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

WHAMR!: Noisy and Reverberant Single-Channel Speech Separation

Maciejewski, Matthew; Wichern, Gordon; McQuinn, Emmett; Le Roux, Jonathan

TR2020-042 April 11, 2020

Abstract

While significant advances have been made with respect to the separation of overlapping speech signals, studies have been largely constrained to mixtures of clean, near anechoic speech, not representative of many real-world scenarios. Although the WHAM! dataset introduced noise to the ubiquitous wsj0-2mix dataset, it did not include reverberation, which is generally present in indoor recordings outside of recording studios. The spectral smearing caused by reverberation can result in significant performance degradation for standard deep learning-based speech separation systems, which rely on spectral structure and the sparsity of speech signals to tease apart sources. To address this, we introduce WHAMR!, an augmented version of WHAM! with synthetic reverberated sources, and provide a thorough baseline analysis of current techniques as well as novel cascaded architectures on the newly introduced conditions.

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

WHAMR!: NOISY AND REVERBERANT SINGLE-CHANNEL SPEECH SEPARATION

Matthew Maciejewski[1,2], *Gordon Wichern*[1], *Emmett McQuinn*[3], *Jonathan Le Roux,*[1]

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
[2]Johns Hopkins University, Baltimore, MD, USA [3]Whisper.ai, San Francisco, CA, USA
mmaciej2@jhu.edu, {wichern, leroux}@merl.com, emmett@whisper.ai

ABSTRACT

While significant advances have been made with respect to the separation of overlapping speech signals, studies have been largely constrained to mixtures of clean, near anechoic speech, not representative of many real-world scenarios. Although the WHAM! dataset introduced noise to the ubiquitous wsj0-2mix dataset, it did not include reverberation, which is generally present in indoor recordings outside of recording studios. The spectral smearing caused by reverberation can result in significant performance degradation for standard deep learning-based speech separation systems, which rely on spectral structure and the sparsity of speech signals to tease apart sources. To address this, we introduce WHAMR!, an augmented version of WHAM! with synthetic reverberated sources, and provide a thorough baseline analysis of current techniques as well as novel cascaded architectures on the newly introduced conditions.

Index Terms— speech separation, speech enhancement, cocktail party problem, reverberation

1. INTRODUCTION

In recordings produced in natural settings with multiple speakers present, it often occurs that more than one person will speak at the same time. The resulting overlapped speech can cause a severe degradation in the performance of speech processing technologies designed for only a single speech signal, such as automatic speech recognition and speaker identification. Moreover, overlapped speech can be difficult to understand for human listeners as well. Speech separation systems aim to solve this problem by producing multiple waveforms, each estimating the clean speech of a single speaker, from recordings of overlapped speech.

Great advancements have been made in recent years on solving the speech separation problem through deep learning-based techniques [1–6]. However, the overwhelming majority of research conducted thus far has used the wsj0-2mix dataset [1], which consists of synthetically-mixed studio recordings of read utterances from the WSJ0 corpus [7] and is not representative of many real-world scenarios in which overlapped speech may be present [8]. In many cases where multiple people are speaking at the same time, they are not speaking directly into the microphone, and are instead captured by a microphone placed at some distance away in the room, as in meetings or in home settings. In these far-field conditions, the distance from the source to the microphone can lead to a relative increase in noise compared to the speech and to increased reverberation [9], neither of which are present in the most common deep learning-based speech separation evaluations. The addition of noise not only masks the speech signal but also corrupts phase information, while reverberation causes spectral smearing of the source. These phenomena could be challenging for separation systems which rely on the spectral structure of speech in the time-frequency domain [10]. The introduction of the WHAM! dataset [11], consisting of two speaker mixtures from the wsj0-2mix dataset together with real ambient noise, was a first step in the direction of more realism. It did not however consider reverberation or more generally spatialization of the speech signals, despite the noise samples being recorded in stereo.

