Recurrent Phase Reconstruction Using Estimated Phase Derivatives from Deep Neural Networks

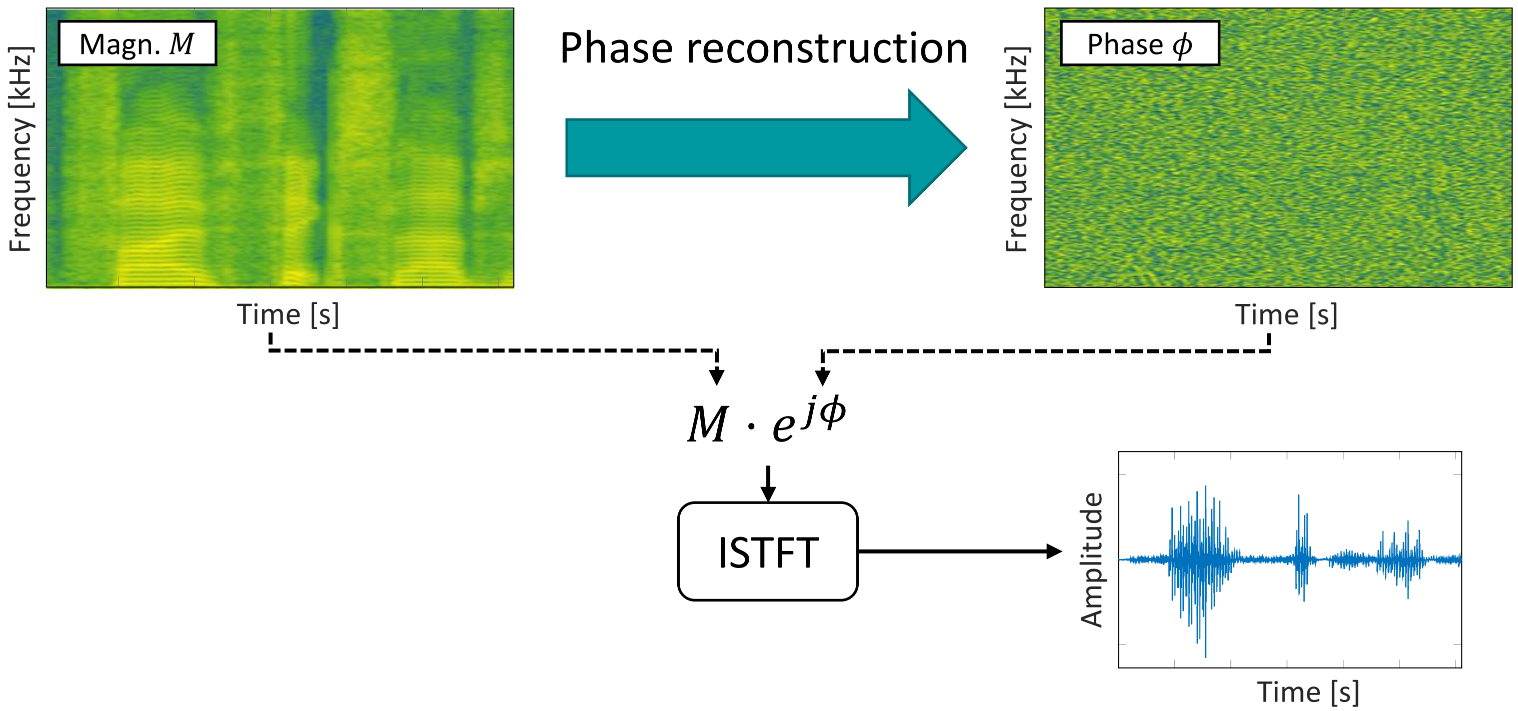
Lars Thieling, Daniel Wilhelm, Peter Jax

IEEE ICASSP, 2021

Institute of Communication Systems
Prof. Dr.-Ing. Peter Jax
RWTH AACHEN UNIVERSITY

