SPEECH DEREVERBERATION BASED ON PROBABILISTIC MODELS OF SOURCE AND ROOM ACOUSTICS

Tomohiro Nakatani[†‡] *Biing-Hwang Juang*[†‡] *Keisuke Kinoshita*[†] *Masato Miyoshi*[†]

[†]NTT Communication Science Labs., NTT Corporation
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237 Japan
[‡]School of ECE, Georgia Institute of Technology, USA
nak@cslab.kecl.ntt.co.jp juang@ece.gatech.edu

ABSTRACT

This paper proposes a new single channel speech dereverberation method, in which the features of source signals and room acoustics are represented by probabilistic density functions (pdf) and the source signals are estimated by maximizing a likelihood function defined based on the pdfs. Two types of pdfs are introduced for the source signals, based on two essential speech signal features, harmonicity and sparseness, while the pdf for the room acoustics is defined based on an inverse filtering operation. The EM algorithm is used to solve this maximum likelihood problem efficiently. The resultant algorithm elaborates the initial source signal estimate given solely based on its source signal features by integrating them with the room acoustics feature through the EM iteration. The effectiveness of the present method is shown in terms of the energy decay curves of the dereverberated impulse responses.

1. INTRODUCTION

Speech signals captured by a distant microphone in an ordinary room inevitably contain reverberation, which has detrimental effect on the perceived quality and intelligibility of the speech signals and degrades the performance of automatic speech recognition (ASR) systems. It is reported, for example, that the recognition performance cannot be improved when the reverberation time is longer than 0.5 sec even when using acoustic models that have been trained under a matched reverberant condition [1]. Dereverberation of the speech signal is essential, whether it is for high quality recording and playback or for ASR.

Although blind dereverberation of a speech signal is still a challenging problem, several techniques have recently been proposed. Some researchers have proposed techniques that decorrelate the observed signal while preserving the correlation within a short time segment of the signal [2, 3]. Others proposed methods that estimate and equalize the poles in the acoustic response of the room [4, 5]. Also proposed were two new approaches based on essential features of speech signals, namely harmonicity (Harmonicity based dERverBeration, HERB [6]) and sparseness (Sparseness Based Dereverberation, SBD [7]). These methods make extensive use of the respective speech features in their initial estimate of the source signal. The initial source signal estimate and the observed reverberant signal are then used together for estimating the inverse filter for dereverberation, which allows further refinement of the source signal estimate. To obtain the initial source estimate, HERB utilizes an adaptive harmonic filter, and SBD utilizes a spectral subtraction based on minimum statistics. It has been shown experimentally that these methods greatly improve the ASR performance of the observed reverberant signals if the signals are sufficiently long.

Although HERB and SBD effectively utilize speech signal features in obtaining dereverberation filters, they do not provide analytical frameworks within which their performance can be optimized. In this paper, we reformulate them as a maximum likelihood (ML) estimation problem, in which the source signal is determined as one that maximizes the likelihood function given the observed signals. For this purpose, we introduce two probabilistic density functions (pdf) for the initial source estimates and the dereverberation filter, and maximize the likelihood function based on the Expectation-Maximization (EM) algorithm. Experimental results show that the performance of HERB and SBD can be further improved in terms of the energy decay curves of the dereverberated impulse responses given the same number of observed signals.

Section 2 defines the Fourier spectra used in this paper. Sections 3 and 4 describes the present method. Section 5 reports the experimental results. Section 6 summarizes our conclusions.

2. SHORT- AND LONG-TIME FOURIER SPECTRA

With our approach, it is important to integrate information on speech signal features, which account for the source characteristics, and on room acoustics features, which account for the reverberation effect. In general, the successive application of short-time frames of the order of tens of milliseconds is useful for analyzing such time-varying speech features, while a relatively long-time frame of the order of thousands of milliseconds is often required to compute room acoustics features. In this paper, we introduce two types of Fourier spectra based on these two analysis frames, a short-time Fourier spectrum (STFS) and a long-time Fourier spectrum (LTFS). We denote the respective frequency components in the STFS and in the LTFS by a symbol with a suffix "$^{(r)}$" as $s^{(r)}_{l,m,k}$ and a symbol without a suffix as $s_{l,k}$, where l and m are indices of long-time and short-time frames, respectively, and k is the frequency index. As shown in the following, a short-time frame can be taken as a component of a long-time frame; therefore, a frequency component in an STFS has both suffixes, l and m. The two spectra are defined as

