INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

A Semi-blind Source Separation Approach for Speech Dereverberation

Ziteng Wang, Yueyue Na, Zhang Liu, Yun Li, Biao Tian, Qiang Fu

Machine Intelligence Technology, Alibaba Group
{ziteng.wzt, tianbiao.tb}@alibaba-inc.com

Abstract

This paper presents a novel semi-blind source separation approach for speech dereverberation. Based on a time independence assumption of the clean speech signals, direct sound and late reverberation are treated as separate sources and are separated using the auxiliary function based independent component analysis (Aux-ICA) algorithm. We show that the dereverberation performance is closely related to the underlying source probability density prior and the proposed approach generalizes to the popular weighted prediction error (WPE) algorithm, if the direct sound follows a Gaussian distribution with time-varying variances. The efficacy of the proposed approach is fully validated by speech quality and speech recognition experiments conducted on the REVERB Challenge dataset.

Index Terms: speech dereverberation, blind source separation, REVERB challenge

1. Introduction

In many speech processing applications, the microphone signal is degraded by reverberation. Reverberation is caused by the cumulation of multiple reflections in the acoustic enclosure when the signal travels from source to the sensor. The reverberant signal is often expressed as a linear convolution of the source signal and an acoustic impulse response (AIR) relating the source pose and the microphone pose. A schematic illustration of an AIR is provided in Figure 1. The AIR is divided into three successive parts: the direct path sound as the first peak, followed by early echos, and a collection of many reflected sounds which is termed late reverberation. While early echos are found beneficial for human perception and even for automatic speech recognition (ASR), late reverberation is generally detrimental and needs to be suppressed [1, 2].

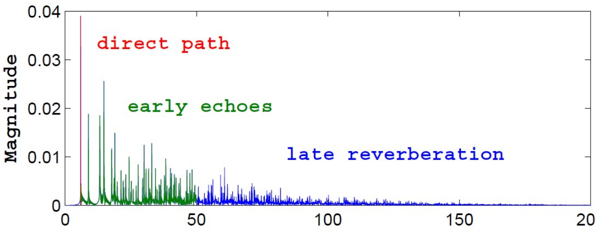

Figure 1: *Magnitude of an example acoustic impulse response in time (ms).*

Speech dereverberation algorithms in the literature broadly fall into two categories, one based on spectral enhancement and the other based on linear filtering. The first category of algorithms design a varying spectral gain to be applied to the signal spectrum based on, e.g. estimated late reverberation power [3], coherent-to-diffuse ratio (CDR) [4], or pretrained deep neural networks [5, 6]. The linear filtering based speech dereverberation algorithms include spatial beamforming approaches [7, 8], Kalman filter based approaches [9] and multi-channel linear prediction (MCLP) approaches [10, 11]. In the recent REVERB Challenge [1] and CHiME challenges [12], an efficient implementation of MCLP in the short time Fourier transform (STFT) domain, often referred to as the weighted prediction error (WPE) algorithm [10], consistently improved distant speech recognition accuracy and gained popularity over time [13, 14].

In [10], the reverberant signal is modeled as an autoregressive process and a delayed multi-channel linear predictor is used to predict the desired clean speech. The predictor is optimized under a maximum likelihood criterion assuming that the direct sound follows a Gaussian distribution with time-varying variances. It is claimed that the speech signal model or source *prior* plays an essential part in the dereverberation performance. Different speech signal models, e.g. Laplacian distribution [15], complex generalized Gaussian distribution [16, 17], and Student's t-distribution [18] have later been investigated to improve the performance of MCLP algorithms.

Speech dereverberation could alternatively be addressed from a blind source separation perspective, as have been proposed in the TRINICON [19] algorithm and in [20]. In [20], speech dereverberation is taken as a sub-problem in a unified source separation framework for joint dereverberation and echo cancellation. Clean speech is separated from the echoed and reverberant mixture based on independent component analysis (ICA). This work could be further explored by considering the later widely used auxiliary function based (Aux-)ICA approaches [21, 22]. Specifically, we propose an Aux-ICA solution to the speech dereverberation problem in this paper. Then, for the first time, we mathematically relates the seemingly different source separation based dereverberation and the MCLP based dereverberation algorithms, by applying the block matrix inverse formula [23] on the proposed solution. Blind source separation naturally relies on the underlying source probability density function (PDF) and it is straightforward to see different source priors leading to different dereverberation performance. If a non-stationary Gaussian PDF is assumed for the desired source, the proposed approach generalizes to the classical WPE algorithm.

