)}\right)
\end{aligned}
\tag{5.21}
$$

ここで，$\boldsymbol{y}_{\mathrm{LC}}$は教師データとしての音声特徴量の局所内共分散ベクトル系列，$(\boldsymbol{y}_{\mathrm{LC}})_t$は教師データとしての音声特徴量の時間フレーム$t$における局所内共分散ベクトル，$(\boldsymbol{y}_{\mathrm{LC}})_t^{(d_1)}$は教

師データとしての次元d_1の音声特徴量についての時間フレームtにおける局所内共分散ベクトル，$(y_{\mathrm{LC}})_t^{(d_1,d_2)}$は教師データとしての次元$d_1$の音声特徴量と次元$d_2$の音声特徴量の時間フレーム$t$における局所内共分散，$L_{\mathrm{LC}}$は後方参照時間フレーム数，$R_{\mathrm{LC}}$は前方参照時間フレーム数である．また，$\boldsymbol{y}_{\mathrm{LC}}$と同様に，LC 損失関数で予測データとしての音声特徴量から算出される局所内共分散を次式で定義する．

$$
\begin{aligned}
\hat{\boldsymbol{y}}_{\mathrm{LC}} &= [(\hat{\boldsymbol{y}}_{\mathrm{LC}})_1^{\top},\cdots,(\hat{\boldsymbol{y}}_{\mathrm{LC}})_t^{\top},\cdots,(\hat{\boldsymbol{y}}_{\mathrm{LC}})_T^{\top}]^{\top} \\
(\hat{\boldsymbol{y}}_{\mathrm{LC}})_t &= \left[(\hat{\boldsymbol{y}}_{\mathrm{LC}})_t^{(1)},\cdots,(\hat{\boldsymbol{y}}_{\mathrm{LC}})_t^{(d_1)},\cdots,(\hat{\boldsymbol{y}}_{\mathrm{LC}})_t^{(D)}\right] \\
(\hat{y}_{\mathrm{LC}})_t^{(d_1)} &= \left[(\hat{y}_{\mathrm{LC}})_t^{(d_1,1)},\cdots,(\hat{y}_{\mathrm{LC}})_t^{(d_1,d_2)},\cdots,(\hat{y}_{\mathrm{LC}})_t^{(d,D)}\right] \\
(\hat{y}_{\mathrm{LC}})_t^{(d_1,d_2)} &= \frac{1}{-L_{\mathrm{LV}}+R_{\mathrm{LV}}+1}\sum_{\tau=L_{\mathrm{LC}}}^{R_{\mathrm{LC}}}\left(\hat{y}_{t+\tau}^{(d_1)}-\bar{\hat{y}}_t^{(d_1)}\right)\left(\hat{y}_{t+\tau}^{(d_2)}-\bar{\hat{y}}_t^{(d_2)}\right)
\end{aligned}
\tag{5.22}
$$

ここで，$\hat{\boldsymbol{y}}_{\mathrm{LC}}$は予測データとしての音声特徴量の局所内共分散ベクトル系列，$(\hat{\boldsymbol{y}}_{\mathrm{LC}})_t$は予測データとしての音声特徴量の時間フレーム$t$における局所内共分散ベクトル，$(\hat{\boldsymbol{y}}_{\mathrm{LC}})_t^{(d_1)}$は予測データとしての次元$d_1$の音声特徴量についての時間フレーム$t$における局所内共分散ベクトル，$(\hat{y}_{\mathrm{LC}})_t^{(d_1,d_2)}$は予測データとしての次元$d_1$の音声特徴量と次元$d_2$の音声特徴量の時間フレーム$t$における局所内共分散である．LC 損失関数は$\boldsymbol{y}_{\mathrm{LC}}$と$\hat{\boldsymbol{y}}_{\mathrm{LC}}$の平均絶対誤差で定義される．

$$
\begin{aligned}
(e_{\mathrm{LC}})_t^{(d_1,d_2)} &= \left|(y_{\mathrm{LC}})_t^{(d_1,d_2)}-(\hat{y}_{\mathrm{LC}})_t^{(d_1,d_2)}\right| \\
\mathcal{L}_{\mathrm{LC}}(\boldsymbol{y},\hat{\boldsymbol{y}}) &= \frac{1}{TD^2}\sum_{t=1}^{T}\sum_{d_1=1}^{D}\sum_{d_2=1}^{D}(e_{\mathrm{LC}})_t^{(d_1,d_2)}
\end{aligned}
\tag{5.23}
$$

ここで，$(e_{\mathrm{LC}})_t^{(d_1,d_2)}$は$(\hat{y}_{\mathrm{LC}})_t^{(d_1,d_2)}$の$(y_{\mathrm{LC}})_t^{(d_1,d_2)}$に対する絶対誤差である．$(\hat{y}_{\mathrm{LC}})_t^{(d_1,d_2)}$は$\hat{y}_{t+\tau}^{(d_1)}\ (L_{\mathrm{LC}}\le\tau\le R_{\mathrm{LC}})$と$\hat{y}_{t+\tau}^{(d_2)}\ (L_{\mathrm{LC}}\le\tau\le R_{\mathrm{LC}})$から算出されるため，$(e_{\mathrm{LC}})_t^{(d_1,d_2)}$は$\hat{y}_{t+\tau}^{(d_1)}\ (L_{\mathrm{LC}}\le\tau\le R_{\mathrm{LC}})$と$\hat{y}_{t+\tau}^{(d_2)}\ (L_{\mathrm{LC}}\le\tau\le R_{\mathrm{LC}})$に関連するモデルパラメータの学習に寄与する．このようにすることで，短区間$[t+L_{\mathrm{LC}},\,t+R_{\mathrm{LC}}]$における音声特徴量の共分散を学習できる．

5.3.6. 系列内分散の損失関数

GV 損失関数は音声特徴量の系列全体における分散の誤差を計算することによって，DNN に系列全体における音声特徴量の振幅の大きさや，時間変動の程度を教える．教師データとしての音声特徴量から算出される系列内分散を次式で定義する．

$$
\begin{aligned}
\boldsymbol{y}_{\mathrm{GV}} &= \left[y_{\mathrm{GV}}^{(1)},\cdots,y_{\mathrm{GV}}^{(d)},\cdots,y_{\mathrm{GV}}^{(D)}\right] \\
y_{\mathrm{GV}}^{(d)} &= \frac{1}{T}\sum_{t=1}^{T}\left(y_t^{(d)}-\bar{y}^{(d)}\right)^2 \\
\bar{y}^{(d)} &= \frac{1}{T}\sum_{t=1}^{T}y_t^{(d)}
\end{aligned}
\tag{5.24}
$$

ここで，$\boldsymbol{y}_{\mathrm{GV}}$は教師データとしての音声特徴量の系列内分散ベクトル，$y_{\mathrm{GV}}^{(d)}$は教師データとしての次元dの音声特徴量の系列内分散，$\bar{y}^{(d)}$は教師データとしての次元の音声特徴量の系列内平均である．また，$\boldsymbol{y}_{\mathrm{GV}}$と同様に，GV 損失関数で予測データとしての音声特徴量から算出される系列内分散を次式で定義する．

$$\begin{aligned}\hat{\boldsymbol{y}}_{\mathrm{GV}} &= \left[\hat{y}_{\mathrm{GV}}^{(1)}, \cdots, \hat{y}_{\mathrm{GV}}^{(d)}, \cdots, \hat{y}_{\mathrm{GV}}^{(D)}\right] \\ \hat{y}_{\mathrm{GV}}^{(d)} &= \frac{1}{T}\sum_{t=1}^{T}\left(\hat{y}_t^{(d)} - \bar{\hat{y}}^{(d)}\right)^2 \\ \bar{\hat{y}}^{(d)} &= \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t^{(d)}\end{aligned} \tag{5.25}$$

ここで，$\hat{\boldsymbol{y}}_{\mathrm{GV}}$は予測データとしての音声特徴量の系列内分散ベクトル，$\hat{y}_{\mathrm{GV}}^{(d)}$は予測データとしての次元$d$の音声特徴量の系列内分散，$\bar{\hat{y}}^{(d)}$は予測データとしての次元の音声特徴量の系列内平均である．GV 損失関数は$\boldsymbol{y}_{\mathrm{GV}}$と$\hat{\boldsymbol{y}}_{\mathrm{GV}}$の平均絶対誤差で定義される．

$$\begin{aligned}e_{\mathrm{GV}}^{(d)} &= \left|y_{\mathrm{GV}}^{(d)} - \hat{y}_{\mathrm{GV}}^{(d)}\right| \\ \mathcal{L}_{\mathrm{GV}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= \frac{1}{D}\sum_{d=1}^{D} e_{\mathrm{GV}}^{(d)}\end{aligned} \tag{5.26}$$

ここで，$e_{\mathrm{GV}}^{(d)}$は$\hat{y}_{\mathrm{GV}}^{(d)}$の$y_{\mathrm{GV}}^{(d)}$に対する絶対誤差である．$\hat{y}_{\mathrm{GV}}^{(d)}$は$\hat{y}_t^{(d)}$ $(1 \le t \le T)$から算出されるため，$e_{\mathrm{GV}}^{(d)}$は$\hat{y}_t^{(d)}$ $(1 \le t \le T)$に関連する DNN のモデルパラメータの学習に寄与する．このようにすることで，系列全体における音声特徴量の分散を学習できる．

5.3.7. 系列内共分散の損失関数

GC 損失関数は，メルケプストラムのような多次元の音声特徴量に対して利用する損失関数であり，音声特徴量の系列全体における共分散の誤差を計算することによって，DNN に系列全体における音声特徴量の相関関係を教える．GC 損失関数で教師データとしての音声特徴量から算出される系列内共分散を次式で定義する．

$$\begin{aligned}\boldsymbol{y}_{\mathrm{GC}} &= \left[\boldsymbol{y}_{\mathrm{GC}}^{(1)}, \cdots, \boldsymbol{y}_{\mathrm{GC}}^{(d_1)}, \cdots, \boldsymbol{y}_{\mathrm{GC}}^{(D)}\right] \\ \boldsymbol{y}_{\mathrm{GC}}^{(d_1)} &= \left[y_{\mathrm{GC}}^{(d_1,1)}, \cdots, y_{\mathrm{GC}}^{(d_1,d_2)}, \cdots, y_{\mathrm{GC}}^{(d_1,D)}\right] \\ y_{\mathrm{GC}}^{(d_1,d_2)} &= \frac{1}{T}\sum_{t=1}^{T}\left(y_t^{(d_1)} - \bar{y}^{(d_1)}\right)\left(y_t^{(d_2)} - \bar{y}^{(d_2)}\right)\end{aligned} \tag{5.27}$$

ここで，$\boldsymbol{y}_{\mathrm{GC}}$は教師データとしての音声特徴量の共分散ベクトル，$\boldsymbol{y}_{\mathrm{GC}}^{(d_1)}$は教師データとしての次元$d_1$の音声特徴量の共分散ベクトル，$y_{\mathrm{GC}}^{(d_1,d_2)}$は教師データとしての次元$d_1$の音声特徴量と次元$d_2$の音声特徴量の共分散である．また，$\boldsymbol{y}_{\mathrm{GC}}$と同様に，GC 損失関数で予測データとしての音声特徴量から算出される系列内共分散を次式で定義する．

$$
\begin{aligned}
\hat{\boldsymbol{y}}_{\mathrm{GC}} &= \left[\hat{\boldsymbol{y}}_{\mathrm{GC}}^{(1)}, \cdots, \hat{\boldsymbol{y}}_{\mathrm{GC}}^{(d_1)}, \cdots, \hat{\boldsymbol{y}}_{\mathrm{GC}}^{(D)}\right] \\
\hat{\boldsymbol{y}}_{\mathrm{GC}}^{(d_1)} &= \left[\hat{y}_{\mathrm{GC}}^{(d_1,1)}, \cdots, \hat{y}_{\mathrm{GC}}^{(d_1,d_2)}, \cdots, \hat{y}_{\mathrm{GC}}^{(d_1,D)}\right] \\
\hat{y}_{\mathrm{GC}}^{(d_1,d_2)} &= \frac{1}{T}\sum_{t=1}^{T}\left(\hat{y}_t^{(d_1)} - \bar{\hat{y}}^{(d_1)}\right)\left(\hat{y}_t^{(d_2)} - \bar{\hat{y}}^{(d_2)}\right)
\end{aligned}
\tag{5.28}
$$

ここで，$\hat{\boldsymbol{y}}_{\mathrm{GC}}$は予測データとしての音声特徴量の共分散ベクトル，$\hat{\boldsymbol{y}}_{\mathrm{GC}}^{(d_1)}$は予測データとしての次元$d_1$の音声特徴量の共分散ベクトル，$\hat{y}_{\mathrm{GC}}^{(d_1,d_2)}$は予測データとしての次元$d_1$の音声特徴量と次元$d_2$の音声特徴量の共分散である．GC 損失関数は$\boldsymbol{y}_{\mathrm{GC}}$と$\hat{\boldsymbol{y}}_{\mathrm{GC}}$の平均絶対誤差で定義される．

$$
\begin{aligned}
e_{\mathrm{GC}}^{(d_1,d_2)} &= \left|y_{\mathrm{GV}}^{(d_1,d_2)} - \hat{y}_{\mathrm{GV}}^{(d_1,d_2)}\right| \\
\mathcal{L}_{\mathrm{GV}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= \frac{1}{D^2}\sum_{d_1=1}^{D}\sum_{d_2=1}^{D} e_{\mathrm{GC}}^{(d_1,d_2)}
\end{aligned}
\tag{5.29}
$$

ここで，$e_{\mathrm{GC}}^{(d_1,d_2)}$は$\hat{y}_{\mathrm{GV}}^{(d_1,d_2)}$の$y_{\mathrm{GV}}^{(d_1,d_2)}$に対する絶対誤差である．$\hat{y}_{\mathrm{GV}}^{(d_1,d_2)}$は$\hat{y}_t^{(d_1)}\ (1 \le t \le T)$と$\hat{y}_t^{(d_2)}\ (1 \le t \le T)$から算出されるため，$e_{\mathrm{GC}}^{(d_1,d_2)}$は$\hat{y}_t^{(d_1)}\ (1 \le t \le T)$と$\hat{y}_t^{(d_2)}\ (1 \le t \le T)$に関連する DNN のモデルパラメータの学習に寄与する．このようにすることで，系列全体における音声特徴量の共分散を学習できる．

5.4. 実験方法

5.4.1. 音声特徴量予測部の学習条件

実験に用いた音声特徴量予測部，DNN，損失関数，勾配法の組み合わせを表 5.1 に示す．FFNN-MSE と FFNN-MGE は損失関数が異なるだけで，対象とする音声特徴量予測部の構成は同じである．勾配法はいずれも Adam 法であり，Adam 法のパラメータについては，学習率を 0.001，β_1を 0.9，β_2を 0.999，微小量を10^{-7}，学習率減衰を 0.0 とした．エポック数は 20 とし，バッチサイズは 1 文ごとの音声特徴量の時間フレーム数とした．言語特徴量の正規化法は 4.2.4 で提案した 2 つの言語特徴量の属性値の比を取る正規化法を使用した．学習データセットと評価データセットはそれぞれ 2.3 で説明した$\mathbb{U}_{2000}$と$\mathbb{U}_{\mathrm{s}}$を使用した．

表 5.1　音声特徴量予測部，DNN，損失関数，勾配法の組み合わせ

識別名	音声特徴量予測部の構成	DNN の構成	損失関数	勾配法
FFNN-MSE	FFNN MLPG ケプストラム強調 （3.2.1）	全結合層×5 （FFNN-3.2.1）	$\mathcal{L}_{\mathrm{MSE}}(\boldsymbol{\mu},\hat{\boldsymbol{\mu}})$ （5.2.2）	Adam 法
FFNN-MGE	FFNN MLPG ケプストラム強調 （3.2.1）	全結合層×5 （FFNN-3.2.1）	$\mathcal{L}_{\mathrm{MGE}}(\boldsymbol{y},\hat{\boldsymbol{\psi}})$ （5.2.3）	Adam 法
RNN-MSE	RNN ケプストラム強調 （3.2.2）	LSTM 層 再帰層 （RNN-3.2.2）	$\mathcal{L}_{\mathrm{MSE}}(\boldsymbol{y},\hat{\boldsymbol{y}})$ （5.2.1）	Adam 法
FFNN-MATS	FFNN （3.2.3）	全結合層×5 （FFNN-3.2.3）	$\mathcal{L}_{\mathrm{MATS}}(\boldsymbol{y},\hat{\boldsymbol{y}})$ （5.3）	Adam 法

5.4.2. 聴取実験の方法

各音声特徴量予測部で予測した音声特徴量を比較するために, MUSHRA 法による合成音声の聴取実験で主観評価した．隠れ参照とアンカーを用いた複数刺激の聴取実験法（MUSHRA 法：Multi-Stimulus listening test using the Hidden Reference and Anchor 法 [39]）による聴取実験の手順を図 5.1 に示す．MUSHRA 法では，複数の評価群に加えて，参照群とアンカー群を用意する．参照群は実験における最高品質の音声，アンカー群は実験における最低品質の音声とする．参照群とアンカー群を使用することで，各刺激音声を採点する際の上限と下限の評価基準を設けることができる．ただし，参照群やアンカー群の音声がどの刺激音声に割り当てられているかは知らされない．参加者は基準音声と刺激音声を比較したり，刺激音声同士を比較したりして，基準音声に対する刺激音声の評価を表 5.2 に従い採点する．また，基準音声と同じと判断される刺激音声は必ず 100 点で採点する．採点するにあたり，基準音声や刺激音声は何度も聴くことができる．各群の合成音声がどの刺激音声に割り当たるかは，セッションごとにランダムで決めた．参加者の平均評点は次式に従って集計した．

$$\mathbb{V}=\left\{ v_i^{(G)} \;\middle|\; v_i^{(G)}=\frac{1}{S}\sum_{s=1}^{S}(v_i)_s^{(G)} \right\} \tag{5.30}$$

ここで，$v_i^{(G)}$は参加者iの評価群Gの平均評点，$(v_i)_s^{(G)}$はs回目のセッションにおける参加者iの評価群Gの評点，Sはセッション数である．

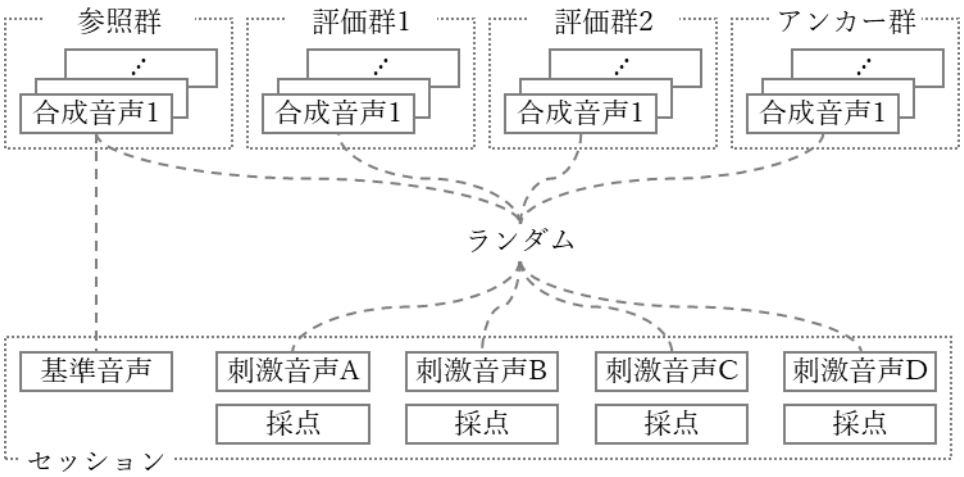

図 5.1　MUSHRA 法による聴取実験の手順

図は評価群が 2 つの場合の例である．評価群がN_Gの場合，1 セッションあたりの刺激音声の数は$N_G + 2$となる．ここで，N_Gは評価群の総数である．

表 5.2　MUSHRA 法の評点

評点	説明
80～100 点	基準音声との違いが分からない
60～80 点	基準音声との違いが分かるが気にならない
40～60 点	基準音声との違いが少し気になる
20～40 点	基準音声との違いが気になる
0～20 点	基準音声との違いがとても気になる

5.4.3. 予測誤差の算出方法

聴取実験の結果を裏付けるために音声特徴量の予測誤差を計算する．ただし，音声特徴量の予測誤差と合成音声の品質との因果関係は絶対的なものではないため，音声特徴量の予測誤差は聴取実験の結果を補足するために用いる．本章において，3 つの音声特徴量の予測誤差を計算した．1 つめは時間フレームごとの音声特徴量の絶対誤差，2 つめは音声特徴量の系列内分散の平方根の絶対誤差，3 つめは音声特徴量の変調スペクトルの絶対誤差である[40]．音声特徴量の平均絶対誤差を次式で定義する．

$$\mathbb{E}_{\mathrm{DC}} = \left\{ \varepsilon_{\mathrm{DC}} \;\middle|\; \varepsilon_{\mathrm{DC}} = \frac{1}{TD} \sum_{t=1}^{T} \sum_{d=1}^{D} \left| y_t^{(d)} - \hat{y}_t^{(d)} \right| \right\} \quad (\boldsymbol{y} \in \mathbb{U}) \tag{5.31}$$

ここで，$\mathbb{U}$は評価データセット，$\boldsymbol{y}$は$\mathbb{U}$に含まれる原音声の音声特徴量，$y_t^{(d)}$は時間フレームtにおけるd次の原音声の音声特徴量，$\hat{y}_t^{(d)}$は$y_t^{(d)}$に対応する時間フレームtにおけるd次の予測した音声特徴量，$\varepsilon_{\mathrm{DC}}$は$\boldsymbol{y}$についての平均絶対誤差，$\mathbb{E}_{\mathrm{DC}}$は$\mathbb{U}$についての$\varepsilon_{\mathrm{DC}}$の集合である．系列内分散の平方根の平均絶対誤差を次式で定義する．

$$\mathbb{E}_{\mathrm{GV}} = \left\{ \varepsilon_{\mathrm{GV}} \;\middle|\; \varepsilon_{\mathrm{GV}} = \frac{1}{D}\sum_{d=1}^{D}\left|\sqrt{y_{\mathrm{GV}}^{(d)}} - \sqrt{\hat{y}_{\mathrm{GV}}^{(d)}}\right| \right\} \quad (\boldsymbol{y} \in \mathbb{U}) \tag{5.32}$$

ここで，$y_{\mathrm{GV}}^{(d)}$はd次の原音声の音声特徴量の系列内分散，は$y_{\mathrm{GV}}^{(d)}$に対応するd次の予測した音声特徴量の系列内分散，$\varepsilon_{\mathrm{GV}}$は$\boldsymbol{y}$についての系列内分散の平均絶対誤差，$\mathbb{E}_{\mathrm{GV}}$は$\mathbb{U}$についての$\varepsilon_{\mathrm{GV}}$の集合である．変調スペクトルの平均絶対誤差を次式で定義する．

$$\begin{aligned}
&\mathbb{E}_{\mathrm{MS}} = \left\{ \varepsilon_{\mathrm{MS}} \;\middle|\; \varepsilon_{\mathrm{MS}} = \frac{1}{TDH}\sum_{t=1}^{T}\sum_{d=1}^{D}\sum_{j=1}^{H}\left|(y_{\mathrm{MS}})_t^{(j,d)} - (\hat{y}_{\mathrm{MS}})_t^{(j,d)}\right| \right\} \quad (\boldsymbol{y} \in \mathbb{U}) \\
&(y_{\mathrm{MS}})_t^{(j,d)} = \mathcal{F}_{\mathrm{MS}}(\boldsymbol{y} \mid t, d, L_{\mathrm{MS}}, R_{\mathrm{MS}}) \quad (j = 1, 2, \cdots, H) \\
&(\hat{y}_{\mathrm{MS}})_t^{(j,d)} = \mathcal{F}_{\mathrm{MS}}(\hat{\boldsymbol{y}} \mid t, d, L_{\mathrm{MS}}, R_{\mathrm{MS}}) \quad (j = 1, 2, \cdots, H) \\
&H = \frac{-L_{\mathrm{MS}} + R_{\mathrm{MS}} + 1}{2} + 1
\end{aligned} \tag{5.33}$$

ここで，$(y_{\mathrm{MS}})_t^{(j,d)}$は時間フレーム$t$における$d$次の音声特徴量の$j$番目の周波数ビンの変調スペクトル，$L_{\mathrm{MS}}$は前方参照時間フレーム数，$R_{\mathrm{MS}}$は後方参照時間フレーム数，$\varepsilon_{\mathrm{MS}}$は$\boldsymbol{y}$についての変調スペクトルの平均絶対誤差，$\mathbb{E}_{\mathrm{MS}}$は$\mathbb{U}$についての$\varepsilon_{\mathrm{MS}}$の集合である．ただし，$L_{\mathrm{MS}}$は負数であり-64 とした．$R_{\mathrm{MS}}$は正数であり 63 とした．また，$\mathcal{F}_{\mathrm{MS}}$は$(y_{\mathrm{MS}})_t^{(j,d)}$を算出する関数であり，次式で定義される．

$$\begin{aligned}
&\mathcal{F}_{\mathrm{MS}}(\boldsymbol{y} \mid t, d, L_{\mathrm{MS}}, R_{\mathrm{MS}}) \equiv 20\log_{10}\left|\mathfrak{F}\left(\boldsymbol{y}_{(t,L_{\mathrm{MS}},R_{\mathrm{MS}})}^{(d)}\right)\right| \\
&\boldsymbol{y}_{(t,L_{\mathrm{MS}},R_{\mathrm{MS}})}^{(d)} = \left[y_{t+L_{\mathrm{MS}}}^{(d)}h_{L_{\mathrm{MS}}}, \cdots, y_{t+\tau}^{(d)}h_{\tau}, \cdots, y_{t+R_{\mathrm{MS}}}^{(d)}h_{R_{\mathrm{MS}}}\right] \\
&h_\tau = \frac{h'_\tau}{h'} \\
&h'_\tau = 0.5 - 0.5\cos\left(\frac{2\pi(\tau + H - 0.5)}{-L_{\mathrm{MS}} + R_{\mathrm{MS}} + 1}\right) \\
&h' = \sum_{\tau=L_{\mathrm{MS}}}^{R_{\mathrm{MS}}} h'_\tau
\end{aligned} \tag{5.34}$$

ここで，$\mathfrak{F}$は離散フーリエ変換，$\boldsymbol{y}_{(t,L_{\mathrm{MS}},R_{\mathrm{MS}})}^{(d)}$は時間フレーム$t$を中心とする短区間$[t + L_{\mathrm{MS}}, t + R_{\mathrm{MS}}]$における窓関数を適用した$d$次の音声特徴量ベクトル，$h_\tau$は正規化されたハン窓の係数である．

5.5. 対数基本周波数についての実験結果

MATS 損失関数はモデル化する音声特徴量を適切に捉えるために，対象とする音声特徴量ごとに各損失関数のパラメータの設定を調整する必要がある．そのため，まず，MATS 損失関数の各損失関数のパラメータの設定の調整法について述べる．次に，対数基本周波数に適した設定をした MATS 損失関数を用いたときの聴取実験と予測誤差について述べる．

