An Advanced Expectation-Maximization Algorithm for Multimicrophone Speech Dereverberation and Noise Reduction

Liu Han[1, *] **and Zhongfu Ye**[2]

[1]School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China
[2]School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China

*Corresponding author e-mail: 820261611@qq.com

Abstract. In speech communication scenes, the microphone signals are often weakened by reverberation and ambient noise. Reverberation can be divided into an early part that is comprised of the direct speech and some early ones and a late part. In this paper, an advanced expectation-maximization (EM) algorithm for multimicrophone speech dereverberation and noise reduction, extending EM by taking the noise statistics into account, is proposed. This method aims at estimating the early parts filtered by early transform functions (ETFs). The EM algorithm is executed in two steps. In the E-step, the filtered early speech is calculated. Then, in the M-step, the ETFs, the power spectral density (PSD) of the filtered early speech and the late reverberation, and the spatial coherence matrix of the late reverberation are estimated. This algorithm is evaluated using an open-source room impulse responses (RIRs) database with a reverberation time of 0.36s and 0.61s under different signal-to-noise (SNR) conditions. It is proved an improvement compared with the original EM method.

1. Introduction

With the development of science and technology, we have ushered in the era of artificial intelligence. To achieve human-computer interaction in artificial intelligence, automatic speech recognition (ASR) is crucial. In practical application scenarios, the signal received by microphone arrays often contains noise and reverberation formed by the reflection of the wall which degrade speech intelligibility and accuracy of ASR systems. Highly reverberant speech is equally difficult for machines and people to understand and can easily lead to hearing fatigue. With the increase in demand for high-quality voice, dereverberation and noise reduction has become essential speech enhancement technologies. Existing joint dereverberation and noise reduction methods can be mainly divided into spectral enhancement techniques, multichannel equalization techniques [1~6], and probabilistic model techniques [7].

Spectral enhancement techniques can be implemented in a two-stage approach. A common two-level approach is based on multichannel Wiener filter (WCMF), which divides the MCWF into a minimum variance distortionless response (MVDR) beam former (BF) and a single-channel post filter(PF)[8]. The MVDR BF is usually implemented in a generalized sidelobe canceller (GSC) structure which consists of three parts in two branches. The fixed beam former (FBF) is the upper branch responsible for preserving the response required for the signal of interest. The lower branch is separated into two parts, a blocking matrix (BM) and a noise canceller (NC). The BM blocks the signal of interest while the NC eliminates the interference. The generalization of MVDR BF is linear constrained minimum variance BF (LCMV BF). LCMV BF can be applied to construct a beam model that satisfies a series of directional constraints while minimizing the power of output noise and can also be used to achieve dereverberation and noise reduction [9].

Multichannel equalization techniques based on multichannel inverse theorem (MINT) [10] are in the category of reverberation cancellation. These methods aim to reconstruct the RIRs between the source and the microphone array. In [11], two joint dereverberation and noise reduction time-domain techniques are proposed. The first technique explicitly considers the noise data in regularized partial MINT (RPMINT). Besides the regularization parameters in RPMINT, it introduces additional weight parameters that weigh the performance between dereverberation and noise reduction. The second technique combined with dereverberation and noise reduction MCWF (MCWF-DNR) takes into account both speech and noise data, and uses the RPMINT filter to calculate a dereverberated correlation signal for the MCWF.

Based on probabilistic model techniques, acoustic transport system is usually modeled as an auto-regressive process [12, 13] or using the convolution transfer function. The spectral coefficients of clean speech are modeled as a Gaussian distribution [12]. Dereverberation is then achieved by maximizing the likelihood functions of all unknown model parameters. Expectation-maximization (EM) algorithm is one of the probabilistic model technologies.

In our paper, an advanced EM algorithm for multimicrophone speech dereverberation and noise reduction is proposed. Late reverberation, anechoic speech and additive noise are assumed to be mutually independent and modeled by zero-mean multidimensional Gaussian components. The early reverberation is modeled by multiplication of ETFs and the anechoic speech. The PSD matrix of late reverberation is time-variant while the spatial coherence matrix of that is time-invariant since the locations of source and microphone array are known and fixed. The anechoic speech, the late reverberation and the noise are defined as the hidden data. In E-step, the estimation of anechoic speech is completed by a MCWF. The other parameters are estimated in M-step following. In previous works, the PSD matrix of noise is assumed known. As in the practical applications, the noise is unknown. We have incorporated the PSD matrix of noise into the hidden data, and showing by doing so, the noise reduction performance can indeed improved.

