ANALYSIS OF ROBUSTNESS OF DEEP SINGLE-CHANNEL SPEECH SEPARATION USING CORPORA CONSTRUCTED FROM MULTIPLE DOMAINS

Matthew Maciejewski[1], *Gregory Sell*[2], *Yusuke Fujita*[1,3], *Leibny Paola Garcia-Perera*[1], *Shinji Watanabe*[1], *Sanjeev Khudanpur*[1,2]

[1] Center for Language and Speech Processing, The Johns Hopkins University, USA
[2] Human Language Technology Center of Excellence, The Johns Hopkins University, USA
[3] Hitachi, Ltd. Research & Development Group, Japan

ABSTRACT

Deep-learning based single-channel speech separation has been studied with great success, though evaluations have typically been limited to relatively controlled environments based on clean, near-field, and read speech. This work investigates the robustness of such representative techniques in more realistic environments with multiple and diverse conditions. To this end, we first construct datasets from the Mixer 6 and CHiME-5 corpora, featuring studio interviews and dinner parties respectively, using a procedure carefully designed to generate desirable synthetic overlap data sufficient for evaluation as well as for training deep learning models. Using these new datasets, we demonstrate the substantial shortcomings in mismatched conditions of these separation techniques. Though multi-condition training greatly mitigated the performance degradation in near-field conditions, one of the important findings is that both matched and multi-condition training have significant gaps from the oracle performance in far-field conditions, which advocates a need for extending existing separation techniques to deal with far-field/highly-reverberant speech mixtures.

Index Terms— single-channel speech separation, deep learning, far-field speech

1. INTRODUCTION

A common situation that arises in audio featuring multiple speakers at meeting or casual conversation scenarios is that those speakers will inevitably speak simultaneously [1–4]. This can lead to a breakdown of performance in speech technologies, such as automatic speech recognition (ASR) and speaker identification, as the models are unable to tease apart the speech from the different sources. Speech separation seeks to solve this problem by producing non-overlapping waveforms for each speaker from a recording in which multiple speakers are talking at the same time. Speech separation studies have been initiated from computational auditory scene analysis based on the human auditory system [5] and extended to statistical modeling based on independent component analysis and/or non-negative matrix factorization [6, 7].

Recently, supervised speech separation has become a more powerful alternative due to the development of machine learning techniques, including deep learning [8–15]. These supervised methods require preparation of parallel training data, typically generated synthetically, of overlapping speech mixtures and their corresponding source speech signals or masks. The bulk of recent work conducted on deep-learning based speech separation has been done mainly using the mixture of the Wall Street Journal (WSJ0) [16] corpus. This dataset consists of read speech of news utterances, recorded in a clean environment on close-talking microphones, which are then synthetically added to form overlapped speech [8]. Obviously, this environment is not representative of our practical speech separation scenarios that inevitably include noise and reverberation. This paper investigates the robustness of the speech separation techniques utterance-level Permutation Invariant Training (uPIT) [11], Recurrent Selective Attention Network (RSAN) [14], and Deep Clustering (DPCL) [8] in more realistic environments with multiple and diverse conditions.

To this end, this paper first establishes the process of creating multi-domain datasets which allow clean/noisy and near/far-field comparisons. Since evaluation metrics and model training require ground truth single-speaker speech, we created a system to isolate high-quality single-speaker speech regions from the real multi-talker corpora. Finally, we analyze the robustness of the above three different deep-learning based methods under the varying conditions using the multi-domain datasets.

2. CORPUS SELECTION AND DATA PREPARATION

This section describes the construction of multi-domain datasets of two-speaker mixtures that are effective for quantitative analysis of speech separation techniques: (1) selecting corpora consisting of differing difficulties for speech separation, (2) extracting single-speaker segments from noisy corpora, and (3) generating mixture lists which match the pre-existing WSJ0 dataset as closely as possible. We are releasing the code and resulting mixture lists used in our experiments for reproducibility and use in further studies[1,2].

