修士論文

非負値自己符号化器に基づく スペクトルモデリングを用いた DNN音声合成

2022 年 1 月 27 日

指導教員　齋藤 大輔 准教授

東京大学大学院工学系研究科
電気系工学専攻　37-206477

五来 丈瑠

内容梗概

テキスト音声合成 (TTS) とはテキストからそれに対応する音声を推定する技術であり，アナウンスやスマートスピーカーなど，様々なアプリケーションに用いられている．近年，テキストから音声信号等を直接推定する end-to-end 音声合成モデルが提案され，高い自然性を達成している．それに対し従来の統計的パラメトリック音声合成 (SPSS) はいくつかのモジュールに分かれた音声合成システムであり，DNN 等により言語特徴量からスペクトル包絡や基本周波数などの音響特徴量を推定し，ボコーダによって波形生成を行うという枠組みが一般的である．SPSS は比較的少量のデータで動作する，話者性の変更が容易である等の利点があるが，合成音声の品質は end-to-end 音声合成に劣る．

品質低下の大きな要因としてボコーダの利用が挙げられる．ボコーダとは音声信号から分析が容易な音響特徴量を抽出し，また逆に音響特徴量から音声信号を生成する技術であるが，分析及び合成の過程で劣化を生ずる．また，ボコーダパラメータのうち声道特徴に対応するスペクトル包絡は高次元で冗長な特徴量であるため，音響モデルの学習時に低次元の特徴量に圧縮することが一般的である．メルケプストラムはこのような低次元特徴量として広く用いられている声道パラメータであるが，スペクトル包絡の微細構造を過剰に平滑化しているという欠点があり，合成音声の劣化の一因となっている．

本研究では SPSS における合成音声の品質向上を目指し，スペクトルモデリングに非負値自己符号化器 (NAE) を用いた TTS システムを提案する．NAE の潜在変数の非負性により，スペクトル包絡の微細構造が失われにくいような次元削減が行われることが期待される．単一話者の音声データを用いた評価実験により提案法の有効性が示された．またその応用として，話者ごとに別々に NAE を学習することによる複数話者モデリングについても検討を行った．評価実験では，NAE のみを話者ごとに学習することによって話者適応が適切に行われることを確認した．

さらに本研究では，ボコーダによる波形生成の過程で起こる合成音声の品質低下を改善するため，音響特徴量から音声信号の推定にニューラルボコーダの一種である Neural Source Filter (NSF) を用いたモデルについて検討した．まず自然音声を用いて学習した NSF を提案した TTS システムにそのまま適用した．実験により NSF の学習時には自然音声，推論時には音響モデルにより合成した特徴量を入力することが原因で自然性が低下することが明らかになった．そこで学習時と推論時における入力特徴量のミスマッチを改善するため，自然音声に対し NAE による再構成を適用し合成した音響特徴量に近づけたものを NSF の学習データとして用いるという枠組みを検討した．評価実験では合成音声の品質の向上が示された．

目次

第 1 章　序論 **1**
1.1　本研究の背景 2
1.2　本研究の目的 2
1.3　本論文の構成 3

第 2 章　テキスト音声合成に関する基礎技術 **4**
2.1　統計的音声合成の概要 5
2.2　音響特徴量の抽出 6
2.2.1　短時間フーリエ変換による音声分析 6
2.2.2　ソースフィルタモデルに基づく音声分析 6
2.2.3　ボコーダによる音声分析 8
2.2.4　声道特徴量の次元削減 8
2.3　音声信号の生成 9
2.3.1　WORLD における波形生成 9
2.3.2　ニューラルボコーダによる波形生成 9
2.4　複数話者音声合成 10

第 3 章　NAE を用いた音声合成 **12**
3.1　はじめに 13
3.2　NMF とその音声合成への応用 13
3.3　NMF を用いた TTS 13
3.4　Non-negative Autoencoder 13
3.5　NAE を用いたテキスト音声合成 14
3.6　評価実験 15
3.6.1　実験条件 15
3.6.2　実験結果 17

第 4 章　NAE に基づく複数話者モデリング **18**
4.1　NAE に基づく話者適応 19
4.2　評価実験 19
4.2.1　実験条件 19
4.2.2　実験結果 20
4.3　話者コードを用いた場合の実験 20
4.3.1　実験条件 21
4.3.2　実験結果 21
4.4　基本周波数及び非周期性指標についても推定を行う場合の実験 21

4.4.1 実験条件 . . . 21
4.4.2 実験結果 . . . 22

第 5 章 Neural Source Filter を用いた合成音声の高品質化 23
5.1 はじめに . . . 24
5.2 Neural Source Filter . . . 24
5.3 評価実験 . . . 24
5.3.1 実験条件 . . . 24
5.3.2 実験結果 . . . 26
5.3.3 まとめ . . . 27
5.4 非周期性指標を補助特徴量に加えた場合の実験 . . . 27
5.4.1 実験条件 . . . 28
5.4.2 実験結果 . . . 28
5.5 WORLD を用いた合成音声の最小位相化 . . . 29
5.5.1 実験条件 . . . 29
5.5.2 実験結果 . . . 29

第 6 章 NSF における音響的ミスマッチの低減 32
6.1 はじめに . . . 33
6.2 合成した音響特徴量を NSF の学習に用いた場合の実験 . . . 33
6.2.1 実験条件 . . . 33
6.2.2 結果及び考察 . . . 33
6.3 NAE により再構成した音響特徴量を NSF の学習に用いた場合の実験 . . . 33
6.3.1 用いた手法 . . . 34
6.3.2 実験条件 . . . 34
6.3.3 実験結果 . . . 36

第 7 章 結論 38
7.1 まとめ . . . 39
7.2 今後の展望 . . . 39

謝辞 41

参考文献 42

発表文献 46

図目次

2.1 統計的パラメトリック音声合成の概念図 6
2.2 短時間フーリエ変換 7
2.3 メルケプストラムへの変換による平滑化 8
2.4 DAE をスペクトルモデリングに用いた TTS システムの学習手順 9
2.5 dilated causal convolution の構造 10
2.6 DNN 音声合成における話者適応のアプローチ 11

3.1 NMF を用いた TTS の概念図 14
3.2 NAE と DNN 音響モデルの同時学習 15
3.3 提案する TTS システムの全体像 (テスト時) 16
3.4 合成音声の自然性に関する主観評価実験の結果 17

4.1 提案手法における複数話者モデリング 19
4.2 合成音声の自然性及び話者性に関する主観評価実験の結果 20

5.1 Neural Source Filter モデルの全体像 25
5.2 NSF による合成音声の品質に関する主観評価実験の結果 (自然音声) 27
5.3 NSF による合成音声の品質に関する主観評価実験の結果 (TTS) 27
5.4 非周期性指標を加えた実験における合成音声の品質に関する主観評価実験の結果 . 28
5.5 WORLD の再合成による波形の変化．男性話者の評価用音声 (p226_351) より抽出した． 30
5.6 NSF の合成音声に対する WORLD による最小位相化の効果 (自然音声) 31
5.7 NSF の合成音声に対する WORLD による最小位相化の効果 (TTS) 31

6.1 合成した声道特徴量を用いて学習した NSF の性能の評価 34
6.2 提案する TTS システムの概要 35
6.3 NSF による合成音声の品質に関する主観評価実験の結果 36

表目次

3.1 メルケプストラム歪みによる合成音声の品質の評価 17

4.1 メルケプストラム歪みによる合成音声の品質の評価 20
4.2 メルケプストラム歪みによる合成音声の品質の評価 21
4.3 F_0，bap，V/UV についても推定を行う実験における客観評価の結果 22

5.1 誤差の算出時に用いた STFT のパラメータ 26
5.2 NSF による合成音声の客観評価結果 (自然音声) 26
5.3 NSF による合成音声の客観評価結果 (TTS) 26
5.4 非周期性指標を加えた実験における合成音声の客観評価結果 28

6.1 自然音声，NAE の出力，音響モデルの出力のメルケプストラム距離 36
6.2 NSF による合成音声の客観評価結果 37

第1章

序論

1.1　本研究の背景

テキスト音声合成 (Text-to-Speech Synthesis; TTS) とはテキストからそれに対応する音声を推定する技術であり，アナウンスやスマートスピーカーなど，様々なアプリケーションに用いられている．近年，テキストから音声信号等を直接推定する end-to-end 音声合成モデルが高い自然性を達成している [1, 2].

それに対し従来の統計的パラメトリック音声合成 (Statistical Parametric Speech Synthesis; SPSS) はいくつかのモジュールに分かれた音声合成システムであり，DNN 等により言語特徴量からスペクトル包絡や基本周波数などの音響特徴量を推定し，ボコーダによって波形生成を行うという枠組みが一般的である．SPSS は比較的少量のデータで動作する，話者性の変更が容易である等の利点があるが，合成音声の品質は end-to-end 音声合成に劣る，

品質低下の原因の一つにボコーダの利用が挙げられる，ボコーダとは，音声信号から分析が容易な音響特徴量を抽出し，また逆に音響特徴量から音声信号を生成する技術である，従来の信号処理技術に基づくボコーダは波形生成時に品質低下を生ずる，この問題を解消するため，WaveNet ボコーダ [3, 4] をはじめとするニューラルボコーダが提案され，より肉声に近い音声が合成可能となったが，学習及び推論に時間を要する，学習が容易でないなどの欠点が存在する，

また，ボコーダにより抽出されたスペクトル包絡は高次元で冗長な特徴量であるため，前段で適切な低次元特徴量に圧縮することが音響モデルを精緻に学習するうえで必要であることが示唆されている [5]．低次元特徴量として最も一般的なものがメルケプストラムである [6]．メルケプストラムは対数スペクトルに離散フーリエ変換を行い低次元成分を抽出することで得られる特徴量であり，スペクトル包絡を低次元のパラメータで効率的に表現できることが知られているが，スペクトル包絡の微細構造を過剰に平滑化しているという欠点があり，合成音声の劣化の一因となっている，

そこで，スペクトルモデリングをより精緻に行うための統計的手法についていくつか提案がされている．その一つに deep autoencoder (DAE) を用いた手法 [5] が挙げられる．この手法では，DAE の潜在変数を音響特徴量として用いることによりメルケプストラムよりも音声合成に適した次元削減を行っている．

また，非負値行列因子分解 (Non-negative Matrix Factorization; NMF) を用いた手法 [7] も提案されている．この手法では NMF によりスペクトル包絡をスペクトルのテンプレートである基底行列と，そのスパースな重みであるアクティベーションの積に分解し，アクティベーションを音響特徴量として用いることでスペクトルの微細な構造を失わずに合成を行うことが可能となっている，

1.2　本研究の目的

本研究では，SPSS における合成音声の品質向上を目指し，非負値自己符号化器 (Non-negative AutoEncoder; NAE) [8] によるスペクトルモデリングを用いた TTS を提案する．NAE とは NMF を autoencoder の形に置き換えることでニューラルネットワークに拡張したモデルであり，より柔軟に動作することが知られている．また，DNN に組み込むことができるため，NAE による低次元特徴量の抽出と DNN 音響モデルを同時に学習できるという利点もある．

またその応用として，話者ごとに別々に NAE を学習することによる複数話者モデリングについても検討を行う，

さらに NAE に基づく TTS と Neural Source Filter モデルを組み合わせることにより自然性の向上を図る，

1.3　本論文の構成

本論文は，全 7 章で構成される，第 2 章では，統計的音声合成に関する基礎技術と課題を主に音響的側面から述べる，第 3 章では，音響特徴量の次元削減に NAE を用いた TTS を提案し，実験的評価を行う，第 4 章では，NAE に基づく複数話者モデリングについて検討する，第 5 章では，NAE を用いた TTS システムにおいて，ボコーダを Neural Source Filter (NSF) モデルに置き換えることにより合成音声の品質の改善を試みる，第 6 章では，学習データとテストデータの音響的性質が異なることにより NSF の合成音声が劣化する問題を解決するため，NAE により再構成された音響特徴量を用いて NSF を学習する手法について検討する，最後に第 7 章で本研究での成果をまとめ，今後の展望について述べる，

第2章

テキスト音声合成に関する基礎技術

2.1　統計的音声合成の概要

テキスト音声合成 (Text-to-Speech Synthesis; TTS) は，テキストとそれに対応する音声の組があるとき，与えられたテキストに対応する音声信号を推定する技術である．

コーパスベースのテキスト音声合成には，波形接続に基づく手法 [9] と統計的手法がある．波形接続に基づく手法では，大規模な音声データベースに含まれる自然音声の素片を接続することで合成音声が得られる．そのため自然性は高いものの，話者性や発話スタイルの制御は難しい．

一方，統計的音声合成は音声データにより統計モデルを学習し直接合成音声を得る手法であり，話者性などの変更が可能である．近年，深層学習の発展などにより盛んに研究が行われており，合成音声の自然性が大きく向上している．統計的音声合成の基本問題は，学習データに含まれるテキスト $\mathcal{W}$ と音声信号 $\mathcal{X}$ の組を用いて，テストデータのテキスト $\boldsymbol{w}$ から音声信号 $\boldsymbol{x}$ を推定することであるが，そのまま音声信号の予測分布をモデリングすることは困難である．そこで統計的パラメトリック音声合成 (Statistical Parametric Speech Synthesis; SPSS) ではこの基本問題を次式の比較的容易に計算できる副問題に分解する [10].

$$\hat{\mathcal{O}} = \arg\max_{\mathcal{O}} p(\mathcal{X} \mid \mathcal{O}) \qquad \text{(音響特徴量の抽出)} \tag{2.1}$$

$$\hat{\mathcal{L}} = \arg\max_{\mathcal{L}} P(\mathcal{L} \mid \mathcal{W}) \qquad \text{(言語特徴量の抽出)} \tag{2.2}$$

$$\hat{\lambda} = \arg\max_{\lambda} p(\hat{\mathcal{O}} \mid \hat{\mathcal{L}}, \lambda) \qquad \text{(音響モデルの学習)} \tag{2.3}$$

$$\hat{\boldsymbol{l}} = \arg\max_{\boldsymbol{l}} P(\boldsymbol{l} \mid \boldsymbol{w}) \qquad \text{(言語特徴量の抽出)} \tag{2.4}$$

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} p(\hat{\boldsymbol{o}} \mid \hat{\boldsymbol{l}}, \hat{\lambda}) \qquad \text{(音響特徴量の推定)} \tag{2.5}$$

$$\hat{\boldsymbol{x}} \sim p(\boldsymbol{x} \mid \hat{\boldsymbol{o}}) \qquad \text{(音声信号の生成)} \tag{2.6}$$

