Phase reconstruction from amplitude spectrograms based on von-Mises-distribution deep neural network

Shinnosuke Takamichi[*1], Yuki Saito[*1], Norihiro Takamune[*1], Daichi Kitamura[*2], and Hiroshi Saruwatari[*1]

*1: Univ. of Tokyo, Japan, *2: National Institute of Tech., Kagawa College, Japan

Hypothesis: phase reconstruction from amplitude spectrograms based on DNNs

Phase reconstruction
- Audio signal processing often processes amplitude spectrograms.
- Speech synthesis is shifting from vocoder params. to amplitudes.

➡ Griffin-Lim phase reconstruction method[1] **provides unnatural artifacts in speech.**

Griffin-Lim method [1]
A phase reconstruction method by iterating STFT and inverse STFT.

DNN-based phase reconstruction
- Can we train DNNs to predict the phase?
- **Isotropic-Gaussian-distribution DNN (mean squared error training) is not suitable because phase is a periodic variable.**

Our approach
- Propose a novel **DNN that has the von Mises distribution** which is a probability distribution for a periodic variable.
- Introduce **group delays** that has strong relationship to the amplitude of speech.

1) DNN can predict group delay accurately more than phases, and
2) our methods achieve better speech quality than the conventional Griffin-Lim method.

Proposed method: von-Mises-distribution DNN-based phase reconstruction

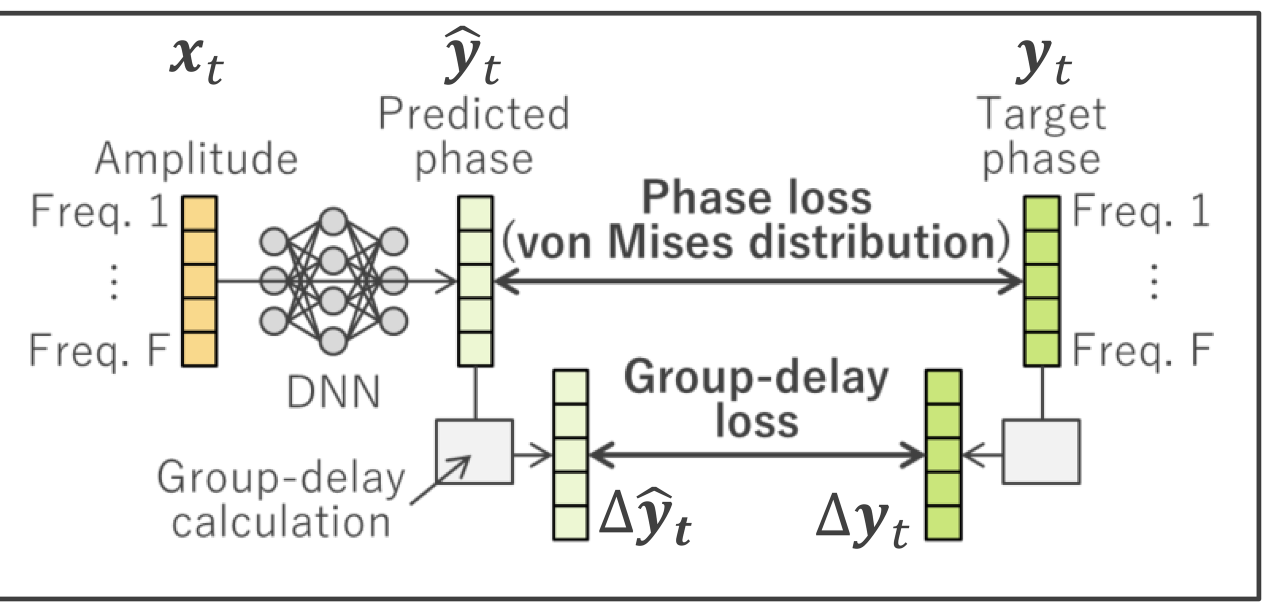
von Mises distribution and DNN-based phase reconstruction

von Mises (vM) distribution[3]
- $P(y;\mu,\kappa)=e^{\kappa\cos(y-\mu)}/2\pi I_0(\kappa)$
  - μ: mean, κ: cnocentration
  - $I_0(\cdot)$: Modified Bessel function

Negative log likelihood (μ: parameter)
- $-\log P(y;\mu,\kappa)\propto-\cos(y-\mu)$

DNN-based phase reconstruction
- DNN that convert an amplitude $\boldsymbol{x}_t$ to phase $\boldsymbol{y}_t$ (t is the frame index.)

Loss for DNN training
- **Phase loss** derived from vM dist.
- **Group-delay loss**



1) Phase loss
- Maximum likelihood estimation of vM distribution.

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

$y_{t,f}$: phase at t-th frame and f-th freq. bin

2) Group-delay loss
- Approximate group delay with 1st-order difference.

