Blind dereverberation of single-channel speech signals using an ICA-based generative model

Jong-Hwan Lee[1], Sang-Hoon Oh[2] and Soo-Young Lee[3]

[1] Brain Science Research Center and
Department of Electrial Engineering & Computer Science
Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea
Phone: +82-42-869-5431, Fax: +82-42-869-8490, Email: jhlee@neuron.kaist.ac.kr
[2] Department of Information Communication Engineering
Mokwon University, Daejeon, 302-729, Republic of Korea
[3] Brain Science Research Center
Department of BioSystems and Dept. of Electrical Engineering & Computer Science
Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea

Abstract. In this paper, an adaptive blind dereverberation method based on speech generative model is presented. Our ICA-based speech generative model can decompose speeches into independent sources. Experimental results show that the proposed blind dereverberation model successfully performs even in non-minimum phase channels.

1 Introduction

In real room environments, sounds are corrupted with delayed versions of themselves reflected from walls. This room reverberation severely degrades intelligibility of speeches and performance of automatic speech recognition system [1]. In some applications, it is necessary to recover an unknown source signal using only observed signal through an unknown convolutive channel. This problem is called the blind deconvolution and also known as the blind dereverberation when convolving channels are room impulse responses. Almost every methods for the blind deconvolution are developed under the assumption that a source signal is independent identically distributed (IID) and non-Gaussian [2–8]. When an IID non-Gaussian source signal is convolved with a multi-path channel, the probability density function (p.d.f.) of the received signal approaches to Gaussian due to the central limit theorem. Deconvolution can then be accomplished by adapting a deconvolution filter which makes the p.d.f. of deconvolved signal away from Gaussian [2–4]. When sources are not IID such as speeches, the existing algorithms cannot be directly applied.

In this paper, we make a generative model of speeches, which linearly decompose them into independent components. In the first stage of our model, we extract independence transform matrix using independent component analysis (ICA) of natural human speech signals. Using the independence transformation of speeches, we derive blind dereverberation learning rules based on the Least Square (LS) method [6, 7].

2 ICA-based speech generative model

We adopt ICA algorithms to find efficient representations of speech signals such that their sample by sample redundancy is reduced significantly. This redundancy reduction leads nonstationary correlated speech signals to IID-like signals.

ICA assumes a source vector $\mathbf{s}$ whose components $s_i (i = 1, \cdots, N)$ are mutually independent. We can only observe linear combinations

$$\mathbf{x} = \mathbf{A}_I \mathbf{s} \tag{1}$$

where $\mathbf{A}_I$ is an $N \times N$ mixing matrix and their columns are called as basis vectors. After ICA adaptation which minimizes the mutual information among unknown sources [3, 5], estimated sources will be as independent as possible. If the observation vector is a frame of speech, we can find an independent signal vector and related basis vectors. Here we will call $\mathbf{W}_I = \mathbf{A}_I^{-1}$ as the "independence transform matrix".

To learn $\mathbf{W}_I$ from natural human speech signals, we used 10 sentences from one speaker (mcpm0), which corresponds to DR1 New England dialect of the train set in the TIMIT continuous speech corpus. 8kHz sampling was used to reduce computation time. We assumed 16 basis vectors for the ICA-based speech generative model and each speech frame were composed of 16 samples, i.e. 2ms time interval. Figure 1 shows a diagram of the speech generative model. A part of mcpm0's sentence, 'she had your dark suit', can be generated with independent sources through trained 16 basis vectors.

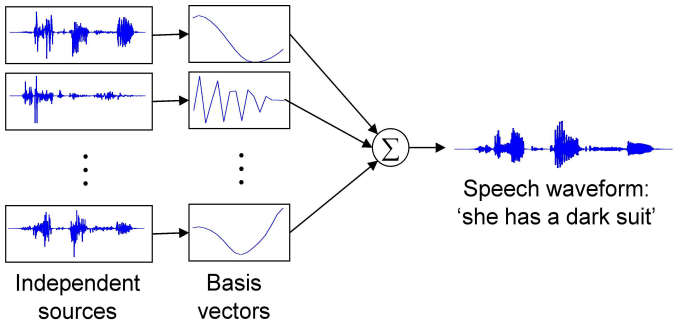


Fig. 1. Diagram of speech generative model with 16 trained basis vectors.