Problem Statement And Motivation

- Problem: Reconstruct phase from given magnitude spectrum

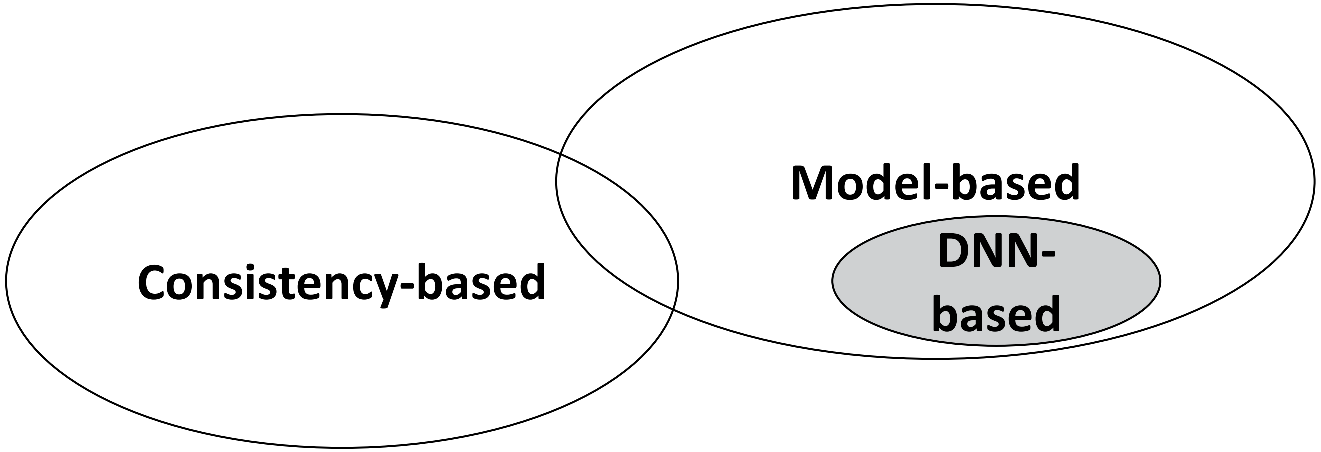


- Many algorithms only process/generate the magnitude spectrum of speech, e.g. in
  - Speech enhancement and speech separation
  - Speech synthesis and voice conversion

ISTFT: Inverse short-time Fourier transform

Introduction

- Phase reconstruction approaches
  - Consistency-based approaches
    - Exploit properties of overlapping frames within STFT, e.g., [Griffin et al., ICASSP1984]
  - Model-based approaches
    - Based on models of the target signal

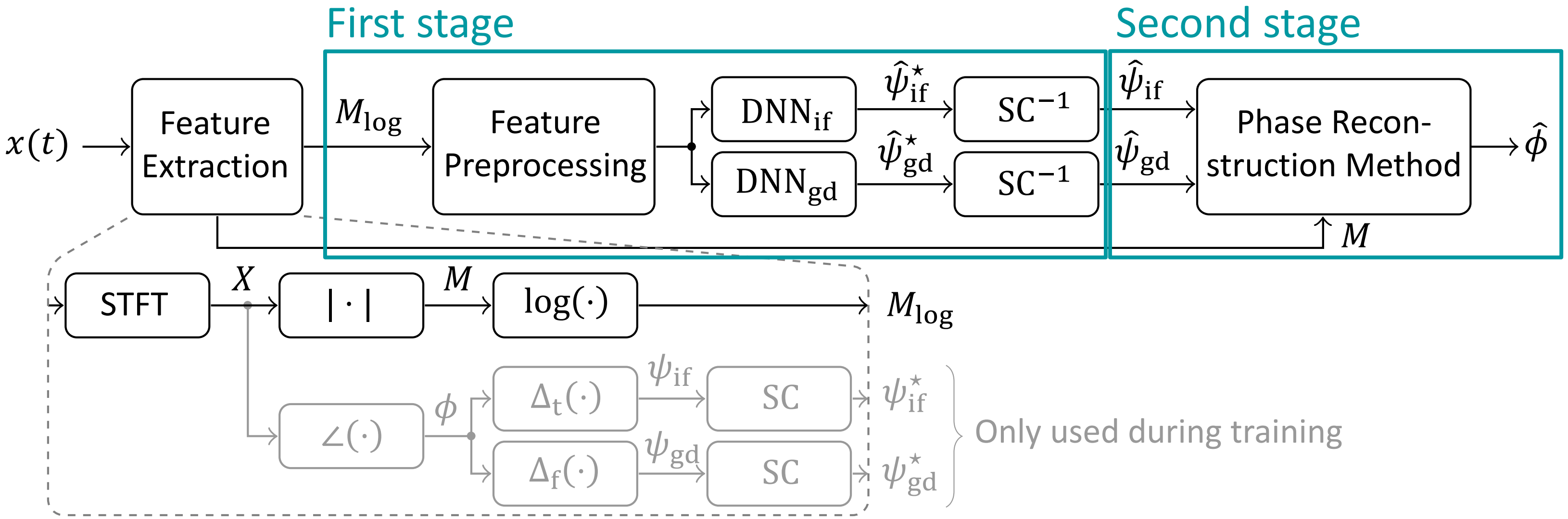


- Deep neural network (DNN)-based approaches, that estimate
  - Discretized phase [Takahashi et al., Interspeech2018]
  - Continuous phase [Takamichi et al., IWAENC2018]
  - Complex-valued spectrum [Oyamada et al., EUSIPCO2018]

- **Here:** Two-stage phase reconstruction system based on [Masuyama et al., ICASSP2020]
  1. Estimate phase derivatives using DNNs
  2. Reconstruct phase spectrum from its estimated derivatives

