PHASE RECONSTRUCTION WITH LEARNED TIME-FREQUENCY REPRESENTATIONS FOR SINGLE-CHANNEL SPEECH SEPARATION

Gordon Wichern *Jonathan Le Roux*

Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA

ABSTRACT

Progress in solving the cocktail party problem, i.e., separating the speech from multiple overlapping speakers, has recently accelerated with the invention of techniques such as deep clustering and permutation free mask inference. These approaches typically focus on estimating target STFT magnitudes and ignore problems of phase inconsistency. In this paper, we explicitly integrate phase reconstruction into our separation algorithm using a loss function defined on time-domain signals. A deep neural network structure is defined by unfolding a phase reconstruction algorithm and treating each iteration as a layer in our network. Furthermore, instead of using fixed STFT/iSTFT time-frequency representations, we allow our network to learn a modified version of these representations from data. We compare several variants of these unfolded phase reconstruction networks achieving state of the art results on the publicly available wsj0-2mix dataset, and show improved performance when the STFT/iSTFT-like representations are allowed to adapt.

Index Terms— Source separation, iterative phase reconstruction, learnable representations

1. INTRODUCTION

Progress in separating multiple overlapping speakers using a single microphone, often referred to as the cocktail party problem, has accelerated greatly with the advent of deep neural network based techniques [1]. In particular, discriminative time-frequency masking techniques such as deep clustering [2], permutation-free mask inference [2, 3, 4], and their combination under the chimera framework [5, 6] have significantly advanced the state of the art. These methods only modify magnitude spectrograms, and use the mixture phase to synthesize the separated speech via a simple inverse short-time Fourier transform (iSTFT).

As magnitude processing improves and begins to approach oracle performance, i.e., the upper bound of source separation performance using the noisy phase, interest in processing phase for source separation has increased. Recent proposed approaches include avoiding phase processing by using time domain waveforms as the input/output of learned separation networks [7, 8, 9], or estimating complex (real + imaginary) time-frequency masks [10]. However, at the present time, techniques operating on magnitude spectrograms are the state of the art in speaker separation and other audio source separation tasks, e.g., speech enhancement [11] and music separation [12]. Thus, an approach that incorporates phase processing into these magnitude estimation networks would be highly valuable.

When combined with the noisy phase, separated source magnitudes may be inconsistent, i.e., no corresponding time-domain signal may exist [13, 14]. Iterative reconstruction techniques such as Griffin-Lim [15] and multiple input spectrogram inversion (MISI) [16] attempt to recover each source's clean phase by fixing its magnitude estimate and running alternating STFT and iSTFT iterations starting from the noisy phase. Applying iterative phase reconstruction as a post processing to a magnitude enhancement network often results in modest improvements in source separation performance [17, 6]. This has inspired recent techniques where phase reconstruction is computed by a generative network [18] or by a phase subnetwork trained in tandem with a magnitude subnetwork for audiovisual speech enhancement [19].

This paper presents an extension to recent work by Wang et al. [9], which learns a time-frequency mask estimation network by training through multiple *unfolded* MISI phase reconstruction iterations. Following the deep unfolding framework [20], each phase reconstruction iteration is treated as a neural network layer. In [9], the phase reconstruction layers are fixed, i.e., they have no learnable parameters implementing the STFT, iSTFT, absolute value, and angle operations necessary to perform phase reconstruction. Similarly to [7, 8] in which STFT-like encoders are implemented as convolutional layers (and iSTFT-like decoders are implemented as transposed convolution layers), we here allow the DFT/iDFT basis matrices of the STFT/iSTFT layers to be adapted based on data. The learning of these phase reconstruction layers happens in tandem with a mask inference network for estimating the separated source magnitudes, and the entire system can be trained end-to-end. This allows the magnitude estimation network to produce outputs suitable for subsequent iterative phase reconstruction. We evaluate multiple variations of this underlying approach and show improvements over the fixed unfolded phase reconstruction approach of [9] on the public wsj0-2mix corpus, leading to a new state of the art.

