Speech Dereverberation Based on Sparse Matrix Decomposition

Miao Fan[1, a] , Liyang Liu[2,b] and Weifeng Li[3,c*]

[1]Department of Electronic Engineer/Graduate School at Shenzhen,Tsinghua University,China

[2]Shenzhen Key Laboratory of Information Science and Technology,Guangdong,China

[a]email:fanm14@mails.tsinghua.edu.cn

[b]email:Li.Weifeng@sz.tsinghua.edu.cn

Keywords: RPCA,dereverberation,.sparse matrix,low rank matrix

Abstract. Because of the increasingly demands of high quality audio signal, speech dereverberation,as the preliminary processing of speaker recognition and automatic speech recognition(ASR), becomes more and more important. The speech obtained from microphones is always distorted by reverberation. Conventional approaches always build a model to dereverberate speech. However, in different environments, these models may not be effective. For this reason, we propose an algorithm which does not base on any environment model assumptions so that it can be used for all speech. A piece of clean speech can be represented through a sparse matrix. The reverberated speech matrix can be decomposed into two matrices, clean speech matrix and reverberated noise matrix, to capture the sparse components of the speech using Robust Principal Component Analysis (RPCA). Evaluations via many different criterions show that the new approach preserves the clean speech's information well and dereverberate the speech well.

Introduction

Speech dereverberation is vital in many conditions such as automatic speech recognition and speaker recognition. The speech captured by microphones is generally smeared by different kinds of noise. In this paper we pay our attention to multiplicative noise. Reverberated speech is the result of a room impulse response convolving with a piece of clean speech. A reverberated noise is the subtraction between reverberated speech and clean speech. The spectral subtraction algorithm [1] dereverberates the speech by subtracting the reverberated noise from reverberated speech via some prior knowledge. Gaussian Mixture Models (GMM) and contextdependent microphones for acoustic modeling are used in LSA method [2] to dereverberate the speech. The Multiple Regression [3] algorithm dereverberates the speech by using multiple distributed microphones to evaluate a adaptive nonlinear regression. Unlike the above models, we use a matrix decomposition model to dereverberate speech. We assume that clean speech matrix is a sparse matrix and reverberated noise matrix is a low rank matrix. A sparse matrix and a low rank matrix can be decomposed from a reverberated speech matrix by Robust Principal Component Analysis (RPCA) [4] effectively. RPCA have been used in speech enhancing recent years. [5] proposes a novel speech enhancement system based on decomposing the spectrogram into sparse activation and a low-rank background model. In this paper, we use RPCA to dereverberate speech.

Suppression of Reverberation Based on Matrix Decomposition

A. Basic model

The reverberated speech can be described as follow. The x(t) delegates our observed speech with reverberated noise, then

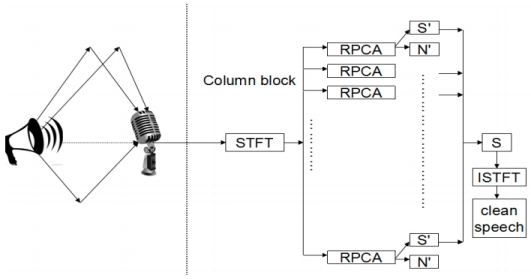
$$x(t) = a(t) * s(t), (1)$$

where a(t) and s(t) are room impulse responses and clean speech. And □ is convolution operator. Performing short-time Fourier transform (STFT), we can obtain the time-frequency signal X(n, k), S(n, k), A(n, k), which are corresponding to x(t), s(t) and a(t) respectively. Now we are supposed to

obtain S(n, k) from X(n, k). We propose a assumption that the clean speech matrix can be decomposed into two matrices as follow

$$X(n,k) = S(n,k) + N(n,k), (2)$$

where N(n, k) denotes the STFT spectrum of reverberated noise. In this paper, we pay our attention to late reverberation which occurs at approximate 50ms to 80ms after the arrival of direct signal. The direct signal is the first arrival from a distanttalker to microphones. The reverberated noise represents the noise caused by late reverberation. S(n, k) denotes the STFT spectrum of direct signal and the noise caused by early reverberation. Early reverberation occurs at approximate 50ms to 80ms before the arrival of direct signal. Our approach concentrate on removing reverberated noise, which seriously corrupt the speech signal. The flow chart shows in Fig 1.



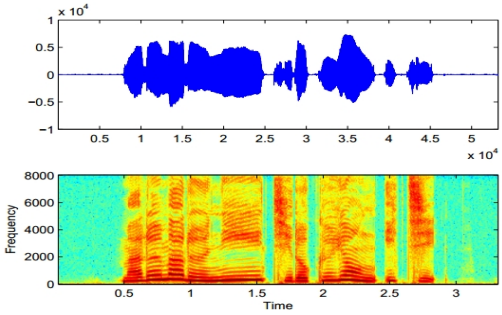


Fig. 1 The operation flow chart.

