MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Reverberation as Supervision for Speech Separation

Aralikatti, Rohith; Boeddeker, Christoph; Wichern, Gordon; Subramanian, Aswin Shanmugam; Le Roux, Jonathan

TR2023-016 April 14, 2023

Abstract

This paper proposes reverberation as supervision (RAS), a novel un- supervised loss function for single-channel reverberant speech sepa- ration. Prior methods for unsupervised separation required the syn- thesis of mixtures of mixtures or assumed the existence of a teacher model, making them difficult to consider as potential methods ex- plaining the emergence of separation abilities in an animal's audi- tory system. We assume the availability of two-channel mixtures at training time, and train a neural network to separate the sources given one of the channels as input such that the other channel may be predicted from the separated sources. As the relationship be- tween the room impulse responses (RIRs) of each channel depends on the locations of the sources, which are unknown to the network, the network cannot rely on learning that relationship. Instead, our proposed loss function fits each of the separated sources to the mix- ture in the target channel via Wiener filtering, and compares the resulting mixture to the ground-truth one. We show that minimiz- ing the scale-invariant signal-to-distortion ratio (SI-SDR) of the pre- dicted right-channel mixture with respect to the ground truth implic- itly guides the network towards separating the left-channel sources. On a semi-supervised reverberant speech separation task based on the WHAMR! dataset, using training data where just 5% (resp., 10%) of the mixtures are labeled with associated isolated sources, we achieve 70% (resp., 78%) of the SI-SDR improvement obtained when training with supervision on the full training set, while a model trained only on the labeled data obtains 43% (resp., 45%).

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2023

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

REVERBERATION AS SUPERVISION FOR SPEECH SEPARATION

Rohith Aralikatti[1,2]*, Christoph Boeddeker*[1,3]*, Gordon Wichern*[1]*, Aswin Subramanian*[1]*, Jonathan Le Roux*[1]

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
[2]University of Maryland, College Park, MD, USA [3]Paderborn University, Paderborn, Germany

ABSTRACT

This paper proposes reverberation as supervision (RAS), a novel unsupervised loss function for single-channel reverberant speech separation. Prior methods for unsupervised separation required the synthesis of mixtures of mixtures or assumed the existence of a teacher model, making them difficult to consider as potential methods explaining the emergence of separation abilities in an animal's auditory system. We assume the availability of two-channel mixtures at training time, and train a neural network to separate the sources given one of the channels as input such that the other channel may be predicted from the separated sources. As the relationship between the room impulse responses (RIRs) of each channel depends on the locations of the sources, which are unknown to the network, the network cannot rely on learning that relationship. Instead, our proposed loss function fits each of the separated sources to the mixture in the target channel via Wiener filtering, and compares the resulting mixture to the ground-truth one. We show that minimizing the scale-invariant signal-to-distortion ratio (SI-SDR) of the predicted right-channel mixture with respect to the ground truth implicitly guides the network towards separating the left-channel sources. On a semi-supervised reverberant speech separation task based on the WHAMR! dataset, using training data where just 5% (resp., 10%) of the mixtures are labeled with associated isolated sources, we achieve 70% (resp., 78%) of the SI-SDR improvement obtained when training with supervision on the full training set, while a model trained only on the labeled data obtains 43% (resp., 45%).

Index Terms— Speech separation, semi-supervised learning, room impulse response, wiener filtering

1. INTRODUCTION

The approaches of deep clustering [1] and permutation-invariant training [2, 3] facilitated an explosion of interest in learning to separate overlapped speech signals, a research field commonly known as speech separation. Additional advances included time domain models [4–6] and powerful multi-channel systems [7–9]. Speech separation is now a standard front-end supporting speaker diarization [10–12] and automatic speech recognition [13, 14] amongst other applications. However, full emulation of the human ability to solve the cocktail party problem, i.e., effortlessly focus on target speech in a noisy, multi-talker environment remains far off.