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

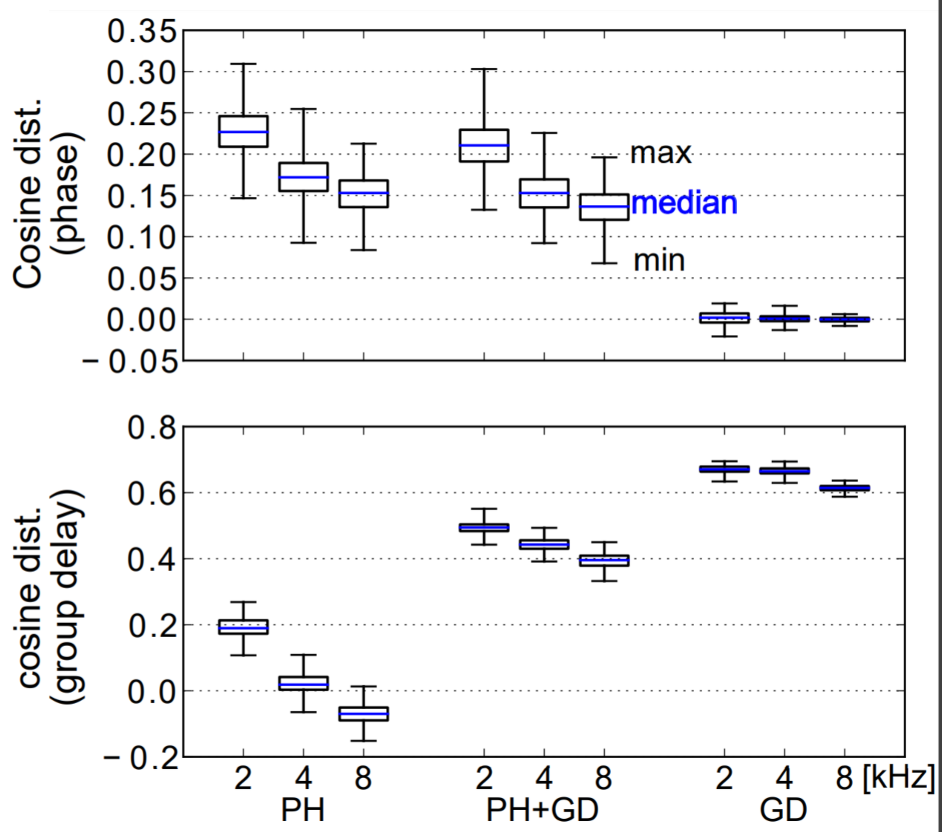
*Group delay and phase of AR models have strong relationship [3].

$\Delta y_{t,f}=-(y_{t,f+1}-y_{t,f})$: group delay at t-th frame and f-th freq. bin

Evaluation: prediction accuracy, effects to speech quality, and improvements by group delay

Contents	Value/Settings
Training / test data	JSUT speech corpus[4] 5000 / 300 utts.
Sampling freq.	16 kHz
Frame shift, FFT taps	80 samples (5 ms), 512 samples
DNN input	Log amplitudes at current ± 2 frames
DNN output	Phase (3 types: 0-2kHz, 0-4kHz, 0-8kHz)
DNN architecture	Feed-Forward w/ gated activation units
Post-process	Phase refinement by the Griffin-Lim method

1) Prediction accuracy
- Compared systems
  - PH: Phase loss only
  - GD: Group-delay loss only
  - PH+GD: Multi-task learning
- Evaluation method
  - Cosine distance
- Results
  - **Group delay is estimated accurately more than phase.**



2) Speech quality
- Evaluation methods
  - Preference AB tests on speech quality
- Results
  - **Better than Griffin-Lim**
  - **Multi-task learning achieves better in all settings**

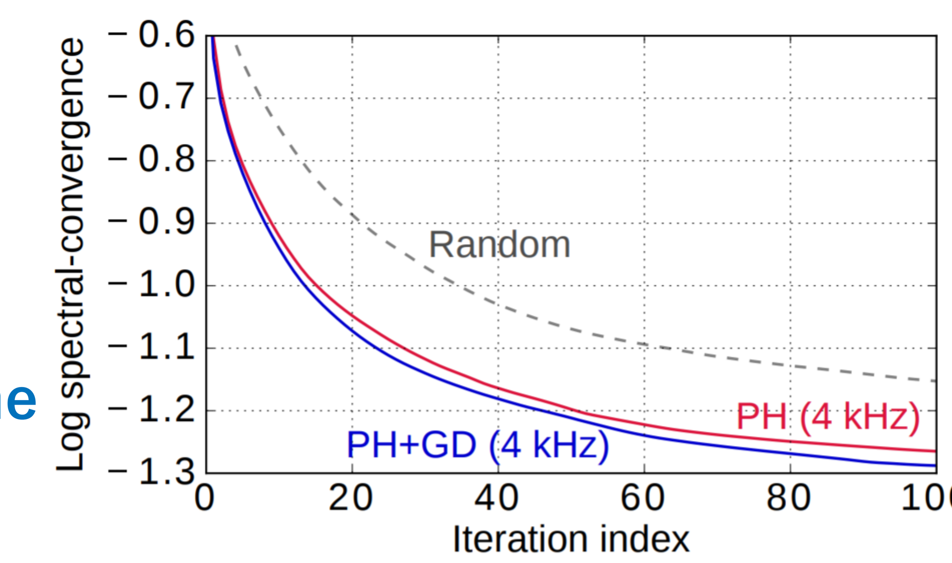
Preference AB tests by 30 listeners on crowdsourcing

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)

3) Improvements by group-delay loss
- Speech quality
  - Better than phase loss

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)

- Convergence in post-process
  - Spectral convergence[5]: reconstruction performance through STFT & inverse STFT
  - **Group-delay loss provides phases that are closest to the perfect reconstruction**



Reference
[1] Griffin et al., IEEE Trans., 1984. [2] Bishop, Springer, 2006. [3] Itakura et al., Proc. ICASSP, 1987. [4] Sonobe et al., arXiv, 2007.
[5] Sturmel et al., Intl. Conf. on Digital Audio Effects, 2011.