$$s^{(r)}_{l,m,k} = 1/K^{(r)} \sum_{n=0}^{K^{(r)}-1} g^{(r)}[n] s[t_{l,m}+n] e^{-j2\pi kn/K^{(r)}},$$

$$s_{l,k} = 1/K \sum_{n=0}^{K-1} g[n] s[t_l+n] e^{-j2\pi kn/K},$$

where $s[n]$ is a digitized waveform signal, $g^{(r)}[n]$ and $g[n]$, $K^{(r)}$ and K, and $t_{l,m}$ and t_l are window functions, the number of discrete Fourier transformation (DFT) points, and time indices for an STFS and an LTFS, respectively. We set the relationship between $t_{l,m}$ and t_l as $t_{l,m} = t_l + m\tau$ for $m = 0$ to $M-1$ where τ is a frame shift between successive short-time frames. Furthermore we introduce the following normalization condition

$$\begin{aligned} K &= \kappa K^{(r)}, \\ g[n] &= \kappa \sum_{m=0}^{M-1} g^{(r)}[n - m\tau]. \end{aligned}$$

where κ is an integer constant. With this, the following equation holds between STFS, $s^{(r)}_{l,m,k}$ and LTFS, $s_{l,k'}$ where $k' = \kappa k$:

$$s_{l,k'} = \sum_{m=0}^{M-1} s^{(r)}_{l,m,k} \eta^{-m}, \tag{1}$$

where $\eta = e^{j2\pi k\tau / K^{(r)}}$. We also define an inverse operation, denoted by $\mathrm{LS}_{m,k}\{\cdot\}$, that transforms a set of LTFS bins $s_{l,k'}$ for $k' = 1 \sim K$ at a long-time frame l, denoted by $\{s_{l,k'}\}_l$, to an STFS bin at a short-time frame m and a frequency index k as

$$s^{(r)}_{l,m,k} = \mathrm{LS}_{m,k}\{\{s_{l,k'}\}_l\}.$$

This transformation can be implemented by cascading an inverse long-time Fourier transformation and a short-time Fourier transformation. Obviously, $\mathrm{LS}_{m,k}\{\cdot\}$ is a linear operator.

3. PROBABILISTIC MODELS OF SOURCE AND ROOM ACOUSTICS

Let us define the following terms:

$x^{(r)}_{l,m,k}$	:	STFS of the observed reverberant signal
$s^{(r)}_{l,m,k}$	:	STFS of the unknown source signal
$\hat{s}^{(r)}_{l,m,k}$	:	STFS of the initial source signal estimate
$w_{k'}$	:	LTFS of the unknown inverse filter ($k' = \kappa k$)

We assume that $x^{(r)}_{l,m,k}$, $s^{(r)}_{l,m,k}$, $\hat{s}^{(r)}_{l,m,k}$ and $w_{k'}$ are the realizations of random processes $X^{(r)}_{l,m,k}$, $S^{(r)}_{l,m,k}$, $\hat{S}^{(r)}_{l,m,k}$ and $W_{k'}$, respectively, and that $\hat{s}^{(r)}_{l,m,k}$ is given from the observed signal based on the features of a speech signal such as harmonicity and sparseness.

Now, assume $x^{(r)}_{l,m,k}$ and $\hat{s}^{(r)}_{l,m,k}$ are given for a certain time duration and let $z^{(r)}_k = \{\{x^{(r)}_{l,m,k}\}_k, \{\hat{s}^{(r)}_{l,m,k}\}_k\}$ where $\{\cdot\}_k$ represents the time series of STFS bins at a frequency index k. With this, we assume that speech can be dereverberated by estimating a source signal that maximizes a likelihood function defined at each frequency index k as

$$\begin{aligned} \theta_k &= \arg\max_{\Theta_k} \log p\{z^{(r)}_k | \Theta_k\} \\ &= \arg\max_{\Theta_k} \log \int p\{w_{k'}, z^{(r)}_k | \Theta_k\} dw_{k'}, \end{aligned} \tag{2}$$

where $\Theta_k = \{S^{(r)}_{l,m,k}\}_k$, $\theta_k = \{s^{(r)}_{l,m,k}\}_k$, and $k' = \kappa k$ is a frequency index for LTFS bins. Note that the integral in (2) is a simple double integral on the real and imaginary parts of $w_{k'}$. To analyze this function, we further assume $\{\hat{S}^{(r)}_{l,m,k}\}_k$ and the joint