The remainder of this paper is organized as follows. In Section 2-1.1, we formulate the problem. In Section 2-1.2, the EM algorithm is derived. In Section 3, the results and the evaluation of the proposed approach are present in the form of a chart. Section 4 is allocated for concluding the whole works.

2. Configuration And Notation

2.1. Problem Formulation

We consider a reverberant system in a noisy environment. There is a sound source and I microphones settled as a line array. The signal received by the ith microphone is $y_i(n)$, i=1,...,N, and when time subscript is n:

$$\begin{aligned} y_{\mathrm{i}}(n) &= h_{\mathrm{i}}(n) \otimes s(n) + v_i(n) \\ &= x_{\mathrm{i}}(n) + v_{\mathrm{i}}(n) \end{aligned} \tag{1}$$

Here $h_{\mathrm{i}}(n)$ is the RIRs between the sound source and the microphone array, $s(n)$ is clean speech, $x_{\mathrm{i}}(n)$ is reveberant speech, and $v_{\mathrm{i}}(n)$ is additive noise. Let $h_{e,i}(n)$ denotes the direct path and early reflections of RIRs and $h_{r,m}(n)$ denotes the late reverberation data of the RIRs, the signal received by the ith microphone in (1) can be written as:

$$\begin{aligned} y_i(n) &= h_{e,i}(n) \otimes s(n) + h_{r,i}(n) \otimes s(n) + v_i(n) \\ &= x_{e,i}(n) + x_{r,i}(n) + v_i(n) \end{aligned} \tag{2}$$

Taking STFT to (2), we get:

$$Y_i(m,k) = X_{e,i}(m,k) + R_i(m,k) + V_i(m,k) \tag{3}$$

Here m represents the time index, k represents the frequency index. $X_{e,i}(m,k)$ can be modeled as the product of anechoic speech and ETFs:

$$\mathrm{X}_{e,i}(m,k) = G_{e,i}(k)S(m,k) \tag{4}$$

Here $G_{e,i}(k)$ is the ETFs, $S(m,k)$ is the anechoic speech. Combining I microphone signals in a vector form yields:

$$\mathbf{y}(m,k) = \mathbf{x}_e(m,k) + \mathbf{r}(m,k) + \mathbf{v}(m,k) \tag{5}$$

$$\mathbf{x_e}(m,k) = \mathbf{g_e}(m,k)S(m,k) \tag{6}$$

Where

$\mathbf{y}(m,k) = [Y_1(m,k) \;\; Y_2(m,k)...Y_N(m,k)]^T$, $\mathbf{x}_e(m,k) = [X_{e,1}(m,k) \;\; X_{e,2}(m,k)...X_{e,\mathrm{N}}(m,k)]^T$

$\mathbf{r}(m,k) = [R_1(m,k) \;\; R_2(m,k)...R_N(m,k)]^T$, $\mathbf{v}(m,k) = [V_1(m,k) \;\; V_2(m,k)...V_N(m,k)]^T$

$\mathbf{g}_e(k) = [G_{e,1}(k) \;\; G_{e,2}(k)...G_{e,\mathrm{N}}(k)]^T$

Replace (6) with:

$$\mathbf{x}_e(m,k) = \overline{\mathbf{g}}_{\mathrm{e}}(k)S_F(m,k) \tag{7}$$

Where

$S_F(m,k) = \mathbf{q}^H(k)\mathbf{g}_e(k)S(m,k)$

$\bar{\mathbf{g}}_e(k) = \dfrac{\mathbf{g}_e(k)}{\mathbf{q}^H(k)\mathbf{g}_e(k)}$

Reverberation, anechoic speech and additive noise are assumed to be mutually independent and modeled by zero-mean multidimensional Gaussian components. Let PSD of the reverberation is time-varying and spatially coherent matrix is time-invariant, the probability density function (p.d.f) of late reverberation can be modeled as:

$$f(\mathbf{r}(m,k);\phi_R(m,k),\Gamma(k)) = N^c(\mathbf{r}(m,k);\mathbf{0};\phi_R(m,k)\Gamma(k)) \tag{8}$$

Here:

$N^c(\mathbf{x};\mathbf{0},\Phi) = \dfrac{1}{\pi^N|\Phi|}\exp(-\mathbf{x}^H\Phi^{-1}\mathbf{x})$

x represents the Gaussian vector, Φ is the PSD matrix, and the invariant spatial correlation matrix $\Gamma(k)$ describes the spatial characteristics of the late reverberation field, $\phi_R(m,k)$ represents the time-varying PSD of late reverberation. So:

$$\phi_R(m,k) = \frac{1}{N}\sum_{i=1}^{N} E\{|R_i(m,k)|^2\}$$

$$\phi_{S_F}(m,k) = E\{|S_F(m,k)|^2\},\ f(S_F(m,k);\phi_{S_F}(m,k)) = N^c(S_F(m,k);0,\phi_{S_F}(m,k))$$

Define $\phi_R(k) = [\phi_R(1,k),...,\phi_R(M,k)]$ and $\phi_{S_F}(k) = [\phi_{S_F}(1,k),...,\phi_{S_F}(M,k)]$. The entire parameter set for the problem is:

$$\boldsymbol{\theta}(k) = \{\phi_{S_F}(k), \bar{\mathbf{g}}_e(k), \phi_R(k), \Gamma(k), \Phi_v(k)\} \tag{9}$$

Where M is the total number of frames. Define $\bar{\mathbf{y}}(k) = [\mathbf{y}^T(1,k)\ \ ...\ \ \mathbf{y}^T(M,k)]$, the PSD is:

$$f(\bar{\mathbf{y}}(k);\boldsymbol{\theta}(k)) = \prod_{m=1}^{M} N^c(\mathbf{y}(m,k);\mathbf{0},\Phi_{\mathbf{y}}(m,k)) \tag{10}$$

$$\Phi_{\mathbf{y}}(m,k) = \phi_{S_F}(m,k)\bar{\mathbf{g}}_e{}^H(k) + \phi_R(m,k)\Gamma(k) + \Phi_v(k) \tag{11}$$

Our aim is to maximize (15):

$$\boldsymbol{\theta}_{ML}(k) = \arg\max_{\boldsymbol{\theta}} f(\bar{\mathbf{y}}(k);\boldsymbol{\theta}(k)) \tag{12}$$

2.2. EM Algorithm

In order to use the EM algorithm, hidden variables must be defined. We define $S_F(m,k)$, $\mathbf{r}(m,k)$ and $\mathbf{v}(m,k)$ as hidden data. In step E, auxiliary functions, such as the joint log-likelihood expectations for the observed and hidden variables, need to be derived. In step M, the auxiliary functions are maximized according to the relationship between the parameters. This process converges to the local optimum of the likelihood function. In the following description, in order to make the formula look simpler, the frequency subscript k is ignored.

Define the hidden data as:

$$\mathbf{d}(m) \overset{\Delta}{=} [S_F(m)\quad \mathbf{r}^T(m)\quad \mathbf{v}^T(m)]^T \tag{13}$$

The expression (5) can be written as: $\mathbf{y}(m) = \mathbf{Hd}(m)$, $\mathbf{H} \overset{\Delta}{=} [\bar{\mathbf{g}}_e\quad \mathbf{I}_{N\times N}\quad \mathbf{I}_{N\times N}]$. In order to achieve the E step, several estimates are added:

1) $\hat{\mathbf{d}}(m) \overset{\Delta}{=} E\{\mathbf{d}(m)|\mathbf{y}(m);\theta^{(l)}\}$

2) $\hat{\boldsymbol{\Psi}}_d(m) \overset{\Delta}{=} \widehat{\mathbf{d}(m)\mathbf{d}^H(m)} = E\{\mathbf{d}(m)\mathbf{d}(m)^H|\mathbf{y}(m);\boldsymbol{\theta}^{(l)}\}$

Then, we can deduce:

$\hat{\mathbf{r}}(m)=\hat{\mathbf{d}}_{\{2:N+1\}}(m)$, $\hat{S}_F(m)=\hat{\mathbf{d}}_{\{1\}}(m)$, $\hat{\mathbf{v}}(m)=\hat{\mathbf{d}}_{\{N+2:2N+1\}}(m)$, $\widehat{\mathbf{r}(m)\mathbf{r}^H(m)}=\widehat{\Psi}_{d,\{2:N+1,2:N+1\}}(m)$, $\widehat{\mathbf{v}(m)\mathbf{v}^{\mathbf{H}}(m)}=\widehat{\Psi}_{d,\{N+2:2N+1,N+2:2N+1\}}$, $\widehat{\left|S_F(m)\right|^2}=\widehat{\Psi}_{d,\{1,1\}}(m)$, $\widehat{S^*{}_F(m)\mathbf{r}(m)}=\widehat{\Psi}_{d,\{2:N+1,1\}}(m)$

