Frequency-domain Dereverberation on Speech Signal using Surround Retinex

Mingming Zhang[1,2], Weifeng Li[1,2], Longbiao Wang[3], Jianguo Wei[4], Zhiyong Wu[1,2], Qingmin Liao[1,2]
[1]Shenzhen Key Lab. of Information Sci&Tech/Shenzhen Engineering Lab. of IS&DRM
[2]Department of Electronic Engineering/Graduate School at Shenzhen, Tsinghua University, China
[3]Nagaoka University of Technology, Japan
[4]School of Computer Science and Technology, Tianjin University, China

Abstract

In this paper, we propose a novel and practical single channel dereverberation scheme, which utilizes surround retinex model in frequency domain. A dereverberation filter is derived by the proposed method for suppressing the environmental reflections. The proposed algorithm can achieve effective dereverberation with a more reasonable computational complexity than conventional methods. Experimental results also reveal an improvement in automatic speech recognition (ASR) performance even in severely reverberation environments.

Index Terms: speech dereverberation, surround retinex, image illumination

1. Introduction

Reverberation is a kind of noise produced in a closed space, such as in the car, lounge, meeting room, office and so on, which is from the speech reflection from walls and other objects. Speech reverberations degrade the speech fidelity and the performance of automatic speech recognition. Speech dereverberation is widely used in hand-free telephone, hearing aid, tele-conferencing system, as well in high fidelity voice-recording system and automatic speech recognition (ASR).

Speech dereverberation methods, also called reverberation cancelling, are generally divided into two categories: single-microphone and multi-microphone cancellation. Generally, multi-microphone cancellation methods can achieve better performance, but it needs microphone array and is easy to come true, while the single microphone system has simple hardware. In this paper, we focus on single-microphone speech dereverberation.

One direct method in speech dereververation is inverse filtering [1,2], which needs to obtain the impulse response from the response to a known signal, which is a non-trivial task for the impulse response does not have stable reversibility. Another method is Cepstral Filtering techniques [3], but the deconvolution when reconstructing with cepstral is difficult. Therefore some improved methods are proposed. In [4], the author described a modified cepstral filtering to estimate the impulse response. [5] proposed an idea on speech dereverberation using backward estimation of the late reverberation spectral variance.

In this paper, we focus on estimating the channel distortion or impulse response in frequency domain by using surround Retinex (SR) model, which has been used for the image enhancement. The feasibility is that a speech spectrogram which can be considered as an image, and a corrupt speech spectrogram by the channel distortion or the reflections is similar to an image with the scene illumination contaminated. SR model has been proved to be an effective tool for image enhancement for its capability of estimating the image illumination source from a degraded image. We utilized this model for estimating the channel distortion or impulse response (We name it as "reverberation estimation") from a reverberant-speech spectrogram. Figure 1 is the flow chart of our algorithm, and the key block is "reverberation estimation".

This paper is organized as follows: In section 2 we will introduce the Surround Retinex model from a point of view of image processing. In section 3 we present our method of speech dereverberation using the Surround Retinex model. The evaluation criterions and our experimental results on real-recorded reverberant speech database will be presented in section 3. Finally, we draw our conclusions in section 4.

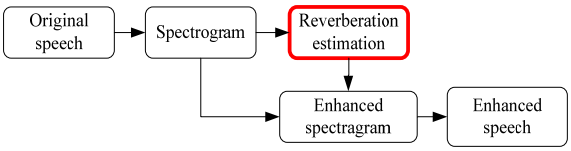

Figure 1: Flow chart of our algorithm

2. Surround Retinex Model

"Retinex" is a synthetic word with two words, Retina and Cortex, which is an image enhancement theory based human visual system and developed and presented by Land and McCann in 1971.This theory states that the perceived image in a natural scene has a strong correlation with reflectance, even though the amount of visual light reaching the eye depends on the reflectance and illumination [6]. In other words, the human visual system can perceive colors even in difficult illumination conditions by relying on the reflectance of the scene and neglecting the scene illumination. This theory is based on the reflectance image model, which can be formulated as follows:

$$F(x,y) = R(x,y) \cdot I(x,y) \qquad (1)$$

where $F(x,y)$ denotes an perceived image of natural scene, and $R(x,y)$ denotes the reflectance and $I(x,y)$ denotes the illumination. $R(x,y)$ only depends on the reflectivity of the