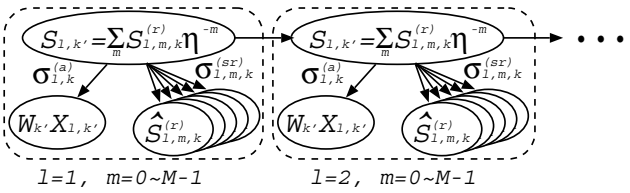

Fig. 1. Graphical model for speech dereverberation

event of $\{X^{(r)}_{l,m,k}\}_k$ and $W_{k'}$ are statistically independent given $\{S^{(r)}_{l,m,k}\}_k$ as shown by the graphical model in Fig. 1. With this, $p\{w_{k'}, z_k|\Theta_k\}$ in (2) can be divided into two functions as

$$p\{w_{k'}, z_k|\Theta_k\} = p\{w_{k'}, \{x^{(r)}_{l,m,k}\}_k|\Theta_k\} p\{\{\hat{s}^{(r)}_{l,m,k}\}_k|\Theta_k\}. \tag{3}$$

The former is a pdf related to room acoustics, that is, the joint pdf of the observed signal and the inverse filter given the source signal. The latter is a pdf related to the information provided by the initial estimation, that is, the pdf of the initial source estimate given the source signal. We can also interpret the second component as being the probabilistic presence of the speech features given the true source signal. We refer to them as acoustics and source pdfs, respectively. Ideally, the inverse transfer function $w_{k'}$ transforms $x_{l,k'}$ into $s_{l,k'}$, that is, $w_{k'} x_{l,k'} = s_{l,k'}$. However, in a real acoustical environment, this equation may contain a certain error $\varepsilon^{(a)}_{l,k'} = w_{k'} x_{l,k'} - s_{l,k'}$ for such reasons as insufficient inverse filter length and fluctuation of room transfer function. Therefore, the acoustics pdf can be considered as a pdf for this error as $p\{w_{k'}, \{x^{(r)}_{l,m,k}\}_k|\Theta_k\} = p\{\{\varepsilon^{(a)}_{l,k'}\}_{k'}|\Theta_k\}$. Similarly, the source pdf can be considered as a pdf for the error $\varepsilon^{(sr)}_{l,m,k} = \hat{s}^{(r)}_{l,m,k} - S^{(r)}_{l,m,k}$ as $p\{\{\hat{s}^{(r)}_{l,m,k}\}_k|\Theta_k\} = p\{\{\varepsilon^{(sr)}_{l,m,k}\}_k|\Theta_k\}$, or the difference between the source signal and the feature-based signal.

In this paper, for the sake of simplicity, we assume these errors to be sequentially independent random processes given $\{S^{(r)}_{l,m,k}\}_k$. We further assume that the real and imaginary parts of the above two error processes are mutually independent with the same variances and can individually be modeled by Gaussian random processes with zero means. With these assumptions, the error pdfs are represented as

$$\begin{aligned} p\{\{\varepsilon^{(a)}_{l,k'}\}_{k'}|\Theta_k\} &= \prod_l b^{(a)}_{l,k} \exp\left\{-\frac{|\varepsilon^{(a)}_{l,k'}|^2}{2\sigma^{(a)}_{l,k'}}\right\}, \\ p\{\{\varepsilon^{(sr)}_{l,m,k}\}_k|\Theta_k\} &= \prod_l \prod_m b^{(sr)}_{l,m,k} \exp\left\{-\frac{|\varepsilon^{(sr)}_{l,m,k}|^2}{2\sigma^{(sr)}_{l,m,k}}\right\}, \end{aligned}$$

where $\sigma^{(a)}_{l,k'}$ and $\sigma^{(sr)}_{l,m,k}$ are, respectively, variances for the two pdfs, hereafter referred to as acoustics and source uncertainties. In this paper, we assume these two values to be given based on the features of the speech signals and room acoustics.

4. SOLUTION BASED ON EM ALGORITHM

One effective way to solve (2) is to use the Expectation-Maximization (EM) algorithm. With this approach, the expectation step (E-step) with an auxiliary function $Q(\Theta_k|\theta_k)$ and the maximization step

(M-step), respectively, are defined for speech dereverberation as

$$Q(\Theta_k|\theta_k) = E_{|\theta}\left[\log p\left\{\log p\{W_{k'}, z_k^{(r)}|\Theta_k\}\right\}\right], \quad (4)$$

$$= \int \left[\log p\{w_{k'}, z_k^{(r)}|\Theta_k\}\right] p\{w_{k'}, z_k^{(r)}|\theta_k\} dw_{k'},$$

$$\tilde{\theta}_k = \arg\max_{\Theta_k} Q(\Theta_k|\theta_k). \quad (5)$$

According to the EM algorithm, the log-likelihood $\log p\{z_k^{(r)}|\theta_k\}$ increases by updating θ_k with $\tilde{\theta}_k$ obtained through an EM iteration, and it converges to a stationary point solution by repeating the iteration.