$\mathcal{O}$ 及び $\boldsymbol{o}$ は音響特徴量と呼ばれ，一般にボコーダにより抽出した，音声の特徴を表すパラメータが用いられる．$\mathcal{L}$ 及び $\boldsymbol{l}$ は言語特徴量と呼ばれ，テキストから音素や品詞，アクセントなどの情報を抽出しベクトル化した特徴量である．音響モデル λ は言語特徴量から音響特徴量への変換を行う統計モデルである．従来は隠れマルコフモデル (hidden Markov model; HMM) が用いられていたが [11]，DNN を音響モデルに用いた手法が品質を大幅に向上させたことから [12]，近年では DNN に基づく手法が広く用いられている．SPSS ではこれらの副問題の最適化を逐次的に行う．概念図を図 2.1 に示す．

ここで言語特徴量から音響特徴量へのマッピングを学習するにあたり，各音素ラベル音響特徴量のどのフレームに対応しているかを決定し，言語特徴量に付与する必要がある．そのため一般的な SPSS システムでは音響モデルとは別に各音素の長さを推定する継続長モデルが用いられる．ただし，本研究では音響特徴量の分析及び合成に主眼を置くため，予めアライメントを行ったテキストラベルを用いることとする．次節以降では，音響特徴量の抽出と音声信号の生成についてより詳細に述べる．

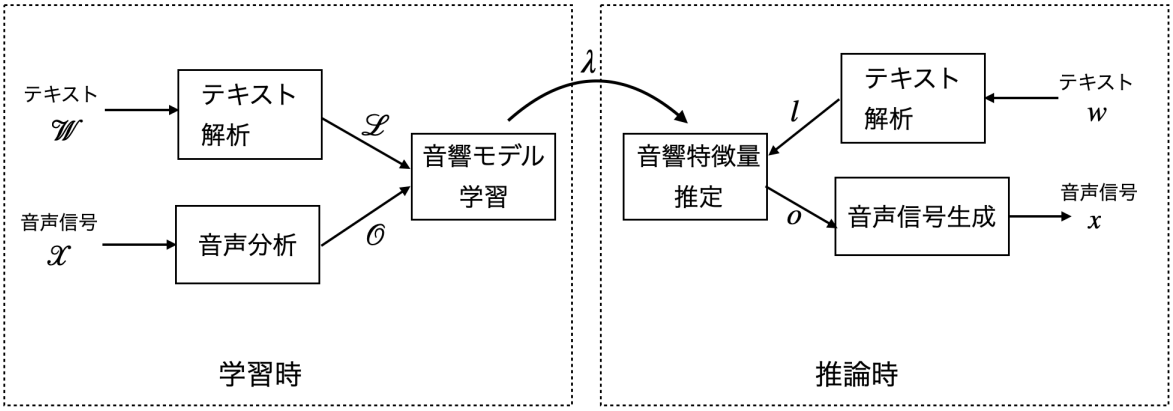


図 2.1: 統計的パラメトリック音声合成の概念図

2.2　音響特徴量の抽出

2.2.1　短時間フーリエ変換による音声分析

音声処理の分野において，音声信号に対し短時間フーリエ変換 (Short-Time Fourier Transform; STFT) を行い時間周波数領域の特徴量 (スペクトログラム) に変換して分析を行うことが一般的である．STFT は図 2.2 に示す手順で行われる．まず元の音声信号から短時間の信号の素片 (フレーム) を窓関数を掛けることにより切り出す．次に切り出した信号を離散フーリエ変換 (Discrete Fourier Transform; DFT) によりスペクトルに変換する．この処理を窓の位置をずらしながら反復することでスペクトログラムが得られる．このときのフレームの長さをフレーム長，窓をずらす間隔をフレームシフトと呼ぶ．DFT により得たスペクトルは複素数の値を持つが，人間の位相特性に対する聴覚の感度が低いことから振幅スペクトルに変換して分析を行うことが一般的である．

2.2.2　ソースフィルタモデルに基づく音声分析

人間の音声は，声帯振動で生じた音源信号（ソース）が声道を通過する過程で音色付け（フィルタ）されることで生成される．ソースフィルタモデルとは，この生成過程に着目した数理モデルであり，ソースとフィルタの独立性を仮定することで前節で述べたスペクトログラムをより扱いやすい因子に分解できる．

声帯振動はおよそ周期的な信号であることが知られている．周期を T，声帯振動一回分の応答を $g(t)$ とすると，音源信号は次式で近似できる．

$$g(t) \otimes \sum_{n=-\infty}^{\infty} \delta\left(t - nT\right) \tag{2.7}$$

$g(t)$ のスペクトルを $G(\omega)$，声道フィルタのスペクトルを $H(\omega)$ とすると音声信号は周期 T のパルス列と次式で表される $Y(\omega)$ の畳み込みによって表現される [13].

$$Y(\omega) = G(\omega)H(\omega) \tag{2.8}$$

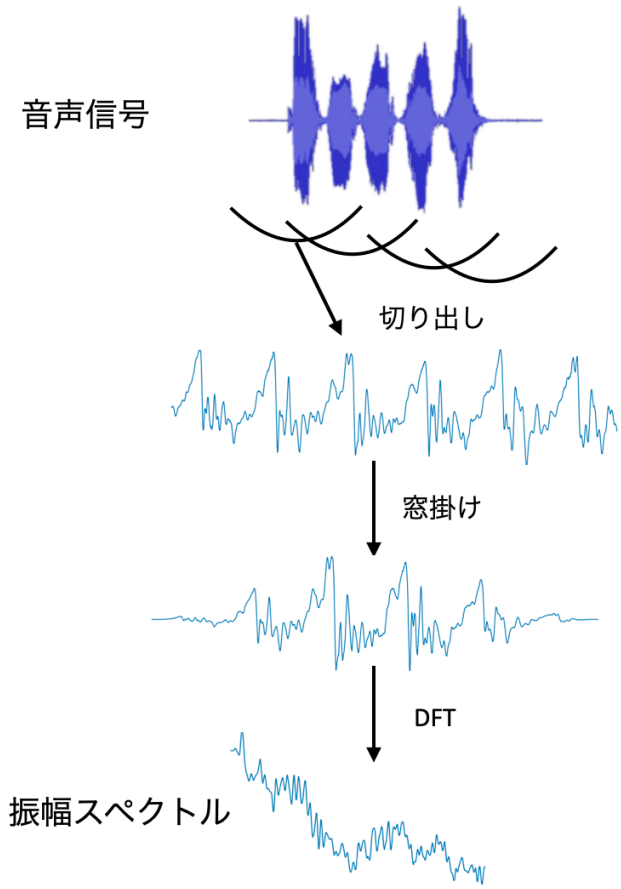


図 2.2: 短時間フーリエ変換

$Y(\omega)$ はスペクトル包絡と呼ばれ，スペクトルから声帯振動による櫛形の微細構造を取り除いた概形を表す特徴量である．スペクトル包絡は音声の音色に対応している．また，周期 T の逆数は基本周波数 (F_0) と呼ばれ，声の高さに対応した特徴量である．

ここまでで音源信号は完全に周期的な信号として扱ったが，実際には非周期的な雑音成分が含まれている．そこで信号全体のパワーのうち非周期成分が占める割合がパラメータとして導入されており，非周期性指標と呼ばれる．この比は帯域に依存するため，非周期性指標を $A_p(\omega)$ と表すとスペクトル包絡との間には次式の関係が存在する．

$$Y(\omega) = Y(\omega)(1 - A_p(\omega)) + Y(\omega)A_p(\omega) \tag{2.9}$$

ソースフィルタモデルでは，周期性成分 $Y(\omega)(1 - A_p(\omega))$ とパルス列の畳み込みと，非周期性成分 $Y(\omega)A_p(\omega)$ とホワイトノイズの畳み込みの和で音声信号を近似することが一般的である．以上のモデル化により音声信号は基本周波数，スペクトル包絡，非周期性指標の 3 つのパラメータ系列に分解可能である．

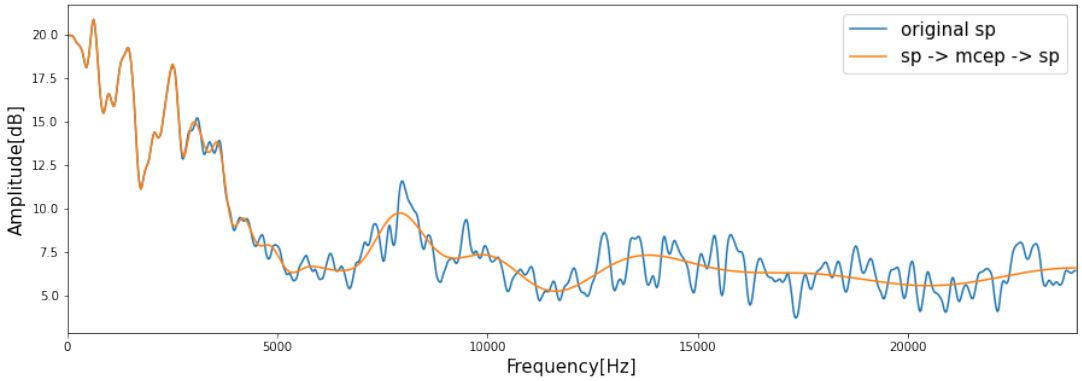


図 2.3: メルケプストラムへの変換によるスペクトルの微細構造の平滑化．青線は元のスペクトル，オレンジ線は一度メルケプストラムに圧縮した後元に戻したスペクトルを表す．

2.2.3　ボコーダによる音声分析

ボコーダとは，音声信号を分析が容易なパラメータへと符号化し，また逆にパラメータから音声信号を生成する技術である．SPSS における音響特徴量には STRAIGHT [14] や WORLD [15] など，ソースフィルタモデルに基づいたボコーダにより抽出されたパラメータが一般に用いられる．これらのボコーダでは主に信号処理技術を用いて音声信号から基本周波数，スペクトル包絡，非周期性指標が抽出される．

2.2.4　声道特徴量の次元削減

前節で述べた音響特徴量のうちスペクトル包絡は高次元なパラメータであるため，より低次元の特徴量になるよう圧縮を行い DNN 等の学習を行うことが一般的である．

SPSS システムにおいて広く用いられている声道特徴量がメルケプストラム (Mel-Frequency Cepstrum; MCEP) である．ケプストラムは対数振幅スペクトルに対し逆フーリエ変換を行い，その低次元成分を抽出することで得られる特徴量であり，スペクトルを少量のパラメータで表現可能である．メルケプストラムはケプストラムを算出する際にメル尺度に従い周波数伸縮を行ったものである．メル尺度とは低周波領域では感度が高く，高周波領域では感度が低いという人間の聴覚特性を考慮した尺度である．したがってメルケプストラムは聴覚上の品質低下をなるべく伴わないよう声道特徴を効率良く圧縮したパラメータであるといえる．そのため，音声の分析及び合成の研究において広く用いられている [16]．しかし，図 2.3 に示すようにメルケプストラムはスペクトル包絡の微細構造を過剰に平滑化するという欠点があり，合成音声の品質低下の要因となる．

そこで，統計的アプローチを用いてより適切なスペクトルパラメータを得る手法がいくつか提案されている．[5] では deep autoencoder (DAE) を用いて低次元の声道特徴量を抽出している．この手法ではまず DAE が入力した対数スペクトル包絡を再構成するように学習を行う．次に言語特徴量から DAE により抽出された低次元の潜在変数への変換を DNN 音響モデルにより学習する．最後に学習済みの DAE の Decoder と音響モデルを連結し，スペクトル包絡を直接推定するネットワークとして fine-tuning を行う．概念図を図 2.4 に示す．評価実験では，スペクトル包絡の次元削減及び連結学習が適切に行われたため，メルケプストラムを用いたモデルと比較して自然性の高い音声が合成されたことが示されている．

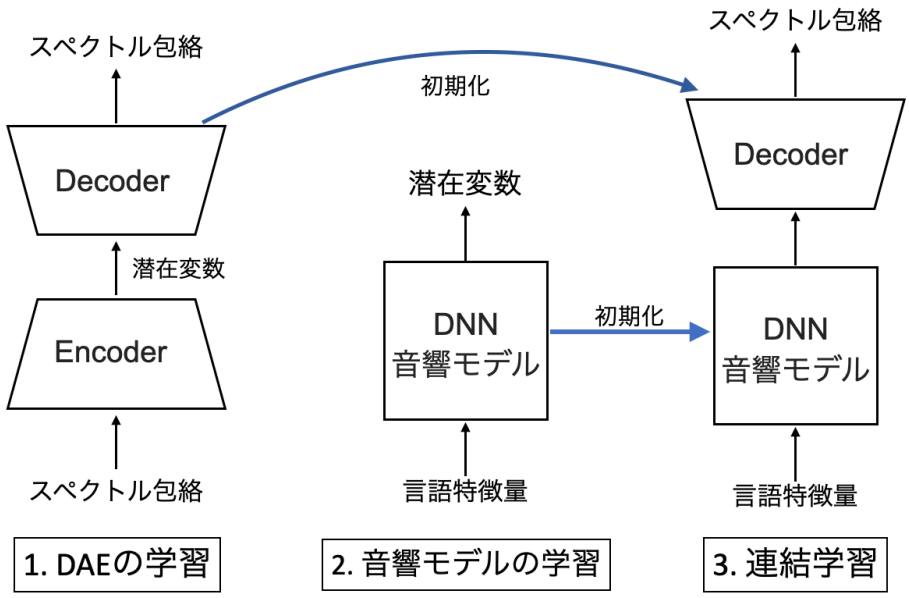


図 2.4: DAE をスペクトルモデリングに用いた TTS システムの学習手順

また，NAE によるスペクトル包絡の次元圧縮を用いた TTS 手法 [7] も提案されている．この手法では NMF によりスペクトル包絡をスペクトルのテンプレートである基底行列と，そのスパースな重みであるアクティベーションの積に分解し，アクティベーションを音響特徴量として用いることでスペクトルの微細な構造を失わずに合成を行うことが可能となっている．

2.3　音声信号の生成

2.3.1　WORLD における波形生成

2.2.3 節で述べたソースフィルタに基づく音声分析合成システムにおける波形生成について説明する．

励起信号にスペクトル包絡を周波数特性として持つフィルタを畳み込むことにより音声信号を合成する．ここで周期成分については基本周波数に対応するパルス列を，非周期性分についてはホワイトノイズを励起信号として用いる．フィルタの位相情報は音声の分析時に失われているので，何らかの位相スペクトルを与える必要がある．WORLD では因果性を満たし，パワーがパルス付近に集中する最小位相応答が用いられており，比較的高品質な音声が合成可能になっている．

2.3.2　ニューラルボコーダによる波形生成

前節で述べた信号処理に基づくボコーダでは分析窓長が一定である，時不変な線形フィルタを用いている，声帯振動のモデル化が単純であるなど様々な近似が用いられている．また，音声分析の際に位相情報が失われる．そのため音響特徴量から再合成した音声には自然性の低下が起こる．