To aid in the development and evaluation of speech separation systems in even more realistic conditions, we introduce the WHAMR! dataset that adds reverberation to WHAM!'s noise augmentation of wsj0-2mix. We have generated realistic room parameters which are used to generate room impulse responses that can produce reverberant audio waveforms for each source in a manner similar to the multi-channel version of wsj0-2mix introduced in [12], but with the microphone geometry constrained by the binaural recording setup used to collect the WHAM! noise corpus. Although some noisy and reverberant speech separation datasets were introduced in [13], they are constructed using actual recordings of noisy and reverberant speech. As such, they lack ground truth for clean and anechoic speech. WHAMR! provides a contrasting and complementary data paradigm; similarly to other WSJ0-based speech separation datasets, WHAMR! is constructed synthetically, with artificially-mixed speech plus noise and artificial reverberation. This synthetic construction provides the ground truth of all speech signals with and without reverberation, which is necessary to effectively train and evaluate deep learning-based systems.

In this paper, we investigate the performance of various systems for clean, noisy, reverberant, and noisy plus reverberant separation as well as enhancement (denoising and dereverberation) tasks based on the WHAMR! dataset, establishing strong baselines and proposing new cascaded combination systems that can be trained end-to-end.

2. WHAMR! DATASET

The WHAMR! dataset[1] is an extension of the WHAM! dataset [11], which is a noise-augmented version of the wsj0-2mix dataset [1]. The wsj0-2mix dataset consists of mixtures of utterances from the WSJ0 corpus, combined with random gain between 0 and 5 dB to create overlapping speech. There are four configurations: a *min* condition where the mixture is trimmed to the length of the shorter utterance and the corresponding non-trimmed *max* condition, both available at 8 kHz and 16 kHz sampling rate. The mixtures are partitioned into training, validation, and test sets of 20,000, 5,000, and 3,000 mixtures respectively. In the WHAM! dataset, each speech mixture from the wsj0-2mix corpus was associated to a randomly

This work was performed while M. Maciejewski was an intern at MERL.

[1]Available at: http://wham.whisper.ai

Table 1. Room impulse response parameter sampling distributions. Units for all parameters are meters with the exception of reverberation time (T_{60}) which is in seconds and angles in radians.

Room	L	$\mathcal{U}(5, 10)$
	W	$\mathcal{U}(5, 10)$
	H	$\mathcal{U}(3, 4)$
$\mathbf{T_{60}}$	high	$\mathcal{U}(0.4, 1.0)$
	med.	$\mathcal{U}(0.2, 0.6)$
	low	$\mathcal{U}(0.1, 0.3)$

Mic. Center	L	$\frac{L_{\text{Room}}}{2} + \mathcal{U}(-0.2, 0.2)$
	W	$\frac{W_{\text{Room}}}{2} + \mathcal{U}(-0.2, 0.2)$
	H	$\mathcal{U}(0.9, 1.8)$
Mic. Array	sep.	noise mic. separation
	θ	$\mathcal{U}(0, 2\pi)$
Sources	H	$\mathcal{U}(0.9, 1.8)$
	dist.	$\mathcal{U}(0.66, 2)$
	θ	$\mathcal{U}(0, 2\pi)$

sampled excerpt from noises recorded with binaural microphones in various urban environments throughout the San Francisco Bay Area, and mixed such that the louder speaker was at a randomly selected SNR between -6 and $+3$ dB relative to the noise [11].

WHAMR! extends WHAM! by introducing reverberation to the speech sources in addition to the existing noise. Room impulse responses were generated and convolved using pyroomacoustics [14] according to the random room configurations shown in Table 1. Reverberation times were chosen to approximate domestic and classroom environments [9] (as we expect these to be similar to the restaurants and coffee shops where the WHAM! noise was collected), and further classified as high, medium, and low reverberation based on a qualitative assessment of the mixture's noise recording.