5.5.1. MATS 損失関数のパラメータ設定

対数基本周波数のモデル化において利用可能な損失関数は DC 損失関数，TD 損失関数，

GV 損失関数，LV 損失関数である．対数基本周波数に対するこれらの損失関数の挙動を確認するため，約 40 通りの損失関数のパラメータの組み合わせを試した．試した損失関数のパラメータの組み合わせの一部の結果を図 5.18 に示す．1 列目の「条件」の見出しは表 5.11 と対応する．DC1 が示すように，DC 損失関数だけでは，滑らかな対数基本周波数を予測する DNN のモデルパラメータを獲得できない．TD1 から TD5 が示すように，DC 損失関数と TD 損失関数では，w_2を大きくしても，十分に滑らかな対数基本周波数を予測する DNN のモデルパラメータを獲得できない．一方で，TD6 から TD10 が示すように，TD 損失関数を使用して，w_2を大きくすることで，滑らかな対数基本周波数を予測する DNN のモデルパラメータを獲得できた．ただし，w_2を大きくすることで，$\mathbb{E}_{\mathrm{MS}}$は減少したが，$\mathbb{E}_{\mathrm{GV}}$は増加した．GV1 から GV4 が示すように，GV 損失関数を追加することで，対数基本周波数の GV も考慮した DNN のモデルパラメータを獲得できた．ただし，ω_{GV}を大きくすると対数基本周波数の軌跡が崩れてしまうことがわかった．また，LV1 から LV12 が示すように，LV 損失関数にもわずかではあるが，$\mathbb{E}_{\mathrm{GV}}$を減少させる傾向がみられた．

対数基本周波数のモデル化においては，合成音声の韻律が滑らかな変化し，抑揚が単調にならないようにするため，対数基本周波数が滑らかに変化することと，対数基本周波数の系列内分散が小さくならないようすることに注意した．この方針と図 5.18 の実験で得た知見に基づいて，TD 損失関数と GV 損失関数が MATS 損失関数の中核となるようにパラメータの調整を行った．パラメータは試行錯誤を繰り返すことで調整した．DNN を学習し，数個の合成音声を聴いて韻律を確認するということを繰り返し行った．韻律に不具合が生じたら，その原因と考えられるパラメータの値の調整や，その不具合を抑制すると考えられる損失関数を追加した．対数基本周波数については，TD 損失関数と GV 損失関数で良好な DNN のモデルパラメータが学習できるが，GV 損失関数との相乗効果を期待して LV 損失関数も使用することにした．最終的に，対数基本周波数を MATS 損失関数で学習するときのパラメータは$\omega_{\mathrm{TD}} = 1$，$L_{\mathrm{TD}} = -1$，$R_{\mathrm{TD}} = 0$，$w_1 = 1$，$w_2 = 20$，$\omega_{\mathrm{GV}} = 1$，$\omega_{\mathrm{LV}} = 2$，$L_{\mathrm{LV}} = -8$，$R_{\mathrm{LV}} = 8$となった．

5.5.2. 聴取実験の結果

MUSHRA 法による聴取実験の結果を図 5.2 に示す．実験に用いた合成音声は 2.1.2 のボコーダの合成部で生成した．対数基本周波数を予測するときは，原音声の継続長から算出した時間フレーム情報を付与した言語特徴量を使用した．FFNN-MSE，FFNN-MGE，RNN-MSE，FFNN-MATS で予測した対数基本周波数から合成した音声を評価群の音声とした．評価群の音声は予測した対数基本周波数と，その対数基本周波数に対応する原音声のスペクトル包絡と非周期性指標から合成した．参照音声は原音声の分析再合成音声である．アンカー音声は FFNN-MATS において DC 損失関数のみで学習した FFNN で予測した対数基本周波数と，その対数基本周波数に対応する原音声のスペクトル包絡と非周期性指標から合成した．参加者は合成音声の韻律や音質の違いに敏感な 10 名である．合成音声の韻律を

評価するため，合成音声のアクセントや抑揚に注目して評価するように指示をした．また，セッション数は 100 であるため，参加者を適宜休憩させた．

図 5.2 の評点を Tukey-Kramer 法によって比較した結果を表 5.3 に示す．その結果，FFNN-MATS 群と FFNN-MSE 群，FFNN-MATS 群と FFNN-MGE 群，FFNN-MGE 群と RNN-MSE 群の評点には有意差が認められなかった．このため，対数基本周波数の予測については，FFNN-MATS は FFNN-MSE と FFNN-MGE と同等であり，RNN-MSE よりは優れているといえる．特に，FFNN-MSE 群と FFNN-MATS 群の合成音声には，評点が 95 点を超え，参照群の合成音声とほぼ同じ品質のものがあった．

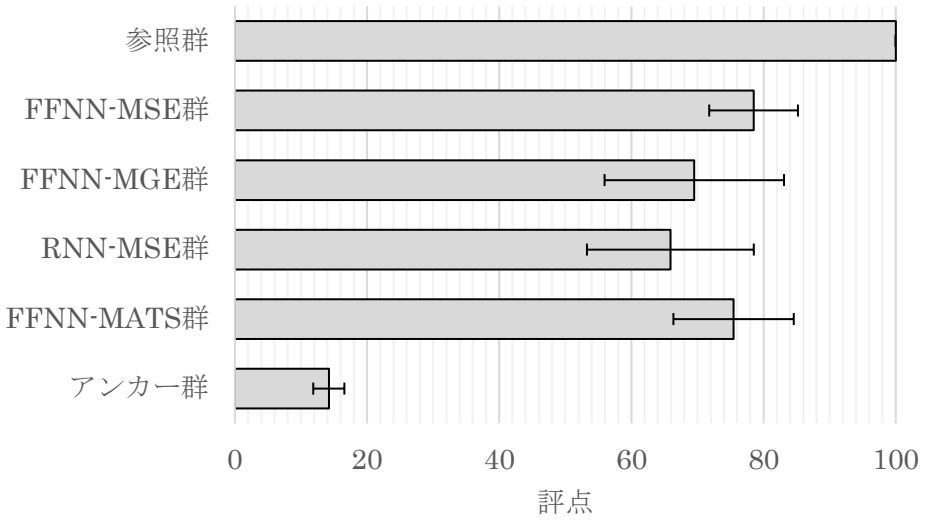

図 5.2　韻律の品質についての MUSHRA 法による聴取実験の結果

表 5.3　Tukey-Kramer 法による聴取実験の$\mathbb{V}$の平均値の比較結果

表中の数値はスチューデント化された範囲分布の q 値と p 値である．群数は 6，自由度は 54，信頼区間は 95%である．

群 1	群 2	q 値	p 値
参照群	FFNN-MSE 群	14.18	0.001
参照群	FFNN-MGE 群	20.10	0.001
参照群	RNN-MSE 群	22.48	0.001
参照群	FFNN-MATS 群	16.18	0.001
参照群	アンカー群	56.56	0.001
FFNN-MSE 群	FFNN-MGE 群	5.91	0.002
FFNN-MSE 群	RNN-MSE 群	8.30	0.001
FFNN-MSE 群	FFNN-MATS 群	1.99	0.696
FFNN-MSE 群	アンカー群	42.37	0.001
FFNN-MGE 群	RNN-MSE 群	2.39	0.540
FFNN-MGE 群	FFNN-MATS 群	3.92	0.078
FFNN-MGE 群	アンカー群	36.46	0.001
RNN-MSE 群	FFNN-MATS 群	6.30	0.001
RNN-MSE 群	アンカー群	34.07	0.001
FFNN-MATS 群	アンカー群	40.38	0.001

5.5.3. 予測誤差の結果

各音声特徴量予測部で予測した対数基本周波数の代表例を図 5.3 から図 5.6 までに示す．いずれの対数基本周波数パターンも起伏の大きさは異なるものの，原音声の対数基本周波数パターンと類似していた．予測した対数基本周波数の系列内分散の平方根は原音声の対数基本周波数の系列内分散の平方根よりも約 0.04 小さかった．FFNN-MATS の対数基本周波数は 2.5 秒や 2.8 秒においてわずかに不連続であった．また，RNN-MSE の対数基本周波数は 1.0 秒から 1.4 秒や，2.0 秒から 2.2 秒の区間において不規則に変動していた．これらの不連続や不規則な変動は，10 Hz 以上の帯域の変調スペクトルのレベルを上昇させたが，合成音声を聴いても知覚できなかった．これは，対数基本周波数の変調スペクトルの主成分が 10 Hz 以下の帯域にあり，10 Hz 以上の帯域の変調スペクトルと主成分の差が数十 dB 以上あったためである．一方で，FFNN-MSE と FFNN-MGE の変調スペクトルは原音声の変調スペクトルと同じであり，原音声の対数基本周波数と同じように滑らかだった．これは，MLPG の平滑化によるものである．

各音声特徴量予測部で予測した対数基本周波数の$\mathbb{U}_s$についての$\mathbb{E}_{DC}$，$\mathbb{E}_{GV}$，$\mathbb{E}_{MS}$をそれぞれ，図 5.7，図 5.8，図 5.9 に示す．また，これらの対数基本周波数の$\mathbb{U}_s$についての予測誤差$\mathbb{E}_{DC}$，$\mathbb{E}_{GV}$，$\mathbb{E}_{MS}$の平均値を Tukey-Kramer 法で比較した結果を表 5.4，表 5.5，表 5.6 に

示す．FFNN-MATS の$\mathbb{E}_{\mathrm{DC}}$の平均値は，FFNN-MGE の$\mathbb{E}_{\mathrm{DC}}$の平均値よりも有意に小さく，FFNN-MSE と RNN-MSE の$\mathbb{E}_{\mathrm{DC}}$の平均値との有意差はなかった．FFNN-MATS の$\mathbb{E}_{\mathrm{DC}}$の中央値と FFNN-MSE，FFNN-MGE，RNN-MGE の$\mathbb{E}_{\mathrm{DC}}$の中央値の差は約 0.005 以下であり，対数基本周波数の値や聴取実験の評点を考慮すると，これらの差は合成音声において無視できる程度のものである．

FFNN-MATS の$\mathbb{E}_{\mathrm{GV}}$の平均値は，FFNN-MSE，FFNN-MGE，RNN-MSE の$\mathbb{E}_{\mathrm{GV}}$の平均値よりも有意に小さかった．FFNN-MATS の$\mathbb{E}_{\mathrm{GV}}$の中央値と FFNN-MSE，FFNN-MGE，RNN-MSE の$\mathbb{E}_{\mathrm{GV}}$の中央値の差は約 0.02 以下であり，対数基本周波数の系列内分散の平方根の値や聴取実験の評点を考慮すると，これらの差は合成音声において無視できる程度のものである．

FFNN-MATS の$\mathbb{E}_{\mathrm{MS}}$の平均値は，FFNN-MSE，FFNN-MGE，RNN-MSE の$\mathbb{E}_{\mathrm{MS}}$の平均値よりも有意に大きかった．FFNN-MATS の$\mathbb{E}_{\mathrm{MS}}$の中央値は FFNN-MSE，FFNN-MGE，RNN-MSE の$\mathbb{E}_{\mathrm{MS}}$の中央値よりもそれぞれ約 8 dB，約 7 dB，約 3 dB 大きかった．しかし，これらの差は，対数基本周波数の代表例の変調スペクトルについて述べた 10 Hz 以上の帯域における誤差によるものであり，合成音声の品質を大きく損ねるものではない．また，RNN-MSE の$\mathbb{E}_{\mathrm{MS}}$の中央値は，FFNN-MSE，FFNN-MGE の$\mathbb{E}_{\mathrm{MS}}$の中央値よりもそれぞれ約 5 dB，約 4 dB 大きいが，FFNN-MATS の$\mathbb{E}_{\mathrm{MS}}$と同様に，対数基本周波数の代表例の変調スペクトルについて述べた 10 Hz 以上の帯域における誤差によるものであり，合成音声の品質を大きく損ねるものではない．

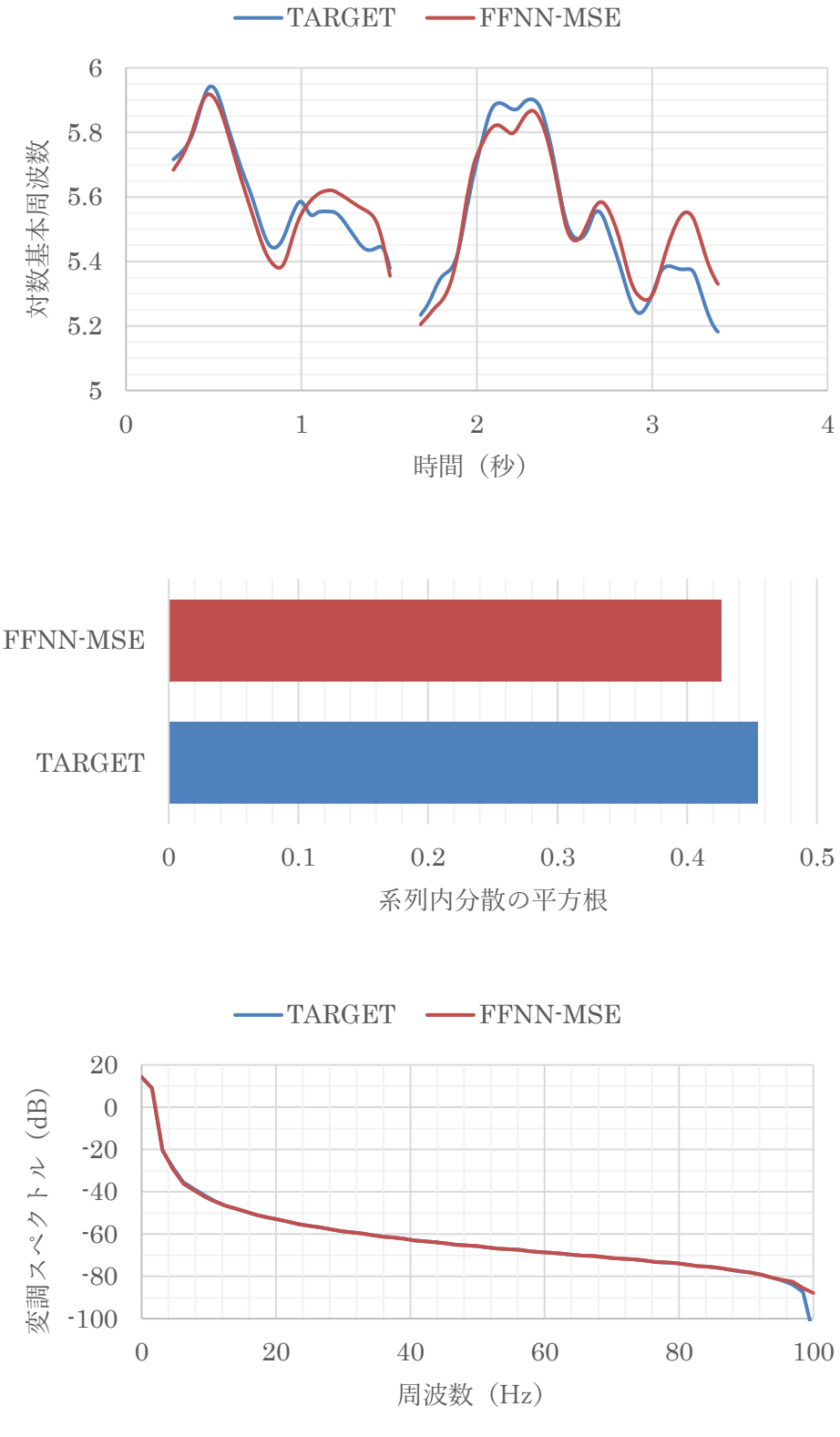

図 5.3　FFNN-MSE の対数基本周波数の代表例

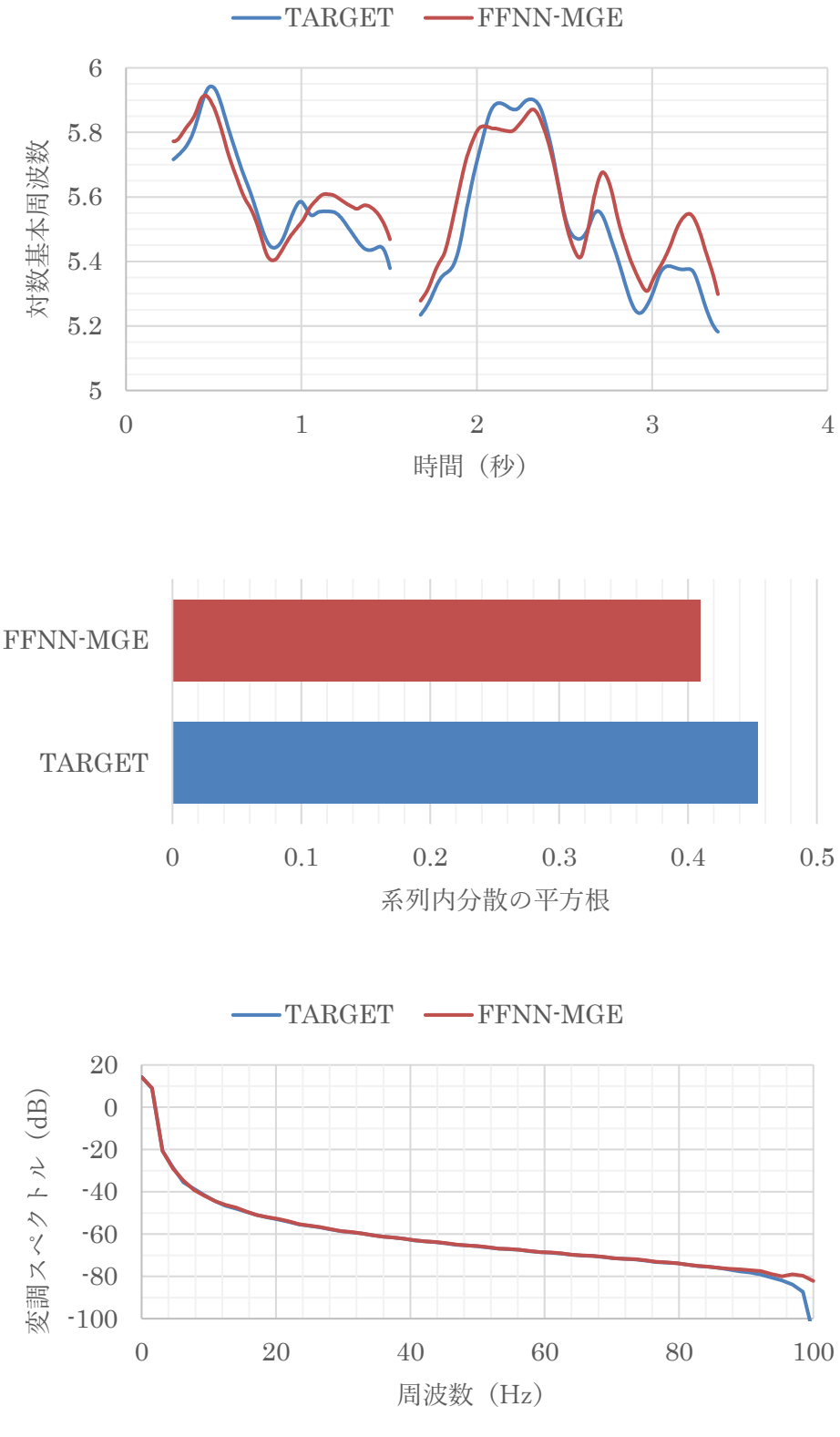

図 5.4　FFNN-MGE の対数基本周波数の代表例

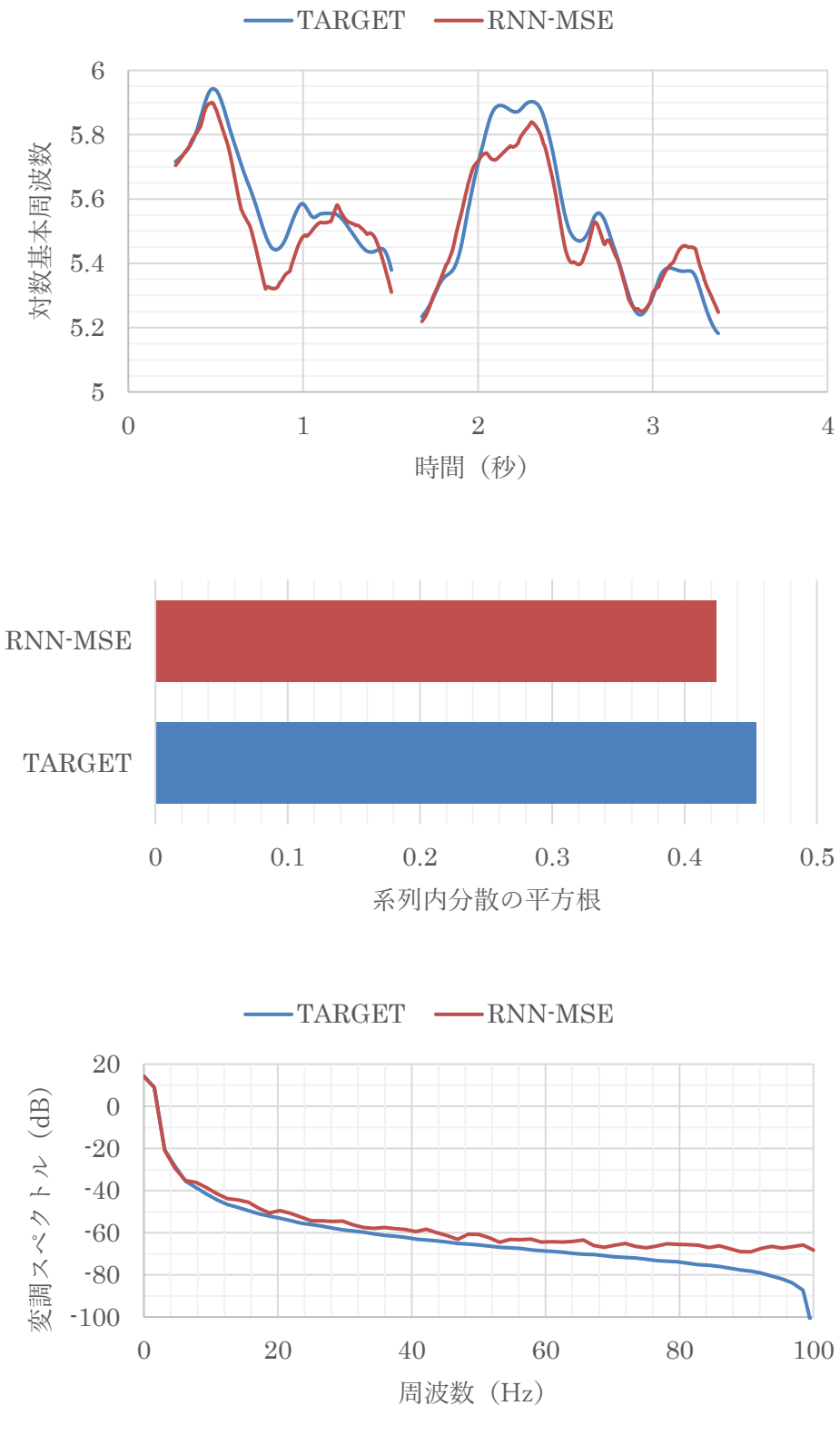

図 5.5　RNN-MSE の対数基本周波数の代表例

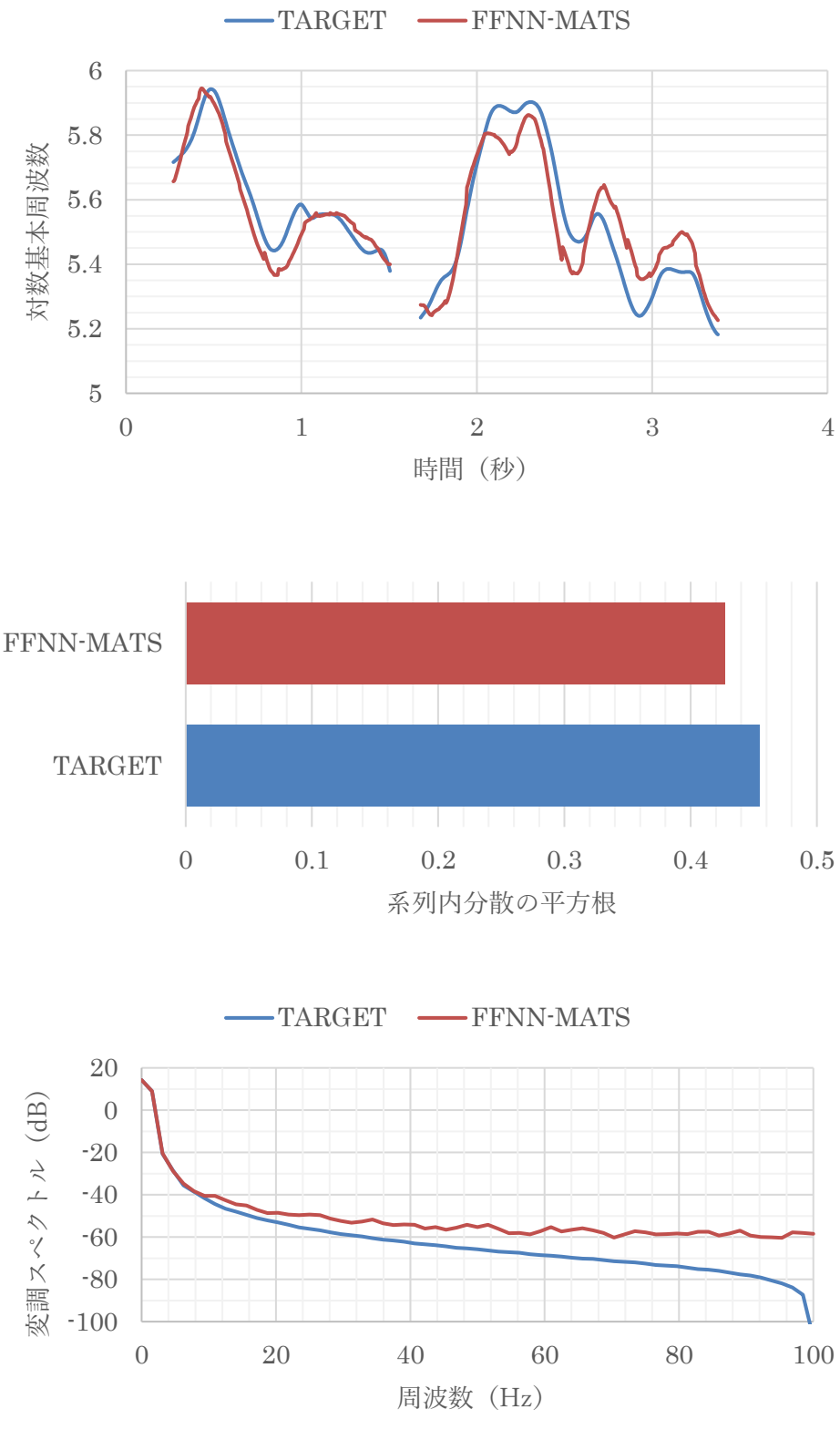

図 5.6　FFNN-MATS の対数基本周波数の代表例

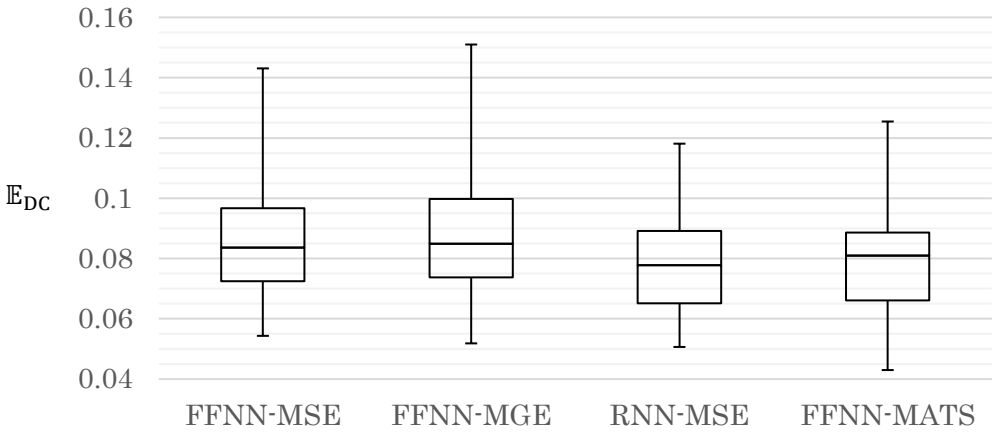

図 5.7　対数基本周波数の$\mathbb{E}_{DC}$

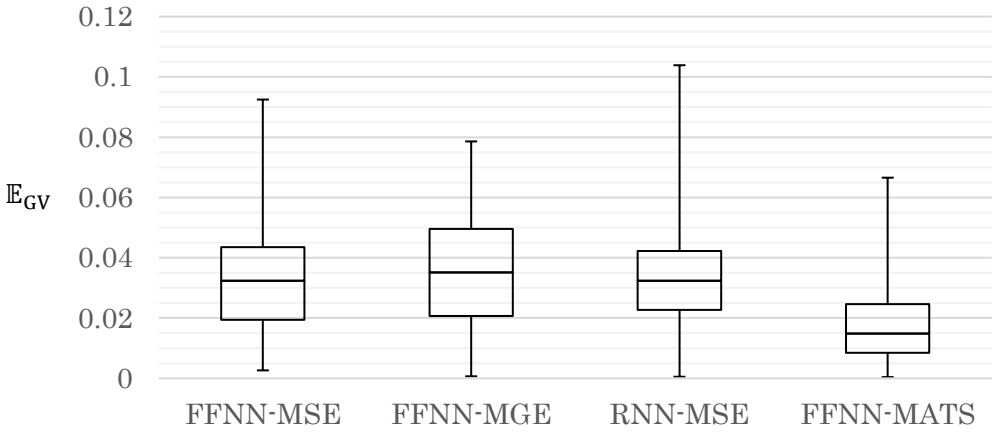

図 5.8　対数基本周波数の$\mathbb{E}_{GV}$

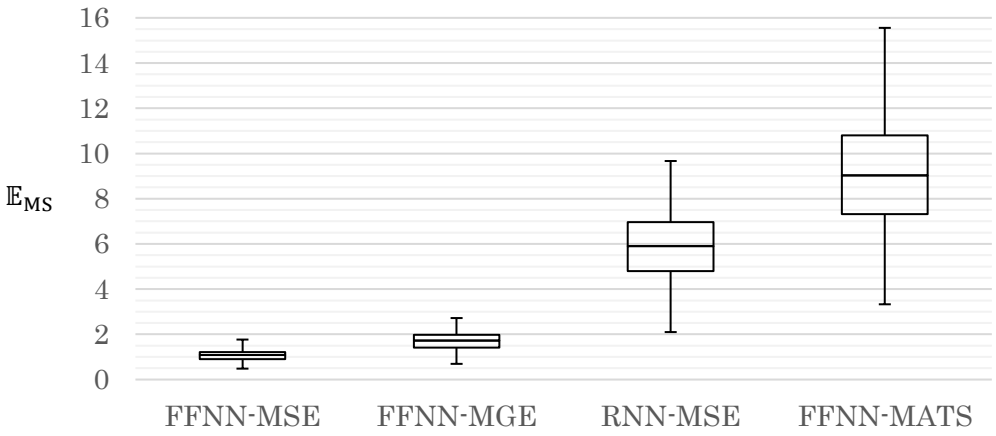

図 5.9　対数基本周波数の$\mathbb{E}_{MS}$（dB）

表 5.4　Tukey-Kramer 法による対数基本周波数の$\mathbb{E}_{DC}$の平均値の比較結果
表中の数値はスチューデント化された範囲分布の q 値と p 値である．群数は 4，自由度は 396，信頼区間は 95%である．