Part of the reason speech separation approaches remain far off from human-level performance is the reliance on fully-supervised approaches requiring vast quantities of labeled mixture data, which is typically generated synthetically. It is challenging to record both the mixture and the ground-truth speech signals in a real-world environment as it is difficult to suppress cross-talk and background noise. However, synthetically generated mixtures cannot accurately model many real-world phenomena such as reverberation and diffraction, material properties of acoustic environments, moving sources and listeners, etc. This domain mismatch between synthetic data and the actual data distribution often causes a reduction in performance. Hence, there is great value in formulating techniques that can leverage recorded, unlabeled mixtures to address this data gap.

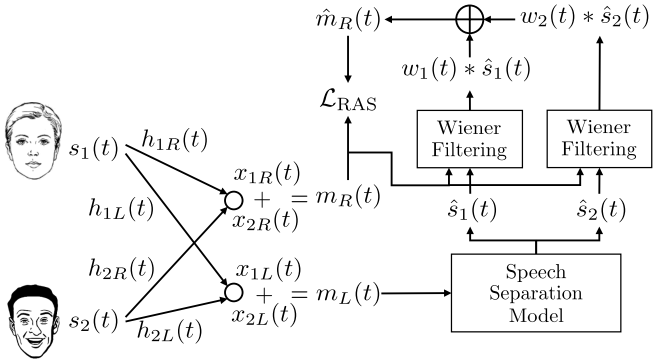

Fig. 1: Illustration of the proposed reverberation as supervision approach.

Recently, MixIT [15] demonstrated great success in unsupervised speech separation by creating "mixtures of mixtures," i.e., summing together multiple overlapped speech signals, separating them, and then using a permutation matrix to reconstruct the original mixtures which are compared to the input mixtures as a training objective. Further extensions of MixIT for real-world meeting data [16] and RemixIT [17] for speech enhancement have confirmed its value for speech separation. However, despite the successes, there is something unsatisfying about MixIT from a biological perspective, i.e., it is hard to imagine the human brain learning to separate sounds by creating mixtures of mixtures. Humans use a variety of cues - contextual and perceptual cues, speaker lip movements, and spatial information to extract the required speech. In this work, we explore an approach where difference in the observed reverberation between two microphones (analogous to two ears) can be used as a supervision signal for speech signals as illustrated in Figure 1.

Prior approaches for unsupervised speech separation based on spatial cues [18–20] use features such as inter-channel phase difference computed from two channel speech mixtures to create pseudo labels for training supervised speech separation models. Similar to the discussion on MixIT above, the existence of a teacher model is difficult to consider as a potential method explaining the emergence of separation abilities in an animal's auditory system. The approach in [21] does not make use of a teacher-student setup, but uses a large number of channels to learn a mask-based beamformer. Inspired by [22], which uses the strong inductive biases of an audio rendering system to train a multi-channel separation network in an

This work was performed while R. Aralikatti and C. Boeddeker were interns at MERL.

unsupervised manner, we here propose reverberation as supervision (RAS) to train a neural network to separate speech sources given one channel in a two-channel mixture such that the other channel may be predicted from the separated sources. By fitting each of the separated sources to the mixture in the opposite channel via differentiable Wiener filtering [23], we use the difference between the reconstructed and ground-truth opposite channel as our loss function.

Through experiments on the WHAMR! dataset [24] of two-channel reverberant two-speaker mixtures, we demonstrate the value of RAS in a semi-supervised setting. We also enumerate several design considerations necessary to obtain strong performance, such as the use of independent Wiener filters for each speaker, and filtering samples whose geometry is uninformative for the RAS objective. While we consider the two-channel two-source case for simplicity, the method could be readily extended to more channels (either to train a single-channel system, or a K channel system from a dataset of M channel mixtures, for $K < M$) and more sources.

