STUDY ON SPEECH DEREVERBERATION WITH AUTOCORRELATION CODEBOOK

Tomohiro Nakatani[†], *Biing-Hwang Juang*[†‡], *Takafumi Hikichi*[†], *Takuya Yoshioka*[†], *Keisuke Kinoshita*[†], *Marc Delcroix*[†], *Masato Miyoshi*[†]

[†]NTT Communication Science Labs., NTT Corporation, Kyoto, Japan
[‡]School of ECE, Georgia Institute of Technology, GA, USA
nak@cslab.kecl.ntt.co.jp

ABSTRACT

This paper proposes a new speech dereverberation approach based on a statistical speech model. An autocorrelation codebook is introduced as a model that can represent time-varying short-time speech characteristics corresponding to the cepstrum and harmonics. The speech dereverberation is formulated as a likelihood maximization problem, in which the quality of a speech signal is recovered by turning the signal into one that is probabilistically more like a clean speech. Two dereverberation algorithms are derived based on different scenarios, regularized inversion and inverse filter estimation. Experimental results show that the proposed approach allows us to reduce both reverberation and noise with the regularized inversion, and to estimate inverse filters that can dereverberate signals effectively from just a small number of observed signals.

Index Terms— Dereverberation, Autocorrelation codebook, Likelihood maximization, Inverse filtering, Regularization

1. INTRODUCTION

Speech signals captured in an enclosure such as a conference room will inevitably contain reverberant components due to reflections from the walls, the floor or the ceiling. These reverberant components are detrimental to the quality of the signal and often cause serious degradation in many applications including automatic speech recognition.

A number of techniques have been proposed to mitigate the reverberation problem. Microphone arrays have been used to focus on sound sources in the "look" direction, while suppressing reflected signals from other directions [1, 2]. Deconvolution by inverting the room impulse response, which can be considered an aggregate of all the reflections with corresponding delays, has also been suggested [3, 4, 5]. However, the real time estimation and tracking of a room impulse response from the source (which may be moving as in a meeting) to the microphone (which may or may not be fixed) remain elusive.

Recently, recognizing the fact that the signal of interest is often a speech signal that manifests certain characteristics (e.g., harmonicity in voiced sounds), many have suggested using these strong source attributes to aid the estimation of a dereverberation filter to suppress the reverberant components in the microphone signal (e.g., [6]). The use of source characteristics led to a new formulation of dereverberation as a problem of probabilistic modeling in which the objective is to design a filter (as part of an overall probabilistic model) which would turn the reverberant speech into a signal that is probabilistically more like a clean speech. The maximum likelihood estimation can be employed to solve the resultant optimization problem.

The probabilistic model formulation of the dereverberation problem has another interesting technical implication. If we look at dereverberation as a speech enhancement problem (as opposed to an inversion problem or a blind deconvolution problem), past experiences indicate that better results can be expected if the solution adapts to the time-varying characteristics of the speech signal. Therefore, it is reasonable to include the instantaneous "state" of the speech signal, which gives proper information about the nature of the speech characteristics at a particular time, in the overall objective function as useful source model constraints. It is thus the purpose of this paper to propose a generalized probabilistic formulation of the dereverberation problem and a solution thereto. A side benefit of this particular formulation is that it allows us to relate speech dereverberation to other conventional speech enhancement approaches such as Wiener filtering.

This paper is organized as follows: Section 2 presents the probabilistic formulation of the speech dereverberation with a statistical speech model based on an autocorrelation (AC) codebook. Two dereverberation scenarios are discussed according to this formulation, and two new dereverberation methods are designed based on regularized inversion and inverse filter estimation. In section 3, the effectiveness of the present methods is examined by preliminary experiments. We show the present regularized inversion is capable of reducing reverberation and noise simultaneously and that the inverse filter estimation enables us to achieve high quality dereverberation with only a few seconds observation. Concluding remarks are provided in section 4.