群 1	群 2	q 値	p 値
FFNN-MSE	FFNN-MGE	0.50	0.900
FFNN-MSE	RNN-MSE	4.53	0.008
FFNN-MSE	FFNN-MATS	3.30	0.092
FFNN-MGE	RNN-MSE	5.03	0.002
FFNN-MGE	FFNN-MATS	3.80	0.037
RNN-MSE	FFNN-MATS	1.23	0.798

表 5.5　Tukey-Kramer 法による対数基本周波数の$\mathbb{E}_{GV}$の平均値の比較結果
表中の数値はスチューデント化された範囲分布の q 値と p 値である．群数は 4，自由度は 396，信頼区間は 95%である．

群 1	群 2	q 値	p 値
FFNN-MSE	FFNN-MGE	1.75	0.592
FFNN-MSE	RNN-MSE	0.61	0.900
FFNN-MSE	FFNN-MATS	7.80	0.001
FFNN-MGE	RNN-MSE	1.14	0.833
FFNN-MGE	FFNN-MATS	9.55	0.001
RNN-MSE	FFNN-MATS	8.41	0.001

表 5.6　Tukey-Kramer 法による対数基本周波数の$\mathbb{E}_{MS}$の平均値の比較結果
表中の数値はスチューデント化された範囲分布の q 値と p 値である．群数は 4，自由度は 396，信頼区間は 95%である．

群 1	群 2	q 値	p 値
FFNN-MSE	FFNN-MGE	4.57	0.007
FFNN-MSE	RNN-MSE	33.89	0.001
FFNN-MSE	FFNN-MATS	56.94	0.001
FFNN-MGE	RNN-MSE	29.32	0.001
FFNN-MGE	FFNN-MATS	52.37	0.001
RNN-MSE	FFNN-MATS	23.05	0.001

5.6. メルケプストラムについての実験結果

MATS 損失関数はモデル化する音声特徴量を適切に捉えるために，対象とする音声特徴量ごとに各損失関数のパラメータの設定を調整する必要がある．そのため，まず，MATS 損失関数の各損失関数のパラメータの設定の調整法について述べる．次に，メルケプストラムに適した設定をした MATS 損失関数を用いたときの聴取実験と予測誤差について述べる．

5.6.1. MATS 損失関数のパラメータ設定

メルケプストラムのモデル化において利用可能な損失関数は DC 損失関数，TD 損失関数，DD 損失関数，GV 損失関数，GC 損失関数，LV 損失関数，LC 損失関数である．メルケプストラムに対するこれらの損失関数の挙動を確認するため，約 100 通りの損失関数のパラメータの組み合わせを試した．試した損失関数のパラメータの組み合わせの一部の結果を図 5.19 に示す．1 列目の「条件」の見出しは表 5.12 と対応する．DC1 が示すように，DC 損失関数だけでは，複雑な時間構造を持つメルケプストラムを予測する DNN のモデルパラメータを獲得できない．TD1 から TD3 が示すように，TD 損失関数を使用すると，平滑化されたメルケプストラムを予測する DNN のモデルパラメータが学習される．また，w_2 が大きくなるにつれて，$\mathbb{E}_{\mathrm{GV}}$や$\mathbb{E}_{\mathrm{MS}}$も大きくなった．GV1 から GV3 が示すように，GV 損失関数を使用すると，メルケプストラムの GV を考慮した DNN のモデルパラメータが学習される．また，ω_{GV}が大きくなるにつれて，$\mathbb{E}_{\mathrm{GV}}$や$\mathbb{E}_{\mathrm{MS}}$は小さくなったが，$\mathbb{E}_{\mathrm{DC}}$は大きくなった．LV1 から LV12 が示すように，LV 損失関数を使用すると，複雑な時間構造を持つメルケプストラムを予測する DNN のモデルパラメータが学習される．L_{LV}からR_{LV}までの範囲やω_{LV}を大きくするにつれて，$\mathbb{E}_{\mathrm{GV}}$や$\mathbb{E}_{\mathrm{MS}}$は小さくなったが，適切な値にしないと，自然な時間構造を無視して，不自然に振動するだけのメルケプストラムを予測する DNN のモデルパラメータが学習されてしまう．GC1 から GC3 が示すように，GC 損失関数を使用しても予測されるメルケプストラムはほとんど変化しなかった．LC1 から LC12 が示すように，LC 損失関数を使用しても予測されるメルケプストラムはほとんど変化しなかった．DD1 から DD3 が示すように，DD 損失関数を使用しても予測されるメルケプストラムはほとんど変化しなかった．GC 損失関数，LC 損失関数，DD 損失関数は DC 損失関数との組み合わせではほとんど機能しなかったが，GV 損失関数や LV 損失関数と組み合わせることで，GV 損失関数を使用することで生じる局所的なパワーの増大や，LV 損失関数を使用することで生じる不自然な振動を抑制する効果はあった．これらの損失関数は，能動的に機能せず，他の損失関数の制約として機能する．

メルケプストラムのモデル化においては，統計手法による過剰平滑の問題を回避するために，メルケプストラムの系列内分散が小さくならないようにすること，高次のメルケプストラムが滑らかに変化しすぎないようにすることに注意した．この方針と図 5.19 の実験で得た知見に基づいて，GV 損失関数と LV 損失関数が MATS 損失関数の中核となるようにパラメータの調整を行った．パラメータは試行錯誤を繰り返すことで調整した．DNN を学

習し，数個の合成音声を聴いて音質を確認するということを繰り返し行った．音質に不具合が生じたら，その原因と考えられるパラメータの値の調整や，その不具合を抑制すると考えられる損失関数を追加した．メルケプストラムについては DC 損失関数と GV 損失関数である程度良好な DNN のモデルパラメータを学習できるが，これ以上の音質の改善を行う場合は LV 損失関数が必要になる．しかし，LV 損失関数を使用するとメルケプストラムが不自然に振動してしまい，合成音声の音質を劣化させてしまう．これを抑制するために，LC 損失関数や DD 損失関数を使用した．GC 損失関数は LC 損失関数よりも過剰平滑を招きやすかったため使用しなかった．これはメルケプストラムの系列全体の分布が正規分布に従わないためだと考える．LV 損失関数による不具合を改善するために，最終的に，TD 損失関数も利用した．TD 損失関数はメルケプストラムが滑らかに変化するようにするため，LV 損失関数とは対照的である．しかし，LV 損失関数ではメルケプストラムの自然な時間構造を捉えることが困難であったため，不自然に振動するような時間構造にならないように TD 損失関数を使用することにした．最終的に，メルケプストラムを MATS 損失関数で学習するときのパラメータは$\omega_{\mathrm{TD}}=2$，$L=-1$，$R=0$，$w_1=1$，$w_2=2$，$\omega_{\mathrm{DD}}=2$，$\omega_{\mathrm{GV}}=1$，$\omega_{\mathrm{LV}}=3$，$L_{\mathrm{LV}}=-4$，$R_{\mathrm{LV}}=4$，$\omega_{\mathrm{LC}}=3$，$L_{\mathrm{LC}}=-4$，$R_{\mathrm{LC}}=4$となった．

5.6.2. 聴取実験の結果

MUSHRA 法による聴取実験の結果を図 5.10 に示す．実験に用いた合成音声は 2.1.2 のボコーダの合成部で生成した．FFNN-MSE，FFNN-MGE，RNN-MSE，FFNN-MATS の音声は各音声特徴量予測部で予測したメルケプストラムと，そのメルケプストラムに対応する原音声の基本周波数と非周期性指標から合成した．参照音声は原音声の分析再合成音声である．アンカー音声は FFNN-MATS において DC 損失関数のみで学習した FFNN で予測したメルケプストラムと，そのメルケプストラムに対応する原音声の基本周波数と非周期性指標から合成した．参加者は合成音声の韻律や音質の違いに敏感な 10 名である．合成音声の音質を評価するため，合成音声のアクセントや抑揚に注目しないように指示をした．セッション数は 100 であるため，参加者を適宜休憩させた．

図 5.10 の評点を Tukey-Kramer 法によって比較した結果を表 5.7 に示す．その結果，FFNN-MSE 群と FFNN-MGE 群，FFNN-MSE 群と RNN-MSE 群，FFNN-MGE 群と RNN-MSE 群の評点には有意差が認められなかった．FFNN-MSE 群，FFNN-MGE 群，RNN-MSE 群の音声はどれも同じような音質であり，ケプストラム強調によって過剰平滑による音質の問題は解決したが，ダウンサンプリングした音声のように音声の帯域が狭まったような音質であった．FFNN-MATS の音声の音質は，FFNN-MSE 群，FFNN-MGE 群，RNN-MSE 群の音声よりも音声の帯域が広がったような音質だった．対数基本周波数の聴取実験では，評価群の音声の中には参照群の音声と同じ程度の品質のものがあったが，メルケプストラムの聴取実験では，各評価群の音声と参照群の音声との音質の差は大きかった．この結果は，メルケプストラムの学習法には改善の余地があることを示している．

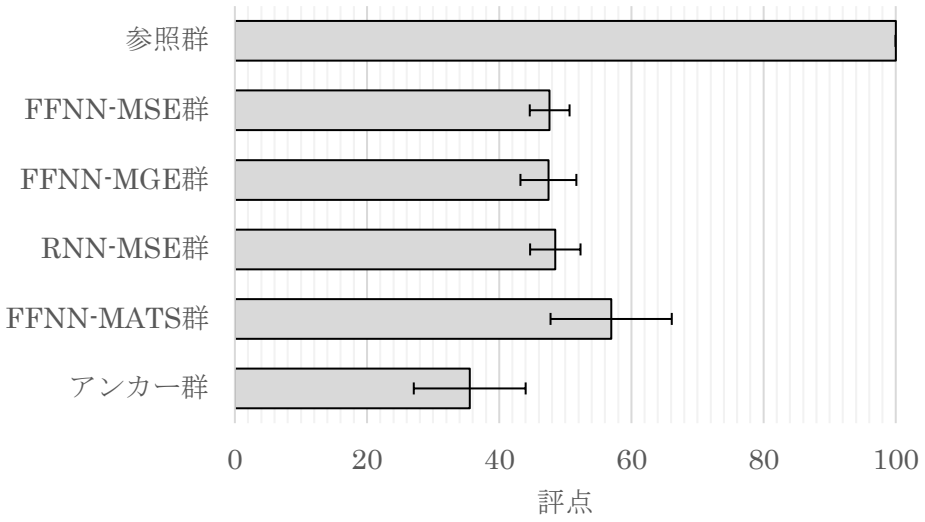

図 5.10 MUSHRA 法による合成音声の音質についての聴取実験の結果

表 5.7　Tukey-Kramer 法による聴取実験の𝕍の平均値の比較結果
表中の数値はスチューデント化された範囲分布の q 値と p 値である．群数は 6，自由度は 54，信頼区間は 95%である．

群 1	群 2	q 値	p 値
参照群	FFNN-MSE 群	53.70	0.001
参照群	FFNN-MGE 群	53.89	0.001
参照群	RNN-MSE 群	52.82	0.001
参照群	FFNN-MATS 群	44.15	0.001
参照群	アンカー群	66.10	0.001
FFNN-MSE 群	FFNN-MGE 群	0.19	0.900
FFNN-MSE 群	RNN-MSE 群	0.88	0.900
FFNN-MSE 群	FFNN-MATS 群	9.54	0.001
FFNN-MSE 群	アンカー群	12.40	0.001
FFNN-MGE 群	RNN-MSE 群	1.07	0.900
FFNN-MGE 群	FFNN-MATS 群	9.74	0.001
FFNN-MGE 群	アンカー群	12.21	0.001
RNN-MSE 群	FFNN-MATS 群	8.67	0.001
RNN-MSE 群	アンカー群	13.28	0.001
FFNN-MATS 群	アンカー群	21.95	0.001

5.6.3. 予測誤差の結果

各音声特徴量予測部で予測したメルケプストラムの代表例を図 5.11 から図 5.14 までに示す．いずれの 15 次のメルケプストラムも原音声の 15 次のメルケプストラムのような複雑な時間構造を再現するには至っていないが，概ね形状は一致していた．いずれの 15 次のメルケプストラムの系列内分散の平方根も，原音声の 15 次のメルケプストラムの系列内分散の平方根よりも約 0.01～0.02 小さかった．ただし，FFNN-MSE，FFNN-MGE，RNN-MSE のメルケプストラムはケプストラム強調により係数を 1.4 倍されているため，ケプストラム強調前の系列内分散の平方根は図示されたものより約1.4^{-1}倍小さいことになる．ケプストラム強調がなくても FFNN-MATS はケプストラム強調を適用したメルケプストラムと同等の系列内分散を持つメルケプストラムを予測できたといえる．また，いずれの 15 次のメルケプストラムの変調スペクトルも 16 Hz 以上から徐々に原音声の 15 次のメルケプストラムの変調スペクトルとの差が大きくなり，その差は最大で約 10～15 dB となった．

各音声特徴量予測部で予測したメルケプストラムの$\mathbb{U}_s$についての$\mathbb{E}_{DC}$，$\mathbb{E}_{GV}$，$\mathbb{E}_{MS}$をそれぞれ，図 5.15，図 5.16，図 5.17 に示す．また，これらのメルケプストラムの$\mathbb{U}_s$についての$\mathbb{E}_{DC}$，$\mathbb{E}_{GV}$，$\mathbb{E}_{MS}$の平均値を Tukey-Kramer 法で比較した結果を表 5.8，表 5.9，表 5.10 に示す．FFNN-MATS の$\mathbb{E}_{DC}$の平均値は，FFNN-MSE，FFNN-MGE，RNN-MSE の$\mathbb{E}_{DC}$の平均値よりも有意に大きかった．FFNN-MATS の$\mathbb{E}_{DC}$の中央値と FFNN-MSE，FFNN-MGE，RNN-MSE の$\mathbb{E}_{DC}$の中央値との差は約 0.007 以下であり，ケプストラム係数の値や聴取実験の評点を考慮すると，これらの差は合成音声において無視できる程度のものである．

FFNN-MATS の$\mathbb{E}_{GV}$の平均値は，FFNN-MSE，FFNN-MGE，RNN-MSE の$\mathbb{E}_{GV}$の平均値よりも有意に小さかった．FFNN-MATS の$\mathbb{E}_{GV}$の中央値と FFNN-MSE，FFNN-MGE，RNN-MSE の$\mathbb{E}_{GV}$の中央値との差は約 0.07 であり，ケプストラム係数の値や聴取実験の評点を考慮すると，これらの差は合成音声の音質に影響する程度のものである．

FFNN-MATS の$\mathbb{E}_{MS}$の平均値は，FFNN-MSE，FFNN-MGE，RNN-MSE の$\mathbb{E}_{MS}$の平均値よりも有意に小さかった．FFNN-MATS の$\mathbb{E}_{MS}$の中央値は，FFNN-MSE，FFNN-MGE，RNN-MSE の$\mathbb{E}_{MS}$の中央値よりもそれぞれ約 7 dB，約 3.5 dB，約 3 dB 小さかった．これは，局所内分散を明示的に学習したことによるものである．聴取実験の評点を考慮すると，これらの差は，FFNN-MATS の合成音声の品質に影響を与える程度のものである．