この問題を解決するため WaveNet [3] をはじめとした，ニューラルネットワークを用いて直接

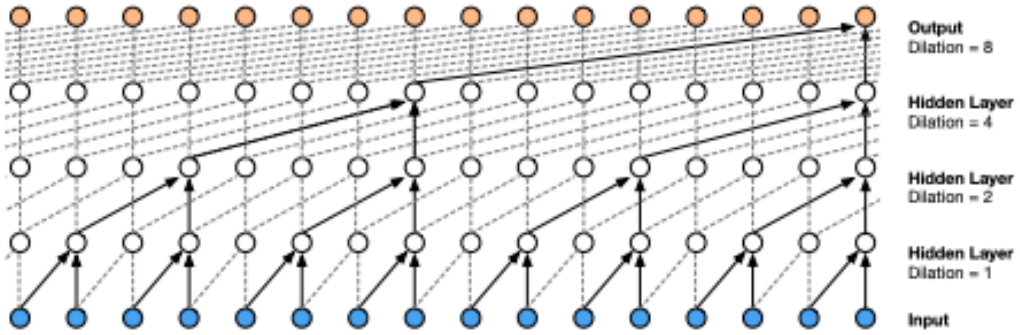


図 2.5: dilated causal convolution の構造．[3] より引用．

音声信号の生成を行うニューラルボコーダが提案されている．WaveNet では音声信号 $x_1, ..., x_N$ の生起確率を各サンプルの自己回帰確率の積としてモデル化する．

$$p(x_1, ..., x_N) = \prod_{n=1}^{N} p(x_n \mid x_1, ..., x_{n-1}) \tag{2.10}$$

過去の出力を推定に用いるため，図 2.5 に示すような dilated causal convolution と呼ばれるネットワーク構造が用いられている．

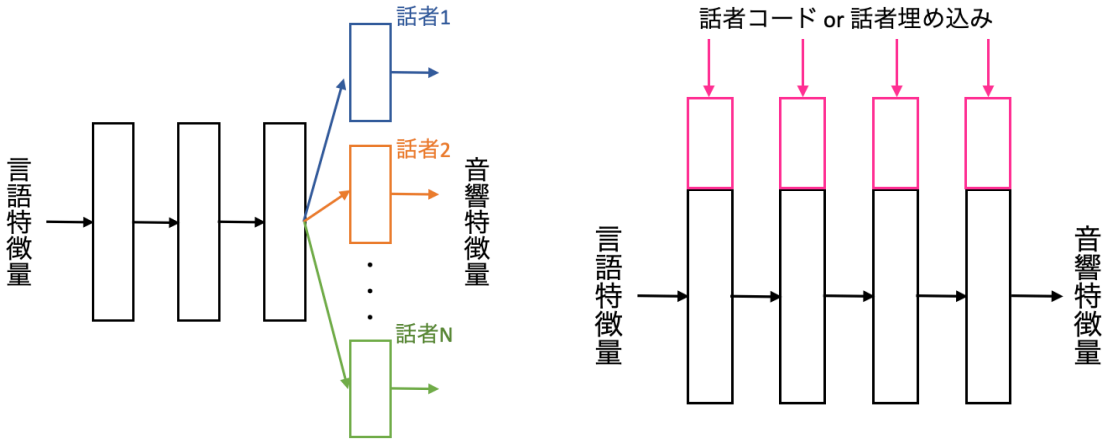
また，補助特徴量を付加して自己回帰確率を条件付けることが可能である． [3] では，言語特徴量を条件付けることで音声信号を直接生成している．WaveNet ボコーダと呼ばれる，補助特徴量にボコーダパラメータを用いたモデル [4] も提案されている．WaveNet により合成音声の品質は大きく向上したが，サンプルを逐次的に生成する必要があるため合成時に膨大な計算時間を要するという欠点がある．

この課題を解決するため様々なモデルが検討されている． [17] では知識蒸留を用いて推論の高速化を実現している．WaveGlow [18] は，Glow [19] と呼ばれるサンプルの対数尤度を直接最大化する生成モデルを利用した手法であり，自己回帰構造を持たないため高速に推論が可能である．また，Parallel WaveGan [20] は，敵対的生成ネットワーク (Generative Adversarial Networks; GAN) に基づく波形生成モデルであり，誤差関数として対数振幅スペクトログラムの二乗誤差と識別器の出力を用いている．

ここまでで述べた DNN に基づく波形生成モデルは人間の音声に関する物理的な仮定を置かず，統計的な性質のみでモデル化を行っている．それに対し，Neural Source Filter [21] は従来のソースフィルタモデルに基づくボコーダに WaveNet のアーキテクチャを導入した手法である．自己回帰や知識蒸留を用いていないため，学習及び推論を高速に行うことが可能である．また，韻律の制御が比較的容易であり，SPSS との相性も良いと考えられる．

2.4　複数話者音声合成

一般的な話者依存の音声合成システムは，音響モデルの学習を話者ごとに別々に行う．それに対し，複数話者音声合成では一つのモデルで複数の話者性を持つ音声を合成する．複数話者モデ



(a) DNN の最終層を話者ごとに学習する手法　　(b) DNN の最終層を話者ごとに学習する手法

図 2.6: DNN 音声合成における話者適応のアプローチ

ルでは学習データ量を実質的に増やすことができるため，十分な量の音声データを話者ごとに用意できない場合に話者依存モデルに比べ高品質な音声を合成できる．

話者適応のアプローチは大きく二つに分けられる．それぞれの手法の概要を図 2.6 に示す．一つは DNN 音響モデルの中間層を話者共通にし，最終層を話者ごとに学習する手法 [22] である．

もう一つは，入力層または隠れ層に話者情報をベクトルとして付加する手法である．[23] では話者コード，すなわち学習データ内の話者の ID を one-hot ベクトルで表現したものを補助特徴量として用いている．また，学習データに含まれない話者に対し話者適応を行うため，話者性を特徴量ベクトルで表現した話者埋め込みを用いる手法も提案されている [24]．

第3章

NAEを用いた音声合成

3.1　はじめに

前章では，統計的パラメトリック音声合成における課題を音響特徴量の観点からいくつか挙げた．本章ではその中でもスペクトル包絡の次元削減に焦点を置く．まず関連研究として NMF による次元削減に基づく TTS 手法について述べる．この手法を受け，NMF を NAE に置き換えた TTS モデルを提案する．

3.2　NMF とその音声合成への応用

NMF [25] とは次式のように非負行列 $Y \in \mathbb{R}^{\geq 0, K \times N}$ を非負行列 $H \in \mathbb{R}^{\geq 0, K \times M}$ と $U \in \mathbb{R}^{\geq 0, M \times N}$ の積で近似する手法である．

$$\boldsymbol{Y} \simeq \boldsymbol{H}\boldsymbol{U} \tag{3.1}$$

ここで，行列 $\boldsymbol{H}$ は基底行列，行列 $\boldsymbol{U}$ はアクティベーションと呼ばれる．基底数 M を N や K よりも十分小さい値に設定すると NMF は低ランク近似と解釈でき，$\boldsymbol{Y}$ を低次元な特徴量 $\boldsymbol{U}$ へと次元削減していると見なすことができる．また，非負制約によりアクティベーションはスパースになることが知られている．

音声信号に用いる場合は一般に振幅スペクトログラムに適用され，スペクトルのテンプレートとそれらの重みの積に分解していると解釈することができる．このとき基底行列は話者などデータセットの性質に強く依存する．また，アクティベーションは言語情報に強く依存し，話者性などにあまり依存しないという傾向がある．この性質を利用した，NMF による話者変換 [26] や帯域拡張 [27] が提案されている．

3.3　NMF を用いた TTS

NMF を声道特徴量の抽出に用いた統計的パラメトリック音声合成が提案されている [7]．概念図を図 3.1 に示す．この手法では，まず訓練用音声のスペクトル包絡に対し NMF を適用し基底行列とアクティベーションに分解する．次に言語特徴量からアクティベーションへの変換を DNN 音響モデルによって学習する．推論時は DNN を用いて与えられた言語特徴量に対応するアクティベーションを推定し，学習時に生成された基底行列を掛けることにより目的のスペクトル包絡を得る．

ここでアクティベーションは基底の生起状態を表したスパースな特徴量であるため，その推定はクラス分類に近い問題と考えられる．そのため，アクティベーションを和が 1 となるよう正規化しカテゴリカル分布とみなしたものと，アクティベーションの L1 ノルムに分けて推定を行っている．この手法では，アクティベーションが基底スペクトルのスパースな重みとなることを生かし，スペクトル包絡の微細な構造を保持したまま次元削減を行っている．評価実験では，音響モデルに入力する特徴量としてメルケプストラムを用いた場合や次元削減を用いずに対数スペクトル包絡を用いた場合と比較して合成音声の自然性の向上が見られたことが報告されている．

3.4　Non-negative Autoencoder

Non-negative autoencoders (NAE) [8] は，NMF の枠組みをニューラルネットワークに拡張したモデルである．

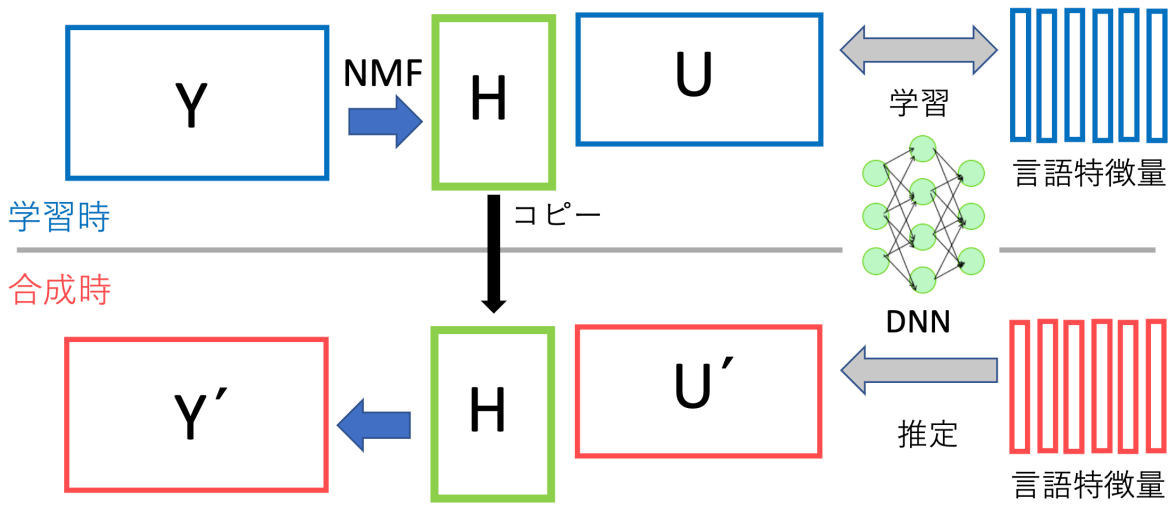


図 3.1: NMF を用いた TTS の概念図

NMF において，2 つの非負行列への分解と再構成を線形 autoencoder として解釈すると式 (3.2) のように書き換えられる．

$$\begin{aligned} Encoder &: U = H^{\ddagger}Y \\ Decoder &: Y = HU \end{aligned} \tag{3.2}$$

$H^{\ddagger}$ は H の擬似逆行列である．すなわち，$H^{\ddagger}$ と H がそれぞれエンコーダ，デコーダとなり，U が潜在変数になっていると解釈できる．

これをニューラルネットワークに適した形に定式化したものが NAE であり，式 (3.3) で表される．

$$\begin{aligned} Encoder &: \boldsymbol{z} = g(W_1\boldsymbol{y}) \\ Decoder &: \hat{\boldsymbol{y}} = g(W_2\boldsymbol{z}) \end{aligned} \tag{3.3}$$

g は出力の非負性を保証する活性化関数であり，softplus などが用いられる． [8] ではカルバック・ライブラー情報量 (Kullback-Leibler Divergence; KLD) を用いて再構成誤差を計算している．

$$D_{KL}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \sum_i (y_i(log(y_i) - log(\hat{y}_i)) - y_i + \hat{y}_i) \tag{3.4}$$

[8] における実験では，スパース正則化を行うことでデコーダの重み行列が非負に近づくことが示されている．NMF を autoencoder の形式に拡張することで，より柔軟に動作する，深層学習に組み込むことができるなどの利点が生まれる．

3.5　NAE を用いたテキスト音声合成

本研究では NAE を用いた統計的音声合成を提案する．この手法では図 3.2 に示すように NAE の再構成の学習と言語特徴量から音響特徴量を推定する DNN の学習を同時に行う．3.3 節で述べ

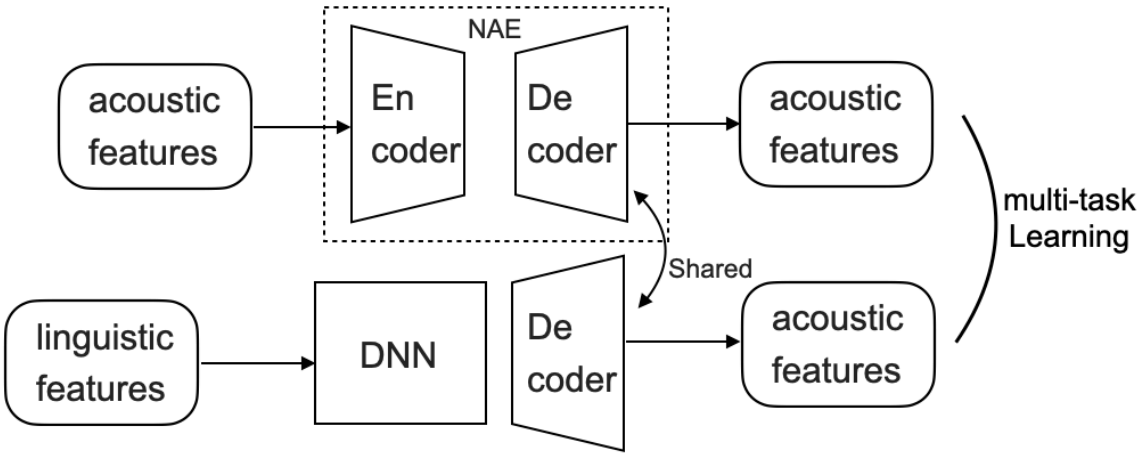


図 3.2: NAE と DNN 音響モデルの同時学習

た NMF に基づく TTS では，NMF と音響モデルの学習は独立に行われていた．それに対し提案手法では NAE と DNN 音響モデルを同時に学習することで，言語情報を考慮した音響特徴量の抽出が NAE によって行われることが期待される．