2. MODEL BASED SPEECH DEREVERBERATION

Suppose a single speech source is captured by two microphones with a certain amount of observation noise. Let s_t, $x_t^{(l)}$, and $d_t^{(l)}$ be digitized sequences of the source, the observed, and the noise signals, respectively, where t and l are the time and microphone indices, respectively. Further let $\bar{s}_t$, $\bar{x}_t$, and $\bar{d}_t$ be the corresponding vector representations with lengths of K, $2L$, and $2L$, respectively, defined by

$\bar{s}_t = [s_t\ s_{t-1}\ \dots\ s_{t-K+1}]^T$,
$\bar{x}_t = [(\bar{x}_t^{(1)})^T\ (\bar{x}_t^{(2)})^T]^T$, where $\bar{x}_t^{(l)} = [x_t^{(l)} x_{t-1}^{(l)} \dots x_{t-L+1}^{(l)}]^T$, and
$\bar{d}_t = [(\bar{d}_t^{(1)})^T\ (\bar{d}_t^{(2)})^T]^T$, where $\bar{d}_t^{(l)} = [d_t^{(l)} d_{t-1}^{(l)} \dots d_{t-L+1}^{(l)}]^T$.

Then, the observation process can be modeled by

$$\bar{x}_t = \mathbf{a}\bar{s}_t + \bar{d}_t, \tag{1}$$

where $\mathbf{a}$ is a stationary convolution matrix ($2L \times K$) defined based on a 2-channel room impulse response (RIR). Let $\bar{a} = [(\bar{a}^{(1)})^T\ (\bar{a}^{(2)})^T]^T$

where $\bar{a}^{(l)} = [a_1^{(l)}\ a_2^{(l)}\ \dots\ a_M^{(l)}]^T$ is the single channel RIR from the source to the l-th microphone. Then $\mathbf{a}$ is represented as

$$\mathbf{a} = \begin{bmatrix} \mathbf{a}^{(1)} \\ \mathbf{a}^{(2)} \end{bmatrix}, \text{ and}$$

$$\mathbf{a}^{(l)} = \begin{bmatrix} a_1^{(l)} & \dots & a_M^{(l)} & 0 & \dots & 0 \\ 0 & a_1^{(l)} & \dots & a_M^{(l)} & 0 & \vdots \\ \vdots & \ddots & \ddots & & \ddots & 0 \\ 0 & \dots & 0 & a_1^{(l)} & \dots & a_M^{(l)} \end{bmatrix}$$

In this paper, we take the above signals, $\bar{x}_t$, $\bar{s}_t$, $\bar{d}_t$, and $\bar{a}$ as realizations of random variables $\bar{X}_t$, $\bar{S}_t$, $\bar{D}_t$, and $\bar{A}$, respectively, and formulate speech dereverberation as the problem of finding a certain parameter set θ that maximizes a log likelihood function defined as

$$\begin{aligned} \mathcal{L}(\theta) &= \sum_t \log p(\bar{X}_t, \bar{S}_t | \bar{A}, \bar{H}_t; \theta), \\ &= \sum_t \log p(\bar{X}_t | \bar{S}_t, \bar{A}; \theta) + \sum_t \log P(\bar{S}_t | \bar{H}_t; \theta), \quad (2) \end{aligned}$$

where $\bar{H}_t$ is a random variable that represents the state of speech at time t that we discuss in detail in the following sections. Because $\bar{s}_t$ is the signal to be estimated, it is taken as a parameter included in θ throughout this paper. The probability density function (pdf) $p(\bar{X}_t | \bar{A}, \bar{S}_t; \theta)$ can be specified from the pdf of $\bar{D}_t$, i.e., $p(\bar{D}_t; \theta)$, because of the relationship (1). Similar to most spectral subtraction scenarios, hereafter we assume that $p(\bar{D}_t; \theta)$ is given in advance (or can be estimated on line).

