Efficient Neural and Numerical Methods for High-Quality Online Speech Spectrogram Inversion via Gradient Theorem

Andres Fernandez, Juan Azcarreta, Çağdaş Bilen, Jesus M. Alvarez

a.fernandez@uni-tuebingen.de, jsmalvarez@meta.com

Meta Reality Labs

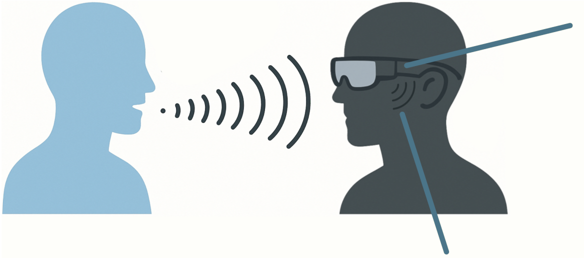
INTERSPEECH 2025

17-21 AUGUST – ROTTERDAM

Meta

Motivation

Meta



Wishlist:

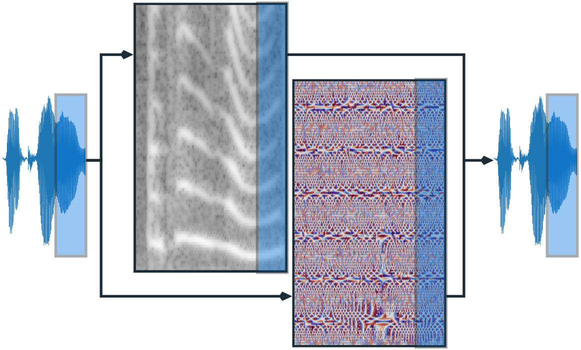
- Minimal **latency**
- Minimal **energy/arithmetic/memory**
- Good **quality** and **clarity**

Tasks:
- Speech enhancement
- Live translation
- . . .

Here: **Spectrogram-based methods**

Waveforms & Spectrograms

Meta



- Spectrograms $\rightarrow$ nice!
- Phases $\rightarrow$ messy! (irregular & 2π-periodic)
- Missing phase $\rightarrow$ Inverse STFT not trivial
- Modified spectrograms may be inconsistent

Here: **Real-Time Spectrogram Inversion (RSI)**

Efficient Neural and Numerical Methods

Meta

…for **real-time** spectrogram inversion. Improvements on previous 2-stage work:

	GL	RTISI	SPSI	GT+DL	Ours
Low-latency	✗	✓	✓	✓	✓
Low-compute	✗	✗	✓	✗	✓
High-quality	✗	✗	✗	✓	✓

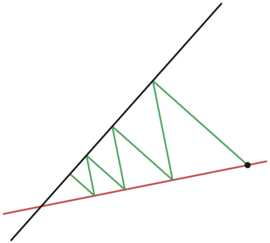
- $\sim$30x smaller/faster causal CNN
- Extra 2x at cost of 1 hop in latency
- Linear-complexity least-squares solver

Background

Consistency and Correctness

Meta

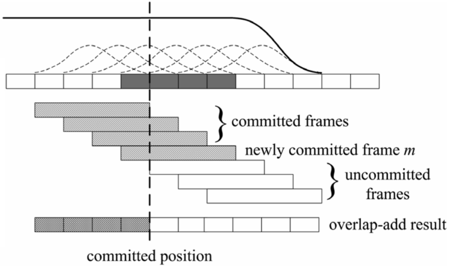
Griffin-Lim (Griffin et al. 1983)



Alternating GL projections (from Peer et al. 2022)

- Initialize: $\hat{Y} \leftarrow (|Y|, \phi)$ for some phase ϕ
- Consistency: $\hat{Y} \leftarrow \text{STFT} \circ \text{iSTFT} \circ \hat{Y}$
- Correctness: $\hat{Y} \leftarrow |Y| \frac{\hat{Y}}{|\hat{Y}|}$
- Recovery: $\hat{y} \leftarrow \text{iSTFT} \circ \hat{Y}$

RTISI (Beauregard et al. 2005)



GL on the current frame alone (from Zhu et al. 2007)

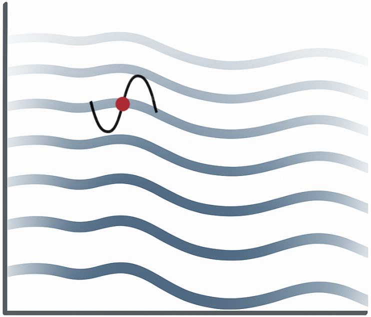
Real-time, but...