4.1. Solution

Instead of directly calculating the E- and M-steps, we first analyze $Q(\Theta_k|\theta_k) - Q(\theta_k|\theta_k)$ that has its maximum value at the same Θ_k as $Q(\Theta_k|\theta_k)$. After a certain arrangement of $Q(\Theta_k|\theta_k) - Q(\theta_k|\theta_k)$ and only extracting the terms that involves Θ_k, we obtain the following function.

$$Q_\Theta\{\Theta_k|\theta_k\}$$
$$= \sum_l \left\{ \frac{-|\tilde{w}_{k'} x_{l,k'} - S_{l,k'}|^2}{2\sigma_{l,k'}^{(a)}} + \sum_m \frac{-|\hat{s}_{l,m,k}^{(r)} - S_{l,m,k}^{(r)}|^2}{2\sigma_{l,m,k}^{(sr)}} \right\}. \quad (6)$$

where

$$\tilde{w}_{k'} = \frac{\sum_l s_{l,k'} x_{l,k'}^* / \sigma_{l,k'}^{(a)}}{\sum_l x_{l,k'} x_{l,k'}^* / \sigma_{l,k'}^{(a)}}, \quad (7)$$

and "$*$" means a complex conjugate. It is important to note that the Θ_k that maximizes $Q_\Theta\{\Theta_k|\theta_k\}$ also maximizes $Q(\Theta_k|\theta_k)$, and the Θ_k that makes $Q_\Theta\{\Theta_k|\theta_k\} > Q_\Theta\{\theta_k|\theta_k\}$ also makes $Q(\Theta_k|\theta_k) > Q(\theta_k|\theta_k)$. We can obtain the Θ_k that maximizes $Q_\Theta\{\Theta_k|\theta_k\}$ by differentiating it with $S_{l,m,k}^{(r)}$, setting it at zero, and solving the resultant simultaneous equations. However, the computational cost of obtaining the solution is rather high because we need to solve this equation with M unknown variables for each l and k.

Instead, to maximize (6) in a more efficient way, we introduce the following assumption: The power of an LTFS bin can be approximated by the sum of the power of the STFS bins that compose the LTFS bin based on (1), that is,

$$|s_{l,k'}|^2 \simeq \sum_{m=0}^{M-1} |s_{l,m,k}^{(r)}|^2. \quad (8)$$

With this assumption, (6) can be rewritten as

$$Q_\Theta\{\Theta_k|\theta_k\} = \sum_l \sum_m \frac{-|\mathrm{LS}_{m,k}\{\{\tilde{w}_{k'} x_{l,k'}\}_l\} - S_{l,m,k}^{(r)}|^2}{2\sigma_{l,k'}^{(a)}}$$
$$+ \sum_l \sum_m \frac{-|\hat{s}_{l,m,k}^{(r)} - S_{l,m,k}^{(r)}|^2}{2\sigma_{l,m,k}^{(sr)}}.$$

By differentiating the above equation and setting it at zero, we can obtain a closed form solution for (5) as follows

$$\tilde{s}_{l,m,k}^{(r)} = \frac{\sigma_{l,m,k}^{(sr)} \mathrm{LS}_{m,k}\{\{\tilde{w}_{k'} x_{l,k'}\}_l\} + \sigma_{l,k'}^{(a)} \hat{s}_{l,m,k}^{(r)}}{\sigma_{l,k'}^{(a)} + \sigma_{l,m,k}^{(sr)}}. \quad (9)$$

4.2. Discussion

- With this approach, the dereverberation is achieved by repeatedly calculating (7) and (9) in turn.
- $\tilde{w}_{k'}$ in (7) corresponds to the dereverberation filter obtained by the conventional HERB and SBD approaches given the source signal estimates $s_{l,k'}$ and the observed signals $x_{l,k'}$.
- Equation (9) updates the source estimate by a weighted average of the initial source estimate $\hat{s}_{l,m,k}^{(r)}$ and the source estimate obtained by multiplying $x_{l,k'}^{(r)}$ by $\tilde{w}_{k'}$. The weight is determined according to the source and acoustics uncertainties. In other words, one EM iteration elaborates the source estimate by integrating two types of source estimates obtained based on source and room acoustics properties.
- While we can reduce the computational cost with the approximation (8), an advantageous feature of the EM algorithm, namely, the monotonic increase in the log-likelihood may be degraded. We examine this issue experimentally.

