Vol.2018-MUS-119 No.54
Vol.2018-SLP-122 No.54
2018/6/17

von Mises分布DNNに基づく振幅スペクトログラムからの位相復元

高道 慎之介[1,a)]　齋藤 佑樹[1]　高宗 典玄[1]　北村 大地[2]　猿渡 洋[1]

概要：本稿では，deep neural network (DNN) に基づく振幅スペクトログラムからの位相復元について述べる．音声音響信号処理では振幅スペクトログラムに対する処理がしばしば行われ，その位相スペクトログラムは得られない場合が多い．これに対し Griffin-Lim 法は，無矛盾性に基づき振幅スペクトログラムから位相を復元するが，生成音声に対して不自然なアーティファクトをもたらす．この問題に対処するために，本論文では von Mises 分布 DNN を導入する．この DNN は，位相のような周期変数の確率密度関数である von Mises 分布を条件付き分布として有する深層生成モデルであり，そのモデルパラメータは最尤基準で学習される．我々は，これを振幅スペクトログラムからの位相復元に適用し，更に，推定された位相の群遅延を自然な群遅延に近づけるための DNN 学習基準を導入する．実験結果より，(1) DNN は，位相そのものより群遅延を高精度に推定できること，また，(2) 提案法は，従来の Griffin-Lim 法を超える音質を達成できることを示す．

Shinnosuke Takamichi[1,a)]　Yuki Saito[1]　Norihiro Takamune[1]　Daichi Kitamura[2]
Hiroshi Saruwatari[1]

1. はじめに

音源分離や音声強調などの音響信号処理ではしばしば，短時間フーリエ変換 (short-term Fourier transform: STFT) による振幅スペクトログラムに対する処理が行われる．また，近年の統計的音声合成 [1] は，ボコーダパラメータを生成する枠組みから振幅スペクトルを直接的に生成する枠組み [2], [3], [4] に移行しつつある．これらの技術により最終的な音声を生成する場合，与えられた振幅スペクトログラムに対応する位相スペクトログラムが必要だが，その位相スペクトログラムは得られない場合が多い．Griffin-Lim 法 [5] は，振幅スペクトログラムから位相スペクトログラムを復元する手法であり，STFT と逆 STFT を通して位相を反復的に推定する．この方法は，事前学習を必要としないため高いポータビリティを持つが，最終的な生成音声に対して不自然なアーティファクトをもたらす．この問題に対し本論文では，生成モデルを用いた事前学習に基づく位相復元に取り組む．

[1] 東京大学 大学院情報理工学系システム情報学専攻, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
[2] 香川高等専門学校 電気情報工学科, 355 Chokushi-cho, Takamatsu, Kagawa, 761-8058, Japan.
[a)] shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp

Deep neural network (DNN) は強力な生成モデルの 1 つであり，その確率分布はノンパラメトリック [6], [7], [8], [9] とパラメトリックなものに大別される．本稿では，後者のパラメトリックな確率分布を扱う．その代表例として，等方性多変量ガウス分布 [10] や，動的特徴量制約付きガウス分布 [11], [12]（統計的音声合成では，トラジェクトリ DNN として知られる）がある．振幅スペクトログラムから位相を推定する単純な方法はこれらのモデルを使用することだが，ガウス分布は，2π の周期を持つ周期変数である位相の確率分布のモデル化に適さない．

本稿では，von Mises 分布 DNN に基づく，振幅スペクトログラムからの位相復元法を提案する．von Mises 分布 [13] は，円周上の変数をモデル化する確率密度関数であり，周期変数の確率分布のモデル化に適している．von Mises 分布 DNN は，この von Mises 分布を条件付き確率密度関数として有する深層生成モデルである．von Mises 分布を持つ浅い neural network は Nabney [14] らによって提案されており，本稿では，これを発展させた von Mises 分布 DNN を振幅スペクトログラムからの位相推定に利用する（**図 1** 参照）．DNN 学習のための損失関数（位相ロス）は，von Mises 分布の負の対数尤度を最小化するように定義さ

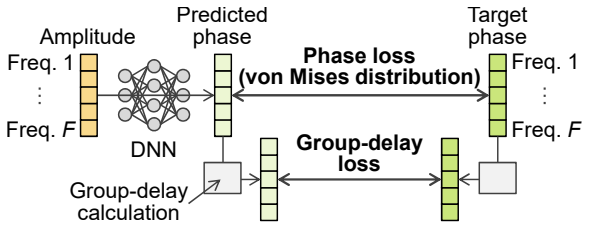

図 1 提案法の構成．図の簡略化のため，マルチフレームや系列単位の推定ではなく，フレーム毎の位相推定を示している．

Fig. 1 Overview of proposed phase reconstruction method. This figure shows frame-by-frame phase prediction rather than multi-frame or sequence-wise prediction for clear illustration.

