TABLE III
WORD RECOGNITION PERFORMANCE(%) FOR 1022 ISOLATED KOREAN WORDS WITH THE PROPOSED MCDHMM FOR DIFFERENT NUMBERS OF MIXTURES BETWEEN STATE DURATION-INDEPENDENT OBSERVATION PROBABILITY AND STATE DURATION-DEPENDENT OBSERVATION PROBABILITY (NO. OF TEST WORDS WAS 2810)

A	B	Parameter sets	Proposed MCDHMM
6	1	$\hat{C}$	59.08(73.03)
		$\hat{C}+\Delta\hat{C}$	72.85(84.74)
6	2	$\hat{C}$	61.79(74.35)
		$\hat{C}+\Delta\hat{C}$	76.29(88.12)
6	3	$\hat{C}$	62.64(75.42)
		$\hat{C}+\Delta\hat{C}$	77.27(88.30)
9	1	$\hat{C}$	60.55(74.07)
		$\hat{C}+\Delta\hat{C}$	75.48(86.77)
9	2	$\hat{C}$	62.28(75.92)
		$\hat{C}+\Delta\hat{C}$	77.76(88.44)
9	3	$\hat{C}$	63.57(76.70)
		$\hat{C}+\Delta\hat{C}$	78.72(89.21)
12	1	$\hat{C}$	60.48(74.36)
		$\hat{C}+\Delta\hat{C}$	74.84(86.69)
12	2	$\hat{C}$	63.19(76.10)
		$\hat{C}+\Delta\hat{C}$	79.19(89.26)
12	3	$\hat{C}$	64.86(77.56)
		$\hat{C}+\Delta\hat{C}$	80.34(89.47)
15	1	$\hat{C}$	59.34(72.69)
		$\hat{C}+\Delta\hat{C}$	75.02(86.27)
15	2	$\hat{C}$	62.54(75.60)
		$\hat{C}+\Delta\hat{C}$	79.55(89.19)
15	3	$\hat{C}$	64.43(77.53)
		$\hat{C}+\Delta\hat{C}$	79.58(89.65)

A : THE NUMBER OF MIXTURES IN STATE DURATION-INDEPENDENT OBSERVATION PROBABILITY
B : THE NUMBER OF MIXTURES IN STATE DURATION-DEPENDENT OBSERVATION PROBABILITY

We have found that when compared with the conventional HMM, the proposed MHMM yields improvements in recognition performance. It has indirectly been shown that the transient and timing information in the spectra not only characterize the time-varying vocal tract but also play an important role in human perception. It has been shown that MHMM's can estimate the state duration-dependent parameters by Viterbi decoding in a straightforward manner.

In the MCDHMM, detailed acoustic modeling will increase the number of parameters that will be estimated from observations and require a large amount of training utterances. Thus, we have proposed a solution to the conflict between detailed acoustic modeling and insufficient training data. The solution is the use of a different number of mixtures between state duration-independent and state duration-dependent observation probability. In case of insufficient training data, the duration-dependent parameters can be estimated by postprocessing.

REFERENCES

[1] S. Furui and M. Sondhi, *Advances in Speech Signal Processing*. New York: Dekker, 1992, p. 509.

[2] Y. Guedon and C. Cocozza-Thivent, "Explict state occupancy modeling by hidden semi-Markov models: Application of Derin's scheme," *Comput. Speech Language*, vol. 4, pp. 167–192, 1990.

[3] L. Deng, P. Kenny, M. Lennig, and P. Mermelstein, "Modeling acoustic transitions in speech by state interpolation hidden Markov models," *IEEE Trans. Signal Processing*, vol. 40. no. 2, Feb. 1992.

[4] L. Deng, "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal," *Signal Processing*, vol. 27, no. 1, pp. 65–78, Apr. 1992.

[5] O. Ghitza and M. Sondhi, "Hidden Markov models with templates as nonstationary states: An application to speech recognition," *Comput. Speech Language*, vol. 7, no. 2, pp. 101–119, 1993.

[6] J. G. Wilpon and L. R. Rabiner, "A modified k-means clustering algorithm for use in isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 587–594, June 1985.

[7] Y. J. Chung and C. K. Un, "Use of different numbers of mixtures in continuous density hidden Markov models," *IEE Electron. Lett.*, vol. 29, no. 9, pp. 824–825, Apr. 29, 1993.