Two-Stage Phase Reconstruction System

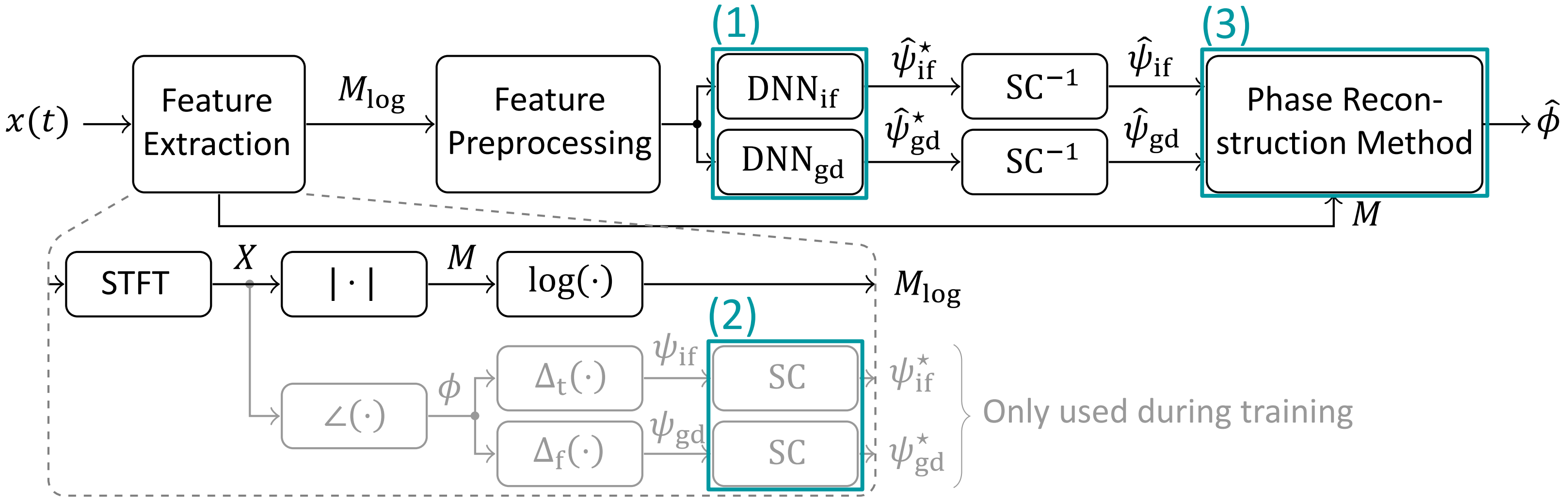
Block diagram of the overall system



ψ_{if}: Instantaneous frequency (IF)
ψ_{gd}: Group delay (GD)
M: Magnitude spectrum

Two-Stage Phase Reconstruction System

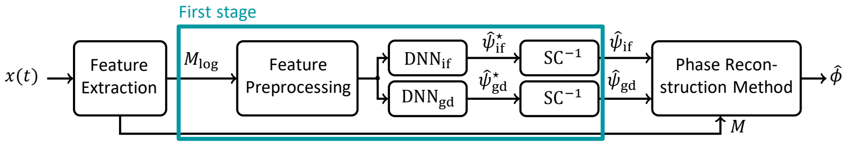
Block diagram of the overall system



- Proposed improvements:
  - (1) A novel regularized cosine loss function
  - (2) Shift correction (SC) as a pre-processing step for the phase derivatives during training
  - (3) A novel phase reconstruction method

ψ_{if}: Instantaneous frequency (IF)
ψ_{gd}: Group delay (GD)
M: Magnitude spectrum

Two-Stage Phase Reconstruction System



Phase derivatives estimation

- For discrete-time signals the phase derivatives can be approximated by:
  - Instantaneous frequency (IF):

$$\psi_{\mathrm{if}}(k,m) \coloneqq \Delta_{\mathrm{t}}\, \phi(k,m) = \phi(k,m) - \phi(k,m-1)$$

  - Group delay (GD):

$$\psi_{\mathrm{gd}}(k,m) \coloneqq \Delta_{\mathrm{f}}\, \phi(k,m) = \phi(k,m) - \phi(k-1,m)$$

- Two equally structured and simultaneously trained DNNs using combined loss:

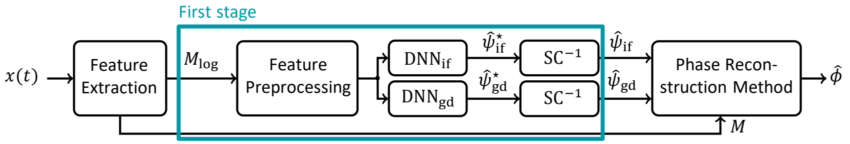
$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}\big(\psi_{\mathrm{if}} - \hat{\psi}_{\mathrm{if}}\big) + \mathcal{L}\big(\psi_{\mathrm{gd}} - \hat{\psi}_{\mathrm{gd}}\big)$$

- Phase and its derivatives are periodic variables and are typically wrapped to $[-\pi, \pi)$