- Requires iterations
- Artifacts

Single-Pass Spectrogram Inversion

Meta

SPSI (Beauregard et al. 2015)



- Assume **Instantaneous Frequency**
- Initialize frame: $\hat{\boldsymbol{Y}}_\tau \leftarrow (|\boldsymbol{Y}|_\tau, \phi_\tau)$
- Inst. Freq.: $\omega \leftarrow$ spectral peaks in $\hat{\boldsymbol{Y}}_\tau$
- Propagate phase: $\phi_{\tau+1} \leftarrow \phi_\tau + \partial\tau \cdot \omega$

Iteration-free, but strong assumption $\rightarrow$ **artifacts**

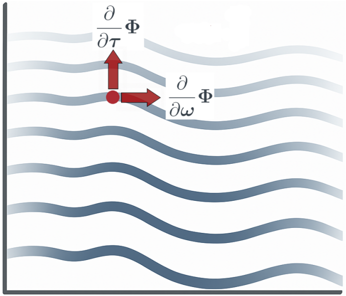
Increasing Quality with Better Assumptions

Meta

Gradient Theorem (Portnoff 1979): for Gaussian STFT window $\varphi_\lambda(t) := e^{-\pi \frac{t^2}{\lambda}}$,

$$\left.\begin{aligned} \frac{\partial}{\partial \omega} \operatorname{Arg}(Y_{y,\varphi_\lambda}(\omega, t)) &= -\lambda \frac{\partial}{\partial t} \log|Y_{y,\varphi_\lambda}(\omega, t)| \\ \frac{\partial}{\partial t} \operatorname{Arg}(Y_{y,\varphi_\lambda}(\omega, t)) &= \frac{1}{\lambda} \frac{\partial}{\partial \omega} \log|Y_{y,\varphi_\lambda}(\omega, t)| + 2\pi\omega \end{aligned}\right\}$$

Efficient numerical integration via **RTPGHI** algorithm (Průša et al. 2016)



Powerful!

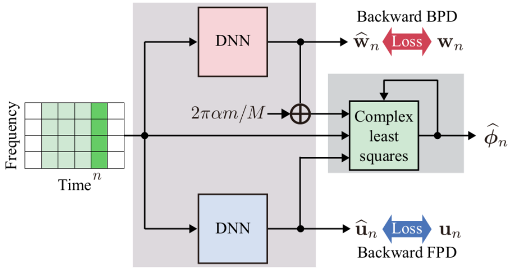
- Minimal latency, ∂ is local
- No assumptions on ***y***, only φ
- Still, error due to discretization and non-Gaussian φ

Two-Stage Framework with Deep Learning

Meta

Two stages (Masuyama et al. 2023):

- Predict $\partial\Phi$ from $\partial|\boldsymbol{Y}|$ using DL
- Reconstruct Φ from $\partial\Phi$ via complex least-squares



(ii) Detail of each stage

Two-stage GT+DL framework from Masuyama et al. 2023

Complex least-squares:

$$\boldsymbol{z}^{(\natural)} = \arg\min_{\boldsymbol{z}} \underbrace{\|\boldsymbol{z} - (\boldsymbol{Y}[\omega, \tau\text{-}1] \odot \mathfrak{v}_\tau)\|_\Lambda^2}_{\tau\text{-term}} + \underbrace{\|\boldsymbol{D}_\tau \boldsymbol{z}\|_\Gamma^2}_{\omega\text{-term}}$$

$$= \underbrace{\left(\Lambda + \boldsymbol{D}_\tau^{\mathsf{H}} \Gamma \boldsymbol{D}_\tau\right)}_{\boldsymbol{A}}{}^{-1} \underbrace{\Lambda(\boldsymbol{Y}[\omega, \tau\text{-}1] \odot \mathfrak{v}_\tau)}_{\boldsymbol{b}}$$