Cepstrum-Based Deconvolution for Speech Dereverberation

Suresh Subramaniam, Athina P. Petropulu, and Christopher Wendt

***Abstract*— We present a blind deconvolution-based approach for the restoration of speech degraded by the acoustic environment. The proposed scheme processes the outputs of two microphones using cepstra operations and the theory of signal reconstruction from phase only. Under mild assumptions, it reconstructs the room impulse response associated with each microphone and restores the speech signal.**

I. INTRODUCTION

We consider the problem of restoring speech that has been degraded through addition of multiple echoes. This problem appears very often in hands-free telephony when the microphone is placed far from the speaker. The sound wave at the microphone location consists of the direct path wave and multiple delayed and attenuated versions of it due to reflections on the room walls and other surfaces. Depending on the microphone location, the energy of the echoed speech might be large enough to degrade the intelligibility of the speech the far-end listener hears. Reverberation heavily affects the performance of any automatic speech recognition system. Real applications demand that the performance of a speech recognition system not be affected by changes in the environment. However, it is well known that when a recognition system is tested under conditions different than those used to train it, its recognition rate drops dramatically.

There are many sources of distortion that can degrade the accuracy of a speech recognition system. Generally, the sources of degradation are clustered into two categories: 1) additive noise sources that could be due to machinery, interference from other speakers, etc. and 2) convolutional noise sources due to the acoustical properties of the environment or due to the impulse response of the microphones used. Several types of array processing strategies have been applied to speech recognition systems. One approach is the delay-and-sum beamformer [2], where steerings delays are applied at the outputs of the microphones to compensate for arrival time differences at the two microphones, reinforcing the desired signal over other background signals. Another approach is based on adaptive minimization of the mean square error [11]. These algorithms provide nulls in the direction of undesired noise sources and reinforce sensitivity in the direction of the desired signal. The key assumption in these algorithms is that the desired speech signal is statistically independent of all sources of degradation, which means that the distortion cannot be due to delayed versions of the desired signal, as is the case in a reverberant room. All these techniques require multiple microphones and calibration of the microphone array.

Manuscript received November 14, 1994; revised April 1, 1996. The associate editor coordinating the review of this paper and approving it for publication was Dr. James H. Snyder.

The authors are with the Electrical and Computer Engineering Department, Drexel University, Philadelphia, PA 19104 USA.

Publisher Item Identifier S 1063-6676(96)06711-9.

This correspondence focuses on restoring clean speech degraded by convolutional noise. We apply a variant of the blind deconvolution scheme proposed in [8] on speech signals collected by two desktop microphones in order to identify the reverberation associated with each measurement and reconstruct the clean speech signal. The microphones can be placed anywhere in the room as long as they are apart from each other. Although the magnitude spectrum of speech suffices for recognition algorithms, to preserve speech quality, both magnitude and phase are important. In this correspondence, we will address the more general problem of reconstructing both magnitude and phase. The performance of the proposed algorithm is demonstrated on real speech recordings.

II. Background

Acoustic signals radiated inside a room are distorted by wall reflections. The distortion can be well modeled by a linear filtering operation. At the ith microphone, the sampled wave $x_i(n)$ can be expressed as

$$x_i(n) = s(n) * h_i(n), \qquad i = 1, 2 \tag{1}$$

where

- $s(n)$ clean speech
- $h_i(n)$ reverberation (room impulse response as seen by the ith microphone)
- n sampled time
- "*" convolution.

Under realistic conditions, the room impulse response is nonminimum phase [5]. Our objective is to reconstruct the speech signal $s(n)$ given the signals $x_1(n)$ and $x_2(n)$. This is a typical blind deconvolution problem and, as such, is ill posed. For a unique solution to exist, the signals $s(n)$ and $h_i(n), i = 1, 2$ must satisfy certain conditions. Under the following conditions

- A1) $h_1(n)$ and $h_2(n)$ are finite extent sequences (FIR) whose Z transforms have no common zeros
- A2) there are no zero-pole cancellations between $s(n)$ and $h_i(n), i = 1, 2$
- A3) the channels $h_i(n), i = 1, 2$ have no zeros on the unit circle

it was shown in [8] that (1) can be solved for $s(n), h_1(n)$, and $h_2(n)$ (within a time delay and a scalar factor). In this correspondence, we employ a variant of the scheme originally proposed in [8]. The basic blind deconvolution scheme is based on the following propositions. Let

$$H_i(z) = cz^{-p_i^-} H_i^{\min}(z^{-1}) H_i^{\max}(z), \qquad i = 1, 2 \tag{2}$$

where $H_i^{\min}(z^{-1})$ and $H_i^{\max}(z)$ are the minimum and maximum phase parts of $H_i(z)$, p_i^- is the number of zeros of $H_i(z)$ outside the unit circle, and c is a constant.