We created spatialized versions—*anechoic* and *reverberant*—of all components of the original WHAM! dataset, except noise, which was recorded spatialized. The anechoic sources (i.e., direct path signals) serve as targets to reverberated sources for models involving dereverberation, allowing them to be trained without needing to account for the time delay of the spatialized sources. In spatializing the audio, we generated a two-channel version of the dataset, using microphone spacing from the WHAM! noise metadata, but in this study we focus on single-channel separation and use only the left channel. The spatialized audio was rescaled to remove attenuation, such that the non-spatialized WHAM! and anechoic WHAMR! differ only by small time delays, and we found negligible performance differences when training and testing models using the two datasets. While the results for non-reverberant conditions in Section 4 use anechoic WHAMR!, they are directly comparable with WHAM! [11].

Since all source, noise, and reverberated components and their combinations are included in the corpus, several enhancement, separation, and joint enhancement-separation tasks are enabled for training and evaluation. For example, in separating noisy and reverberant speech, we may want to produce either two clean, anechoic recordings or two clean, reverberant recordings, leaving dereverberation to post-processing. We choose to define four core separation tasks:

- **clean** – anechoic clean mixture to anechoic sources
- **noisy** – anechoic noisy mixture to anechoic sources
- **reverberant** – reverberant clean mixture to anechoic sources
- **noisy and reverberant** – reverberant noisy mixture to anechoic sources

All other configurations are only considered and evaluated as subcomponents to the above tasks. Since each condition has its own unprocessed signal-to-distortion ratio (SDR), comparisons across tasks can be difficult. By restricting to the above tasks, where the targets are the same in all four conditions, raw SDR can be thought of as a directly comparable, "objective" quality metric of the output sources across tasks. SDR *improvement* also brings insight by reporting how much improvement a system has made to the signal.

3. EXPERIMENTAL CONFIGURATIONS

3.1. Network Configurations

For our experiments, we use four basic network configurations, all under the same paradigm. First, the waveforms are projected to a spectro-temporal representation. Next, an internal network takes the spectral representation and produces a spectral mask with values from 0 to 1. Finally, this spectral mask is applied to the original representation, suppressing interfering signals, before the representation is projected back to produce an estimated source waveform. In enhancement, the internal masking network produces a single mask, attempting to suppress noise and/or reverberation. In separation, the masking network produces a mask for each speech signal, attempting to suppress the interfering speakers from each target speaker.

The four configurations we use are the possible combinations of two spectral feature extractors and two internal masking networks. The feature extractors we compare are a standard short-time Fourier transform (STFT) and a TasNet-style learned basis transform [5, 15], which consists of projecting sliding-window subsegments of the waveform onto a set of learned basis functions. The resulting weights can be applied to a reconstruction set of basis functions and summed together along the same sliding window to reconstruct the signal under a similar paradigm to overlap-and-add for the STFT. For internal masking, we evaluate both bi-directional long short-term memory (BLSTM) networks (the typical internals of earlier deep learning-based speech separation systems [1–4,11,15]) and temporal convolutional networks (TCN) [16] with dilated convolutions (popular in recent state-of-the-art separation techniques [5,6]).

For consistency with the prior WHAM! work [11], our BLSTM architecture has four BLSTM layers with 600 units in each direction followed by a fully-connected layer for each output mask. A dropout of 0.3 is applied on each BLSTM layer output except the last. The TCN architecture was chosen to match the best system reported in [5]. It consists of a 128-dimensional bottleneck, 128-dimensional skip-connection paths, and 512 channels in the convolutional blocks, with kernel size 3, 8 blocks per repeat, and 3 repeats.

The STFT features are also chosen to be consistent with [11], with a window length of 32 ms and hop size of 8 ms. The log of the magnitude spectrum is used as input to the internal masking network. The learned basis feature parameters are also chosen to be consistent with [11], with a 10 ms window and 5 ms hop, with 500 learned basis vectors. While the original BLSTM TasNet [15] used a gated convolutional encoder, in this work we use a single learned encoder and ReLU nonlinearity as in Conv-TasNet [5] for both the BLSTM and TCN masking networks with learned bases.

