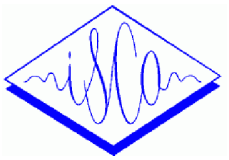
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Model-based Single-Channel Dereverberation in Noisy Acoustical Environments

Xulei Bao[1], *Jie Zhu*[2]

[1]Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China)
[2]Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China)

qunzhong@sjtu.edu.cn, zhujie@sjtu.edu.cn

Abstract

This paper illustrates a new system for recovering clean speech signals from noisy acoustical environments using one microphone. At the beginning of this paper, we propose an assumption that the background noise is comprised of reverberant noise and direct-path noise. And a novel late reverberant spectral variance (LRSV) estimator is generated referring to this assumption, which can be used in the noisy acoustical environments. What's more, the shape parameters of this LRSV estimator are updated by taking some frames of previous late reverberation into account. At last, a new spectral process system is developed to help making the LRSV estimator more efficient. The experimental results show the benefits of our new system when it is used to reduce the interference in both noise-free and noisy acoustical environments.

Index Terms: statistical reverberation model; late reverberant spectral variance estimator; spectral processing

1. Introduction

Speech signals captured by a distant microphone in a confined room are often corrupted by both reverberation and background noise. This distortion is detrimental to the perceived quality and intelligibility and often cause serious degradation in many speech applications, such as automatic speech recognition (ASR) [1, 2, 3]. It has been reported that the ASR performance cannot be improved even applying the acoustic models that have been trained with a matched reverberation condition when the reverberation time (T_{60}) is more than 0.5 seconds [4]. Therefore, the dereverberation of such speech signals is essential for speech applications.

Fortunately, a lot of work has been done on the development of dereverberation methods during the past two decades, especially in model-based method. The main idea of this method is assuming that the early and late reverberant speech component are mutually independent, and the suppression of late reverberant spectral variance (LRSV) is commonly carried out in the short-time Fourier transform (STFT) domain using so-called spectral enhancement algorithm. In the last decade, several LRSV estimators have been developed [5, 6, 7, 8]. In [5], the LRSV estimator is derived under the condition that the source-microphone distance is larger than the critical distance. However, this LRSV estimator may overestimate the LRSV when the source-microphone distance is smaller than the critical distance. To overcome this problem, Habets derives a more general LRSV estimator using a statistical reverberation model that takes the energy contribution of the direct-path into account [8]. These model-based LRSV estimators are limited to modeling the room impulse responses (RIRs) as time-invariant realizations of a stochastic process. In order to extend the LRSV estimator to time-varying environments, Erkelens proposes a statistical model for time-varying RIRs and derives the estimator under only a few mild conditions [7]. However, this LRSV estimator is not accurate enough since the shape parameter κ which is used for estimating the LRSV is sensitive to the denoised signal in noisy environments.

In this paper, we derive a new LRSV estimator in noisy and reverberant environments by applying the same statistical model as that proposed by Erkelens in [7], but using a novel assumption that the background noise is comprised of reverberant noise and direct-path noise corresponding to the reverberation and direct-path speech respectively. To achieve high accuracy of the LRSV estimator, its parameter κ is updated by taking some frames of previous late reverberation into account. In our work, a new spectral process system is also developed for the proposed LRSV estimator.

The paper is organized as follows. In Section 2, the statistical reverberation model is introduced and the LRSV estimator is derived. A new spectral process system for dereverberation and noise reduction is developed in Section 3. The performance of the proposed system is analyzed with practical RIRs in Section 4. Finally, in Section 5, we present our concluding comments.

2. Signal Model and LRSV Estimator

Both statistical reverberant models for RIRs and these model-based LRSV estimators have been widespread reported as mentioned above, for these model-based estimators can lead to a simple expression for the LRSV depending on past values of the spectral variance of the noise-free reverberant signals. When noisy and reverberant signals are captured, the background noise is first mitigated by the speech enhancement method, and the LRSV is then estimated by the LRSV estimator according to the denoised signals. It seems that the model-based method is effective for such speech signals. However, we should note that the background noise cannot be eliminated completely and the speech enhancement algorithm may also distort the reverberant speech, which will degrade the performance of the LRSV estimator. Because of this, we investigate LRSV estimator further for noisy and reverberant signals.