The reminder of this paper is organized as follows. Speech dereverberation is reformulated as a semi-blind source separation problem in Section 2. The proposed Aux-ICA solution is then presented in Section 3, along with its relation to previous work and WPE. The experiments are in Section 4 and conclusions are drawn in Section 5.

2. Problem formulation

Consider an acoustic scenario where a single speech source is captured by M microphones. Let $s_{t,f}$ denote the clean speech signal in the STFT domain with time index t and frequency bin index f. The reverberant signal observed at the m-th micro-

http://dx.doi.org/10.21437/Interspeech.2020-1307

phone can be represented using a convolutive transfer function model as

$$x_{m,t,f} = \sum_{\tau=0}^{L_h-1} h_{m,\tau,f} s_{t-\tau,f} + e_{m,t,f} \tag{1}$$

where $h_{m,\tau,f}$ with length L_h time frames models the frequency domain transfer function between the speech source and the m-th microphone, and $e_{m,t,f}$ represents the modeling error and ambient noise. As in [10], by assuming the addictive term $e_{m,t,f} = 0$, the signal at an arbitrarily chosen microphone (e.g. $m = 1$) can be written in the MCLP form as

$$x_{t,f} = d_{t,f} + \sum_{m=1}^{M} \sum_{\tau=0}^{L-1} a^*_{m,\tau,f} x_{m,t-\triangle-\tau,f} \tag{2}$$

where L is the order of the delayed multi-channel linear predictor $a_{m,\tau,f}$, $\triangle$ is the prediction delay and $(\cdot)^*$ denotes conjugate. The prediction delay is determined by the chosen boundary between the desired speech signal $d_{t,f}$ and late reverberation.

The signal model in (2) can be written in matrix notation as

$$\begin{bmatrix} x_{t,f} \\ \mathbf{x}_{\triangle,f} \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{a}_f^H \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} d_{t,f} \\ \mathbf{x}_{\triangle,f} \end{bmatrix} \tag{3}$$

where

$$\begin{aligned} \mathbf{x}_{\triangle,f} = [&x_{1,t-\triangle,f}, ..., x_{1,t-\triangle-L+1,f}, ... \\ &x_{M,t-\triangle,f}, ..., x_{M,t-\triangle-L+1,f}]^T \end{aligned} \tag{4}$$

is the vector of STFT samples delayed by $\triangle$, **0** is a zero vector of length ML, **I** is a unit matrix of order ML and

$$\begin{aligned} \mathbf{a}_f = [&a_{1,0,f}, ..., a_{1,L-1,f}, ... \\ &a_{M,0,f}, ..., a_{M,L-1,f}, ...]^T \end{aligned} \tag{5}$$

is the MCLP prediction coefficient vector. $(\cdot)^T$ denotes transpose and $(\cdot)^H$ denotes Hermitian transpose. Equation (3) clearly represents a non-singular mixing process from the source separation perspective, so the desired speech signal can be separated from the observation vector as

$$\begin{bmatrix} \hat{d}_{t,f} \\ \mathbf{x}_{\triangle,f} \end{bmatrix} = \mathbf{B}_f \begin{bmatrix} x_{t,f} \\ \mathbf{x}_{\triangle,f} \end{bmatrix} \tag{6}$$

where $\hat{(\cdot)}$ denotes the estimate of a variable, and $\mathbf{B}_f$ is termed the unmixing matrix.

Speech dereverberation is now reformulated by (3) and (6) as a semi-blind source separation problem, since $\mathbf{x}_{\triangle,f}$ is already known. Furthermore, by assuming that $\{d_{t,f}, \mathbf{x}_{\triangle,f}\}$ are mutually independent, the unmixing matrix $\mathbf{B}_f$ can be uniquely determined by

$$\mathbf{B}_f = \begin{bmatrix} 1 & \mathbf{b}_f^H \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \tag{7}$$

with

$$\begin{aligned} \mathbf{b}_f = [&b_{1,0,f}, ..., b_{1,L-1,f}, ... \\ &b_{M,0,f}, ..., b_{M,L-1,f}, ...]^T. \end{aligned} \tag{8}$$

The independence assumption is valid if $\{d_{t,f}\}$ are assumed independent across time, and a proof of this independence exchange lemma can be found in [20].