2. REVERBERATION AS SUPERVISION (RAS)

2.1. Motivation

We consider two dry source speech signals $s_1(t)$ and $s_2(t)$ originating from two speakers in different locations in a room, and a listener consisting of a two-microphone array at another location. For clarity of presentation and to stress the analogy with a binaural biological system, we refer to the individual channels as Left (L) and Right (R). The room impulse response from speaker $k \in \{1, 2\}$ to microphone $c \in \{L, R\}$ is denoted as $h_{kc}(t)$. The source-image signal at each microphone is obtained as

$$x_{kc}(t) = h_{kc}(t) * s_k(t), \tag{1}$$

where $*$ denotes convolution, and the observed mixture as

$$m_c(t) = \sum_k x_{kc}(t). \tag{2}$$

The direct-path (anechoic) signal corresponding to $x_{kc}(t)$ is denoted as $d_{kc}(t)$. For simplicity of presentation, we will consider the left channel as the input to the system, and the right channel as the supervision. At training time, we can alternately reverse their roles.

We are interested in deriving a loss function that pushes a neural network to separate the input mixture in the left channel into its constituent sources, by utilizing as an implicit measure of separation performance how well the observed mixture in the right channel can be predicted from the separated outputs. Because the network has no knowledge of the location of the sources, this prediction requires a relative impulse response to be applied in order to match the right channel. One way to do so is to apply a Wiener filter to the estimates.

In order to assess the viability of using the predicted fit to the right channel as an implicit measure of separation of the left channel, we first consider oracle experiments in which we fit the left-channel signals of the original input mixture as well as various versions of the ground-truth sources to the right-channel mixture via a Wiener filter, and evaluate the prediction performance. When fitting the two sources, the Wiener filter for each source is computed independently, and the filtered sources are added to form the mixture estimate $\hat{m}_R$. We use a Wiener filter with 512 coefficients, out of which 100 are non-causal. Non-causal filter weights are required as the right-channel signal may have a negative delay with respect to the left-channel signal if a source is closer to the right channel. Table 1 shows the signal-to-distortion ratio (SDR) of the right-channel mixture estimate $\hat{m}_R$ with respect to the ground truth m_R over the WHAMR! validation set. We see that there is an SDR improvement of 5 to 6 dB when estimating the right-channel mixture from some version of the ground-truth left-channel sources as compared to estimating it from the left-channel mixture m_L itself. We believe this difference is sufficient to push the neural network towards separating the sources when we optimize the speech separation network through the Wiener filter to reduce the SDR of the predicted right-channel mixture. We also see that the direct-path signals lead to better reconstruction than the reverberant source-image signals, indicating that our objective may also incentivize some amount of dereverberation.

Table 1: Oracle SDR [dB] when predicting the right channel mixture from various left channel signals using a Wiener filter. We see that the filter models the output mixture more accurately when separated speech signals are given as input. Predicting the right channel mixture from the left channel mixture leads to a reduction of roughly 5 dB in SDR.

Wiener filter input	SDR(m_R, $\hat{m}_R$)
Mixture m_L	6.9
Reverberant source-image signals x_{1L}, x_{2L}	12.0
Dry source signals s_1, s_2	13.2
Direct-path signals d_{1L}, d_{2L}	13.9

2.2. Filter design considerations

As the network has no knowledge of the relative Room Impulse Response (RIR) of each source from the left channel to the right channel, we need to allow some filtering on the separated signals to fit the right-channel mixture. We need to carefully design the filter in order to allow the right amount of flexibility, as too much flexibility would allow any signal to fit the right-channel mixture, while too little flexibility would prevent any meaningful fit, which in both cases would result in an uninformative objective. Design choices to consider include in particular the filter length and the split between causal and non-causal coefficients, and the way the filter is optimized. The filters we consider are FIR filters.

Causality and filter length: The longer the filter, the better it can fit a target signal with less dependence on the signal being filtered. We are not concerned with perfectly reproducing the relative RIR between the left and right channels, and focus on reproducing the effect of its early parts. As a source may be closer to the right channel than the left one, we also need to allow for non-causal coefficients so that the filtered signal may start earlier than the signal being filtered. We thus assume that a linear filter $w_k(t)$ lies between a good estimate for the kth speaker $\hat{x}_{kL}(t)$ of the left microphone and the signal of the same speaker at the other microphone $x_{kR}(t)$:

$$\hat{x}_{kR}(t) = (w_k * \hat{x}_{kL})(t) = \sum_{\tau=-\tau_{\text{nc}}}^{\tau_{\text{c}}-1} w_k(\tau)\hat{x}_{kL}(t-\tau), \tag{3}$$

where τ_{nc} denotes the number of non-causal coefficients, and τ_{c} the number of causal coefficients. In practice, we set $\tau_{\text{nc}} = 100$ and $\tau_{\text{c}} = 412$.

