

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, 齋藤 佑樹, 高宗 典弘 (東京大学), 北村 大地 (香川高専), 猿渡 洋 (東京大学)

概要：振幅スペクトログラムからの位相復元

- 近年の音声音響信号処理における位相復元の価値
 - 音源分離や音声強調では、振幅スペクトルを処理対象にすることが多い
 - 統計的音声合成でも、ボコーダパラメータ生成から振幅スペクトル生成へ
- DNN位相復元
 - DNNに基づいて、事前学習可能な位相復元器を作れるか？
 - **位相は周期変数**なので、**ガウス分布DNN (平均二乗誤差最小化による学習) は使えない。**
- 提案法
 - 周期変数の確率密度関数である**von Mises分布を持つDNN**を導入。位相ロスを最小化して学習。
 - 音声の振幅スペクトルと強い関係性を持つ**群遅延**を導入。群遅延ロスも最小化して学習。

Griffin-Lim法^[1]による位相復元が多いが
生成音声に不自然なアーティファクトが発生

Griffin-Lim法 [1] —
与えられた振幅スペクトルから、短時間フーリエ変換(STFT)と逆STFTを繰り返して、位相を推定する方法。無矛盾位相復元とも呼ばれる。

(1) DNNは位相より群遅延を推定しやすいこと、(2) Griffin-Lim法を超える音質を出せることを明らかに

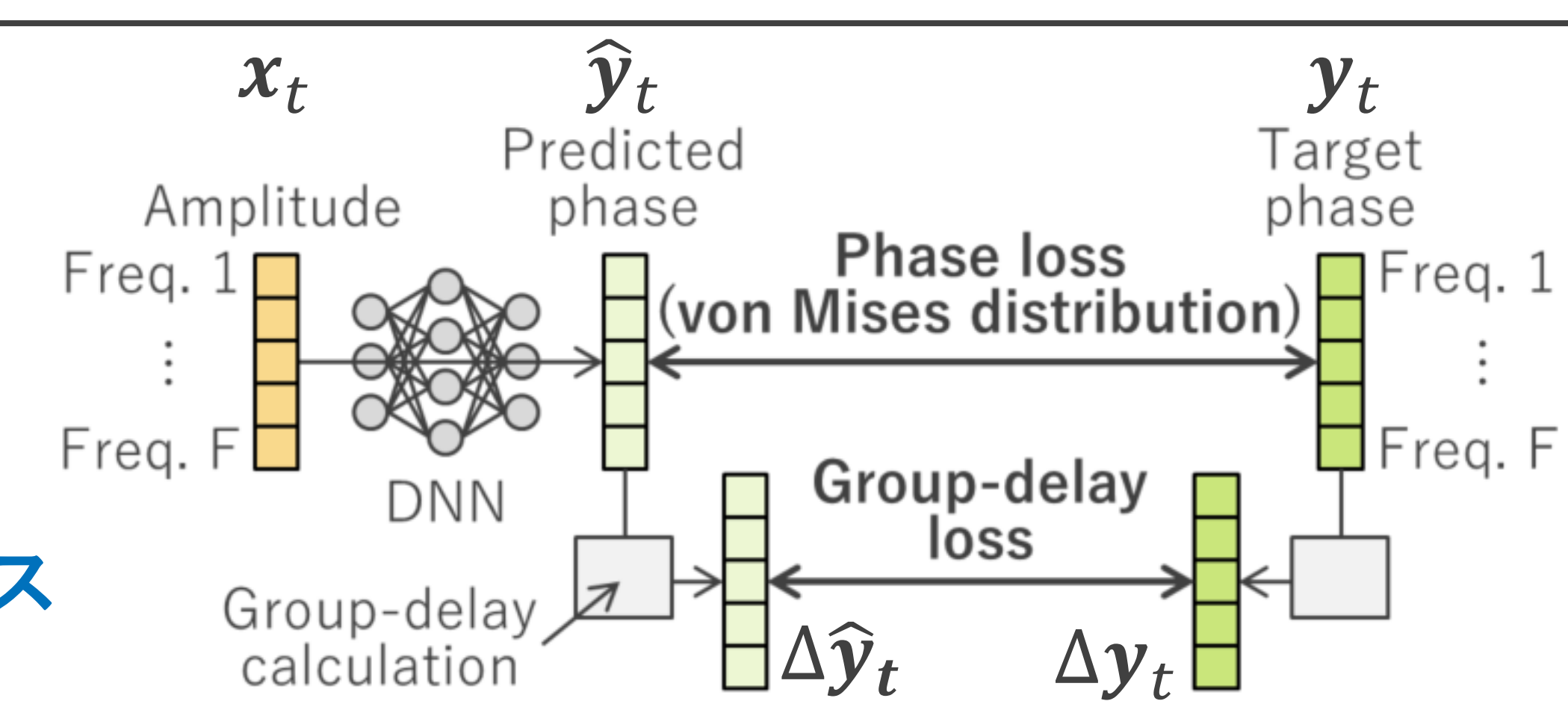
提案法：Von Mises 分布DNNに基づく位相復元

von Mises分布とDNN位相復元

- von Mises分布^[3]
 - $P(y; \mu, \kappa) = e^{\kappa \cos(y - \mu)} / 2\pi I_0(\kappa)$
 μ : 平均, κ : 形状パラメータ
 $I_0(\cdot)$: 0次の第1種変形Bessel関数