3. The proposed approach

3.1. The Aux-ICA solution

To estimate the unmixing matrix, an independence measure is firstly defined by employing the Kullback-Leibler divergence as

$$\begin{aligned} J(\mathbf{B}_f) &= \int_{d_{t,f}} \int_{\mathbf{x}_{\triangle,f}} p(d_{t,f}, \mathbf{x}_{\triangle,f}) \log \frac{p(d_{t,f}, \mathbf{x}_{\triangle,f})}{q(d_{t,f}, \mathbf{x}_{\triangle,f})} \\ &= H(d_{t,f}) + H(\mathbf{x}_{\triangle,f}) - H(d_{t,f}, \mathbf{x}_{\triangle,f}) \\ &= const. + E[G(d_{t,f})] - \log|\det \mathbf{B}_f| \end{aligned} \tag{9}$$

where $p(\cdot)$ represents the source PDF, $q(\cdot)$ the product of approximated PDF of individual sources, $H(\cdot)$ the entropy function, and $E[\cdot]$ denotes the expectation operation. $G(d_{t,f})$ is called the contrast function and has a relationship $G(d_{t,f}) = -\log p(d_{t,f})$.

Minimizing (9) is a nonlinear optimization problem. In the auxiliary function technique, a function $Q(\mathbf{B}_f, \mathbf{C}_f)$ is designed such that

$$J(\mathbf{B}_f) = \min_{\mathbf{C}_f} Q(\mathbf{B}_f, \mathbf{C}_f). \tag{10}$$

Then instead of directly minimizing the objective function $J(\mathbf{B}_f)$, the auxiliary function $Q(\mathbf{B}_f, \mathbf{C}_f)$ is minimized in terms of $\mathbf{B}_f$ and $\mathbf{C}_f$, alternatively.

If we assume a general super-Gaussian PDF of the clean speech signal, which requires that the contrast function $G(r)$ is continuous and $G'(r)/r$ is monotonically decreasing in $r \geq 0$, with $(\cdot)'$ denoting the derivative operator. Note that the speech signal models introduced in [16, 18] all satisfy this assumption. Then the following inequality

$$G(d_{t,f}) \leq \frac{G'(r_0)}{2r_0}|d_{t,f}|^2 + G(r_0) - \frac{r_0 G'(r_0)}{2} \tag{11}$$

holds for any $d_{t,f}$ and r_0 [21]. The equality sign is satisfied if and only if $r_0 = |d_{t,f}|$. On the basis of (11), we have the auxiliary function

$$Q(\mathbf{B}_f, \mathbf{C}_f) = \sum_{i=1}^{ML+1} \mathbf{w}_{i,f}^H \mathbf{C}_{i,f} \mathbf{w}_{i,f} + const. \tag{12}$$

where $\mathbf{w}_{i,f}^H$ is the i-th row vector of $\mathbf{B}_f$ and the auxiliary variable $\mathbf{C}_{i,f}$ is defined by

$$\mathbf{C}_{i,f} = E[\frac{G'(r_{i,t,f})}{r_{i,t,f}} \mathbf{x}_{t,f} \mathbf{x}_{t,f}^H] \tag{13}$$

with $\mathbf{x}_{t,f} = [x_{t,f}, \mathbf{x}_{\triangle,f}^T]^T$ the observation vector and $r_{i,t,f} = |\mathbf{w}_{i,f}^H \mathbf{x}_{t,f}|$ the i-th separated source.