2. MASK INFERENCE NETWORK

An overall block diagram of our system is shown in Figure 1. In this section, following [6, 9], we describe how to train a mask inference network to estimate a magnitude representation of high enough quality that a subsequent iterative phase reconstruction network can lead to improved performance. Let $X \in \mathbb{C}^{F\times T}$ be the complex spectrogram containing the mixture of C sources $S_c \in \mathbb{C}^{F\times T}$ for $c = 1, \ldots, C$. Our goal is to estimate a real valued mask for each source $\hat{M}_c \in \mathbb{R}^{F\times T}$ by learning a nonlinear function $\Phi(\cdot)$ from an input feature space (typically log-magnitude, independently whitened in each dimension), i.e.,

$$\hat{M}_1, \ldots, \hat{M}_C = \Phi(\log|\tilde{X}|). \quad (1)$$

Here $\Phi(\cdot)$ is learned by the mask inference network in Figure 1 to minimize the truncated phase sensitive approximation (tPSA) objective [11] in a permutation-free manner [2, 3, 4]:

$$\mathcal{L}_{\text{tPSA}} = \min_{\pi\in\mathcal{P}} \sum_c \Big\| \hat{M}_{\pi(c)} \odot |X| - \mathrm{T}_0^{\gamma|X|}\left(|S_c| \odot \cos(\angle S_c - \angle X)\right) \Big\|_1, \quad (2)$$

where $\mathcal{P}$ is the set of all possible permutations over the set of sources $\{1, \ldots, C\}$, $\odot$ denotes element-wise product, $\angle S_c$ is the true phase of source spectrogram c, and $\angle X$ is the mixture phase. The ℓ_1 norm is used in (2) as opposed to the mean square error because it was found to empirically perform better in [6]. The truncation function in (2) is defined as $\mathrm{T}_a^b(x) = \min(\max(x, a), b)$, where $a = 0$ and $b = \gamma|X|$. For the sigmoid and softmax activation functions often used for mask inference [4, 6], γ is typically set to $\gamma = 1$, however, setting $\gamma > 1$ can account for phase cancellation, but requires a modified activation function for the output of the mask layer. The study in [9] noted improved performance by setting $\gamma = 2$, and defining a *convex softmax* nonlinearity where the final fully-connected layer uses a softmax to output a probability distribution $[p_0, p_1, p_2]$ over a set of potential mask values $\{0, 1, 2\}$ for each time frequency bin of each source, which is then used to compute the expected mask value $y = [p_0, p_1, p_2][0, 1, 2]^T$. Intuitively, the potential mask values $\{0, 1, 2\}$ correspond to three outcomes: (1) the source is not present ($y = 0$), (2) the source is dominant ($y = 1$), or (3) the source is involved in phase cancellation ($y = 2$).

As proposed in [6], separation performance can be further improved by adding a deep clustering branch to the mask network and using the multi-task training objective

$$\mathcal{L}_{\text{chi}^{++}_{\alpha}} = \alpha\mathcal{L}_{\text{DC}} + (1-\alpha)\mathcal{L}_{\text{tPSA}} \quad (3)$$

where $\mathcal{L}_{\text{DC}}$ is the whitened k-means objective defined in [6] and the weight α is typically set to a high value, e.g., 0.975. Although the deep clustering branch is not used during inference time, it acts as an effective regularizer during training.