  ➡ $\mathcal{L}$ should consider this ambiguity of 2π

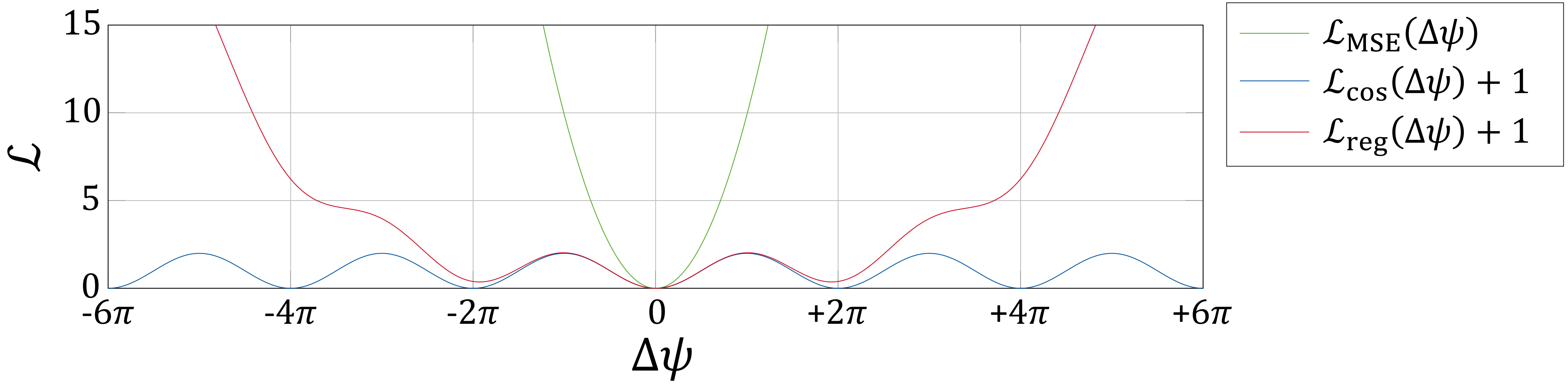
  ➡ $\mathcal{L}$ should have a limited solution space

ϕ: Phase spectrum
k: Frequency bin
m: Frame index



Novelty (1): Regularized Cosine Loss Function

Used loss functions $\mathcal{L}$



- Regularized cosine function (**Novel!**):

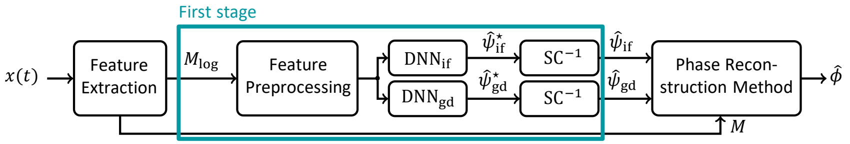
$$\mathcal{L}_{\mathrm{reg}}(\Delta\hat{\psi}) := \underbrace{\sum_{k,m} -\cos\left(\Delta\hat{\psi}(k,m)\right)}_{\mathcal{L}_{\cos}(\Delta\hat{\psi})} + \lambda \cdot \left(\Delta\hat{\psi}(k,m)\right)^4$$

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

Loss	Ambiguity of 2π	Solution space
$\mathcal{L}_{\mathrm{MSE}}$	☹	☺
$\mathcal{L}_{\cos}$	☺	☹
$\mathcal{L}_{\mathrm{reg}}$	☺	☺

$\mathcal{L}_{\mathrm{MSE}}$: Mean-squared error
$\mathcal{L}_{\cos}$: Negative cosine func. [Takamichi et al., IWAENC2018]

Novelty (2): Pre-Processing Via Shift Correction (SC)



Shift correction (SC)

- For the IF, a systematic offset can be described by the shift theorem of the discrete Fourier transform (DFT):

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

$S = \frac{N}{4}$: Window shift
N: DFT size

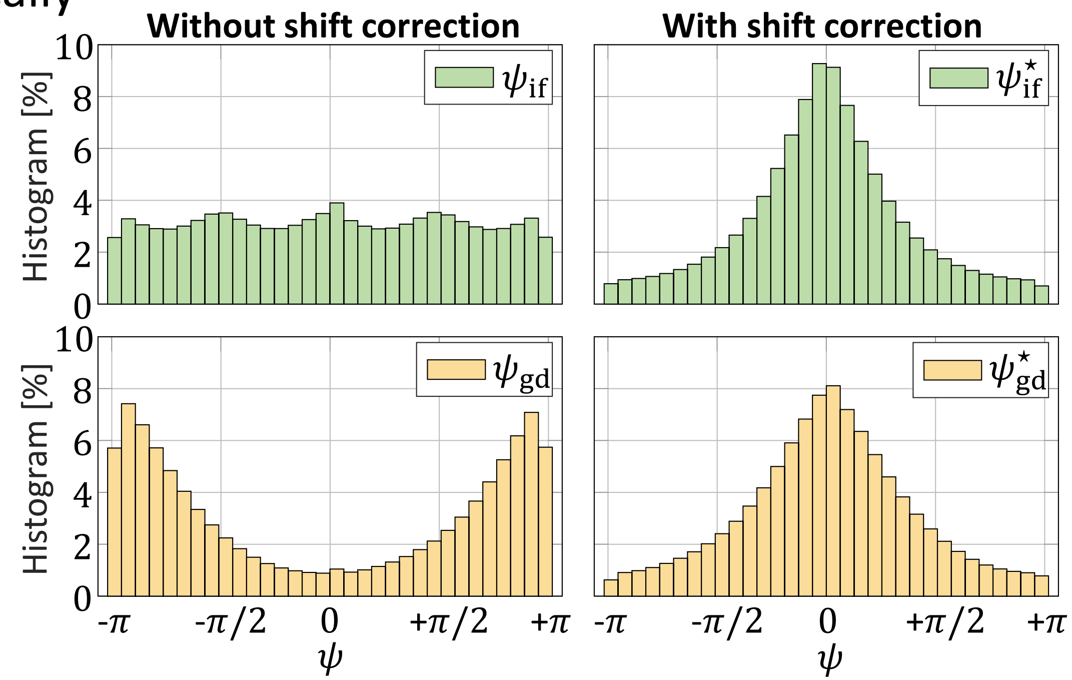
- For the GD, a systematic shift of π can be observed empirically