For the semi-blind source separation task at hand, only the first demixing row vector needs to be estimated. The update rule is thus obtained by minimizing $Q(\mathbf{B}_f, \mathbf{C}_f)$ in terms of $\mathbf{w}_{1,f}$ as

$$\mathbf{w}_{1,f} = [\mathbf{B}_f \mathbf{C}_{1,f}]^{-1} \mathbf{i}_1 \tag{14}$$

$$\mathbf{w}_{1,f} \leftarrow \mathbf{w}_{1,f} / \mathbf{w}_{1,f,1} \tag{15}$$

where $\mathbf{i}_1 = [1, 0, ..., 0]^T$ is a $ML + 1$ dimensional vector and $\mathbf{w}_{1,f}$ is normalized with respect to its first element. (14) could be further simplified as

$$\begin{aligned} \mathbf{w}_{1,f} &= \mathbf{C}_{1,f}^{-1} \mathbf{B}_f^{-1} \mathbf{i}_1 \\ &= \mathbf{C}_{1,f}^{-1} \mathbf{i}_1. \end{aligned} \tag{16}$$

Consequently, the proposed Aux-ICA solution consists of iteratively applying (13)(16)(15) and (6) in order until convergence.

3.2. Relation to prior work

For now, the Aux-ICA solution for speech dereverberation seems different from the MCLP based approaches, though they are based on the same signal mode in (2). Their relationship is firstly revealed as in the following.

Rewrite the auxiliary variable $\mathbf{C}_{1,f}$ in block form as

$$\mathbf{C}_{1,f} = \begin{bmatrix} c_{11} & \mathbf{r}_{\triangle,f}^H \\ \mathbf{r}_{\triangle,f} & \mathbf{R}_{\triangle,f} \end{bmatrix} \tag{17}$$

where

$$\begin{aligned} \mathbf{R}_{\triangle,f} &= E[\frac{G'(r)}{r}\mathbf{x}_{\triangle,f}\mathbf{x}_{\triangle,f}^H], \\ \mathbf{r}_{\triangle,f} &= E[\frac{G'(r)}{r}\mathbf{x}_{\triangle,f}x_{t,f}^*]. \end{aligned} \tag{18}$$

Then apply block matrix inversion

$$\mathbf{C}^{-1} = \begin{bmatrix} (c_{11} - \mathbf{r}^H\mathbf{R}^{-1}\mathbf{r})^{-1} & -(c_{11} - \mathbf{r}^H\mathbf{R}^{-1}\mathbf{r})^{-1}\mathbf{r}^H\mathbf{R}^{-1} \\ -\mathbf{R}^{-1}\mathbf{r}(c_{11} - \mathbf{r}^H\mathbf{R}^{-1}\mathbf{r})^{-1} & \mathbf{R}^{-1} + \mathbf{R}^{-1}\mathbf{r}(c_{11} - \mathbf{r}^H\mathbf{R}^{-1}\mathbf{r})^{-1}\mathbf{r}^H\mathbf{R}^{-1} \end{bmatrix}. \tag{19}$$

And by taking (19) into (16) and (15), we have the demixing coefficients

$$\mathbf{b}_f = -\mathbf{R}_{\triangle,f}^{-1}\mathbf{r}_{\triangle,f} \tag{20}$$

while the separated desired source is given by

$$\hat{d}_{t,f} = x_{t,f} + \mathbf{b}_f^H\mathbf{x}_{\triangle,f}. \tag{21}$$

The consequent solution now consists of applying (18)(20) and (21) iteratively until convergence, the same as in the MCLP approaches [10, 11], but different in the correlation weighting factor, which is determined by the underlying source prior. If the speech signal follows a non-stationary Gaussian PDF as in the WPE algorithm [10]

$$p(d_{t,f}) \sim \mathcal{N}_{\mathbb{C}}(d_{t,f}; 0, \lambda_{t,f}) \propto \exp(-\frac{|d_{t,f}|^2}{\lambda_{t,f}}), \tag{22}$$

then we get $G'(r)/r = 1/\lambda_{t,f}$ with $\lambda_{t,f}$ the time-varying variance. Here the Aux-ICA solution and WPE turns out equivalent only mathematically, because (22) is not super-Gaussian.