Time-frequency objectives such as (2) and (3) do not account for phase inconsistencies, so [9] proposed the waveform

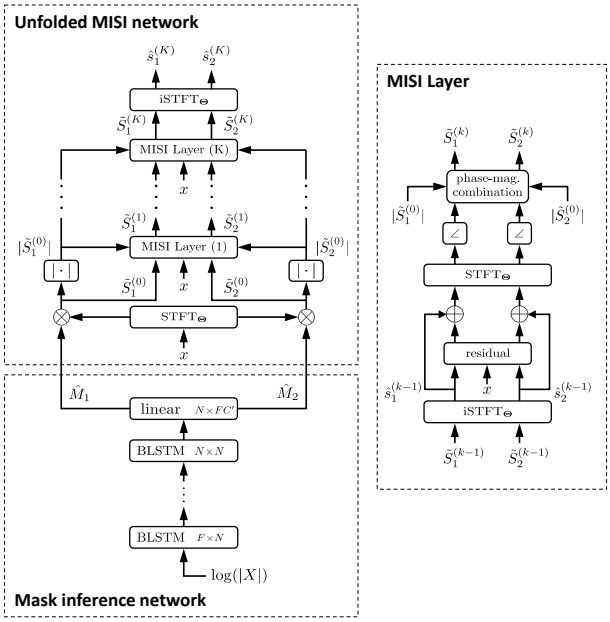

Fig. 1. Proposed speech separation system. C' is equal to $3C$ for convex softmax and to C for other nonlinearities.

approximation objective

$$\mathcal{L}_{\text{WA}} = \min_{\pi\in\mathcal{P}} \sum_c \Big\| \hat{s}_{\pi(c)} - s_c \Big\|_1, \quad (4)$$

that operates directly on the reconstructed time domain signals $\hat{s}_1, \ldots, \hat{s}_C$. The time domain signals can be obtained via a single iSTFT or by further using an unfolded MISI network as described next.

3. ITERATIVE PHASE RECONSTRUCTION NETWORK

The MISI algorithm modifies the classic Griffin-Lim approach to phase reconstruction specifically for source separation by enforcing the constraint that the sum of the separated and reconstructed sources should equal the mixture. Algorithm 1 describes an extension of the MISI algorithm considered here, where the STFT and iSTFT operations are generalized to STFT-like and iSTFT-like operations that incorporate parameters. These parameters are part of a parameter set Θ.

3.1. STFT/iSTFT Convolution Layers

For real-valued sequences such as audio signals, an N point DFT has $N/2+1$ unique complex coefficients. The DFT can be implemented using only real valued operations by stacking the real and imaginary components and defining the elements of the basis matrix $W \in \mathbb{R}^{(N+2)\times N}$ as

$$W_{i,n} = \begin{cases} w(n)\cos(2\pi ki/N), & i \in [\![1, \frac{N}{2}+1]\!] \\ -w(n)\sin(2\pi ki/N), & i \in [\![\frac{N}{2}+2, 2N+2]\!] \end{cases} \quad (5)$$

where we have incorporated the analysis window $w(n)$ into the basis matrix. By treating W as the weight matrix in a

one-dimensional convolution layer, and setting the stride parameter of this layer equal to the hop size, we can efficiently create STFT-like layers with learnable basis matrices. The inverse DFT matrix can be defined similarly to (5) by using the synthesis window and accounting for the appropriate normalization terms. We can again implement a trainable iSTFT-like layer using transposed convolutions.

3.2. Unfolded MISI

Depending on the number of unique $\mathrm{STFT}_{\Theta}^{(k)}$ and $\mathrm{iSTFT}_{\Theta}^{(k)}$ operators in Algorithm 1, we can implement the unfolded phase reconstruction network from Figure 1 in the following variations:

- **Post** [6]: A mask inference network is trained using the chimera objective (3), and MISI is only used as a post processing step, i.e., no back propagation through MISI;
- **Fixed** [9]: The mask inference network of Figure 1 is trained using the objective (4) while keeping the STFT and iSTFT layers fixed;
- **Tied** (Proposed): Together with the mask inference network, the DFT/iDFT matrices of the phase reconstruction network are also adapted after being initialized as in (5). Only one DFT-like weight matrix and one iDFT-like weight matrix are learned for the entire network and shared across the unfolded phase reconstruction layers;
- **Untied** (Proposed): The forward and inverse transform parameters are untied and independently learned for each STFT-like and iSTFT-like layer.