- 負の対数尤度 (μ をモデル変数)
 - $-\log P(y; \mu, \kappa) \propto -\cos(y - \mu)$

DNN位相復元

- 振幅 x_t から位相 y_t を推定するDNN
 - t は時間インデックス
- DNN学習のための損失関数
 - von Mises分布から導出される**位相ロス**
 - 群遅延を一致させる**群遅延ロス**



①位相ロス：位相を予測するための損失関数

- von Mises分布を条件付き分布とみなし最尤推定

$$L_{PH}(y_t, \hat{y}_t) = - \sum_f \cos(y_{t,f} - \hat{y}_{t,f})$$

$y_{t,f}$: y_t における周波数 f の位相

②群遅延ロス：群遅延を予測するための損失関数

- 群遅延を近傍周波数との位相差分で表現し最尤推定

$$L_{GD}(y_t, \hat{y}_t) = - \sum_f \cos(\Delta y_{t,f} - \Delta \hat{y}_{t,f})$$

$\Delta y_{t,f} = -(y_{t,f+1} - y_{t,f})$: y_t における周波数 f の群遅延

*ARモデルの振幅と位相は強い関係 [3]

実験的評価：予測の精度・群遅延の効果・音質改善への効果

項目	値／設定
学習データ／評価データ	JSUT音声コーパス ^[4] 5000文／同300文
サンプリング周波数	16kHz
シフト長, FFT長	80 samples (5 ms), 512 samples
DNN入力	当該フレーム & ±2フレームの対数振幅
DNN出力	位相 (0-2kHz, 0-4kHz, 0-8kHzの3パターン)
DNN	Feed-Forward w/ gated linear hidden units
後処理	DNNで推定後に Griffin-Lim法で補正