Let also $X_i(k), H_i(k)$ be the discrete Fourier transform of $x_i(n), h_i(n)$, respectively.

Proposition 1 [8]: Let

$$\phi_{\min}(k) = \operatorname{Im}\left[\sum_{\lambda=0}^{N-1} (\hat{X}_1(\lambda) - \hat{X}_2(\lambda)) W(k-\lambda)\right] \tag{3}$$

where

$$\hat{X}_i(\lambda) = \log X_i(\lambda), \quad \text{and}$$
$$W(\lambda) = \sum_{m=0}^{N/2-1} e^{-j(2\pi/N)m\lambda}, \qquad \lambda = 0, \cdots, N-1.$$

Then

$$\phi_{\min}(k) = \arg\{H_{\min}(k)\} - \frac{2\pi}{N} p_2^+ k \tag{4}$$

where

$$H_{\min}(k) = H_1^{\min}(k)(H_2^{\min}(k))^*. \tag{5}$$

Proposition 2 [8]: Let

$$\phi_{\max}(k) = \operatorname{Im}\left[\sum_{\lambda=0}^{N-1} (\hat{X}_2(\lambda) - \hat{X}_1(\lambda)) W^*(k-\lambda)\right]. \tag{6}$$

Then

$$\phi_{\max}(k) = \arg\{H_{\max}(k)\} - \frac{2\pi}{N} p_2^- k, \tag{7}$$

where

$$H_{\max}(k) = (H_1^{\max}(k))^* H_2^{\max}(k). \tag{8}$$

The region of support of $h_{\min}(n)$ and $h_{\max}(n)$ is $[-p_2^+, p_1^+]$ and $[-p_2^-, p_1^-]$, respectively, where p_i^+ denotes the number of zeros of $H_i(z)$ that are inside the unit circle. Due to assumption A1), $h_{\min}(n)$ is FIR and has no zeros in conjugate reciprocal pairs; hence, it can be reconstructed within a scalar constant from its Fourier phase only [3]. Similarly, $h_{\max}(n)$ can be recovered from its Fourier phase. The phases of $h_{\min}(n)$ and $h_{\max}(n)$ can be computed from the data only, within the linear phase terms $(2\pi/N)p_2^+ k$ and $(2\pi/N)p_2^- k$, respectively (see (3) and (6)).

Once $h_{\min}(n)$ and $h_{\max}(n)$ have been reconstructed, the complex cepstra of $h_1(n)$ and $h_2(n)$ can be recovered as

$$\hat{h}_1(n) = \hat{h}_{\min}(n), n > 0, \quad \hat{h}_1(-n) = \hat{h}_{\max}(n), n > 0$$
$$\hat{h}_2(n) = \hat{h}_{\min}(-n), n > 0, \quad \hat{h}_2(-n) = \hat{h}_{\max}(n), n > 0 \tag{9}$$

where $\hat{h}_{\min}(n), \hat{h}_{\max}(n)$ denote the cepstra of $h_{\min}(n)$ and $h_{\max}(n)$, respectively.

Thus, the problem of estimating the channels breaks down to estimating $h_{\min}(n)$ and $h_{\max}(n)$ from $\phi_{\min}(k)$ and $\phi_{\max}(k)$, respectively. There exist several iterative schemes for the reconstruction of an FIR sequence from its Fourier phase only [3]. The general iteration consists of successive application of the phase substitution and time-limiting operator, which assumes that the sequence's length is available and that its phase is exactly known.