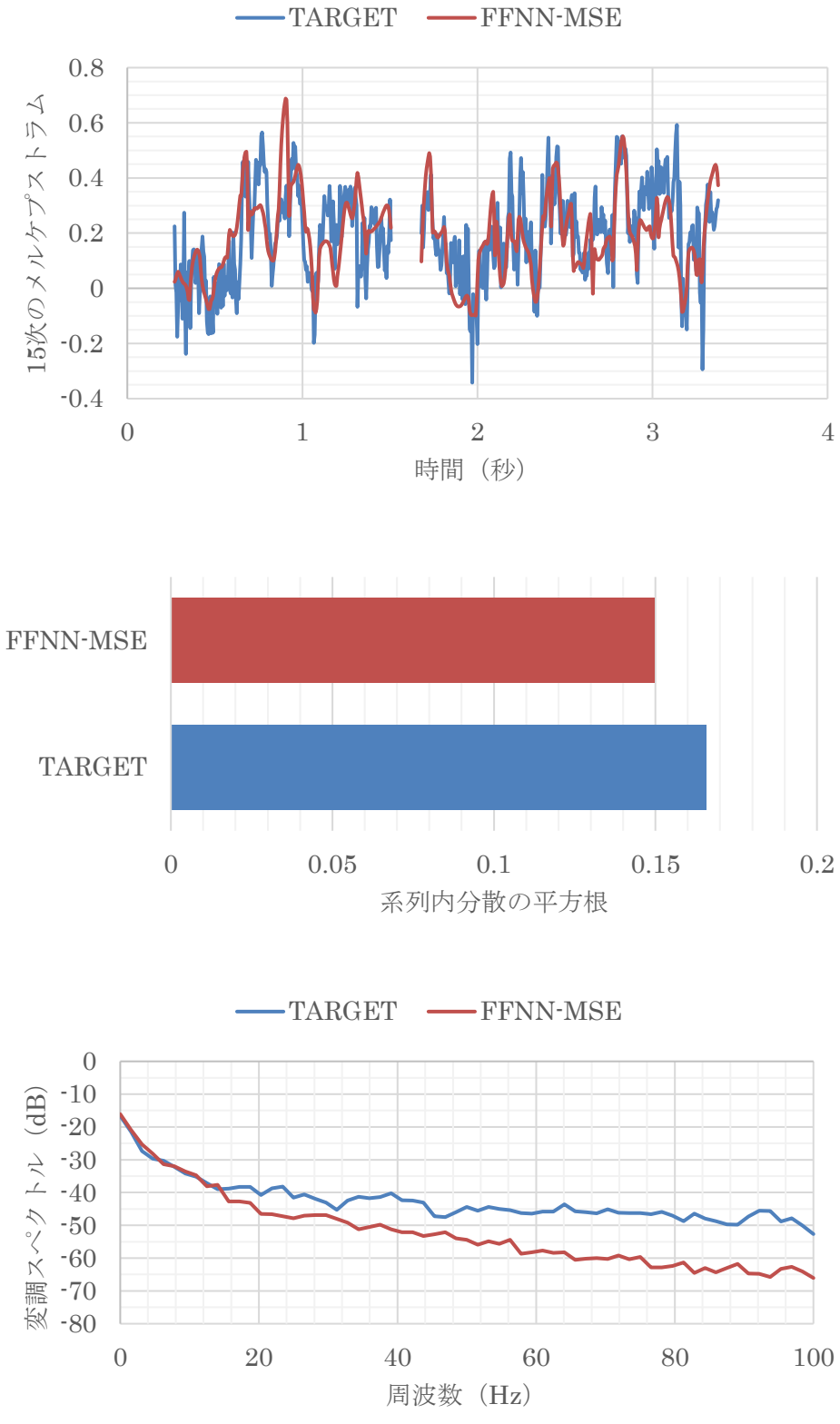

図 5.11 FFNN-MSE のメルケプストラムの代表例

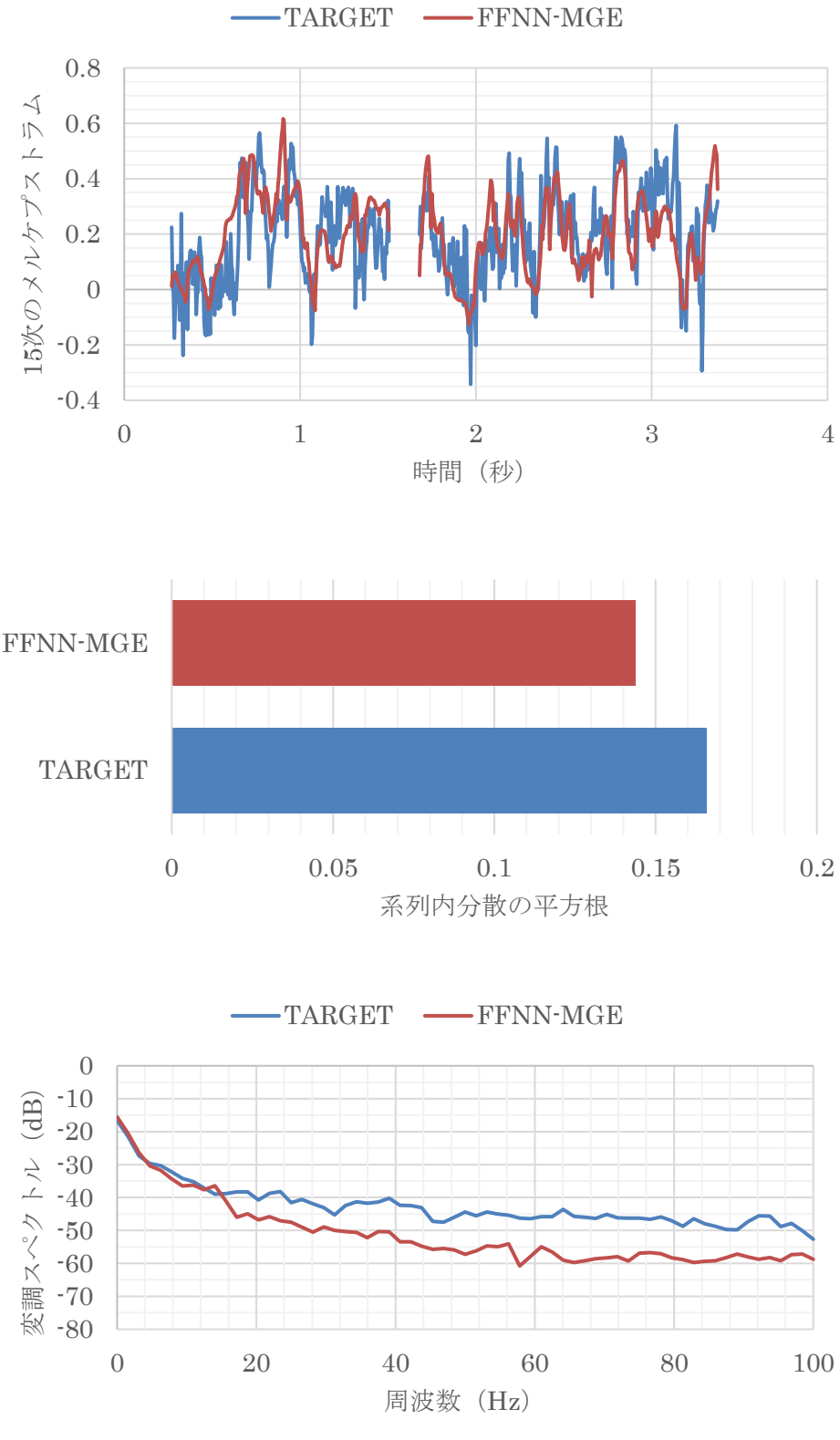

図 5.12 FFNN-MGE のメルケプストラムの代表例

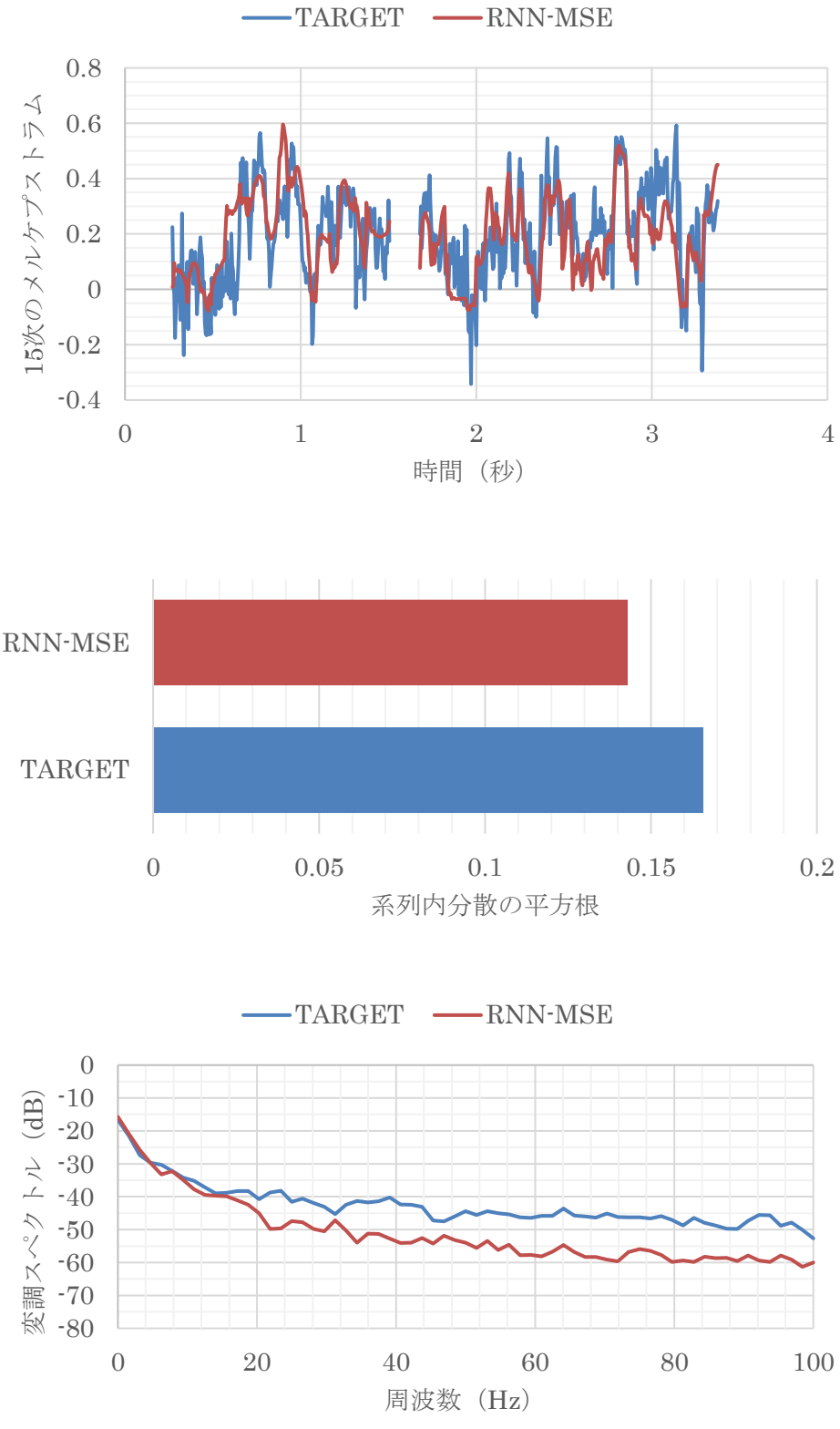

図 5.13 RNN-MSE のメルケプストラムの代表例

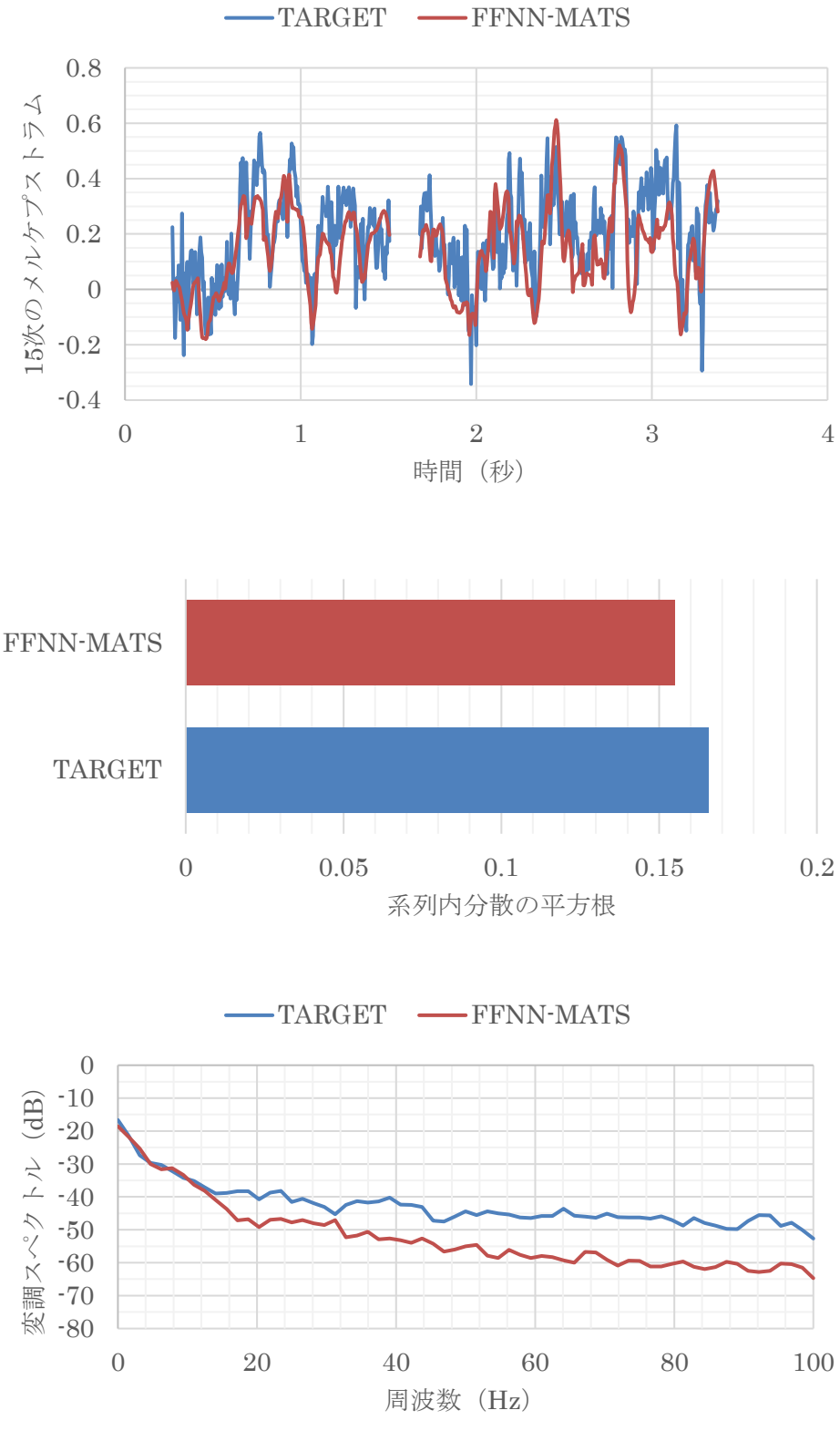

図 5.14 FFNN-MATS のメルケプストラムの代表例

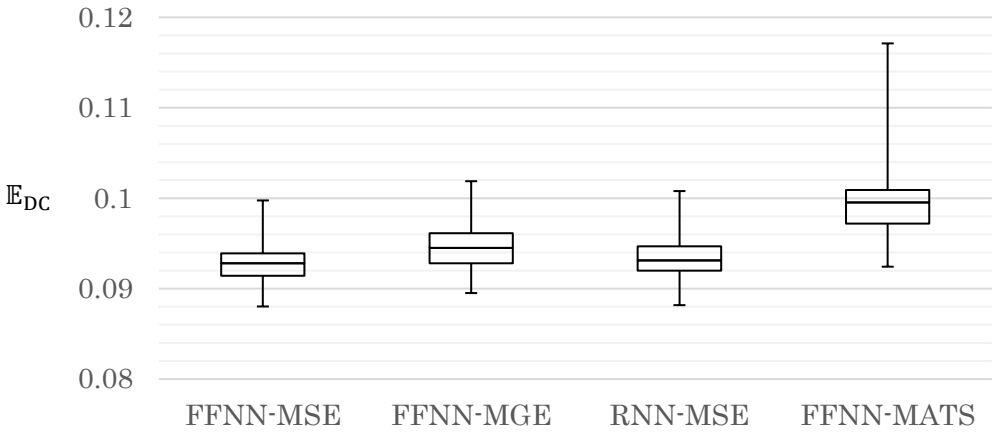

図 5.15 メルケプストラムの平均絶対誤差

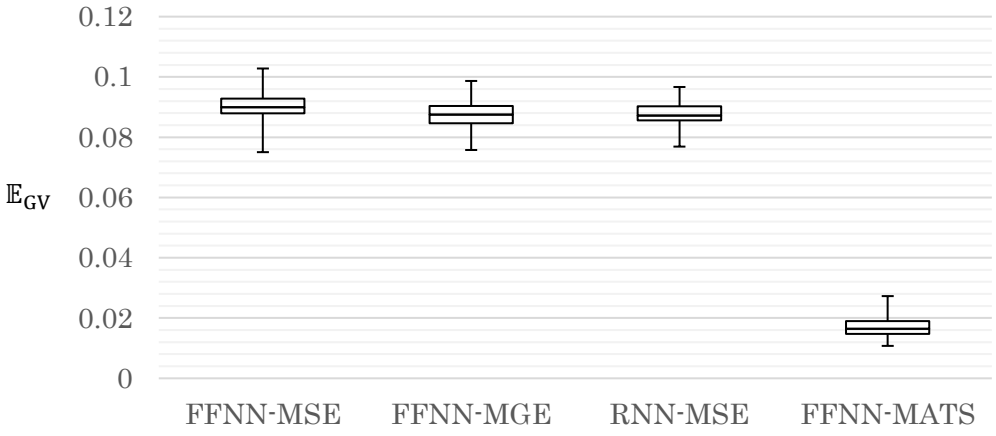

図 5.16 メルケプストラムの系列内分散の平方根の平均絶対誤差

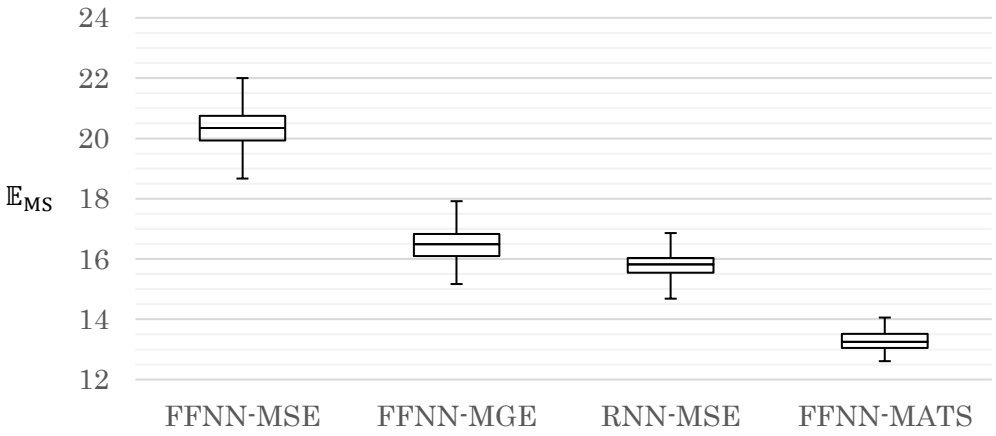

図 5.17 メルケプストラムの変調スペクトルの平均絶対誤差（dB）

表 5.8　Tukey-Kramer 法によるメルケプストラムの$\mathbb{E}_{\mathrm{DC}}$の平均値の比較結果
表中の数値はスチューデント化された範囲分布の q 値と p 値である．群数は 4，自由度は 396，信頼区間は 95%である．

群 1	群 2	q 値	p 値
FFNN-MSE	FFNN-MGE	6.88	0.001
FFNN-MSE	RNN-MSE	1.74	0.595
FFNN-MSE	FFNN-MATS	25.36	0.001
FFNN-MGE	RNN-MSE	5.14	0.002
FFNN-MGE	FFNN-MATS	18.48	0.004
RNN-MSE	FFNN-MATS	23.62	0.005

表 5.9　Tukey-Kramer 法によるメルケプストラムの$\mathbb{E}_{\mathrm{GV}}$の平均値の比較結果
表中の数値はスチューデント化された範囲分布の q 値と p 値である．群数は 4，自由度は 396，信頼区間は 95%である．

群 1	群 2	q 値	p 値
FFNN-MSE	FFNN-MGE	8.83	0.001
FFNN-MSE	RNN-MSE	9.04	0.001
FFNN-MSE	FFNN-MATS	157.19	0.001
FFNN-MGE	RNN-MSE	0.21	0.900
FFNN-MGE	FFNN-MATS	148.36	0.001
RNN-MSE	FFNN-MATS	148.15	0.001

表 5.10　Tukey-Kramer 法によるメルケプストラムの$\mathbb{E}_{\mathrm{MS}}$の平均値の比較結果
表中の数値はスチューデント化された範囲分布の q 値と p 値である．群数は 4，自由度は 396，信頼区間は 95%である．

群 1	群 2	q 値	p 値
FFNN-MSE	FFNN-MGE	77.01	0.001
FFNN-MSE	RNN-MSE	90.70	0.001
FFNN-MSE	FFNN-MATS	141.21	0.001
FFNN-MGE	RNN-MSE	13.69	0.001
FFNN-MGE	FFNN-MATS	64.19	0.001
RNN-MSE	FFNN-MATS	50.50	0.001

5.7. 考察

FFNN でも時間構造を捉えて音声特徴量系列を生成できるようにするために，時系列の複数の属性を考慮した新しい学習法を提案した．従来の学習法と比較した結果，FFNN であっても提案法を用いることで，知覚的に優れた対数基本周波数系列やメルケプストラム系列を生成できることが示された．

各音声特徴量予測部で予測した対数基本周波数はいずれも滑らかであったが，合成音声の韻律の品質についての聴取実験の評点には差があった．この差は，対数基本周波数の動的特徴量を明示的に学習したかによって生じたと考える．対数基本周波数の動的特徴量は，基本周波数の相対的変化を表す特徴量であり，日本語の音高の知覚に深く関与している．FFNN-MATS と FFNN-MSE は対数基本周波数の動的特徴量を直接学習し，FFNN-MGE は MLPG を介して対数基本周波数の動的特徴量を間接的に学習し，RNN-MSE は対数基本周波数の動的特徴量を学習しない．対数基本周波数のモデル化においては，対数基本周波数の動的特徴量を明示的かつ重点的に学習することが，合成音声の韻律の品質向上に大きく寄与している．

FFNN-MSE，FFNN-MGE，RNN-MSE は，モデル構造や学習法が異なっているにも関わらず，変調スペクトルの誤差を除いた予測誤差と聴取実験の評点はほぼ同じであった．従来の損失関数は，時間フレームごとの音声特徴量の二乗誤差のみであり，音声特徴量の構造を捉えるには不十分である．RNN-MSE の実験結果が示すように，単純な構造である対数基本周波数であれば，従来の損失関数であっても，RNN の再帰構造によってその構造を捉えることができる．一方で，複雑な構造であるメルケプストラムでは，従来の損失関数でも RNN の再帰構造でもその構造を捉えることができない．DNN のモデルパラメータは損失関数が算出した誤差に基づいて更新されるため，音声特徴量の構造を捉えた DNN のモデルパラメータを得るには，複数の損失関数で音声特徴量の多角的な誤差を算出することが重要である．FFNN-MATS は MATS 損失関数でメルケプストラムの多角的な誤差を算出したため，FFNN-MSE，FFNN-MGE，RNN-MSE よりも知覚的に優れたメルケプストラムを予測する FFNN のモデルパラメータを学習できた．

MATS 損失関数のパラメータを調整する際に得た知見を述べる．1 つめは，数学的には GC の対角成分は GV と同じであるにも関わらず，GC 損失関数は GV 損失関数を内包するようには機能しないことである．これは GC の大半を占める非対角成分の損失が，対角成分の損失よりも支配的になるためである．このことは，LC 損失関数と LV 損失関数にも同様のことがいえる．2 つめは，音高や音質の知覚に関する特徴量を重点的に学習することである．対数基本数は数の動的特徴量を明示的かつ重点的に学習することで，合成音声の韻律の品質は向上した．また，メルケプストラムでは GV や LV に重点を置き学習することとで，合成音声の音質が向上したが，未だに原音声の分析再合成音声との差は大きかった．メルケプストラムのどの特徴が音質の知覚に深く関係しているかを発見し，損失関数に組み込むことが合成音声の音質を向上させる鍵になる．

MATS 損失関数の戦略は，経験や知見に基づく規則を損失関数に組み込み，その規則を明示的かつ直接的に学習するものである．これにより，比較的に容易に合成音声の品質を向上させることができる．この戦略は DNN が入力と出力の関係を自動で獲得するという一般的な戦略に反するものである．しかし，一般的な学習法では，DNN の構造やハイパーパラメータと音声特徴量の関係は直接的に関連付かないため，各音声特徴量に最適なハイパーパラメータを調整することは難しい．一方で，MATS 損失関数のパラメータは音声特徴量と直接的に関連付くため理解しやすい．MATS 損失関数は，音声特徴量を学習する際に経験や知見に基づく規則があれば，それを直接的かつ明示的に学習できる利点があるといえる．

5.8. まとめ

FFNN が時系列の時間構造を考慮したモデルパラメータを獲得できるようにするため，時系列の複数の属性を考慮した損失関数による FFNN の学習法を提案した．対数基本周波数とメルケプストラムを対象として，合成音声の韻律と音質を評価する聴取実験と，対数基本周波数とメルケプストラムの予測誤差により，提案法と従来法を比較した．その結果，提案法は従来法と同等以上の知覚的に優れた対数基本周波数やメルケプストラムの予測を可能にした．これにより，3.2.3 で述べた計算資源が限られた音声特徴量予測部の FFNN による合成音声の音質の問題を解決した．

表 5.11 対数基本周波数についての MATS 損失関数の各損失関数の挙動を確認したときのパラメータの組み合わせ

条件	パラメータ一覧（表記のないものは使用していない）
DC1	$\omega_{\mathrm{DC}} = 1$
TD1	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{TD}} = 1, w_2 = 1, L_{\mathrm{TD}} = -1, R_{\mathrm{TD}} = 0$　（TD6 との比較用．DC 損失関数と TD 損失関数の併用は禁止）
TD2	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{TD}} = 1, w_2 = 5, L_{\mathrm{TD}} = -1, R_{\mathrm{TD}} = 0$　（TD7 との比較用．DC 損失関数と TD 損失関数の併用は禁止）
TD3	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{TD}} = 1, w_2 = 10, L_{\mathrm{TD}} = -1, R_{\mathrm{TD}} = 0$　（TD8 との比較用．DC 損失関数と TD 損失関数の併用は禁止）
TD4	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{TD}} = 1, w_2 = 15, L_{\mathrm{TD}} = -1, R_{\mathrm{TD}} = 0$　（TD9 との比較用．DC 損失関数と TD 損失関数の併用は禁止）
TD5	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{TD}} = 1, w_2 = 20, L_{\mathrm{TD}} = -1, R_{\mathrm{TD}} = 0$　（TD10 との比較用．DC 損失関数と TD 損失関数の併用は禁止）
TD6	$\omega_{\mathrm{TD}} = 1, w_1 = 1, w_2 = 1, L_{\mathrm{TD}} = -1, R_{\mathrm{TD}} = 0$
TD7	$\omega_{\mathrm{TD}} = 1, w_1 = 1, w_2 = 5, L_{\mathrm{TD}} = -1, R_{\mathrm{TD}} = 0$
TD8	$\omega_{\mathrm{TD}} = 1, w_1 = 1, w_2 = 10, L_{\mathrm{TD}} = -1, R_{\mathrm{TD}} = 0$
TD9	$\omega_{\mathrm{TD}} = 1, w_1 = 1, w_2 = 15, L_{\mathrm{TD}} = -1, R_{\mathrm{TD}} = 0$
TD10	$\omega_{\mathrm{TD}} = 1, w_1 = 1, w_2 = 20, L_{\mathrm{TD}} = -1, R_{\mathrm{TD}} = 0$
GV1	$\omega_{\mathrm{TD}} = 1, w_1 = 1, w_2 = 20, L_{\mathrm{TD}} = -1, R_{\mathrm{TD}} = 0, \omega_{\mathrm{GV}} = 1$
GV2	$\omega_{\mathrm{TD}} = 1, w_1 = 1, w_2 = 20, L_{\mathrm{TD}} = -1, R_{\mathrm{TD}} = 0, \omega_{\mathrm{GV}} = 2$
GV3	$\omega_{\mathrm{TD}} = 1, w_1 = 1, w_2 = 20, L_{\mathrm{TD}} = -1, R_{\mathrm{TD}} = 0, \omega_{\mathrm{GV}} = 4$
GV4	$\omega_{\mathrm{TD}} = 1, w_1 = 1, w_2 = 20, L_{\mathrm{TD}} = -1, R_{\mathrm{TD}} = 0, \omega_{\mathrm{GV}} = 8$
LV1	$\omega_{\mathrm{TD}} = 1, w_1 = 1, w_2 = 20, L_{\mathrm{TD}} = -1, R_{\mathrm{TD}} = 0, \omega_{\mathrm{LV}} = 1, L_{\mathrm{LV}} = -4, R_{\mathrm{LV}} = 4$
LV2	$\omega_{\mathrm{TD}} = 1, w_1 = 1, w_2 = 20, L_{\mathrm{TD}} = -1, R_{\mathrm{TD}} = 0, \omega_{\mathrm{LV}} = 1, L_{\mathrm{LV}} = -8, R_{\mathrm{LV}} = 8$
LV3	$\omega_{\mathrm{TD}} = 1, w_1 = 1, w_2 = 20, L_{\mathrm{TD}} = -1, R_{\mathrm{TD}} = 0, \omega_{\mathrm{LV}} = 1, L_{\mathrm{LV}} = -12, R_{\mathrm{LV}} = 12$
LV4	$\omega_{\mathrm{TD}} = 1, w_1 = 1, w_2 = 20, L_{\mathrm{TD}} = -1, R_{\mathrm{TD}} = 0, \omega_{\mathrm{LV}} = 2, L_{\mathrm{LV}} = -4, R_{\mathrm{LV}} = 4$
LV5	$\omega_{\mathrm{TD}} = 1, w_1 = 1, w_2 = 20, L_{\mathrm{TD}} = -1, R_{\mathrm{TD}} = 0, \omega_{\mathrm{LV}} = 2, L_{\mathrm{LV}} = -8, R_{\mathrm{LV}} = 8$

表 5.11 対数基本周波数についての MATS 損失関数の各損失関数の挙動を確認したときのパラメータの組み合わせ

LV6	$\omega_{\mathrm{TD}}=1, w_1=1, w_2=20, L_{\mathrm{TD}}=-1, R_{\mathrm{TD}}=0, \omega_{\mathrm{LV}}=2, L_{\mathrm{LV}}=-12, R_{\mathrm{LV}}=12$
LV7	$\omega_{\mathrm{TD}}=1, w_1=1, w_2=20, L_{\mathrm{TD}}=-1, R_{\mathrm{TD}}=0, \omega_{\mathrm{LV}}=4, L_{\mathrm{LV}}=-4, R_{\mathrm{LV}}=4$
LV8	$\omega_{\mathrm{TD}}=1, w_1=1, w_2=20, L_{\mathrm{TD}}=-1, R_{\mathrm{TD}}=0, \omega_{\mathrm{LV}}=4, L_{\mathrm{LV}}=-8, R_{\mathrm{LV}}=8$
LV9	$\omega_{\mathrm{TD}}=1, w_1=1, w_2=20, L_{\mathrm{TD}}=-1, R_{\mathrm{TD}}=0, \omega_{\mathrm{LV}}=4, L_{\mathrm{LV}}=-12, R_{\mathrm{LV}}=12$
LV10	$\omega_{\mathrm{TD}}=1, w_1=1, w_2=20, L_{\mathrm{TD}}=-1, R_{\mathrm{TD}}=0, \omega_{\mathrm{LV}}=8, L_{\mathrm{LV}}=-4, R_{\mathrm{LV}}=4$
LV11	$\omega_{\mathrm{TD}}=1, w_1=1, w_2=20, L_{\mathrm{TD}}=-1, R_{\mathrm{TD}}=0, \omega_{\mathrm{LV}}=8, L_{\mathrm{LV}}=-8, R_{\mathrm{LV}}=8$
LV12	$\omega_{\mathrm{TD}}=1, w_1=1, w_2=20, L_{\mathrm{TD}}=-1, R_{\mathrm{TD}}=0, \omega_{\mathrm{LV}}=8, L_{\mathrm{LV}}=-12, R_{\mathrm{LV}}=12$