For separation, we evaluate learned basis configurations only, as they have been shown to outperform STFT-based methods on clean data, and performed best in preliminary experiments. However, we perform full comparisons of the differing features for enhancement, for which TasNet-like systems have only rarely been evaluated [17].

We train all networks using permutation invariant training [1, 3] with the scale-invariant signal-to-distortion ratio (SI-SDR, also referred to as SI-SNR) waveform-level training objective [2, 15, 18]. SI-SDR is also the evaluation metric and allows for end-to-end joint training of cascaded enhancement and separation models:

$$\text{SI-SDR} = 10\log_{10}(\|\alpha s\|^2/\|\alpha s - \hat{s}\|^2),\ \alpha = \langle \hat{s}, s\rangle/\|s\|^2. \quad (1)$$

Because the loss is scale-invariant and the outputs are not constrained to sum up to the mixture, the outputs may be in a different dynamic range as the mixture, which as we will see can lead to problems with the cascaded models proposed in this work.

3.2. Cascaded Models

In addition to training single models for each of the WHAMR! core tasks, we evaluate combinations of models in which enhancement (i.e., denoising and/or dereverberation) and separation systems are cascaded, with the output of one system being fed into the next. The main motivation is that jointly separating and enhancing may be too difficult for a single network to learn, and modularization may allow the networks to focus on specific tasks. Two-stage approaches have previously been explored for denoising plus dereverberation [19,20], separation plus dereverberation [21], and denoising plus separation [11].

The cascaded configurations we consider consist of an optional pre-enhancement system cascaded into a separation network cascaded into an optional post-enhancement system. We evaluate all combinations where noise is removed by either pre-enhancement or the separator, and reverberation is removed by either pre-enhancement, post-enhancement, or the separator. Post-separation denoising is not considered, as separation-without-denoising is a somewhat ill-defined task: noise does not 'belong' to either speech signal, so it is unclear how the network should distribute the noise when not removing it.

For cascaded systems, the sub-models are trained with appropriate input and targets for each sub-task. For example, in the system consisting of denoising followed by separation then dereverberation, the networks are trained as follows: pre-enhancement is trained with noisy reverberant mixtures as input and clean reverberant mixtures as output; the separator with reverberant mixtures as input and reverberant sources as output; and post-enhancement with single reverberant sources as input and single anechoic sources as output.

As mentioned above, due to the scale-invariant loss function, each model's outputs have no constraint to be within any particular dynamic range, and we thus observe strong degradation in performance in cascaded systems when sub-models are trained separately, due to the scaling mismatch between the output of one model and the training data of the next. To address this problem, we scale each output $\hat{s}$, obtained from an input mixture x as an estimate for a target source s, to make it consistent with the scaling of s in x. Because s is unknown, we need to rely on $\hat{s}$ and x alone. If we assume that the interfering signal $n = x - s$ is orthogonal to s, which is generally approximately the case, and that the direction of $\hat{s}$ is close to that of s, then a reasonable choice for the rescaling factor $\beta(\hat{s}|x)$ is that obtained by ensuring that $\beta(\hat{s}|x)\hat{s}$ is orthogonal to the residual $\hat{n} = x - \beta(\hat{s}|x)\hat{s}$. This results in a scaling factor

$$\beta(\hat{s}|x) = \frac{\langle x, \hat{s} \rangle}{\|\hat{s}\|^2}. \tag{2}$$

As the estimate $\hat{s}$ improves (i.e., $\hat{s}$ and s become more colinear), the scaling factor improves as well.