Joint vs independent filter estimation: To estimate the filter $w_k(t)$, a common objective is the sum of the squared errors

$$\underset{w_k(t)}{\arg\min} \sum_t |(w_k * \hat{x}_{kL})(t) - x_{kR}(t)|^2, \tag{4}$$

which yields the solution of the Wiener-Hopf equation.

This estimation requires supervision, but when we jointly estimate the filters for each speaker (i.e., we sum the estimates), we could use the observation $m_R(t)$ as target in the estimation:

$$\underset{w_1(t),w_2(t)}{\arg\min} \sum_t \left| \sum_k (w_k * \hat{x}_{kL})(t) - m_R(t) \right|^2 . \tag{5}$$

This equation is well suited to estimate the filters $w_k(t)$, but has issues with the gradients for $\hat{x}_{kL}(t)$, which are necessary for the Neural Network (NN) training. Indeed, any linear combination of the estimates $\hat{x}_{kL}(t)$ will yield the same score of the objective, if the linear equation system has a full rank (i.e., the transformation matrix is invertible).

To partially address this issue, we consider an independent estimation:

$$\underset{w_k(t)}{\arg\min} \sum_t |(w_k * \hat{x}_{kL})(t) - m_R(t)|^2, \forall k, \tag{6}$$

while keeping the estimation of $\hat{m}_R(t)$ as:

$$\hat{m}_R(t) = w_1(t) * \hat{x}_{1L}(t) + w_2(t) * \hat{x}_{2L}(t). \tag{7}$$

It can be shown that this independent estimation promotes some notion of uncorrelatedness between the estimates, which is a stronger constraint than with the joint estimation. It however does not by itself fully constrain the network to output statistically independent estimates $\hat{x}_{kL}$ that are close to each source-image signal, which is why we will assume the availability of a small amount of supervised data in order to nudge the network towards the proper solution. More advanced techniques for obtaining statistically independent estimates, such as independent component analysis (ICA) [25], may be considered in future works to introduce a stronger inductive bias and avoid this issue.

2.3. RAS objective

Our goal is to perform single-channel speech separation to estimate the separated signals $x_{1L}(t)$ and $x_{2L}(t)$ given an observed mixture at the left channel $m_L(t)$. In our proposed method, we first obtain estimates $\hat{x}_{kL}(t)$ from a neural network given $m_L(t)$ as input. Wiener filtering is used to independently fit each $\hat{x}_{kL}(t)$ to the right-channel mixture $m_R(t)$ as in Eq. (6), using the filtered signals to obtain an estimate $\hat{m}_R(t)$ of the right-channel mixture as in Eq. (7). The Reverberation-as-supervision (RAS) loss is then defined as:

$$\mathcal{L}_{\text{RAS}}(m_R, \hat{x}_{1L}, \hat{x}_{2L}) = \mathcal{L}_{\text{SI-SDR}}(m_R(t), \hat{m}_R(t)), \tag{8}$$

where $\mathcal{L}_{\text{SI-SDR}}$ is the SI-SDR loss [4, 26, 27] defined as

$$\mathcal{L}_{\text{SI-SDR}}(x, \hat{x}) = -10 \log_{10} \frac{\|\alpha x\|^2}{\|\alpha x - \hat{x}\|^2}, \ \alpha = \frac{\langle x, \hat{x} \rangle}{\|x\|^2}. \tag{9}$$

As no information about the ground-truth isolated signals is used in $\mathcal{L}_{RAS}$, it can be applied to mixtures for which no ground-truth signals are available. Assuming the availability of a small amount of labeled data (single-channel mixtures with corresponding isolated signals), we can combine $\mathcal{L}_{\text{SI-SDR}}$ on the labeled data with $\mathcal{L}_{\text{RAS}}$ on unlabeled data (two-channel mixtures) to obtain a semi-supervised approach for training speech separation models. In our experiments, we show that this semi-supervised approach leads to impressive results when compared to a fully-supervised setting with just a fraction (5 % - 10 %) of supervised data.