One innovative point in this paper is to adopt a posteriori pdf $p(\bar{S}_t | \bar{H}_t; \theta)$ determined based on the statistics of the clean speech data. For this purpose, we extract a set of feature vectors, referred to as codewords in an autocorrelation (AC) codebook, each of which contains an autocorrelation function (ACF) of a short-time speech segment, from certain speech database in advance, and use them to represent the pdf. As consequence, we expect, the parameter set θ that maximizes the likelihood function reflects both the condition (1) and the speech model statistics.

2.1. Model of speech statistics with autocorrelation codebook

Although it is often assumed that a speech signal follows a super-Gaussian distribution, such as a Laplacian distribution, it cannot represent time-varying short-time speech characteristics such as the cepstrum and harmonic structure. Therefore, it is difficult to recover such speech characteristics precisely based only on this assumption. In order to construct a speech model that can reasonably represent short-time speech characteristics, we introduce the following assumptions.

1. Each short-time segment of a speech signal at time index t with length N ($N \ll K$), $\bar{\bar{s}}_t = [s_t\ s_{t-1}\ \dots\ s_{t-N+1}]^T$, can be categorized into one of a finite number of states, $H_t = h_t$, where $1 \le h_t \le N_s$, and N_s is the assumed number of distinctive states of speech;
2. In each state, the waveform of the signal is a stationary random process that can be modeled by a Gaussian pdf with an autocorrelation (AC) matrix $\mathbf{r}_h \approx E\{\bar{\bar{s}}_t \bar{\bar{s}}_t^T\}$; i.e.,

$$p(\bar{\bar{S}}_t = \bar{\bar{s}}_t | H_t = h_t) = \mathcal{N}(\bar{\bar{s}}_t; 0, \mathbf{r}_{h_t})^1. \quad (3)$$

[1] Hereafter, we often denote a pdf omitting the names of random variables, for example, denote $p(\bar{S}_t = \bar{s}_t | \bar{H}_t = \bar{h}_t)$ by $p(\bar{s}_t | \bar{h}_t)$.

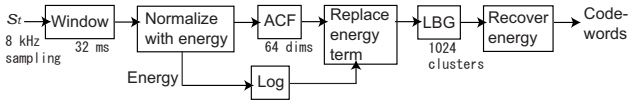

Fig. 1. A method for generating an autocorrelation codebook

According to the above assumptions, the speech statistics are modeled by a set of AC matrices $\mathbf{r}_h$ for $h = 1$ to N_s, or equivalently a set of autocorrelation functions (ACFs), referred to as the codewords in an AC codebook. The time-variation of the speech characteristics can be represented by appropriately switching the AC codewords frame by frame. With dereverberation, because h_t is not given in advance, it is considered to be a parameter that is determined through likelihood maximization.

It is important to note that we have introduced two different time segments $\bar{s}_t$ and $\bar{\bar{s}}_t$ in (1) and (3). The sequence $\bar{s}_t$ in (1) is a "long-time" segment of length K, roughly corresponding to the length of the RIR; the sequence $\bar{\bar{s}}_t$ in (3) is a "short-time" segment of length N from which the short-time speech characteristics of interest is extracted. In terms of their relationship, a long-time segment, $\bar{s}_t$, is equal to a cascade of short-time segments, $\bar{\bar{s}}_t$. Accordingly, we denote the state of a long-time segment, $\bar{h}_t$, as a cascaded sequence of states corresponding to each of the short-time segments in $\bar{s}_t$. Then, by assuming $\bar{\bar{S}}_t$ to be sequentially independent given H_t, the second term in (2) can be rewritten as

$$\sum_t \log p(\bar{S}_t | \bar{H}_t; \theta) = \sum_t \log p(\bar{\bar{S}}_t | H_t; \theta),$$

as long as the corresponding range of summation in t is maintained.

