

On The Compensation Between Magnitude and Phase in Speech Separation

Wang, Zhong-Qiu; Wichern, Gordon; Le Roux, Jonathan

TR2021-137 November 13, 2021

Abstract

Many recent studies optimize loss functions defined solely in the time or complex domain, without including a loss on magnitude. Although such loss functions typically produce better scores if the evaluation metrics are objective time-domain metrics, they however produce worse scores on speech quality and intelligibility metrics and usually lead to worse speech recognition performance, compared with including a loss on magnitude. While this phenomenon has been experimentally observed by many studies, it is often not accurately explained and there lacks a thorough understanding on its fundamental cause. This paper provides a novel view from the perspective of the implicit compensation between estimated magnitude and phase.

IEEE Signal Processing Letters 2021

On The Compensation Between Magnitude and Phase in Speech Separation

Zhong-Qiu Wang, Gordon Wichern, and Jonathan Le Roux

Abstract—Deep neural network (DNN) based end-to-end optimization in the complex time-frequency (T-F) domain or time domain has shown considerable potential in monaural speech separation. Many recent studies optimize loss functions defined solely in the time or complex domain, without including a loss on magnitude. Although such loss functions typically produce better scores if the evaluation metrics are objective time-domain metrics, they however produce worse scores on speech quality and intelligibility metrics and usually lead to worse speech recognition performance, compared with including a loss on magnitude. While this phenomenon has been experimentally observed by many studies, it is often not accurately explained and there lacks a thorough understanding on its fundamental cause. This paper provides a novel view from the perspective of the implicit compensation between estimated magnitude and phase. Analytical results based on monaural speech separation and robust automatic speech recognition (ASR) tasks in noisy-reverberant conditions support the validity of our view.

Index Terms—End-to-end optimization, speech enhancement, speaker separation, phase estimation, deep learning.

I. INTRODUCTION

DEEP learning has elevated the performance of speech separation in the past decade. Early DNN based approaches operated in the magnitude domain, where DNNs were trained based on magnitude features to predict target magnitudes directly or via estimating T-F masks, and the mixture phase was used for signal re-synthesis [1]. Popular T-F masks include the ideal binary/ratio mask [1], ideal amplitude mask (IAM) [2], and phase-sensitive mask (PSM) [3]. To obtain better phase for re-synthesis, subsequent studies estimated it by using DNN-estimated magnitudes to drive iterative phase reconstruction (IPR) algorithms [4]–[7]. Building upon deep learning based end-to-end optimization, recent studies implicitly estimate phase by predicting the real and imaginary (RI) components of target speech from the mixture RI components [8]–[14], or predicting target waveforms from the mixture waveform [15]–[19]. This end-to-end approach has shown large improvements over magnitude-domain approaches. Many studies along this line define their loss functions only in the time domain to optimize for example SI-SDR [17], or in the complex domain to minimize an L_p -norm distance between predicted and target RI components [10], [20]. Some studies add a loss on the magnitude of the predicted RI components or waveforms [5], [9], [11]–[14], [18], [19], [21]–[26]. This loss is reported to produce clear improvements in speech quality, intelligibility, and ASR scores with slightly worse

SI-SDR results [11]–[14]. Although this phenomenon has been experimentally observed in many studies, its fundamental cause is often misinterpreted or not thoroughly analyzed.

Our study provides a novel explanation to this observation. Our insight is that, since phase is always difficult to estimate accurately, if the loss is defined solely in the complex or time domain, the magnitude of the estimated speech will tend to compensate for an inaccurate phase estimate, leading to a less accurate magnitude compared with the one obtained by including a magnitude loss, or, alternatively, training a magnitude-domain model for direct magnitude estimation. We shall point out that many techniques in this paper have been proposed before. Our contribution is a novel view on why they work well and under what conditions they would work less well. Such a view can facilitate understanding and guide algorithmic design in speech separation.

II. END-TO-END SPEECH SEPARATION

We review two popular end-to-end approaches for speech separation. Given a monaural mixture, the physical model relating the mixture y , target s , and non-target signals v can be formulated in the time domain as $y[n] = s[n] + v[n]$, where n indexes discrete time. In the short-time Fourier transform (STFT) domain, the physical model is formulated as

$$Y(t, f) = S(t, f) + V(t, f), \quad (1)$$

where Y , S , and V respectively denote the STFT spectra of y , s , and v , and t and f index time and frequency.