れる．本稿では更に，振幅スペクトルと強い関係 [15] を持つ群遅延を利用し，群遅延ロスと呼ぶ別の損失関数を提案する．この群遅延ロスは，推定された位相の群遅延をターゲットの群遅延に近づける．群遅延及び群遅延ロスは位相により微分可能であるため，DNN 学習は，標準的な backpropagation アルゴリズムにより行われる．本論文では，提案法の有効性を示すために客観・主観評価を実施する．実験結果より，(1) DNN は，位相そのものよりも群遅延を高精度に推定できること，また，(2) 提案法は，従来の Griffin-Lim 法を超える音質を達成できることを示す．

2. Griffin-Lim 法

本節では，従来の Griffin-Lim 法 [5] による位相復元を概説する．Griffin-Lim 法は，振幅スペクトログラムから位相を復元する，信号処理ベースの反復アルゴリズムである．ここで，$\boldsymbol{x} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t, \cdots, \boldsymbol{x}_T]$ と $\boldsymbol{y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_t, \cdots, \boldsymbol{y}_T]$ をそれぞれ，振幅・位相スペクトログラムとする．$\boldsymbol{x}_t = [x_{t,0}, \cdots, x_{t,f}, \cdots, x_{t,F}]$ と $\boldsymbol{y}_t = [y_{t,0}, \cdots, y_{t,f}, \cdots, y_{t,F}]$ はそれぞれ，フレーム t における振幅及び位相である．f は周波数ビンのインデックスであり，F はナイキスト周波数に対応する．$x_{t,f}$ と $y_{t,f}$ は実数値であり，$y_{t,f}$ は，2π の周期をもつ周期変数である．Griffin-Lim 法ではまず $\boldsymbol{y}$ を乱数で初期化する．その後，(1) 逆 STFT を施し $\boldsymbol{x}$ と $\boldsymbol{y}$ から波形を生成，(2) その波形に対する STFT により $\boldsymbol{x}$ と $\boldsymbol{y}$ を再取得，(3) 再取得した $\boldsymbol{x}$ を元の $\boldsymbol{x}$ に置換しステップ (1) に戻る．これらの逆 STFT と STFT は，収束するまで反復的に行われる．この手法は，与えられた振幅スペクトログラムに対し矛盾のない位相を生成できるが，$\boldsymbol{y}$ の不適切な初期化により，生成波形に対して残響等の不自然なアーティファクトをもたらす．

3. von Mises 分布 DNN による位相復元

本節では，von Mises 分布 DNN を導入し，更に，DNN 位相復元のための二つの損失関数を提案する．

3.1 von Mises 分布

von Mises 分布 $P^{(\mathrm{vM})}(\cdot)$ [13] は，周期変数のための確率密度関数であり，次式で与えられる．

$$P^{(\mathrm{vM})}(y_{t,f}; \mu, \kappa) = \frac{\exp(\kappa \cos(y_{t,f} - \mu))}{2\pi I_0(\kappa)} \tag{1}$$

ここで，μ は，平均（正規分布の平均に対応），κ は集中度パラメータ（正規分布の精度に対応），$I_0(\cdot)$ は 0 次の第 1 種変形 Bessel 関数である．$\boldsymbol{y}_t$ が与えられた時の負の対数尤度は次式で与えられる．

$$-\log P^{(\mathrm{vM})}(\boldsymbol{y}_t; \mu, \kappa) \propto -\sum_{f=0}^{F} \cos(y_{t,f} - \mu) + \mathrm{Const.} \tag{2}$$