- Both shifts can be corrected:

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

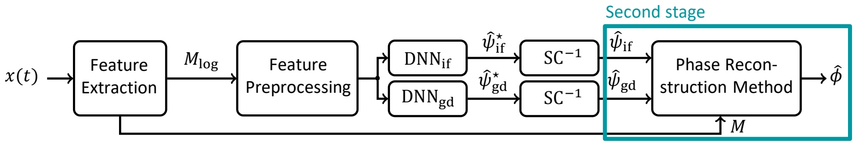
➡ Reduced standard deviation and mean close to 0

➡ Number of values near $\pm\pi$ is reduced



$\mathcal{W}$: Wrapping operator

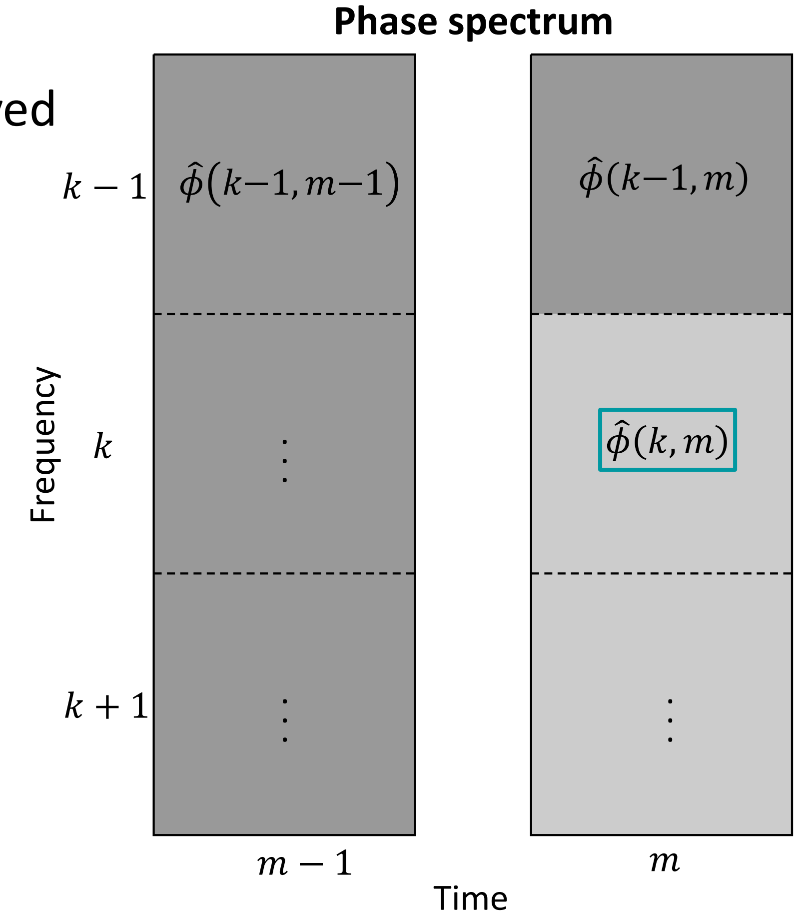
Novelty (3): Phase Reconstruction Method



Phase reconstruction from its estimated derivatives

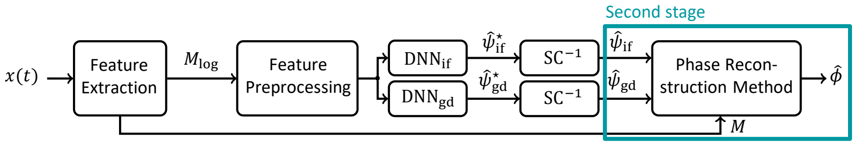
- Combine $\hat{\psi}_{\mathrm{if}}$ and $\hat{\psi}_{\mathrm{gd}}$ such that a consistent phase spectrum $\hat{\phi}$ is achieved
- Averaging of weighted estimates from P paths (**Novel!**):