2.1. Corpus Selection

To assess varying difficulties for speech separation, we selected the WSJ0, CHiME-5 [3], and Mixer 6 [17] speech corpora. The WSJ0 corpus was selected due to the pre-existence of a synthetic overlap dataset [8], a standard of speech separation evaluation. This dataset has been effectively used in a number of speech separation research experiments, and so its composition was the model for our dataset generation pipeline.

The CHiME-5 corpus was chosen to serve as the most challenging, "realistic" condition. The corpus consists of dinner parties

[1] https://github.com/mmaciej2/kaldi/tree/chime5-single-speaker-generation/egs/chime5/single_speaker_generation

[2] https://github.com/mmaciej2/kaldi/tree/mixer6-single-speaker-generation/egs/mixer6/single_speaker_generation

Table 1: Synthetic overlap dataset statistics. 'mean utt. usage' refers to the average number of times a single-speaker segment is used in a synthetic mixture, giving a sense of how much repeated speech is present in overlap mixtures.

corpus	set	spk. count	mix. count	total length	mean utt. usage	mean mix. length
WSJ0	train	101	20k	30.4 h	4.6	5.5 s
	dev.	101	5k	7.7 h	2.8	5.5 s
	eval.	18	3k	4.8 h	3.4	5.8 s
Mixer 6	train	451	20k	28.3 h	1.0	5.1 s
	dev.	50	5k	6.1 h	1.0	4.4 s
	eval.	45	3k	4.1 h	1.0	4.9 s
	train 100k	453	100k	98.3 h	1.3	3.5 s
CHiME-5	train	32	20k	12.7 h	3.4	2.3 s
	dev.	8	5k	3.3 h	8.1	2.4 s

recorded with microphone arrays placed around the apartment as well as binaural microphones, allowing us to generate parallel near-field and far-field datasets with identical utterances. This condition resulted in a number of unique challenges in the audio, such as naturally occurring non-speech noises, multiple simultaneous speakers, and time-varying locations.

The Mixer 6 corpus was chosen to serve as a middle ground between the WSJ0 and CHiME-5 corpora. Including interviews recorded with 14 microphones in a constructed recording room, the Mixer 6 corpus allows a similar near- and far-field comparison, but in a more controlled environment with stationary speakers, consistent channel, and relatively minimal noise.

2.2. Cleanup Methods

To ensure the source data is single-speaker, we used a pipeline implemented with the Kaldi Speech Recognition Toolkit [18] following two stages:

Stage 1) Run a speech activity detection (SAD) system to produce reasonable utterances. The SAD system used is a Time-Delay Neural Network-based system with statistics pooling trained as in [19] with reverberated LibriSpeech [20] data and added noise from MUSAN [21]. The SAD output is then merged with single-speaker region labeling, which comes from the reference transcription for CHiME-5 and an energy-based analysis for the Mixer 6.

Stage 2) Perform segment verification by removing utterances which are too short, have incorrect speaker labels, or are non-speech vocalizations. We used a state-of-the-art speaker identification setup with x-vectors [22] and a probabilistic linear discriminant analysis (PLDA) scoring backend [23,24]. The models were trained using the VoxCeleb [25] and VoxCeleb2 [26] corpora augmented with MUSAN [21] and reverberated with the simulated room impulse responses described in [27]. We scored utterance embeddings against embeddings extracted from all speech by its speaker and rejected the utterances below a qualitatively-tuned score threshold.

2.3. Mixture List Generation

For consistency, we generated the mixture lists to be compatible with the MERL scripts for generating overlap[3] and with similar properties to the WSJ0 mixtures. Still, there was a lot of freedom in how to pair utterances from the base corpus to generate mixtures. As a result, we created the mixture lists algorithmically according to a set of desirable criteria selected to maximize data diversity and efficiency:

1. avoiding mixtures of two utterances by the same speaker
2. minimizing repeated usage of single utterances
3. maximizing speaker diversity within pairs using an utterance
4. pairing utterances of similar length

We selected two microphone conditions from each corpus for use in our experiments. For the far-field CHiME-5 condition, we selected the first channel of the first microphone array. For the near-field CHiME-5 condition, we used the left channel of the binaural microphone for the speaker corresponding to each utterance. For the far-field Mixer 6 condition, we chose channel 9, which is the microphone placed farthest from the speaker. For the near-field Mixer 6 condition, we chose channel 2, which is the lapel microphone for the subject.