② 音質への効果

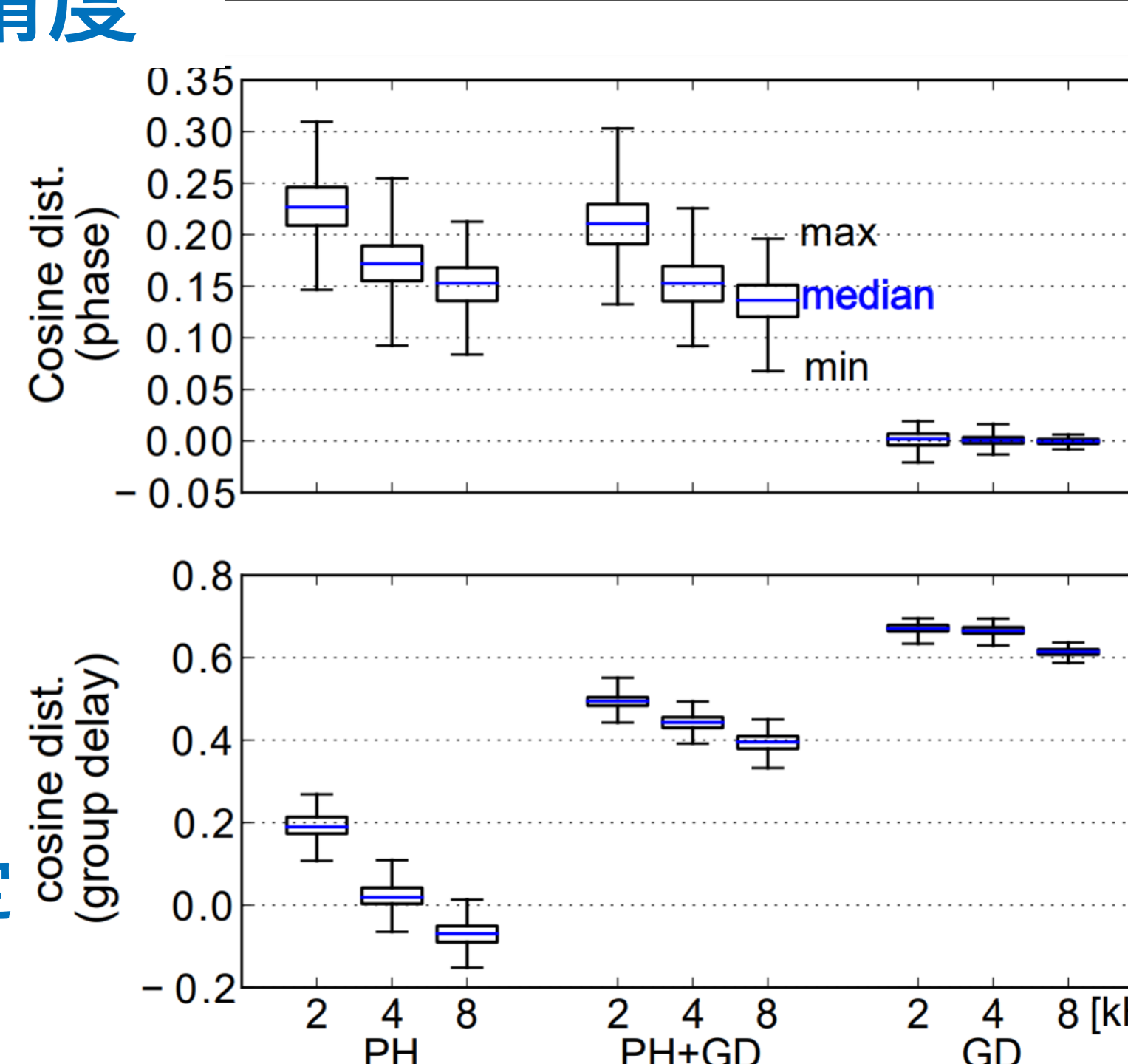
- 評価法
 - ・音質に関するABテスト (主観評価試験)
- 結果
 - ・全設定で従来法より良い
 - ・マルチタスクの場合, 全設定で有意差あり。

プリファレンスABテスト(30人)の結果, クラウドソーシング評価システムで実施.

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)