ここで，Const. は，μ に依存しない定数項である．この式では，$y_{t,f}$ のみならず μ も 2π の周期を持つ．

3.2 DNN 学習

von Mises 分布を条件付き確率分布として有する DNN を学習する．分布の平均は，各フレーム・周波数毎に $\boldsymbol{x}$ から推定される．ここで DNN を $\boldsymbol{G}(\cdot)$ とすると，推定位相（平均）$\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1, \cdots, \hat{\boldsymbol{y}}_t, \cdots, \hat{\boldsymbol{y}}_T]$ は，$\hat{\boldsymbol{y}} = \boldsymbol{G}(\boldsymbol{x})$ として与えられる．以降では，$\boldsymbol{G}(\cdot)$ のモデルパラメータを推定するための 2 つの損失関数（位相ロス $L_{\mathrm{ph}}(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t)$ と群遅延ロス $L_{\mathrm{gd}}(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t)$）を提案し，更に，それらを組み合わせたマルチタスク学習法を提案する．

3.2.1 位相ロス

位相ロス $L_{\mathrm{ph}}(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t)$ は，式 (2) から導出される．

$$L_{\mathrm{ph}}(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) = \sum_{f=0}^{F} -\cos(y_{t,f} - \hat{y}_{t,f}) \tag{3}$$

$\boldsymbol{G}(\cdot)$ のモデルパラメータは，この関数を最小化するように backpropagation を用いて反復的に推定される．関数を最小とする $\hat{y}_{t,f}$ は周期的であり，$\hat{y}_{t,f} = y_{t,f} \pm 2\pi N$ として得られる．N は任意の整数値である．

3.2.2 群遅延ロス

音声の群遅延は，音声認識 [15]・話者認識 [16] において有効な特徴量である．群遅延は，位相の周波数微分の負値として定義される．一般に，人間の声道を通じた音声生成過程は，$A(\omega)\exp(j\phi(\omega))$ で定義される全極フィルタで効率的にモデル化できる．ここで，$A(\omega)$ と $\phi(\omega)$ はそれぞれ，振幅及び位相関数である．$\omega = \pi f / F$ は角周波数である．そのフィルタが，P 個の複素極 $z_p = r_p \exp(j\omega_p)$ $(p = 1, ..., P)$ を持つとき，$\log A(\omega)$ は次式で与えられる．

$$\log A(\omega) = \log \prod_{p=1}^{P} A_p(\omega) = 2 \sum_{p=1}^{P} \sum_{n=1}^{\infty} \frac{r_p^n}{n} \cos n(\omega - \omega_p) \tag{4}$$

ここで，$A_p(\omega)$ は，p 番目の極による単一極モデルの振幅である [15]．このとき，この群遅延は次式で与えられる．

$$-\frac{d\phi(\omega)}{d\omega} = c\sum_{p=1}^{P}\sum_{n=1}^{\infty} n\cos n(\omega-\omega_p)\int_{-\pi}^{\pi}\log A_p(\omega)\cos(n\omega)d\omega \tag{5}$$

ここで，c は定数を表す．この式は，群遅延と振幅スペクトルに強い関係性があることを示している．そのため，DNN による位相モデリングにおいて，正則化項としての群遅延の効果が期待される．

本項では，群遅延を一次差分で近似する．

$$\Delta y_{t,f} = -(y_{t,f+1} - y_{t,f}) \tag{6}$$

フレーム t・周波数 f における群遅延 $\Delta y_{t,f}$ もまた周期変数であるため，群遅延ロスは，式 (3) と同様の形式で得られる．

$$L_{\mathrm{gd}}\left(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t\right) = \sum_{f=0}^{F} -\cos\left(\Delta y_{t,f} - \Delta\hat{y}_{t,f}\right) \tag{7}$$

この群遅延ロスは，$\Delta\hat{y}_{t,f}$ を $\Delta y_{t,f}$ に一致させる役割を持つ．$\hat{\boldsymbol{y}}_t$ に対する式 (6) は，$\hat{\boldsymbol{y}}_t$ の線形変換で得られるため，**3.2.1 節**と同様に backpropagation を利用できる．

3.2.3 マルチタスク学習

マルチタスク学習法に基づき，位相ロスと群遅延ロスの両方を考慮して DNN を学習する．損失関数 $L\left(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t\right)$ は次式で得られる．

$$L\left(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t\right) = L_{\mathrm{ph}}\left(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t\right) + \alpha L_{\mathrm{gd}}\left(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t\right) \tag{8}$$