Figure 1 illustrates a method that we adopted to generate the AC codebook for the experiments in this paper. Speech signals were first divided into short-time segments by windowing, then ACFs were calculated for individual segments, and finally the LBG algorithm [7] was used to cluster the ACFs and to generate the codewords. The significance of the signal level with respect to codeword clustering was separated from that of the spectral shape; that is, the distance was measured separately for the energy and the shape of the ACFs. For this purpose, the ACF was calculated after each segment had been normalized by its energy, and the ACF coefficient at time lag zero was replaced with the log of the energy. A Euclidean distance between the modified ACF vectors was employed in clustering. (An alternative to normalization and the choice of distance measure is the likelihood distortion based on residual-normalized ACFs.) After clustering, the energy of each codeword was restored based on its coefficient at time lag zero. The AC matrices in (3) can be generated by forming Toeplitz matrices based on the individual codewords. In the following, we discuss two speech dereverberation scenarios that involve AC matching.

2.2. Regularized inversion – dereverberation in noisy environment with given room impulse response

If we can assume that the source signal is a white Gaussian process with zero mean and unity variance, and that the RIR can be measured in advance, the maximum likelihood solution to (2) becomes

$$\bar{s}_t = (\mathbf{a}^T \mathbf{a} + \sigma_d^2 I)^{-1} \mathbf{a}^T \bar{x}_t.$$

Here, we also assumed that the noise pdf, $p(D_t; \theta)$, is given by $\mathcal{N}(0, \sigma_d^2 I)$. The above equation corresponds to Tikhonov regularization [8], and can greatly reduce the noise that is undesirably am-

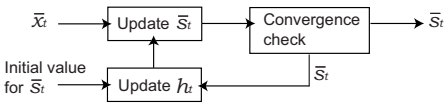

Fig. 2. Processing flow of regularized inversion

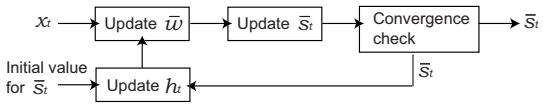

Fig. 3. Processing flow of inverse filter estimation

plified as the result of inverse filtering when a Moore-Penrose inverse of $\mathbf{a}$ is simply applied to $\bar{x}_t$ in (1).

With the AC codebook, the performance of the above technique can be further improved because of the additional source statistics. Here, we take $\bar{s}_t$ and $\bar{h}_t$ in (2) as unknown parameters to be optimized and $\bar{x}_t$, $\bar{a}$, and $p(D_t;\theta)$ as given in advance, similar to the conventional regularized inversion. After certain mathematical manipulations, we obtain the following equations that maximize the likelihood function in terms of h_t and $\bar{s}_t$, respectively.

$$\hat{h}_t = \arg\max_{h_t} p(\bar{\bar{s}}_t|h_t;\theta) \rightarrow h_t, \quad (4)$$

$$\hat{\bar{s}}_t = (\mathbf{a}^T\mathbf{a} + \sigma_d^2\bar{\mathbf{r}}_t^{-1})^{-1}\mathbf{a}^T\bar{x}_t \rightarrow \bar{s}_t. \quad (5)$$

Here, $\bar{\mathbf{r}}_t$ is a block diagonal matrix that contains, as its diagonal components, AC matrices $\mathbf{r}_{h_t}$ corresponding to the state sequence h_t that represents the additional source information in the long-time segment $\bar{s}_t$. According to (4) and (5), starting from certain initial values, the likelihood function can be maximized up to a stationary point by iteratively updating the sequences of the source states and the source estimates in turn. Figure 2 summarizes the processing flow.

Note that (5) can be viewed not only as inverse filtering but also as a form of Wiener filtering in the AC domain, offering the possibility of reducing both reverberation and additive noise simultaneously. This can be confirmed by the fact that the equation becomes equal to the Moore-Penrose inverse when $\sigma_d = 0$ (i.e., the noise-free case), while it can be rewritten as $\bar{s}_t = (\bar{\mathbf{r}}_t + \sigma_d^2 I)^{-1}\bar{\mathbf{r}}_t\bar{x}_t$ when we assume $\mathbf{a}$ to be an identity matrix (i.e., the reverberation-free case).