When the best-performing system of a WHAMR! task is a cascaded model, we also evaluate the system with additional end-to-end tuning. Since all component systems are waveform-to-waveform, we can tune the entire system by performing additional training through all cascaded sub-models directly. End-to-end joint training of sub-models has been shown to be successful in joint training of automatic speech recognition with enhancement and separation [22–25].

3.3. Training Configurations

All networks are trained on 4 second segments using the Adam algorithm [26]. The learning rate is decreased by a factor of 2 if validation loss does not improve for 3 consecutive epochs. Gradient clipping is applied with a maximum ℓ_2 norm of 5. Models are trained for 100 epochs with an initial learning rate of 10^{-3}, with the exception of cascaded model tuning, during which we train the models for 25 epochs with a learning rate of 10^{-4}. Because the SI-SDR loss is undefined for silent sources, training models on the *max* data subset is cumbersome, as the 4 s segments randomly sampled during training occasionally fall within regions where only one speaker is talking. Thus, for the 16 kHz *max* condition, we train on 16 kHz *min*. Unless otherwise noted, all results are for the 8 kHz *min* condition.

Table 2. SI-SDR [dB] results for a single separation network. Highlighted rows represent new WHAMR! conditions.

Input			Conv-TasNet		TasNet-BLSTM	
Noise	Reverb	Input	Output	Δ	Output	Δ
		0.0	12.9	12.9	**14.2**	**14.2**
✓		−4.5	7.0	11.5	**7.5**	**12.0**
	✓	−3.3	4.3	7.6	**5.6**	**8.9**
✓	✓	−6.1	2.2	8.3	**3.0**	**9.2**

Table 3. ΔSI-SDR [dB] comparison of our implementations with the best Conv-TasNet number in [5] and the corresponding learned feature configuration of 512 bases, window length 16, window shift 8.

TasNet-BLSTM	Conv-TasNet	Conv-TasNet [5]
16.6	14.4	15.3

4. EXPERIMENTAL RESULTS

For all experiments, we report results using scale-invariant source-to-distortion ratio (SI-SDR) [18], which is also the training objective. Furthermore, because the input SI-SDR between tasks is highly variable, we also report the SI-SDR improvement (Δ), i.e., the difference between output and input SI-SDR.

Table 2 shows the results of our core systems, without cascade. Reverberation seems to be more challenging than noise as reflected by the lower SI-SDR. While the noisy and clean conditions are comparable in terms of SI-SDR improvement, they still differ significantly in terms of raw SI-SDR. Interestingly, we observe consistently better performance from the BLSTM model over the TCN model, which is somewhat unexpected. Indeed, although the BLSTM contains many more parameters than the TCN, this result contradicts prior results in the literature [5, 15]. A comparison of clean separation models with a smaller basis window is shown in Table 3, confirming that the performance difference is not due to the window parameters.

In addition, we note that the TasNet-BLSTM numbers in the first two rows are considerably better than the corresponding numbers in the original WHAM! paper [11]. The newer network uses the same configuration, but is trained with more aggressive gradient clipping and stagnation learning rate adjustment, which supports the findings regarding training optimizer parameters reported in [5, 17].

Table 4 shows experimental results with enhancement networks. We use denoising and dereverberation of two-speaker mixtures as a proxy for all other enhancement conditions. Since performance trends are consistent across these two tasks, we think this is reasonable evidence to conclude that the learned feature BLSTM model (TasNet-BLSTM) is the best architecture for enhancement. While the learned basis TCN and BLSTM perform similarly, we see significant drops in performance moving from learned basis to STFT features. This suggests that the benefits shown in speech separation are also likely present in speech denoising and dereverberation.

Table 5 shows the results of the cascaded model experiments. In accordance with the previous results, all sub-models are TasNet-

Table 4. SI-SDR [dB] for two-speaker enhancement tasks.