提案手法についてより詳細な説明を行う．NAE の入力・出力特徴量には各周波数成分の全ての和が 1 になるように正規化されたスペクトル包絡を用いる．DNN 音響モデルは言語特徴量を入力とし，NAE の潜在変数とスペクトル包絡のパワーを推定する．音響モデルにより推定された潜在変数はデコーダに入力され，その出力を元に TTS の誤差が計算される．いずれも KLD を誤差関数とする．全体の誤差関数は式 (3.5) で表される．

$$D_{KL}(\boldsymbol{y}, d(\boldsymbol{z}_{enc})) + D_{KL}(\boldsymbol{y}, d(\boldsymbol{z}_{tts}))) + D_{KL}(p, \hat{p}) \tag{3.5}$$

ここで $\boldsymbol{y}$ は正規化された振幅スペクトル包絡，$\boldsymbol{z}_{enc}$ は NAE に $\boldsymbol{y}$ を入力した時のエンコーダの出力，$\boldsymbol{z}_{tts}$ は DNN 音響モデルにより推定された潜在変数，d はデコーダのネットワーク，パワー項 p は元のスペクトル包絡の L1 ノルムである．

音響特徴量の抽出にボコーダを用いた場合，合成時におけるシステムの全体像は図 3.3 のようになる．DNN 音響モデルは声道特徴量と同時に基本周波数や非周期性指標などのボコーダのパラメーターを推定し，最後に波形生成を行う．

3.6　評価実験

3.6.1　実験条件

音声のデータセットとして ATR 日本語音声データベース [28] を用いた． [28] は A-J セット，合計 503 文の発話で構成されており，そのうち 450 文を訓練用データ，53 文をテスト用データとした．音声サンプルは HTS-demo [29] にある男性の発話データを用いた．サンプリング周波数は 48kHz である．

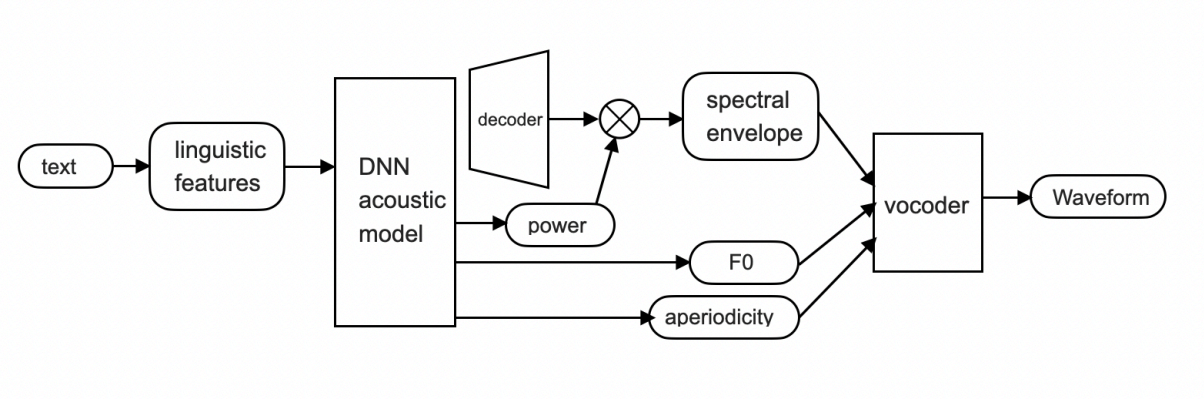


図 3.3: 提案する TTS システムの全体像 (テスト時)

音声の分析・合成には，音声分析変換合成システム WORLD [15](D4C edition [30]) を用い，音響特徴量であるスペクトル包絡は 1025 次元とした．本研究では声道特徴量の推定に焦点を当てているため，基本周波数，非周期性指標については元の音声から抽出したものを用いた．

音響モデルの入力特徴量には HTS-demo にあるフルコンテキストラベルを元にフレームごとに計算された言語特徴量を [0.01,0.99] の値を取るように正規化したものを用いた．言語特徴量の次元は 675 次元とした．

提案手法について，NAE の入力及び出力にはフレームごとに L1 ノルムが 1 になるように正規化された 1025 次元のスペクトル包絡を用いた．隠れ層は 1 層，潜在変数の次元数は 200 次元とした．隠れ層及び出力層の活性化関数は softplus とし，隠れ層は和が 1 になるよう正規化した．DNN 音響モデルには FeedForward 型の DNN を用いた．隠れ層は 6 層・1024 次元，隠れ層の活性化関数は tanh とした．出力は NAE の潜在変数 200 次元にスペクトル包絡のパワー項を加えた 201 次元とし，活性化関数は潜在変数は softmax，パワー項は softplus とした．TTS の学習時は**??**節で述べた手法で NAE と音響モデルを同時学習し，NAE の重みを固定してから再度学習を行った．

DAE を用いた手法について，DAE の入力・出力には [5] と同様に対数スペクトル包絡を 0-1 に正規化したものを用いた．潜在変数の次元数は提案手法と同様 200 次元とした．また，デコーダの重み行列にはエンコーダの重みの転置行列を用い重みを共有した．DNN 音響モデルの出力層は 200 次元，活性化関数が tanh であり，出力層以外は提案手法の条件と同じである．学習は次の手順で行った．まず DAE による再構成のみを学習し，学習済みの DAE の重みを固定した状態で DNN 音響モデルの学習をする．最後にネットワーク全体の fine-tuning を行う．

本実験では主観評価を行うため，合成音声の自然性に関するプリファレンス AB テストを行った．このテストでは被験者が比較する 2 手法による合成音声のうち自然性が高いと感じた方を選択する．各ペアにつき被験者は 25 人とし，各被験者は評価用の音声からランダムに選択された 10 文を評価した．被験者はクラウドソーシングによって募集した．

また，客観評価指標としてメルケプストラム歪み (Mel-Cepstral Distortion; MCD) を用いた．MCD とは合成音声と元の自然音声の声道特徴量の距離を表した指標であり，値が小さいほど品質が高いことを示す．c_m を m 次元目のメルケプストラム係数とすると次式で計算される．

$$MCD[dB] = \frac{10}{\ln 10}\sqrt{2\sum_{k=1}^{M}(c_m - \hat{c}_m)^2} \tag{3.6}$$

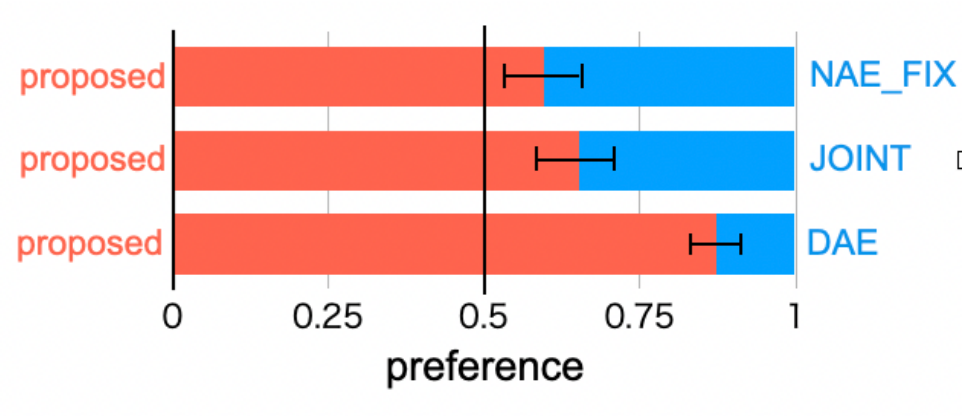


図 3.4: 合成音声の自然性に関する主観評価実験の結果．エラーバーは 95%信頼区間を表す．

表 3.1: メルケプストラム歪みによる合成音声の品質の評価

Method	MCD (dB)
proposed	5.389 ± 0.014
JOINT	5.503 ± 0.014
NAE_FIX	5.477 ± 0.014
DAE	5.856 ± 0.015

3.6.2　実験結果

本実験では，提案手法と 3 手法 (*NAE_FIX*, *JOINT*, *DAE*) の比較を行った．*NAE_FIX* ではまず NAE の再構成の学習のみを行い，デコーダの重みを固定して DNN 音響モデルの学習を行った．*JOINT* では NAE の再構成の学習を行わず，提案手法と同様の構造で TTS の学習のみを行った．*DAE* の条件は 3.6.1 節に述べた通りである．

主観実験の結果を図 3.4 に示す．また，各条件での MCD を表 3.1 に示す．提案手法による合成音声は *NAE_FIX* のものと比較して高い自然性を示した．これは同時学習を行うことで言語情報を考慮した，より TTS の学習に適した特徴量が得られたことを示している．次に *JOINT* と比較しても提案手法の優位性が示された．これは NAE が再構成される制約を加えたことで効率的な次元削減が行われたためであると考えられる． また，*DAE* と比較して提案手法による合成音声の自然性が高いという結果が得られた．これは NAE の非負制約により，スペクトル包絡の微細構造が失われにくい次元削減が行われたためであると推測される．MCD による客観評価についても主観評価結果と整合する結果が得られた．

第4章

NAEに基づく複数話者モデリング

4.1　NAE に基づく話者適応

[22] のように音響モデルには話者共通の DNN を用い話者ごとに NAE を分ける．DNN 音響モデルは話者非依存の音響特徴量を推定し，話者ごとに学習した NAE によって話者適応を行う．このとき NAE の非負制約により話者情報と話者非依存な潜在表現の分離が効率的に行われることが期待される．概念図を図 4.1 に示す．DNN 音響モデルを話者非依存な形で学習することでデータの拡張ができ，単一話者のみで音響モデルを構築した場合と比較して精度が向上することが期待される．

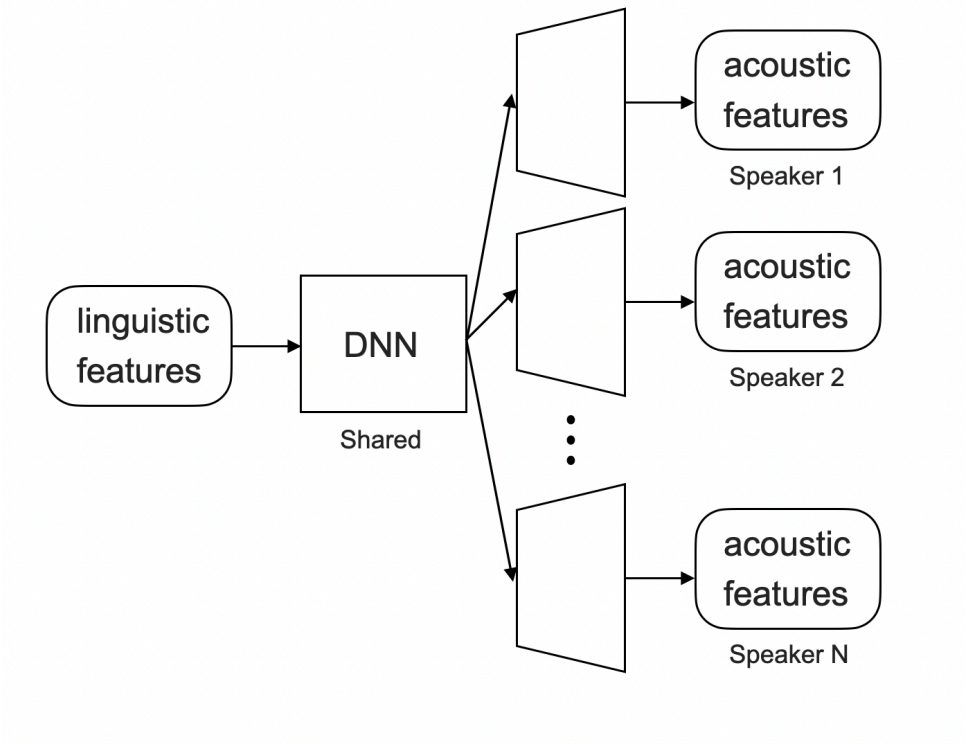


図 4.1: 提案手法における複数話者モデリング

4.2　評価実験

4.2.1　実験条件

音声のデータセットには VCTK コーパス [31] を用い，そのうち 100 話者 (男性 42 名，女性 58 名) の音声を用いた．各話者につき 300 発話を訓練用，20 発話をテスト用のデータとした． サンプリング周波数は 48kHz である．

音声の分析・合成は 3.6.1 節で述べた実験と同一条件で行った．言語特徴量は 420 次元とした．各ネットワークの構造は DNN 音響モデルの入力次元を除いて単一話者実験のものと同一である．NAE は 4.1 節で述べたように話者ごとに異なる NAE を用いて学習を行った．NAE の学習率は DNN 音響モデルの学習率の 10 倍に設定した．

構築した複数話者モデルの評価を行うため，単一話者モデルとの比較を主観評価実験によって行った．評価実験には学習に用いた 100 話者のうち p226 (男性)，p227 (男性)，p228 (女性)，p229

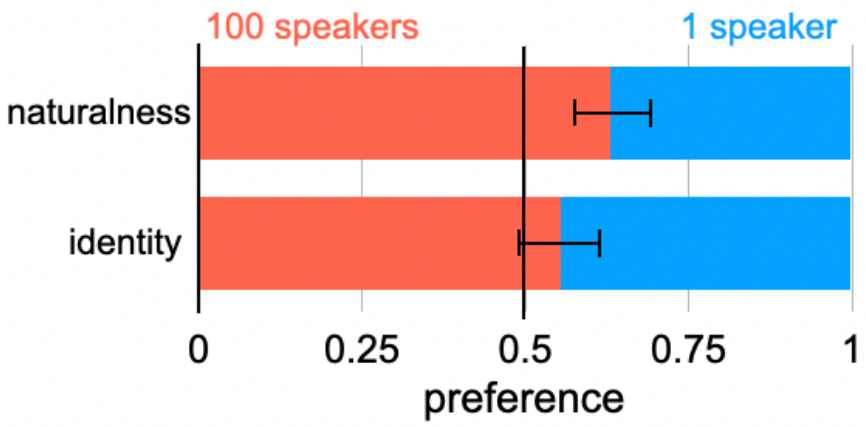


図 4.2: 合成音声の自然性及び話者性に関する主観評価実験の結果．エラーバーは 95%信頼区間を表す．

表 4.1: メルケプストラム歪みによる合成音声の品質の評価

Method	MCD (dB)
Speaker dependent	6.342 ± 0.042
Multi-speaker	6.206 ± 0.037

(女性) の 4 話者の音声を用い，単一話者モデルは各話者ごとに独立に学習した．本実験では 3.6.1 節で述べた自然性に関するプリファレンス AB テストに加え，話者性を評価するための ABX テストを行った．このテストでは被験者が比較する 2 手法の合成音声のうち参照音声に近いと感じた方を選択する．

4.2.2　実験結果

主観実験の結果を図 4.2 に示す．MCD による客観評価の結果を表 4.1 に示す．自然性については複数話者モデルの優位性が示された．話者性についてはわずかに複数話者モデルが高い評価を示していたが，大きな差は見られなかった．客観評価においても複数話者モデルの優位性を示していた．