2.3. Inverse filter estimation in noise-free environment

As our second example, we discuss an inverse filter estimation method based on the AC codebook. For the sake of simplicity, we deal with a condition with no observation noise, and rewrite (1) using a 2-channel inverse filter $\bar{w} = [(\bar{w}^{(1)})^T\ (\bar{w}^{(2)})^T]^T$ of the RIR $\bar{a}$ as

$$\bar{\bar{s}}_t = \mathbf{x}_t\bar{w}, \quad (6)$$

where $\mathbf{x}_t$ is a matrix representation of $\bar{x}_t$, which is defined as $\mathbf{x}_t = [\bar{x}'_t\ \bar{x}'_{t-1}\ \ldots\ \bar{x}'_{t-N+1}]^T$ and $\bar{x}'_t = [x_t^{(1)}\ x_{t-1}^{(1)}\ \ldots\ x_{t-M+1}^{(1)}\ x_t^{(2)}\ x_{t-1}^{(2)}\ \ldots\ x_{t-M+1}^{(2)}]^T$. By definition, the inverse filter $\bar{w}$ should satisfy $\bar{a}^{(1)}(z)\bar{w}^{(1)}(z) + \bar{a}^{(2)}(z)\bar{w}^{(2)}(z) = 1$. The existence of such an inverse filter is guaranteed under the condition that the impulse responses corresponding to the channels, $\bar{a}^{(1)}(z)$ and $\bar{a}^{(2)}(z)$, do not share common zeros [9].

With regard to the likelihood function, the first term in (2) can be discarded because of the noise-free assumption, and thus it can be written as

$$\mathcal{L}(\theta) = \sum_t \log p(\bar{\bar{S}}_t = \mathbf{x}_t\bar{w}|H_t;\theta).$$

We adopt $\bar{w}$ and h_t as parameters to be optimized for inverse filter estimation and take $\bar{\bar{s}}_t$ as the objective of the joint optimization with $\bar{w}$ according to (6). A certain constraint on $\bar{w}$ is necessary to avoid a self-evident solution $\bar{w} = 0$. For this purpose, we introduce $w_1^{(1)} = 1$ assuming the 1st microphone is physically the closest to the source location and the gain of the observed signal is appropriately normalized. Then, the following two equations are derived as those that maximize the likelihood function in terms of h_t and $\bar{w}$, respectively.

$$\hat{h}_t = \arg\max_h p(\bar{\bar{s}}_t|h;\theta) \rightarrow h_t,$$

$$\hat{\bar{w}} = \begin{bmatrix} 1 \\ -\mathbf{r}_{2..2M,2..2M}^{-1}\mathbf{r}_{2..2M,1..1} \end{bmatrix} \rightarrow \bar{w},$$

where $\mathbf{r}$ is a square matrix ($2M \times 2M$) defined as

$$\mathbf{r} = \sum_t \mathbf{x}_t^T\mathbf{r}_{h_t}^{-1}\mathbf{x}_t,$$

and $\mathbf{r}_{n_1..n_2,n_3..n_4}$ denotes a submatrix of $\mathbf{r}$ ranging from the n_1-th to n_2-th rows and from the n_3-th to n_4-th columns. The likelihood function can be maximized again up to a stationary point by iteratively updating the sequence of the source states and the inverse filter in turn from certain initial values. Figure 3 summarizes the processing flow.

It should be noted that the above estimation method can be viewed as a variation of multi-channel linear prediction (MCLP) [10]. Equation (6) becomes identical to that of the MCLP when we introduce an additional constraint $w_1^{(2)} = 0$ and set $p(\bar{\bar{S}}_t|H;\theta) = \mathcal{N}(0,\sigma_s^2 I)$. In other words, the AC codebook allows us to dereverberate speech signals using the MCLP without whitening them.