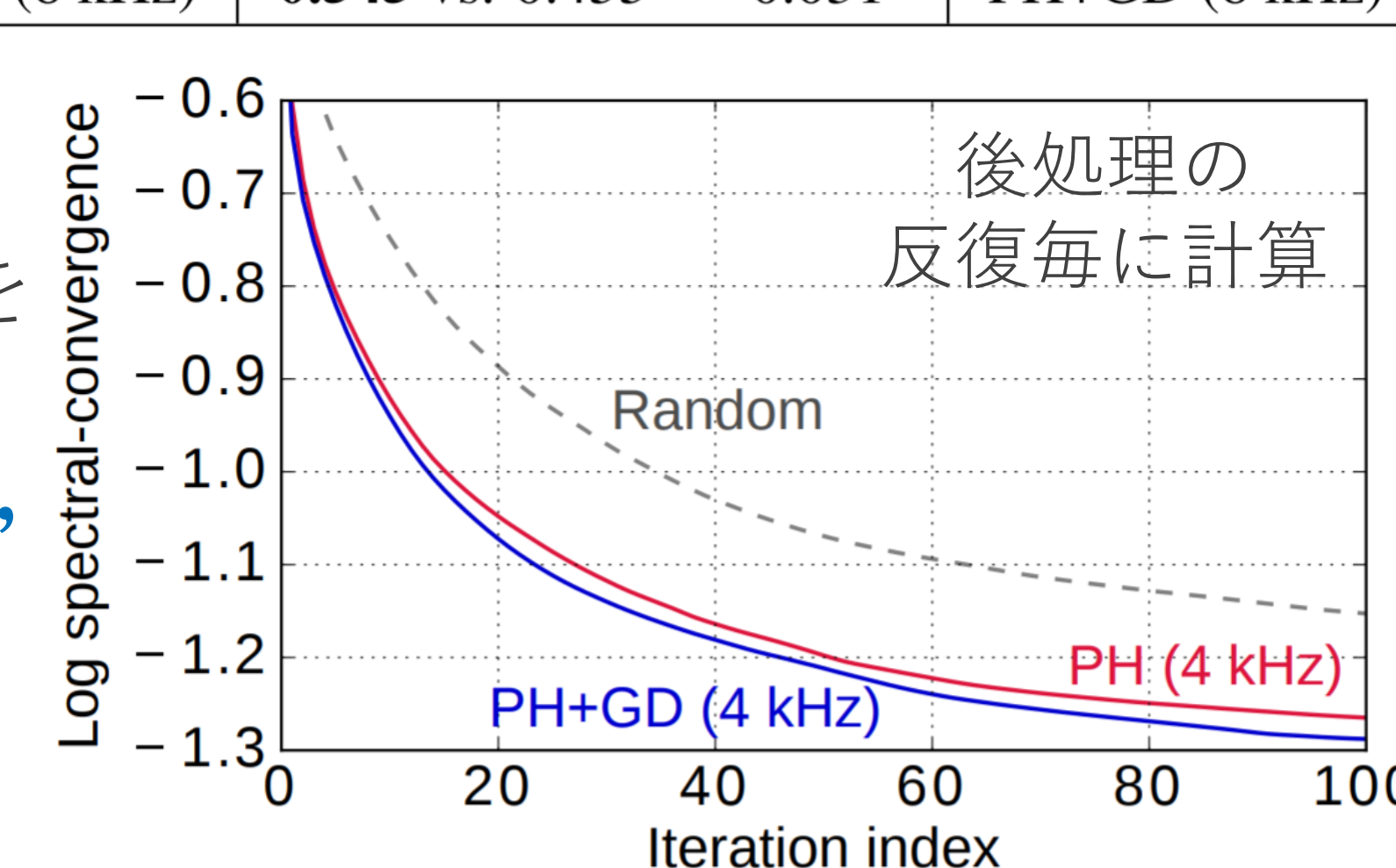
① 位相・群遅延の予測精度

- 比較手法
 - ・PH：位相ロスのみ
 - ・GD：群遅延ロスのみ
 - ・PH+GD：マルチタスク学習
- 評価法
 - ・正解とのコサイン距離
- 結果
 - ・位相より群遅延を高精度に推定
 - ・マルチタスクで、位相推定と群遅延推定の両立が可能



③ 群遅延ロスの効果

- 音質への影響
 - ・位相ロスよりわずかに**高品質** (有意差なし)
- 後処理での収束性能
 - ・STFT/逆STFTの再構成性能を図る spectral convergence^[5]
 - ・群遅延ロスを考慮することで、より完全再構成により近づく



Reference