In the case of $h_{\min}(n)$, where the length is unknown and the phase $\arg\{H_{\min}(k)\}, k = 0, \cdots, N-1$ is known within the unknown linear phase $(2\pi/N)p_2^+ k$, the iteration can be modified as follows:

$$h_{\min}^{(i+1)}(n) = T_{lm}[h_{\min}^{(i)}(n)] = T_2 T_1[h_{\min}^{(i)}(n)], \qquad i = 0, 1, \cdots,$$
$$H_{\min}^{(0)}(k) = 1, \qquad k = 0, \cdots, N-1 \tag{10}$$

where $h_{\min}^{(i)}(n)$ is the sequence reconstructed at the ith iteration, and $H_{\min}^{(i)}(k)$ is the corresponding discrete Fourier transform. The operators $T_2[\cdot]$ and $T_1[\cdot]$, which are successively applied on $h_{\min}^{(i)}(n)$, are defined as

$$T_1[h_{\min}^{(i)}(n)] = y(n),$$

where

$$Y(k) = |H_{\min}^{(i)}(k)| e^{j\phi_{\min}(k) + j(2\pi/N)mk},$$
$$k = 0, \cdots, N-1 \text{ (phase substitution)} \tag{11}$$

and

$$T_2[y(n)] = \begin{cases} y(n) & n \in [0, l] \\ 0 & \text{otherwise} \end{cases} \quad \text{(time limiting)} \tag{12}$$

where l is an estimate of $L_{\min}$, and m is an estimate of p_2^+. As i increases, the iteration either converges, or the mean-squared error between the sequences reconstructed during successive operations is nonincreasing [3]. However, even if a low mean-squared error is reached, convergence to the correct solution is not guaranteed. If a

TABLE I
RECONSTRUCTION OF $h_{\min}(n)\phi_{\min}(k)$.

```
for l = 2, ..., length(h1) + length(h2)
    for m = 0, ..., l − 1
        if the iteration h_min^(i+1)(n) = T_lm[h_min^(i)(n)] converges then
            p_2^+ = m
            L_min = l
            STOP
        endif
    end
end
```

FIR sequence corresponding to the particular values of l and m exists, the iteration will converge to it; otherwise, it will not converge. The latter means that although a low mean-square error may be reached, the corresponding sequence will be incorrect since its phase will be different than $\phi_{\min}(k) + (2\pi/N)mk$.

It was shown [9], [8] that $h_{\min}(n)$ can be reconstructed from $\phi_{\min}(k)$ according to the loop of Table I. In other words, while increasing the length l and applying the iteration (10) for $m = 0, \cdots, l-1$, the first convergent solution we encounter is the correct one, and the corresponding l and m equal $L_{\min}(n)$ and p_2^+, respectively.

It can also be shown [9] that a convergent solution may exist for $l > L_{\min}$; however, the resulting sequence contains zero-phase convolutional components, which contradicts assumption A1). On the other hand, no FIR sequence with phase $\phi_{\min}(k)$ exists for $l < L_{\min}$ [9].

The sequence $h_{\max}(n)$ can be reconstructed from $\phi_{\max}(k)$ in a similar manner.

III. IMPLEMENTATION

To reconstruct the two channels $h_1(n)$ and $h_2(n)$, we need to compute first the log spectra of the observations, i.e., $C_1(k)$ and $C_2(k)$, and at this point, a phase unwrapping algorithm is needed. Then, the phases $\phi_{\min}(k)$ and $\phi_{\max}(k)$ are computed based on (3) and (6) and used as input for the reconstruction of $h_{\min}(n)$ and $h_{\max}(n)$ respectively, as shown in Table I. To determine convergence of the iteration a reliable error measure is needed.

Let

$$\begin{aligned} d_{\min}(n) &= c_1(n) - c_2(n), \qquad n > 0 \\ d_{\max}(n) &= c_1(-n) - c_2(-n), \qquad n > 0 \end{aligned} \tag{13}$$

where $c_i(n), i = 1, 2$ are the cepstra of $x_i(n), i = 1, 2$. The imaginary part of Fourier transform of the odd part of $d_{\min}(n), (d_{\max}(n))$ equals the Fourier phases of $h_{\min}(n)(h_{\max}(n))$. [7]. In addition, let

$$\begin{aligned} d^i_{\min}(n; l, m) &= c^i_{\min}(n; l, m) \\ &\quad - c^i_{\min}(-n; l, m), \qquad n > 0 \end{aligned} \tag{14}$$

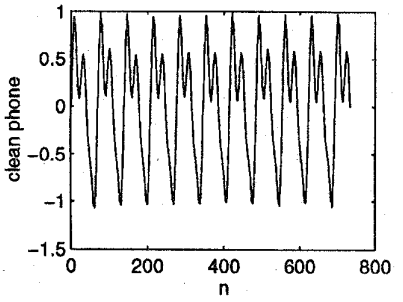
$$\begin{aligned} d^i_{\max}(n; l, m) &= c^i_{\max}(n; l, m) \\ &\quad - c^i_{\max}(-n; l, m), \qquad n > 0 \end{aligned} \tag{15}$$

where $c^i_{\min}(n; l, m), c^i_{\max}(n; l, m)$ are the cepstra of the sequences reconstructed at the ith step of the iteration of (10). The differences in (14) and (15) correspond to the Fourier phases of $h^{(i)}_{\min}(n)$ and $h^{(i)}_{\max}(n)$, respectively.