A. Complex-domain Separation

Complex spectral mapping [8]–[12] predicts target RI components from the mixture RI components, simultaneously modeling magnitude and phase. A typical loss is

$$\mathcal{L}_{\text{RI}} = \|\hat{R} - \text{Real}(S)\|_1 + \|\hat{I} - \text{Imag}(S)\|_1, \quad (2)$$

where \hat{R} and \hat{I} are the predicted RI components, $\text{Real}(\cdot)$ and $\text{Imag}(\cdot)$ extract RI components, and $\|\cdot\|_1$ computes the L_1 norm. The separation result is $\hat{S} = \hat{R} + j\hat{I}$, where j is the imaginary unit. Inverse STFT (iSTFT) is then applied for signal re-synthesis. In [11]–[14], it is suggested that $\text{STFT}(\text{iSTFT}(\hat{S}))$ is very close to \hat{S} , meaning that the magnitude and phase of the estimated complex spectrogram produced by complex spectral mapping are almost consistent with each other.

A magnitude loss can be added [11]–[14]:

$$\mathcal{L}_{\text{RI+Mag}} = \mathcal{L}_{\text{RI}} + \||\hat{R} + j\hat{I}| - |S|\|_1, \quad (3)$$

where $|\cdot|$ extracts magnitude. One can also train through iSTFT and define the loss in the time domain [5], [6], [20]:

$$\mathcal{L}_{\text{RI-iSTFT}} = \|\text{iSTFT}(\hat{S}) - s\|_1. \quad (4)$$

Manuscript received on Jun. 17, 2021; revised on Sep. 14, 2021.

Z.-Q. Wang, G. Wichern, and J. Le Roux are with Mitsubishi Electric Research Laboratories, Cambridge, MA 02139, USA (e-mail: wang.zhongqiu41@gmail.com, {wichern,leroux}@merl.com).

Note that $\text{iSTFT}(\hat{S})$ is the final signal listened to by end users and used for metric computation. To improve its magnitude, a magnitude loss can be included [21]–[24]:

$$\mathcal{L}_{\text{RI-iSTFT+Mag}} = \mathcal{L}_{\text{RI-iSTFT}} + \left| \left| \text{STFT}(\text{iSTFT}(\hat{S})) \right| - |S| \right|_1. \quad (5)$$

An alternative computes the magnitude loss before iSTFT:

$$\mathcal{L}_{\text{Mag+RI-iSTFT}} = \left| \left| \hat{R} + j\hat{I} \right| - |S| \right|_1 + \mathcal{L}_{\text{RI-iSTFT}}. \quad (6)$$

Some studies apply a power compression on the predicted magnitude before computing the loss [22], [27], use a weight between the time- and magnitude-domain loss, or define the loss on magnitude features [28], [29] or on multi-resolution magnitudes [19]. They are out of the scope of this paper.

B. Time-domain Separation

Time-domain approaches predict the target waveform directly from the mixture waveform [15]–[19], [30]. They implicitly model magnitude and phase through end-to-end optimization. The loss is typically defined solely in the time domain, in the form of mean absolute/square error (or their log-compressed versions and SI-SDR [31]) as

$$\mathcal{L}_{\text{Wav}} = \|\hat{s} - s\|_1, \quad (7)$$

where \hat{s} denotes the predicted waveform. Later studies [18], [19] incorporate a magnitude-domain loss

$$\mathcal{L}_{\text{Wav+Mag}} = \mathcal{L}_{\text{Wav}} + \left| \left| \text{STFT}(\hat{s}) \right| - |S| \right|_1. \quad (8)$$

III. COMPENSATION BETWEEN MAGNITUDE AND PHASE

This section describes the compensation problem, and loss functions that lead to better magnitude or phase estimation.

A. The Compensation Problem

Training a model using (2) or (7) essentially tries to find an estimated speech $\hat{S}(t, f)$ that can approximate clean speech $S(t, f)$ at each T-F unit. See Fig. 1(a) for an illustration. Because phase is generally difficult to estimate, $\angle \hat{S}(t, f)$ is typically very different from $\angle S(t, f)$, especially in T-F units with low signal-to-noise ratio (SNR). In such cases, the closest approximation of $S(t, f)$ is the projection of $S(t, f)$ onto the direction determined by $\angle \hat{S}(t, f)$. This approximation, however, is incapable of recovering the clean magnitude, and the difference between the two becomes larger as $\angle \hat{S}(t, f)$ gets away from $\angle S(t, f)$. If $\angle \hat{S}(t, f)$ is more than $\pi/2$ away from $\angle S(t, f)$, the best approximation would lead to a zero magnitude, as illustrated in Fig. 1(b).