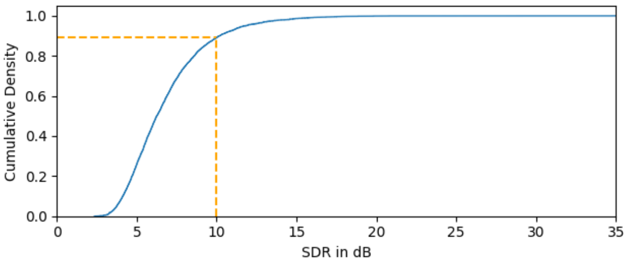

Fig. 2: Cumulative density function (CDF) of $\text{SDR}(m_R(t), \hat{m}_R(t))$ with $\hat{m}_R(t)$ estimated from the left channel mixture $m_L(t)$ using a Wiener filter of size 512 with 100 non-causal weights and $m_R(t)$ as the reference signal. We discard mixtures for which $\text{SDR}(m_R(t), \hat{m}_R(t))$ is above $10\,\text{dB}$ (roughly $10\,\%$ of the mixtures in the training data) during fine-tuning, as the left and right channels are too similar.

2.4. Data selection to remove uninformative examples

While applying RAS, we find it beneficial to filter out mixtures that had a very high correlation between the left and right channel mixtures. Intuitively, if it is easy to predict the right-channel mixture from the left-channel mixture, then there is not much left to gain for the network by separating the sources to further improve the right-channel prediction. These examples may consist for example of situations in which the sources are located in similar positions with respect to the two channels.

We thus first estimate the right-channel mixture from the left-channel mixture using a Wiener filter, and compute the SDR between the estimate and the observed right-channel. We filter out mixtures by thresholding this SDR value. Empirically, we found that using mixtures that had an SDR lower than $10\,\text{dB}$ for fine-tuning gave the best results. The cumulative density function (CDF) of SDR for mixtures present in the validation set is shown in Fig. 2.

3. EXPERIMENTAL VALIDATION

3.1. Dataset and experimental setup

All experiments in this paper are done on the dual-channel (microphone spacing between 15 - 17 cm), noise-free, 8 kHz min version of the WHAMR! [24] speech separation dataset. Realistic speech mixtures in WHAMR! are generated by accurately simulating room reverberation effects and by adding background noise recorded in a variety of environments. Here, we do not add noise, and attempt to separate a reverberant mixture into its constituent reverberant sources. Our proposed approach assumes that a convolutive transfer function maps signals from one channel to the other. The addition of noise would weaken that assumption, and is left to future work.

Because of the statistical independence issue discussed in Section 2.2, we are unable to use the RAS loss in fully unsupervised settings, however, we demonstrate that our proposed method can obtain good speech separation performance with very limited supervised data. We study the impact of the RAS loss in three different data configurations. The WHAMR! training set consists of 20 000 labeled mixtures. In the first two configurations, we restrict the number of supervised training examples (labeled data) to 500 and 1000 samples. These samples are used to train supervised baseline models denoted as “Sup. (500)” and “Sup. (1000),” respectively. Along with these supervised examples, the remaining two-channel mixtures (19 500 and 19 000 samples, respectively) are used for unsupervised fine-tuning using our proposed RAS approach. In a third configuration, we use the full training set of 20 000 labeled mixtures to train a supervised topline model, denoted as “Sup. (full).” This represents

Table 2: Separation and objective speech quality metrics for semi-supervised reverberant speech separation. Metrics are shown for the raw output of the network and after applying a Wiener filter to fit the left-channel mixture m_L, in the format "raw output \| after Wiener filter". SDR and SI-SDR are in dB. Numbers in parentheses denote the amount of supervised samples available for training. For RAS, all other samples are used as unlabeled data.