2.1. LRSV estimator

We suppose the noise-free reverberant speech signal $x(n)$ results from the convolution of a source speech signal $s(n)$ and a possibly time-varying RIRs $h(n)$, and the observed signal $z(n)$ to be the sum of the noise-free reverberant speech signal $x(n)$

and the additive noise $d(n)$, independent of $s(n)$

$$z(n) = x(n) + d(n) = \sum_{l=0}^{\infty} h_n(l)s(n-l) + d(n), \quad (1)$$

where n is the discrete-time sample index, the additive noise $d(n)$ might be stationary or non-stationary. The RIR model is an i.i.d Gaussian noise sequence with exponentially decaying variance for reverberant path and a delta pulse for direct path, as follows [7]

$$h_n(l) = \begin{cases} 1, & \text{for} \quad l = 0 \\ r_n(l)e^{-\delta_n l}, & \text{for} \quad l \geq 1 \end{cases}, \quad (2)$$

where $r_n(l)$ is a zero-mean i.i.d Gaussian process with variance $\sigma_r^2(n) \leq 1$. The decay rate δ_n depends on the reverberation time T_{60} of the room, and the variance $\sigma_r^2(n)$ depends on direct-to-reverberation ratio (DRR) as follows [7]

$$\delta_n = \frac{3\ln(10)}{T_{60}(n)F_s}, \qquad \sigma_r^2(n) = \frac{e^{2\delta_n} - 1}{DRR(n)}, \quad (3)$$

where F_s is the sampling frequency. In this paper, the decay rate δ_n, the reverberation time $T_{60}(n)$, and the variance $\sigma_r^2(n)$ mentioned above are all assumed to change slowly over time at any acoustical environments.

We assume the additive noise $d(n)$ can be separated into two components, i.e., the reverberant noise $d_r(n)$ and the direct-path noise $d_d(n)$. The reverberant noise d_r is assumed to be the convolution of the RIRs $h(n)$ and a random signal $v(n)$

$$d(n) = d_r(n) + d_d(n) = \sum_{l=1}^{\infty} h_n(l)v(n-l) + d_d(n). \quad (4)$$

Note that the random signal $v(n)$ is independent of direct path noise $d_d(n)$ which also means we can divide the background noise into reverberant noise and direct path noise with no doubt, as these two items are uncorrelated.

Under this assumption, the observed signal $z(n)$ can be rewritten as

$$\begin{aligned} z(n) &= s(n) + \sum_{l=1}^{\infty} h_n(l)s(n-l) + d_r(n) + d_d(n) \\ &= s(n) + \sum_{l=1}^{\infty} h_n(l)s'(n-l) + d_d(n) \\ &= s(n) + d_d(n) + z_r(n), \end{aligned} \quad (5)$$

where $s'(n) = s(n) + v(n)$ denotes the speech $s(n)$ contaminated by the signal $v(n)$, $z_r(n) = \sum_{l=1}^{\infty} h_n(l)s'(n-l)$ is the reverberant component of the observed signal. Suppose the reverberant component is consist of late reverberation $z_l(n)$ and early reverberation $z_e(n)$

$$\begin{aligned} z_r(n) &= z_e(n) + z_l(n) \\ &= \sum_{l=1}^{L} h_n(l)s'(n-l) + \sum_{l=L+1}^{\infty} h_n(l)s'(n-l), \end{aligned} \quad (6)$$

where L is the interval after which the late reverberation is assumed to start.