ここで，α は副次タスク（本稿では群遅延ロス）の重みを表す．式 (3) と 式 (7) の取りうる値のレンジは等しいため，スケールを正規化する項は不要である．

3.3 考察

von Mises 分布の一般形は，一般化ハート型分布 (generalized cardioid distribution) [17] として次式で与えられる．

$$P^{(\mathrm{GC})}\left(y_{t,f}; \mu, \kappa, \psi\right) = \frac{\left(\cosh\left(\kappa\psi\right)\right)^{1/\psi}\left(1+\tanh\left(\kappa\psi\right)\cos\left(y_{t,f}-\mu\right)\right)^{1/\psi}}{2\pi P_{1/\psi}\left(\cosh\left(\kappa\psi\right)\right)} \tag{9}$$

ここで，$P_{1/\psi}$ は 0 次のルジャンドル陪関数である．この分布は，$\psi \to 0$ のときに von Mises 分布と等価であり，また，$\psi = 1$ もしくは $\psi = -1$ のときにそれぞれ，ハート型分布 (cardioid distribution) と巻き込み Cauchy 分布 (wrapped Cauchy distribution) と等価である．μ を変数としたハート型分布と巻き込み Cauchy 分布の負の対数尤度は，式 (2) と一致するため，これらの分布を条件付き分布としてもつ DNN は，本稿と同様の枠組みで学習される．本研究の更なる展開としては，この一般化ハート型分布を用いた位相モデリングが考えられる．同様に，正弦摂動非対称分布 [18] と混合分布も，考えられる展開である．

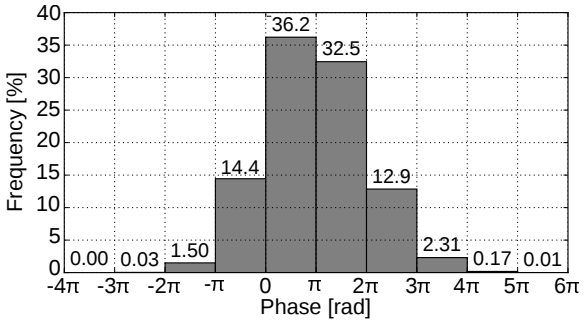

図 2 推定位相のヒストグラム．ターゲットの位相は $[0, 2\pi]$ のレンジに存在するが，推定位相は $[-4\pi, 6\pi]$ のレンジに存在する．この図の頻度は，実験的評価で使用した評価データの全てのフレームと周波数ビンを用いて計算した．

Fig. 2 Histogram of predicted phases. The target phases have a range of $[0, 2\pi]$, but the predicted phases have a range of $[-4\pi, 6\pi]$. All frames and frequency bins of the evaluation data used in evaluation are represented in this figure.

3.2.1 節において述べたように，位相ロスは任意の N において $\hat{y}_{t,f} = y_{t,f} \pm 2\pi N$ となるときに最小化される．故に，von Mises 分布 DNN は，$\hat{y}_{t,f}$ の値の爆発が懸念される．これについて調査するため，**図 2** に推定位相 $\hat{y}_{t,f}$ のヒストグラムを示す．推定位相のレンジは，ターゲットの位相のレンジ $[0, 2\pi]$ よりも広がっているが，値の爆発は見られない．

4. 実験的評価

4.1 実験条件

実験的評価は，単一話者による読み上げ音声コーパス JSUT [19] を用いて実施した．学習データは，サブセット BASIC5000 に含まれる 5,000 文（約 6 時間），評価データは，サブセット ONOMATOPEE300 に含まれる 300 文である．サンプリング周波数は 16 kHz である．フレーム分析における窓長，シフト長，フーリエ変換長はそれぞれ，400 サンプル (25 ms)，80 サンプル (5 ms)，及び 512 サンプルとする．使用した窓関数は，ハミング窓である．DNN への入力特徴量は，当該フレーム及びその前後 2 フレームの対数振幅スペクトルを連結したベクトルである．入力特徴量は，学習時に平均 0・分散 1 に正規化する．DNN のアーキテクチャは，Feed-Forward 型であり，3 層・1024 ユニットの gated linear hidden unit [20] を持つ．予備実験において，隠れ層として ReLU [21] や LeakyReLU [22] 隠れ層を用いた Feed-Forward 型 DNN と比較した結果，本稿の設定による位相推定精度が他の設定を顕著に上回った．