The auxiliary function technique was once applied in [11] to derive a generalized WPE algorithm, but under a different measure called the *Hadamard-Fischer mutual correlation*, assuming a clean speech signal has auto-correlation coefficients of nearly zero for time lags larger than tens of milliseconds. The mathematical equivalence derived here is not too surprising, given the auxiliary function technique turns both the original objective functions in [11] and (9) into quadratic forms that depends on the second-order statistics of the observed signal. Nevertheless, the Aux-ICA solution highlights the roll of statistical modeling of speech, which has been persistently studied in the literature [24, 25], and has recently been found important also in echo cancellation [26] and beamforming [27].

We consider one unified contrast function

$$G(d_{t,f}) = (\frac{d_{t,f}}{\eta})^\beta \tag{23}$$

which covers most speech signal models in the literature, with $\eta > 0$ and $0 < \beta \leq 2$ the scaling and shape parameters. A choice of $\beta \in [0.2, 0.4]$ is suggested for speech separation tasks [25].

4. Experiments

Speech dereverberation experiments are conducted on the the REVERB Challenge dataset [1]. The scenario is listening to a single stationary distant-talking speaker in reverberant rooms. The SimuData test set includes three typical reverberant conditions: a small room, a medium-size room and a large-size room with RT60 0.25s, 0.5s, and 0.7s respectively. Clean utterances are convolved with measured AIRs at two source-microphone distances (near = 0.5m and far = 2.0m). Real-recorded background noise is then added to the simulated data at signal-to-noise ratio (SNR) of 20dB. The RealData test set is collected in a different meeting room with reverberation time of 0.7s, where source-microphone distances are set at approximated 1.0m (near) and 2.5m (far). All the utterances are provided with 1-channel (1ch), 2-channel (2ch) and 8-channel (8ch) formats.

The proposed Aux-ICA speech dereverberation algorithm is evaluated with typical contrast function parameters $\eta = 1$ and $\beta = 0.2, 0.4, 1$ in (22). STFT is performed in 512 length and 128 samples shift. The MCLP signal model parameters are set $\triangle = 3$ and $L = 10$, following that of WPE optimized on this dataset [1]. Both the Aux-ICA algorithm and the baseline WPE algorithm perform utterance-based processing for 5 iterations.

4.1. Quality evaluation

Several objective measures, including cepstrum distance, log likelihood ratio, frequency-weighted segmental SNR, and speech-to-reverberation modulation energy ratio (SRMR) [28] are recommended for objective evaluation in the challenge. The SRMR scores are finally reported in Table 1, for its strong correlation to the dereverberation task at hand.

Compared with the unprocessed recordings, SRMR gains are observed on all the processed data. The best gains are respectively 0.41 dB, 0.96 dB and 1.92 dB on the 1ch, 2ch and 8ch RealData. The capability of the speech dereverberation algorithms clearly increases as more microphones are available. For each room, higher gains are achieved in the far-distance cases than that in the near-distance cases. The objective scores assuming different source PDF shape parameters are close. Nevertheless, the choice of $\beta = 0.4$ gives the best results on all the real datasets and most of the simulated datasets, which extends the argument made specifically for speech separation in [25] to speech dereverberation tasks.

4.2. Recognition evaluation

A state-of-the-art speech recognition system is built up using time delay neural networks (TDNNs) for acoustic modeling. The word error rates (WERs) are summarized in Table 2.

There exist clear gaps between the results on SimData and that on RealData. On the simulated data, the difference between assuming different source priors are not significant. In the following analysis, we focus on the RealData part, which is more indicative of the real performance of the evaluated algorithms. The Aux-ICA based speech dereverberation algorithm shows superiority on the 1ch and 2ch test cases, on average reducing the WER from 19.41% to 16.90% and 15.90%, respectively. While on the 8ch RealData, the vanilla WPE algorithm performs best and the error rate differences are 0.09% and 0.44% compared with the Aux-ICA ($\beta = 0.2$) algorithm. Checking again with the speech quality tests, the general trend is that higher SRMR scores relate to lower recognition errors.