Since ***y*** (***m***) and ***d*** (***m***) are Gaussian vectors, ***d*** (***m***) can be estimated by MCWF:

$$\mathbf{d}(m)=E\{\mathbf{d}(m)\mathbf{y}^H(m)\}\times(E\{\mathbf{y}(m)\mathbf{y}^H(m)\})^{-1}=\Phi_d^{(l)}(m)(\mathbf{H}^{(l)})^H(\Phi_y^{(l)}(m))^{-1}\mathbf{y}(m) \tag{14}$$

With:

$$\Phi_d^{(l)}(m)=\begin{bmatrix} \phi_{S_F}^l(m) & \mathbf{0}_{1\times N} & \mathbf{0}_{1\times N} \\ \mathbf{0}_{1\times N} & \phi_R^l(m)\Gamma^{(l)} & \mathbf{0}_{1\times N} \\ \mathbf{0}_{1\times N} & \mathbf{0}_{1\times N} & \Phi_{\mathrm{v}} \end{bmatrix} \tag{15}$$

$$\Phi_y^{(l)}(m)=\mathbf{H}^{(l)}\Phi_d^{(l)}(m)(\mathbf{H}^{(l)})^H \tag{16}$$

In step M, the auxiliary functions $Q_{MAP}(\boldsymbol{\theta};\boldsymbol{\theta}^{(l)})$ are maximized by linking the problem parameters.

$$\phi_{S_F}^{(l+1)}(m)=\widehat{\left|S_F(m)\right|^2},\bar{\mathbf{g}}_e^{(l+1)}=\frac{\sum_{m=1}^{M}\widehat{S_F^*}(m)\mathbf{y}(m)-\widehat{S_F^*(m)\mathbf{r}(m)}}{\sum_m\widehat{\left|S_F(m)\right|^2}},\mathbf{\Gamma}^{(l+1)}=\frac{1}{M}\sum_{m=1}^{M}(\phi_R^{(l)}(m))^{-1}\widehat{\mathbf{r}(m)\mathbf{r}^H(m)},$$

$$\phi_R^{(l+1)}(m)=\gamma\phi_R^{(0)}(m)+(1-\gamma)\frac{1}{N}Tr[\widehat{\mathbf{r}(m)\mathbf{r}^H(m)}(\mathbf{\Gamma}^{(l+1)})^{-1}],\Phi_v^{(l+1)}=\frac{1}{M}\sum_{m=1}^{M}\widehat{\mathbf{v}(m)\mathbf{v}^H(m)}$$

3. Results

The clean speech library used in this paper is TIMIT, the speech sampling frequency is 16KHz. The reverberation data comes from the open source RIRs database. The microphone array is uneven 4 linear arrays, the distance between the microphones [0.03, 0, 08, 0, 03] m and the sound source is placed 2m in front of the linear array. The speech quality is evaluated by computing the PESQ and LSD. The result is displayed in Table.1.

Table 1. PESQ and LSD for a reverberation time of 0.36s and 0.61s.

		SNR	10dB	15dB	20dB	25dB	30dB
0.36s	PESQ	Unprocessed	1.47	1.71	2.01	2.19	2.38
		OriginalEM	2.10	2.49	2.80	3.01	3.19
		ProposedEM	**2.20**	**2.56**	**2.85**	**3.10**	**3.25**
	LSD	Unprocessed	16.31	12.80	8.30	5.30	3.28
		OriginalEM	6.73	4.38	2.98	2.35	2.10
		ProposedEM	**6.45**	**4.10**	**2.81**	**2.10**	**1.90**
0.61s	PESQ	Unprocessed	1.49	1.68	1.84	1.95	2.00
		OriginalEM	1.91	2.12	2.23	2.32	2.38
		ProposedEM	**1.94**	**2.15**	**2.29**	**2.38**	**2.45**
	LSD	Unprocessed	16.11	12.25	8.90	6.06	4.53
		OriginalEM	7.01	4.85	3.54	3.10	2.93
		ProposedEM	**6.90**	**4.65**	**3.40**	**2.90**	**2.75**

4. Conclusion

In this paper, an advanced EM algorithm was presented to obtain an estimate of a spatially diltered version of the early speech component with suppressing early reflections, late reverberation and ambient noise. We modeled the early speech component as the product of anechoic speech and ETFs. Besides, the late reverberation was assumed to have time-varying PSD and time-invariant spatial characteristics. The hidden data was defined to be the anechoic speech, late reverberation signals and noise vectors. The algorithm was tested in simulation with a reverberation time of 0.36s and 0.61s for several SNR levels. The proposed algorithm performs better than the original method in PESQ and LSD.