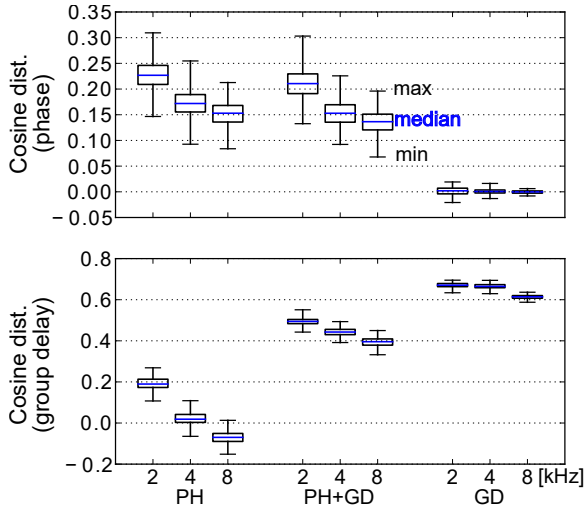

図 3 DNN により推定された位相・群遅延とターゲットの位相・群遅延間のコサイン距離の箱ひげ図．箱は，第 1・第 3 四分位点を表す．

Fig. 3 Box plots of cosine distances between target and predicted phases (upper) and group delays (lower). The box indicates the first and third quartiles.

DNN のモデルパラメータは乱数により初期化する．最適化アルゴリズムには，AdaGrad [23] を利用する．

本稿では，従来の Griffin-Lim 法と次の 3 つの提案法を比較する．

PH：位相ロス (式 (3)) のみ

GD：群遅延ロス (式 (7)) のみ

PH+GD：マルチタスク学習 (式 (8))

Griffin-Lim 法では，位相を乱数で初期化する．位相推定の反復回数は 100 とする．マルチタスク学習における重み α は 0.1 とする．提案法では，低周波数帯域における位相を DNN で推定し，残りの周波数の位相を乱数で与える．位相推定の周波数帯域を，0-2 kHz (96 次元)，0-4 kHz（128 次元)，0-8 kHz（257 次元）の 3 種類とする．また，DNN により位相を推定した後，Griffin-Lim 法により位相を補正する．補正のための反復回数は 100 とする．

4.2 位相と群遅延の推定精度

提案法による位相・群遅延の推定精度を評価する．**図 3** に，推定された位相・群遅延とターゲット（自然音声）の位相・群遅延間のコサイン距離の箱ひげ図を示す．この距離は，全てのフレームと全周波数ビン (0-2, 0-4, もしくは 0-8 kHz) で平均した．

“PH (2 kHz)” の位相推定精度は 0.15 から 0.31 の値をとっており，その分布は対称的である．また，周波数帯域が広がる (“PH (4, 8 kHz)”) と，推定精度は減少することが分かる．これは，高周波数成分の位相がフレーム分析の時間位置によって容易に変化することに鑑みると，妥当な傾向である．一方で，群遅延ロスのみを用いた場合 (“GD”) の群遅延の推定精度は，位相ロスのみを用いた場合（“PH”）の位相の推定精度を大きく上回る．この結果より，Feed-Forward 型 DNN は，位相そのものよりも群遅延を高精度に推定できることが分かる．最後に，位相ロスと群遅延ロスを組み合わせた場合 (“PH+GD”)，その位相推定精度は “GD” を上回り，また，群遅延推定精度は，“PH” を上回ることが分かる．以上より，マルチタスク学習の有効性を確認できる．

表 1 プリファレンステストの結果（従来の Griffin-Lim 法と提案法の比較）．**太字**は，p 値が 0.05 以下である手法を表す．

Table 1 Results of preference tests: conventional Griffin-Lim method vs. proposed methods. **Bold** indicates preferred method that has a p-value smaller than 0.05