By including a loss on magnitude in (3) or (8), the DNN prediction is encouraged to find a balance between complex- and magnitude-domain approximations. This balance explains why adding a loss on magnitude leads to better perceptual evaluation of speech quality (PESQ) [32] and extended short-time objective intelligibility (eSTOI) [33] scores, because they favor an estimated target time-domain signal that has an accurate magnitude, rather than a compensated, less accurate one. We point out that eSTOI and STOI [33] only look at the magnitude envelope of the predicted signal, and PESQ [32] first time-aligns the predicted signal with the reference segment-wisely, forgiving any segmental time delays, and then looks at their short-time Bark-scale power spectra. This balance also suggests that the degradation in time-domain metrics,

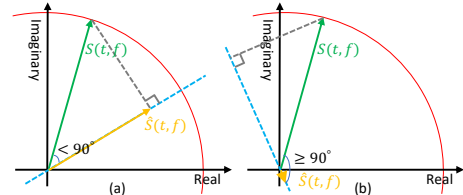


Fig. 1: Complex-plane illustration of magnitude-phase compensation.

such as SI-SDR [31], is due to the magnitude loss pushing $\hat{S}(t, f)$ away from the closest approximation of $S(t, f)$ along the direction of $\angle \hat{S}(t, f)$. Indeed, SI-SDR [31] measures the quality of time-domain sample-level predictions, and hence favors estimated speech with a magnitude spectrogram that compensates for its inaccurate phase spectrogram.

This compensation view is motivated by the PSM (defined as $|S|/|Y|\cos(\angle S - \angle Y)$) [3], proposed to find the magnitude closest to $S(t, f)$ along the direction of $\angle Y(t, f)$ for magnitude-domain speech enhancement. The PSM explicitly considers the compensation so that the SNR of the target estimate can be maximized when $\angle Y$ is used for signal re-synthesis. Our key contribution is to show that this compensation problem, often neglected, implicitly exists in many end-to-end approaches, which usually improve upon the mixture phase but still cannot perfectly reconstruct the clean phase. Our study is also the first to comprehensively analyze the consequences of this compensation problem on multiple popular evaluation metrics, and emphasizes the effectiveness of including a magnitude loss. Such an analysis does not exist in [3] and other previous works.

B. Magnitude Spectrogram Approximation

This compensation view suggests that, in cases where we only need a good estimated magnitude and do not have to estimate or leverage phase, it may be better not modelling magnitude and phase simultaneously. One such scenario is robust ASR based on monaural speech enhancement [34], where the recognition model typically only considers magnitude features. In such cases, direct target magnitude spectrogram approximation (MSA) [3] would likely lead to better magnitude estimation and produce better performance, as the resulting magnitude is not a compensated one. The typical loss function is

$$\mathcal{L}_{\text{MSA}} = \|\hat{M} - |S|\|_1, \quad (9)$$

where \hat{M} denotes estimated magnitude. Eq. (9) can be considered as a teacher-forcing technique [7], [35], assuming that the estimated speech has the same phase as target speech, i.e.,

$$\mathcal{L}_{\text{MSA}} = \|\hat{M}e^{j\angle S} - |S|e^{j\angle S}\|_1. \quad (10)$$

This way, the implicit compensation between magnitude and phase is avoided, because the best approximation of $S(t, f)$ along the direction of $\angle S(t, f)$ is $|S(t, f)|$. One potential issue with MSA is that when signal re-synthesis is needed, the mixture phase is typically used together with the estimated magnitude. This usually leads to sub-optimal SI-SDR, as the best approximation of $S(t, f)$ along $\angle Y(t, f)$ should have a compensated magnitude. In addition, due to the phase inconsistency issue [6], [22], [36], [37], the magnitude of the re-synthesized signal $|\text{STFT}(\text{iSTFT}(\hat{M}e^{j\angle Y}))|$ would not be as good as \hat{M} . This is likely the reason why it is observed

in [34] that extracting ASR features directly from estimated magnitudes, rather than from re-synthesized signals by using the mixture phase, produces better ASR results.

An alternative formulation of MSA uses the RI model of (5), but uses a weight of zero on the time-domain loss [38]:

$$\mathcal{L}_{(\text{RI-iSTFT})\times 0+\text{Mag}} = \left\| \left| \text{STFT}(\text{iSTFT}(\hat{S})) \right| - |S| \right\|_1. \quad (11)$$

We can also do this for the time-domain model in (8) [39]:

$$\mathcal{L}_{\text{Wav}\times 0+\text{Mag}} = \left\| \left| \text{STFT}(\hat{s}) \right| - |S| \right\|_1. \quad (12)$$

Essentially, the models are trained to produce a time-domain signal with a good magnitude. This signal is likely to have a phase not strictly aligned with the clean phase [38], hence would have worse SI-SDR, but would still have reasonable PESQ, STOI, and WER scores, as its magnitude is good. Such a signal could still be good for human listening [38], as the human auditory system is not sensitive to slight signal shift.