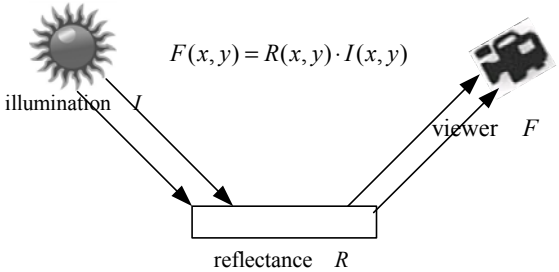

Figure 2: Retinex theory

scene surface, and *L(x,y)* is determined by the illumination source and related to the amount of illumination, just as figure 2. In Retinex model, the image *F(x,y)* is a composite of two parts: the first part *I(x,y)* is light luminance, corresponding to the low frequency, while another part *R(x,y)* is reflectance luminance, corresponding to the high frequency. The two parts are called luminance image and reflectance image [7].

The key of Surround Retinex, which mimics the ability of human visual system in perceiving colors even in difficult illumination conditions as mentioned above, is to estimate the illumination image from the original images. Jobson et al. suggested that *I(x,y)* can be estimated as a blurred version of *F(x,y)* [8,9,10], which can be formulated as:

$$\hat{I}(x,y) = F(x,y) * G(x,y) \qquad (2)$$

where "*" denotes the convolution operator, *G(x,y)* denotes a smoothing kernel. Here, the smoothing kernel takes the form of a Gaussian:

$$G(x,y) = \lambda \cdot e^{\frac{(x^2+y^2)}{2c^2}} \qquad \iint G(x,y)dxdy = 1 \qquad (3)$$

where *c* is filtering radius, the larger *c* is, the more sharpened the image is.

Therefore the reflectance image *R(x,y)* can be described as:

$$\hat{R}(x,y) = \frac{F(x,y)}{\hat{I}(x,y)} \qquad (4)$$

where $\hat{R}(x,y)$ is the recovered illumination-independent component. Taking a face image as example, the result of the surround model is shown in Fig. 3. It can be observed that the illumination component of the face is estimated quite well, and the illumination-invariance component, "clean face", is recovered.

$F(x,y)$ $\hat{I}(x,y)$ $\hat{R}(x,y)$

Figure 3: An example of SR. *F(x,y)* is the original image. $\hat{I}(x,y)$ is the estimated illumination image. $\hat{R}(x,y)$ is the estimated "clean" image.

3. Dereverberation with Surround Retinex

In time-domain, a reverberant signal can be modeled mathematically as follows:

$$x(t) = s(t) * h(t) \qquad (5)$$

where *x(t)* is the perceived speech signal, *s(t)* is the original speech signal representing the "clean" speech, and *h(t)* is the impulse response of channel between microphone and the signal source. Reverberant signal *x(t)* is the convolution of the "clean" speech signal and the impulse response.

In the frequency domain, the model is given by

$$X(\omega) = S(\omega) \cdot H(\omega) \qquad (6)$$

In the discrete speech spectrogram, it has the following form:

$$X(n,f) = S(n,f) \cdot H(n,f) \qquad (7)$$

where *n* is the index of speech frame, and *f* is the discrete frequency.

Compared with the Surround Retinex model, the reverberation signal model has the similar form. The corresponding relationship between them is shown in the figure 4.

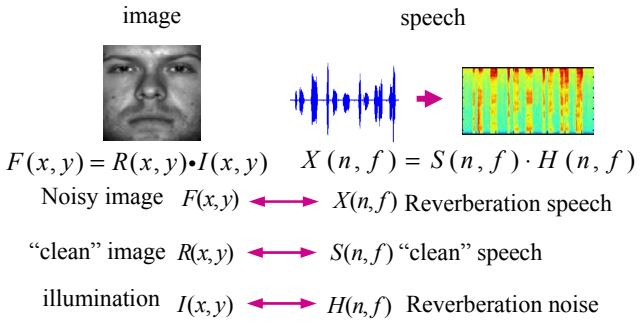

Figure 4: SR model and reverberation signal model

Figure 5 shows our speech dereverberation processing using the SR algorithm. There are five steps for the entire algorithm:

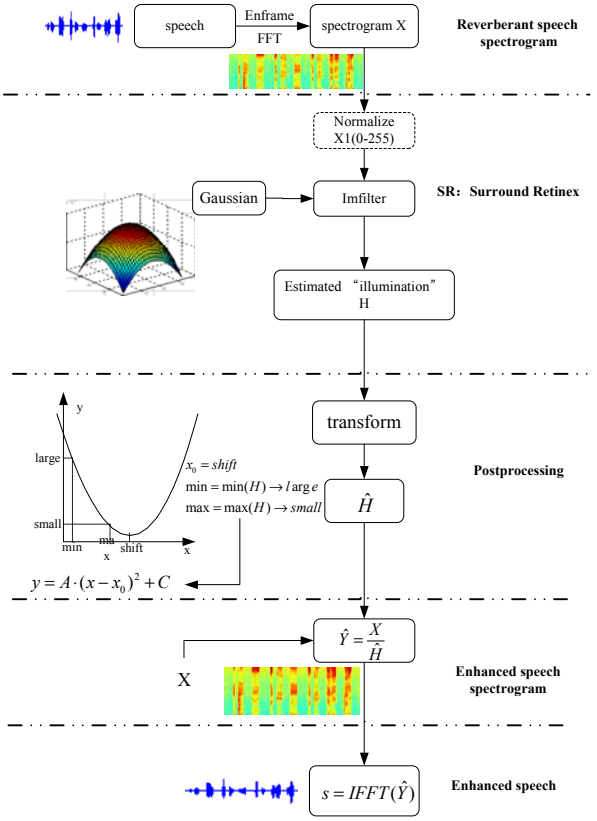

Figure 5: Flow diagram of speech dereverberation

Step 1:

The first step is to obtain the spectrogram *X* of the original speech. We use a hamming window and the frame length of fast Fourier Transform (FFT) is N=128, and an overlap between two successive frame is 50%.

Step 2:

This step is the most important part of the algorithm, because it's the key of SR algorithm. Firstly, the input spectrogram X is normalized to a grey intensity range of 0-255, then the normalized image is filtered by a Gaussian kernel, just as the equation (2) shows. Here, some Gaussian parameters are set like this: the size of Gaussian is 5*5, and the variance is 1. Finally we can obtain the estimated illumination H.

Step 3:

This step is a postprocessing after Gaussian filtering. For a gray-scale face image, the important information like eyes and mouth is expressed with low-value pixels. Different from the gray-scale image, high-value pixels in speech spectrogram stand for the most important information of speech. Therefore, we need to post-process the estimated illumination to address this kind of inverse correspondence, i.e., transforming the small values to the large ones and vice versa. Many transform functions can be used. In this paper, after testing many transform function we experimental select the left half part of the quadratic function as our transform function as Figure 5 shows. There are the following parameters in the quadratic function, the shift of the curve, the largest and smallest values. In this paper, these parameters are set as: *large=1.5*max(H), small=5, and shift=max(H)+0.1*, where *max(H)* is the maximum of H. With such three parameters, we can obtain the quadratic function.

Step 4:

After step 3, we have obtain the estimated reverberation spectrum $\hat{H}$, then we use the equation (4) to obtain the "clean" speech spectrogram $\hat{Y}$.

Step 5:

The final step is to transform the "clean" spectrogram to the enhanced speech using the inverse fast Fourier Transform transformation (IFFT).

4. Experiments

In this section, we evaluate the performance of our novel method, surround Retinex (SR) based speech dereverberation, and present the experimental results.

4.1. Experiment data

The proposed approach was evaluated on a realistic speech recognition task under reverberant environments. The datasets was taken from the CENSREC-4 database [11]. The test set D was recorded in real reverberant environments by 10 human speakers (five females and five males) using two microphones (close-talking and distant-talking), in which the speech recorded by a distant microphone was selected for the evaluation. There were four reverberant environments (in-car, lounge, meeting room, and office). For ASR experiments, the clean training data, in which the total number of the utterance was 8,440 by 110 speakers (55 females and 55 males), were selected for training the acoustic model. For comparisons, a parametric formulation of the generalized spectral subtraction (GSS) [12] was applied.

4.2. Evaluation criterions

Two criterions were used in this paper: objective evaluation of the enhanced speech and the recognition rate of automatic speech recognition (ASR) system.