[1] The WPE implementation and the speech recognition pipeline are available at https://github.com/kaldi-asr/kaldi/tree/master/egs/reverb/s5

Table 1: *SRMR scores under different source probability density function assumptions. Bold figures indicates the best performance in each test condition.*

		SimData							RealData		
		Room1		Room2		Room3			Room1		
		Near	Far	Near	Far	Near	Far	Ave.	Near	Far	Ave.
	Unproc.	4.50	4.58	3.74	2.97	3.57	2.73	3.68	3.17	3.19	3.18
1ch	WPE	4.62	4.84	3.99	3.20	3.85	2.89	3.90	3.43	3.51	3.47
	Aux-ICA, $\beta = 0.2$	4.65	4.91	4.07	3.30	3.95	2.94	3.97	3.49	3.60	3.55
	Aux-ICA, $\beta = 0.4$	**4.66**	**4.95**	**4.13**	3.39	**4.03**	**2.99**	**4.03**	**3.52**	**3.65**	**3.59**
	Aux-ICA, $\beta = 1.0$	4.54	4.77	4.08	**3.41**	4.01	2.95	3.96	3.37	3.56	3.47
2ch	WPE	4.72	5.18	4.31	3.93	4.24	3.36	4.29	3.89	4.08	3.99
	Aux-ICA, $\beta = 0.2$	**4.73**	**5.20**	4.35	4.12	4.31	3.46	4.36	3.97	4.20	4.09
	Aux-ICA, $\beta = 0.4$	4.72	5.18	**4.36**	4.27	**4.34**	3.53	**4.40**	**4.00**	**4.27**	**4.14**
	Aux-ICA, $\beta = 1.0$	4.56	4.94	4.26	**4.40**	4.26	**3.55**	4.33	3.80	4.18	3.99
8ch	WPE	**4.74**	5.19	**4.51**	5.29	**4.63**	4.73	**4.85**	4.85	5.24	5.05
	Aux-ICA, $\beta = 0.2$	4.67	**5.24**	4.47	5.19	4.54	4.55	4.78	4.71	5.12	4.92
	Aux-ICA, $\beta = 0.4$	4.72	5.14	4.50	**5.34**	4.62	**4.75**	4.85	**4.86**	**5.33**	**5.10**
	Aux-ICA, $\beta = 1.0$	4.40	4.96	4.28	5.09	4.37	4.55	4.61	4.50	5.02	4.76

Table 2: *The evaluation set WERs (%) on the baseline TDNN speech recognition system. Bold figures indicates the best performance in each test condition.*

		SimData							RealData		
		Room1		Room2		Room3			Room1		
		Near	Far	Near	Far	Near	Far	Ave.	Near	Far	Ave.
	Unproc.	3.08	3.78	4.64	7.31	4.36	7.09	5.04	18.59	20.22	19.41
1ch	WPE	**3.08**	3.49	4.45	6.83	4.29	6.63	4.80	16.83	18.47	17.65
	Aux-ICA, $\beta = 0.2$	3.08	3.44	**4.40**	6.70	**4.26**	6.37	**4.71**	16.16	17.93	17.05
	Aux-ICA, $\beta = 0.4$	3.15	3.47	4.40	**6.54**	4.35	**6.34**	4.71	16.00	**17.79**	**16.90**
	Aux-ICA, $\beta = 1.0$	3.17	**3.35**	4.58	6.62	4.35	6.51	4.76	**15.62**	18.43	17.03
2ch	WPE	3.50	**3.50**	4.59	6.17	4.23	**6.17**	4.69	15.81	16.78	16.30
	Aux-ICA, $\beta = 0.2$	**3.17**	3.50	4.66	**6.10**	4.26	6.17	**4.64**	**15.55**	**16.24**	**15.90**
	Aux-ICA, $\beta = 0.4$	3.22	3.54	4.64	6.34	4.23	6.24	4.70	15.75	16.85	16.30
	Aux-ICA, $\beta = 1.0$	3.27	3.50	**4.51**	6.47	**4.21**	6.66	4.77	15.78	17.05	16.42
8ch	WPE	3.19	3.57	4.41	5.11	3.57	**3.92**	3.96	**12.62**	**13.27**	**12.95**
	Aux-ICA, $\beta = 0.2$	**3.14**	3.59	**4.36**	**5.06**	3.53	3.97	**3.94**	12.71	13.71	13.21
	Aux-ICA, $\beta = 0.4$	3.14	3.52	4.46	5.08	**3.51**	4.06	3.96	13.22	13.94	13.58
	Aux-ICA, $\beta = 1.0$	3.18	**3.45**	4.46	5.14	3.67	4.08	4.00	13.77	14.72	14.25

5. Conclusions

This paper gives some new insights into speech dereverberation by addressing it from a semi-blind source separation perspective, assuming that clean speech are independent across time. A novel Aux-ICA based solution is proposed under the independence maximization measure, and the proposed solution turns out mathematically equivalent with the popular MCLP algorithms. Our work highlights the role of accurate speech modeling, and a super-Gaussian source prior with shape parameter $\beta \in [0.2, 0.4]$ not only applies to speech separation but also works in the speech dereverberation tasks.