Let $Z(k,m)$, $Z_r(k,m)$, $S(k,m)$ and $D(k,m)$ be complex-valued random variables representing the short-time discrete Fourier transform (DFT) coefficients at frequency index k of the signal frame starting at sample index m from the observed signal, reverberant component, direct-path signal, and the direct path noise. According to (5) we have

$$Z(k,m) = S(k,m) + Z_r(k,m) + D(k,m), \quad (7)$$

where $Z_r(k,m)$ can be written as

$$\begin{aligned} Z_r(k,m) &= Z_l(k,m) + Z_e(k,m), \\ Z_l(k,m) &= \sum_{n=0}^{N-1} w(n) \sum_{l=L+1}^{\infty} h_{m+n}(l)s'(m+n-l)W_{k,n}, \\ Z_e(k,m) &= \sum_{n=0}^{N-1} w(n) \sum_{l=1}^{L} h_{m+n}(l)s'(m+n-l)W_{k,n}, \end{aligned} \quad (8)$$

where N is the frame length, $w(n)$ is an analysis window, and $W_{k,n} = e^{-2\pi ikn/N}$. Z_l and Z_e denote the late reverberation and early reverberation respectively. Using the similar assumption as proposed in [9], we can conclude the spectral variance of the late reverberation can be written as

$$\begin{aligned} &\lambda_l(k,m) \approx e^{-2\tilde{\delta}_m L}(\lambda_e(k,m-L) + \lambda_l(k,m-L)) \\ &= e^{-2\tilde{\delta}_m L}(\kappa_{m-L}\lambda_z(k,m-L) + (1-\kappa_{m-L})\lambda_l(k,m-L)), \end{aligned} \quad (9)$$

where $\lambda_z(k,m)$ and $\lambda_l(k,m)$ are the spectral variance of the observed signal and the spectral variance of the late reverberation. And κ_m is defined as

$$\kappa_m = \lambda_e(k,m)/(\lambda_s(k,m) + \lambda_d(k,m) + \lambda_e(k,m)), \quad (10)$$

where $\lambda_e(k,m)$, $\lambda_s(k,m)$ and $\lambda_d(k,m)$ are the spectral variance of the early reverberation, the spectral variance of the direct-path speech and the spectral variance of the direct-path noise respectively.

Note that, our estimator has one potential advantage over the existing estimators. Formula (4) divides the background into two parts named reverberant noise and direct path noise. Then, these two parts are added into reverberant component of speech and direct path speech respectively in formula (5). When (10) in practice, we can estimate the LRSV (including spectral variance of late reverberant noise and late reverberant speech) directly from the observed signals without estimating and suppressing the background noise first. After dereverberation, the direct path noise remaining in the dereverberanted speech can be suppressed by the speech enhancement method, and a cleaner speech signal is recovered. Compared with the existing estimators that always estimate the LRSV according to the estimation result of the spectral variance of additive noise, our estimator can obtain more precise values without distortion since the estimation of LRSV comes before the noise suppression. Moreover, even if there are some estimation errors produced by the LRSV estimator, these errors can be mitigated by the noise suppression.

2.2. Estimation of the Shape Parameters

The LRSV estimator (10) requires blind estimation of the shape parameters T_{60} and κ_m. Several blind single-channel estimation methods of T_{60} and κ_m have been proposed recently [5, 7, 10]. In this paper, we measure T_{60} with the Schroeder's method as usual, one may refer to [11] for details.

Erkelens [9] has proved that the coefficient κ_m will be a constant if the speech signal is stationary during an interval of

$L+1$ samples and the RIR taps change little during a frame length, or κ_m becomes frequency dependent if RIRs change quickly. In [7], the author makes use of Lebart's estimator to detect whether κ_m is too small, and update κ_m only when its value is either too large or too small, referring to [7] for details.

However, we should note that the estimation of κ_m is influenced by the errors in the estimated reverberant spectral variance. To avoid this, we use the spectral variance of the observed signal instead. The detail steps are presented as follows

1 Let $\lambda_z(m)$ be the spectral variance of the observed signal (including reverberation and additive noise), iterations $K = L/R + 1$, $\lambda_{Rbuf1} = 0, \cdots, \lambda_{RbufK} = 0$ are the estimation values of LRSVs of the previous frames between the frame $m-L$ and the frame m. let $m = 1$ denotes the first frame in STFT domain, and M is the total frame number.