- $\mathfrak{v}$: transition from $\tau - 1$ to τ
- $\boldsymbol{Dz}$: transition from ω to $\omega + 1$
- Weights Λ, Γ to ignore small magnitudes
- Linear solver $\boldsymbol{z}^{(\natural)} = \boldsymbol{A}^{-1}\boldsymbol{b}$ for frame τ
- Desired phase is $\mathrm{Arg}(\mathrm{z}^{(\natural)})$

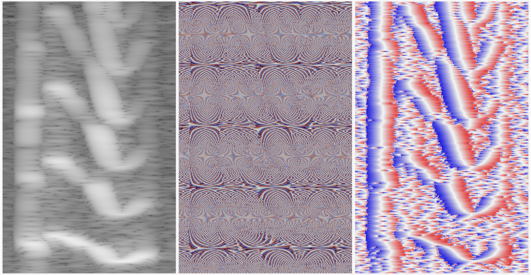
Training the DNN: Phases are not DL-friendly!

Meta

(Recall the GT:)

$$\frac{\partial}{\partial \omega}\mathrm{Arg}(\mathsf{Y}_{\mathsf{y},\varphi_\lambda}(\omega,\mathsf{t})) = -\lambda\frac{\partial}{\partial \mathsf{t}}\log|\mathsf{Y}_{\mathsf{y},\varphi_\lambda}(\omega,\mathsf{t})|$$

$$\frac{\partial}{\partial \mathsf{t}}\mathrm{Arg}(\mathsf{Y}_{\mathsf{y},\varphi_\lambda}(\omega,\mathsf{t})) = \frac{1}{\lambda}\frac{\partial}{\partial \omega}\log|\mathsf{Y}_{\mathsf{y},\varphi_\lambda}(\omega,\mathsf{t})| + 2\pi\omega$$

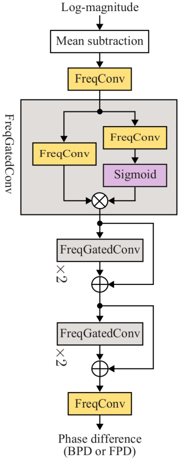


Two main issues → Solutions!

- Irregularity → **Train on derivatives!**
  - Takamichi et al. 2018; Takamichi et al. 2020; Thieling et al. 2021; Thien et al. 2023
- 2π periodicity → **Von-Mises Loss!**
  - $-\sum_\omega \sum_\tau \cos(\boldsymbol{X}[\omega,\tau]-\hat{\boldsymbol{X}}[\omega,\tau])$
  - Takamichi et al. 2018; Thien et al. 2021

Increased computation in Masuyama et al. 2023

DNN: 6 248k params
7.95 GMAC/s



Complex Least-Squares: Solving $\boldsymbol{z} = \boldsymbol{A}^{-1}\boldsymbol{b}$

$$\boldsymbol{z}_0^{(\natural)} = \underbrace{\left(\boldsymbol{\Lambda}_{\tau_0} + \boldsymbol{D}_{\tau_0}\boldsymbol{\Gamma}_{\tau_0}\boldsymbol{D}_{\tau_0}\right)}_{\boldsymbol{A}}{}^{-1} \underbrace{\boldsymbol{\Lambda}_{\tau_0}\left(\boldsymbol{Y}[\omega, \tau_{\text{-}1}] \odot \mathfrak{v}_{\tau_0}\right)}_{\boldsymbol{b}}$$

Solving $\boldsymbol{z} = \boldsymbol{A}^{-1}\boldsymbol{b}$:

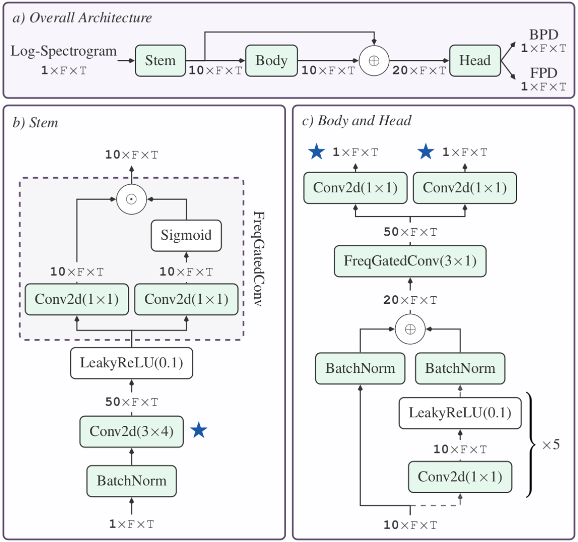
- Memory: $\mathcal{O}(\mathsf{L}^2)$ for STFT window of size $2\mathsf{L}$
- Naive inversion of $\boldsymbol{A}$ is $\mathcal{O}(\mathsf{L}^3)$
- Iterative solvers: $(\kappa(\mathsf{L}+1)^2)$ for κ iterations (Demmel 1997)
- Performed for every frame

Very high quality, but at **increased cost**

Contributions

Faster and Smaller First Stage

Meta



- Cheaper, FFW layers (BN, Conv1x1, LReLU)
- Less residual and gated convs
- Joint FPD and BPD
- Training: Adam with CosineWR schedule

Faster and smaller:

- Params: 248k $\rightarrow$ 8.5k ($\sim$30$\times$)
- GMAC/s: 7.95 $\rightarrow$ 0.27 ($\sim$30$\times$)
- 2x faster, +1hop latency ($\bigstar$)

Linear-Complexity Second Stage

Meta

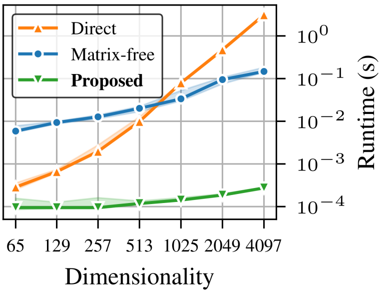
Recall: solving $\boldsymbol{z} = \boldsymbol{A}^{-1}\boldsymbol{b}$ has complexity $\sim \mathcal{O}(\kappa \cdot L^2)$:

$$\boldsymbol{z} = \underbrace{(\Lambda + \boldsymbol{D}^{H}\Gamma\boldsymbol{D})}_{\boldsymbol{A}}{}^{-1} \underbrace{\Lambda(\boldsymbol{Y} \odot \mathfrak{v})}_{\boldsymbol{b}}$$

Observation: $\boldsymbol{A}$ is PSD and tridiagonal!

$$\boldsymbol{D}^{H}\Gamma\boldsymbol{D} = \sum_{l=1}^{L} \gamma_l (\bar{d}_l \boldsymbol{e}_l + \boldsymbol{e}_{l+1})(d_l \boldsymbol{e}_l + \boldsymbol{e}_{l+1})^{T}$$

$$= \sum_{l=1}^{L} \gamma_l \Big(|d_l|^2 \underbrace{\boldsymbol{e}_l \boldsymbol{e}_l^{T}}_{\text{diag.}} + \underbrace{\boldsymbol{e}_{l+1} \boldsymbol{e}_{l+1}^{T}}_{\text{diag.}} \Big) + \sum_{l=1}^{L} \gamma_l d_l \underbrace{\boldsymbol{e}_{l+1} \boldsymbol{e}_l^{T}}_{\text{subdiag.}} + \sum_{l=1}^{L} \gamma_l \bar{d}_l \underbrace{\boldsymbol{e}_l \boldsymbol{e}_{l+1}^{T}}_{\text{superdiag.}}$$

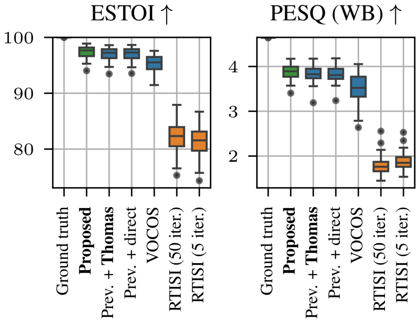


Thomas' Algorithm $\rightarrow \mathcal{O}(L)$ memory and arithmetic!