実験結果から提案手法の複数話者モデルにおいて，DNN 音響モデルによる話者非依存の潜在表現の推定と NAE による話者適応が適切に行われていることが確認された．また，DNN 音響モデルを共有することによるデータ拡張の効果が示された．

4.3　話者コードを用いた場合の実験

本実験では複数話者モデリングの標準的な手法として，話者共通の NAE を用い話者コードのみで話者適応を行う手法との比較を行った．また，デコーダを話者ごとに学習することと話者コードの入力を併用したモデルについても検討を行った．

表 4.2: メルケプストラム歪みによる合成音声の品質の評価

Method	MCD (dB)
SCODE	6.20 ± 0.04
BRANCH	6.13 ± 0.04
HIBRID	6.15 ± 0.04

4.3.1　実験条件

本実験では，3 通りの条件で話者適応を行い (*BRANCH*，*SCODE*，*HYBRID*)，合成音声の品質を比較した．

BRANCH では，4.2 節と同様に NAE を話者ごとに学習することにより話者適応を行った．*SCODE* では，話者コードのみで適応を行った．すなわち入力層に one-hot ベクトルの話者コードを付加し，NAE は話者共通とした．*HIBRID* では，入力層に話者コードを付加し，NAE を話者ごとに分けた．

データセットには VCTK コーパスのうち 10 話者 (男性話者 3 名，女性話者 7 名) を用い，各話者につき 300 発話を訓練用，20 発話をテスト用のデータとした．ネットワークの構成や音声分析合成については 4.2 節で述べた実験と同一の条件で行った．

4.3.2　実験結果

メルケプストラム歪みによる客観評価結果を表 4.2 に示す．*SCODE* と比較してデコーダを分けて話者適応を行った手法がわずかに高い品質を示した．また，*BRANCH* と *HIBRID* では有意差は確認できなかった．

4.4　基本周波数及び非周期性指標についても推定を行う場合の実験

4.1 節で述べた複数話者 TTS システムでは, 基本周波数や非周期性指標などの声道特徴量以外の音響特徴量については話者適応することができない．そこで前節の実験における *HIBRID*，すなわち入力特徴量に話者コードを加えたモデルを用いて，基本周波数，非周期性指標，有声/無声ラベルの推定も同時に行う実験を行った．

4.4.1　実験条件

データセット及び音声分析合成は前節と同一の条件で行った．音響モデルの入力特徴量は話者コードを用いない場合は言語特徴量 420 次元，話者コードを付加する場合は 430 次元とした．出力特徴量は NAE の潜在変数 200 次元とパワー項 1 次元に対数 F_0，非周期性指標 5 次元とそれらの動的特徴量 12 次元，有声/無声フラグを加えた 220 次元とした．対数 F_0 及び非周期性指標について，出力層の活性化関数を線形，誤差関数を最小二乗誤差とした．有声/無声フラグについて，出力層の活性化関数をシグモイド関数，誤差関数を binary cross entropy とした．対数 F_0，非周期性指標については MLPG アルゴリズム [32] により静的特徴量を生成した．その他のネットワークの設定は 4.2 節の実験と同様に行った．

客観評価指標として MCD に加え，対数基本周波数の二乗平均平方根誤差 $\log F_0$-$RMSE$ [dB]，非周期性指標の二乗平均平方根誤差 $BAPD$ [dB]，有声/無声フラグの誤り率 VUV-$error$ [%] を

表 4.3: F_0，bap，V/UV についても推定を行う実験における客観評価の結果

Method	speaker ID	MCD	$\log F_0$-$RMSE$	$BAPD$	VUV-$error$
multi-speaker	average (10 speakers)	6.20	0.168	3.29	10.76
multi-speaker	p226(male)	6.03	0.157	2.74	11.79
speaker-dependent	p226(male)	6.26	0.162	2.79	11.94
multi-speaker	p228(female)	6.18	0.151	3.35	10.95
speaker-dependent	p228(female)	6.37	0.160	3.56	11.11

用いた．

4.4.2　実験結果

客観評価結果を表 4.3 に示す．男性話者と女性話者いずれの場合においても複数話者モデルによる合成音声は話者依存モデルの合成音声に比べ高い品質を示した．音響特徴量に F_0，非周期性指標，有声/無声フラグを含む場合でも適切に話者適応が行われることが示された．データセットに含まれる全 10 話者の MCD の平均は 6.20 であり，前節の実験における *HIBRID* と比較してわずかに品質が低下した．これは声道特徴量以外の音響特徴量についても同時に最適化を行っているためである．本実験により NAE に基づく TTS システムにおいて F_0 や非周期性指標など，声道特徴量以外の音響特徴量も正常に推定可能であることを確認した．

第5章

Neural Source Filterを用いた合成音声の高品質化

5.1　はじめに

本章では従来のボコーダによる波形生成の過程で起こる合成音声の品質低下を改善するため，音響特徴量から音声信号の推定にニューラルボコーダを用いたモデルについて検討する．ニューラルボコーダとして本研究では，学習及び推論を比較的高速に行うことができ，学習が容易な Neural Source Filter [33, 21, 34, 35] を採用する．

評価実験では，3，4 章で提案した NAE を用いた TTS システムのボコーダとして NSF を用いた場合の合成音声を分析，評価した．

5.2　Neural Source Filter

Neural Source Filter (NSF) は，ソースフィルタモデルとニューラルネットワークによる波形生成を組み合わせたモデルである．モデルの概要を図 5.1 に示す．NSF は励振源をモデル化したソースモジュール，声道による音色付けをモデル化したフィルターモジュール，入力特徴量の前処理を行うコンディションモジュールから構成される．

ソースモジュールでは，声帯振動を模した信号の生成を行う．有声音の場合は入力した基本周波数に対応する励起信号 (正弦波や cyclic-noise [35]) を出力し，無声音の場合はガウシアンノイズを出力する．フィルターモジュールでは，ソースモジュールから出力された励起信号に対し dilated convolution layer によるフィルタの畳み込みを行う．入力した音響特徴量は各 filter block に補助特徴量として付加される．

生成された信号から STFT により対数振幅スペクトログラムを計算し，元の音声スペクトログラムとの二乗誤差が小さくなるように学習を行う．このとき複数の解像度で STFT を行いそれぞれ誤差を算出する．振幅スペクトログラムを求める際に位相情報が失われるため，位相スペクトルの学習を陽に行うことができないという欠点がある．

5.3　評価実験

5.3.1　実験条件

DNN 音響モデルの構築及び NSF の学習においてデータセットには VCTK コーパスのうち 10 話者 (男性話者 3 名，女性話者 7 名) を用いた．各話者につき 300 発話を訓練用，20 発話をテスト用のデータとした．サンプリング周波数は 48kHz である．言語特徴量から音響特徴量の推定は 4.4 節の実験におけると同一の条件で行った．

NSF モデルについて，ネットワークの構成及び各パラメータの設定は [35] をもとに行った．誤差を計算する際の STFT のフレーム長，フレームシフト長，窓長については [36] における実験条件を参考にし表 5.1 の設定で行った．また，低周波数帯域を強調するためメル対数振幅スペクトルにおける二乗誤差を誤差関数に加算した．

入力特徴量には harvest [37] により抽出した基本周波数と声道特徴量を用いた．声道特徴量には *mcep*，*melsp*，*logsp* の 3 通りの使用を検討した．*logsp* は WORLD により抽出した 1025 次元のスペクトル包絡の対数をとった特徴量である．*melsp* は *logsp* をメル尺度に従い伸縮し，80 次元にしたものである．*mcep* はスペクトル包絡から抽出した 60 次元のメルケプストラム係数である．各声道特徴量は平均 0，分散 1 となるよう正規化して入力した．それぞれの声道特徴量で学習したモデルを以下 *NSF_SP*，*NSF_MELSP*，*NSF_MCEP* と表す．

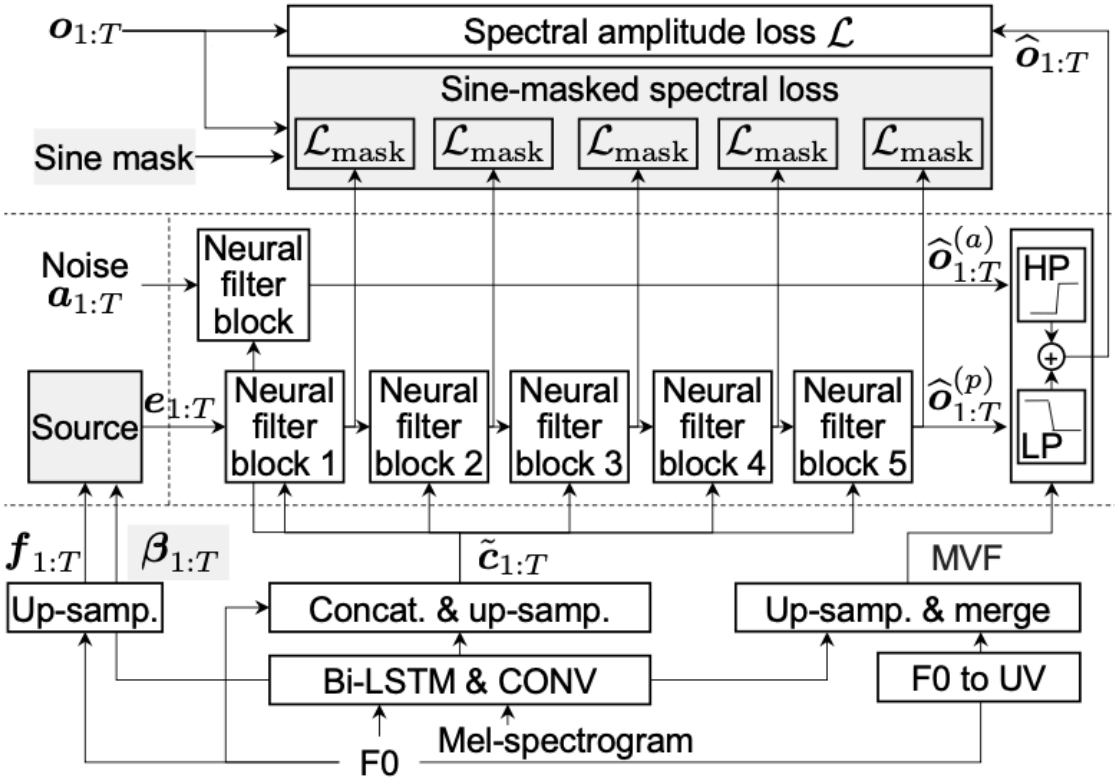


図 5.1: Neural Source Filter モデルの全体像．[35] より引用．

各モデルについて自然音声から抽出した声道特徴量を入力した場合と，音響モデルにより言語特徴量から推定した声道特徴量を入力した場合の 2 通りの音声を生成し，それぞれ評価を行った．また，WORLD ボコーダを用いた場合との比較も行った．音響モデルにより推定した声道特徴量を用いる場合においても基本周波数及び非周期性指標は自然音声から抽出したものを用いた．

合成音声の自然性を評価するため，AB テストによる主観評価及び客観指標による評価を行った．AB テストについては 3.6.1 節で述べた実験と同様の条件で行った．客観指標としてはスペクトル包絡，基本周波数，非周期性指標がどの程度歪んでいるか評価するため，MCD，$\log F_0$-$RMSE$，$BAPD$ を用いた．また，スペクトログラムの歪みを計測するため，対数振幅スペクトログラムの二乗平均平方根誤差 ($logsp$-$RMSE$) を用いた．

$$logsp\text{-}RMSE\ [dB] = \sqrt{\frac{1}{F}\sum_{f=1}^{F}\left(20\log_{10}\frac{|\hat{Y}(f)|}{|Y(f)|}\right)^2} \tag{5.1}$$

ここで Y は音声信号の複素スペクトログラムの 1 フレーム分，F はスペクトルの次元数（本実験の場合は 1025）を表す．すべての指標についてフレーム長 2048，フレームシフト 240 の STFT で求めたスペクトログラムを元に算出し，フレームごとの平均を計測した．

表 5.1: 誤差の算出時における STFT のパラメータ．値は時間信号のサンプル数を示す．

FFT 長	4096	2048	1024	512	256
フレームシフト	400	200	100	50	25
窓長	1600	800	400	200	100
melsp 次元数	640	320	160	80	40

表 5.2: NSF による合成音声の客観評価結果 (自然音声)

Method	*logsp-RMSE*	*MCD*	log F_0-*RMSE*	*BAPD*
WLD	8.11 ± 0.067	3.59 ± 0.025	0.0110 ± 0.0005	0.328 ± 0.0046
NSF_MCEP	8.17 ± 0.035	3.88 ± 0.016	0.0199 ± 0.0010	0.402 ± 0.0055
NSF_MELSP	8.06 ± 0.020	3.74 ± 0.016	0.0128 ± 0.0007	0.401 ± 0.0053
NSF_SP	8.12 ± 0.013	4.17 ± 0.016	0.0144 ± 0.0008	0.385 ± 0.0051

5.3.2　実験結果

自然音声から抽出した音響特徴量から NSF または WORLD を用いて音声を再合成した場合について，客観評価の結果を表 5.2 に，AB テストによる主観評価の結果を図 5.3 に示す．主観評価において NSF の合成音声は WORLD により再合成したものと比較して低い品質を示した．また，*NSF_SP* は，*NSF_MCEP* や *NSF_MELSP* に比べわずかに自然性が低いという結果を示した．これは対数スペクトル包絡の次元数が 1025 次元であり，フィルターモジュールに与える埋め込みベクトルの次元数 (64 次元) と比較して大きいため，声道特徴の付加が精緻に行われなかったことが一因として考えられる．

TTS システムにより推定した音響特徴量を用いた場合について，客観評価の結果を表 5.2 に，主観評価の結果を図 5.3 に示す．*NSF_SP* は自然音声の場合と異なり，*NSF_MCEP* や *NSF_MELSP* に比べ高い自然性を示した．これは NAE を用いた TTS が線形スペクトル領域で 声道特徴量を最適化しているためであると考えられる． また，*WLD* との比較では，*NSF_MCEP* 及び *NSF_MELSP* は有意に自然性が低く，*NSF_SP* は同等の性能を示した．NSF による改善が見られなかった原因として，推定した音響特徴量に対し位相制御や非周期性指標の推定を適切に行うことができなかったことが挙げられる．また，NSF のモデルが自然音声から抽出した声道特徴量に過適合しているために音響モデルにより合成した特徴量に対する推定精度が低下したことも一因であると推測される．

表 5.3: NSF による合成音声の客観評価結果 (TTS)