C. Phase Spectrogram Approximation

The previous sections assume that phase cannot be estimated perfectly. Magnitude also cannot be estimated perfectly. In cases where only phase estimates are needed, one could avoid the influence of inaccurate magnitude estimates by supplying oracle magnitudes and define a phase loss such as

$$\begin{aligned} \mathcal{L}_{\text{Phase}} = & \left\| \text{Real}(|S|e^{j\angle(\hat{R}+j\hat{I})}) - \text{Real}(S) \right\|_1 \\ & + \left\| \text{Imag}(|S|e^{j\angle(\hat{R}+j\hat{I})}) - \text{Imag}(S) \right\|_1. \end{aligned} \quad (13)$$

IV. EXPERIMENTAL SETUP

We validate our ideas on monaural speech enhancement, speaker separation, and ASR tasks. This section describes the datasets, system configurations, and evaluation metrics.

For speech enhancement, we use the WHAMR! corpus [40], designed for noisy-reverberant 2-speaker separation. We tailor the task to noisy-reverberant speech enhancement by removing the second speaker in each mixture. We use the *min* and 16 kHz version of the corpus, and the first channel for training and evaluation. We use the target direct sound as the reference for training and metric computation, and perform joint dereverberation and denoising.

For speaker separation, we use the SMS-WJSJ dataset [41], sampled at 8 kHz. It contains simulated reverberant 2-speaker mixtures. The first channel is used for training and evaluation. We use the direct sound as the training target and perform joint dereverberation, denoising, and separation. For ASR, we use the default Kaldi-based backend trained on single-speaker noisy-reverberant speech data plus its corresponding direct sound data. The window length (WL) and hop length (HL) for ASR feature extraction are 25 ms and 10 ms.

For STFT, our separation models use regular 32/8 ms WL/HL for WHAMR!, but 25/10 ms WL/HL for SMS-WJSJ to align with the ASR backend. For complex spectral mapping and MSA, we use the DenseUNet-TCN architecture [13]. For MSA, we use $|Y|$ as features to directly predict $|S|$. The same architecture is also used for phase-sensitive spectrogram approximation (PSA) [3]. The feature is $|Y|$ and the loss is

$$\mathcal{L}_{\text{PSA}} = \left\| \hat{M} - |S| \text{T}_0^1(\cos(\angle S - \angle Y)) \right\|_1, \quad (14)$$

where $\text{T}_0^1(\cdot)$ truncates the values to the range $[0, 1]$. For time-domain approaches, we employ Conv-TasNet [17], where

the WL/HL are 5/2.5 ms. For speaker separation, the loss functions follow (2)-(14), but we additionally use permutation free (a.k.a. invariant) training [42]–[44].

Our evaluation metrics include SI-SDR [31], eSTOI [33], [45], PESQ [32], [46], and word error rates (WER). We point out that SI-SDR is very sensitive to signal shift, while the other measures are not, as magnitude is not sensitive to slight signal shifts. Additionally, we use magnitude SNR (mSNR) [43] to measure the quality of estimated magnitude,

$$\text{mSNR} = 10 \log_{10} \frac{\sum_{t,f} |S(t,f)|^2}{\sum_{t,f} \left| |S(t,f)| - |\hat{S}(t,f)| \right|^2}, \quad (15)$$

and phase SNR (pSNR) to measure that of estimated phase,

$$\text{pSNR} = 10 \log_{10} \frac{\sum_{t,f} |S(t,f)|^2}{\sum_{t,f} |S(t,f) - |S(t,f)| e^{j\angle\hat{S}(t,f)}|^2}, \quad (16)$$

where oracle magnitude is supplied for metric computation.

V. EVALUATION RESULTS

Table I reports the results on WHAMR!'s enhancement task. The magnitude loss in RI-iSTFT+Mag, Mag+RI-iSTFT, and Wav+Mag is computed using 32/8 ms WL/HL. Comparing RI and RI+Mag, RI-iSTFT and RI-iSTFT+Mag, RI-iSTFT and Mag+RI-iSTFT, and Wav and Wav+Mag, we observe clear improvement in PESQ and eSTOI, better mSNR, and slight degradation in SI-SDR, when a magnitude-domain loss is included and signal re-synthesis is performed. This observation indicates that the estimated magnitude would implicitly compensate for the inaccurate phase estimate when the loss is defined only in the complex- or time-domain, and adding a loss on magnitude would alleviate such compensation. With signal re-synthesis, Mag+RI-iSTFT shows worse performance than RI-iSTFT+Mag. MSA with re-synthesis shows worse SI-SDR than complex-domain models, likely because MSA uses the mixture phase for re-synthesis. However, MSA without re-synthesis obtains the best mSNR at 10.9 dB. This suggests that to obtain a good magnitude, one can consider avoiding modelling magnitude and phase at the same time. Similarly, the Phase model trained with (13) obtains the best pSNR at 12.4 dB. Adding a magnitude loss degrades pSNR (for example 10.8 vs. 10.4 dB for RI vs. RI+Mag). This indicates the improvement over the mixture phase comes from using a strong DNN model for direct complex- or time-domain prediction, rather than from adding the magnitude loss, a key point that is not discussed in [18]. By only optimizing the magnitude loss, (RI-iSTFT) \times 0+Mag shows better PESQ, eSTOI and mSNR than RI-iSTFT+Mag, and worse but still good SI-SDR and pSNR scores, even though no time-domain loss is included. This indicates the complex-domain model implicitly figures out a reasonably-good phase that can produce a good magnitude. (RI-iSTFT) \times 0+Mag produces worse mSNR than MSA (12.9 vs. 13.1 dB), possibly because the magnitude has to be extracted from a time-domain signal, which may limit the model's capability at magnitude estimation. Wav \times 0+Mag obtains good mSNR, PESQ, and eSTOI, but much worse SI-SDR and pSNR scores. This degradation might be due to the small WL in Conv-TasNet. The Wav \times 0+Mag results provide a strong experimental evidence showing that PESQ and eSTOI