Net		Denoise		Dereverb	
Feature	Processor	Output	Δ	Output	Δ
Learned	TCN	10.8	9.6	7.2	3.2
Learned	BLSTM	**11.2**	**10.1**	**8.5**	**4.4**
STFT	TCN	8.4	7.2	4.0	0.0
STFT	BLSTM	9.5	8.4	5.9	1.8
Input SI-SDR:		1.2		4.0	

Table 5. Comparison of cascaded models. A dash indicates speech separation without denoising/dereverberation, while × indicates no enhancement sub-model was used. Results are sorted by increasing performance. The highlighted rows indicate the non-cascaded single-model baseline.

System		SI-SDR	
Pre-Enh. Removes	Separate Speech while Removing	Output	Δ
×	noise	7.5	12.0
noise	–	**8.1**	**12.6**
Input SI-SDR:		−4.5	

(a) noisy condition

System			SI-SDR	
Pre-Enh. Removes	Separate Speech while Removing	Post-Enh. Removes	Output	Δ
×	rev.	×	5.6	8.9
rev.	–	×	6.4	9.7
×	–	rev.	**6.6**	**9.9**
Input SI-SDR:			−3.3	

(b) reverberant condition

System			SI-SDR	
Pre-Enh. Removes	Separate speech while removing	Post-Enh. Removes	Output	Δ
×	noise, rev.	×	3.0	9.2
noise	rev.	×	3.5	9.7
noise, rev.	–	×	3.6	9.7
rev.	noise	×	3.7	9.8
×	noise	rev.	3.7	9.8
noise	–	rev.	**4.0**	**10.1**
Input SI-SDR:			−6.1	

(c) noisy and reverberant condition

BLSTM models. We see that in general, moving the speech enhancement (i.e., denoising and/or dereverberation) tasks to a separate model from separation seems to help performance. From Tables 5(b) and (c), reverberation appears to be particularly difficult for the separation network to remove. We also see that removing reverberation post-separation is slightly better than pre-separation. As two sources will not have the same room impulse response, the dual-source (pre-enhancement) dereverberation network would have to appropriately compensate for two reverberation patterns, while the single-source dereverberation (post-enhancement) network handles only one. The separator network likely has a harder time separating the still-reverberant speech, but this effect appears to be smaller than the difference in single- and double-source dereverberation.

While the cascaded systems do have 2 or 3 times as many parameters as the non-cascaded system, this does not seem to be the sole source of performance improvement, as single models with increased numbers of BLSTM layers provided little performance gain over the results in Table 2. Furthermore, training equivalent cascaded systems from scratch without individual pre-training of the pre-enhancement, separation, and post-enhancement stages provided noticeably less performance improvement over the single network results from Table 2 than the reported cascaded systems in Table 5.

Table 6. SI-SDR comparison of best models with and without additional training. Dashes indicate the best system was not cascaded.

Input			Best System w/o Tuning		Tuned	
Noise	Reverb	Input	Output	Δ	Output	Δ
		0.0	14.2	14.2	–	–
✓		−4.5	8.1	12.6	8.3	12.9
	✓	−3.3	6.6	9.9	7.0	10.3
✓	✓	−6.1	4.0	10.1	4.7	10.8

Table 7. SI-SDR evaluation of 16 kHz conditions using the best model configuration trained on the 16 kHz *min* subset.

Input		16 kHz Min			16 kHz Max		
Noise	Reverb	Input	Output	Δ	Input	Output	Δ
		0.0	12.9	12.9	0.0	12.7	12.7
✓		−4.6	7.8	12.4	−5.8	7.5	13.3
	✓	−3.3	5.6	8.9	−3.4	5.4	8.8
✓	✓	−6.2	3.7	9.9	−7.2	3.5	10.7

Table 6 shows the results of tuning the cascaded systems with additional end-to-end training. Tuning the systems helps, although the performance gains are minor. The noisy and reverberant system, which contains three sub-models in contrast to the others with two, shows the greatest improvement. This suggests training helps with improving the coupling of the connected models.