Even when the number of MISI iterations is zero, we can still learn to adapt the first forward transform representation applied to the mixture and the last inverse transform representation used to reconstruct the estimated waveforms. It is also interesting to point out that the phase reconstruction network structure in Figure 1 is composed of alternating convolutional layers and residual/skip connections, a general structure broadly shared by several recent audio source separation architectures such as those based on WaveNets [21] and, ResNets [12]. This is an exciting side effect of explicitly using the deep unfolding framework [20] to create network architectures. For example, even as we begin to replace portions of the network architecture such as the magnitude and phase functions with more traditional neural network nonlinearities (such as sigmoid and relu as was done in [8]), we retain some intuition as to what the network is learning.

Input: Mixture signal x in the time domain, estimated masks $\hat{M}_c$ for $c = 1, \ldots, C$, and number of iterations K

$X = \mathrm{STFT}_{\Theta}^{(0)}(x)$;

$\tilde{S}_c^{(0)} = \hat{M}_c \odot X$, for $c = 1, \ldots, C$;

for $k = 1, \ldots, K$ **do**

  $\hat{s}_c^{(k-1)} = \mathrm{iSTFT}_{\Theta}^{(k-1)}(\tilde{S}_c^{(k-1)})$, for $c = 1, \ldots, C$;

  $\delta^{(k-1)} = x - \sum_{c=1}^{C} \hat{s}_c^{(k-1)}$;

  $\tilde{S}_c^{(k)} = |\tilde{S}_c^{(0)}| e^{\mathrm{j} \angle \mathrm{STFT}_{\Theta}^{(k)}\left(\hat{s}_c^{(k-1)} + \frac{\delta^{(k-1)}}{C}\right)}$, for $c = 1, \ldots, C$;

end

return $\hat{s}_c^{(K)} = \mathrm{iSTFT}_{\Theta}^{(K)}(\tilde{S}_c^{(K)})$, for $c = 1, \ldots, C$;

Algorithm 1: Unfolded MISI. $\mathrm{STFT}_{\Theta}^{(k)}$ extracts a complex spectrogram of a signal, and $\mathrm{iSTFT}_{\Theta}^{(k)}$ reconstructs a time-domain signal from a complex spectrogram.

4. RESULTS

We use the public wsj0-2mix dataset [2], which contains 30 h of two-speaker mixtures for training, 10 h for validation, and 5 h for testing with a sampling rate of 8 kHz and signal to noise ratios (SNRs) between 0 and 5 dB. We use the same spectral analysis parameters for both the log-magnitude features input to the mask network and for the initial STFT/iSTFT layers in the phase reconstruction network: window size of 32 ms (256 samples) with a stride of 64 samples and square root Hann analysis/synthesis windows designed for perfect reconstruction after overlap-add.

The mask inference network uses the architecture from [6, 9] with four 600 unit BLSTM layers and dropout of 0.3 applied to the first three layers. During training, input sequences are limited to 400 frames (approximately 3.2 s), and optimization is performed using the ADAM algorithm [22]. We train the network for 100 epochs, and if the loss function on the validation set does not decrease for five consecutive epochs the learning rate is decayed by 0.5. Furthermore, we achieved a significant performance boost by following the curriculum learning strategy proposed in [9] for each of the unfolded MISI approaches described in Section 3.2:

1. We first train only the mask network using the chimera objective (3) and an initial ADAM step size of $\alpha = 10^{-3}$.
2. Next, we discard the deep clustering portion of the chimera network and train using the waveform approximation objective (4) by adding forward and inverse time-frequency representation layers (but no MISI layers yet) with an initial ADAM step size of $\alpha = 10^{-4}$.
3. Starting from the network with $k - 1$ MISI layers, we train a network with k MISI layers and an initial ADAM step size of $\alpha = 10^{-4}$. Repeat up to $K = 5$, where performance begins to saturate.