3. EXPERIMENT

We conducted two preliminary experiments to confirm the effectiveness of the dereverberation methods based on the AC codebook. For this purpose, an AC codebook was generated as in fig. 1 using 5240 word utterances produced by a female speaker (FKM) and found in the ATR database. The observed signals were synthesized by convolving the female utterances with a 2-channel RIR (RT60$=0.5$ sec) measured in a reverberant room. We set the sampling rate at 8 kHz, $K = 9216$, $L = 5217$, $M = 4000$, $N = 64$, and $N_s = 1024$.

3.1. Exp-1: performance of regularized inversion

To examine the effect of regularized inversion described in section 2.2, white Gaussian noise was also added to the observed reverberant signals with an average signal-to-noise ratio (SNR) of 10 dB. We compared the dereverberation performances provided by the Moore-Penrose inverse, the conventional regularized inverse, and the proposed method, hereafter referred to as MPI, CRI, and PROP1, respectively. As the initial estimate of $\bar{s}_t$ for PROP1, we adopted the signal dereverberated by CRI. The iteration number of PROP1 was fixed at five. Figure 4 shows spectrograms of speech signals before and after dereverberation. While MPI seriously amplified the noise to obscure the speech completely, CRI adequately dereverberated the signal without amplifying the noise. By contrast, PROP1 reduced not only the reverberation but also the noise. Interestingly,

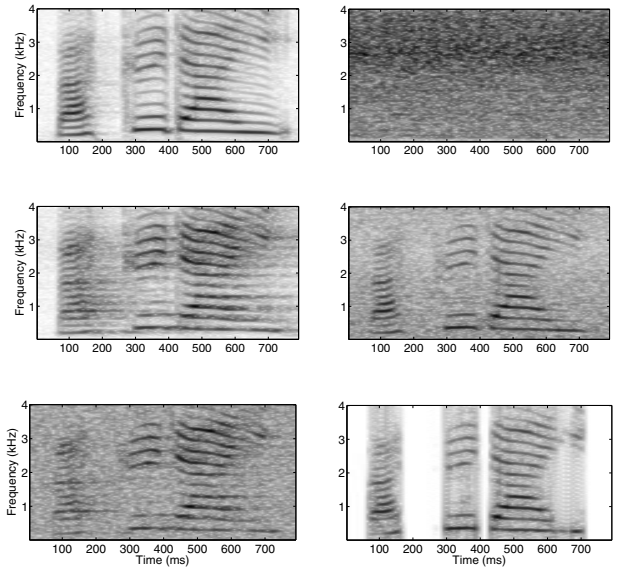

Fig. 4. Spectrograms of the source signal uttering the Japanese word "Ba-Ku-Da-i" (top left), the source with reverberation (center left), the source with reverberation and noise (bottom left), and speech signals processed by MPI, CRI, and PROP1 (top right, center right, and bottom right)

PROP1 eliminated all the signal energy in each time region where the SNR was significantly low. The left panel in fig. 5 depicts the time patterns of the cepstral distances (CDs) between the source signal and the dereverberated signals. PROP1 was the best at reducing the CDs except in the time regions where it eliminated all the signal energy. These results suggest that PROP1 can effectively recover the signal quality depending on the time-varying speech characteristics and SNR.

3.2. Exp-2: performance of inverse filter estimation

We tested the inverse filter estimation method described in section 2.3, hereafter referred to as PROP2, in terms of its dereverberation quality. We set the length of the dereverberation filter at 3000 taps in each channel, which is shorter than that of the RIR, to confirm the robustness of PROP2 with the channel order mismatch. We prepared two sets of observed signals that were composed of one-word and five-word sequences, respectively. In the estimation, we also used these observed signals as the initial estimates of $\bar{s}_t$. The CD between the source and dereverberated signals after five estimation iterations and the spectrograms of the dereverberated signals are shown in the right panel of fig. 5 and the left and right panels of fig. 6, respectively. They clearly show that PROP2 could recover the signal quality very well. In particular, the CD obtained with the five-word observation is almost always below 2 dB. With the one-word observation, the audible sound quality was also well recovered. Even if large CDs are observed at around 200 and 800 ms, it did not affect the audible quality because the signal energy in these regions was sufficiently small.