TABLE I

SI-SDR (dB), PESQ, eSTOI (%), mSNR (dB), and pSNR (dB) on WHAMR! (Enh.)

Approaches	Eq.	WL/HL (ms)	Signal re-syn?	SI-SDR	PESQ	eSTOI	mSNR	pSNR
Unprocessed	-	32/8	-	-2.7	1.53	45.1	-1.6	-2.8
MSA	(9)	32/8	yes	4.4	2.72	78.5	11.3	6.3
MSA	(9)	32/8	no	-	-	-	13.1	-
RI	(2)	32/8	yes	9.1	2.49	80.3	12.7	10.8
RI	(2)	32/8	no	-	-	-	12.6	10.8
RI+Mag	(3)	32/8	yes	8.6	2.92	81.9	12.8	10.4
RI+Mag	(3)	32/8	no	-	-	-	12.8	10.2
RI-iSTFT	(4)	32/8	yes	8.8	2.46	79.0	12.1	10.8
Mag+RI-iSTFT	(6)	32/8	yes	8.4	2.55	80.9	12.5	10.1
Mag+RI-iSTFT	(6)	32/8	no	-	-	-	12.9	9.4
RI-iSTFT+Mag	(5)	32/8	yes	8.6	2.86	81.7	12.6	10.5
(RI-iSTFT) \times 0+Mag	(11)	32/8	yes	7.3	2.91	82.7	12.9	9.1
Phase	(13)	32/8	no	-	-	-	-	12.4
Wav	(7)	5/2.5	yes	7.7	2.20	78.0	11.2	10.1
Wav+Mag	(8)	5/2.5	yes	7.5	2.58	80.1	11.4	9.9
Wav \times 0+Mag	(12)	5/2.5	yes	-9.1	2.67	80.6	11.4	-3.6
PSA	(14)	32/8	yes	5.6	2.36	76.3	10.2	7.2
PSA	(14)	32/8	no	-	-	-	9.9	-
PSM	-	32/8	yes	8.4	3.82	91.1	13.4	9.6
IAM	-	32/8	yes	5.4	3.47	89.7	14.9	6.5

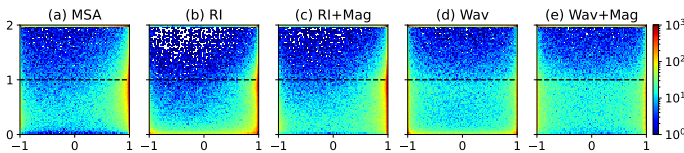


Fig. 2: 2D histograms for various models. The x-axis is the phase difference $\cos(\angle S(t, f) - \angle Y(t, f))$, which gets smaller as $\angle Y(t, f)$ gets away from $\angle S(t, f)$, and the y-axis is the magnitude ratio $|\hat{M}(t, f)|/|S(t, f)|$ (truncated to the range $[0, 2]$). For RI and RI+Mag, $\hat{M} = |\hat{R} + j\hat{I}|$ (i.e., no re-synthesis). For Wav and Wav+Mag, \hat{M} is extracted from the re-synthesized signal. In this example, the mixture mSNR is -2.3 dB, and the output mSNRs are 12.4, 11.0, 12.1, 10.7 and 11.3 dB for the MSA, RI, RI+Mag, WA and WA+Mag models, respectively. Before plotting, we throw away T-F units whose energy in $|S|$ is more than 60 dB weaker than the highest-energy T-F unit.

scores largely depend on the magnitude of the estimated signal.