Acknowledgments

First of all, I would like to express my sincere heartfelt thanks to my tutor, Mr. Ye, for his invaluable advice, constant encouragement and precise modification, and I admire his knowledge and personality.

Then I would also like to express my gratitude to the schoolmates in our lab because they always talk patiently about what I do not understand.

Meanwhile, I would like to thank my parents who always support me in many aspects.

Finally, I would like to thank my roomates who friendly encourage me in writing this paper.

References

[1] M. Miyoshi and Y. Kaneda, Inverse filtering of room acoustics, IEEE Trans. Acoust., Speech, Signal Process., vol.36, no.2, pp. 145 – 152, Feb. 1988.

[2] M. Kallinger and A. Mertins, Multi-channel room impulse response shaping - a study, Proc. I nt. Conf. Acoust., Speech, Signal Process.,Toulouse, France, pp. 101 – 104, May. 2006.

[3] J. O. Jungmann, R. Mazur, M. Kallinger, M. Tiemin, and A. Mertins, Combined acoustic MIMO channel crosstalk cancellation and room impulse response reshaping, IEEE Trans. Audio, Speech, Lang. Process.,vol. 20, no.6, pp. 1829 – 1842, Aug. 2012.

[4] I. Kodrasi, S. Goetze, and S. Doclo, Regularization for partial multichannel equalization for speech dereverberation, IEEE Trans. Audio, Speech, Lang. Process., vol.21, no.9, pp. 1879-1890, Sep. 2013.

[5] F. Lim, W. Zhang, E. A. P. Habets, and P. A. Naylor, Robust multichannel dereverberation using relaxed multichannel least squares, IEEE/ACM Trans. Audio, Speech, Lang. Process., vol.22, no.9, pp. 1379 – 1390, Sep. 2014.

[6] R. S. Rashobh, A. W. H. Khong, and D. Liu, Multichannel equalization in the KLT and frequency domains with application to speech dereverberation, IEEE/AC M Trans. Audio, Speech, Lang. Process., vol.22, no.3, pp. 634 – 646, Mar. 2014.

[7] E.A.P. Habets, S. Gannot, and I. Cohen, Late reverberant spectral variance estimation based on a statistical model, IEEE Signal Process. Lett., vol.16, no.9, pp. 770 – 774, Sep. 2009.

[8] K. U. Simmer, J. Bitzer, and C. Marro, Post-filtering techniques, Microphone Arra ys, M. B randstein and D. Ward, Eds, Berlin, Germany: pringer, 2001

[9] Ofer Schwartz, Sebastian Braun, Sharon Gannot, Emanuël A. P. Habets, Source Separation, Dereverberation and Noise Reduction Using LCMV Beamformer and Postfilter, LVA/ICA, pp 182-191, 2017

[10] M. Miyoshi and Y. Kaneda, Inverse filtering of room acoustics, IEEETrans. Acoust., Speech, Signal Process., vol.36, no.2, pp. 145 – 152, Feb. 1988.

[11] Ina Kodrasi, Simon Doclo, Joint dereverberation and noise reduction based on acoustic multi-channel equalizaiton, IEEE/ACM Transaction on audio,speech,and language peocessing, vol.24, no.4, Apr. 2016.

[12] T. Nakatani, T. Yo shioka, K. Kinoshita, M. Miyoshi, and J. Biing-Hwang, Speech dereverberation based on variance-normalized delayed linear prediction, IEEE Trans. Audio, Speech, Lang. Pro cess., vol.18, no.7, pp. 1717 – 1731, Sep. 2010.

[13] A. Jukic and S. Doclo, Speech dereverberation using weighted prediction error with Laplacian model of the desired signal, Proc. Int. Conf. Acoust., Speech, Signal Process., Florence, Italy, pp. 5172 – 5176, May. 2014.