Fig. 2 Speech wave and STFT spectrum of clean speech.

B. The principle of our approach

The clean speech matrix is a sparse matrix. Every signal can be represented by a sparse matrix through a kind of transform. There are a large number of blank points of clean speech in time domain and the clean speech is combined of only a limited number of frequencies. So we can represent the clean speech as a sparse matrix on spectrogram of STFT. For example, in lounge, we use speech FAK OO21641A to represent the sparsity of the clean speech. The Fig 2 is the speech wave and STFT spectrum of the clean speech. The reverberated noise matrix is a low rank matrix. There are some preliminary experiments proving this character of reverberated noise. Table I and Table II show the proportion.of the eigenvalues which are approximately equal to zero in different environments. When eigenvalues are smaller than one three thousandth of the maximum value, which is approximate 0.02, we consider them as zeros. The result shows that the reverberated noise is low rank. The characteristics of the early reverberation depend strongly on the positions of the speaker and microphones. By contrast, the magnitude of the late reverberation is independent from the positions. In different environments or different periods of time, the reverberated noise matrix is invariant in some degree so that the reverberated noise matrix can be represented by a low rank matrix.

On the other hand, the spectrogram of the reverberation also shows some other information. Higher frequencies may only appear in short period of time. After that time, there are only lower frequencies in the spectrogram. As a result, each frame from STFT spectrum of reverberate speech is a weighted composition of a certain time period of frames from the STFT spectrum of clean speech. Especially, for high frequencies, the time period is approximate 3 or 4 frames before that frame from STFT spectrum of reverberate speech. For low frequencies, the time period is all the frames before that frame from STFT spectrum of reverberated speech. The Fig 3 and 4 is speech wave and STFT spectrum of reverberated speech. High frequencies of STFT spectrum of reverberated speech are shown in the red square as an example. Low frequencies of STFT spectrum of reverberated speech are shown in the blue square. Based on the principle mentioned above, we decompose the reverberate

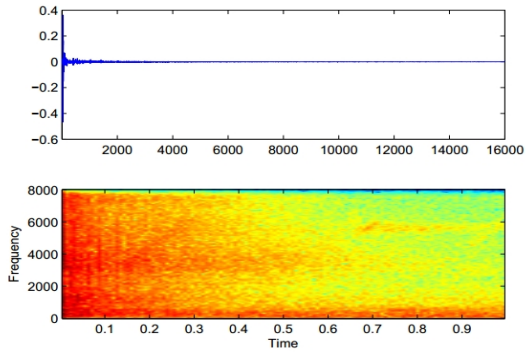


Fig. 3 Impulse responses in time and frequency domain in lounge.

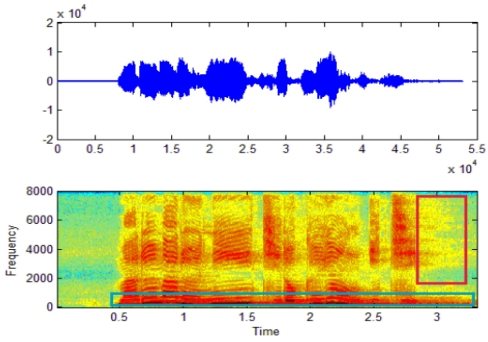


Fig. 4 Speech wave and STFT spectrum of reverberate speech.

speech matrix into a low rank matrix, which delegates the reverberated noise matrix, and a sparse matrix, which delegates the clean speech matrix.

C. RPCA

RPCA is a effective method to decompose a matrix into a sparse matrix and a low rank matrix. As stated above, we want to obtain two components which are the solution of the following equation

$$\min_{N,S} rank(N) + l \times \| S \|_0, s.t. X = S + N, (3)$$

in which X □ Rn×k, S □ Rn×k, N □ Rn×k. $\| \bullet \|_0$ norm is the number of non-zero entries in a matrix and λ is a treadoff parameter between the ||S||0 and the rank(N). However this is a highly non-convex optimization problem and we can obtain an optimization problem by relaxing (3) to the following convex surrogate:

$$\min_{N,S} \| N \|_* + l \times \| S \|_1, s.t. X = S + N, (4)$$

$\| \bullet \|_1$ is the ℓ1 norm which is the summation of absolute values of matrix entries and $\| \bullet \|_*$ denotes the nuclear norm of a matrix which is defined as the summation of all singular values respectively. We expect N to be the reverberated noise and S to be the clean speech. We perform the decomposition as follows: First, we obtain the spectrogram of reverberated speech calculated from the STFT after overlapped framing and half padding. Second, We divide S(n, k) and X(n, k) into some small blocks. Third, we divide the matrix into real part and imaginary part. Then, we use the inexact Augmented Lagrange Multiplier (ALM) method [6] to solve (4) to obtain the real and imaginary parts of noisy spectrogram, separately. Then we synthesize Sr and Si .

$$S_R = S_r + i \times S_i, (5)$$

Finally, we obtain the dereverberated speech. Through this method we can extract clean speech out from reverberate speech.

