3451

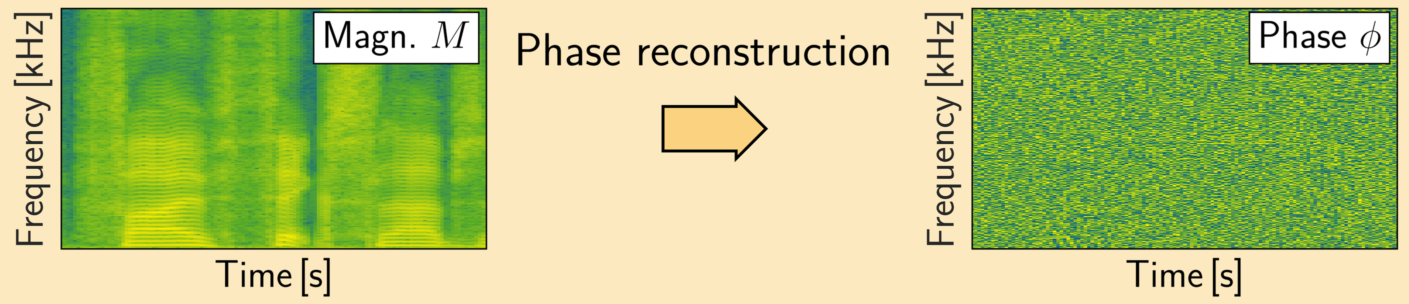
Recurrent Phase Reconstruction Using Estimated Phase Derivatives From Deep Neural Networks

Institute of Communication Systems, Prof. Dr.-Ing. Peter Jax — RWTH Aachen University

Lars Thieling, Daniel Wilhelm, Peter Jax

1 Introduction

Problem: Estimate phase ϕ from given magnitude spectrum M such that a consistent time signal is achieved via inverse short-time Fourier transform (ISTFT)



Applications:

- Speech enhancement and speech separation
- Speech synthesis and voice conversion

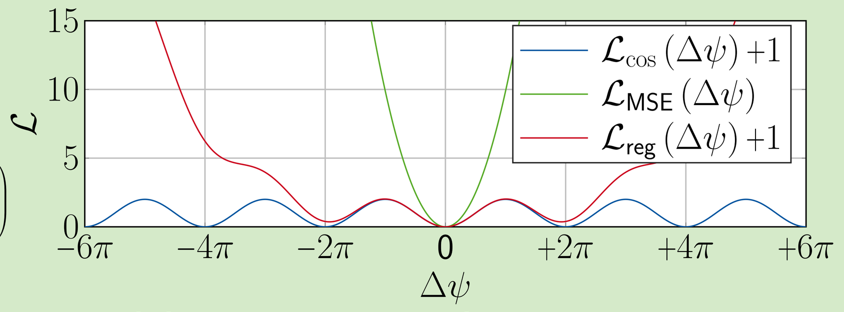
3 Phase Derivatives Estimation

New!

Used Loss Functions

- Train two equally structured DNNs using combined loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}\left(\Delta\hat{\psi}_{\text{if}}\right) + \mathcal{L}\left(\Delta\hat{\psi}_{\text{gd}}\right)$$



- $\mathcal{L}$ should consider 2π ambiguity and have a limited solution space
- **Novelty I** - regularized cosine loss function:

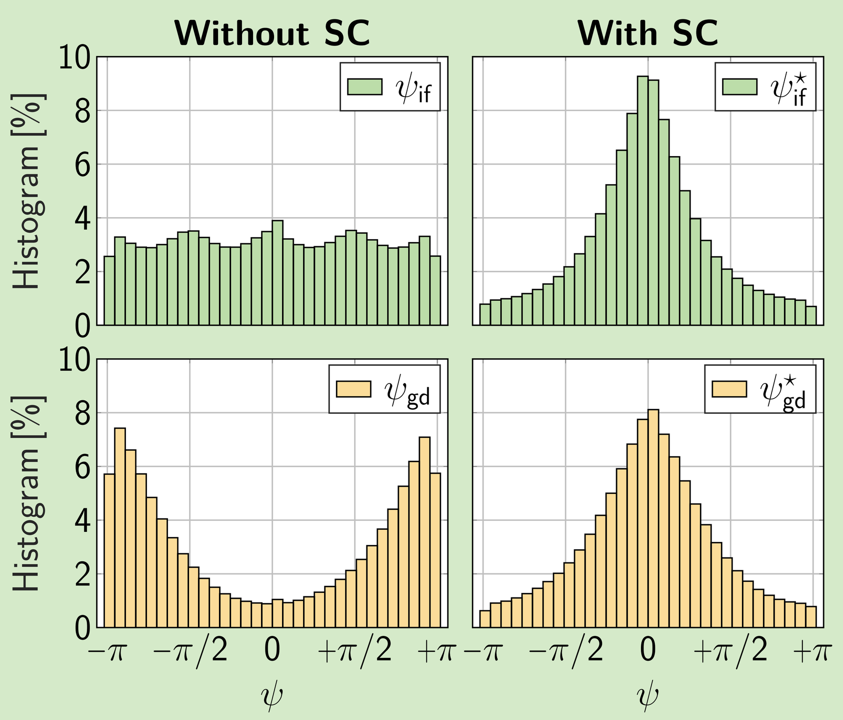
$$\mathcal{L}_{\text{reg}}\left(\Delta\hat{\psi}\right) := \sum_{k,m} -\cos\left(\Delta\hat{\psi}(k,m)\right) + \lambda \cdot \left(\Delta\hat{\psi}(k,m)\right)^4$$

Here: $\lambda = \frac{1}{4000}$

Shift Correction

- Systematic offsets occur in the calculation of ψ_{if} and ψ_{gd}
- Offset in ψ_{if} can be described by the shift theorem of the DFT:

$$x(n-S) \leftrightarrow X(k) \cdot e^{j\frac{2\pi}{N}kS}$$

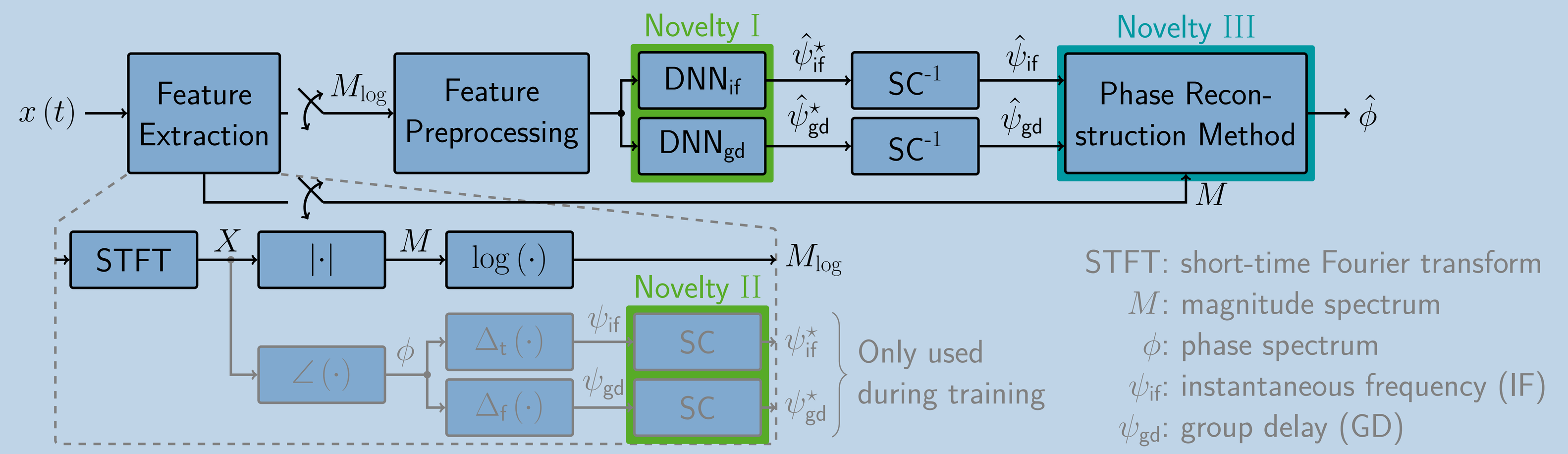


- Systematic shift in ψ_{gd} can be observed empirically
- **Novelty II** - shift correction:

$$\psi^{\star}_{\text{if}}(k,m) = \mathcal{W}\left(\psi_{\text{if}}(k,m) - \frac{\pi}{2}k\right)$$
$$\psi^{\star}_{\text{gd}}(k,m) = \mathcal{W}\left(\psi_{\text{gd}}(k,m) + \pi\right)$$

DFT: discrete Fourier transform
$S = \frac{N}{4}$: window shift
N: DFT size
$\mathcal{W}(\cdot)$: wrapping operator

2 System Overview



STFT: short-time Fourier transform
M: magnitude spectrum
ϕ: phase spectrum
ψ_{if}: instantaneous frequency (IF)
ψ_{gd}: group delay (GD)

Two-stage phase reconstruction system (similar to [1]):

1. Use deep neural networks (DNNs) to estimate phase derivatives 3

$$\psi_{\text{if}}(k,m) := \Delta_t\phi(k,m) = \phi(k,m) - \phi(k,m-1)$$
$$\psi_{\text{gd}}(k,m) := \Delta_f\phi(k,m) = \phi(k,m) - \phi(k-1,m)$$

k: freq. bin index
m: frame index

2. Reconstruct phase from its estimated derivatives 4

Proposed improvements:

I. A novel regularized cosine loss function
II. Shift correction (SC) as a pre-processing step
III. A novel phase reconstruction method

4 Phase Reconstruction Method

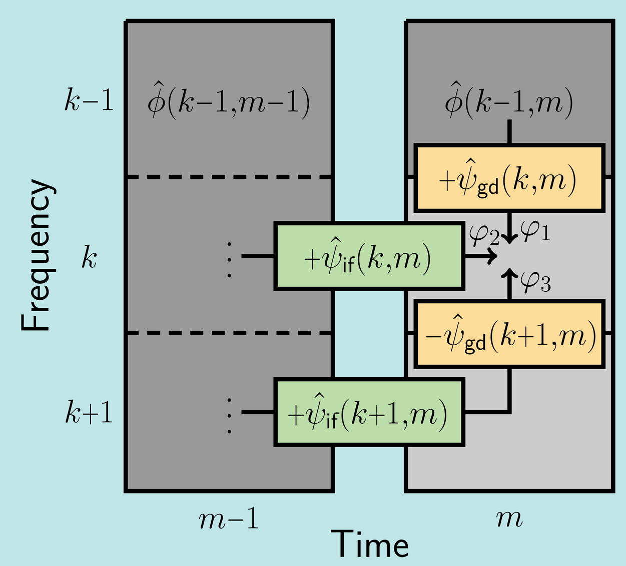
New!

- Combine $\hat{\psi}_{\text{if}}$ and $\hat{\psi}_{\text{gd}}$ such that a consistent $\hat{\phi}$ is achieved
- **Novelty III** - averaging of weighted estimates φ_p from P paths:

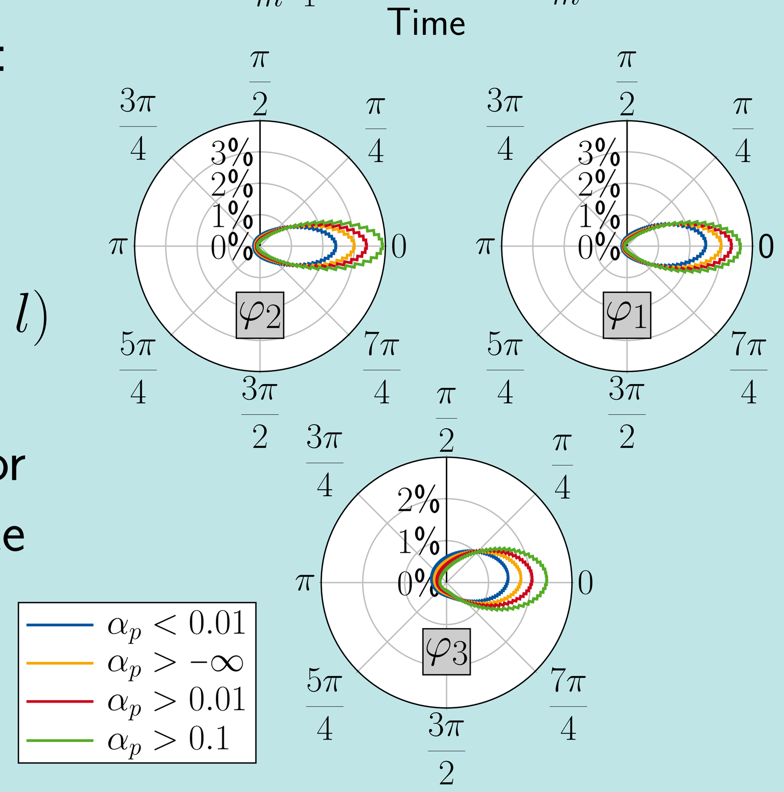
$$\hat{\phi}(k,m) = \angle \sum_{p=1}^{P} \alpha_p(k,m) \cdot e^{j\varphi_p(k,m)}$$

with estimation quality indicators α_p:

$$\alpha_1(k,m) = M(k-1,m)$$
$$\alpha_2(k,m) = M(k,m-1)$$
$$\alpha_3(k,m) = \min_{l=\{-1,0\}} M(k+1,m+l)$$



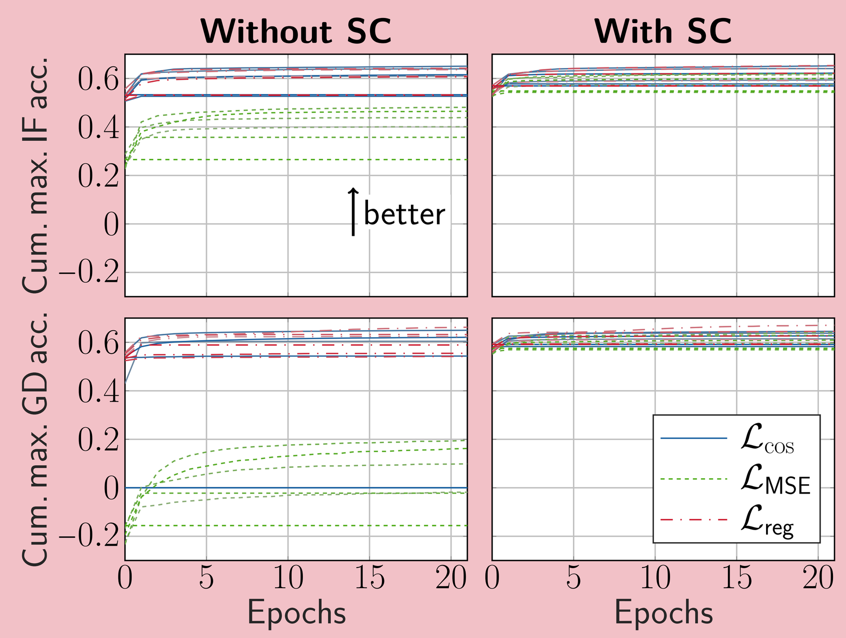
- Polar histograms of path error $\varphi_p(k,m) - \phi(k,m)$ demonstrate suitability of chosen weights



5 Evaluation

Phase Derivatives Estimation

Validation accuracies of different DNN configurations during training:

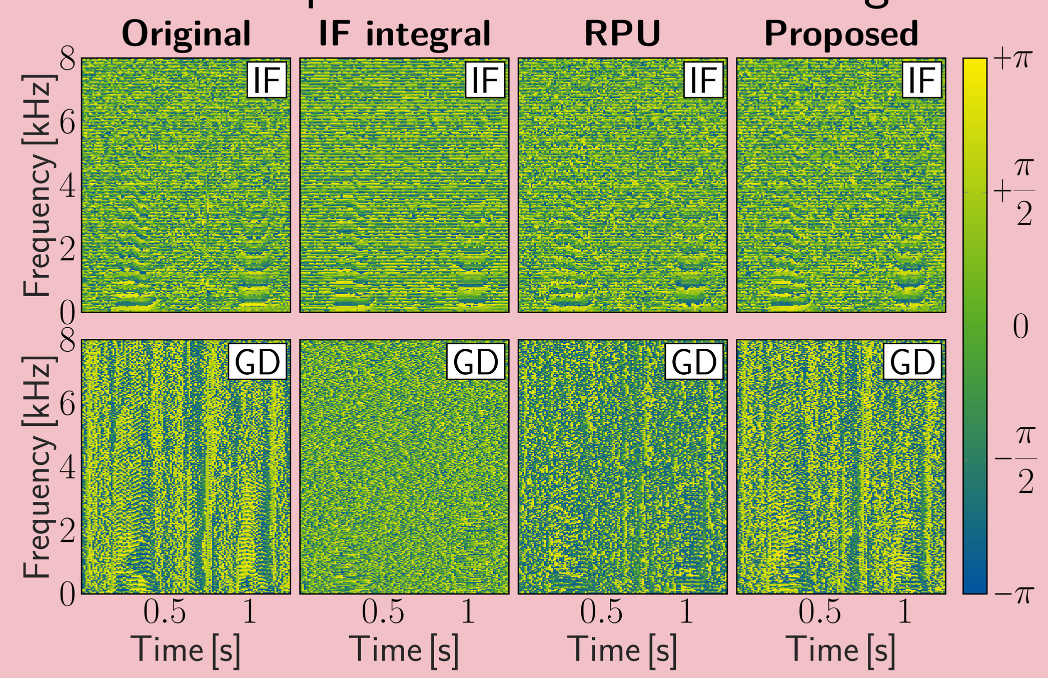


Accuracy: • Mean cosine error
Dataset: • 18.5 hours training data
• 3.5 hours validation data
• 16 kHz sample rate
DNN: • 3 hidden layers à 1024 units
• Normalized log magnitude of frames at ± 2, ± 1 and 0 as input features
• Varying activation function: sigmoid, tanh, ReLU, LeakyReLU, gated linear, gated tanh
STFT: • 640 samples Hann window
• 160 samples window shift
• 640 DFT size

→ $\mathcal{L}_{\text{MSE}}$ is inappropriate
→ $\mathcal{L}_{\text{reg}}$ stabilizes training of GD
→ SC increases accuracy in first epoch
→ SC stabilizes against hyperparameter variations

Phase Reconstruction Method

Results after phase reconstruction using different methods:



Method	STOI	PESQ
IF integral	0.8855	2.703
RPU [1]	0.9451	3.438
Proposed	**0.9852**	**4.197**
Griffin-Lim (10 it.)	0.9642	3.597
Griffin-Lim (100 it.)	**0.9920**	**4.246**

Scan for more audio samples:

http://iks.rwth-aachen.de/qr/icassp2021-rpr

→ Proposed reconstruction method outperforms reference algorithms
→ Similar performance to Griffin-Lim (100 it.) although no iterations are required

6 Conclusion

- ▶ Proposed novelties significantly improve phase reconstruction system
- ▶ Novelty I - regularized cosine loss function stabilizes training
- ▶ Novelty II - shift correction further stabilizes and accelerates training
- ▶ Novelty III - phase reconstruction method outperforms reference algorithms

References

[1] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, "Phase reconstruction based on recurrent phase unwrapping with deep neural networks," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.