2.4. Overlap Dataset Design

In constructing the CHiME-5 and Mixer 6 mixture data, an attempt was made to match the WSJ0 mixture data as closely as possible. The most natural correspondence was to construct train, development, and test sets of equivalent size (20k, 5k, and 3k mixtures respectively). We chose mixture energy ratio levels following the same distribution as well. However, because the size of each source corpus varied, the usage of each speaker and utterance varied as well. Comparison of usage statistics are in Table 1.

In both the CHiME-5 and Mixer 6 mixture datasets, we constructed both near-field and far-field conditions. When doing so, we used identical utterance pairs, as opposed to generating new mixture sets, to reduce the number of confounding factors when comparing performance between near-field and far-field conditions.

In addition, due to the extensive size of the Mixer 6 corpus in comparison to the WSJ0 and CHiME-5 corpora, we were able to construct additional, larger training sets for both Mixer 6 conditions. In these datasets the total size of the training data was increased five-fold to 100k (train 100k), allowing us to do deeper analysis in how quantity of training data affects the model performance.

Finally, we constructed a training set of equivalent size by combining each of the five base datasets (WSJ0, CHiME-5 near and far, Mixer 6 near and far). Two iterations were created. In the first, the combinations were sub-sampled to maintain the size of 20k training examples. In the second, they were fully combined, resulting in 100k examples. These sets allowed us to analyze the potential for producing a robust system based on training on a wide variety of properly manicured data.

3. SPEECH SEPARATION TECHNIQUES

This section describes three representative deep learning mask-based techniques, uPIT [11], RSAN [14], and DPCL [8]. These methods all use bi-directional long short-term memory (BLSTM) recurrent neural networks to process input mixture magnitude spectra with the ultimate goal of producing a spectral mask for each speaker. The mask is applied to the mixture spectrogram, which is then inverted to reconstruct the estimated source waveforms. However, the way in which these masks are generated in each method differs.

[3] http://www.merl.com/demos/deep-clustering/

The uPIT and RSAN networks produce estimated speaker soft masks directly, relying on some mechanism to solve the permutation problem of an order of output masks while training. They are essentially trained with the mean squared error loss between the estimated (masked) and ground truth source magnitude spectra.

$$\text{Loss}^{\text{uPIT,RSAN}}(\hat{\mathbf{M}}, \pi) = \frac{1}{TFS} \sum_{s=1}^{S} \|\hat{\mathbf{M}}_s \circ \mathbf{A}_{mix} - \mathbf{A}_{\pi_s}\|_F^2, \quad (1)$$

where $\hat{\mathbf{M}}_s \in [0,1]^{T\times F}$ is the estimated mask on source s, and $\mathbf{A}_{mix}, \mathbf{A}_s \in \mathbb{R}_{\geq 0}^{T\times F}$ are the short-time Fourier transform (STFT) magnitudes for the mixture and source s, respectively. The summation over S represents the different sources in a mixture. T and F denote the numbers of frames and frequency bins, respectively. π is the permuted source sequence of oracle magnitude spectra, chosen to match the sequence of estimated masks, where π_s returns the s-th element of π, i.e. the ground truth source index matching the s-th estimated mask.

In the uPIT method, the network is set up to produce a fixed number of output masks, where the permutation problem is solved by scoring the output masks against all possible orders of source spectra, and only performing backpropagation on the source order $\hat{\pi}^{\text{uPIT}}$ that results in the lowest loss:

$$\hat{\pi}^{\text{uPIT}} = \arg\min_{\pi\in\mathcal{P}} \text{Loss}^{\text{uPIT,RSAN}}(\hat{\mathbf{M}}, \pi) \quad (2)$$

where $\mathcal{P}$ represents all possible permutations of sources within a mixture.