Model	SDR	SI-SDR	PESQ	STOI
Sup. (500) baseline	2.8 | 2.0	2.2 | 3.0	1.68 | 1.65	0.69 | 0.68
Sup. (500) + RAS	**5.1** | 4.6	4.1 | **5.3**	1.65 | **1.69**	**0.71** | **0.71**
Sup. (1000) baseline	3.9 | 2.6	3.4 | 3.3	**1.91** | 1.85	0.73 | 0.72
Sup. (1000) + RAS	**5.9** | 5.4	4.9 | **5.9**	1.73 | 1.79	**0.74** | **0.74**
Sup. (full) topline	7.5 | 6.4	7.6 | 7.0	2.36 | 2.29	0.83 | 0.82

the best possible speech separation for our model configuration and hyperparameters.

The speech separation model consists of a four-layer bi-directional LSTM with 600 hidden units in each layer. We use dropout with a probability of 0.3 in each layer. The BLSTM predicts a phase-sensitive approximation (PSA) mask [28] for each source. The input to the network is the log of the STFT magnitude of the observed mixture computed with a window size of 32 ms and a hop size of 8 ms. In the RAS case, we first run a pre-training stage on the supervised data, and then a second training stage using the semi-supervised objective combining $\mathcal{L}_{\text{SI-SDR}}$ on the labeled data with $\mathcal{L}_{\text{RAS}}$ on the unlabeled data. Each stage of training is done for a maximum of 100 epochs. The Adam optimizer is used with an initial learning rate of 0.0001. The learning rate is decayed by a factor of 0.5 whenever the validation loss stagnates for more than 5 epochs.

We evaluate performance using BSSEval SDR [27] using the dry ground truth source signals as reference. We also compute SI-SDR [26], PESQ [29], and STOI [30] using the ground-truth reverberant signal as reference. Higher is better for all metrics.

3.2. Results

Table 2 shows the improvement obtained by applying RAS on models trained with limited supervised data. We present metrics on the raw output of the network, as well as after applying a Wiener filter that fits the raw output to the *left-channel* mixture m_L. This is different from what is done at training time (where we have access to the right-channel mixture), so that the inference procedure remains single-channel, only considering the left-channel input. In the supervised case, computing metrics after the Wiener filter causes a mismatch between training and test conditions and can thus be ignored, but we include them in Table 2 for completeness. All metrics are computed using the left-channel signals (either dry or reverberant depending on the metric) as reference. RAS significantly improves separation quality, with a 2 - 3 dB improvement in SDR and SI-SDR. Wiener filtering to fit the left-channel mixture results in a decrease of SDR and an increase in SI-SDR. This is reasonable as SDR compares to the dry source signal and compensates for the channel, so reintroducing a channel effect via Wiener filtering is likely detrimental; on the other hand, SI-SDR compares to the reverberant source and does not compensate for channel mismatch, so fitting the raw output to the mixture via filtering helps match the channel and increase the score. For the speech quality metrics, STOI shows small improvements in both data settings, while PESQ decreases slightly compared to the supervised model trained with 1000 samples. We noticed that RAS helps suppress the interfering speaker, which leads to the observed gain in time-domain SDR metrics, but may at times also over-suppress the interfering speaker which is heavily penalized by PESQ. Wiener filtering to fit the mixture mitigates this issue and reduces the gap. Replacing the SI-SDR loss in (9) with an STFT-based loss function might also improve PESQ.

Table 3: Reverberant speech separation results when using anechoic supervised samples and reverberant unsupervised samples for training. Metrics are on raw output \| after Wiener filter to fit m_L. SDR and SI-SDR are in dB.