Method	*logsp-RMSE*	*MCD*	log F_0-*RMSE*	*BAPD*
WLD	11.27 ± 0.051	6.79 ± 0.024	0.0149 ± 0.0009	0.449 ± 0.0055
NSF_MCEP	11.21 ± 0.041	6.80 ± 0.025	0.0232 ± 0.0011	0.507 ± 0.0069
NSF_MELSP	10.90 ± 0.039	6.70 ± 0.024	0.0169 ± 0.0011	0.671 ± 0.0079
NSF_SP	11.15 ± 0.041	6.87 ± 0.024	0.0181 ± 0.0010	0.479 ± 0.0062

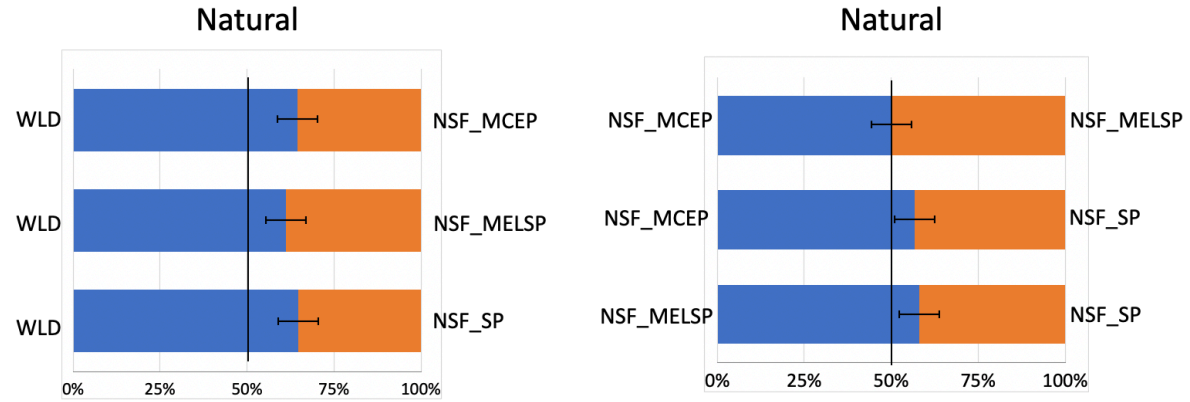


図 5.2: NSF による合成音声の品質に関する主観評価実験の結果 (自然音声から抽出した音響特徴量を使用)．エラーバーは 95%信頼区間を示す．

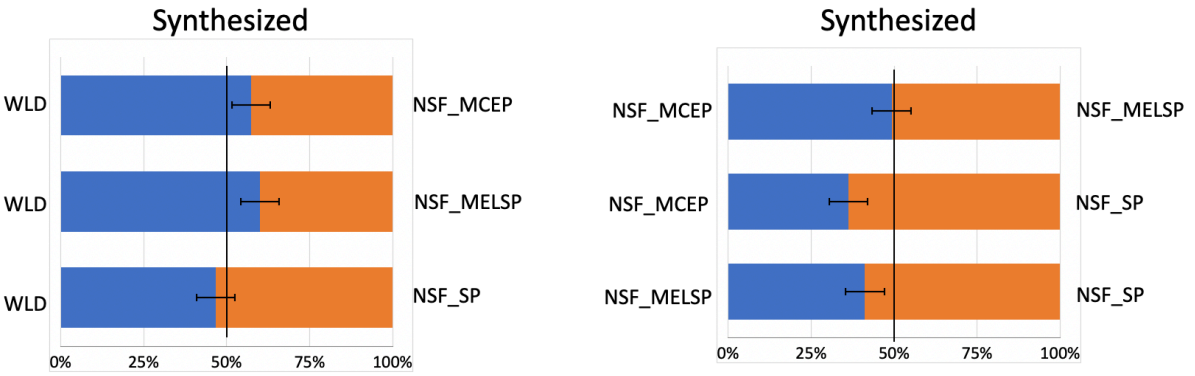


図 5.3: NSF による合成音声の品質に関する主観評価実験の結果 (TTS により推定した音響特徴量を使用)．エラーバーは 95%信頼区間を示す．

5.3.3　まとめ

実験結果を整理すると NSF の合成音声が劣化した主な要因として以下の 3 点が挙げられる．

- 非周期性指標が適切に推定されていない．
- 位相制御が適切に行われていないために合成音声にざらつきを生じている．
- TTS を行った場合に学習時と推論時で入力する声道特徴量にミスマッチが生じている．

以下では，これらの影響について分析し品質の改善を試みる．

5.4　非周期性指標を補助特徴量に加えた場合の実験

5.3 節の評価実験では，品質低下の一因として非周期性指標に当たる情報を付加していないことが挙げられた．そこで，本実験では補助特徴量に非周期性指標を追加して NSF の学習を行い，合成音声の自然性に向上が見られるか検証を行った．

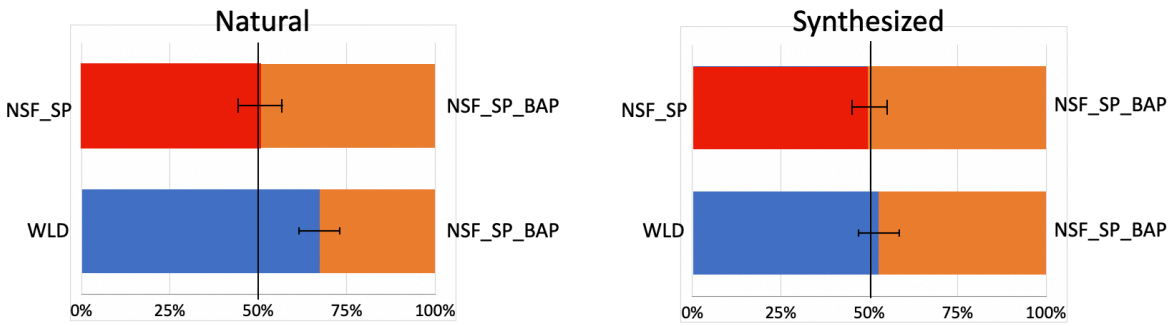


図 5.4: 非周期性指標を加えた実験における合成音声の品質に関する主観評価実験の結果．エラーバーは 95%信頼区間を示す．

表 5.4: 非周期性指標を加えた実験における合成音声の客観評価結果

Method	*logsp-RMSE*	*MCD*	log F_0-*RMSE*	*BAPD*
NSF_SP(自然音声)	8.12 ± 0.013	4.17 ± 0.016	0.0144 ± 0.0008	0.385 ± 0.0051
NSF_SP_BAP(自然音声)	8.13 ± 0.020	4.21 ± 0.015	0.0152 ± 0.0008	0.374 ± 0.0052
WLD(自然音声)	8.11 ± 0.067	3.59 ± 0.025	0.0110 ± 0.0005	0.328 ± 0.0046
NSF_SP(TTS)	11.15 ± 0.041	6.87 ± 0.024	0.0181 ± 0.0010	0.479 ± 0.0062
NSF_SP_BAP(TTS)	11.22 ± 0.041	6.92 ± 0.024	0.0171 ± 0.0009	0.503 ± 0.0067
WLD(TTS)	11.27 ± 0.051	6.79 ± 0.024	0.0149 ± 0.0009	0.449 ± 0.0055

5.4.1　実験条件

NSF_SP は基本周波数と対数スペクトル包絡のみを用いて学習を行ったモデルであり，5.3 節の実験と同一である．*NSF_SP_BAP* は基本周波数，対数スペクトル包絡に加え，WORLD により抽出した 5 次元の非周期性指標を入力特徴量として学習したモデルである．テスト時はスペクトル包絡を自然音声から抽出した場合と音響モデルによって推定した場合の 2 通りでそれぞれ合成を行った．どちらの場合も基本周波数，非周期性指標は自然音声から抽出したものを用いた．二つのモデルの合成音声を AB テストによる主観評価と客観指標評価を用いて比較した．

5.4.2　実験結果

主観評価実験の結果を図 5.4 に，客観評価の結果を表 5.4 に示す．自然音声の分析再合成，TTS いずれの場合においても自然性の向上は見られなかった．また *BAPD* について，自然音声の場合はわずかに改善が見られたが，TTS の場合には精度が低下した．

入力した非周期性指標があまり反映されないという結果を示した．この結果から非周期成分はスペクトル包絡に強く依存しており，主に声道特徴量から推定を行っていることが明らかになった．学習時と推論時において声道特徴量の性質が異なることより，非周期成分の推定精度が低下し，品質低下に寄与していることが推測される．

5.5　WORLD を用いた合成音声の最小位相化

5.3 節の実験では NSF による波形生成において，位相の制御が精緻に行われなかったことが品質低下の一因として挙げられた．[38] では，PSOLA 分析合成においてピッチパルス周辺の振幅の減衰が遅いことがざらつきまたはエコーの要因となることが指摘されている．また同研究では基本波形に対して零位相を施し，ピッチパルス付近にパワーを集中させることでざらつきが改善されることが報告されている．WORLD による分析，再合成を行うことにより同様の効果が得られると考えられる．すなわち，振幅特性は大きく変化せず，周期成分が最小位相化された音声信号が生成されることが予想される．

そこで本実験ではざらつきの影響を検証し改善を行うため，NSF の波形生成の後 WORLD による分析，再合成を行った音声について分析する．

5.5.1　実験条件

データセット及び NSF の学習は 5.3 節の実験と同一の条件で行った．NSF による波形生成の後 WORLD による再合成を行うモデルを *NSF_SP_WR*，*NSF_MELSP_WR*，*NSF_MCEP_WR* と表す．

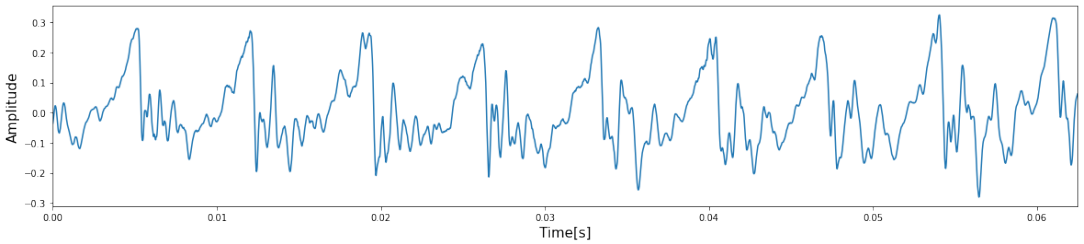
5.5.2　実験結果

自然音声，*NSF_SP* 及び *NSF_SP_WR* の時間波形の例を図 5.5 に示す．これらの波形はすべて同一の発話及び区間を抽出したものである．WORLD による分析，再合成を行うことで周期成分が最小位相化されていることが確認できる．

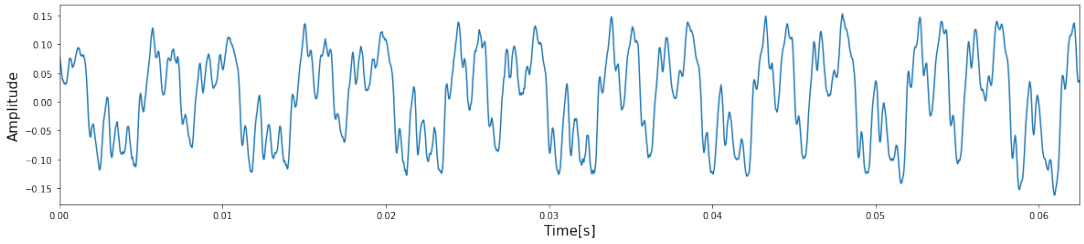
自然音声を分析，再合成した場合の主観評価実験の結果を図 5.6 に示す．NSF の合成音声に対し WORLD の再合成を施すことによる自然性の向上は見られなかった．また，WORLD による合成音声と比較して低い自然性を示した．

TTS を行った場合の主観評価実験の結果を図 5.7 に示す．いずれの声道特徴量を用いた場合でも WORLD の再合成を行うことで品質が改善するという結果が得られた．NSF を用いて合成した特徴量から信号生成を場合において，位相制御が課題となることが明らかになった．

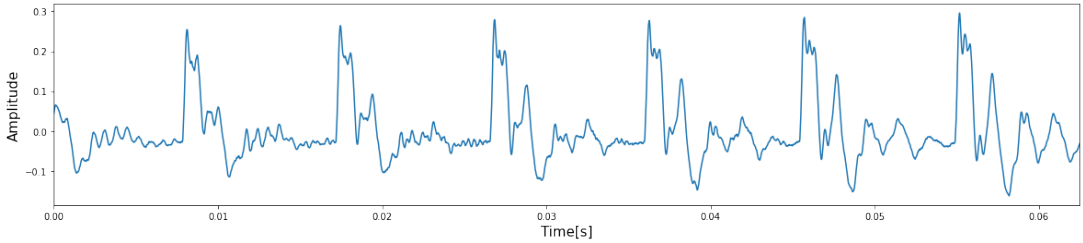
また, *NSF_SP_WR* の品質は *WLD* と比較して有意に高いという結果を示した．*NSF_MELSP_WR* 及び *NSF_MCEP_WR* についても *WLD* と同程度の品質を示した．非周期成分は本来声道特徴量に依存しているが，WORLD のみを用いて波形生成を行う場合は推定された声道特徴量とは独立に非周期性指標を与えているため品質低下が生じたと考えられる．