2 Repeat the following for K times.

$\lambda_{l_i}(m) \approx e^{-2\delta_m L}\{\kappa_m \lambda_z(m-L)+(1-\kappa_m)\lambda_{Rbuf1}\}$,

$\kappa_m^* = (e^{2\hat{\delta}_m L}\lambda_z(m)-\lambda_{Rbuf1})/(\lambda_z(m-L)-\lambda_{Rbuf1})$,

$\kappa_{m_i} = \eta\kappa_m + (1-\eta)\min[\max[\kappa_m^*, 0], 1]$,

$\lambda_{l_i}(m) = \min(\lambda_{l_i}(m), \lambda_z(m))$,

$\lambda_{Rbuf1} = \lambda_{Rbuf2}, \cdots, \lambda_{Rbufk} = \lambda_{RbufK}, \lambda_{RbufK} = \lambda_{l_i}(m)$,

$i = i + 1$.

3 Let $\lambda_l(m) = \lambda_{l_K}(m)$ and $\kappa_{m+R} = \kappa_{m_K}$.

4 if $m < M$, $m = m + 1$, and back to 2.

where i is the index ranging from 1 to K. It is noted that λ_z is the spectral power of the observed signal, which should not be estimated before updating the coefficient κ_m. Moreover, the estimation values of LRSVs of the previous frames between the frame $m-L$ and the frame m are taken into account to estimate the LRSV of the current frame. Therefore, the parameter κ_m is updated for K times in each frame until it converges to a suitable value for calculating the value of LRSV.

3. Spectral Processing in Noisy Reverberant Environment

The noisy and reverberant speech signal $z(n)$, can be written as the sum of the noise-free reverberant speech signal $x(n)$ and additive noise $d(n)$ as

$$z(n) = x(n) + d(n) = h(n) * s(n) + d(n).$$

The goal of dereverberation is to obtain an estimation $\hat{s}(n)$ of the speech signal. Recently, spectral enhancement techniques have been used for speech dereverberation [5, 6, 7]. However, there are two weak points of the proposed methods. First, when the additive noise is small, the value of the noise spectral variance λ_d should be small, and if the estimation of noise variance $\hat{\lambda}_d$ is a little different from λ_d, the $prior$ SNR and $posterior$ SNR may change a lot, which makes the estimation of $\hat{X}(k,m)$ not accuracy. Second, the LRSV estimator relate the late reverberation to the denoised signal, and if the estimation of the denoised speech $\hat{X}(k,m)$ inaccuracy, the performance of the dereverberation might degrade much. To overcome these weak points, the system is modified as shown in Fig.1.

For details, the spectral variance of late reverberation $\hat{\lambda}_l$ is estimated first by the LRSV estimator (10). The noisy and

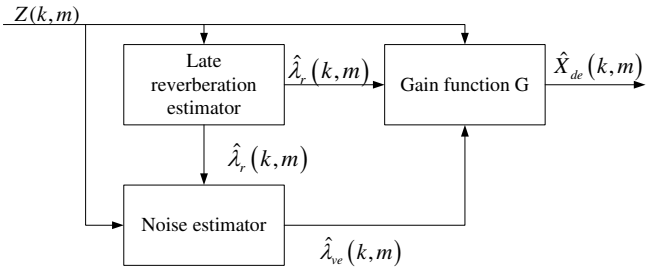


Figure 1: The proposed spectral processing system

dereverberated speech (NDRS) $\hat{Z}_{nd}(k,m)$ is obtained by using the spectral gain function.

$$\hat{Z}_{nd}(k,m) = G(\hat{\xi}_r(k,m)', \hat{\zeta}_r(k,m)')Z(k,m) \tag{11}$$

with

$$\hat{\xi}_r(k,m)' = \frac{\lambda_{nd}(k,m)}{\hat{\lambda}_l(k,m)} \tag{12}$$

and

$$\hat{\zeta}_r(k,m)' = \frac{\left|Z(k,m)\right|^2}{\hat{\lambda}_l(k,m)} \tag{13}$$

where $\lambda_{nd}(k,m)$ is the spectral variance of the NDRS.

After dereverberation, the spectral variance of the direct path noise $\hat{\lambda}_d$ is then estimated by the Data-driven noise tracking method [12].