6. References

[1] 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.

[2] S. Sivasankaran, E. Vincent, and I. Illina, "A combined evaluation of established and new approaches for speech recognition in varied reverberation conditions," *Computer Speech & Language*, vol. 46, pp. 444–460, 2017.

[3] E. A. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770–773, 2009.

[4] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1006–1018, 2015.

[5] B. Wu, K. Li, M. Yang, and C.-H. Lee, "A reverberation-time-aware approach to speech dereverberation based on deep neural networks," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 1, pp. 102–111, 2016.

[6] 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.

[7] E. A. Habets and J. Benesty, "A two-stage beamforming approach for noise reduction and dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 945–958, 2013.

[8] I. Kodrasi and S. Doclo, "Evd-based multi-channel dereverberation of a moving speaker using different retf estimation methods," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*. IEEE, 2017, pp. 116–120.

[9] T. Dietzen, S. Doclo, A. Spriet, W. Tirry, M. Moonen, and T. van Waterschoot, "Low-complexity kalman filter for multi-channel linear-prediction-based blind speech dereverberation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 284–288.

[10] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[11] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[12] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Interspeech 2018-19th Annual Conference of the International Speech Communication Association*, 2018.

[13] J. Caroselli, I. Shafran, A. Narayanan, and R. Rose, "Adaptive multichannel dereverberation for automatic speech recognition." in *Interspeech*, 2017, pp. 3877–3881.

[14] J. Heymann, L. Drude, R. Haeb-Umbach, K. Kinoshita, and T. Nakatani, "Frame-online dnn-wpe dereverberation," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 466–470.

[15] A. Jukić and S. Doclo, "Speech dereverberation using weighted prediction error with laplacian model of the desired signal," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 5172–5176.

[16] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1509–1520, 2015.

[17] T. Taniguchi, A. S. Subramanian, X. Wang, D. Tran, Y. Fujita, and S. Watanabe, "Generalized weighted-prediction-error dereverberation with varying source priors for reverberant speech recognition," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 293–297.

[18] S. R. Chetupalli and T. V. Sreenivas, "Late reverberation cancellation using bayesian estimation of multi-channel linear predictors and student's t-source prior," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1007–1018, 2019.

[19] H. Buchner and W. Kellermann, "TRINICON for dereverberation of speech and audio signals," in *Speech Dereverberation*. Springer, 2010, pp. 311–385.

[20] R. Takeda, K. Nakadai, T. Takahashi, K. Komatani, T. Ogata, and H. G. Okuno, "Efficient blind dereverberation and echo cancellation based on independent component analysis for actual acoustic signals," *Neural computation*, vol. 24, no. 1, pp. 234–272, 2012.

[21] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-gaussian sources," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2010, pp. 165–172.

[22] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2011, pp. 189–192.

[23] T.-T. Lu and S.-H. Shiou, "Inverses of 2 × 2 block matrices," *Computers & Mathematics with Applications*, vol. 43, no. 1-2, pp. 119–129, 2002.

[24] I. Tashev and A. Acero, "Statistical modeling of the speech signal," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010.

[25] N. Ono, "Auxiliary-function-based independent vector analysis with power of vector-norm type weighting functions," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2012, pp. 1–4.

[26] Y. Na, Z. Wang, Z. Liu, Y. Li, B. Tian, and Q. Fu, "A new perspective of auxiliary function based independent component analysis in acoustic echo cancellation," *submitted to Interspeech 2020*.

[27] C. Boeddeker, T. Nakatani, K. Kinoshita, and R. Haeb-Umbach, "Jointly optimal dereverberation and beamforming," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[28] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.