In [13], some objective quality measures for speech enhancement are discussed. We choose the following parameters to evaluate the results: speech, background, and overall quality, which are obtained by linearly combing existing objective measures to form a new measure. These measures can better describe the correlation between the subjective quality measure and the objective quality measure [13], and are formulated as follows:

speech = 3.093 - 1.029*LLR + 0.603*PESQ-0.009*WSS;

background = 1.634 + 0.478 *PESQ - 0.007*WSS + 0.063*segSNR;

overall quality = 1.594 + 0.805*PESQ - 0.512*LLR - 0.007*WSS.

where LLR is log-likelihood ratio, and segSNR is segment Signal-to-Noise ratio, and WSS is weighted spectral slope, and PESQ is perceptual evaluation of speech quality. For speech, background, and overall quality, the larger of the value is, the better performance is.

Another quality measurement is the recognition accuracies of ASR. In our ASR system, the acoustic models consist of 18 phone models that have five states (three states for 'sp'.). we extracted the mel-frequency cepstral coefficients (MFCCs) from the enhanced speech data. The speech signal was windowed with a 20-ms Hamming window every 10 ms (with a pre-emphasis). A 24-channel mel-filter bank (MFB) analysis was applied, and finally the log MFB outputs were converted into 12 MFCCs through Discrete Cosine Transformation (DCT). The feature vectors of ASR system are 39-dimensional vectors consisting of 12-dimensional MFCC parameters and log-energy, along with their delta and delta-delta parameters.

4.3. Experiment results

For each environment, we obtained an average value of all the speech performance (10 persons and about 50 speech utterances for each person). Figure 6 and figure 7 show four speech spectrograms in two environments: lounge and office. From top to down, the spectrograms of clean speech (recorded by the close-talking microphone), noisy speech (recorded by the distant-talking microphone), enhanced speech after GSS method, enhanced speech after our proposed method, are illustrated. It can be observed that the spectrograms of noisy speech are contaminated by the reverberant noise. GSS is effective for reducing the reverberant noise. Compared with GSS, SR method shows a better capacity of removing the reverberant distortions. This demonstrates the effectiveness of our proposed SR method.

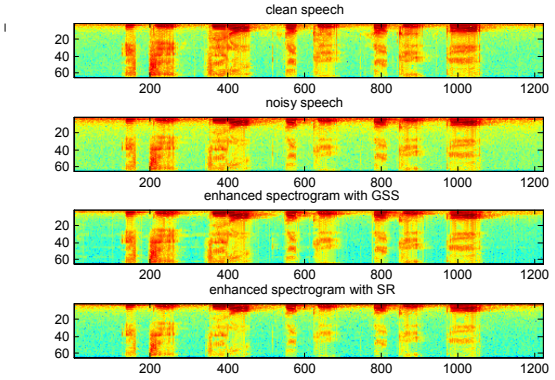

Figure 6: Spectrogram of four kinds of speech in lounge environment

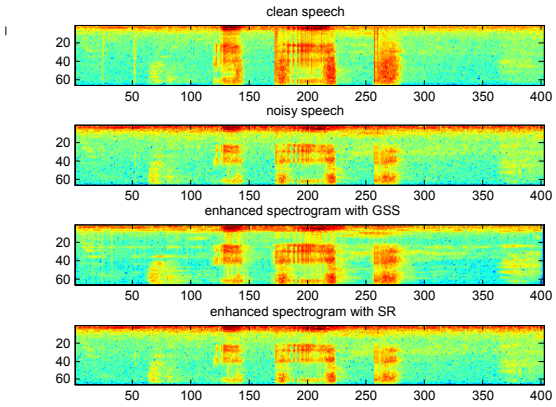

Figure 7: Spectrogram of four kinds of speech in office environment

Table 1 shows the results of different methods using the speech quality measures. Whatever the environment is, the background of our proposed method can obtain better performance, compared with GSS method. In average SR-enhanced speech performs better than both noisy speech and GSS-enhanced speech.

Table 1. Results of quality measures of different methods. "meeting" denotes the meeting-room environment.

		in-car	lounge	meeting	office	Average
speech	noise	4.29	4.34	4.42	4.47	4.38
	GSS	4.1	4.16	4.21	4.38	4.21
	our proposed method (SR)	4.29	4.34	4.44	4.51	4.40
background	noise	2.34	2.41	2.46	2.53	2.44
	GSS	2.68	2.73	2.69	2.87	2.74
	our proposed method (SR)	2.97	2.96	2.97	3.01	2.98
overall quality	noise	3.62	3.67	3.7	3.79	3.69
	GSS	3.54	3.59	3.58	3.8	3.63
	our proposed method (SR)	3.64	3.71	3.73	3.84	3.73

Table 2 shows the recognition results of automatic speech recognition of different methods. The recognition accuracies of noisy speech depend on the reverberation environments. The accuracy can be less than 50% when the environment is seriously reverberant (e.g. "lounge"), while the accuracy of GSS and RS both can reach up to 76%. In the "meeting room" and "office" environments, our proposed SR method can obtain a better result than GSS.