4. Experimental Results

In this section, we present experimental results to show the effectiveness of the developed system. Two preliminary experiments are present. One is to evaluate the performance of proposed parameter estimation method in Section 2.2 in noise-free acoustical environment, and the other is to demonstrate that the developed system is better than the existing systems in noisy acoustical environments.

We take the Microsoft Research Asia (MSRA) Chinese testing corpus with 500 utterances about 0.74h length (Fs=8kHz) for our experiments. The time-invariant RIRs are the down-sampled versions of the measured RIRs from Aachen Impulse Response (AIR) database [13]. And five different kinds of environments from this database named 'straitway', 'meeting', 'Corridor', 'office' and 'bathroom' are taken as examples. The non-stationary background noise is download from $http://www.freesound.com$ named $Cars\_passing.wav$, whose variance is estimated by using the method in [12]. The system uses a frame length N of 256 samples (32ms for a sampling frequency of 8kHz). For the model-based LRSV estimator, 50% overlap between frames ($R = N/2$) is used. Square-root Hanning analysis and synthesis windows are applied. We set $L = N$ in the LRSV estimator, and we set the smoothing parameter $\eta = 0.95$, which is used for updating the parameter κ_m. The gain function for spectral amplitude estimation is assumed to be a generalized Gamma speech prior with parameters $\gamma = 1$ and $\nu = 1$ [7].

The experimental task is to estimate the corresponding clean speech signal in an offline manner. The results of individual trials are evaluated in terms of Mel-frequency cepstral coefficient (MFCC) distance, since the MFCC distance is related to both the audible quality of speech and the automatic

speech recognition performance [14]. In this paper, 13th-order MFCCs(including zeroth order component) are applied, and the MFCC distance D_{MFCC} is defined as

$$D_{MFCC} = \frac{1}{M}\sum_{m=0}^{M-1}\sum_{k'=0}^{12}(c_{m,k'} - \hat{c}_{m,k'})^2, \qquad (14)$$

where M is the number of short-time segments, $c_{m,k'}$ and $\hat{c}_{m,k'}$ ae the k'th MFCCs of $s(n)$ and $\hat{s}(n)$, respectively. Further, word recognition rate (WRR) is also used to demonstrate the effectiveness of the proposed system. For the WRR, 13th-order MFCCs (including zeroth order component) + Δ + $\Delta\Delta$ are employed to characterize the speech signal. Tradition GMM model which is depicted in the HTK tutorial is also used here. Under this model, the WRR of the clean speech is about 66.7%.

In the first experiment, we validate the proposed method can generate more appropriate parameter κ_m for updating the LRSV in noise-free acoustical environments. The time-invariant RIRs are obtained from the AIR database as mentioned above. The T_{60} and DRR are measured by the Schroeder method [11].

Table 1 shows the MFCC distances averaged over 500 utterances for individual system settings in noise-free environment. The estimation of T_{60} of 'stairway', 'meeting', 'corridor', 'office' and 'bathroom' are 0.839, 0.329, 2.049, 0.585 and 0.439 seconds respectively.

Table 1: MFCC distances average over utterances in noise-free environments

Env	observed	Habets	Erkelens	proposed
Corridor	43.336	22.841	23.013	**17.041**
Stairway	28.754	14.898	15.209	**9.867**
Office	24.494	13.499	13.271	**10.526**
Meeting	19.291	12.499	12.808	**10.752**
Bathroom	20.281	10.757	10.672	**7.994**

In the next experiment, we validate the derived LRSV estimator is a good estimator in noisy acoustical environments. The time-invariant RIRs mentioned in the first experiment are also used here. And the noisy environments are generated by adding the practical noise to the reverberant speech. To demonstrate the good performance of the proposed LRSV estimator, the WRR is depicted in Figure 2, and the RIR here is 'Office'. We should note that, the WRR used here is just for comparing. So, the performance of the WRR is not good enough since GMM is only represent the tri-phone and we only use the bi-gram for speech recognition. From this figure, we can find the performance of the proposed LRSV estimator is much better than the classical estimators at any SNR ranges.