To illustrate the compensation of estimated magnitudes when mixture phase $\angle Y(t, f)$ is different from clean phase $\angle S(t, f)$, we provide two-dimensional (2D) histograms of phase difference vs. magnitude ratios in Fig. 2, based on a test mixture of WHAMR!. See the caption for the definitions of the axes and other details. Comparing MSA and RI, we observe that the magnitude ratios by MSA are overall closer to the dashed line, which denotes the case of perfect magnitude estimation. This indicates that MSA produces better magnitude estimation. In addition, for RI, many estimated magnitudes are much smaller than the clean magnitudes (i.e., near the bottom of the plot), while the magnitude ratios in MSA are much less dense near the bottom. This suggests that when $\angle Y(t, f)$ is different from $\angle S(t, f)$ and hence $\angle \hat{S}(t, f)$ likely differs from $\angle S(t, f)$, RI compresses (or sacrifices) the estimated magnitude to better approximate $S(t, f)$, while MSA does not because MSA assumes that $\angle \hat{S}(t, f) = \angle S(t, f)$. Adding a magnitude loss to RI, RI+Mag improves the magnitude estimation, but still produces less accurate magnitude than MSA, as the magnitude ratios are more spread out than MSA when $\angle Y(t, f)$ is different from $\angle S(t, f)$. Similar trends are observed in the WA and WA+Mag plots. These histograms indicate the validity of the magnitude-phase compensation phenomenon, and that direct MSA produces better magnitudes.

Table II reports the results on SMS-WSJ. In this setup, the magnitude loss in RI-iSTFT+Mag and Wav+Mag is computed based on 25/10 ms WL/HL. Adding a magnitude loss leads to

TABLE II

SI-SDR (dB), PESQ, eSTOI (%), mSNR (dB), and WER (%) on SMS-WSJ.

Approaches	Eq.	WL/HL (ms)	Signal re-syn?	SI-SDR	PESQ	eSTOI	mSNR	WER
Unprocessed	-	25/10	-	-5.5	1.50	44.1	-4.4	79.43
MSA	(9)	25/10	yes	0.3	2.20	69.8	8.4	32.84
MSA	(9)	25/10	no	-	-	-	9.6	33.87
RI	(2)	25/10	yes	4.6	1.97	70.1	8.7	42.26
RI	(2)	25/10	no	-	-	-	8.6	40.98
RI+Mag	(3)	25/10	yes	3.2	2.21	70.1	8.6	35.97
RI+Mag	(3)	25/10	no	-	-	-	8.6	36.05
RI-iSTFT	(4)	25/10	yes	4.5	1.87	67.8	8.0	44.58
Mag+RI-iSTFT	(6)	25/10	yes	3.2	1.99	69.4	8.3	39.09
Mag+RI-iSTFT	(6)	25/10	no	-	-	-	9.5	33.98
RI-iSTFT+Mag	(5)	25/10	yes	3.5	2.24	71.3	9.0	33.66
(RI-iSTFT) \times 0+Mag	(11)	25/10	yes	2.0	2.29	72.1	9.0	32.38
Wav	(7)	5/2.5	yes	4.2	1.79	66.3	8.3	47.50
Wav+Mag	(8)	5/2.5	yes	3.4	2.07	68.9	8.3	39.22
Wav \times 0+Mag	(12)	5/2.5	yes	-4.2	2.11	69.6	8.2	37.91
PSA	(14)	25/10	yes	1.4	1.82	65.6	6.7	44.75
PSA	(14)	25/10	no	-	-	-	6.5	48.15
PSM	-	25/10	yes	5.8	3.64	89.7	10.5	5.84
IAM	-	25/10	yes	1.5	3.37	91.1	13.1	5.71
IAM	-	25/10	no	-	-	-	∞	5.46

clearly better WER for RI, RI-iSTFT, and Wav. MSA exhibits strong WER with no re-synthesis at 33.87%, and competitive WER with re-synthesis at 32.84%, possibly because $\angle Y$ used for re-synthesis does not dramatically degrade the magnitude. The trend on the other metrics is similar to that in Table I.

In both tables, we report oracle real-valued T-F masking scores of IAM ($|S|/|Y|$) [2] and PSM [3], both using $\angle Y$ for re-synthesis. IAM makes an aggressive step $|S(t, f)|$ along $\angle Y(t, f)$, while PSM makes a less aggressive step $|S(t, f)|\cos(\angle S(t, f) - \angle Y(t, f))$, considering that $\angle Y(t, f)$ is different from $\angle S(t, f)$. PSM shows better SI-SDR, as the masked mixture spectrum is closer to the clean one. IAM with re-synthesis shows better mSNR and WER than PSM, likely because the magnitude used for re-synthesis is oracle and the re-synthesized signal still has a reasonable magnitude, even though $\angle Y$ is used for re-synthesis. Comparing MSA and PSA, we observe that PSA obtains better SI-SDR while worse scores on the other metrics that favor a good magnitude. The PSA model shows worse performance than end-to-end models.