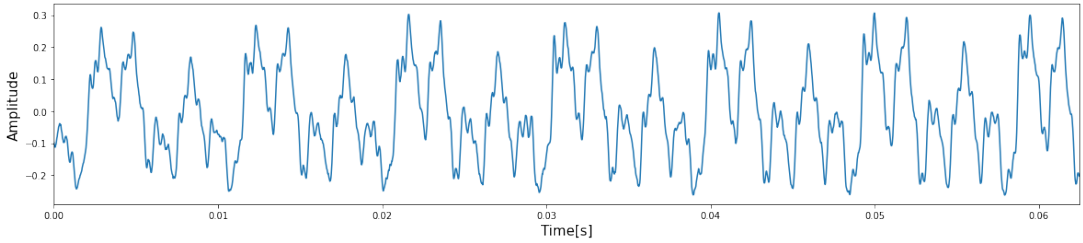
(a) 自然音声の時間波形



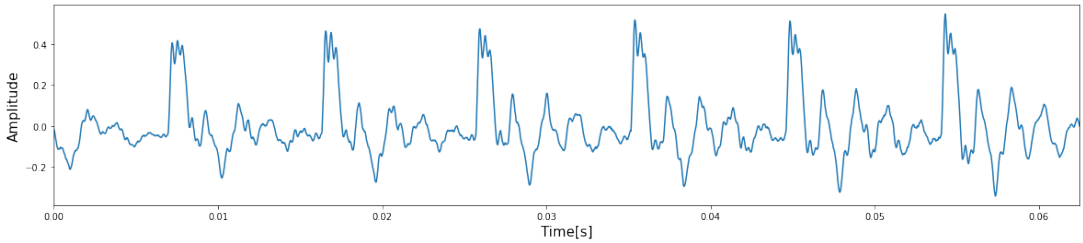
(b) *NSF_SP*(自然音声) の時間波形



(c) *NSF_SP_WR*(自然音声) の時間波形



(d) *NSF_SP*(TTS) の時間波形



(e) *NSF_SP_WR*(TTS) の時間波形

図 5.5: WORLD の再合成による波形の変化．男性話者の評価用音声 (p226_351) より抽出した．

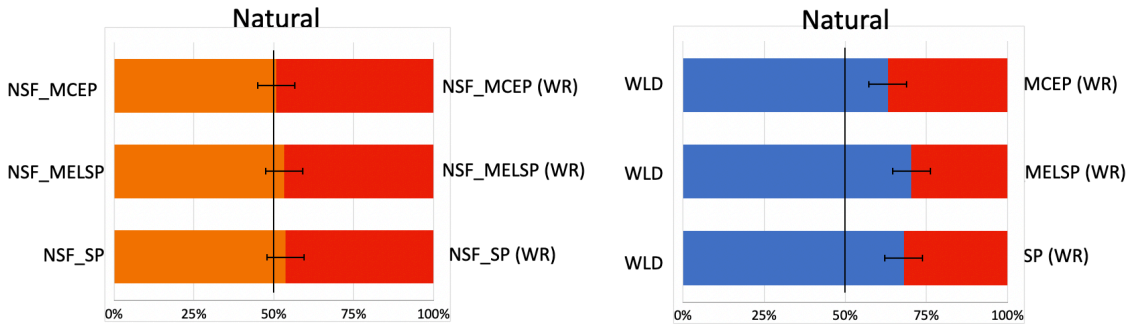


図 5.6: NSF による合成音声の品質に関する主観評価実験の結果 (自然音声より抽出した音響特徴量を使用)．エラーバーは 95%信頼区間を示す．

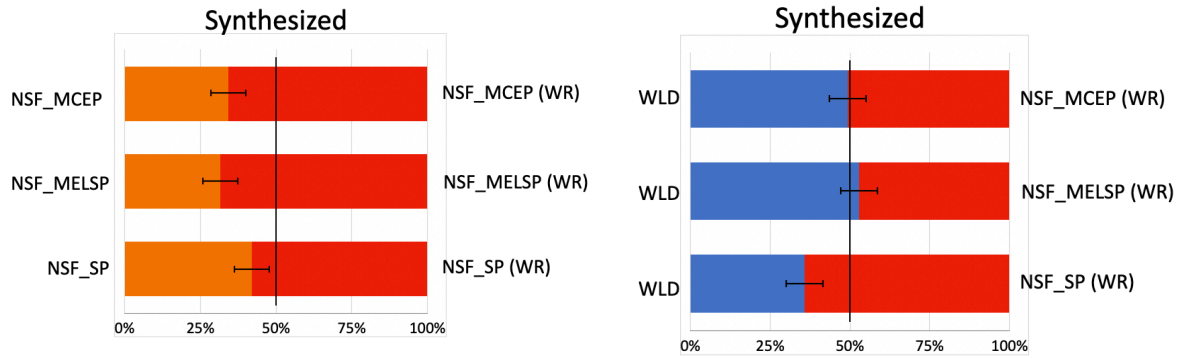


図 5.7: NSF による合成音声に対し WORLD の分析・再合成を行った音声の品質に関する主観評価実験の結果 (TTS により推定した音響特徴量を使用)．いずれもエラーバーは 95%信頼区間を示す．

第6章

NSFにおける音響的ミスマッチの低減

6.1　はじめに

前章で述べた実験ではNSF が自然音声から抽出した音響特徴量に過適合し音響モデルにより推定した音響特徴量に対し十分に汎化できないことが示唆された．本章では，そのような学習時と推論時のミスマッチの改善を試みた実験について述べる．

この問題を改善するため，学習データに用いる声道特徴量の音響的性質を音響モデルで推定したものに近づける必要がある，単純なアプローチとして，自然音声の代わりに音響モデルにより合成した音響特徴量を NSF の学習時に用いるという方法が考えられる．

6.2　合成した音響特徴量を NSF の学習に用いた場合の実験

本実験では，音響モデルにより言語特徴量から推定した声道特徴量を NSF の学習データとすることで，テスト時の合成音声の品質が改善するかを評価実験によって検証した．

6.2.1　実験条件

データセット及び DNN 音響モデルの設定について，5.3 節の実験と同一の条件で行った．NSF の入力特徴量には，訓練用データの言語特徴量から音響モデルにより推定したスペクトル包絡を用いた．基本周波数は自然音声から抽出したものを用いた．テスト時についても音響モデルにより推定した声道特徴量を入力し，評価を行った．NSF の声道特徴量には前章の実験において高い自然性を示した対数スペクトル包絡を用いた．

6.2.2　結果及び考察

自然音声より抽出した音響特徴量を用いて学習した NSF モデル (*TRAIN_NAT*) と本実験で構築したモデル (*TRAIN_SYN*) の合成音声の自然性について AB テストによる比較を行った．結果を図 6.1 に示す．

TRAIN_SYN は *TRAIN_NAT* とほぼ同等の性能を示した．音響モデルにより合成した声道特徴量から音声信号が正常に推定可能であることが確認されたものの自然性の向上は見られなかった．原因として学習に用いた声道特徴量が言語的な誤りを含んでいることが挙げられる． [39] で指摘されているように時間アライメントの不備などにより推定した音響特徴量と自然音声の間には時間構造のミスマッチが存在する．また，言語ラベルに対応する音響特徴量は必ずしも一意に定まらない．このような言語性の不一致を含んだ状態で声道特徴量と音声信号の対応関係を学習したために劣化が生じたと推測される．

6.3　NAE により再構成した音響特徴量を NSF の学習に用いた場合の実験

前節の実験では，学習時に入力した声道特徴量と自然音声の間に言語性や時間構造の不一致が存在したために性能の低下が生じたことが示唆された．問題の解決するためには，言語性と時間構造を保ったまま音響的性質のみを推論時の音響特徴量に近づけた特徴量を NSF の学習に用いる必要があり，自然音声から抽出した音響特徴量からそのような特徴量に変換を行うという方針が有効であると考えられる．

そこで，本研究では自然音声のスペクトル包絡に対して NAE による圧縮及び再構成を行い，生成された声道特徴量を NSF の訓練データとして用いるという枠組みを検討した．類似する手法として，Variational Autoencoder (VAE) を用いた声質変換と WaveNet ボコーダを組み合わせた枠組み [40] が提案されている．この研究では VAE により再構成した音響特徴量を用いて WaveNet を学習することで学習時と推論時のミスマッチが改善されることが報告されている．

6.3.1　用いた手法

学習及び推論の手順を図 6.2 に示す．まず，3 章，4 章で述べた TTS 手法を用いて NAE 及び DNN 音響モデルの学習を行う．次に自然音声のスペクトル包絡に対し NAE による次元削減，再構成を行ったものを入力特徴量として NSF を学習する．このとき基本周波数は自然音声から抽出したものを用いる．推論時は言語特徴量から推定した音響特徴量を学習した NSF に入力し合成音声を得る．

また，NAE のデコーダが自然音声ではなく音響モデルにより推定したスペクトル包絡を再構成するように学習を行う方法についても検討した．すなわち式 (3.5) の代わりに次式の誤差関数を用いて NAE と DNN 音響モデルの同時学習を行った．

$$D_{KL}(d(\boldsymbol{z}_{tts}), d(\boldsymbol{z}_{enc})) + D_{KL}(\boldsymbol{y}, d(\boldsymbol{z}_{tts}))) + D_{KL}(p, \hat{p}) \tag{6.1}$$

より音響モデルの出力に近い特徴量を用いて NSF の学習を行うことで，学習時の推論時の音響的ミスマッチに起因する合成音声の劣化がさらに軽減されることが期待される．また，NAE を用いて音響モデルによる音響特徴量の平滑化を再現することにより，NSF にポストフィルタとしての役割を持たせることが可能である．

6.3.2　実験条件

DNN 音響モデルの学習について，誤差関数に式 (3.5) を用いる場合 (*DECNAT*) と式 (6.1) を用いる場合 (*DECTTS*) の 2 通りを検討した．その他の条件については 4.4 節の実験と同一とした．

NSF について，入力する声道特徴量を除いて前節の実験と同一の条件でモデルの学習を行った．推論時に入力する基本周波数は自然音声から抽出したものを用いた．

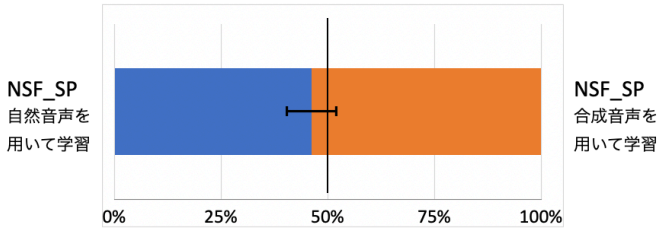
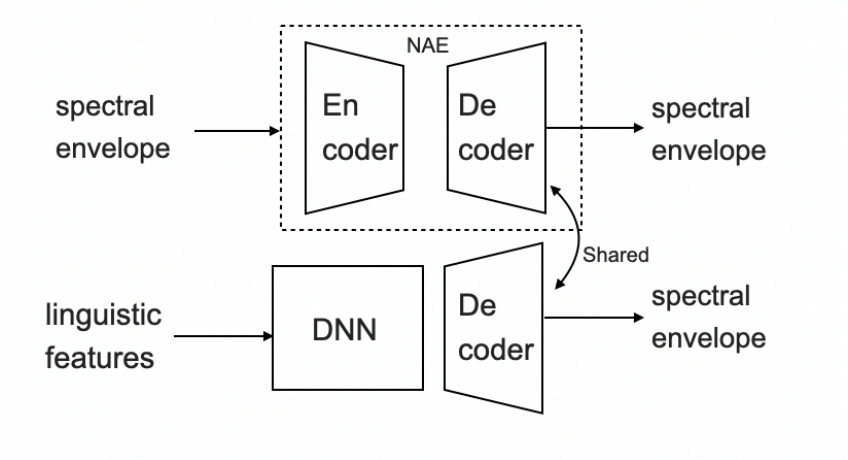
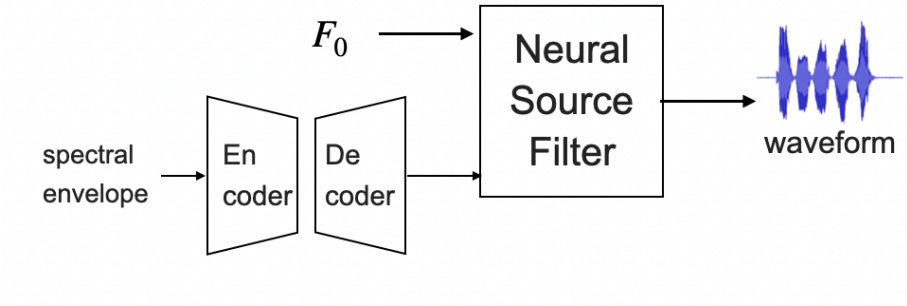


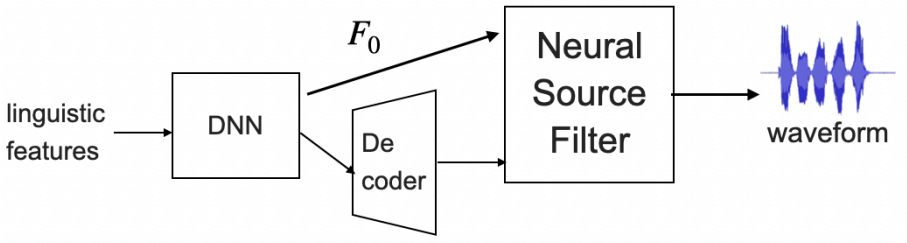
図 6.1: 合成した声道特徴量を用いて学習した NSF の性能の評価．エラーバーは 95%信頼区間を示す．



(a) NAE 及び DNN 音響モデルの学習



(b) NSF の学習



(c) 推論時における音声信号生成

図 6.2: 提案する TTS システムの概要

表 6.1: 自然音声，NAE の出力，音響モデルの出力のメルケプストラム距離

Method	$MCD(y, y_{enc})$	$MCD(y_{enc}, y_{tts})$	$MCD(y, y_{tts})$
DECNAT	1.62	5.79	6.20
DECTTS	5.69	2.62	6.32

6.3.3　実験結果

まず NAE による再構成を行うことで音響モデルが推定した声道特徴量にどの程度近づいたかを確認するため，自然音声から抽出したスペクトル包絡 (y)，それを NAE により再構成したもの (y_{enc})，音響モデルを用いて言語特徴量から推定したもの (y_{tts}) それぞれのメルケプストラム係数の距離 (MCD) を計測した．結果を表 6.1 に示す．*DECNAT* について，$MCD(y_{enc}, y_{tts})$ が $MCD(y, y_{tts})$ に比べ小さくなっており，NAE による次元圧縮を行うことで声道特徴量のミスマッチが改善されることが確認された．*DECTTS* について，y_{enc} と y_{tts}) が大きく改善されておりエンコーダによって音響モデルの出力が十分な精度で再現されていることを示している．また，*DECNAT* と比較して音響モデルの性能がわずかに低下した．