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



φ_p: Estimation of the p^{th} path

Novelty (3): Phase Reconstruction Method



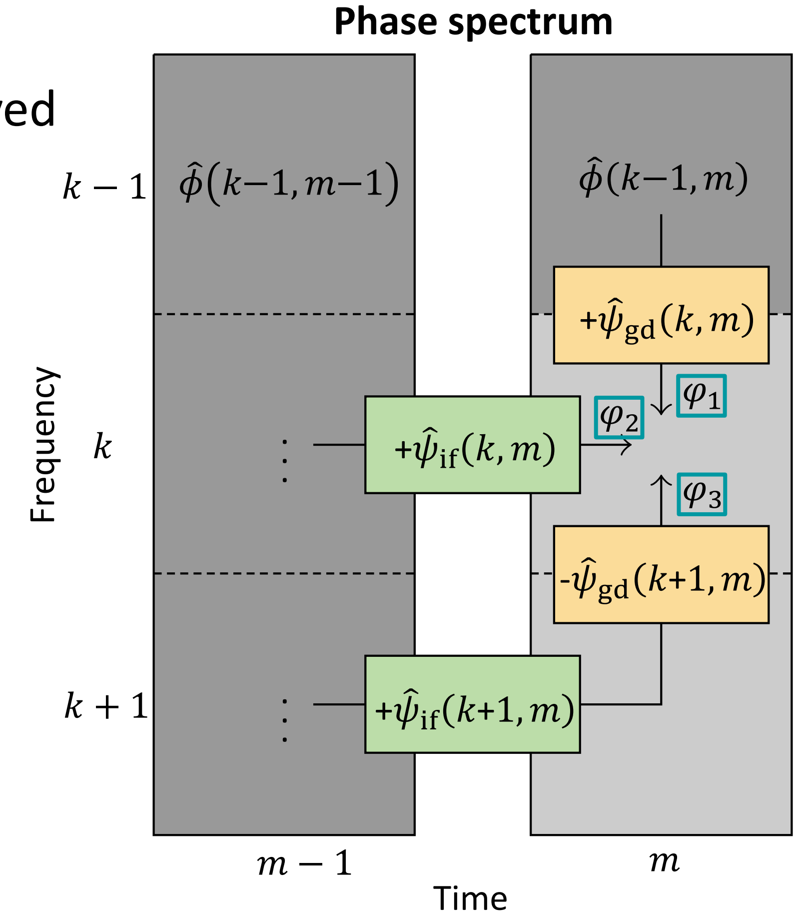
Phase reconstruction from its estimated derivatives

- Combine $\hat{\psi}_{\mathrm{if}}$ and $\hat{\psi}_{\mathrm{gd}}$ such that a consistent phase spectrum $\hat{\phi}$ is achieved
- Averaging of weighted estimates from P paths (**Novel!**):

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

- Quality indicator α_p, e.g., based on magnitude:

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



φ_p: Estimation of the p^{th} path
M: Magnitude spectrum

Evaluation

Experimental setup

Category	Parameters
Dataset	VCTK database [Veaux et al, 2017] preprocessed (e.g. resampled to $16\ \mathrm{kHz}$) • 18.5 hours training data • 3.5 hours validation data
STFT	• 640 samples Hann window • 160 samples window shift • 640 DFT size
DNNs	• Normalized log magnitude of current frame and frames at ± 2, ± 1 as input features • 3 hidden layers • 1024 hidden units per hidden layer • Varying activation function: Sigmoid, tanh, ReLU, LeakyReLU, gated linear, gated tanh • Varying loss function: $\mathcal{L}_{\mathrm{MSE}}$, $\mathcal{L}_{\cos}$, $\mathcal{L}_{\mathrm{reg}}$
Quality Measures	• Accuracy: mean cosine error • Objective: PESQ and STOI