In the RSAN method, each pass through the network produces only a single output mask. The network takes an "attention mask" as input in addition to the mixture spectra. By subtracting each estimated source mask from the attention mask, the mixture can be passed through the network multiple times, being forced to attend to a new source each time. Due to the difference in functionality to the uPIT method, in the RSAN network the permutation problem is solved in a greedy manner, with each successive mask $\hat{\mathbf{M}}_s$ being paired with the remaining unpaired oracle speaker $\hat{\pi}_s^{\text{RSAN}}$ that produces the lowest loss:

$$\begin{aligned} \hat{\pi}_s^{\text{RSAN}} &= \arg\min_{s'\in\tilde{S}} \|\hat{\mathbf{M}}_s \circ \mathbf{A}_{mix} - \mathbf{A}_{s'}\|_F^2 \\ \tilde{S} &= \{1,\ldots,S\} \setminus \{\hat{\pi}_{1:s-1}\}, \end{aligned} \quad (3)$$

where $\tilde{S}$ is the subset of the original source set created by removing the previously selected source indices $\{\hat{\pi}_{1:s-1}\}$.

The DPCL method uses the network to produce an embedding vector for each STFT coefficient, which can then be clustered to produce hard binary masks. The loss function used in DPCL is based on the squared Frobenius norm between oracle and estimated affinity matrices:

$$\text{Loss}^{\text{DPCL}}(\mathbf{V}) = \|\mathbf{V}\mathbf{V}^\top - \mathbf{Y}\mathbf{Y}^\top\|_F^2 \quad (4)$$

The affinity matrix can be generated using the outer product of a matrix $\mathbf{Y} \in \{0,1\}^{(TF)\times S}$ with itself, where the s-th column of $\mathbf{Y}$ is a TF-dimensional binary vector encoding the STFT coefficients belonging to source s. The affinity matrix is estimated with the self-outer product of a matrix $\mathbf{V} \in \mathbb{R}^{(TF)\times D}$ that is produced by the DPCL network, consisting of a D-dimensional embedding vector for each STFT coefficient.

Table 2: Comparison of experimental setup on SDR improvement with the WSJ0 2-speaker mixture dataset.

	method	SDRi
field-reported	uPIT-BLSTM-ST [11]	10.0
	RSAN [14]	8.6
	DPCL [8]	5.8
	DPCL++ [9]	10.8
experimental	uPIT	9.3
	RSAN	9.5
	DPCL	7.7

4. EXPERIMENTS

4.1. Experimental Setup

The spectrograms were generated using a STFT from downsampled 8 kHz audio with a window length of 512 and a step of 128 in the case of the uPIT and RSAN experiments, and a window length of 256 and a step of 64 in the DPCL setup. The input to the networks was the mixture magnitude spectrum. The input speech was a mixture of two speakers, and the systems always output exactly two masks (i.e., $S = 2$ in Section 3).

Both the uPIT and RSAN networks used in our experiments consist of two 600-node BLSTM layers followed by a linear layer, with a sigmoid output. The uPIT network has an input dimension F of 257 with a final output of 514 for two speaker masks, while the RSAN network has an input dimension of 514 to account for the attention mask, with a final output of 257, and is run twice to recursively extract two speaker masks. The DPCL network used in our experiments also used two 600-node BLSTM layers followed by a linear layer, with hyperbolic tangent and ℓ_2-normalization. The input dimension F was 129, and the output dimension was 5,160, corresponding to an embedding dimension of 40 (i.e., $D = 40$). The backend used in the DPCL setup to produce masks was k-means clustering with cosine distance between embedding vectors with $k = 2$ (two speakers). All networks were trained for 200 epochs with an initial learning rate of 0.001 using the Adam [28] optimizer.

For evaluation, we used three standard evaluation metrics: signal to distortion ratio (SDR), signal to interferences ratio (SIR), and signal to artifacts ratio (SAR) [29] implemented in the mir_eval library [30]. Our primary, and most typical speech separation metric, was SDR, while we additionally used SIR and SAR for our initial experimental comparisons. We also provide the SDR of the unprocessed corpus for reference and for computation of SDR improvement.