To check the independence transform property of $\mathbf{W}_I$, joint p.d.f. of two adjacent samples was estimated. Figure 2 (a) shows the contour-plots of joint p.d.f. of two adjacent samples for the mcpm0's sentences. Although adjacent samples in natural human speech signals are highly correlated, their dependencies are very much reduced when the independence transform matrix is applied as shown in Fig.2 (b).

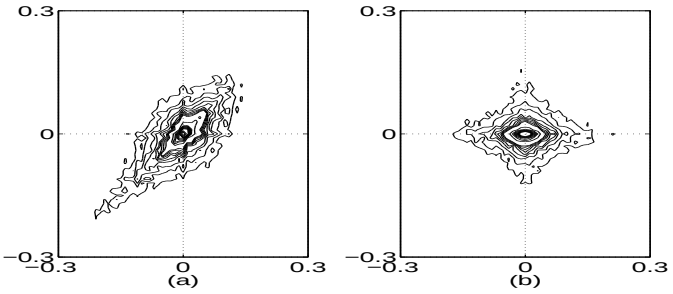


Fig. 2. Contour-plots of joint p.d.f. for mcpm0's sentences. (a) two adjacent samples in original unprocessed speech signal, (b) 1st and 2nd components transformed with $\mathbf{W}_I$.

3 Learning rule for nonminimum-phase channels

Now, we derive the algorithm for non-minimum phase channel based on the LS measure [6, 7]. In the dereverberation block of Fig.3, let's define $\hat{\mathbf{U}}^F \equiv \mathbf{W}_{\text{fft}}\ \hat{\mathbf{u}}$, $\mathbf{X}^F \equiv \mathbf{W}_{\text{fft}}\ \mathbf{x}$, and $\mathbf{W}^F \equiv \mathbf{W}_{\text{fft}}\ \mathbf{w}$, where $\mathbf{W}_{\text{fft}}$ denotes discrete Fourier transform matrix and the superscript ^{F} means frequency domain representation. Now the dereverberated speech signal $\hat{\mathbf{u}}$ can be expressed in the frequency domain as,

$$\hat{\mathbf{U}}^F = \mathbf{W}^F \otimes \mathbf{X}^F, \tag{2}$$

where $\otimes$ means component by component multiplication. IID-like signal $\mathbf{u}$ can be expressed in the frequency domain as,

$$\begin{aligned}\mathbf{U}^F &= \mathbf{W}_{\text{fft}}\ \mathbf{u} = \mathbf{W}_{\text{fft}}\mathbf{W}_I\ \hat{\mathbf{u}} \\ &= \mathbf{W}_{\text{fft}}\mathbf{W}_I\mathbf{W}_{\text{fft}}^{-1}\ (\mathbf{W}^F \otimes \mathbf{X}^F) = \mathbf{W}_\alpha^F\ \hat{\mathbf{U}}^F,\end{aligned} \tag{3}$$

where $\mathbf{W}_\alpha^F \equiv \mathbf{W}_{\text{fft}}\mathbf{W}_I\mathbf{W}_{\text{fft}}^{-1}$.

The LS cost function in the frequency domain corresponds to

$$\mathbf{J}_{LS} \equiv \sum_{\text{all fft points}} |\mathbf{U}^F - \text{fft}\{g(\mathbf{u})\}|^2 = \sum_{\text{all fft point i}} |e_i|^2 \tag{4}$$

where $g(\cdot)$ is the Fisher score function [5] and e_i is the i-th component of $(\mathbf{U}^F - \text{fft}\{g(\mathbf{u})\})$. We can obtain the update rule by minimizing $\mathbf{J}_{LS}$ with respect to $\mathbf{W}^F$. That is, in matrix formulation,

$$\frac{\partial \mathbf{J}_{LS}}{\partial \mathbf{W}^{F\ *}} = \{\mathbf{W}_\alpha^{F\ H}(\mathbf{U}^F - \text{fft}\{g(\mathbf{u}\})\} \otimes \mathbf{X}^{F\ *} \tag{5}$$

where superscript ^{H} denotes the Hermitian operator and * is the complex conjugate. Finally, using the relative gradient ([8]),

$$\begin{aligned}\Delta\mathbf{W}^F &\propto -\frac{\partial \mathbf{J}_{LS}}{\partial \mathbf{W}^{F\ *}} \otimes \mathbf{W}^{F\ *} \otimes \mathbf{W}^F \\ &= -\{\mathbf{W}_\alpha^{F\ H}(\mathbf{U}^F - \text{fft}\{g(\mathbf{u})\})\} \otimes \hat{\mathbf{U}}^{F\ *} \otimes \mathbf{W}^F.\end{aligned} \tag{6}$$