Experimental Results

A.Database

We evaluate our approach using the CENSREC [7] databases, which are distant-talking speech's corpus and evaluation frameworks. In those databases, they simulate various environments by convolving the clean speech with room impulse responses in real environments. We use testsetB and testsetD to do our experiments. TestsetB (lounge, J-s room, meeting room and J-s bath) is a simulated reverberated speech by convolving the impulse responses with the clean speech. There are 4,004 utterances by 104 speakers (including 52 females and 52 males). TestsetD (office , in-car , lounge , meeting room) is recorded data in real environments. This data set was recorded by ten human speakers (five females and five males). There are 2,536 utterances (2,536 files) in testsetD. We choose a part of samples randomly in each environment.

B. Experimental criterions

In our experiments, we choose $\lambda = 12$ and every block has 70 columns. Larger λ means our approach presents the sparsity of clean speech matrix better. Smaller λ means our approach presents the character of reverberated noise matrix better. When $\lambda = 12$, it balances the characters of clean speech and reverberated noise well. We divide reverberated speech matrix into blocks which have 70 columns. The experimental results demonstrate that this way to divide matrix can remove the late reverberation effectively. We use some criterions to evaluate the results of our experiments. llr mean means the distortion degree of the signal. Lower llr mean indicates less distortion degree of the signal. SegSNR [8] denotes the efficacy of removing reverberated noise. Higher SegSNR indicates less power of noise. wss dist is weighted spectrum slope. Lower wss dist indicates less distortion degree of the signal. pesq mos [9] means how well the dereverated speech preserve the information of the clean speech and higher pesq mos indicates that the dereverberated speech is closer to the clean speech [10].

C. Experimental results

We compare our method with four different approaches. ”Spectral subtraction” means speech dereverberation by spectral subtraction [1], ”LSA” means using a minimum meansquare error log-spectral amplitude estimator [11] to dereverberate speech, ”Multiple Regression” means using nonlinear multiple regression to dereverberate speech [3]. Table III, Table V, Table VI, Table IV, Table VII show the results of our method in testB. Fig 5 and Fig 6 show the results of our method in testD. The results indicate the efficacy of our method in terms of preserving speech information for the llr mean is low, the pesq mos is high and wss dist is low. SegSNR is high, which proves our approach’s efficacy in terms of dereverberating speech. As we can obtain from the criterions above, our method is useful to dereverberate speech and has stability in different environments. Our method is the best in the four approaches.

TABLE I
THE RATIO OF LARGE EIGENVALUE , IN TESTB

sparse ratio B	J-s bath	J-s room	lounge	meeting room
ratio	0.51	0.52	0.51	0.50

TABLE II
THE RATIO OF LARGE EIGENVALUE , IN TESTD

sparse ratio D	incar	lounge	meeting_room	office
ratio	0.51	0.52	0.52	0.52

TABLE III
THE RESULT OF USING LOW RANK METHOD TO ENHANCE SPEECH IN TESTB

lowrankB	llr_mean	segSNR	wss_dist	pesq_mos
J-s bath	0.74	-1.81	58.81	2.38
J-s room	1.22	-1.78	72.28	2.00
lounge	1.04	-2.19	60.82	2.72
meeting room	0.66	-1.58	60.10	2.45

TABLE IV
THE RESULT OF USING MULTIPLE REGRESSION METHOD TO ENHANCE SPEECH IN TESTB

MRB	llr_mean	segSNR	wss_dist	pesq_mos
J-s bath	0.79	-3.15	63.45	2.31
J-s room	1.26	-2.90	76.90	2.01
lounge	1.16	-2.81	71.30	2.50
meeting room	0.73	-1.02	51.01	1.82

TABLE V
THE RESULT OF USING SUBTRACTION SPECTRUM METHOD TO ENHANCE SPEECH IN TESTB

SSB	llr_mean	segSNR	wss_dist	pesq_mos
J-s bath	0.94	-4.11	71.58	2.38
J-s room	1.42	-3.68	80.62	2.00
lounge	1.16	-3.53	67.69	2.71
meeting room	0.90	-3.22	69.22	2.45