5. EXPERIMENTS

We performed simple experiments with the aim of confirming the performance with the present method. We adopted the same source signals of word utterances and the same impulse responses with RT60 times of 0.1, 0.2, 0.5, and 1.0 s as those in [6]. The observed signals were synthesized by convolving the source signals with the impulse responses. We prepared two types of initial source estimates that were the same as those used for HERB and SBD, that is, $\hat{s}_{l,m,k}^{(r)} = \mathcal{H}\{x_{l,m,k}^{(r)}\}$ and $\hat{s}_{l,m,k}^{(r)} = \mathcal{N}\{x_{l,m,k}^{(r)}\}$, where $\mathcal{H}\{\cdot\}$ and $\mathcal{N}\{\cdot\}$ are, respectively, a harmonic filter used for HERB [6] and a noise reduction filter used for SBD [7]. We determined the source uncertainty $\sigma_{l,m,k}^{(sr)}$ in relation to a voicing measure, $v_{l,m}$, which is used with HERB to decide the voicing status for each short-time frame of the observed signals. According to this measure, a frame is determined as voiced when $v_{l,m} > \delta$ for a fixed threshold δ. Specifically, $\sigma_{l,m,k}^{(sr)}$ was determined in our experiments as

$$\sigma_{l,m,k}^{(sr)} = \begin{cases} G\left\{\frac{v_{l,m}-\delta}{\max_i\{v_{l,m}\}-\delta}\right\} & \text{if } v_{l,m} > \delta \text{ and } k \text{ is a harmonic frequency,} \\ \infty & \text{if } v_{l,m} > \delta \text{ and } k \text{ is not a harmonic frequency,} \\ G\left\{\frac{v_{l,m}-\delta}{\min_i\{v_{l,m}\}-\delta}\right\} & \text{if } v_{l,m} \le \delta, \end{cases}$$

where $G\{u\}$ is a non-linear normalization function that we defined as $G\{u\} = e^{-160(u-0.95)}$. On the other hand, we set $\sigma_{l,k'}^{(a)}$ at a constant value of 1. As a consequence, the weight for $\hat{s}_{l,m,k}^{(r)}$ in (9) becomes a sigmoid function that varies from 0 to 1 as u in $G\{u\}$ moves from 0 to 1. For each experiment, the EM steps were iterated four times. In addition, we also introduced the repetitive dereverberation filter estimation scheme that was used in [6] and [7]. As analysis conditions, we adopted $K^{(r)} = 42$ ms, $K = 10.9$ s, $\tau = 1$ ms, and a 12 kHz sampling frequency.

5.1. Energy decay curves

Figure 2 shows the energy decay curves of the room impulse responses and impulse responses dereverberated by HERB and SBD with and without the EM algorithm using 100 word utterances

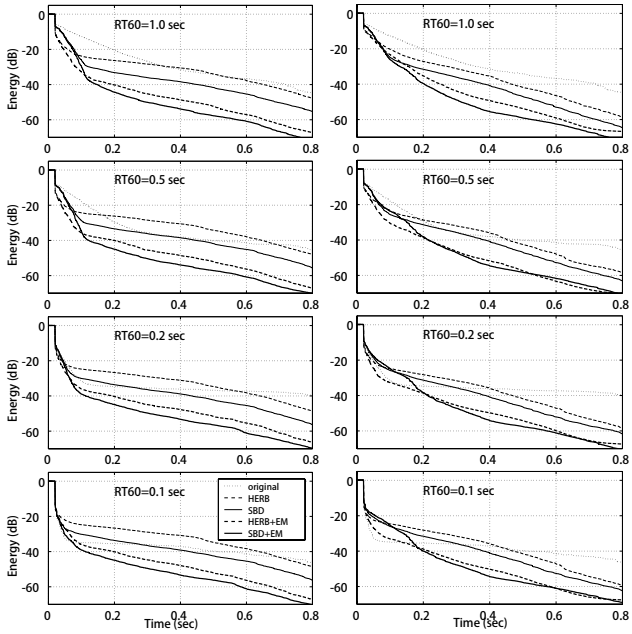

Fig. 2. Reverberation curves of room impulse responses and impulse responses dereverberated by HERB, HERB+EM, SBD and SBD+EM using 100 word observed signals uttered by a woman (left panels) and a man (right panel).