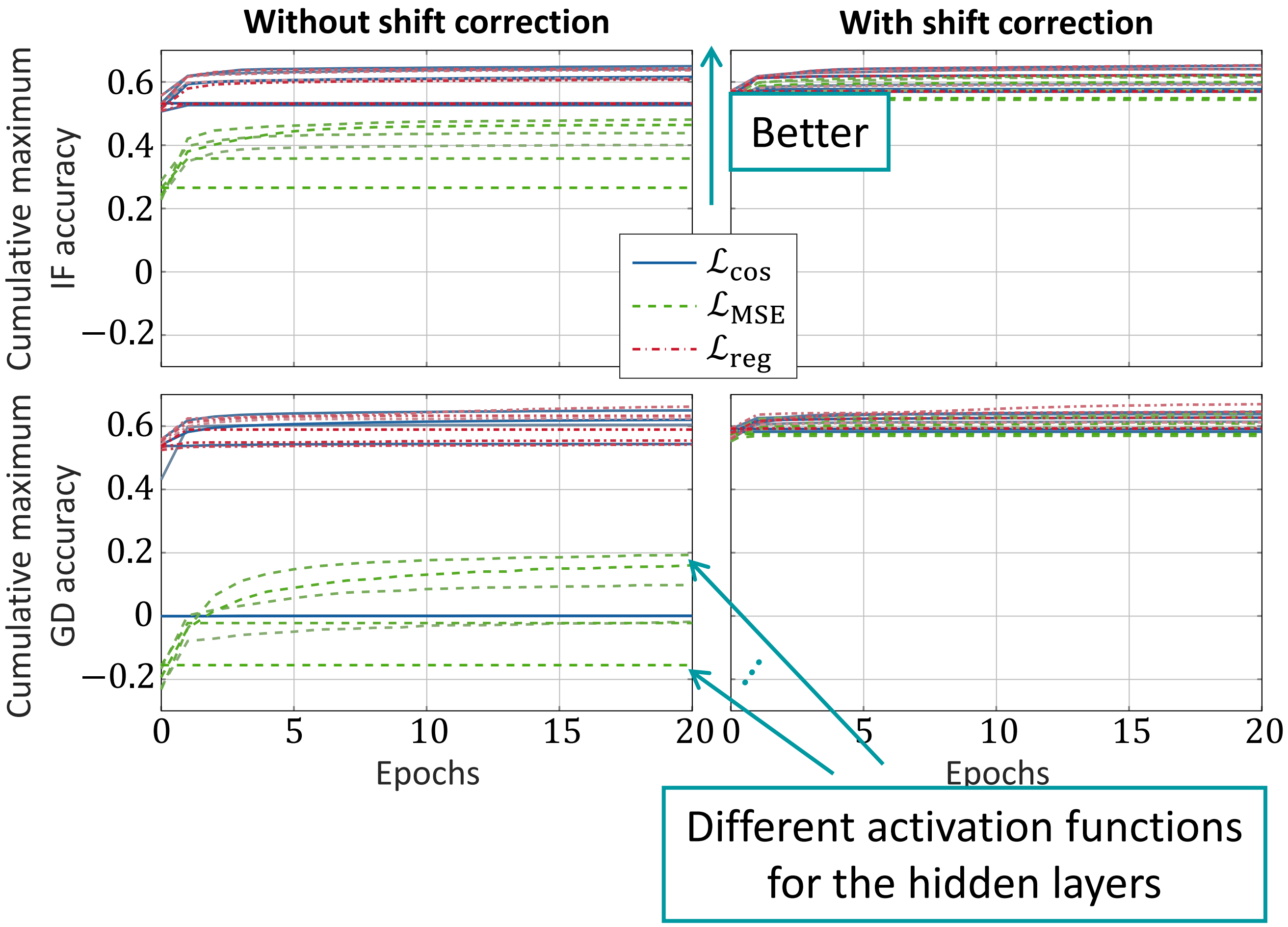
Evaluation – First Stage

Influence of the loss function

- $\mathcal{L}_{\mathrm{MSE}}$ is inappropriate for phase estimation
- $\mathcal{L}_{\mathrm{cos}}$ fails in two cases for GD estimation
- $\mathcal{L}_{\mathrm{reg}}$ stabilizes training of GD compared to $\mathcal{L}_{\mathrm{cos}}$

Influence of the shift correction

- Drastically increases accuracy in first epoch
  - Faster convergence
- Stabilizes against hyperparameter variations
  - All configurations reach very similar accuracies
  - Enables usage of $\mathcal{L}_{\mathrm{MSE}}$