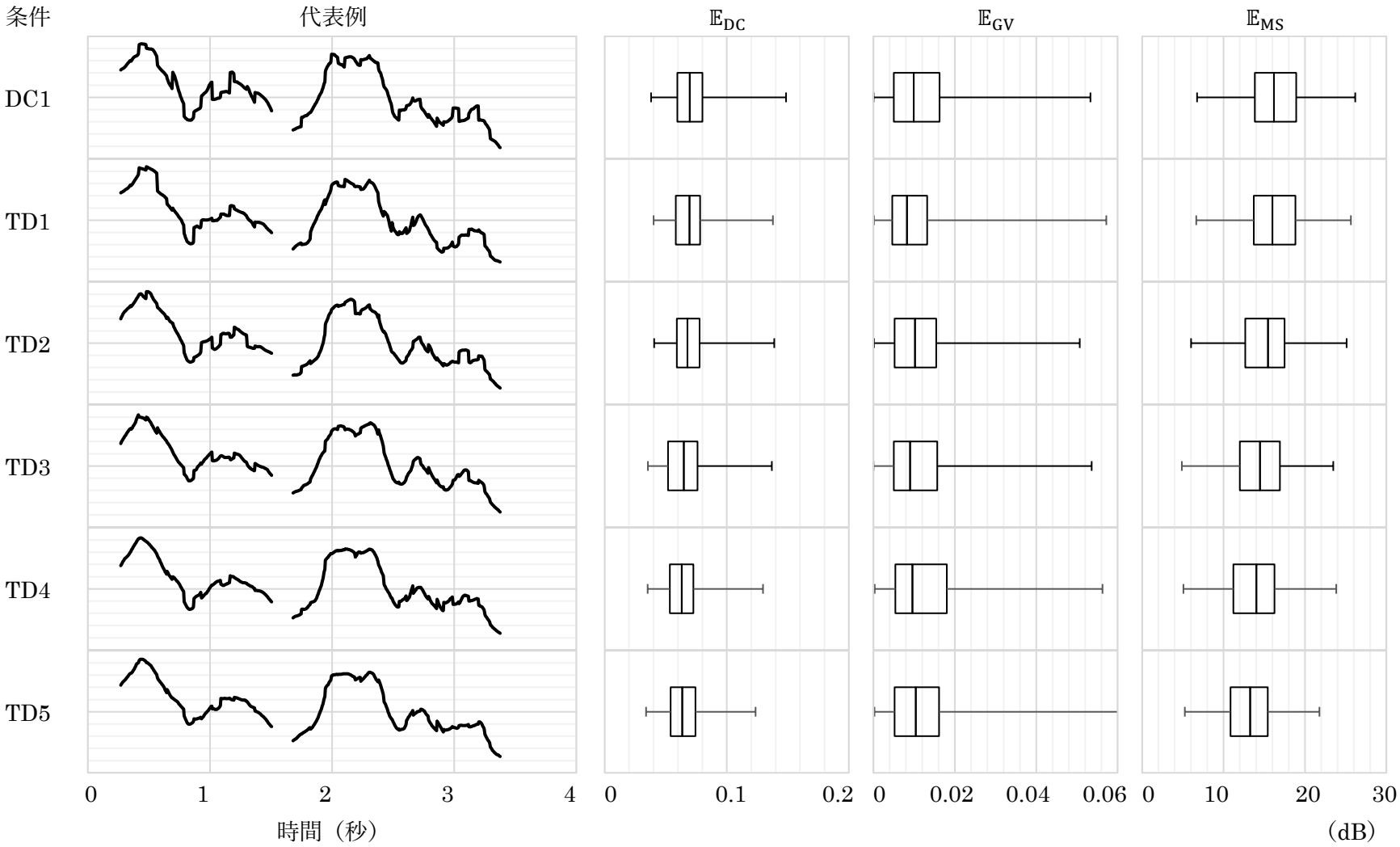

図 5.18 対数基本周波数についての MATS 損失関数の各損失関数の挙動を確認したときの結果

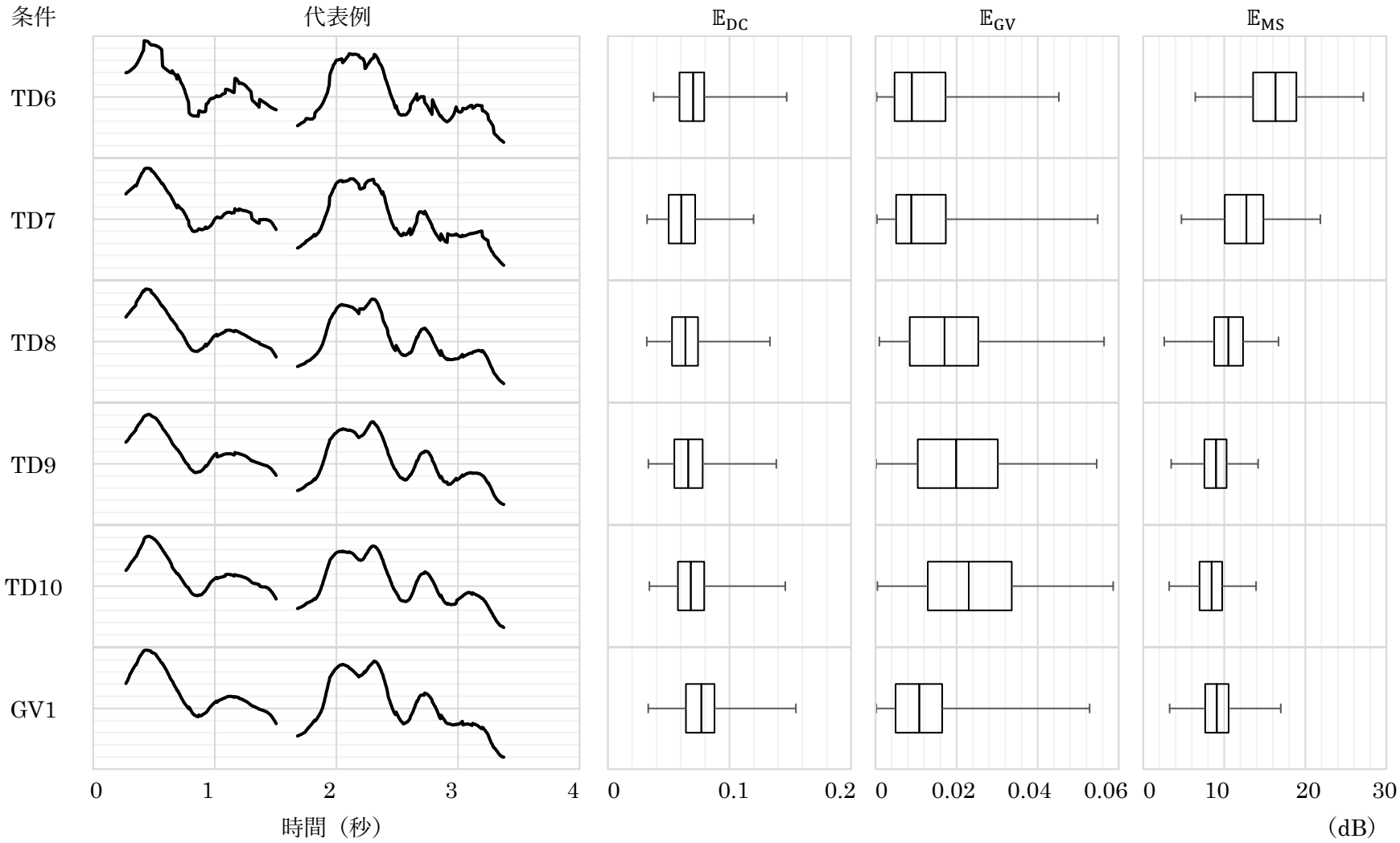

図 5.18 対数基本周波数についての MATS 損失関数の各損失関数の挙動を確認したときの結果

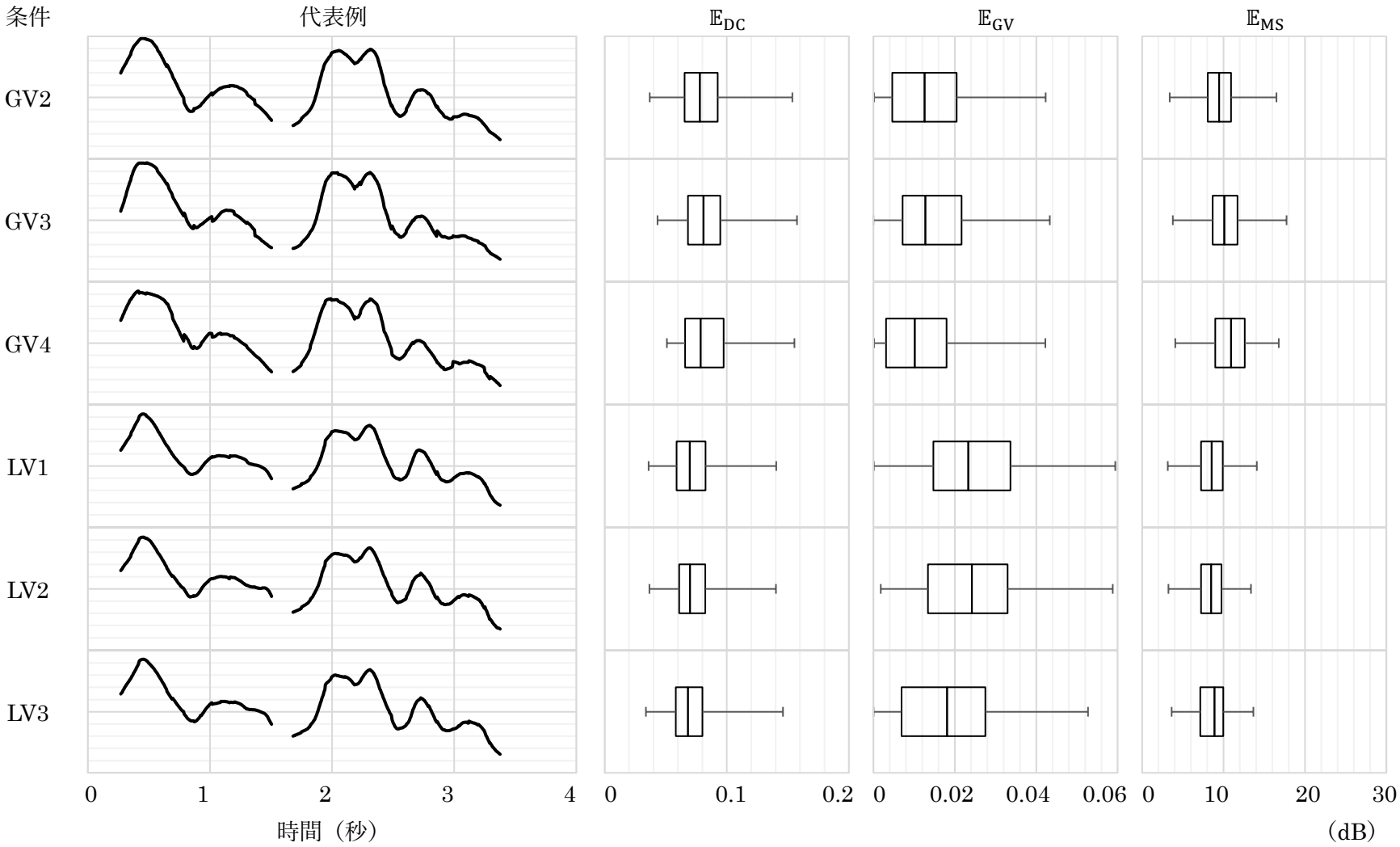

図 5.18 対数基本周波数についての MATS 損失関数の各損失関数の挙動を確認したときの結果

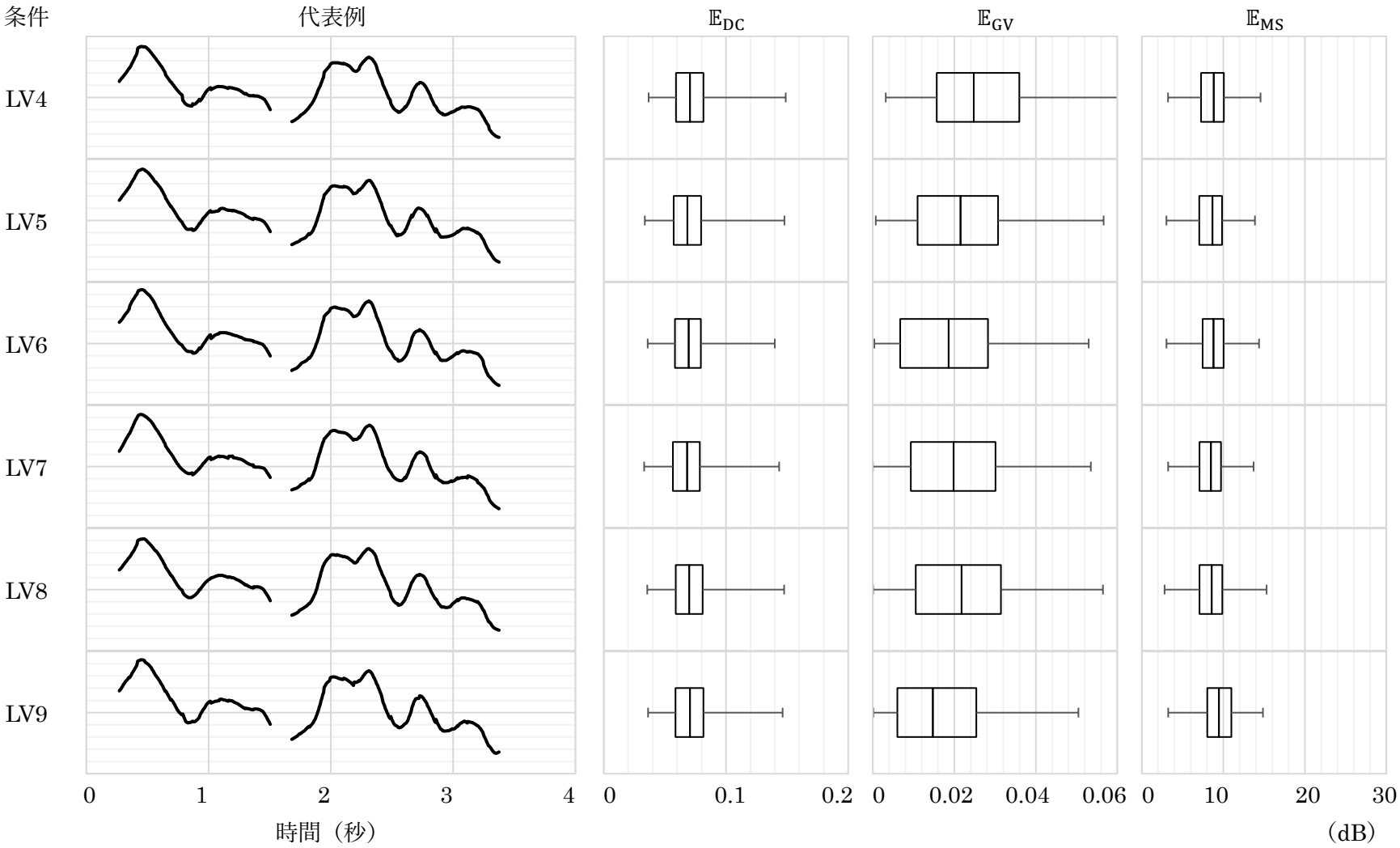

図 5.18 対数基本周波数についての MATS 損失関数の各損失関数の挙動を確認したときの結果

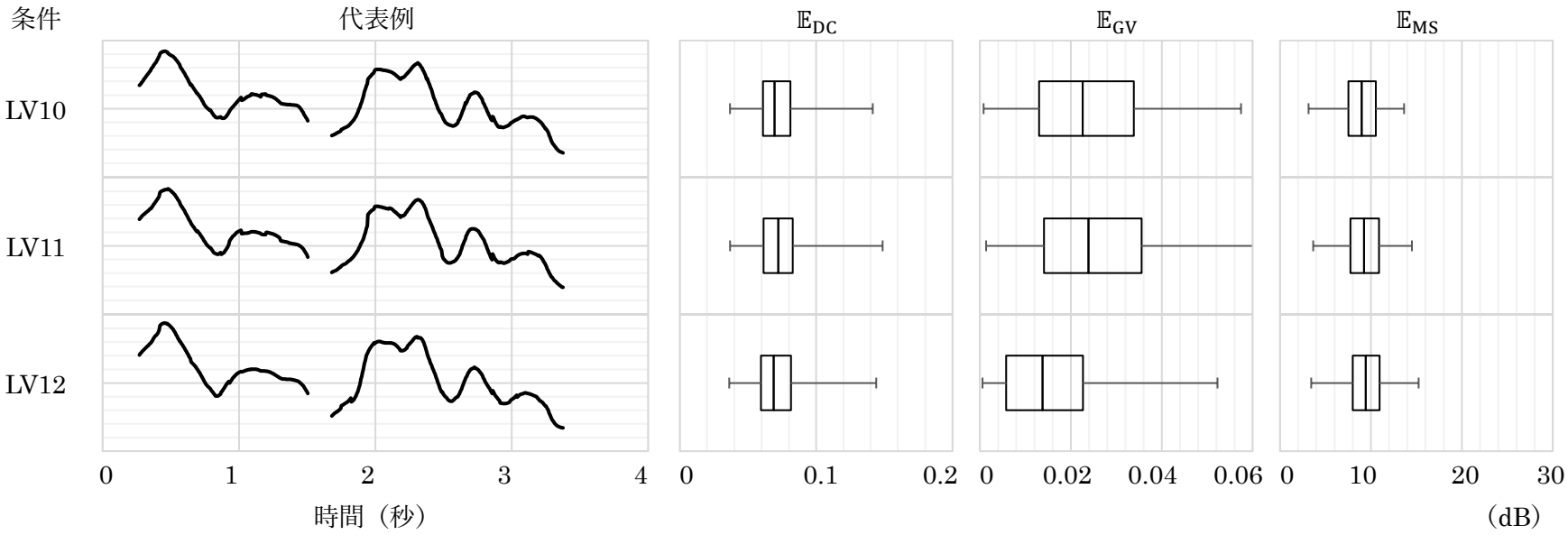

図 5.18 対数基本周波数についての MATS 損失関数の各損失関数の挙動を確認したときの結果

表 5.12 メルケプストラムについての MATS 損失関数の各損失関数の挙動を確認したときのパラメータの組み合わせ

条件	パラメータ一覧（表記のないものは使用していない）
DC1	$\omega_{\mathrm{DC}} = 1$
TD1	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{TD}} = 1, w_2 = 0.5, L_{\mathrm{TD}} = -1, R_{\mathrm{TD}} = 0$
TD2	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{TD}} = 1, w_2 = 1, L_{\mathrm{TD}} = -1, R_{\mathrm{TD}} = 0$
TD3	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{TD}} = 1, w_2 = 2, L_{\mathrm{TD}} = -1, R_{\mathrm{TD}} = 0$
GV1	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{GV}} = 1$
GV2	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{GV}} = 2$
GV3	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{GV}} = 4$
LV1	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{LV}} = 1, L_{\mathrm{LV}} = -1, R_{\mathrm{LV}} = 1$
LV2	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{LV}} = 1, L_{\mathrm{LV}} = -2, R_{\mathrm{LV}} = 2$
LV3	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{LV}} = 1, L_{\mathrm{LV}} = -4, R_{\mathrm{LV}} = 4$
LV4	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{LV}} = 1, L_{\mathrm{LV}} = -8, R_{\mathrm{LV}} = 8$
LV5	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{LV}} = 2, L_{\mathrm{LV}} = -1, R_{\mathrm{LV}} = 1$
LV6	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{LV}} = 2, L_{\mathrm{LV}} = -2, R_{\mathrm{LV}} = 2$
LV7	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{LV}} = 2, L_{\mathrm{LV}} = -4, R_{\mathrm{LV}} = 4$
LV8	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{LV}} = 2, L_{\mathrm{LV}} = -8, R_{\mathrm{LV}} = 8$
LV9	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{LV}} = 4, L_{\mathrm{LV}} = -1, R_{\mathrm{LV}} = 1$
LV10	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{LV}} = 4, L_{\mathrm{LV}} = -2, R_{\mathrm{LV}} = 2$
LV11	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{LV}} = 4, L_{\mathrm{LV}} = -4, R_{\mathrm{LV}} = 4$
LV12	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{LV}} = 4, L_{\mathrm{LV}} = -8, R_{\mathrm{LV}} = 8$
GC1	$\omega_{\mathrm{DC}} = 1, \omega_{\mathrm{GC}} = 1$

表 5.12 メルケプストラムについての MATS 損失関数の各損失関数の挙動を確認したときのパラメータの組み合わせ

GC2	$\omega_{\mathrm{DC}}=1,\omega_{\mathrm{GC}}=2$
GC3	$\omega_{\mathrm{DC}}=1,\omega_{\mathrm{GC}}=4$
LC1	$\omega_{\mathrm{DC}}=1,\omega_{\mathrm{LC}}=1,L_{\mathrm{LC}}=-1,R_{\mathrm{LC}}=1$
LC2	$\omega_{\mathrm{DC}}=1,\omega_{\mathrm{LC}}=1,L_{\mathrm{LC}}=-2,R_{\mathrm{LC}}=2$
LC3	$\omega_{\mathrm{DC}}=1,\omega_{\mathrm{LC}}=1,L_{\mathrm{LC}}=-4,R_{\mathrm{LC}}=4$
LC4	$\omega_{\mathrm{DC}}=1,\omega_{\mathrm{LC}}=1,L_{\mathrm{LC}}=-8,R_{\mathrm{LC}}=8$
LC5	$\omega_{\mathrm{DC}}=1,\omega_{\mathrm{LC}}=1,L_{\mathrm{LC}}=-1,R_{\mathrm{LC}}=1$
LC6	$\omega_{\mathrm{DC}}=1,\omega_{\mathrm{LC}}=1,L_{\mathrm{LC}}=-2,R_{\mathrm{LC}}=2$
LC7	$\omega_{\mathrm{DC}}=1,\omega_{\mathrm{LC}}=1,L_{\mathrm{LC}}=-4,R_{\mathrm{LC}}=4$
LC8	$\omega_{\mathrm{DC}}=1,\omega_{\mathrm{LC}}=1,L_{\mathrm{LC}}=-8,R_{\mathrm{LC}}=8$
LC9	$\omega_{\mathrm{DC}}=1,\omega_{\mathrm{LC}}=1,L_{\mathrm{LC}}=-1,R_{\mathrm{LC}}=1$
LC10	$\omega_{\mathrm{DC}}=1,\omega_{\mathrm{LC}}=1,L_{\mathrm{LC}}=-2,R_{\mathrm{LC}}=2$
LC11	$\omega_{\mathrm{DC}}=1,\omega_{\mathrm{LC}}=1,L_{\mathrm{LC}}=-4,R_{\mathrm{LC}}=4$
LC12	$\omega_{\mathrm{DC}}=1,\omega_{\mathrm{LC}}=1,L_{\mathrm{LC}}=-8,R_{\mathrm{LC}}=8$
DD1	$\omega_{\mathrm{DC}}=1,\omega_{\mathrm{DD}}=1,(w_{\mathrm{DD}})_d^{(m)}$は式（2.9）の「freqt」に従う（$D_1=60,\alpha_1=0.55,D_2=1025,\alpha_2=0.0$）.
DD2	$\omega_{\mathrm{DC}}=1,\omega_{\mathrm{DD}}=2,(w_{\mathrm{DD}})_d^{(m)}$は式（2.9）の「freqt」に従う（$D_1=60,\alpha_1=0.55,D_2=1025,\alpha_2=0.0$）.
DD3	$\omega_{\mathrm{DC}}=1,\omega_{\mathrm{DD}}=4,(w_{\mathrm{DD}})_d^{(m)}$は式（2.9）の「freqt」に従う（$D_1=60,\alpha_1=0.55,D_2=1025,\alpha_2=0.0$）.