Retaining High Quality

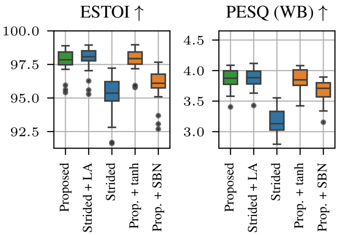
Meta

Intelligibility & Quality



- Inversion of LibriSpeech consistent spectrograms
- Consistently good results on both axes
- Strided version also competitive
- Variation study supports design choices

More results & samples



Thank you!

Meta

Conclusion:

- Low latency and high quality from DL + Gradient Theorem
- Tiny causal CNN for joint BPD/FPD
  - 2x inference at 1-hop extra latency
- Linear-complexity LSTSQ phase recovery

Future work:

- Subjective metrics
- Inconsistent/modified spectrograms
- Noisy phase as prior during inference
- Differentiable second stage
  - $\mathbf{\Lambda}, \mathbf{\Gamma}$ as ℓ_2 regularizers for DNN

Jesus M. Alvarez
Meta RL (Spain)

Juan Azcarreta
Meta RL (UK)

Çağdaş Bilen
Meta RL (UK)

References

Beauregard, Gerald T., Mithila Harish, and Lonce Wyse (2015). "Single Pass Spectrogram Inversion". In: IEEE International Conference on Digital Signal Processing (DSP).

Beauregard, Gerald T, Xinglei Zhu, and Lonce Wyse (2005). "An efficient algorithm for real-time spectrogram inversion". In: DAFx.

Demmel, James W. (1997). Applied Numerical Linear Algebra. Society for Industrial and Applied Mathematics.

Griffin, D. and Jae Lim (1983). "Signal estimation from modified short-time Fourier transform". In: ICASSP.

Masuyama, Yoshiki et al. (2023). "Online Phase Reconstruction via DNN-Based Phase Differences Estimation". In: TASLP 31, pp. 163–176.

Peer, Tal, Simon Welker, and Timo Gerkmann (2022). "Beyond Griffin-Lim: Improved Iterative Phase Retrieval for Speech". In: 2022 International Workshop on Acoustic Signal Enhancement (IWAENC), pp. 1–5. DOI: 10.1109/IWAENC53105.2022.9914686.

Portnoff, M. (1979). "Magnitude-phase relationships for short-time Fourier transforms based on Gaussian analysis windows". In: ICASSP.

Průša, Zdeněk and Peter L. Søndergaard (2016). "Real-Time Spectrogram Inversion Using Phase Gradient Heap Integration". In: DAFx.

Takamichi, Shinnosuke et al. (2018). "Phase Reconstruction from Amplitude Spectrograms Based on Von-Mises-Distribution Deep Neural Network". In: IWAENC.

— (Apr. 2020). "Phase reconstruction from amplitude spectrograms based on directional-statistics deep neural networks". In: Elsevier Signal Processing 169.C.

Thieling, Lars, Daniel Wilhelm, and Peter Jax (2021). "Recurrent Phase Reconstruction Using Estimated Phase Derivatives from Deep Neural Networks". In: ICASSP.

Thien, Nguyen Binh et al. (2021). "Two-stage phase reconstruction using DNN and von Mises distribution-based maximum likelihood". In: APSIPA ASC.