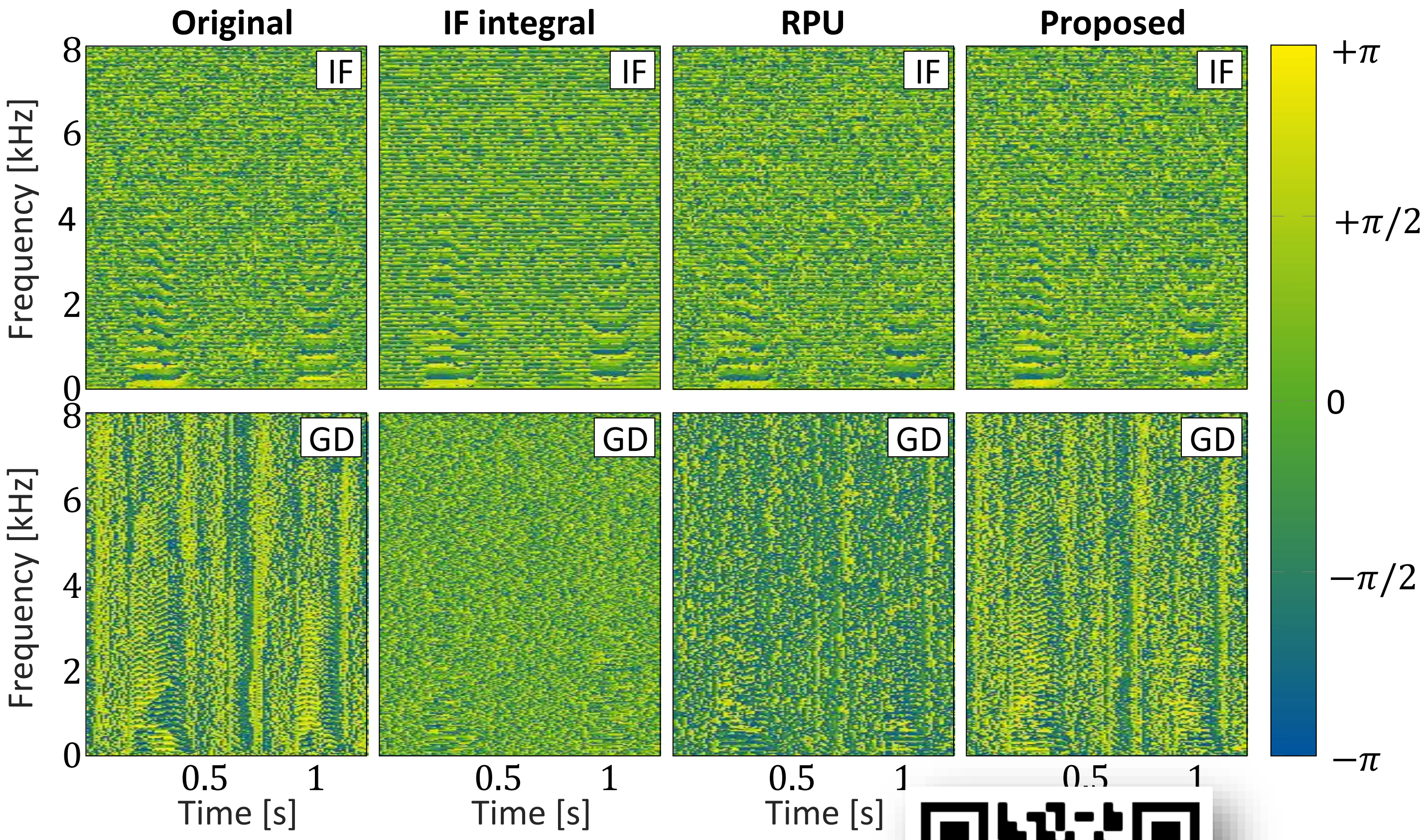


Evaluation – Second Stage

Influence of the phase reconstruction method

- Reference algorithms
  - Integration of IF considered in [Engel et al., ICLR2019]
  - Recurrent Phase Unwrapping (RPU) [Masuyama et al., ICASSP2020]

Method	STOI	PESQ
IF integral	0.8855	2.703
RPU	0.9451	3.438
Proposed	**0.9852**	**4.197**

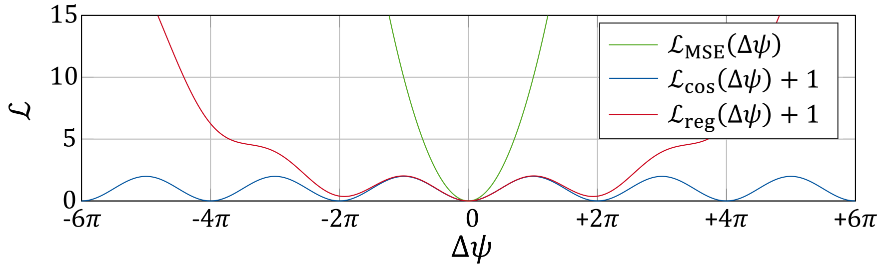


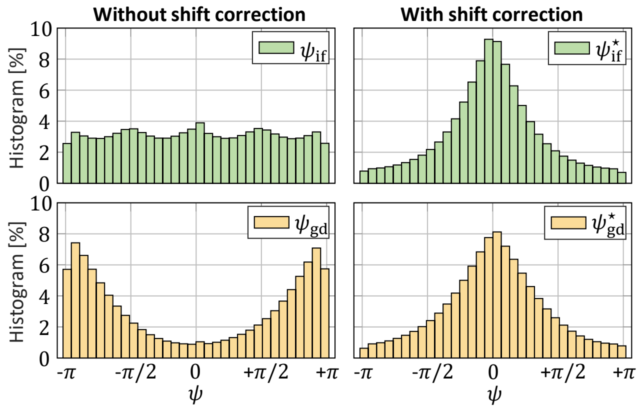
- Audio samples available at: http://iks.rwth-aachen.de/qr/icassp2021-rpr

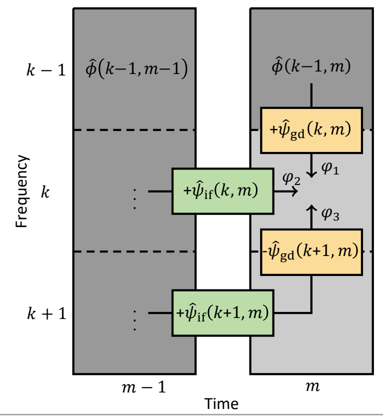
- Further evaluations available in paper

Conclusions

- Novelty (1): Regularized cosine loss function
  - Prevents arbitrary large/small predictions
  - Considers 2π ambiguity
  - → Reduces risk of diverging gradients and stabilizes training
- Novelty (2): Shift correction
  - → Stabilizes training against hyperparameter variations
  - → Reduces training duration
  - → Enables usage of $\mathcal{L}_{\mathrm{MSE}}$
- Novelty (3): Phase reconstruction method
  - Is simple but very effective
  - → Outperforms reference algorithms







iKS Institute of Communication Systems

RWTH AACHEN UNIVERSITY