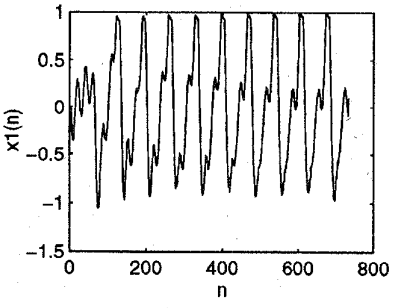
Based on the above definitions, the errors

$$e^i_{\min}(l, m) = \sum_{n=1}^{q} [d_{\min}(n) - d^i_{\min}(n; l, m)]^2 \tag{16}$$

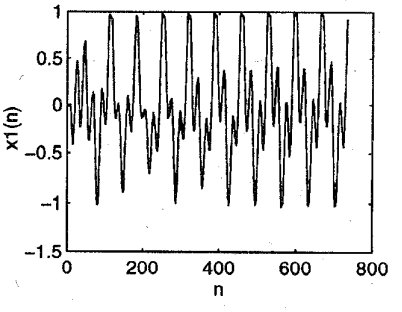
$$e^i_{\max}(l, m) = \sum_{n=1}^{q} [d_{\max}(n) - d^i_{\max}(n; l, m)]^2 \tag{17}$$



(a)

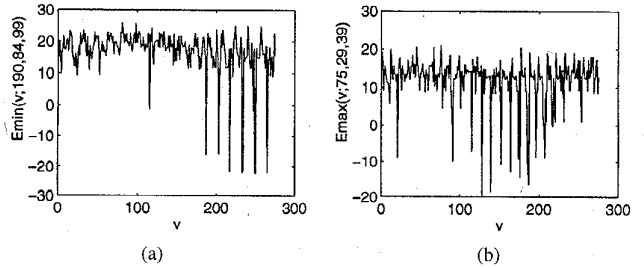


(b)



(c)

Fig. 1. (a) Clean Phoneme /i/. (b) Reverberated signal $x_1(n)$. (c) Reverberated signal $x_2(n)$.



(a) (b)

Fig. 2. Errors. (a) $\mathcal{E}_{\min}(v; 190, 84, 99)$ in decibels. (b) $\mathcal{E}_{\max}(v; 75, 29, 39)$ in decibels.

quantify the closeness of the phases of the sequences reconstructed at each iteration and the given phases. If

$$e^i_{\min}(l, m) \le \epsilon_1, \quad e^i_{\max}(l, m) \le \epsilon_2 \tag{18}$$

where ϵ_1 and ϵ_2 are some thresholds, then we can safely assume that the iteration has converged. Since the errors in (16) and (17) are not uniformly decreasing with i, we can resort to them after the mean-