5. Conclusions

We proposed an LRSV estimator for noisy and reverberant environment. The proposed LRSV estimator can also lead to the same expression for the LRSV under the noisy acoustical environments. Furthermore, we proposed a new method to estimate the parameter κ_m. The experimental results show that the estimation value of κ_m by this method is more appropriate. Additionally, we proposed a new spectral processing system for the reverberation and noise suppression. Our experiments show the proposed system is suitable for both noisy and noise-free reverberant environments.

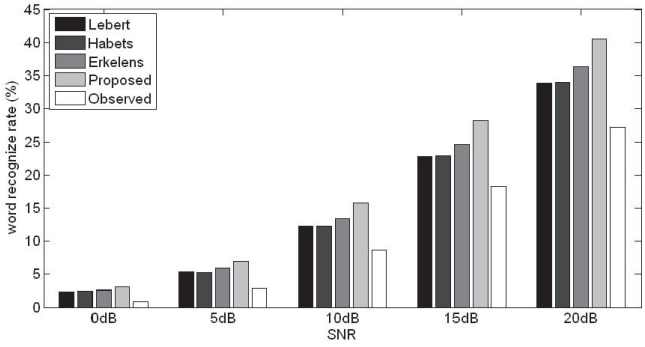


Figure 2: Word recognition rate in noisy 'Office' environments at different SNRs

6. References

[1] Kumar K. and Stern R., "Maximum-likelihood-based cepstral inverse filtering for blind speech dereverberation", Proc. IEEE Internat. Conf. Acoust. Speech, Signal Process., 4282–4285, 2010.

[2] Sehr A., Maas R., and Kellermann W., "Reverberation model-based decoding in the logmelspec domain for robust distant-talking speech recognition", IEEE Trans. Audio, Speech and Lang. Process., 18(7):1676–1691, 2010

[3] Gomez R. and Kawahara T., "Optimization of dereverberation parameters based on likelyhood of speech recognizer", Proc.Interspeech, 1223–1226, 2009.

[4] Kingsbury B. and Morgan N., "Recognizing reverberant speech with RASTA-PLP", Proc. IEEE Internat. Conf. Acoust., Speech, Signal Process., 1259–1262, 1997.

[5] Lebert K., Boucher J.M., and Denbigh P., "A new method based on spectral substraction for speech dereverberation", Acta Acoustica, 87(3):359–368, 2001.

[6] Habets E. A. P., Gaubitch N.D., and Naylor P.A., "Temporal selective dereverberation of noise speech using one microphone", Proc. IEEE Internat. Conf. Acoust., Speech, Signal Process., 4577–4580, 2008.

[7] Erkelens J. S. and Heusdens R., "Noise and late-reverberation suppression in time-varing acoustical environments", Proc. IEEE Internat. Conf. Acoust., Speech, Signal Process., 4706–4709, 2010.

[8] Habets E., Gannot S., and Cohen I.,"Late reverberant spectral variance estimation based on a statistical model", IEEE Signal Process. Lett., 16(9):770–773, 2009.

[9] Erkelens J. S. and Heusdens R., "Correlation-based and model-based blind single-channel late-reverberation suppression in noisy time-varying acoustical environments", IEEE Trans. Audio, Speech and Lang. Process., 18(7):1746–1765, 2010.

[10] Wen J. Y. C, Habets E. A. P., and Naylor P. A., "Blind estimation of reverberation time based on the distribution of signal decay rates", Proc. of DSP, 329–332, 2008.

[11] Schroeder M. R., "New method of measuring reverberation time", J.Acoust. Soc. Am., 37(3):409–412, 1965.

[12] Erkelens, J. S. and Heusdens, R., "Tracking of nonstationary noise based on data-driven recursive noise power estimation", IEEE Trans. Audio, Speech and Lang. Process., 16(6):1112–1123, 2008.

[13] Jeub M., Schafer M. and Vary P., "A binaural room impulse response database for the evaluation of dereverberation algorithms", Proc. Internat. Conf. Digtial Signal Process., 1–5, 2010.

[14] Yoshioka T., Nakatani T., and Miyoshi M., "Integrated speech enhancement method using noise suppression and dereverberation", IEEE Trans. Audio, Speech and Language Process., 17(2):231-246, 2009.