Method A	Scores	p-value	Method B
Griffin-Lim	0.497 vs. 0.503	0.871	PH (2 kHz)
Griffin-Lim	0.280 vs. **0.720**	$< 10^{-9}$	PH (4 kHz)
Griffin-Lim	0.277 vs. **0.723**	$< 10^{-9}$	PH (8 kHz)
Griffin-Lim	0.453 vs. **0.547**	0.022	PH+GD (2 kHz)
Griffin-Lim	0.233 vs. **0.767**	$< 10^{-9}$	PH+GD (4 kHz)
Griffin-Lim	0.247 vs. **0.753**	$< 10^{-9}$	PH+GD (8 kHz)
Griffin-Lim	0.447 vs. **0.553**	0.009	GD (2 kHz)
Griffin-Lim	0.463 vs. 0.537	0.073	GD (4 kHz)
Griffin-Lim	0.490 vs. 0.510	0.619	GD (8 kHz)

4.3 Griffin-Lim 法と提案法の比較

提案法の有効性を確認するため，Griffin-Lim 法と提案法による音声品質を比較する．比較のために，我々のクラウドソーシング型評価システムにおけるプリファレンス AB テストを実施した．各評価に対し 30 人が参加し，各評価者に対し 50 円を支払った．評価者には，高音質の音声サンプルを選択させた．各手法の音声サンプルはランダムに提示した．これらの設定は，以降の主観評価でも同様である．

表 1 に主観評価結果を示す．全ての損失関数・周波数帯域の設定において，提案法は従来法を上回ることが分かる．特に，マルチタスク学習（“PH+GD”）は，全ての周波数帯域の設定において有意に従来法を上回る．以上の結果より，音質改善効果における提案法の有効性を確認できる．

4.4 位相補正の効果

4.1 節に述べたとおり，提案法は，DNN により位相を推定した後に Griffin-Lim 法による位相補正を行う．ここでは，その位相補正の効果を検証する．**図 4** に推定された位相・群遅延と補正後の位相・群遅延間のコサイン距離の箱ひげ図を示す．全体の傾向は**図 3** と共通しており，DNN 学習時に位相ロスを用いた場合 (“PH” と “PH+GD”)，その位相は補正後でも比較的保持され，同様に，DNN 学習時に群遅延ロスを用いた場合 (“GD” と “PH+GD”)，その群遅延は補正後でも比較的保持されることが分かる．位相

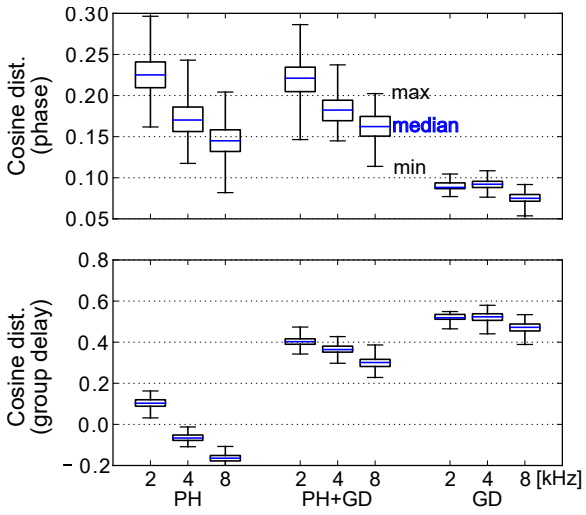

図 4 DNN により推定された位相・群遅延と補正後の位相・群遅延間のコサイン距離の箱ひげ図．箱は，第 1・第 3 四分位点を表す．

Fig. 4 Box plots of cosine distances between predicted and refined phases (upper) and group delays (lower). The box indicates the first and third quartiles.

補正に関する予備実験として，補正ありと補正なし（すなわち，DNN で推定された位相を直接，最終的な音声波形の生成に用いる）の音質を比較した．その結果，補正なしの音質が，補正ありの音質より顕著に悪化することを確認している．

4.5 周波数帯域による影響

同一の損失関数 (“PH”, “PH+GD”, もしくは “GD”) を用いた場合に，推定位相の周波数帯域が音質に与える影響を調査する．**表 2** にプリファレンス AB テストの結果を示す．“PH” と “PH+GD” では，0-4 kHz 帯域の利用が，0-2 kHz 帯域の利用よりも高品質音声を生成できる．また，その音質は，0-8 kHz 帯域を利用した音声と同程度である．この結果より，少なくとも 0-4 kHz 帯域の位相を推定し，それ以上の周波数帯域の位相は乱数とすればよいことが分かる．この傾向は，harmonics plus noise model [24] と同様である．興味深い点として，“GD” では，0-8 kHz 帯域の利用が，0-4 kHz 帯域の利用よりも顕著に音質を悪化させることが挙げられる．今後は，この理由を調査する．