Table 2. Recognition accuracies (as percentages) of ASR obtained from different methods. "meeting" denotes the meeting room environment

	in-car	lounge	meeting	office	Average
noise	76.27	43.83	89.12	85.18	73.6
GSS	86.32	76.67	85.65	83.29	82.98
our proposed method (SR)	86.48	76.51	88.96	85.46	84.35

5. Conclusions

One of the main challenges of speech dereverberation algorithm is the estimation of the reverberation noise. In this paper, we proposed a method to estimate the speech reverberation by using a image processing algorithm, Surround Retinex. with the rationality that Surround Retinex for image enhancement can be used to speech dereverberation and also obtain the better results than GSS. However, our results reveal that we have not obtained significant improvement in some reveberant environments. The step 4, postprocessing, plays an important role in this algorithm, and the quadratic function may not be the best. Therefore this part should continue to be improved.

Our research provides a novel idea for speech dereverberation by using image processing algorithm. Other image enhance algorithms except Surround Retinex are worth exploring in order to improve speech dereverberatin performance.

6. Acknowledgements

This work was supported in part by Shenzhen Basic Research Grant JCYJ20120831165730913, in part by National Natural Science Foundation of China (No. 61175016) and in part by National Basic Research Program of China (No. 2013CB329305).

7. References

[1] S. T. Neely, J. B. Allen, "Invertibility of a room impulse response," J. Acoust. Soc. Amer., 1979, Vol.66, pp.165-169.

[2] M. Miyoshi, M. and Kaneda, Y. "Inverse filtering of room acoustics," IEEE Trans. ASSP, 1988, vol.36(2). pp.145-152.

[3] D. Bees, M. Blostein, and P. Kabal, "Reverberant speech enhancement using cepstral processing," ICASSP-91, pp.977-980, 1991.

[4] BEES D., BLOSTEIN M., KABALP, "Reverberant speech enhancement using cepstral processing," ICASSP-91, 1991: 977-980, 1991.

[5] E. A. P. Habets, S. Gannot, and I. Cohen, "Speech dereverberation using backward estimation of the late reverberant spectral variance," In Proc. IEEE Conf. Electrical & Electronics Engineers in Israel (IEEEI), Israel, Dec. 2008, pp. 384–388.

[6] Štruc V. and Pavešić N.: *Photometric normalization techniques for illumination invariance*, IGI-Global, pp. 8-12, 2011. (The chapter is based on the internal report entitled: *Performance evaluation of photometric normalization techniques for illumination invariant face recognition).*

[7] G. Orsini, G. Ramponi, P. Carrai, et al., "A Modified Retinex for Image Contrast Enhancement and Dynamics Contro," Image Processing, 2003, 3: 14—17.

[8] D. H. Choi, I. H. Jang, M. H. Kim, and N. C. Kim, "Color Image Enhancement Based on Single-Scale Retinex with a JND-Based Nonlinear Filter," in Proc. in Proc. of IEEE International Symposium on Circuits and Systems, 2007, pp. 3948-3951.

[9] D. J. Jobson, Z. Rahman , G. A. Woodell, "Properties and Performance of a Center/Surround Retinex," IEEE Transactions on Image Proces2sing, 1997 ,6 (3): 4512-4621.

[10] Z. Rahman, D. J. Jobson, G. A. Woodell, "Multi-scale Retinex for Color Image Enhancement," Lausanne, Switzerland: Proceedings of International Conference on Image Processing, 1996, pp.10032-10061.

[11] T. Nishiura, R. Gruhn, and S. Nakamura, "Evaluation frame-work for distant-talking speech recognition under reverberant environments," in Interspeech-2008, 2008.

[12] B. L. Sim, Y. C. Tong, J. S. Chang, and C.T. Tan, "A parametric formulation of the generalized spectral subtraction method,"

IEEE Transactions on Speech and Audio Processing, vol.6, no. 4, pp. 328 –337, 1998.

[13] Hu Y. and Loizou P. (2006), “evaluation of objective measures for speech enhancement,” Proceedings of INTERSPEECH-2006, Philadelphia, PA, September 2006.