TABLE VI
THE RESULT OF USING LSA METHOD TO ENHANCE SPEECH IN TESTB

LSAB	llr_mean	segSNR	wss_dist	pesq_mos
J-s bath	0.93	-4.16	68.43	2.37
J-s room	1.40	-3.72	79.40	2.00
lounge	1.15	-3.55	66.30	2.70
meeting room	0.87	-3.29	67.85	2.43

TABLE VII
THE RESULT OF NOT USING ANY METHOD IN TESTB

RSB	llr_mean	segSNR	wss_dist	pesq_mos
J-s bath	0.75	-4.26	59.04	2.38
J-s room	1.21	-3.97	73.34	2.00
lounge	1.06	-3.63	61.45	2.72
meeting room	0.66	-3.19	59.68	2.45

Conclusions

We proposed an approach based on RPCA to dereverberate speech. Experimental results demonstrate that our method outperforms many traditional methods in most environments. And this method has robustness in preserving the clean speech’s information and dereverberating speech. Even if in some environments some criterions aren’t good enough, they are approximately equal to the criterions in other methods (in testB). In general, our method is the best. In the future, we will do some other listening tests to further examine the presented method. And with some post processing the dereverberated speech will be more effective.

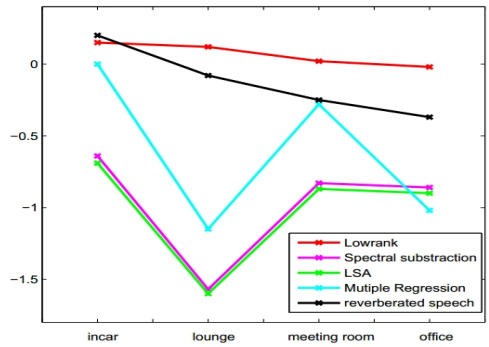


Fig. 5 SegSNR in different environments using different ways.

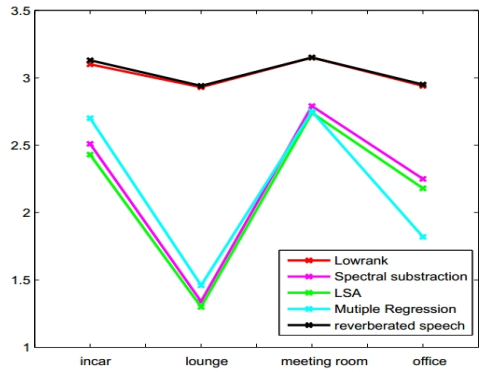


Fig. 6 PESQ in different environments using using different ways.

References

[1] S. Boll, Suppression of acoustic noise in speech using spectral subtraction, Acoustics, Speech and Signal Processing, IEEE Transactions on 27 (2) (1979) 113–120.

[2] R. Martin, I. Wittke, P. Jax, Optimized estimation of spectral parameters for the coding of noisy speech, in: Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on, Vol. 3, IEEE, 2000, pp. 1479–1482.

[3] L. Weifeng, C. Miyajima, T. Nishino, I. Katsunobu, K. Takeda, F. Itakura, Adaptive nonlinear regression using multiple distributed microphones for in-car speech recognition, IEICE transactions on fundamentals of electronics, communications and computer sciences 88 (7) (2005) 1716–1723.

[4] E. J. Candes, X. Li, Y. Ma, J. Wright, Robust principal component ` analysis?, Journal of the ACM (JACM) 58 (3) (2011) 11.

[5] Z. Chen, D. P. Ellis, Speech enhancement by sparse, low-rank, and dictionary spectrogram decomposition, in: Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on, IEEE, 2013, pp. 1–4.

[6] J. Wright, A. Ganesh, S. Rao, Y. Peng, Y. Ma, Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization, in: Advances in neural information processing systems, 2009, pp. 2080–2088.

[7] M. Nakayama, T. Nishiura, Y. Denda, N. Kitaoka, K. Yamamoto, T. Yamada, S. Tsuge, C. Miyajima, M. Fujimoto, T. Takiguchi, et al., Censrec-4: Development of evaluation framework for distant-talking speech recognition under reverberant environments, in: Ninth Annual Conference of the International Speech Communication Association, 2008.

[8] Y. Hu, P. C. Loizou, Evaluation of objective quality measures for speech enhancement, Audio, Speech, and Language Processing, IEEE Transactions on 16 (1) (2008) 229–238.

[9] T. H. Falk, W.-Y. Chan, A non-intrusive quality measure of dereverberated speech, in: Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC), 2008.

[10] Y. Hu, P. C. Loizou, Evaluation of objective measures for speechenhancement., in: Interspeech, 2006.

[11] Y. Ephraim, H. L. Van Trees, A signal subspace approach for speech enhancement, Speech and Audio Processing, IEEE Transactions on 3 (4) (1995) 251–266.