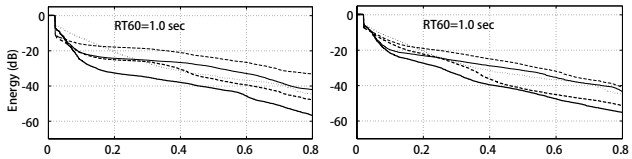

Fig. 3. Reverberation curves of original room impulse responses and impulse responses dereverberated by HERB, HERB+EM, SBD and SBD+EM using 10 word observed signals uttered by a woman (left panels) and a man (right panel).

as the observed signals. The figure clearly demonstrates that the EM algorithm can effectively reduce the reverberation energy with both HERB and SBD. Figure 3 shows the results when only 10 word utterances were used as the observed signals. Here, only the results for RT60=1.0 sec are shown because of the limited space. Although the performance of each method was degraded compared with that achieved based on 100 word utterances, the EM algorithm still worked effectively to reduce reverberation energies in each method.

5.2. Increase of log-likelihood function

Instead of directly examining whether or not the log-likelihood function (2) actually increases, we counted the ratio by which each EM iteration increases $Q_\Theta(\Theta|\theta)$ in (6). Figure 4 shows the resultant ratios at each EM iteration. In all cases, the ratio was more than 0.95, which means that in most cases each EM iteration increases the log-likelihood function. In addition, except for the first step, the more the iteration is repeated, the more steadily the function increases. These results show that approximation (8) works

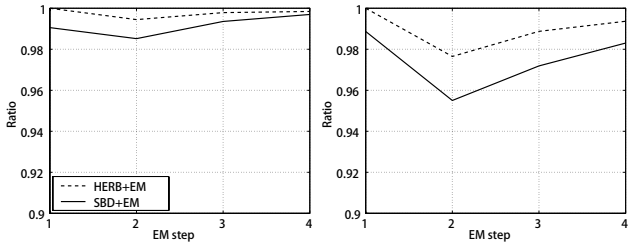

Fig. 4. Ratio of $Q_\Theta(\Theta|\theta)$ becoming larger than $Q_\Theta(\theta|\theta)$ by an EM iteration with 100 words (left panel) and 10 words (right panel) observations.

rather well for our dereverberation purpose.

6. CONCLUSION

This paper proposed a new dereverberation method, in which features of source signals and room acoustics are represented by means of Gaussian pdfs, and the source signals are estimated as signals that maximize the likelihood function defined based on these pdfs. We employed the EM algorithm to solve this optimization problem efficiently. The experimental results showed that the present method can greatly improve the performance of the two dereverberation methods based on speech signal features, HERB and SBD, in terms of the energy decay curves of the dereverberated impulse responses. Since HERB and SBD are effective in improving the ASR performance for speech signals captured in a reverberant environment, we expect the present method also improves the performance with fewer observed signals.

Future work will include the extension and integration of the present method to multi-channel signal processing technologies, its application to adaptive filtering, and the elaboration of probabilistic optimization frameworks such as the optimization of source and room acoustics uncertainties.

REFERENCES

[1] B. Kingsbury and N. Morgan, "Recognizing reverberant speech with rasta-plp," *Proc. 1997 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP-97)*, vol. 2, pp. 1259–1262, 1997.

[2] B. W. Gillespie and L. E. Atlas, "Strategies for improving audible quality and speech recognition accuracy of reverberant speech," *Proc. 2003 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP-2003)*, vol. 1, pp. 676–679, 2003.

[3] H. Buchner, R. Aichner, and W. Kellermann, "Trinicon: a versatile framework for multichannel blind signal processing," *Proc. 2004 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP-2004)*, vol. III, pp. 889–892, May 2004.

[4] T. Hikichi and M. Miyoshi, "Blind algorithm for calculating common poles based on linear prediction," *Proc. of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, vol. IV, pp. 89–92, May 2004.

[5] J. R. Hopgood and P. J. W. Rayner, "Blind single channel deconvolution using nonstationary signal processing," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 5, pp. 476–488, Sep. 2003.

[6] T. Nakatani, and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," *Proc. ICASSP-2003*, vol. 1, pp. 92–95, Apr., 2003.

[7] K. Kinoshita, T. Nakatani and M. Miyoshi, "Efficient blind dereverberation framework for automatic speech recognition," *Proc. Interspeech-2005*, Sep., 2005.