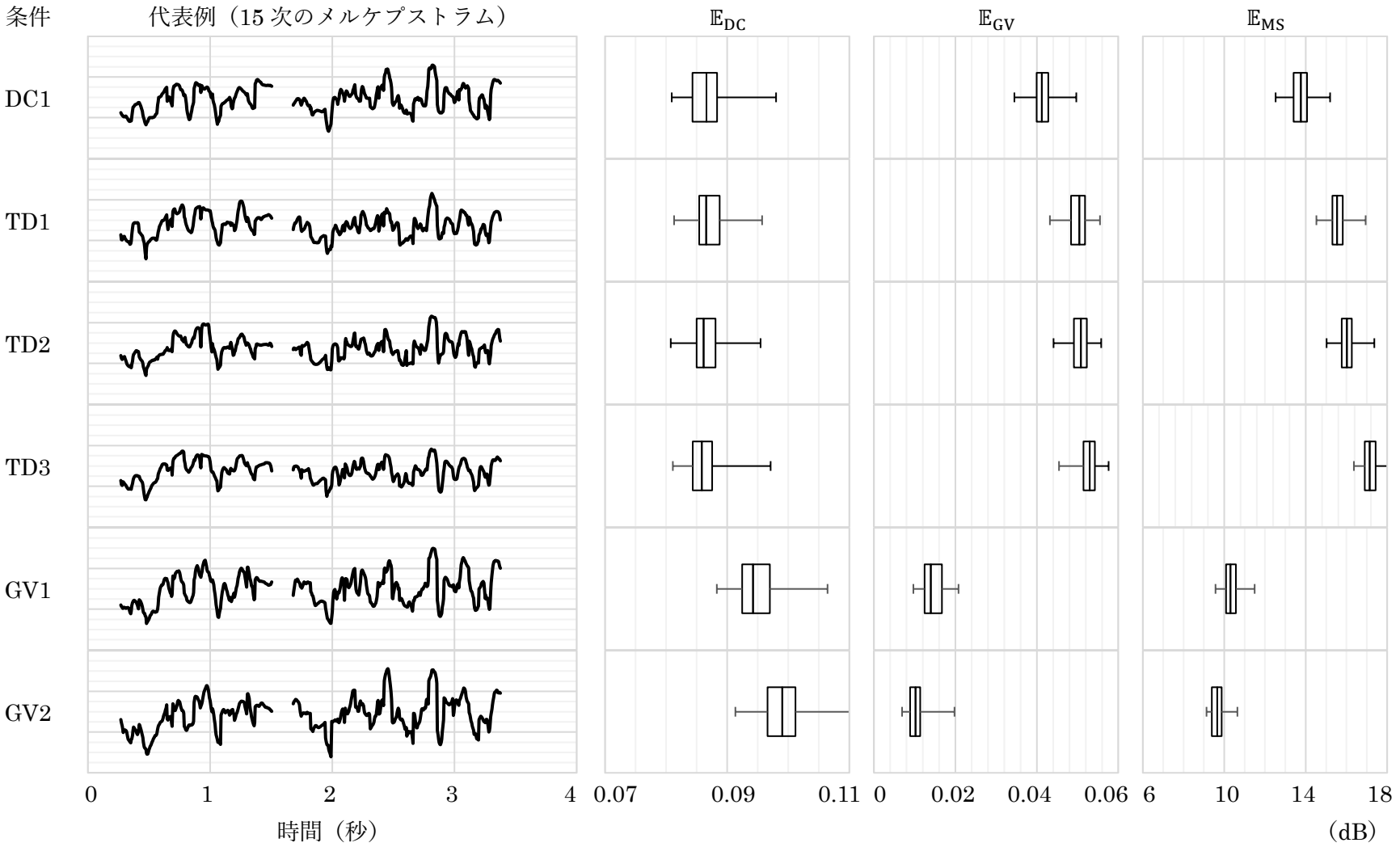

図 5.19 メルケプストラムについての MATS 損失関数の各損失関数の挙動を確認したときの結果

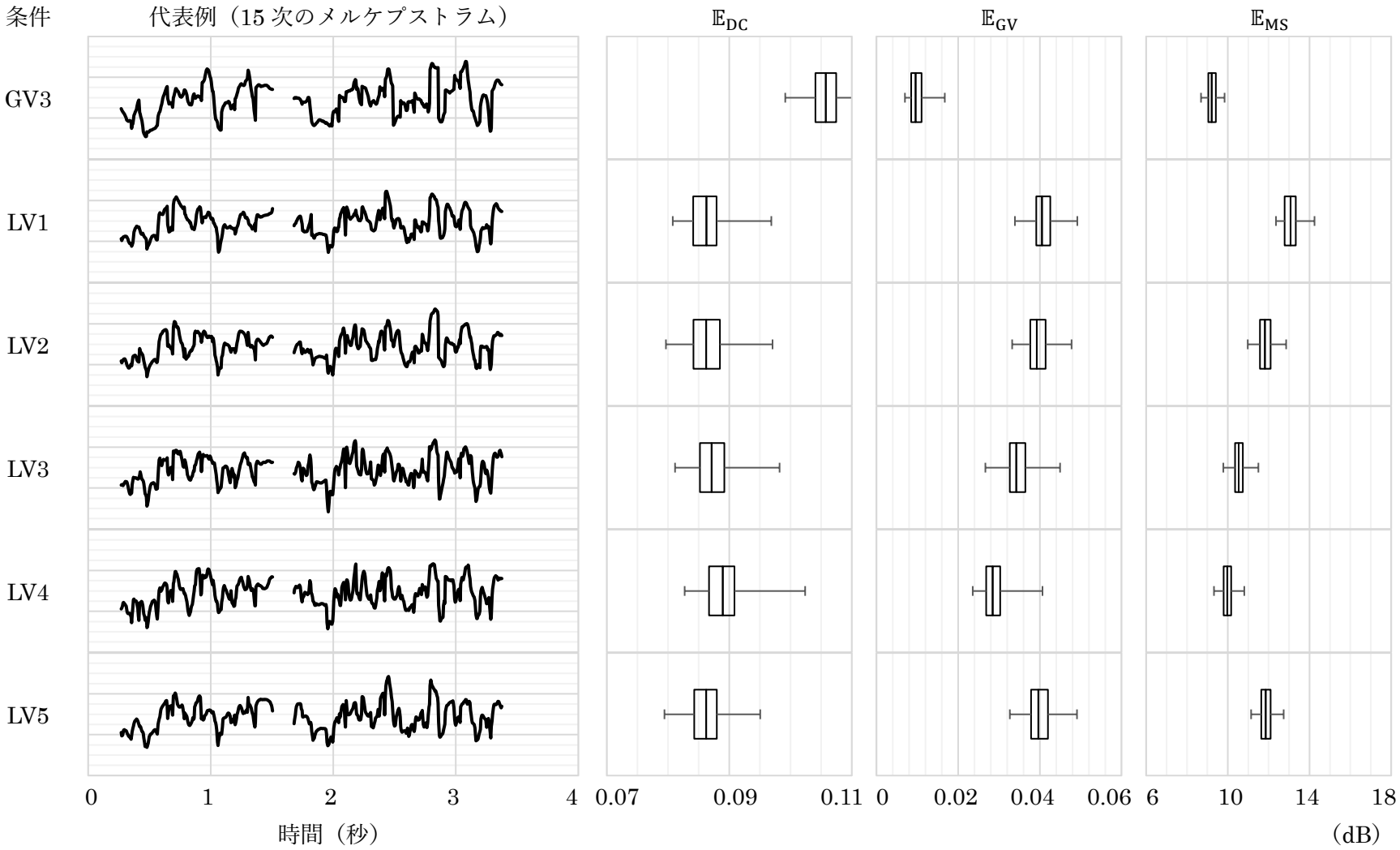

図 5.19 メルケプストラムについての MATS 損失関数の各損失関数の挙動を確認したときの結果

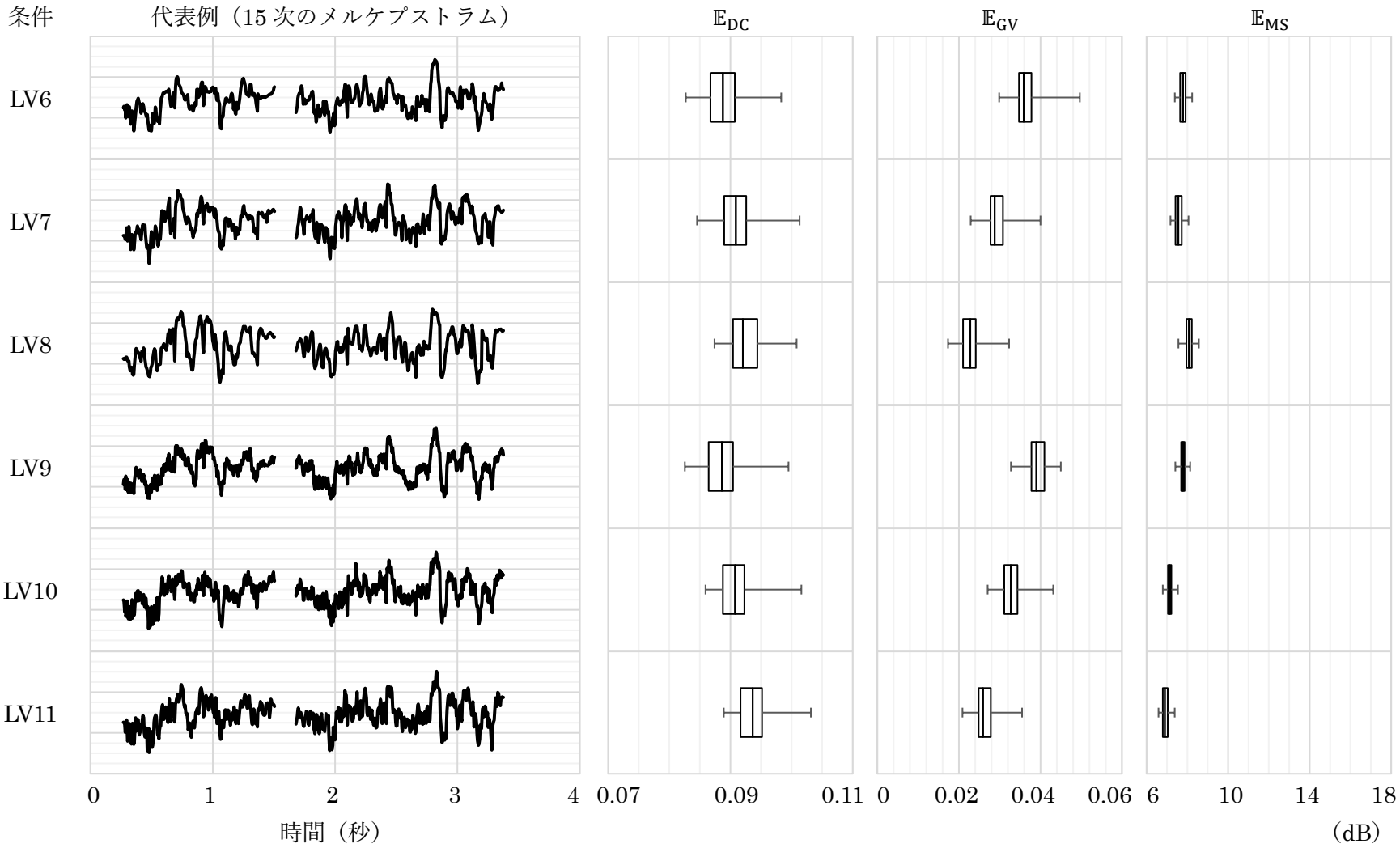

図 5.19 メルケプストラムについての MATS 損失関数の各損失関数の挙動を確認したときの結果

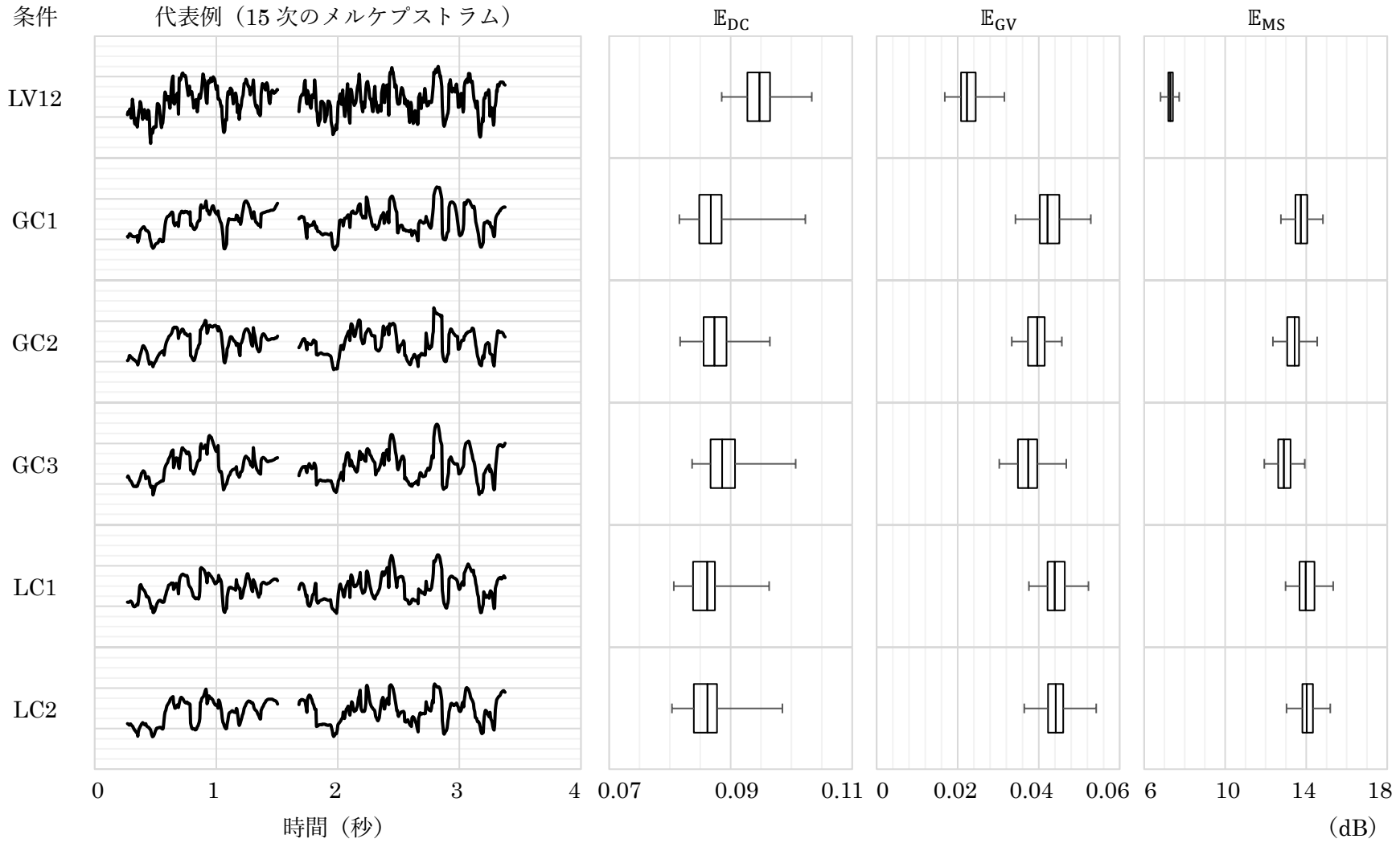

図 5.19 メルケプストラムについての MATS 損失関数の各損失関数の挙動を確認したときの結果

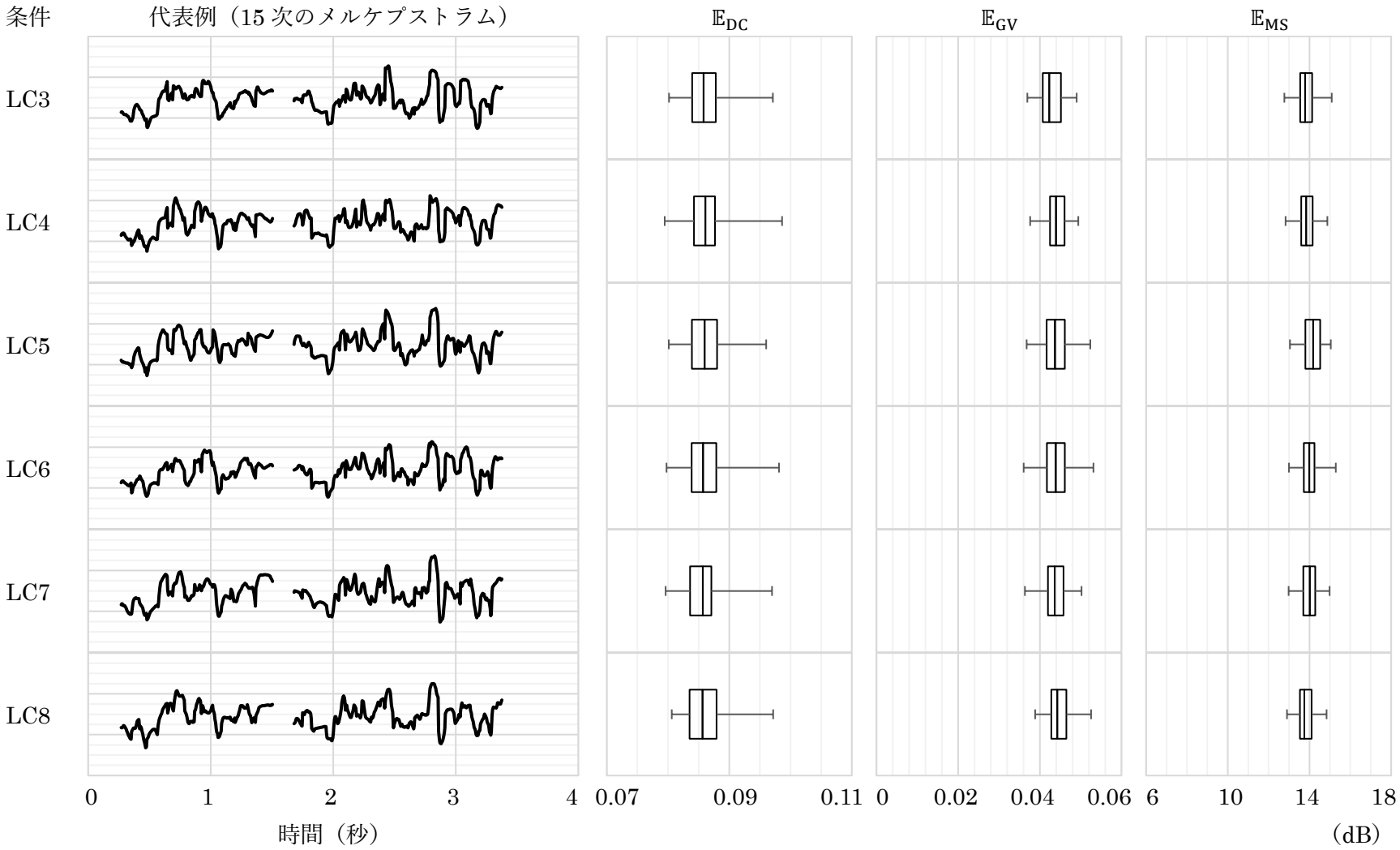

図 5.19 メルケプストラムについての MATS 損失関数の各損失関数の挙動を確認したときの結果

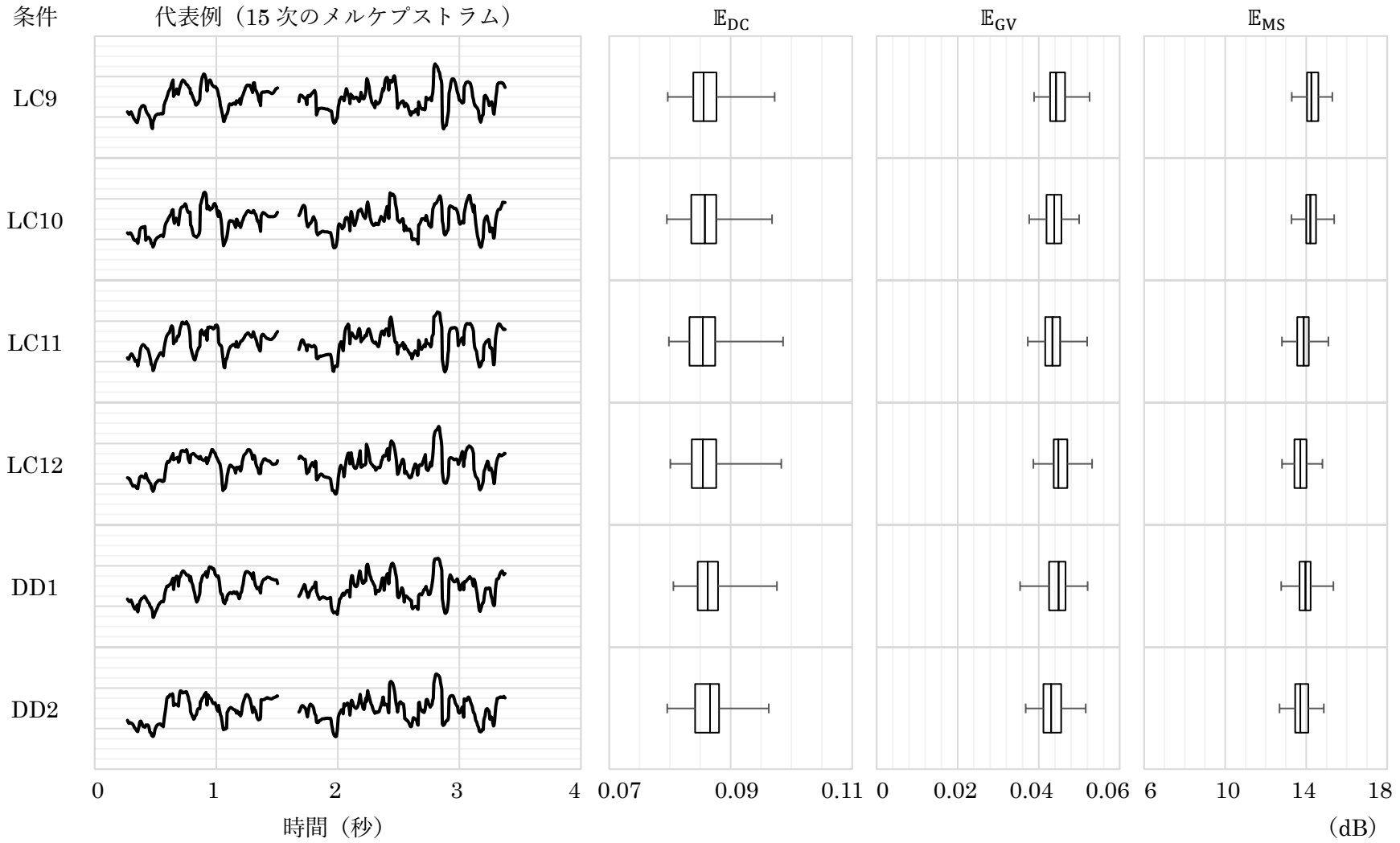

図 5.19 メルケプストラムについての MATS 損失関数の各損失関数の挙動を確認したときの結果

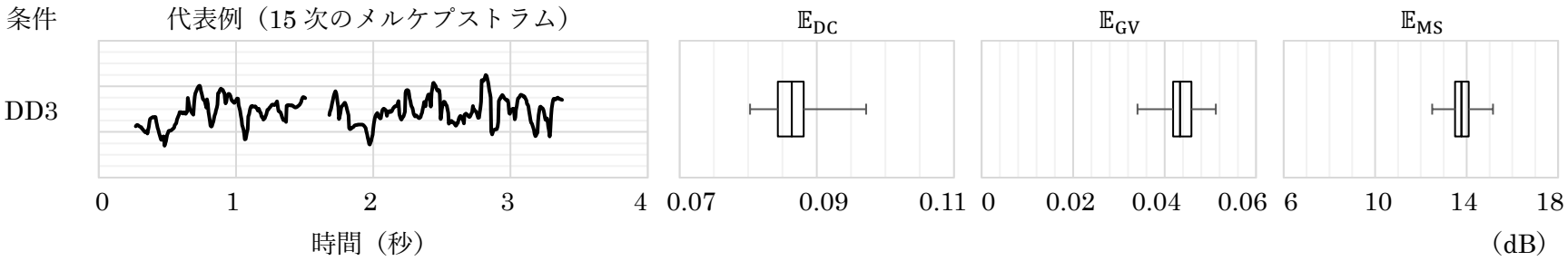

図 5.19 メルケプストラムについての MATS 損失関数の各損失関数の挙動を確認したときの結果

6. 時系列を考慮した生成的敵対ネットワークによる FFNN の学習法

6.1. はじめに

3 章では，後処理を用いない、FFNN のみで構成された音声特徴量予測部が最も高速であることを明らかにした．しかし，FFNN が時間フレームごとに独立して音声特徴量をモデル化するため，合成音声の品質が低下する問題がある．これに対し，5 章では MATS 損失関数による学習法を提案し，FFNN が単独で知覚的に優れた音声特徴量を予測することを可能にした．MATS 損失関数は経験や知見に基づいて複数の損失関数を定義する必要があった．そこで，本章では，複雑な時間構造を持つために，MATS 損失関数の学習において多くの損失関数を必要としたメルケプストラムを対象として，できる限り経験や知見を必要とせず，FFNN が音声特徴量の時間構造を考慮したモデルパラメータを獲得できるようする敵対的ネットワークによる学習法を提案する．

6.2. 生成的敵対ネットワーク

生成的敵対ネットワーク（GAN：Generative Adversarial Network [41] [42]）による学習法の構成を図 6.1 に示す．この学習法では，言語特徴量から音声特徴量を予測する DNN に加えて，原音声の音声特徴量か予測された音声特徴量かを識別する DNN を利用する．GAN による学習法では，言語特徴量から音声特徴量を予測する DNN を生成モデル，原音声の音声特徴量か予測された音声特徴量かを識別する DNN を識別モデルと呼ぶ. 識別モデルは，生成モデルが予測した音声特徴量が原音声の音声特徴量かの真偽を判定する．生成モデルは識別モデルを欺こうと学習され，識別モデルは正確に真偽を判定できるように学習される．このように，GAN による学習法では，互いのモデルが敵対するようにして生成モデルを学習することで，生成モデル単体での学習よりも生成モデルの予測性能を向上させる．また，識別モデルは学習時にのみ使用されるので，識別モデルの構成には制約はない．

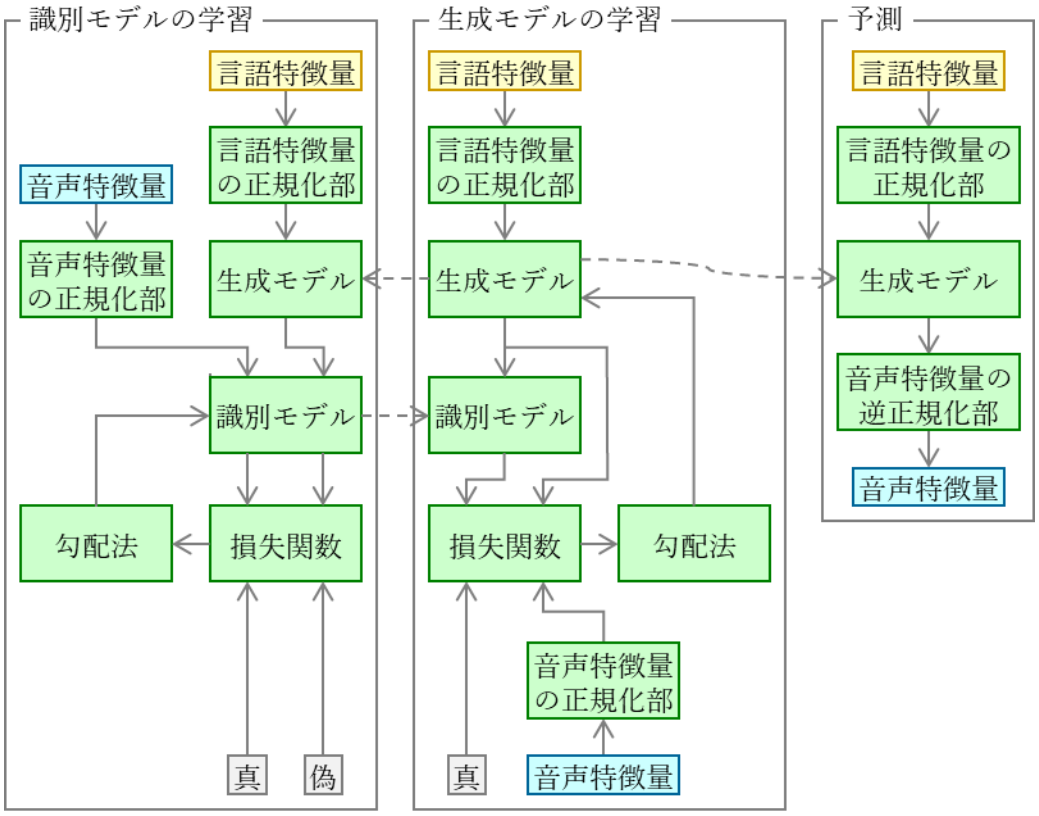

図 6.1　生成的敵対ネットワークによる学習法の構成

学習データセットの言語特徴量と音声特徴量を次式で定義する．

$$\begin{aligned} \boldsymbol{x} &= [\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_T^\top]^\top \\ \boldsymbol{x}_t &= \left[x_t^{(1)}, \cdots, x_t^{(k)}, \cdots, x_t^{(K)}\right] \end{aligned} \tag{6.1}$$

$$\begin{aligned} \boldsymbol{y} &= [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_t^\top, \cdots, \boldsymbol{y}_T^\top]^\top \\ \boldsymbol{y}_t &= \left[y_t^{(1)}, \cdots, y_t^{(d)}, \cdots, y_t^{(D)}\right] \end{aligned} \tag{6.2}$$

ここで，$\boldsymbol{x}$は言語特徴量ベクトル系列，$\boldsymbol{x}_t$は時間フレームtにおける言語特徴量ベクトル，$x_t^{(k)}$は時間フレームtにおける次元kの言語特徴量，$\boldsymbol{y}$は原音声の音声特徴量ベクトル系列，$\boldsymbol{y}_t$は時間フレームtにおける原音声の音声特徴量ベクトル，$y_t^{(d)}$は時間フレームtにおける次元dの原音声の音声特徴量，Kは言語特徴量の次元数，Dは音声特徴量の次元数，Tは時間フレーム数である．$\boldsymbol{y}$に対応する生成モデルで$\boldsymbol{x}$から予測した音声特徴量を次式で定義する．

$$\begin{aligned} \hat{\boldsymbol{y}} &= \mathcal{G}(\boldsymbol{x}) \\ &= [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_t^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top \\ \hat{\boldsymbol{y}}_t &= \left[\hat{y}_t^{(1)}, \cdots, \hat{y}_t^{(d)}, \cdots, \hat{y}_t^{(D)}\right] \end{aligned} \tag{6.3}$$

ここで，$\mathcal{G}$は生成モデル，$\hat{\boldsymbol{y}}$は予測した音声特徴量ベクトル系列，$\hat{\boldsymbol{y}}_t$は時間フレームtにおける予測した音声特徴量ベクトル，$\hat{y}_t^{(d)}$は時間フレームtにおける次元dの予測した音声特徴量である．$\boldsymbol{y}$や$\hat{\boldsymbol{y}}$に対応する識別モデルの教師データとして用いる真値や偽値を次式で定義する．

$$
\begin{aligned}
\boldsymbol{z} &= \begin{cases} \boldsymbol{z}^{(\mathcal{R})} \\ \boldsymbol{z}^{(\mathcal{F})} \end{cases} \\
&= [z_1, \cdots, z_t, \cdots, z_T]^\top \\
\boldsymbol{z}^{(\mathcal{R})} &= \left[z^{(\mathcal{R})}, z^{(\mathcal{R})}, \cdots, z^{(\mathcal{R})}\right]^\top \\
\boldsymbol{z}^{(\mathcal{F})} &= \left[z^{(\mathcal{F})}, z^{(\mathcal{F})}, \cdots, z^{(\mathcal{F})}\right]^\top
\end{aligned}
\tag{6.4}
$$

ここで，$\boldsymbol{z}$は教師データとしての真偽値系列，z_tは時間フレームtにおける真偽値，$\boldsymbol{z}^{(\mathcal{R})}$は真値系列，$z^{(\mathcal{R})}$は真値，$\boldsymbol{z}^{(\mathcal{F})}$は偽値系列，$z^{(\mathcal{F})}$は偽値である．$\boldsymbol{z}$に対応する$\boldsymbol{y}$や$\hat{\boldsymbol{y}}$に対する識別モデルの出力を次式で定義する．

$$
\begin{aligned}
\hat{\boldsymbol{z}} &= \begin{cases} \mathcal{D}(\boldsymbol{y}) \\ \mathcal{D}(\hat{\boldsymbol{y}}) \end{cases} \\
&= [\hat{z}_1, \cdots, \hat{z}_t, \cdots, \hat{z}_T]^\top
\end{aligned}
\tag{6.5}
$$

ここで, $\mathcal{D}$は識別モデル, $\hat{\boldsymbol{z}}$は$\mathcal{D}$の識別値系列, $\hat{z}_t$は時間フレームtにおける$\mathcal{D}$の識別値である．音声特徴量の生成誤差は$\boldsymbol{y}$と$\hat{\boldsymbol{y}}$の平均絶対誤差で定義される．

$$
\mathcal{L}_{\mathcal{G}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{TD} \sum_{t=1}^{T} \sum_{d=1}^{D} \left(y_t^{(d)} - \hat{y}_t^{(d)} \right)^2 \tag{6.6}
$$

ここで，$\mathcal{L}_{\mathcal{G}}$は生成誤差を求める損失関数である．また，$\boldsymbol{y}$や$\hat{\boldsymbol{y}}$の識別誤差は$\boldsymbol{z}$と$\hat{\boldsymbol{z}}$の交差エントロピーで定義される．

$$
\mathcal{L}_{\mathcal{D}}(\boldsymbol{z}, \hat{\boldsymbol{z}}) = -\frac{1}{T} \sum_{t=1}^{T} (z_t \log(\hat{z}_t) + (1 - z_t) \log(1 - \hat{z}_t)) \quad \begin{cases} \boldsymbol{z} = \boldsymbol{z}^{(\mathcal{R})}, \hat{\boldsymbol{z}} = \mathcal{D}(\boldsymbol{y}) \\ \boldsymbol{z} = \boldsymbol{z}^{(\mathcal{F})}, \hat{\boldsymbol{z}} = \mathcal{D}(\hat{\boldsymbol{y}}) \end{cases} \tag{6.7}
$$

ここで，$\mathcal{L}_{\mathcal{D}}$は識別誤差を求める損失関数である．GAN に基づく学習は生成モデルと識別モデルの学習を交互に繰り返す．識別モデルの学習では，$\boldsymbol{y}$の$\hat{\boldsymbol{z}}$に対するする$\boldsymbol{z}$を$\boldsymbol{z}^{(\mathcal{R})}$とし，$\hat{\boldsymbol{y}}$の$\hat{\boldsymbol{z}}$に対する$\boldsymbol{z}$を$\boldsymbol{z}^{(\mathcal{F})}$として, それぞれの場合における$\mathcal{L}_{\mathcal{D}}$に基づいて識別モデルのモデルパラメータを更新する．生成モデルの学習では，$\mathcal{L}_{\mathcal{G}}$と，$\hat{\boldsymbol{y}}$の$\hat{\boldsymbol{z}}$に対する$\boldsymbol{z}$を$\boldsymbol{z}^{(\mathcal{R})}$としたときの$\mathcal{L}_{\mathcal{D}}$との和の誤差に基づいて生成モデルのモデルパラメータを更新する．$\mathcal{L}_{\mathcal{G}}$と$\mathcal{L}_{\mathcal{D}}$の和の誤差は次式で定義される．

$$
\begin{aligned}
\mathcal{L}_{\mathcal{A}}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}, \boldsymbol{z}^{(\mathcal{R})}\right) &= \mathcal{L}_{\mathcal{G}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \frac{E_{\mathcal{G}}}{E_{\mathcal{D}}} \mathcal{L}_{\mathcal{D}}\left(\boldsymbol{z}^{(\mathcal{R})}, \hat{\boldsymbol{z}}\right) \\
&= \mathcal{L}_{\mathcal{G}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \frac{E_{\mathcal{G}}}{E_{\mathcal{D}}} \mathcal{L}_{\mathcal{D}}\left(\boldsymbol{z}^{(\mathcal{R})}, \mathcal{D}(\hat{\boldsymbol{y}})\right)
\end{aligned}
\tag{6.8}
$$