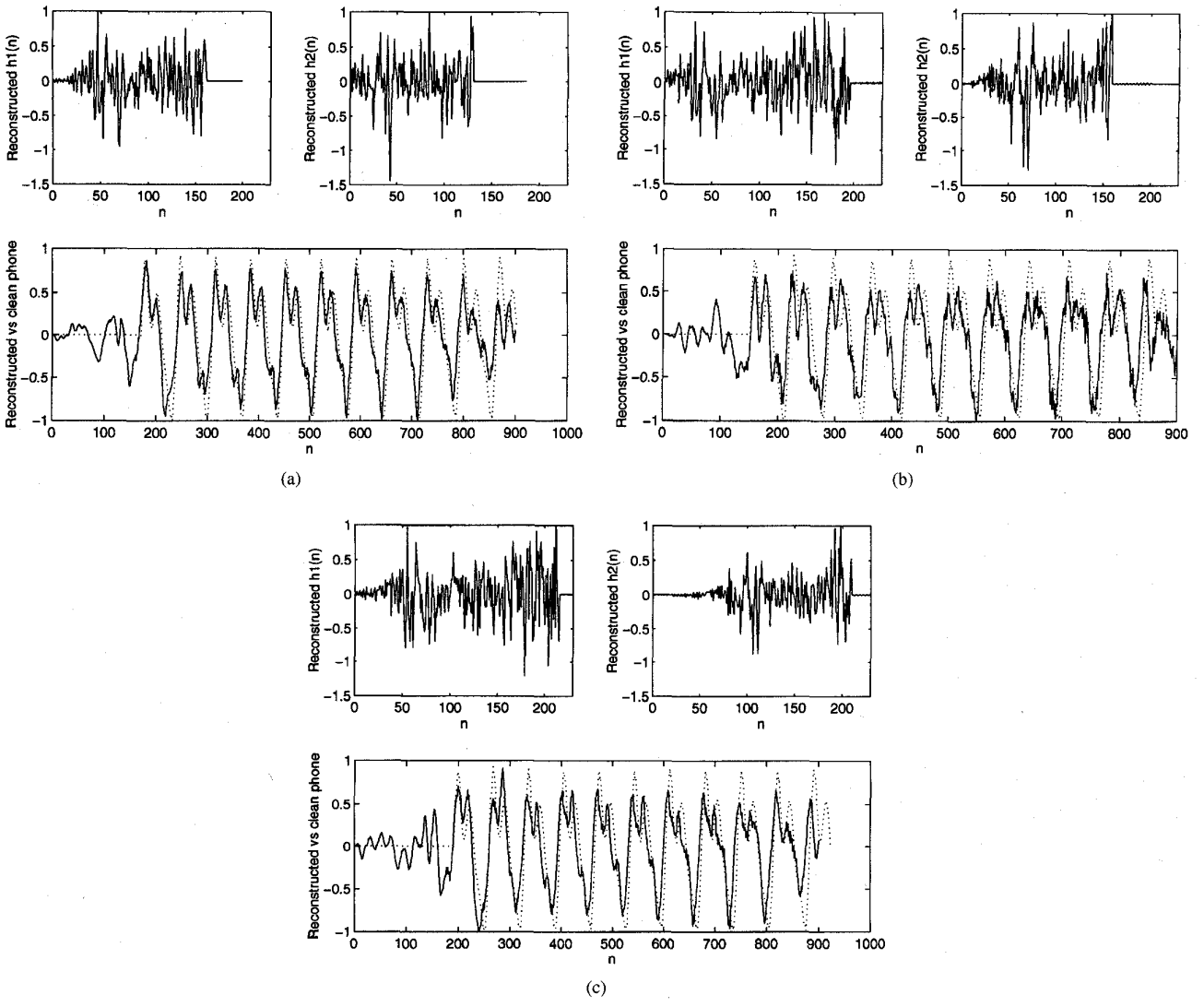


Fig. 3. Reconstructed channels and the corresponding recovered phone (solid line) against the clean one (dotted line) for (a) $L_{\min} = 203, p_2^+ = 93, L_{\max} = 86, p_2^- = 36$. (b) $L_{\min} = 253, p_2^+ = 93, L_{\max} = 100, p_2^- = 64$. (c) $L_{\min} = 285, p_2^+ = 129, L_{\max} = 148, p_2^- = 104$.

squared error between sequences reconstructed during successive iterations has reached some low value. According to Table I, the iteration must stop the first time the thresholds have been crossed.

IV. Experimental Results

In this section, we demonstrate the performance of the proposed algorithm using experimental data. The phoneme /i/ was recorded using a pair of high quality microphones in a relatively quite environment in a room (25, 15, 7) (ft). We used an experimental system that was able to record a spoken phone at two different positions in the room. This setup included a computer running software that was able to record and play back digital audio simultaneously, a high-quality speaker and amplifier, and a pair of high-quality microphones and preamplifiers. This equipment was chosen to ensure that results were as accurate as possible. For our experiment, a phoneme was chosen and was digitally recorded in a reflection-free environment. This was considered to be the "clean" phone. These samples were then played through the speaker into a room and simultaneously digitally recorded through two microphones in different positions in the room. The experiment was conducted in a medium size room with a very low reverberation time due to carpet and acoustic ceiling tiles. The source and the microphone locations were (1, 4, 4), (5, 7, 4), and (8, 5, 4) (ft). A low sampling rate, 4 kHz, which was then downsampled and filtered to 2 kHz, was used with an 8-b resolution. The clean phone /i/ and the reverberated signals $x_1(n)$ and $x_2(n)$ are shown in Figs. 1(a)–(c), respectively. All horizontal axes are in discrete time samples.