VI. CONCLUSION

We have provided a novel view on the implicit compensation between estimated magnitude and phase in DNN based speech separation. This view provides a fundamental understanding of the performance differences between including and not including a magnitude-domain loss for training. This understanding can benefit the design of many source separation algorithms and has broad implications. For example, when using time-domain models as benchmarks and PESQ, STOI, or WER as the evaluation metrics, a study should consider training its time-domain models with a magnitude-domain loss in combination with a time-domain loss. Another such example is monaural sound event detection that performs separation before detection, where many studies train the separator to optimize SI-SDR [47], [48]. One could train the separator by including a magnitude loss, as the detector is usually trained on energy features, or train the separator jointly with the detector to only optimize the detection loss, similarly to the joint frontend and backend training [49], [50] in robust ASR.

REFERENCES

- [1] D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, pp. 1702–1726, 2018.
- [2] Y. Wang, A. Narayanan, and D. Wang, "On Training Targets for Supervised Speech Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, pp. 1849–1858, 2014.
- [3] 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. ICASSP*, 2015, pp. 708–712.
- [4] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning Spectral Mapping for Speech Dereverberation and Denoising," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 982–992, 2015.
- [5] Z.-Q. Wang, J. Le Roux, D. Wang, and J. R. Hershey, "End-to-End Speech Separation with Unfolded Iterative Phase Reconstruction," in *Proc. Interspeech*, 2018, pp. 2708–2712.
- [6] Z.-Q. Wang, K. Tan, and D. Wang, "Deep Learning Based Phase Reconstruction for Speaker Separation: A Trigonometric Perspective," in *Proc. ICASSP*, 2019, pp. 71–75.
- [7] Y. Zhao, Z.-Q. Wang, and D. Wang, "Two-Stage Deep Learning for Noisy-Reverberant Speech Enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 53–62, 2019.
- [8] D. S. Williamson, Y. Wang, and D. Wang, "Complex Ratio Masking for Monaural Speech Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, pp. 483–492, 2016.
- [9] S.-W. Fu, T.-Y. Hu, Y. Tsao, and X. Lu, "Complex Spectrogram Enhancement By Convolutional Neural Network with Multi-Metrics Learning," in *Proc. MLSP*, 2017, pp. 1–6.
- [10] K. Tan and D. Wang, "Learning Complex Spectral Mapping With Gated Convolutional Recurrent Networks for Monaural Speech Enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 380–390, 2020.
- [11] Z.-Q. Wang, P. Wang, and D. Wang, "Complex Spectral Mapping for Single-and Multi-Channel Speech Enhancement and Robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1778–1787, 2020.
- [12] Z.-Q. Wang and D. Wang, "Deep Learning Based Target Cancellation for Speech Dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 941–950, 2020.
- [13] —, "Multi-Microphone Complex Spectral Mapping for Speech Dereverberation," in *Proc. ICASSP*, 2020, pp. 486–490.
- [14] Z.-Q. Wang, P. Wang, and D. Wang, "Multi-Microphone Complex Spectral Mapping for Utterance-Wise and Continuous Speaker Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2021.
- [15] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN : Speech Enhancement Generative Adversarial Network," in *Proc. Interspeech*, 2017.
- [16] D. Rethage, J. Pons, and X. Serra, "A WaveNet for Speech Denoising," in *Proc. ICASSP*, 2018, pp. 5069–5073.
- [17] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [18] A. Pandey and D. Wang, "A New Framework for CNN-Based Speech Enhancement in the Time Domain," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 1179–1188, 2019.
- [19] A. Défossez, G. Synnaeve, and Y. Adi, "Real Time Speech Enhancement in the Waveform Domain," in *Proc. Interspeech*, 2020.
- [20] Y. Liu and D. Wang, "Divide and Conquer: A Deep CASA Approach to Talker-Independent Monaural Speaker Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2092–2102, 2019.
- [21] U. Isik, R. Giri, N. Phansalkar, J. M. Valin, K. Helwani, and A. Krishnaswamy, "PoCoNet: Better Speech Enhancement with Frequency-Positional Embeddings, Semi-Supervised Conversational Data, and Biased Loss," in *Proc. Interspeech*, 2020, pp. 2487–2491.