Model	SDR	SI-SDR	PESQ	STOI
Sup. (500) baseline	3.2 | 2.1	2.7 | 2.7	**1.86** | 1.80	**0.72** | 0.71
Sup. (500) + RAS	**5.3** | 4.9	4.3 | **5.5**	1.68 | 1.72	**0.72** | **0.72**
Sup. (1000) baseline	4.0 | 2.9	3.5 | 3.5	**1.83** | 1.79	0.73 | 0.72
Sup. (1000) + RAS	**5.8** | 5.1	4.7 | **5.7**	1.74 | 1.78	0.73 | **0.74**
Sup. (full) topline	6.9 | 5.8	6.6 | 6.0	2.23 | 2.21	0.81 | 0.80

Table 4: Ablation experiments for RAS loss function. Metrics are on the raw output of the network using 1000 supervised examples.

Ablation	SDR	SI-SDR	PESQ	STOI
Filter: indep. ⇒ joint	3.1	1.5	1.72	0.70
Loss: SI-SDR ⇒ SNR	5.5	4.5	1.67	0.71
Threshold: 10 ⇒ ∞	5.8	4.8	**1.73**	**0.74**
Proposed	**5.9**	**4.9**	**1.73**	**0.74**

In Table 3, we show that RAS can be used for domain adaptation from anechoic to reverberant data. In these experiments, we only use the supervised anechoic data (which in practice is typically easier to obtain than labeled reverberant mixtures) from WHAMR! and finetune on reverberant mixtures using RAS. When comparing the "Sup. (full) topline" rows between Table 2 and Table 3, as expected, we observe better performance on reverberant test data when using reverberant training data (Table 2). However, when using limited amounts of supervised data, we see that models trained on anechoic data (Table 3) actually perform better than those trained on reverberant data (Table 2), possibly because when learning to separate with limited training data, anechoic mixtures are easier to learn a model of speech from. The trends between RAS and the supervised models are similar between Table 2 and Table 3.

Table 4 shows ablation results for our proposed method. If we replace the independent Wiener filter estimation for the two estimated sources with joint filter estimation, we observe a drop in all metrics. This is consistent with the discussion in Section 2.2 regarding the difficulty of obtaining consistent gradient directions for joint filter estimation. Replacing the SI-SDR loss in Eq. (8) with SNR (i.e., removing scale invariance) also reduces separation quality. While the Wiener filter could be expected to absorb any scale ambiguities, in practice the added flexibility from the scale-invariant loss proves valuable. Filtering out mixtures used for fine-tuning (using a threshold on SDR($m_R(t)$, $\hat{m}_R(t)$) as described in Section 2.3) leads to a slight boost in SDR, but it is also beneficial as it reduces training time by not including unsupervised examples for which RAS separation is not expected to work well.

4. CONCLUSION

We proposed reverberation as separation (RAS), a semi-supervised approach for reverberant speech separation. We have shown that RAS significantly improves speech separation when there is limited labeled data, and that it is a viable approach for domain adaptation. Future work includes improving the design of the filtering procedure to introduce a stronger inductive bias that further promotes separation of the source estimates, and exploring the possibility of a fully unsupervised version of RAS.

5. REFERENCES

[1] J. R. Hershey, Z. Chen, and J. Le Roux, “Deep clustering: Discriminative embeddings for segmentation and separation,” in *Proc. ICASSP*, Mar. 2016, pp. 31–35.

[2] M. Kolbæk, D. Yu, Z.-H. Tan, J. Jensen *et al.*, “Multitalker 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.

[3] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe *et al.*, “Single-channel multi-speaker separation using deep clustering,” in *Proc. Interspeech*, Sep. 2016, pp. 545–549.

[4] 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.

[5] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude *et al.*, “Demystifying tasnet: A dissecting approach,” in *Proc. ICASSP*, May 2020, pp. 6359–6363.

[6] D. Ditter and T. Gerkmann, “A multi-phase gammatone filter-bank for speech separation via tasnet,” in *Proc. ICASSP*, May 2020, pp. 36–40.

[7] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, “Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation,” in *Proc. ICASSP*, Apr. 2018, pp. 1–5.