In order to avoid any poles or zeros close to the unit circle, both sequences $x_1(n)$ and $x_2(n)$ were windowed using an exponential window α^n, where $\alpha = 0.985$ for the first experiment. Since moving some zeros away from the unit circle causes some other zeros to come closer the unit circle, the following procedure was employed for the selection of an appropriate value for the window parameter α:

i) Given the two observations $x_1(n)$ and $x_2(n)$, compute $x_3(n) = x_1(n) * x_2(n)$.

ii) Apply an exponential window α^n on the signals, and compute the corresponding complex cepstra.

iii) Let $c_1^w(n), c_2^w(n)$, and $c_3^w(n)$ be the complex cepstra of the windowed sequences $\alpha^n x_1(n), \alpha^n x_2(n)$ and $\alpha^n x_3(n)$, respectively. If α is properly selected, then it holds that $c_3^w(n) = c_1^w(n) + c_2^w(n)$.

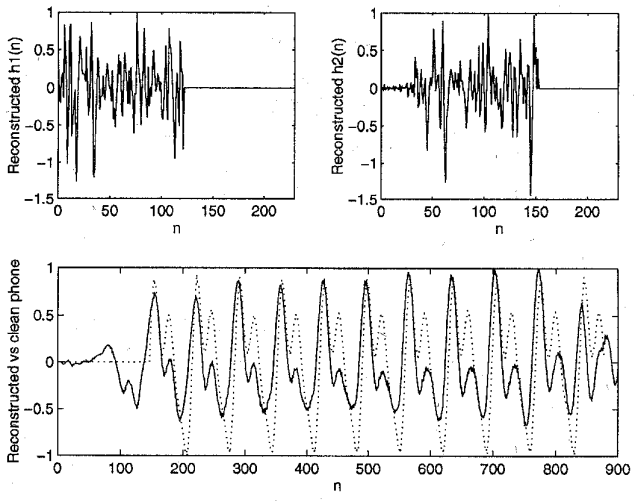


Fig. 4. Reconstructed channels and deconvolved phone (solid line) versus clean one (dotted line) in the cases of underestimating the channel lengths, i.e., $L_{\min} = 201, p_2^+ = 95$ and $L_{\max} = 72, p_2^+ = 56$.

The blind deconvolution algorithm was applied to the windowed data. Using the procedure outlined in Table I, the errors $e_{\min}^{100}(l,m)$ and $e_{\max}^{100}(l,m)$ were computed for $l \geq L$ and $M_L \leq m \leq M_H$.

For demonstration purposes, we use the 1-D vectors $\mathcal{E}_{\min}(v; L, M_L, M_H)$ and $\mathcal{E}_{\max}(v; L, M_L, M_H)$ formed by appending the rows of the matrices $e_{\min}^{100}(l,m)$ and $e_{\max}^{100}(l,m)$, respectively. The indices l, m, v are related as

$$l = \lfloor \frac{v}{M} \rfloor M + L; \quad m = v - \lfloor \frac{v}{M} \rfloor M + M_L;$$
$$M = M_H - M_L + 1 \quad v = (l - L)M + m - M_L. \qquad (19)$$

The errors $\mathcal{E}_{\min}(v; 190, 84, 99)$ and $\mathcal{E}_{\max}(v; 75, 29, 39)$ are illustrated in Fig. 2(a) and (b) for $v = 1, \cdots, 300$. The first significant error drop in $\mathcal{E}_{\min}(v, 190, 84, 99)$ occurs at $v = 217$, which, via (19), corresponds to $l = 203$ $(= L_{\min})$ and $m = 93$ $(= p_2^+)$. Although relatively low errors also appear at $v = 187$, corresponding to $l = 201, m = 95$, after plotting $d_{\min}(n)$ against $d_{\min}^{100}(n; 201, 95)$, we can see that the matching is not very good at the beginning of the sequences. Since, in these cepstra-based functions, the most important information is contained in the first few samples, a small mismatch at the beginning, although it can yield small mean-squared error, does not quantify correctly the similarity between the two sequences. This may indicate that a better error measure than the mean-squared error (see (16) and (17)) should be considered. This is subject of further investigation.

The first significant error drop in $\mathcal{E}_{\max}(v; 75, 29, 39)$ (Fig. 2(b)) occurs at $v = 128$, which, via (19), corresponds to $l = 86$ $(= L_{\max})$ and $m = 36$ $(= p_2^-)$. Based on $L_{\min} = 203, p_2^+ = 93$, and $L_{\max} = 96, p_2^- = 36$, the reconstructed channels are shown in Fig. 3(a).