システム全体の合成音声の品質について，主観評価結果を図 6.3 に示す．また，客観評価の結果を表 6.2 に示す．

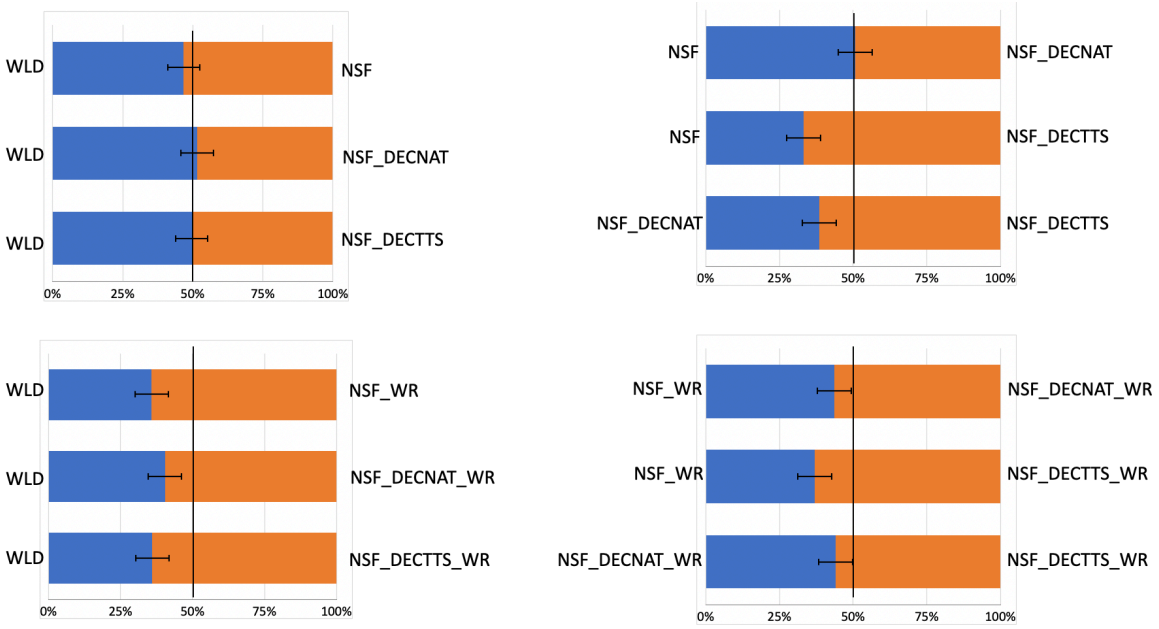


図 6.3: NSF による合成音声の品質に関する主観評価実験の結果．エラーバーは 95%信頼区間を示す．

NSF_DECNAT について，*NSF* と比較してわずかに自然性が向上した．*NSF_DECTTS* は *NSF* や *NSF_DECNAT* に比べ高い自然性を示し，*logsp-RMSE* にも向上が見られた．音響モデルにより平滑化されたスペクトル包絡を学習に用いたことにより，NSF がポストフィルタとして働いたためである．また BAPD も大きく改善しており，学習時と合成時のミスマッチが軽減されたこ

表 6.2: NSF による合成音声の客観評価結果

Method	*logsp-RMSE*	*MCD*	$\log F_0$-*RMSE*	*BAPD*
NSF	11.15 ± 0.041	6.87 ± 0.024	0.0181 ± 0.0010	0.479 ± 0.0062
NSF_DECNAT	11.31 ± 0.043	7.04 ± 0.025	0.0168 ± 0.0009	0.497 ± 0.0067
NSF_DECTTS	10.83 ± 0.039	7.31 ± 0.029	0.0157 ± 0.0008	0.399 ± 0.0057
NSF_WR	11.38 ± 0.039	7.09 ± 0.024	0.0184 ± 0.0010	0.555 ± 0.0062
NSF_DECNAT_WR	11.53 ± 0.042	7.26 ± 0.025	0.0186 ± 0.0010	0.583 ± 0.0071
NSF_DECTTS_WR	11.10 ± 0.037	7.48 ± 0.028	0.0174 ± 0.0009	0.471 ± 0.0060

とにより非周期成分の推定がより精緻に行われたことが示された.

第7章

結論

7.1　まとめ

本研究では，従来の統計的パラメトリック音声合成における合成音声の品質低下の原因としてスペクトログラムの冗長性とボコーダによる劣化の 2 点を挙げ，それらの問題の改善を目指し，NAE を用いた DNN 音声合成の枠組みを提案した．

まず NAE の潜在変数を音響特徴量として用いる TTS モデルについて検討した．単一話者のデータを用いた評価実験では，従来の DAE を用いた手法と比較して高い性能を示し，非負制約によりスペクトル包絡の微細構造が失われにくい次元削減が行われたことが示唆された．また，NAE の次元圧縮と音響モデルを同時に学習することで合成音声の品質が向上することが示された．

また，NAE と音響モデルの同時学習の枠組みを応用した複数話者 TTS についても検討を行った．提案する枠組みでは，NAE を話者ごとに学習することにより話者適応を行う．主観評価実験では話者依存モデルと比較して自然性が向上するという結果を示し，各 NAE のデコーダの重みが話者性を適切に反映するように学習されることが明らかになった．

さらにボコーダによる波形生成の過程で起こる劣化を緩和するため，従来のボコーダの代用として Neural Source Filter (NSF) の適用を検討した．NAE を用いた複数話者 TTS のボコーダとして NSF を用いる実験では，入力に用いる声道特徴量として対数スペクトログラムが適していることが明らかになった．ボコーダとして WORLD を用いた場合との比較実験では，NSF の優位性は確認できなかった．劣化の主な原因として次の 2 点が示唆された．一つは NSF モデルで位相が精緻に推定できていないことである．もう一つは NSF を自然音声から抽出した音響特徴量を用いて学習した場合に自然音声の声道特徴量に過適合してしまい合成した声道特徴量を入力した際の精度が低下することである．

位相制御の問題について，NSF による波形生成の後に WORLD の再合成による最小位相化を行うことでざらつきが改善することが示された．また，NSF を用いることにより音響モデルによる平滑化が改善され，WORLD のみを用いた場合と比較して明瞭性が向上することが明らかになった．

また，学習時と推論時の声道特徴量のミスマッチの問題を緩和するため，音響モデルにより推定した特徴量を用いて NSF の学習を行った．評価実験の結果，言語ラベルの誤りや時間構造の不一致により NSF の学習を精緻に行うことができないという知見を得た．そこで，NAE による圧縮及び再構成を自然音声に適用し音響モデルにより推定した声道特徴量に近づけたものを NSF の訓練データとして用いるという枠組みを検討した．主観評価実験において，自然音声の音響特徴量で学習した NSF を用いた場合と比較して性能が向上することを確認した．

7.2　今後の展望

NAE に基づく複数話者 TTS では，NAE の潜在変数が話者非依存となるように学習が行われた．この枠組みを応用したノンパラレル声質変換が検討可能である．すなわち，ある話者のエンコーダを用いて音声から潜在変数を抽出し，別の話者のデコーダに入力することで話者性のみを変換することが可能である．

また，本研究では NSF による波形生成の後に WORLD による最小位相化を行い，ざらつきを改善した．しかし，このような位相制御は NSF による波形生成の過程でデータドリブンに行うことが合理的であると言える．例えば最小位相応答に近づけるため，ピッチパルス直後に信号のエネルギーが集中する制約を加えるなどの検討が可能である．また，WaveNet や Parallel WaveGan などのより直接的に位相の推定を行った波形生成手法との性能の比較を行う必要がある．

NSF の学習時と推論時の入力特徴量のミスマッチを緩和する枠組みではテキストラベルが与えられた音声を用いて NSF を学習したが，学習済みの NAE を用いることでラベルが付与されていない音声に対しても変換を行い，音響モデルの出力を再現することが可能である．そこで音声データは大量に存在するがテキストラベルがそのうち一部にのみ付与されているという状況で，音響モデルを少量のデータで学習し，大量のデータで学習した NSF を用いて品質を上げるといった応用が考えられる．

謝辞

指導教員である本学大学院工学系研究科の齋藤大輔准教授には，学部生時代から熱心なご指導をいただき，また多くの助言を賜りました。ご多忙の中打ち合わせのため頻繁に時間を割いてくださり、研究を進めるにあたって大きな助けとなりました。論文執筆、発表練習等でも多くの助言をいただきました。また，研究室のサーバ管理を始めとして私たちの研究の環境を整え支えてくださいました。心より感謝申し上げます。本学大学院工学系研究科の峯松信明教授には，第二の指導教員として研究に関わってくださり、特にミーティングや発表練習では適切な助言をいただきました。深く感謝いたします。

齋藤研究室博士 3 年の須田仁志氏、小谷岳氏をはじめとする先輩方には、大変お世話になりました。特に学部生時代に論文の書き方や研究を進める上で生じた疑問への回答など熱心にご指導してくださりました。また、研究室の同期や後輩からは刺激を貰い、多くの学びを得ました。高橋登技術専門員，事務補佐員である押田美智子氏および池上恵氏には研究生活を送る上でお世話になりました。ありがとうございました。

最後にここまで私を育ててくださった家族と友人に心から感謝いたします。

2022 年 1 月 27 日
五来 丈瑠

参考文献

[1] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., "Tacotron: Towards end-to-end speech synthesis", *INTERSPEECH*, pp. 4006–4010, 2017.

[2] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions", *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4779–4783, 2018.

[3] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio", *arXiv preprint arXiv:1609.03499*, 2016.

[4] Akira Tamamori, Tomoki Hayashi, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda, "Speaker-dependent wavenet vocoder", *INTERSPEECH*, pp. 1118–1122, 2017.

[5] Shinji Takaki, SangJin Kim, Junichi Yamagishi, and JongJin Kim, "Multiple feed-forward deep neural networks for statistical parametric speech synthesis", *INTERSPEECH*, pp. 2242–2246, 2015.

[6] Takashi Masuko, Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai, "Speech synthesis using HMMs with dynamic features", *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 389–392, 1996.

[7] Shunsuke Goto, Daisuke Saito, and Nobuaki Minematsu, "DNN-based Statistical Parametric Speech Synthesis Incorporating Non-negative Matrix Factorization", *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 148–153, 2019.

[8] Paris Smaragdis and Shrikant Venkataramani, "A neural network alternative to non-negative audio models", *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 86–90, 2017.

[9] A.J. Hunt and A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database", *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, Vol. 1, pp. 373–376, 1996.

[10] 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.

[11] Keiichi Tokuda, Heiga Zen, and Alan W Black, "An HMM-based speech synthesis system applied to English", *IEEE Workshop on Speech Synthesis*, pp. 227–230, 2002.

[12] Heiga Zen, Andrew Senior, and Mike Schuster, "Statistical parametric speech synthesis using deep neural networks", *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7962–7966, 2013.

[13] 森勢将雅, 音声分析合成, コロナ社, 2018.

[14] Hideki Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds", *Acoustical science and technology*, Vol. 27, No. 6, pp. 349–353, 2006.

[15] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications", *IEICE TRANSACTIONS on Information and Systems*, Vol. 99, No. 7, pp. 1877–1884, 2016.

[16] Keiichi Tokuda, Takao Kobayashi, Satoshi Imai, and Takeshi Chiba, "Spectral estimation of speech by mel-generalized cepstral analysis", *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, Vol. 76, No. 2, pp. 30–43, 1993.

[17] Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al., "Parallel wavenet: Fast high-fidelity speech synthesis", *International conference on machine learning*, pp. 3918–3926, 2018.

[18] Ryan Prenger, Rafael Valle, and Bryan Catanzaro, "Waveglow: A flow-based generative network for speech synthesis", *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3617–3621, 2019.

[19] Diederik P Kingma and Prafulla Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions", *arXiv preprint arXiv:1807.03039*, 2018.

[20] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram", *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6199–6203, 2020.

[21] Xin Wang, Shinji Takaki, and Junichi Yamagishi, "Neural source-filter waveform models for statistical parametric speech synthesis", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 402–415, 2019.

[22] Yuchen Fan, Yao Qian, Frank K Soong, and Lei He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis", *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4475–4479, 2015.

[23] Nobukatsu Hojo, Yusuke Ijima, and Hideyuki Mizuno, "DNN-based speech synthesis using speaker codes", *IEICE TRANSACTIONS on Information and Systems*, Vol. 101, No. 2, pp. 462–472, 2018.

[24] Ye Jia, Yu Zhang, Ron J Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis", *Advances in neural information processing systems*, pp. 4480–4490, 2018.

[25] Daniel D Lee and H Sebastian Seung, "Algorithms for non-negative matrix factorization", 2001.

[26] Zhizheng Wu, Tuomas Virtanen, Tomi Kinnunen, Eng Siong Chng, and Haizhou Li, "Exemplar-based voice conversion using non-negative spectrogram deconvolution.", *8th ISCA Speech Synthesis Workshop*, pp. 201–206, 2013.

[27] Dhananjay Bansal, Bhiksha Raj, and Paris Smaragdis, "Bandwidth expansion of narrow-band speech using non-negative matrix factorization.", *INTERSPEECH*, pp. 1505–1508, 2005.

[28] Akira Kurematsu, Kazuya Takeda, Yoshinori Sagisaka, Shigeru Katagiri, Hisao Kuwabara, and Kiyohiro Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis", *Speech communication*, Vol. 9, No. 4, pp. 357–363, 1990.

[29] `http://hts.sp.nitech.ac.jp/`.

[30] Masanori Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis", *Speech Communication*, Vol. 84, pp. 57–65, 2016.

[31] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al., "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit", 2016.

[32] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis", *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, Vol. 3, pp. 1315–1318, 2000.

[33] Xin Wang, Shinji Takaki, and Junichi Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis", *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5916–5920, 2019.

[34] Xin Wang and Junichi Yamagishi, "Neural harmonic-plus-noise waveform model with trainable maximum voice frequency for text-to-speech synthesis", *arXiv preprint arXiv:1908.10256*, 2019.

[35] Xin Wang and Junichi Yamagishi, "Using cyclic noise as the source signal for neural source-filter-based speech waveform model", *arXiv preprint arXiv:2004.02191*, 2020.

[36] Po-chun Hsu and Hung-yi Lee, "WG-WaveNet: Real-time high-fidelity speech synthesis without GPU", *INTERSPEECH*, pp. 210–214, 2020.

[37] Masanori Morise, et al., "Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals.", *INTERSPEECH*, pp. 2321–2325, 2017.

[38] Nobuaki Minematsu and Seiichi Nakagawa, "Quality improvement of PSOLA analysis-synthesis using partial zero-phase conversion.", *INTERSPEECH*, pp. 779–782, 2000.

[39] Yi-Chiao Wu, Patrick Lumban Tobing, Kazuki Yasuhara, Noriyuki Matsunaga, Yamato Ohtani, and Tomoki Toda, "A cyclical post-filtering approach to mismatch refinement of neural vocoder for text-to-speech systems", *INTERSPEECH*, 2020.

[40] Wen-Chin Huang, Yi-Chiao Wu, Hsin-Te Hwang, Patrick Lumban Tobing, Tomoki Hayashi, Kazuhiro Kobayashi, Tomoki Toda, Yu Tsao, and Hsin-Min Wang, "Refined wavenet vocoder for variational autoencoder based voice conversion", *27th European Signal Processing Conference*, pp. 1–5, 2019.

発表文献

国内研究会・全国大会

[1] 五来丈瑠, 須田仁志, 齋藤大輔, 峯松信明, “テキスト音声合成における劣化音声を活用したデータ拡充に関する検討”, 情報処理学会音声言語情報処理研究会, pp. 1-6, 2020. (優秀発表賞)

[2] 五来丈瑠, 齋藤大輔, 峯松信明, “統計的音声合成のための非負値自己符号化器を用いた音響モデリングの検討”, 日本音響学会秋季研究発表会, 1-3P-13, 2021.

学位論文

“テキスト音声合成における劣化音声を活用したデータ拡充に関する検討”, 東京大学工学部電気電子工学科卒業論文, 2020.