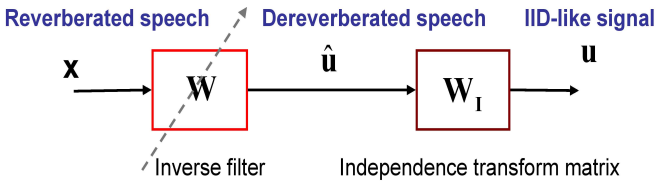


Fig. 3. Proposed blind dereverberation method with speech generative model.

4 Experimental results

We conducted blind dereverberation experiments using simulated room impulse response. During this deconvolution phase the independence transform matrix is fixed to the previously-trained values. To get the simulated room impulse response we used the commercial software 'Room Impulse Response v2.5' which assumes a rectangular enclosure with a source-to-receiver impulse response calculated using a time-domain image expansion method [9].

We assumed that the room dimensions are 4 m × 5 m × 3 m, a sound speed of 345 m/s, and reflection coefficients for 4 walls are 0.9, ceiling and floor are 0.7. Volume of the room is 60 m^3, and the reverberation time is 0.56 s.

Three different reverberant channels regarding the position of the source and receiver were used for experiments. The position of the source was fixed at (2 m, 2 m, 1 m), and the positions of three receivers were at (2 m, 1.7 m, 1 m), (2 m, 1.5 m, 1 m) and (2 m, 1 m, 1 m). The length of room impulse response was truncated by 512 samples. Fig.4 shows obtained three different room impulse responses. Channel distortions are much heavier as the distances are increased.

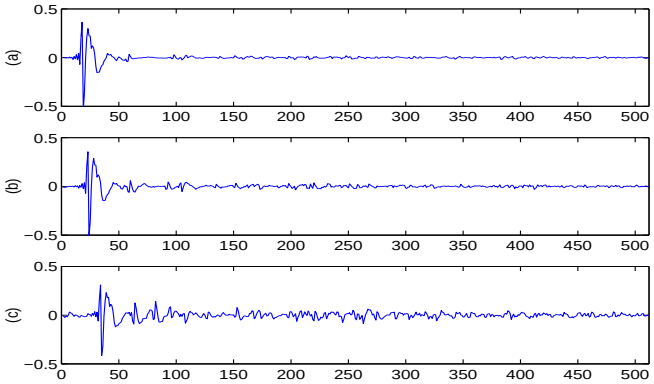


Fig. 4. Three different simulated room impulse responses. The distances between the source and receiver are (a) 0.3 m (channel 1), (b) 0.5 m (channel 2) and (c) 1 m (channel 3).

Equation (6) was used to update inverse filter $\mathbf{W}$ in Fig.3. 1024-tap delayed causal FIR (finite impulse response) filter was used for the inverse filter, and the delay was 512 samples. Ten sentences of mcpm0's speaker were used for blind dereverberation. Signal-to-reverberant component ratio (SRR) and inverse of inter-symbol interference (IISI) were used as performance measure. SRR is defined as:

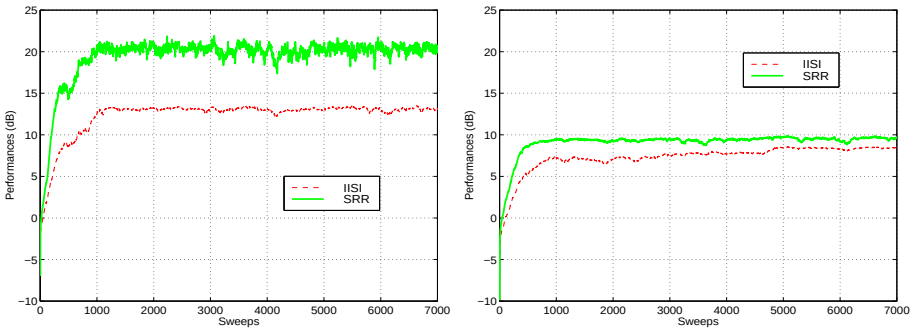
$$\text{SRR (dB)} = 10 \ \log\left(\frac{\sum_n \hat{s}_n^2}{\sum_n (\hat{s}_n - \hat{u}_n)^2}\right) \tag{7}$$