[8] Z.-Q. Wang and D. Wang, “Integrating spectral and spatial features for multi-channel speaker separation.” in *Proc. Interspeech*, Sep. 2018, pp. 2718–2722.

[9] R. Gu, S.-X. Zhang, L. Chen, Y. Xu *et al.*, “Enhancing end-to-end multi-channel speech separation via spatial feature learning,” in *Proc. ICASSP*, May 2020, pp. 7319–7323.

[10] K. Kinoshita, M. Delcroix, S. Araki, and T. Nakatani, “Tackling real noisy reverberant meetings with all-neural source separation, counting, and diarization system,” in *Proc. ICASSP*, May 2020, pp. 381–385.

[11] Z. Chen, T. Yoshioka, L. Lu, T. Zhou *et al.*, “Continuous speech separation: Dataset and analysis,” in *Proc. ICASSP*, May 2020, pp. 7284–7288.

[12] D. Raj, P. Denisov, Z. Chen, H. Erdogan *et al.*, “Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis,” in *Proc. SLT*, Jun. 2021, pp. 897–904.

[13] S. Settle, J. Le Roux, T. Hori, S. Watanabe *et al.*, “End-to-end multi-speaker speech recognition,” in *Proc. ICASSP*, Apr. 2018, pp. 4819–4823.

[14] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe *et al.*, “Far-field automatic speech recognition,” *Proceedings of the IEEE*, vol. 109, no. 2, pp. 124–148, 2021.

[15] S. Wisdom, E. Tzinis, H. Erdogan, R. J. Weiss *et al.*, “Unsupervised sound separation using mixture invariant training,” in *Proc. NeurIPS*, vol. 33, Dec. 2020, pp. 3846–3857.

[16] A. Sivaraman, S. Wisdom, H. Erdogan, and J. R. Hershey, “Adapting speech separation to real-world meetings using mixture invariant training,” in *Proc. ICASSP*, May 2022, pp. 686–690.

[17] E. Tzinis, Y. Adi, V. K. Ithapu, B. Xu *et al.*, “Remixit: Continual self-training of speech enhancement models via bootstrapped remixing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1329–1341, 2022.

[18] L. Drude, D. Hasenklever, and R. Haeb-Umbach, “Unsupervised training of a deep clustering model for multichannel blind source separation,” in *Proc. ICASSP*, 2019, pp. 695–699.

[19] P. Seetharaman, G. Wichern, J. Le Roux, and B. Pardo, “Bootstrapping single-channel source separation via unsupervised spatial clustering on stereo mixtures,” in *Proc. ICASSP*, 2019, pp. 356–360.

[20] E. Tzinis, S. Venkataramani, and P. Smaragdis, “Unsupervised deep clustering for source separation: Direct learning from mixtures using spatial information,” in *Proc. ICASSP*, 2019, pp. 81–85.

[21] L. Drude, J. Heymann, and R. Haeb-Umbach, “Unsupervised Training of Neural Mask-Based Beamforming,” in *Proc. Interspeech*, 2019, pp. 1253–1257.

[22] D. Arteaga and J. Pons, “Multichannel-based learning for audio object extraction,” in *Proc. ICASSP*, 2021, pp. 206–210.

[23] C. Boeddeker, W. Zhang, T. Nakatani, K. Kinoshita *et al.*, “Convolutive transfer function invariant SDR training criteria for multi-channel reverberant speech separation,” in *Proc. ICASSP*, 2021, pp. 8428–8432.

[24] M. Maciejewski, G. Wichern, E. McQuinn, and J. Le Roux, “WHAMR!: Noisy and reverberant single-channel speech separation,” in *Proc. ICASSP*, 2020, pp. 696–700.

[25] A. Hyvarinen, J. Karhunen, and E. Oja, “Independent component analysis,” *Studies in informatics and control*, vol. 11, no. 2, pp. 205–207, 2002.

[26] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR–half-baked or well done?” in *Proc. ICASSP*, May 2019.

[27] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[28] 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*, Apr. 2015, pp. 708–712.

[29] 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, May 2001, pp. 749–752.

[30] 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, 2011.