4. CONCLUSION

This paper proposed an autocorrelation (AC) codebook as a model of speech statistics for speech dereverberation. The AC codebook

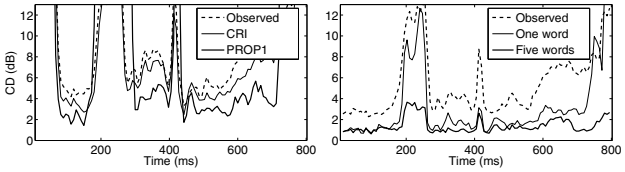

Fig. 5. Time patterns of cepstral distances (CDs) from the source signal to the observed and dereverberated signals, including signals processed by CRI and PROP1 in Exp-1 (left) and those processed by PROP2 (one-word and five-word observations) in Exp-2 (right)

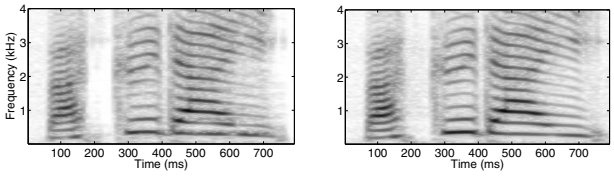

Fig. 6. Spectrograms of the dereverberated signals in Exp-2 obtained based on one-word (left) and five-word (right) observations.

can represent short-time and time-varying speech characteristics precisely, and thus allows us to formulate speech dereverberation in a reasonable manner. We formulated two dereverberation approaches, regularized inversion and inverse filter estimation, using the AC codebook with the maximum likelihood estimation framework. Preliminary experiments showed that the AC codebook enabled the regularized inversion to reduce both reverberation and noise effectively, and the inverse filter estimation to achieve precise dereverberation with only a few seconds observation. Future work will include the integration of inverse filter estimation and regularized inversion, and a comprehensive evaluation of the performance of the AC codebook based speech dereverberation.

5. REFERENCES

[1] J.L. Flanagan, J. Johnston, R. Zahn, and G. Elko, "Computer-steered microphone arrays for sound transduction in large rooms, *J. Acoust. Soc. Am.*, vol. 78, pp. 1508–1518, 1985.

[2] G.W. Elko, "Superdirective Microphone Arrays," in *Acoustic Signal Processing for Telecommunication*, S.L. Gay and J. Benesty, eds., pp. 181-235, Kluwer Academic Press, 2000.

[3] B.W. Gillespie and L.E. Atlas, "Strategies for improving audible quality and speech recognition accuracy of reverberant speech," *Proc. ICASSP-2003*, vol. 1, pp. 676–679, 2003.

[4] P.A. Naylor, "Speech dereverberation," *Proc. IWAENC-05*, 2005.

[5] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Spectral subtraction steered by multi-step forward linear prediction for single channel speech dereverberation," Proc. ICASSP-2006, vol.I, pp.817–820, May, 2006.

[6] T. Nakatani, B.H. Juang, K. Kinoshita, and M. Miyoshi, "Speech dereverberation based on probabilistic models of source and room acoustics," *Proc. ICASSP-06*, vol. I, pp. 821–824, May, 2006.

[7] Y. Linde, A. Buzo, and R.M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. COM-28, No. 1, pp. 84-95, Jan., 1980.

[8] C.R. Vogel, *Computational Methods for Inverse Problems*, Soc for Industrial & Applied Math, June, 2002.

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

[10] M. Delcroix, T. Hikichi, and M. Miyoshi, "Dereverberation of speech signals based on linear prediction," *Proc. ICSLP-2004*, 2004.