4.2. Results and Analysis

We analyzed the robustness of the speech separation techniques based on our implementations of RSAN, uPIT, and DPCL with multiple datasets, as introduced in Section 2. We used widely-reported SDR improvement on the WSJ-2mix [8] dataset to verify that our implementations used are within the range of state-of-the-art performance, reflected in other reports, as shown in Table 2.

Matched and Mismatched Conditions

Results of experiments containing models trained purely on in-domain data are presented in Table 3. Overall, performance de-

Table 3: Comparison of SDR, SIR, and SAR in matched-condition training and evaluation sets

dataset		uPIT					RSAN					DPCL				
		wsj	mx6 near	ch5 near	mx6 far	ch5 far	wsj	mx6 near	ch5 near	mx6 far	ch5 far	wsj	mx6 near	ch5 near	mx6 far	ch5 far
metric	**SDR**	9.41	6.86	7.12	4.05	2.21	9.69	6.92	7.20	3.53	1.78	7.80	3.24	2.37	-3.11	-2.85
	SIR	14.18	10.28	10.10	5.90	4.10	14.45	10.54	10.47	5.57	3.76	16.32	9.94	8.46	2.70	3.58
	SAR	11.79	10.72	11.32	10.40	9.30	12.04	10.77	11.21	10.69	10.35	10.31	8.59	9.13	8.64	8.45

Table 4: SDR with 20k-mixture train sets and varying test conditions. To emphasize the difference between near and far conditions, the numbers greater than 5.0 are highlighted, with boldface used for the best result per evaluation condition.

eval		RSAN				
		wsj	mx6 near	ch5 near	mx6 far	ch5 far
train	wsj	**9.69**	4.99	6.12	1.05	1.13
	mx6 near	7.53	**6.92**	7.01	2.23	1.01
	ch5 near	7.13	5.64	**7.20**	2.31	1.70
	mx6 far	2.38	2.99	3.53	**3.53**	0.45
	ch5 far	1.53	-0.55	1.13	-1.14	**1.78**
oracle		13.98	13.18	13.06	9.59	10.86
corpus		0.15	0.17	0.33	0.34	0.32

Table 5: 20k- and 100k-mixture train sets SDR comparison

eval		RSAN				
		wsj	mx6 near	ch5 near	mx6 far	ch5 far
train 20k	mx6 near	7.53	6.92	7.01	2.23	1.01
	mx6 far	2.38	2.99	3.53	3.53	0.45
	combo	7.45	5.45	6.22	2.84	2.15
train 100k	mx6 near	7.99	7.48	7.53	2.67	1.63
	mx6 far	3.39	3.64	4.24	4.49	1.25
	combo	9.01	6.80	7.53	4.08	3.07
oracle		13.98	13.18	13.06	9.59	10.86
corpus		0.15	0.17	0.33	0.34	0.32

grades as we move from clean to more noisy conditions as well as from near- to far-field. We also see that the uPIT and RSAN networks produce similar results due to their similar separation framework, as discussed in Section 3, while the DPCL shows stronger degradation. This may be due to a lack of tuning the speech/noise threshold parameter, and also is not representative of the more advanced and better-performing DPCL++ [9] method reflected in Table 2, which includes improvements to signal reconstruction and soft masking. Similar trends are reflected across all three metrics, so we chose to report only the standard SDR metric for all other experiments. For similar reasons, we restrict our results to the RSAN method, chosen due to being one of our best-performing methods.

Experimental results of all train–test configurations using training sets of size 20k are shown in Table 4. Interestingly, the dataset mismatch among clean and near-field conditions did not cause a serious degradation despite the noisy and speaking-style variations across the datasets. However, we observed a large degradation in any combination of training and test data when including far-field conditions. Although the oracle performance computed from the ideal ratio mask, shown in Table 4, presents intrinsic difficulties of far-field conditions compared to near-field conditions, the observed degradation in using speech separation was even greater.

Effect of Multi-Condition Training and Training Data Size

We see that the training conditions comprised of a combination of all corpora (combo), presented in Table 5, result in performance near that of matched training for each condition. Increasing the amount of training data five-fold (train 100k combo) improves performance further. This result suggests multi-condition training, which is widely used in speech processing, is still effective for deep-learning based speech separation. However, the performance in far-field conditions is quite poor, even with multi-condition training or increased quantity of training data.