ここで，$\mathcal{L}_{\mathcal{A}}$は生成モデルを学習するときに用いる損失関数，$E_{\mathcal{G}}$は$\mathcal{L}_{\mathcal{G}}$の期待値，$E_{\mathcal{D}}$は$\mathcal{L}_{\mathcal{D}}$の期待値である．$E_{\mathcal{G}}$と$E_{\mathcal{D}}$で，$\mathcal{L}_{\mathcal{G}}$の生成誤差と$\mathcal{L}_{\mathcal{D}}$の識別誤差のスケールの違いを調整する．$E_{\mathcal{G}}$と$E_{\mathcal{D}}$は生成モデルと識別モデルのパラメータが更新されるたびに計算する．

6.3. 識別モデル

本章では, 3.2.3 の音声特徴量予測部の構成を対象とするため, 生成モデルの構成は表 3.1 の FFNN-3.2.3 となる．生成モデルは固定であるため，ここでは 3 つの識別モデルについ

て述べる．1 つめは従来法で，FFNN による識別モデルである．2 つめは従来法で，CNN による識別モデルである．3 つめは提案法で，時系列の相関関係を考慮する識別モデルである．

6.3.1. 従来法：FFNN の識別モデル

文献 [42]では，生成モデルは MGE 学習による FFNN，識別モデルは FFNN である．識別モデルは FFNN であるため，音声特徴量の時間構造を考慮せず，時間フレームごとに独立して原音声の音声特徴量か予測された音声特徴量かを識別する．このため，生成モデルの学習に用いられる識別誤差には音声特徴量の時間構造に関する情報は含まれない．ただし，MGE 学習により，生成モデル自体は隣接する時間フレーム間の音声特徴量の動的特徴量の関係を学習できる．

本章では，表 3.1 の FFNN-3.2.3 を生成モデルとした 3.2.3 の音声特徴量予測部の構成を対象とする．MLPG を用いないため，生成モデルも識別モデルも音声特徴量の時間構造を考慮しない．この構成の GAN による学習法が，GAN による学習を利用しないで，生成誤差だけに基づいて学習された生成モデルよりも優れた生成モデルを学習できるかを確認する．

6.3.2. 従来法：CNN の識別モデル

文献 [43]では，生成モデルは RNN，識別モデルは CNN である．識別モデルは畳み込み層により，原音声の音声特徴量か予測された音声特徴量かを識別するための特徴量を複数の時間フレームの音声特徴量から抽出する．また，生成モデルは再帰構造により音声特徴量の時間構造を捉えることができる．この構成の GAN による学習法では，生成モデルは，自身の再帰構造で音声特徴量の時間構造を捉えながら，音声特徴量の平均二乗誤差と，CNN で抽出された特徴量の情報が含まれた識別誤差に基づいて学習される．

本章では，表 3.1 の FFNN-3.2.3 を生成モデルとした 3.2.3 の音声特徴量予測部の構成を対象とする．生成モデルは FFNN であるため，識別モデルが音声特徴量の時間構造を捉える役割を果たす．この構成の GAN による学習法が，6.3.1 の構成の GAN による学習法よりも優れた生成モデルを学習できるかを確認する．

6.3.3. 提案法：時系列の相関関係を考慮する識別モデル

従来法の生成モデルは MGE 学習や RNN により，生成モデル自体が音声特徴量の時間構造を考慮できる状態で学習される．しかし，本章では，表 3.1 の FFNN-3.2.3 を生成モデルとした 3.2.3 の音声特徴量予測部の構成を対象としている．生成モデルは音声特徴量の時間構造を考慮しないため，識別モデルが音声特徴量の時間構造を捉える必要がある．6.3.2 の識別モデルは CNN であるが，畳み込みの幅の制限があるため，音声特徴量から抽出される特徴量に系列全体の特徴は含まれていない．

そこで本章では，系列全体の特徴を表す音声特徴量のグラム行列を用いた識別モデルを提案する．従来の識別モデルは時間フレームごとの音声特徴量を識別するが，提案する識別モデルは系列全体を表すグラム行列を識別する．これにより，生成モデルは音声特徴量の時間フレームごとの特徴を捉える役割を担い，識別モデルは音声特徴量の系列全体の特徴を捉える役割を担う．このようにして，提案する GAN による学習法は生成モデルと識別モデルを別々の基準で学習することで，音声特徴量を多角的に捉えることを可能にする．原音声の音声特徴量のグラム行列を次式で定義する．

$$\boldsymbol{g} = \frac{\boldsymbol{y}^{\top}\boldsymbol{y}}{T} \tag{6.9}$$

ここで，$\boldsymbol{g}$は原音声の音声特徴量の$D \times D$のグラム行列である．また，生成モデルで予測された音声特徴量のグラム行列を次式で定義する．

$$\hat{\boldsymbol{g}} = \frac{\hat{\boldsymbol{y}}^{\top}\hat{\boldsymbol{y}}}{T} \tag{6.10}$$

ここで，$\hat{\boldsymbol{g}}$は原音声の音声特徴量の$D \times D$のグラム行列である．これらのグラム行列をTで除算するのは，音声特徴量の時間フレーム数による影響をなくすためである．グラム行列から真偽値を出力するまでの畳み込み層の構成は，画像生成において高性能な GAN である畳み込みニューラルネットワークによる GAN（DC-GAN：Deep Convolutional GAN [44]）の識別モデルと同じである．この GAN による学習法はグラム行列と DC-GAN を利用するため，グラム行列による畳み込みニューラルネットワークによる GAN（GDC-GAN：Gram matrix DC-GAN）と命名した．

6.4. 実験方法

6.4.1. 生成モデルと識別モデル

実験に用いた識別モデルの構成を表 6.1 に示す．FFNN-DIS は，ユニット数がそれぞれ 128，64，32，活性化関数が漏洩型正規化線形関数（Leaky ReLU 関数：Leaky Rectified Linear Unit 関数 [45]）の 3 層の全結合層と，ユニット数が 1 で，活性化関数がシグモイド関数の全結合層で構成される．CNN-DIS は，フィルタ数がそれぞれ 128，64，32，ステップ幅が 1，活性化関数が Leaky ReLU 関数の 3 層の 1 次元畳み込み層と，ユニット数が 1 で，活性化関数がシグモイド関数の全結合層で構成される．GDC-DIS は，グラム行列算出層と，フィルタ数がそれぞれ 8，16，32，ステップ幅が 2，活性化関数が Leaky ReLU 関数の 2 次元畳み込み層と，ユニット数が 1 で，活性化関数がシグモイド関数の全結合層で構成される．FFNN-DIS と CNN-DIS は，時間フレーム数がTの音声特徴量に対して，時間フレーム数がTの識別値を出力するが，GDC-DIS は時間フレーム数がTの音声特徴量からグラム行列を算出し，そのグラム行列に対して，1 つの識別結果を出力する．

また，実験に用いた GAN の構成を表 6.2 に示す．生成モデルは表 3.1 の FFNN-3.2.3 である．生成モデルの損失関数は$\mathcal{L}_{\mathcal{A}}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}, \boldsymbol{z}^{(\mathcal{R})}\right)$である．$\mathcal{L}_{\mathcal{A}}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}, \boldsymbol{z}^{(\mathcal{R})}\right)$のうち生成誤差を算出す

るのは$\mathcal{L}_{\mathcal{G}}(\boldsymbol{y},\hat{\boldsymbol{y}})$であり，$\mathcal{L}_{\mathcal{G}}(\boldsymbol{y},\hat{\boldsymbol{y}})$は 5.2.1 で述べた$\mathcal{L}_{\mathrm{MSE}}(\boldsymbol{y},\hat{\boldsymbol{y}})$と同じである．生成モデルの勾配法は Adam 法であり，Adam 法のパラメータについては，学習率を 0.001，β_1を 0.9，β_2を 0.999，微小量を10^{-7}，学習率減衰を 0.0 とした．識別モデルの勾配法は Adam 法であり, Adam 法のパラメータについては, 学習率を10^{-6}, β_1を 0.9, β_2を 0.999, 微小量を10^{-7}, 学習率減衰を 0.0 とした．エポック数は 20 とし，バッチサイズは 1 文ごとの音声特徴量の時間フレーム数とした. 言語特徴量の正規化法は 4.2.4 で提案した 2 つの言語特徴量の属性値の比を取る正規化法を使用した. 学習データセットと評価データセットはそれぞれ 2.3 で説明した$\mathbb{U}_{2000}$と$\mathbb{U}_{\mathrm{s}}$を使用した.

表 6.1　識別モデルの構成

識別名	FFNN-DIS	CNN-DIS	GDC-DIS
1 層目	全結合層 128 units Leaky ReLU	1 次元畳み込み層 filter：$(5 \times D)$ 128 filters 1 step Leaky ReLU	グラム行列算出層
2 層目	全結合層 64 units Leaky ReLU	1 次元畳み込み層 filter：(5×128) 64 filters 1 step Leaky ReLU	2 次元畳み込み層 filter：$(D \times D)$ 8 filters 2 steps Leaky ReLU
3 層目	全結合層 32 units Leaky ReLU	1 次元畳み込み層 filter：(5×64) 32 filters 1 step Leaky ReLU	2 次元畳み込み層 filter：$(D/2 \times D/2)$ 16 filters 2 steps Leaky ReLU
4 層目	全結合層 1 unit Sigmoid	全結合層 1 unit Sigmoid	2 次元畳み込み層 filter：$(D/4 \times D/4)$ 32 filters 2 steps Leaky ReLU
5 層目			全結合層 1 unit Sigmoid

表 6.2　生成的敵対ネットワークの構成

識別名	FFNN-GAN	CNN-GAN	GDC-GAN
音声特徴量予測部の構成	FFNN （3.2.3）	FFNN （3.2.3）	FFNN （3.2.3）
生成モデルの構成	全結合層×5 （FFNN-3.2.3）	全結合層×5 （FFNN-3.2.3）	全結合層×5 （FFNN-3.2.3）
生成モデルの損失関数	$\mathcal{L}_{\mathcal{A}}(\boldsymbol{y},\hat{\boldsymbol{y}},\boldsymbol{z}^{(\mathcal{R})})$	$\mathcal{L}_{\mathcal{A}}(\boldsymbol{y},\hat{\boldsymbol{y}},\boldsymbol{z}^{(\mathcal{R})})$	$\mathcal{L}_{\mathcal{A}}(\boldsymbol{y},\hat{\boldsymbol{y}},\boldsymbol{z}^{(\mathcal{R})})$
生成モデルの勾配法	Adam 法	Adam 法	Adam 法
識別モデルの構成	FFNN-DIS	CNN-DIS	GDC-DIS
識別モデルの損失関数	$\mathcal{L}_{\mathcal{D}}(\boldsymbol{z},\hat{\boldsymbol{z}})$	$\mathcal{L}_{\mathcal{D}}(\boldsymbol{z},\hat{\boldsymbol{z}})$	$\mathcal{L}_{\mathcal{D}}(\boldsymbol{z},\hat{\boldsymbol{z}})$
識別モデルの勾配法	Adam 法	Adam 法	Adam 法

6.4.2. 聴取実験方法

聴取実験は 5.4.2 で述べた MUSHRA 法を用いた. 学習データセットと評価データセットはそれぞれ 2.3 で述べた$\mathbb{U}_{2000}$と$\mathbb{U}_{\mathrm{s}}$を使用した．実験に用いた合成音声は 2.1.2 のボコーダの合成部で生成した．メルケプストラムを予測するときは，原音声の継続長から算出した時間フレーム情報を付与した言語特徴量を使用した．各評価群の音声は予測したメルケプストラムと，そのメルケプストラムに対応する原音声の基本周波数と非周期性指標から合成した．参照群の音声は，原音声の分析再合成音声である．アンカー群の音声は，GAN による学習法を適用していない生成モデルで予測したメルケプストラムと，そのメルケプストラムに対応する原音声の基本周波数と非周期性指標から合成した．参加者は合成音声の韻律や音質の違いに敏感な 10 名である．合成音声の音質を評価するため，合成音声のアクセントや抑揚に注目しないように指示をした．セッション数は 100 であるため，参加者を適宜休憩させた．

6.4.3. 予測誤差の算出方法

予測誤差は 5.4.3 と同様にメルケプストラムの平均絶対誤差$\mathbb{E}_{\mathrm{DC}}$，メルケプストラムの系列内分散の平方根の平均絶対誤差$\mathbb{E}_{\mathrm{GV}}$，メルケプストラムの変調スペクトルの平均絶対誤差$\mathbb{E}_{\mathrm{MS}}$を用いて，各 GAN の生成モデルで予測したメルケプストラムを評価した．評価データセットは$\mathbb{U}_{\mathrm{s}}$を使用した．

6.5. 実験結果

6.5.1. 聴取実験結果

MUSHRA 法による聴取実験の結果をに図 6.2 示す．また，図 6.2 の評点を Tukey-

Kramer 法によって比較した結果を表 6.3 に示す．その結果，CNN-GAN 群とアンカー群には有意差が認められなかった．FFNN-GAN 群と GDC-GAN 群の音声はどちらも過剰平滑の問題が解決され，アンカー群の音声よりも品質が向上していた．FFNN-GAN 群の音声と GDC-GAN 群の音声の違いはわずかではあったが，FFNN-GAN 群の音声の方が，音声の帯域が広い印象を受けた．CNN-GAN 群の音声は粗造性嗄声のような印象を受けた．以上より，評価群においては，GDC-GAN 群の音声の品質が最も高かった．

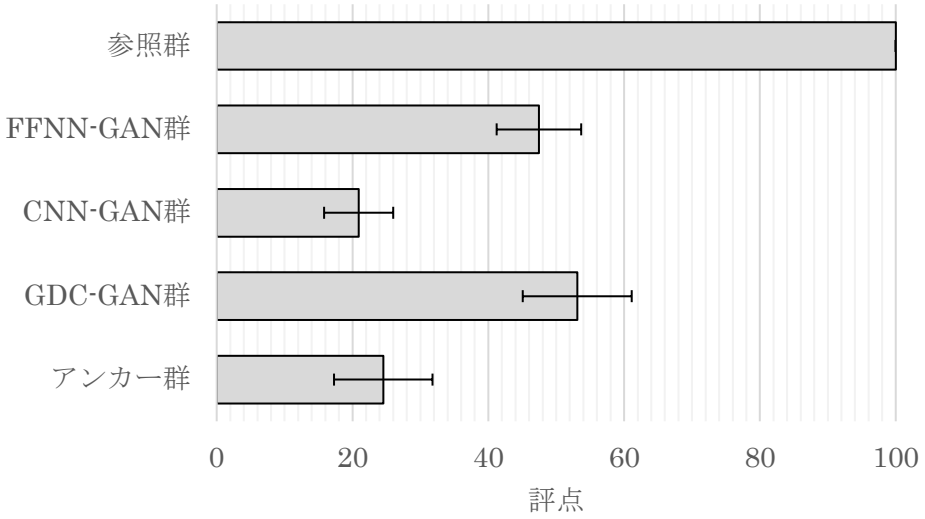

図 6.2　MUSHRA 法による合成音声の音質についての聴取実験の結果

表 6.3　Tukey-Kramer 法による聴取実験の𝕍の平均値の比較結果
表中の数値はスチューデント化された範囲分布の q 値と p 値である．群数は 6，自由度は 54，信頼区間は 95%である．

群 1	群 2	q 値	p 値
参照群	FFNN-GAN 群	51.26	0.001
参照群	CNN-GAN 群	77.16	0.001
参照群	GDC-GAN 群	45.76	0.001
参照群	アンカー群	73.63	0.001
FFNN-GAN 群	CNN-GAN 群	25.90	0.001
FFNN-GAN 群	GDC-GAN 群	5.50	0.003
FFNN-GAN 群	アンカー群	22.37	0.001
CNN-GAN 群	GDC-GAN 群	31.40	0.001
CNN-GAN 群	アンカー群	3.53	0.109
GDC-GAN 群	アンカー群	27.87	0.001

6.5.2. 予測誤差の結果

各 GAN により学習した生成モデルで予測した 15 次のメルケプストラムの代表例を図 6.3 から図 6.5 までに示す．これらの 15 次のメルケプストラムについて，原音声の 15 次のメルケプストラムとの比較を述べる．FFNN-GAN の 15 次のメルケプストラムについては，複雑な時間構造は現れていないが，起伏は概ね一致した．系列内分散の平方根は約 0.02 小さかった．変調スペクトルは約 30 Hz 以上の帯域で約 10 dB 小さかった．複雑な時間構造が現れており，起伏も概ね一致した．系列内分散の平方根は約 0.05 小さかった．変調スペクトルは概ね一致した．GDC-GAN の 15 次のメルケプストラムについては，複雑な時間構造が現れており，起伏も概ね一致した．系列内分散の平方根は約 0.02 小さかった．変調スペクトルは約 30 Hz 以上の帯域で約 10 dB 小さかった．

各 GAN により学習した生成モデルで予測したメルケプストラムの$\mathbb{U}_s$についての$\mathbb{E}_{DC}$，$\mathbb{E}_{GV}$，$\mathbb{E}_{MS}$をそれぞれ，図 6.6，図 6.7，図 6.8 に示す．また，これらのメルケプストラムの$\mathbb{U}_s$についての$\mathbb{E}_{DC}$, $\mathbb{E}_{GV}$, $\mathbb{E}_{MS}$の平均値を Tukey-Kramer 法で比較した結果を表 6.4, 表 6.5, 表 6.6 に示す．GDC-GAN の$\mathbb{E}_{DC}$の平均値は，FFNN-GAN，CNN-GAN の$\mathbb{E}_{DC}$の平均値よりも有意に大きかった．GDC-GAN の$\mathbb{E}_{DC}$の中央値と FFNN-GAN，CNN-GAN の$\mathbb{E}_{DC}$の中央値との差は約 0.006 以下であり，メルケプストラム係数の値を考慮すると，これらの差は合成音声において無視できる程度のものといえる．

GDC-GAN の$\mathbb{E}_{GV}$の平均値は，FFNN-GAN，CNN-GAN の$\mathbb{E}_{GV}$の平均値よりも有意に小さかった．GDC-GAN の$\mathbb{E}_{GV}$の中央値は，FFNN-GAN，CNN-GAN の$\mathbb{E}_{GV}$の中央値よりもそれぞれ約 0.008，約 0.016 小さかった．これらの差は，メルケプストラムの系列内分散の平方根の値や聴取実験の評点を考慮すると，合成音声の品質に影響と及ぼす程度のものといえる．

GDC-GAN の$\mathbb{E}_{MS}$の平均値は，FFNN-GAN の$\mathbb{E}_{MS}$の平均値よりも有意に小さく，CNN-GAN の$\mathbb{E}_{MS}$の平均値よりも有意に大きかった．GDC-GAN の$\mathbb{E}_{MS}$の中央値は，FFNN-GAN の$\mathbb{E}_{MS}$の中央値よりも約 2 dB 小さく，CNN-GAN の$\mathbb{E}_{MS}$の中央値よりも 1.8 dB 大きかった．FFNN-GAN の$\mathbb{E}_{MS}$との差は，聴取実験の評点を考慮すると，合成音声の品質に影響を及ぼす程度のものといえる．ただし，CNN-GAN については，その聴取実験の評点を考慮すると，系列内分散を改善しないまま，変調スペクトルを改善しても合成音声の音質を改善することができないといえる．