Deconvolution for the recovery of the phone was performed by using the optimum delay least squares filter technique [4] for the inversion of a mixed-phase channel. Fig. 3(a) illustrates the recovered phone (solid line) against the clean one (dotted line).

To determine the effect of overestimating the lengths $L_{\min}, L_{\max}$, we reconstructed channels corresponding to $L_{\min} = 253, p_2^+ = 93$ and $L_{\max} = 100, p_2^- = 64$, which correspond to low errors $\mathcal{E}_{\min}(v = 1017; 190, 84, 99) = -30$ db and $\mathcal{E}_{\max}(v = 310; 75, 29, 39) = -25$ db. The reconstructed channels and the deconvolved phone is shown in Fig. 3(b).

One more case was considered with $L_{\min} = 285, p_2^+ = 129, (\mathcal{E}_{\min}(27726; 151, 69, 129) = -41$ dB and $(L_{\max} = 148, p_2^- = 104, \mathcal{E}_{\max}(8314; 41, 29, 105) = -36$ dB). The reconstructed channels and the recovered phone for each case are shown in Fig. 3(c).

From Fig. 3, we see that the reconstructed phoneme /i/ is good for lengths $L_{\min} = 203, L_{\max} = 86$ and $p_2^+ = 93, p_2^- = 36$ (see Fig. 3(a)). By considering even lower errors, the reconstructed signals $h_{\min}(n)$ and $h_{\max}(n)$ are longer (see Figs. 3(b) and (c)) and according to theory, they introduce common zeros to the reconstructed channels. However, the reconstruction is still reasonably good, which indicates that overestimating $L_{\min}$ and $L_{\max}$ is not a very sensitive issue.

It is interesting to note that the channels $h_1(n)$ and $h_2(n)$ reconstructed in all the above cases exhibit very similar time domain characteristics everywhere except in their tail parts.

To see the effect of underestimating the channel lengths, the iteration was stopped at $L_{\min} = 201, p_2^+ = 95$ and $L_{\max} = 72, p_2^+ = 56$. The reconstructed channels $\hat{h}_1(n)$ and $\hat{h}_2(n)$ and the deconvolved phone are shown in Fig. 4. According to theory [8], if indeed 203, 86 are the true values for $L_{\min}$ and $L_{\max}$, then no convergent solution is expected when $L_{\min} < 203$ and $L_{\max} < 86$. Indeed, the reconstruction was very poor, as indicated by Fig. 4. This has the intuitive explanation that when the channel length is taken to be less than the true length, we are actually missing significant late-arriving reverberant energy.

V. Conclusion

In this correspondence, we have tested a variant of the blind deconvolution algorithm proposed in [8] for the restoration of clean speech signal from two reverberated signals. The experimental results supported the theory, and reconstruction of the time domain clean phone was achieved. The additive noise case is currently under investigation, and the performance of the algorithm is going to be evaluated for continuous speech.

References

[1] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, 1979.
[2] J. L. Flanagan *et al.*, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Amer.*, vol. 78, pp. 1508–1518, Nov. 1985.
[3] M. H. Hayes, J. S. Lim, and A. V. Oppenheim, "Signal reconstruction from the phase or magnitude," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 672–680, Dec. 1980.
[4] J. Mourjopoulos, P. M. Clarkson, and K. J. Hammond, "A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals," in *Proc. Intl. Conf. Acoust., Speech Signal Processing*, 1982, pp. 1858–1861.
[5] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Amer.*, pp. 165–169, July 1979.
[6] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
[7] A. P. Petropulu and C. L. Nikias, "Signal reconstruction from the phase of the bispectrum," *IEEE Trans. Signal Processing*, vol. 40, no. 3, pp. 601–610, Mar. 1992.
[8] ——, "Blind deconvolution using reconstruction from partial higher order cepstra information," *IEEE Trans. Signal Processing*, vol. 41, no. 6, pp. 2088–2095, June 1993.
[9] ——, "Blind deconvolution of colored signals based on higher order cepstra and data fusion," *Proc. Inst. Elec. Eng.*, pt. F. vol. 140, no. 6, Dec. 1993.
[10] A. P. Petropulu and S. Subramaniam, "Cepstrum based deconvolution for speech dereverberation," in *Proc. Int. Conf. Acousti., Speech Signal Processing, ICASSP '94*, Adelaide, Australia.
[11] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1985.