From our experiments, we can conclude that current speech separation techniques are reasonably robust across the datasets in near-field conditions. However, these experiments also reveal that both matched and multi-condition training have significant degradation in far-field conditions, a differing result from other learning-based speech processing, notably automatic speech recognition [31, 32].

5. CONCLUSIONS

In this work we demonstrated shortcomings of supervised speech separation techniques in mismatched conditions which were not captured by previous evaluation conditions. We provided new synthetic overlap datasets that expand the domain of single-channel speech separation evaluation from clean, near-field conditions to far-field, realistic conversational speech. Using these datasets, we showed that performance degradation in far-field conditions is largely unsolved. Though the lack of robustness can be mitigated by training models on more data from multiple conditions, there remains a significant gap from the oracle performance in far-field conditions, which advocates a need for extending separation techniques to deal with far-field speech mixtures.

Acknowledgments

We would like to thank Xuankai Chang for providing comments useful to producing the final version of this manuscript.

6. REFERENCES

[1] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, *et al.*, "The ICSI meeting corpus," in *Proc. of ICASSP*, vol. 1, 2003, pp. 364–367.

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

[3] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The Fifth CHiME Speech Separation and Recognition Challenge: Dataset, task and baselines," in *Proc. of Interspeech*, 2018.

[4] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First DIHARD challenge evaluation plan," 2018, tech. Rep., 2018. Available: https://zenodo.org/record/1199638.

[5] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

[6] S. Makino, T.-W. Lee, and H. Sawada, *Blind speech separation*. Springer, 2007, vol. 615.

[7] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[8] J. R. Hershey, J. Le Roux, Z. Chen, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. of ICASSP*, 2016, pp. 31–35.

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

[10] D. Yu, M. Kolbæk, Z. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multitalker speech separation," in *Proc. of ICASSP*, 2017, pp. 241–245.

[11] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen, "Multitalker 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, Oct 2017.

[12] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. of ICASSP*, 2017, pp. 246–250.

[13] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787–796, 2018.

[14] K. Kinoshita, L. Drude, M. Delcroix, and T. Nakatani, "Listening to each speaker one by one with recurrent selective hearing networks," in *Proc. of ICASSP*, 2018, pp. 5064–5068.

[15] Y. Luo and N. Mesgarani, "TasNet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. of ICASSP*, 2018, pp. 696–700.

[16] J. Garofolo, D. Graff, D. Paul, and D. Pallett, *CSR-I (WSJ0) Complete LDC93S6A*. Philadelphia: Linguistic Data Consortium, 1993.

[17] L. Brandschain, D. Graff, and K. Walker, *Mixer-6 Speech LDC2013S03*. Philadelphia: Linguistic Data Consortium, 2013.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannermann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi Speech Recognition Toolkit," in *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[19] P. Ghahremani, V. Manohar, D. Povey, and S. Khudanpur, "Acoustic modelling from the signal domain using CNNs." in *Proc. of Interspeech*, 2016, pp. 3434–3438.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. of ICASSP*, 2015, pp. 5206–5210.

[21] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.

[22] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition Using Data Augmentation," in *Proc. of ICASSP*, 2018.

[23] S. J. D. Prince and J. H. Elder, "Probabilistic Linear Discriminant Analysis for Inferences About Identity," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[24] S. Ioffe, "Probabilistic Linear Discriminant Analysis," in *ECCV*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Springer-Verlag, 2006, pp. 531–542.

[25] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *Proc. of Interspeech*, 2017.

[26] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. of Interspeech*, 2018.

[27] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. of ICASSP*, 2017, pp. 5220–5224.

[28] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[29] E. Vincent, R. Gribonval, and C. Févotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Transactions on Acoustics, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[30] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, "MIR_EVAL: A transparent implementation of common MIR metrics," in *ISMIR*, 2014.

[31] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, *et al.*, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.

[32] B. Li, T. N. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Pundak, K. K. Chin, *et al.*, "Acoustic modeling for google home." in *Proc. of Interspeech*, 2017, pp. 399–403.