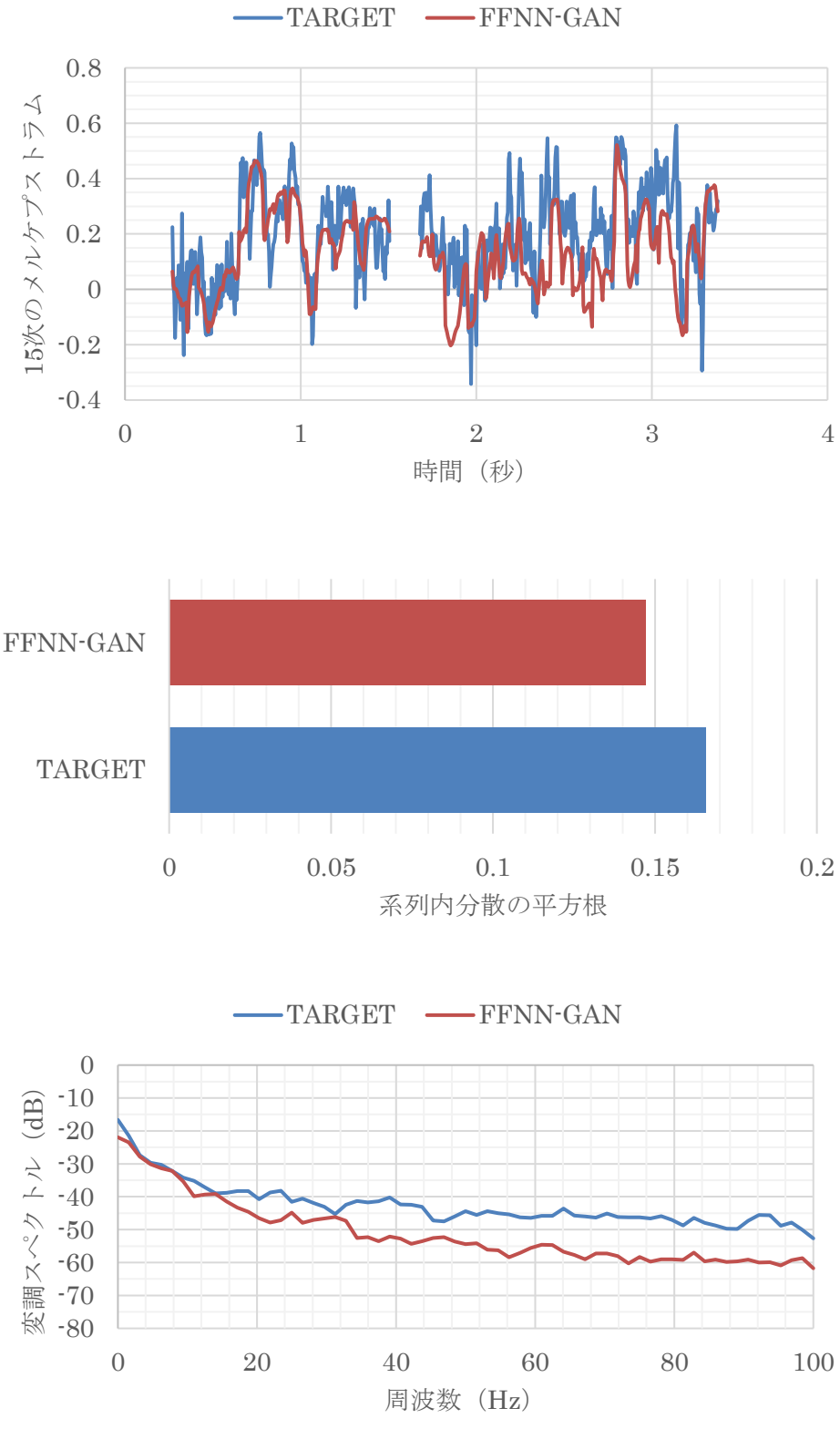

図 6.3　FFNN-GAN のメルケプストラムの代表例

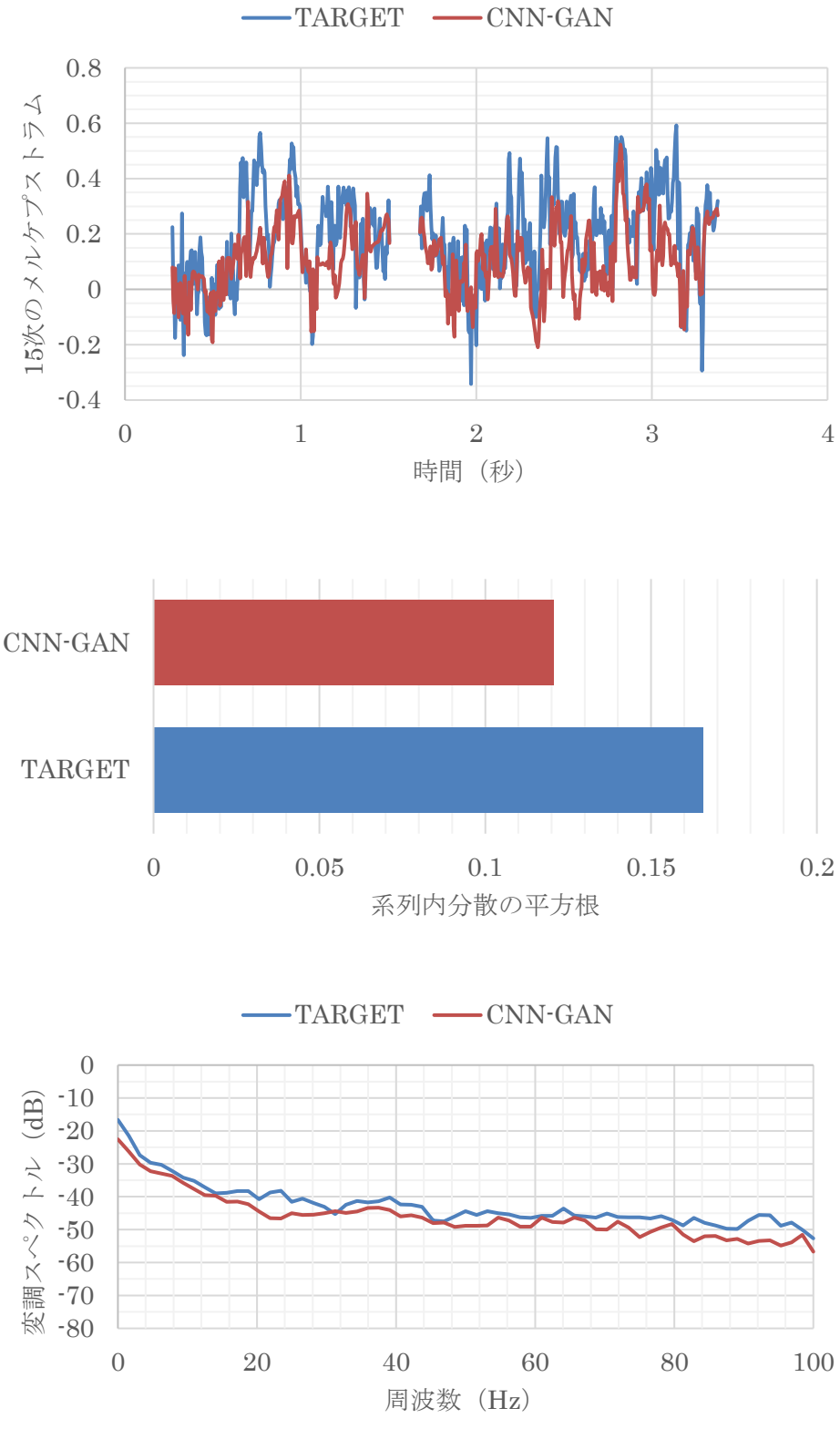

図 6.4　FFNN-GAN のメルケプストラムの代表例

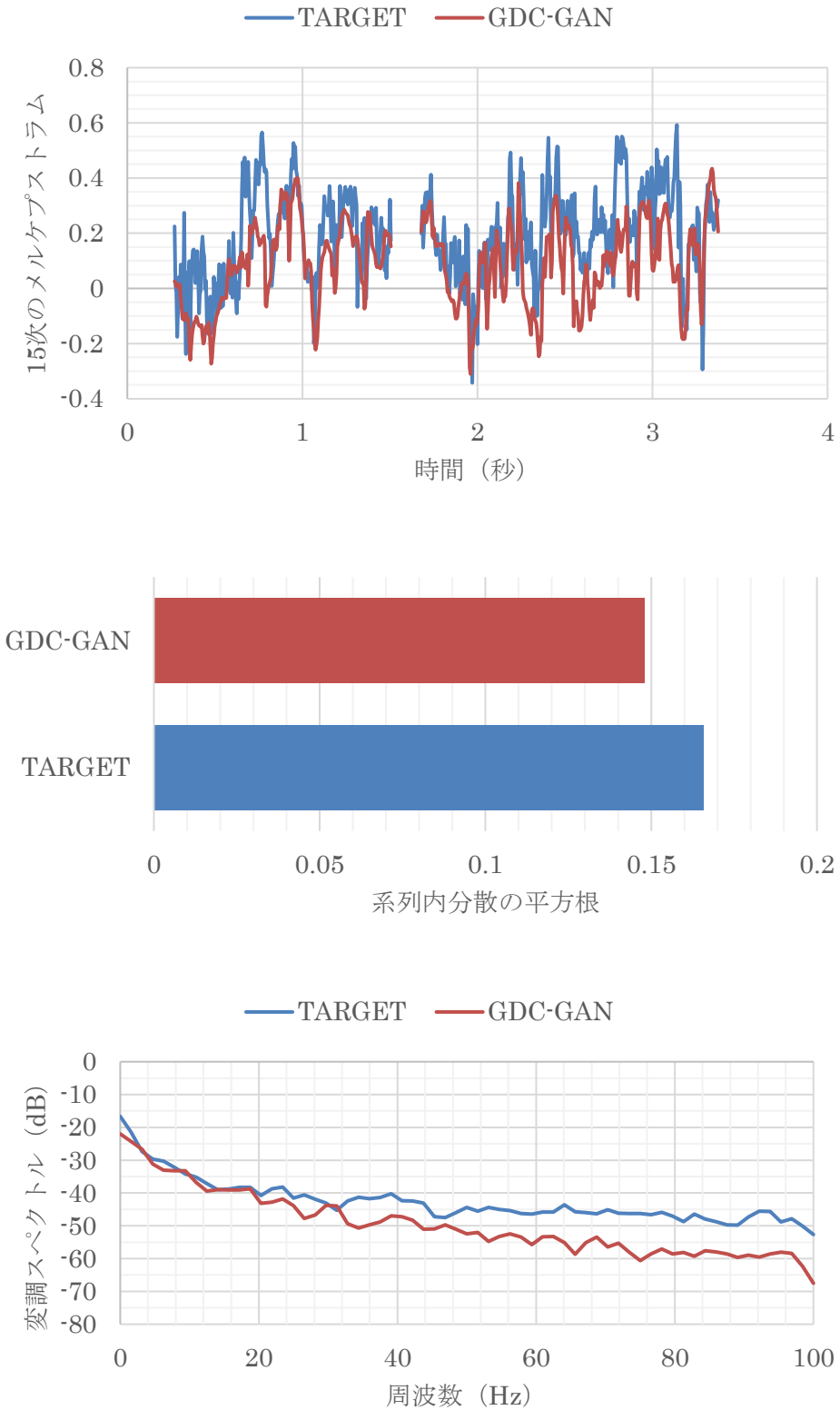

図 6.5　FFNN-GAN のメルケプストラムの代表例

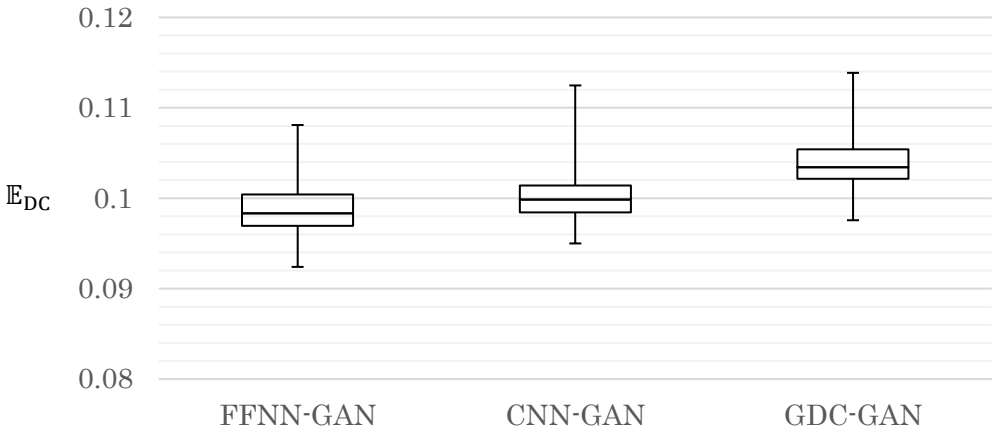

図 6.6　メルケプストラムの平均絶対誤差

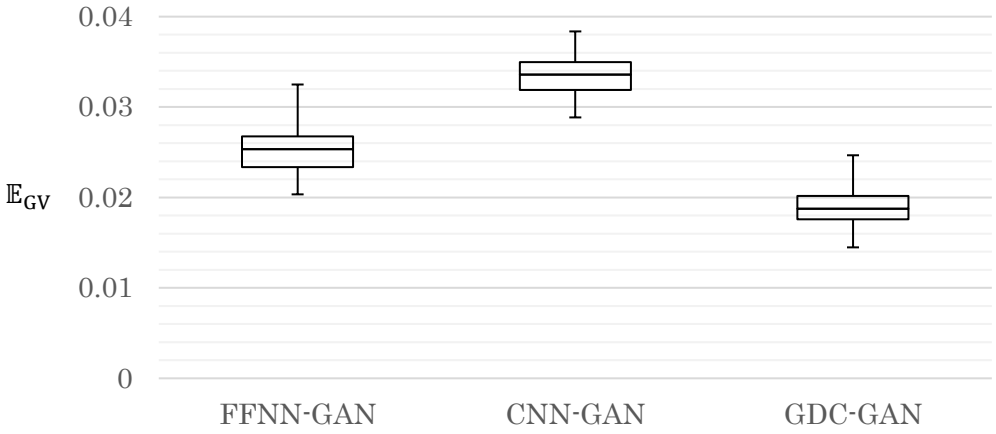

図 6.7　メルケプストラムの系列内分散の平方根の平均絶対誤差

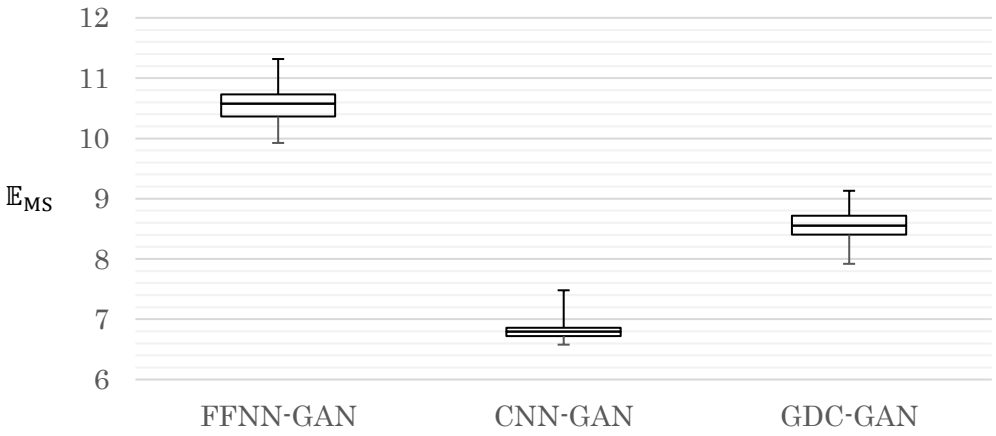

図 6.8　メルケプストラムの変調スペクトルの平均絶対誤差（dB）

表 6.4　Tukey-Kramer 法によるメルケプストラムの$\mathbb{E}_{DC}$の平均値の比較結果
表中の数値はスチューデント化された範囲分布の q 値と p 値である．群数は 3，自由度は 296，信頼区間は 95%である．

群 1	群 2	q 値	p 値
FFNN-GAN	CNN-GAN	5.06	0.002
FFNN-GAN	GDC-GAN	18.77	0.001
CNN-GAN	GDC-GAN	13.71	0.001

表 6.5　Tukey-Kramer 法によるメルケプストラムの$\mathbb{E}_{GV}$の平均値の比較結果
表中の数値はスチューデント化された範囲分布の q 値と p 値である．群数は 3，自由度は 296，信頼区間は 95%である．

群 1	群 2	q 値	p 値
FFNN-GAN	CNN-GAN	38.66	0.001
FFNN-GAN	GDC-GAN	29.61	0.001
CNN-GAN	GDC-GAN	68.27	0.001

表 6.6　Tukey-Kramer 法によるメルケプストラムの$\mathbb{E}_{MS}$の平均値の比較結果
表中の数値はスチューデント化された範囲分布の q 値と p 値である．群数は 3，自由度は 296，信頼区間は 95%である．

群 1	群 2	q 値	p 値
FFNN-GAN	CNN-GAN	165.93	0.001
FFNN-GAN	GDC-GAN	88.36	0.001
CNN-GAN	GDC-GAN	77.57	0.001

6.6. 考察

グラム行列の対角成分は GV に相当するものであり，GDC-DIS が GV を考慮できるため，実験で比較した 3 つの GAN の中で最も$\mathbb{E}_{GV}$が小さかった．また，グラム行列によりメルケプストラムの系列全体の相関関係が考慮されたことの副次的な効果として変調スペクトルの改善が確認できた．これらの効果により，聴取実験の評点が 3 つの GAN の中で最も改善された．

FFNN-GAN の生成モデルと識別モデルは隣接する時間フレーム間のメルケプストラムの関係やメルケプストラムの GV を考慮しないにも関わらず，GV を改善することができた．これは，FFNN-DIS がケプストラム強調のように機能したためだと考える．ケプストラム強調はメルケプストラムの係数を定数倍することによって，スペクトル包絡のフォルマントを強調する．この処理は時間フレームごとに独立して行われる．この点は，FFNN-DIS も同じである．FFNN-DIS が入力されたメルケプストラムの係数の大きさに基づいて

識別を行っていたとすると，識別誤差にメルケプストラムの係数の大きさに関する情報が含まれるため，生成モデルはその情報に基づき，生成誤差で学習されるときよりもメルケプストラムの係数を大きくするように学習されたと考える．

CNN-DIS は，1 次元畳み込み層によって，隣接する時間フレーム間のメルケプストラムの特徴を抽出できるため，その特徴に基づき識別を行う．これにより，メルケプストラムの変調スペクトルの予測誤差を小さくすることができた．しかし，1 次元畳み込み層では，メルケプストラムの系列全体の特徴を捉えることができないため，メルケプストラムの系列内分散の予測誤差は小さくすることができなかった．系列内分散を改善しないまま，変調スペクトルを改善しても，メルケプストラムは過剰平滑化されたまま揺らぎが付与された状態になる．その結果として，粗造性嗄声のような音質になってしまう．

GAN による学習法では，識別モデルの構造によって，学習される生成モデルのモデルパラメータが異なることがわかった．このため，所望の生成モデルのモデルパラメータを獲得するには，音声特徴量の特徴をどのように捉えるかを考慮して識別モデルの構造を決める必要がある．

実験において各 GAN を効果的に機能させるには識別モデルのハイパーパラメータや，勾配法の学習率を調整する必要があった．これらのハイパーパラメータの調整は直感的に行うことが難しく，試行錯誤を繰り返した．試行錯誤により得た知見としては，識別モデルの層数やユニット数を減らして，識別モデルの性能を低くすると比較的安定して生成モデルを学習できた．また，CNN-GAN のハイパーパラメータも何十回と試行錯誤したが，効果的に機能するように調整することはできなかった．ハイパーパラメータの調整については，文献 [46]でも指摘されており，GAN による学習法を効果的に機能させるには，最適なパイパーパラメータの探索が必要となる．

5 章の時系列の複数の属性を考慮した損失関数による学習法と，本章の時系列の相関関係を考慮した GAN による学習法は，どちらも 3.2.3 節で述べた計算資源が限られた音声特徴量予測部に用いられる FFNN が音声特徴量の時間構造を考慮したモデルパラメータを獲得させるという目標は達成している．また，聴取実験の評点や予測誤差を比較しても大きな差はないため，どちらの学習法も同程度の性能といえる．ここでは，更なる合成音声の音質の向上を図るための考察として，複数の損失関数による学習法と GAN による学習法について議論する．

DNN のモデルパラメータは損失関数で算出された誤差に基づいて更新される．そのため，複数の損失関数による学習法は DNN へ伝搬される誤差を直接制御することができる．特に，GV を直接学習することは，対数基本周波数でもメルケプストラムでも有効であった．このように，この学習法は，教師データの特定の特徴を明示的に学習することで，比較的容易に一定の水準まで音声の品質を向上できる．一方で，この学習法の学習基準は定義した損失関数に制限されてしまう．また，複数の損失関数を定義した場合，損失関数間の整合性をとることが困難である．

GAN による学習法では，生成モデルのモデルパラメータは生成誤差と識別誤差に基づいて更新される．識別モデルは教師データと予測データを識別するために，それぞれのデータから識別に必要な特徴を自動で抽出して学習する．この学習された特徴の情報は識別誤差を起点する誤差逆伝播によって生成モデルへと伝搬する．これにより，GAN による学習法は明示的に教師データから特徴量を抽出する必要なく生成モデルを学習できる．また，識別モデルと生成モデルのモデルパラメータは交互に更新されるため，生成モデルの学習基準は常に更新され続ける．一方で，識別モデルは損失関数のように教師データと予測データの差を学習しているわけではないため，損失関数と同じように機能しない．そのため，教師データの特定の特徴を学習することで合成音声の品質を向上させられると分かっていても，識別モデルはその特徴を直接学習することはできない．

このため，どちらの学習法が優れているとはいえない．ただし，GAN による学習法は，生成誤差の算出法を変更することができるため，拡張性については GAN による学習法が優れている．例えば，生成誤差を複数の損失関数で算出することで，複数の損失関数による学習法と GAN による学習法の利点を活かせると考える．このように，本論文で得た知見を活かして新たな学習法を模索し，合成音声の品質を向上させることが今後の課題である．

6.7. まとめ

経験や知見を必要とせず，FFNN が音声特徴量の構造を考慮したモデルパラメータを獲得できるようにする学習法として，時系列の相関関係を考慮した GAN による学習法を提案した．複雑な時間構造を持つメルケプストラムを対象として，合成音声の音質を評価する聴取実験とメルケプストラムの予測誤差により，提案法と従来法を比較した．その結果，提案法は従来法よりも知覚的に優れたメルケプストラムの予測を可能にした．これにより，3.2.3 節で述べた計算資源が限られた音声特徴量予測部の FFNN による合成音声の音質の問題を解決した．

7. 結論

合成音声システムの保守性や制御性を考慮しつつ，計算資源が限られた計算環境においても，頑健かつ高速に動作する音声合成システムを目指すために，音声合成システムの音声特徴量予測部に用いられる DNN の学習法を提案した．本論文では，音声合成システムの保守性や制御性を考慮して，言語解析部，音声特徴量予測部，波形生成部の 3 つのサブシステムのシステム構成とした．サブシステムのうち，音声特徴量予測部を対象として，計算資源が限られた計算環境を想定して，音声特徴量予測部の処理の高速化，音声特徴量の予測の頑健性の向上，音声特徴量の予測精度の改善に取り組んだ．音声特徴量予測部の処理の高速化については，後処理を削減し，単純な構造の DNN である FFNN のみで音声特徴量予測部を構成する必要があった．音声特徴量の予測の頑健性の向上には，DNN が学習外の外れ値に対して脆弱である問題を解決する必要があった．また，FFNN のみで音声特徴量予測部を構成するためには，FFNN が時間フレームごとに独立して音声特徴量をモデル化することにより音声特徴量の予測精度が低下する問題を解決する必要があった．

音声特徴量予測部の処理の高速化については，後処理を削減し，単純な構造の DNN である FFNN のみで音声特徴量予測部を構成することで，計算資源が限られた計算環境に適した処理時間やモデルサイズになることを明らかにした．DNN が外れ値に対して脆弱である問題に対しては，2 つの言語特徴量の属性値の比を取る正規化法を提案し，正規化後の言語特徴量に外れ値が含まれないようにした．この正規化法は，学習外の外れ値を含む言語特徴量に対して，基本周波数を頑健に予測することを可能にした．FFNN が時間フレームごとに独立して音声特徴量をモデル化することにより合成音声の品質が低下する問題に対しては，時系列の複数の属性を考慮した損失関数による学習法と，時系列の相関関係を考慮した敵対的ネットワークによる学習法を提案した．時系列の複数の属性を考慮した損失関数は，複数の損失関数により音声特徴量を多角的に捉えた誤差を算出することで，FFNN が音声特徴量の時間構造や次元間の関係を考慮したモデルパラメータを獲得できることを可能にした．時系列の相関関係を考慮した敵対的ネットワークによる学習法は，生成モデルが音声特徴量の時間フレームごととの特徴を捉え，識別モデルが音声特徴量の系列全体の特徴を捉えることで，FFNN が音声特徴量の時間構造や次元間の関係を考慮したモデルパラメータを獲得できることを可能にした．

このように，本研究により，計算資源が限られた計算環境においても，頑健かつ高速に音声特徴量を予測する深層学習モデルを用いた音声特徴量予測部を実現することができた．

8. 参考文献

[1] 古井貞熙,“新音響・音声工学,”近代科学社, pp. 102, 163-166, 2006.

[2] D. H. Klatt,“Review of text-to-speech conversion for English,”The Journal of the Acoustic Society of America, vol.82, no.3, pp.737-793, 1987.

[3] F. Charpentier and M. Stella,“Diphone synthesis using an overlap-add technique for speech waveforms concatenation,”ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 2015-2018, Tokyo, Japan, 1986.

[4] M. Morise, F. Yokomori, and K. Osawa,“WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” IEICE Transactions on Information and Systems, vol.E99-D, issue 7, pp. 1877-1884, 2016.

[5] H. Kawahara, M. Morise, and T. Takahashi,“Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation,”ICASSP 2008, IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, 2008.

[6] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura,“Speech parameter generation algorithms for HMM-based speech synthesis,” ICASSP 2000, IEEE International Conference on Acoustics, Speech and Signal Processing, Istanbul, Turkey, 2000.

[7] G. E. Hinton and R. R. Salakhutdinov,“Reducing the Dimensionality of Data with Neural Networks,”Science, vol.313, issue 5786, pp.504-507, 2006.

[8] F. Seide, G. Li, and D. Yu,“Conversational Speech Transcription Using Context-Dependent Deep Neural Networks,” INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, 2011.

[9] H. Zen, A. Senior, and M. Schuster,“Statistical Parametric Speech Synthesis Using Deep Neural Network,”ICASSP 2013, IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, 2013.

[10] T. Mikolob, K. Chen, G. Corrado, and J. Dean,“Efficient Estimation of Word Representations in Vector Space,”arXiv:1301.3781, 2013.

[11] J. Sotelo, S. Mehri, K. Kumar, J. Santos, K.Kastner, A. Courville, and Y. Bengio,“Char2Wav: End-to-End speech synthesis,”ICLR workshop, 2017.

[12] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. Saurous, “Tacotron: A fully end-to-end text-to-speech synthesis model,” Proc. Interspeech, 2017.

[13] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “WaveNet: a generative model for raw audio,” arXiv.1609.03499, 2016.

[14] N. Kalchbrenner, E. Elsen, K. Simonyaan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, “Efficient Neural Audio Synthesis,” arXiv:1802.08435, 2018.

[15] J. Valin and J. Skoglund, “LPCNET: Improving Neural Speech Synthesis through Linear Prediction,” ICASSP 2019, IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, United Kingdom, 2019.

[16] R. Prenger, R. Valle, and B. Catanzaro, “WaveGlow: A Flow-based Generative Network for Speech Synthesis,” arXiv:1811.00002, 2018.

[17] 全炳河, “テキスト音声合成技術の変遷と最先端,” 日本音響学会誌, vol.74, no.7, pp.387-393, 2018.

[18] 山岸順一, 徳田恵一, 戸田智基, みわよしこ, “おしゃべりなコンピュータ ― 音声合成ぎじゅつの現在と未来,” 丸善出版株式会社, 2015.

[19] 工藤 拓, 山本 薫, 松本 裕治, “Conditional Random Fields を用いた日本語形態素解析,” 自然言語処理研究会報告, vol.161, pp.89-96, 2004.

[20] Nara Institute of Science and Technology, “ChaSen,” 2007. [オンライン]. Available: https://chasen-legacy.osdn.jp/.

[21] 匂坂 芳典, 佐藤 大和, “日本語単語連鎖のアクセント規則,” 電子通信学会論文誌 D 66(7), p849-856, 1983.

[22] K. Tokuda, K. Oura, K. Hashimoto, K. Sawada, T. Yoshimura, S. Takaki, H. Zen, J. Yamagishi, T. Toda, T. Nose, S. Sako, and A. W. Black, “HMM/DNN-based Speech Synthesis System (HTS),” 2017. [オンライン]. Available: http://hts.sp.nitech.ac.jp/.

[23] M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis,” Speech Communication, vol. 84, pp. 57-65, Nov. 2016.

[24] L. Bottou, F.E. Curtis, J. Nocedal, “Optimization Methods for Large-Scale Machine Learning,” arXiv:1606.04838, 2016.

[25] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Incorporating a mixed excitation model and postfilter into HMM-based text-to-

speech synthesis,” Systems and Computers in Japan, volume 36, issue 12, 2005.

[26] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Comunication, vol.9, no.4, pp.357-363, 1990.

[27] L. L. Beranek, “Balanced Noise Criterion (NCB) Curves,” The Journal of the Acoustical Society of America, 86(2), p.650-664, 1989.

[28] 高橋遼太, 能勢隆, 伊藤彰則, “HMM 音声合成におけるアクセントラベリング基準が合成音声に与える影響の分析,” 情報処理学会研究報告, vol. 2015-SLP-106, no.1, 2015.

[29] K. Tokuda, T. Kobayashi, T. Fukada, H. Saito, and S. Imai, “Spectral estimation of speech based on mel-cepstral representation,” Journal of IEICE, Vol.J74-A, No.8, pp.1240–1248, 1991.

[30] K. Tokuda, K. Oura, T. Yoshimura, A. Tamamori, S. Sako, H. zen, T. Nose, T. Takahashi, J. Yamagishi, and Y. Nankaku, “Speech Signal Processing Toolkit (SPTK),” 2017. [オンライン]. Available: http://sp-tk.sourceforge.net/.

[31] H. Zen and H. Sak, “Unidirectional Long Short-Term Memory Recurrent Neural Network with Recurrent Output Layer for Low-Latency Speech Synthesis,” ICASSP 2015, IEEE International Conference on Acoustics, Speech and Signal Processing, Brisbane, Australia, 2015.

[32] V. Nair, and G. E. Hinton, “Rectified Linear Units Improve Restricted Boltzmann Machines,” ICML 2010, The 27th International Conference on Machine Learning, Haifa, Israel, 2010.

[33] T. Toda, T. Muramatsu, and H. Banno, “Implementation of computationally efficient real-time voice conversion,” Proc. INTERSPEECH 2012, pp.94-97, USA, 2012.

[34] D. P. Kingma, and J. Ba, “Adam: A Method for Stochastic Optimization,” arXiv:1412.6980, 2014.

[35] M. T. Ribeiro, S. Singh, and C. Guestrin, ““Why Should I Trust You?” Explaining the Predictions of Any Classifier,” KDD 2016, 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, USA, 2016.

[36] Z. Wu and S. King, “Minimum trajectory error training for deep neural networks combined with stacked bottleneck features,” INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, 2015.

[37] T. Nose, V. Chunwijitra, and T. Kobayashi, "A parameter Generation Algorithm Using Local Variance for HMM-Based Speech Synthesis," IEEE Journal of Selected Topics in Signal Processing, vol. 8, no. 2, pp. 221-228, 2014.

[38] T. Toda and K. Tokuda, "A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis," IEICE Transactions on Information and Systems, E90-D (5), pp.816-824, 2007.

[39] International Telecommunication Union, "Recommendation ITU-R BS.1534-1: Methods for subjective assessment of intermediate quality level of coding systems," 2015.

[40] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A postfilter to modify the modulation spectrum in HMM-based speech synthesis," ICASSP 2014, IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, 2014.

[41] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, "Generative Adversarial Nets," NIPS 2014, Neural Information Processing Systems 27, 2014.

[42] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 1, 2018.

[43] S. Yang, L. Xie, X. Chen, X. Lou, X. Zhu, D. Huang, and H. Li, "Statistical Parametric Speech Synthesis Using Generative Adversarial Networks Under A Multi-task Learning Framework," arXiv:1707.01670, 2017.

[44] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," ICLR 2016, International Conference on Learning Representations, San Juan, Puerto Rico, 2016.

[45] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier Nonlinearities Improve Neural Network Acoustic Models," ICML 2013, The 30th International Conference on Machine Learning, Atlanta, USA, 2013.

[46] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, "Are GANs Created Equal? A Large-Scale Study," arXiv:1711.10337, 2017.

9. 謝辞

学位論文をまとめるにあたり，多くの方々にご指導とご助力をいただきました．主査の富山県立大学 平原達也教授には，研究の枠組みについて有益な助言をいただきました．深く感謝申し上げます．副査の北陸先端科学技術大学 赤木正人教授，富山県立大学 神谷和秀教授，富山県立大学 小柳健一教授，富山県立大学 Parham Mokhtari 准教授には，学位論文について有益なご指摘をいただきました．深く感謝します．株式会社エーアイ 吉田大介社長，廣飯伸一副社長には，社会人として博士後期課程への進学および研究全般に渡るご支援を賜りました．深く感謝申し上げます．株式会社エーアイ 大谷大和氏には，研究を遂行するにあたり有益な助言をいただきました．深く感謝申し上げます．最後に，音声コーパスの作成や実験に協力してくださったすべての方々にお礼を申し上げるとともに，日々の生活を支えて下さった妻と両親に感謝の意を表して謝辞といたします．

10. 発表論文リスト

学術論文

[1] 松永悟行, 大谷大和, 平原達也, “深層学習を用いた日本語音声合成における基本周波数に適した言語特徴量の正規化手法,” 電子情報通信学会論文誌 D, Vol.J102-D, No.10, pp.721-729, 2019.

[2] N. Matsunaga, Y. Ohtani, and T. Hirahara, “Loss function considering multiple attributes of a temporal sequence for feed-forward neural networks,” IEICE Transactions, Vol.E103-D, No.12, 2020.

国際学会プロシーディング論文（査読あり）

[1] N. Matsunaga, Y. Ohtani, and T. Hirahara, “Loss function considering temporal sequence for feed-forward neural network – fundamental frequency case,” The 10th ISCA Speech Synthesis Workshop, Vienna, Austria, 2019.

口頭発表

[1] 松永悟行, 大谷大和, 平原達也, “深層学習に基づく日本語音声合成における基本周波数ための言語特徴量の正規化手法の検討,” 日本音響学会 2019 年春季研究発表会, 1-P-21, 2019.

[2] 松永悟行, 大谷大和, 平原達也, “深層学習に基づく音声合成における 2 次統計量を用いたスペクトル特徴量のモデリングの検討,” 日本音響学会 2019 年秋季研究発表会, 1-P-23, 2019.