Table 7 shows the results of our 16 kHz systems. As mentioned earlier, we trained on 16 kHz *min* and evaluated on both the *min* and *max* conditions. Although the performance on 16 kHz data is worse than in the 8 kHz systems, there does not appear to be any significant breakdown in performance. Similarly, performance in the *max* condition is only slightly worse than the *min* condition. Although the SI-SDR improvement in the noisy case is better in *max* than *min*, this is likely due to differences in amount of speech and does not reflect any significant difference in performance.

5. CONCLUSION

We have introduced WHAMR!, an extension of the WHAM! noisy speech separation dataset to include reverberation, with the goal of further promoting the advancement of speech separation technologies towards more realistic conditions. Preliminary results demonstrate that, although noise and reverberation do degrade overall performance, networks with learned basis feature representations are effective not only in separation but also in speech enhancement. We have also demonstrated the value in using cascaded models combining pre-trained separation and enhancement modules, and of further jointly fine-tuning them, establishing strong baseline results for the WHAMR! dataset. Extending the proposed model cascades to stereo is an important topic of future work, and is supported in the WHAMR! scripts available at `http://wham.whisper.ai`.

6. REFERENCES

[1] 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)*, Mar. 2016, pp. 31–35.

[2] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. ISCA Interspeech*, Sep. 2016, pp. 545–549.

[3] 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*, vol. 25, no. 10, pp. 1901–1913, 2017.

[4] 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)*, Apr. 2018.

[5] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[6] Z. Shi, H. Lin, L. Liu, R. Liu, and J. Han, "FurcaNeXt: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks," *arXiv preprint arXiv:1902.04891*, 2019.

[7] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete LDC93S6A," 1993, Web Download. Philadelphia: Linguistic Data Consortium.

[8] S. Bengio and H. Bourlard, *Machine learning for multimodal interaction*. Springer, 2005.

[9] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[10] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*, 1st ed. Wiley Publishing, 2018.

[11] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "WHAM!: Extending speech separation to noisy environments," in *Proc. ISCA Interspeech*, Sep. 2019.

[12] 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. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018.

[13] M. Maciejewski, G. Sell, Y. Fujita, L. P. Garcia-Perera, S. Watanabe, and S. Khudanpur, "Analysis of robustness of deep single-channel speech separation using corpora constructed from multiple domains," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2019.

[14] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 351–355.

[15] 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)*, Apr. 2018.

[16] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 1003–1012.

[17] Y. Luo and N. Mesgarani, "Real-time single-channel dereverberation and separation with time-domain audio separation network." in *Interspeech*, 2018, pp. 342–346.

[18] J. Le Roux, S. T. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019.

[19] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 6, pp. 982–992, 2015.

[20] Y. Zhao, Z.-Q. Wang, and D. Wang, "Two-stage deep learning for noisy-reverberant speech enhancement," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 27, no. 1, pp. 53–62, 2018.

[21] M. Delfarah and D. Wang, "Deep learning for talker-dependent reverberant speaker separation: An empirical study," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 27, no. 11, pp. 1839–1848, 2019.

[22] T. Ochiai, S. Watanabe, T. Hori, J. R. Hershey, and X. Xiao, "Unified architecture for multichannel end-to-end speech recognition with neural beamforming," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1274–1288, Dec. 2017.

[23] S. Settle, J. Le Roux, T. Hori, S. Watanabe, and J. R. Hershey, "End-to-end multi-speaker speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018.

[24] H. Seki, T. Hori, S. Watanabe, J. Le Roux, and J. R. Hershey, "A purely end-to-end system for multi-speaker speech recognition," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, Jul. 2018.

[25] X. Chang, Y. Qian, K. Yu, and S. Watanabe, "End-to-end monaural multi-speaker ASR system without pretraining," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6256–6260.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of the 3rd International Conference on Learning Representations (ICLR)*, 2015.