- [22] S. Wisdom, J. R. Hershey, K. Wilson, J. Thorpe, M. Chinen, B. Patton, and R. A. Saurous, "Differentiable Consistency Constraints for Improved Deep Speech Enhancement," in *Proc. ICASSP*, 2019, pp. 900–904.
- [23] A. Li, W. Liu, C. Zheng, C. Fan, and X. Li, "Two Heads Are Better Than One: A Two-Stage Complex Spectral Mapping Approach for Monaural Speech Enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2021.
- [24] S. Braun, H. Gamper, C. K. A. Reddy, and I. Tashev, "Towards Efficient Models for Real-Time Deep Noise Suppression," in *Proc. ICASSP*, 2021, pp. 656–660.
- [25] A. Aroudi and S. Braun, "DBnet: Doa-Driven Beamforming Network for End-to-End Reverberant Sound Source Separation," in *Proc. ICASSP*, 2021, pp. 211–215.
- [26] R. Sawata, S. Uhlich, S. Takahashi, and Y. Mitsufuji, "All For One And One For All: Improving Music Separation By Bridging Networks," in *Proc. ICASSP*, 2021, pp. 51–55.
- [27] A. Li, C. Zheng, R. Peng, and X. Li, "On The Importance of Power Compression and Phase Estimation in Monaural Speech Dereverberation," *JASA Express Letters*, vol. 1, no. 1, p. 014802, 2021.
- [28] F. G. Germain, Q. Chen, and V. Koltun, "Speech Denoising with Deep Feature Losses," in *Proc. Interspeech*, 2019, pp. 2723–2727.
- [29] P. Manocha, Z. Jin, R. Zhang, and A. Finkelstein, "CDPAM: Contrastive Learning for Perceptual Audio Similarity," in *Proc. ICASSP*, 2021, pp. 196–200.
- [30] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation," in *Proc. ISMIR*, 2018, pp. 334–340.
- [31] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-Baked or Well Done?" in *Proc. ICASSP*, 2019, pp. 626–630.
- [32] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ)-A New Method for Speech Quality Assessment of Telephone Networks and Codecs," in *Proc. ICASSP*, vol. 2, 2001, pp. 749–752.
- [33] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, sep 2011.
- [34] P. Wang and D. Wang, "Enhanced Spectral Features for Distortion-Independent Acoustic Modeling," in *Proc. Interspeech*, 2019.
- [35] A. Courville, I. Goodfellow, and Y. Bengio, *Deep Learning*. MIT Press, 2016. [Online]. Available: <http://www.deeplearningbook.org>
- [36] D. W. Griffin and J. S. Lim, "Signal Estimation from Modified Short-Time Fourier Transform," *IEEE Trans. Audio, Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, 1984.
- [37] J. Le Roux, N. Ono, and S. Sagayama, "Explicit Consistency Constraints for STFT Spectrograms and Their Application to Phase Reconstruction," *Proceedings of SAPA*, 2008.
- [38] M. Kolbæk, Z.-H. Tan, S. H. Jensen, and J. Jensen, "On Loss Functions for Supervised Monaural Time-Domain Speech Enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 825–838, 2020.
- [39] Y. Luo, E. Ceolini, C. Han, S.-C. Liu, and N. Mesgarani, "FaSNet: Low-latency Adaptive Beamforming for Multi-Microphone Audio Processing," in *Proc. WASPAA*, 2019.
- [40] M. Maciejewski, G. Wichern, E. McQuinn, and J. Le Roux, "WHAMR!: Noisy and Reverberant Single-Channel Speech Separation," in *Proc. ICASSP*, 2020.
- [41] L. Drude, J. Heitkaemper, C. Boeddeker, and R. Haeb-Umbach, "SMS-WSJ: Database, Performance Measures, and Baseline Recipe for Multi-Channel Source Separation and Recognition," in *arXiv preprint arXiv:1910.13934*, 2019.
- [42] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep Clustering: Discriminative Embeddings for Segmentation and Separation," in *Proc. ICASSP*, 2016, pp. 31–35.
- [43] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. Interspeech*, Sep. 2016, pp. 545–549.
- [44] 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 Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, 2017.
- [45] P. Manuel, "<https://github.com/mpariente/pystoi>," 2020.
- [46] Ludlows, "<https://github.com/ludlows/python-pesq>," 2020.
- [47] N. Turpault, S. Wisdom, H. Erdogan, J. Hershey, R. Serizel, E. Fonseca, P. Seetharaman, and J. Salamon, "Improving Sound Event Detection in Domestic Environments using Sound Separation," in *Proc. DCASE*, 2020.
- [48] S. Wisdom, H. Erdogan, D. P. W. Ellis, R. Serizel, N. Turpault, E. Fonseca, J. Salamon, P. Seetharaman, and J. R. Hershey, "What's All The Fuss About Free Universal Sound Separation Data?" in *Proc. ICASSP*, 2021, pp. 186–190.
- [49] Z.-Q. Wang and D. Wang, "A Joint Training Framework for Robust Automatic Speech Recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 4, pp. 796–806, 2016.
- [50] X. Chang, W. Zhang, Y. Qian, J. Le Roux, and S. Watanabe, "MIMO-SPEECH: End-to-End Multi-Channel Multi-Speaker Speech Recognition," in *Proc. ASRU*, 2019.