We report scale-invariant SDR (SI-SDR) on the wsj0-2mix test set, as opposed to the version computed using the bss_eval software package for reasons discussed in [23]. Figure 2 compares network architectures for different numbers of MISI layers for both the sigmoid and convex softmax activations. As first reported in [9], training the mask inference network through fixed MISI layers provides a significant boost in performance compared to using phase reconstruction only for post processing. Furthermore, learning the time frequency representations in the phase network provides a boost in performance for all numbers of iterations/layers with untied performing better than tied. Finally, we note that for the

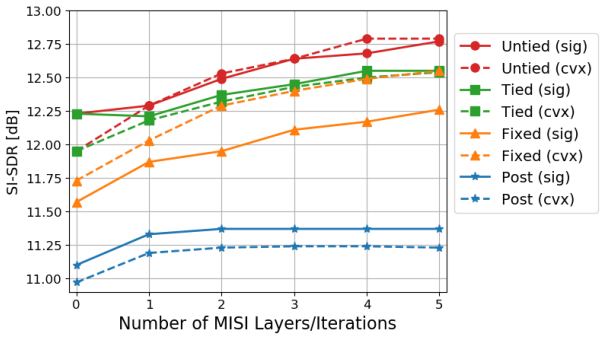

Fig. 2. Performance of different MISI reconstruction configurations. Solid lines correspond to sigmoid activation and dashed lines to convex softmax at the mask network output.

Table 1. SI-SDR (dB) comparison with other recent systems on the wsj0-2mix test set.

Approach	MISI Iterations	SI-SDR [dB]
TasNet-BLSTM [8]	-	11.1
Chimera++ [6]	0	11.2
	5	11.5
Fixed (convex softmax) [9]	0	11.8
	5	12.6
Tied-MISI (sigmoid)	0	**12.2**
	5	12.6
Untied-MISI (sigmoid)	5	**12.8**
Oracle Algorithms		
Magnitude Ratio Mask	0	12.7
	5	13.7
Ideal Binary Mask	0	13.5
	5	13.4
Phase Sensitive Mask	0	16.4
	5	18.3
Ideal Amplitude Mask	0	12.8
	5	26.6

fixed case, as first shown in [9], using the convex softmax activation provides a performance boost compared to sigmoid, as allowing the mask network to predict values greater than 1 allows for time-frequency magnitude values which are closer to those obtained from actual time-domain signals. However, for the conditions with learned time-frequency representations (tied/untied), the difference in performance between activation functions almost disappears. We hypothesize this is because the network learns time-frequency representations that are appropriately scaled for the mask network output, something not possible with fixed STFT/iSTFT representations. The biggest SI-SDR gain for the tied/untied cases also happens when no phase reconstruction is performed (i.e., 0 MISI iterations), achieving over 12.2 dB SI-SDR by learning just one forward transform and one inverse transform. Two MISI iterations are required to match such performance when using the fixed STFT/iSTFT representations.

Table 1 compares the performance of the iterative phase reconstruction algorithms proposed in this paper, with some recent competitive approaches, and oracle approaches both with and without MISI. Our results improve upon the best known previous results from [9] by 0.4 dB for 0 MISI iterations and 0.2 dB for 5 MISI iterations, approaching the performance of some oracle masks.

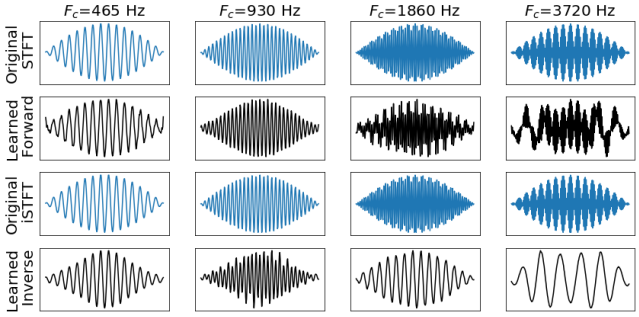

Fig. 3. Adapted forward and inverse filters for the Tied-MISI-5 condition. Labeled by original center frequency (F_c).