where $\hat{\mathbf{s}}$ is unknown clean speech signal and $\hat{\mathbf{u}}$ is dereverberated signal. IISI is a measure of how close the dereverberated impulse response to the delta function. IISI is defined as:

$$\text{IISI (dB)} = 10 \ \log\left(\frac{\sum_k |t_k|^2 - \max_k |t_k|^2}{\max_k |t_k|^2}\right) \tag{8}$$

where $\mathbf{t}$ is the convolution of the reverberant channel and the estimated inverse filter of the channel. Higher SRR and IISI show better result.

Figure 5 show the learning curves for the channels. Dashed and solid lines show the resulting IISI and SRR values respectively. Totally 7000 sweeps were performed, and training converged at about 1000 sweeps.



(a) for channel 1 (b) for channel 3

Fig. 5. Learning curves of IISI (dashed line) and SRR (solid line).

IISI and SRR values at the initial stage and the convergence are shown in Table 1. Final value means the average value at the convergence, and increment means the difference between the final value and the initial value. Performances are very much increased even though the room impulse responses are non-minimum phase and show about 15 ∼17 (dB) improvement in IISI and 20 ∼ 27 (dB) improvement in SRR.

To verify the speech quality before and after dereverberation, and predict the performance improvement in the automatic speech recognition system we compared the spectrograms. Spectrogram of reverberated speeches is blurred

Table 1. IISI and SRR values at the initial stage and the convergence.

	IISI (dB)			SRR (dB)		
	Initial	Final	Increment	Initial	Final	Increment
Channel 1	-4.1	13.0	17.1	-7.0	20.0	27.0
Channel 2	-4.6	12.8	17.4	-7.3	20.0	27.3
Channel 3	-6.8	8.5	15.3	-9.9	9.6	19.5

by the room impulse response especially in the mid and high frequency ranges. Those corrupted frequency structure could be recovered after dereverberation process and we can expect that speech recognition rate would be improved.

5 Conclusion

In this paper, a method for blind dereverberation based on speech generative model was proposed and LS-based learning rule was derived. Proposed blind dereverberation method was successfully applied to the simulated room impulse responses even though it is non-minimum phase and shows around 20 (dB) improvement in SRR and IISI.

Acknowledgment

This research was supported as a Brain Neuroinformatics Research Program by Korean Ministry of Science and Technology.

References

1. Haas, H.: The influence of a single echo on the audibility of speech. Journal of the Audio Engineering Society. **20**(2) (1972) 146–159
2. Shalvi, O., Weinstein, E.: New criteria for blind deconvolution of nonminimum phase systems (channels). IEEE Trans. on Information Theory. **36**(2) (1990) 312–321
3. Bell, A. J., Sejnowski, T. J.: An information-maximization approach to blind separation and blind deconvolution. Neural Computation. **7**(6) (1995) 1004–1034
4. Cichocki, A., Amari, S.: Adaptive blind signal and image processing - Learning algorithms and applications. John Wiley & Sons, Ltd. (2002)
5. Lee, T. W.: Independent component analysis - Theory and applications. Boston: Kluwer Academic Publisher. (1998)
6. Bellini, S.: Bussgang techniques for blind deconvolution and equalization. In Blind Deconvolution (S. Haykin, ed.). Englewood Cliffs, New Jersey: Prentice Hall. (1994) 8–52
7. Godfrey, R., Rooca, F: Zero memory non-linear decvonolution. Geophysical Prospecting. **29** (1981) 189–228
8. Cardoso, J. F., Laheld, B. H.: Equivariant adaptive source separation. IEEE Trans. on Signal Processing. **44**(12) (1996) 3017–3030
9. http://www.dspalgorithms.com/room/room25.html