4.6 群遅延ロスの有効性

位相ロスと比較した場合の群遅延ロスの有効性を検証する．ここではまず，位相補正においてこれらを比較する．**図 5** に，“PH (4 kHz)” と “PH+GD (4 kHz)” の spectral covergence [25] の対数値を示す．また，比較のため，ランダム初期位相による結果（“Random”）も示す．この図より，ランダム初期位相よりも提案法は小さい spectral covergence を持ち，また，“PH+GD (4 kHz)” は “PH (4 kHz)” よりも更に小さい値を持つことが分かる．以上より，群遅延の利用により，完全再構成により近い位相を生成できることが分かる．

最後に，“PH” と “PH+GD” の音質を比較する．**表 3** に，周波数帯域の全設定におけるプリファレンス AB テストの結果を示す．0-8kHz 帯域においては “PH” のスコアが高いものの，それ以外の帯域では “PH+GD” のスコアが高いことが分かる．以上より，群遅延ロスの有効性が分かる．

表 2 プリファレンステストの結果（周波数帯域の影響の調査）．**太字**は，p 値が 0.05 以下である手法を表す．

Table 2 Results of preference tests: proposed methods with different frequency bands. **Bold** indicates preferred method that has a p-value smaller than 0.05

Method A	Scores	p-value	Method B
PH (2 kHz)	0.270 vs. **0.730**	$< 10^{-9}$	PH (4 kHz)
PH (4 kHz)	0.507 vs. 0.493	0.744	PH (8 kHz)
PH+GD (2 kHz)	0.223 vs. **0.777**	$< 10^{-9}$	PH+GD (4 kHz)
PH+GD (4 kHz)	0.493 vs. 0.507	0.744	PH+GD (8 kHz)
GD (2 kHz)	0.513 vs. 0.487	0.514	GD (4 kHz)
GD (4 kHz)	**0.567** vs. 0.433	0.001	GD (8 kHz)

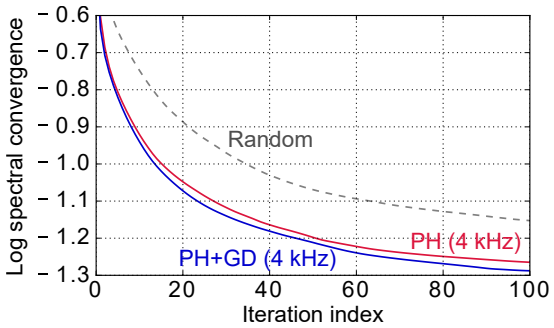

図 5 位相補正による spectral convergence の対数値の変化．この値が $-\infty$ となるとき，STFT と逆 STFT を通した完全再構成が成り立つ．この図は，評価データの一つの結果のみを示しているが，評価データの全てで同様の傾向が得られる．

Fig. 5 Log spectral convergence by phase refinements. When the value is $-\infty$, perfect reconstruction through STFT and inverse STFT is achieved. This is the result of one of the evaluation datasets, but the same tendency was observed in all evaluation datasets.

表 3 プリファレンステストの結果（群遅延ロスの影響の調査）．**太字**は，p 値が 0.05 以下である手法を表す．

Table 3 Results of preference tests: effects of group-delay loss. **Bold** indicates preferred method that has a p-value smaller than 0.05

Method A	Scores	p-value	Method B
PH (2 kHz)	0.487 vs. 0.513	0.514	PH+GD (2 kHz)
PH (4 kHz)	0.486 vs. 0.514	0.500	PH+GD (4 kHz)
PH (8 kHz)	**0.545** vs. 0.455	0.031	PH+GD (8 kHz)

5. まとめ

本稿では，DNN に基づく振幅スペクトログラムからの位相復元法を提案した．周期変数に対応する確率密度関数である von Mises 分布の最尤推定に基づき，DNN 学習の損失関数として位相ロスと群遅延ロスを提案した．実験的評価より，(1) DNN は，位相そのものよりも群遅延を高精度に推定できること，また，(2) 提案法は，従来の Griffin-Lim 法を超える音質の音声を生成できることを明らかにした．今後は，周期変数に対応する他の確率密度関数，他の位相補正法，敵対的ボコーダフリー音声合成 [4] への統合を検討する．

謝辞： 本研究の一部は，セコム科学技術支援財団，JSPS 科研費 18K18100 の助成を受け実施した．

参考文献

[1] H. Zen, K. Tokuda, and A. Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] S. Takaki, H. Kameoka, and J. Yamagishi, “Direct modeling of frequency spectra and waveform generation based on phase recovery for DNN-based speech synthesis,” in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017.