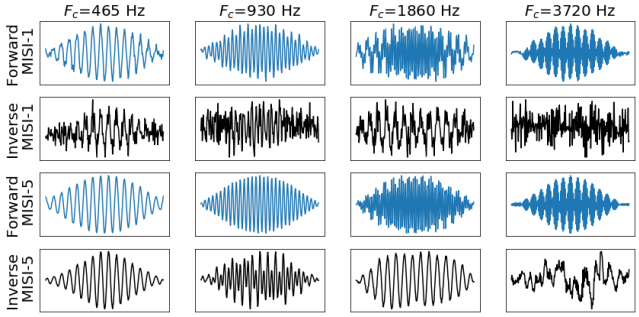

Fig. 4. Example filters for MISI layers from the Untied-MISI-5 condition. Labeled by original center frequency (F_c).

Figure 3 compares some learned time frequency representations to their corresponding original STFT versions. Some changes in the shape of the analysis/synthesis windows can be observed. It also illustrates that some original high frequency representations (e.g., the far right column in Figure 3) learn to focus on low frequency signal components. This is consistent with the results of [7, 8] where many learned time-frequency filters focus on the low frequency range. Figure 4 displays learned filters for two MISI layers in an untied network configuration. While the filters are significantly different from those in the STFT/iSTFT, they are more difficult to interpret.

5. CONCLUSION

We proposed an approach to speech separation based on deep unfolding of the MISI iterative phase reconstruction algorithm. By learning time-frequency representations from data, we obtain quantitative improvements in separation performance. Future work includes improving the initial phase estimates over the noisy phase and exploring the application of the proposed techniques to other domains such as speech enhancement, music, or environmental sound separation.

6. REFERENCES

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," in *arXiv preprint arXiv:1708.07524*, 2017.

[2] J. R. Hershey, Z. Chen, and J. Le Roux, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[3] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. Interspeech*, 2016.

[4] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multi-talker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2017.

[5] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[6] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[7] S. Venkataramani and P. Smaragdis, "End-to-end source separation with adaptive front-ends," in *arXiv preprint arXiv:1705.02514*, 2017.

[8] Y. Luo and N. Mesgarani, "TasNet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[9] Z.-Q. Wang, J. Le Roux, D. Wang, and J. R. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," in *arXiv preprint arXiv:1804.10204*, 2018.

[10] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[11] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[12] N. Takahashi, N. Goswami, and Y. Mitsufuji, "MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *arXiv preprint arXiv:1805.02410*, 2018.

[13] J. Le Roux, N. Ono, and S. Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction," in *Proc. ISCA Workshop on Statistical and Perceptual Audition (SAPA)*, Sep. 2008.

[14] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, 2015.

[15] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, 1984.

[16] D. Gunawan and D. Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures," in *IEEE Signal Processing Letters*, 2010.

[17] Y. Zhao, Z.-Q. Wang, and D. Wang, "A two-stage algorithm for noisy and reverberant speech enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[18] K. Oyamada, H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, and H. Ando, "Generative adversarial network-based approach to signal reconstruction from magnitude spectrograms," in *arXiv preprint arXiv:1804.02181*, 2018.

[19] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *arXiv preprint arXiv:1804.04121*, 2018.

[20] J. R. Hershey, J. Le Roux, and F. Weninger, "Deep unfolding: Model-based inspiration of novel deep architectures," in *arXiv preprint arXiv:1409.2574*, 2014.

[21] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[22] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *arXiv preprint arXiv:1412.6980*, 2014.

[23] J. Le Roux, J. R. Hershey, A. Liutkus, F. Stöter, S. T. Wisdom, and H. Erdogan, "SDR – half-baked or well done?" Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA, Tech. Rep., 2018.