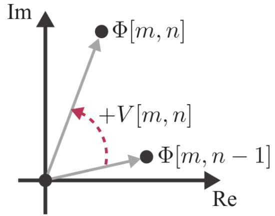
— (2023). "Inter-Frequency Phase Difference for Phase Reconstruction Using Deep Neural Networks and Maximum Likelihood". In: TASLP 31.

Zhu, Xinglei, Gerald T. Beauregard, and Lonce L. Wyse (2007). "Real-Time Signal Estimation From Modified Short-Time Fourier Transform Magnitude Spectra". In: IEEE Transactions on Audio, Speech, and Language Processing 15.5, pp. 1645–1653. DOI: 10.1109/TASL.2007.899236.

Complex Least-Squares from Masuyama et al. 2023: Details

$$|\mathfrak{u}_{\tau_0}| := \frac{|\boldsymbol{Y}[\omega,\tau_0]|}{|\boldsymbol{Y}[\omega\text{-}1,\tau_0]|}, \quad \mathrm{Arg}(\mathfrak{u}_{\tau_0}) := \mathrm{Arg}\left(\frac{\boldsymbol{Y}[\omega,\tau_0]}{\boldsymbol{Y}[\omega\text{-}1,\tau_0]}\right) = \boldsymbol{u}_{\tau_0}$$

$$|\mathfrak{v}_{\tau_0}| := \frac{|\boldsymbol{Y}[\omega,\tau_0]|}{|\boldsymbol{Y}[\omega,\tau_{\text{-}1}]|}, \quad \mathrm{Arg}(\mathfrak{v}_{\tau_0}) := \mathrm{Arg}\left(\frac{\boldsymbol{Y}[\omega,\tau_0]}{\boldsymbol{Y}[\omega,\tau_{\text{-}1}]}\right) = \boldsymbol{v}_{\tau_0}$$



Phase addition schematic from Masuyama et al. 2023

These ratios satisfy $\boldsymbol{Y}[\omega,\tau_0] = \boldsymbol{Y}[\omega\text{-}1,\tau_0] \odot \mathfrak{u}_{\tau_0}$ as well as $\boldsymbol{Y}[\omega,\tau_0] = \boldsymbol{Y}[\omega,\tau_{\text{-}1}] \odot \mathfrak{v}_{\tau_0}$ (assuming all $\boldsymbol{Y}[\omega,\tau] \neq 0$). This allows us to express $\boldsymbol{Y}[\omega,\tau_0]$ as the optimum of the following quadratic objective [21]:

$$\arg\min_{\boldsymbol{z}} \underbrace{\|\boldsymbol{z} - (\boldsymbol{Y}[\omega,\tau_{\text{-}1}] \odot \mathfrak{v}_{\tau_0})\|^2_{\boldsymbol{\Lambda}_{\tau_0}}}_{\tau\text{-term}} + \underbrace{\|\boldsymbol{D}_{\tau_0}\boldsymbol{z}\|^2_{\boldsymbol{\Gamma}_{\tau_0}}}_{\omega\text{-term}}$$

where $\boldsymbol{D}_{\tau_0} \in \mathbb{C}^{L\times(L+1)}$ is a matrix with $-\mathfrak{u}_{\tau_0}$ in the main diagonal, ones in the diagonal above, and zeros elsewhere. Here, $\|\boldsymbol{a}\|^2_{\boldsymbol{X}} := \boldsymbol{a}^{\mathsf{H}}\boldsymbol{X}\boldsymbol{a}$ is a weighted norm with *diagonal nonnegative* matrix $\boldsymbol{X}$, used in [21] to mitigate errors for small magnitudes. Equation 10 admits the following closed-form solution:

$$\boldsymbol{z}_0^{(\natural)} = (\boldsymbol{\Lambda}_{\tau_0} + \boldsymbol{D}^{\mathsf{H}}_{\tau_0}\boldsymbol{\Gamma}_{\tau_0}\boldsymbol{D}_{\tau_0})^{-1}\boldsymbol{\Lambda}_{\tau_0}(\boldsymbol{Y}[\omega,\tau_{\text{-}1}] \odot \mathfrak{v}_{\tau_0})$$