[3] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, “Tacotron: Towards end-to-end speech synthesis,” vol. abs/1609.03499, 2017. [Online]. Available: https://arxiv.org/abs/1703.10135

[4] Y. Saito, S. Takamichi, and H. Saruwatari, “Text-to-speech synthesis using stft spectra based on low-/multi-resolution generative adversarial networks,” in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 5299–5303.

[5] D. W. Griffin and J. S. Lim, “Signal estimation from modified short-time fourier transform,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Proc. NIPS*, pp. 2672–2680, 2014.

[7] 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, pp. 755–767, Jun. 2018.

[8] Y. Li, K. Swersky, and R. Zemel, “Generative moment matching networks,” in *Proc. ICML*, Lille, France, Jul. 2015, pp. 1718–1727.

[9] S. Takamichi, K. Tomoki, and H. Saruwatari, “Sampling-based speech parameter generation using moment-matching network,” in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 3961–3965.

[10] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in *Proc. ICASSP*, Vancouver, Canada, May 2013.

[11] Z. Wu and S. King, “Minimum trajectory error training for deep neural networks, combined with stacked bottleneck features,” in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 309–313.

[12] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “The effect of neural networks in statistical parametric speech synthesis,” in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4455–4459.

[13] K. V. Mardia and P. E. Jupp, *Directional Statistics*. John Wiley & Sons Ltd., 1999.

[14] I. Nabney, C. Bishop, and C. Legleye, “Modelling conditional probability distributions for periodic variables,” in *1995 Fourth International Conference on Artificial Neural Networks*, Calgary, Canada, Jun. 1995, pp. 177–182.

[15] F. Itakura and T. Umezaki, “Distance measure for speech recognition based on the smoothed group delay spectrum,” in *Proc. ICASSP*, Dallas, U.S.A., Apr. 1987, pp. 1257–1260.

[16] R. Padmanabhan, S. H. K. Parthasarathi, and H. A. Murthy, “Robustness of phase based features for speaker recognition,” in *Proc. INTERSPEECH*, Brighton, U. K., Sep. 2009, pp. 2355–2358.

[17] M. C. Jones and A. Pewsey, “A family of symmetric distributions on the circle,” *Journal of the American Statistical Association*, vol. 100, no. 472, pp. 1422–1428, Dec. 2005.

[18] T. Abe and A. Pewsey, “Sine-skewed circular distributions,” *Statistical Papers*, vol. 52, no. 3, pp. 683–707, Aug. 2011.

[19] R. Sonobe, S. Takamichi, and H. Saruwatari, “JSUT corpus: free large-scale japanese speech corpus for end-to-end speech synthesis,” vol. abs/1711.00354, 2017.

[20] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” vol. abs/1612.08083, 2016.

[21] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in *Proc. AISTATS*, Lauderdale, U.S.A., Apr. 2011, pp. 315–323.

[22] L. A. Maas, Y. A. Hannun, and Y. A. Ng, “Rectifier nonlinearities improve neural network acoustic models,” in *Proc. ICML*, vol. 30, no. 1, 2013.

[23] J. Duchi, E. Hazan, and Y. Singer, “Adaptive subgradient methods for online learning and stochastic optimization,” *EURASIP Journal on Applied Signal Processing*, vol. 12, pp. 2121–2159, 2011.

[24] Y. Stylianou, “Applying the harmonics plus noise model in concatenative speech synthesis,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, Jun. 2001.

[25] N. Sturmel and L. Daudet, “Signal reconstruction from STFT magnitude: A state of the art,” in *Proc. of 14th International Conference on Digital Audio Effects DAFx-11*, Paris, France, Sep. 2011, pp. 177–182.