ELSEVIER

Speech Communication 18 (1996) 317–334

SPEECH COMMUNICATION

A microphone array processing technique for speech enhancement in a reverberant space [1]

Qing-Guang Liu [a], Benoît Champagne [a,*], Peter Kabal [a,b]

[a] *INRS-Télécommunications, Université du Québec, 16 Place du Commerce, Verdun, Québec, Canada H3E 1H6*

[b] *Department of Electrical Engineering, McGill University, 3480 University Street, Montreal, Quebec, Canada H3A 2A7*

Received 4 April 1995; revised 11 January 1996

Abstract

In this paper, a new microphone array processing technique is proposed for blind dereverberation of speech signals affected by room acoustics. It is based on the separate processing of the minimum-phase and all-pass components of delay-steered multi-microphone signals. The minimum-phase components are processed in the cepstrum-domain, where spatial averaging followed by low-time filtering is applied. The all-pass components, which contain the source location information, are processed in the frequency-domain by performing spatial averaging and by retaining only the all-pass component of the resulting output. The underlying motivation for the new processor is to use spatio-temporal processing over a single set of synchronous speech segments from several microphones to reconstruct the source speech, such that it is applicable to practical time-variant acoustic environments. Simulated room impulse responses are used to evaluate the new processor and to compare it to a conventional beamformer. Significant improvements in array gain and important reductions of reverberation in listening tests are observed.

Zusammenfassung

Dieser Beitrag beschreibt ein neuartiges Verfahren zur blinden Entzerrung von durch Raumakustik verzerrten Sprachsignalen mit Hilfe einer Anordnung mehrerer Mikrophone. Hierfür werden die Minimalphasen- und die Allpaßkomponenten der durch Verzögerung gesteuerten Multimikrophonsignale getrennt verarbeitet. Die Minimalphasenkomponenten werden im Cepstral-Bereich örtlich gemittelt und kurzzeitgefiltert. Von den Allpaßkomponenten, welche die Ortsinformation der Quelle enthalten, wird lediglich ein im Frequenzbereich örtlich gemitteltes Signal weiterverwendet. Als wesentliches Merkmal des Verfahrens wird das gesuchte Sprachsignal aus synchronen Sprachsegmenten verschiedener Mikrophone rekonstruiert, wodurch praktische Anwendungen in zeitvarianter akustischer Umgebung möglich werden. Zur Beurteilung des Verfahrens und zum Vergleich mit konventionellen Strahlformern werden simulierte Raumstoßantworten verwendet. Sowohl beim Signal zu Störabstand als auch bei Bewertung des Echos in Hörtests werden deutliche Verbesserungen erzielt.

* Corresponding author. E-mail: champagne@inrs-telecom.uquebec.ca.

[1] Audiofiles available. See http://www.elsevier.nl/locate/specome .

PII S0167-6393(96)00011-8

Résumé

Dans cet article, nous présentons une nouvelle technique de traitement pour réseaux de microphones ayant pour but la déréverbération aveugle des signaux de parole altérés par l'acoustique de salle. La méthode repose sur le traitement séparé des composantes en phase-minimale et passe-tout des signaux de sortie des microphones. Les composantes en phase-minimale sont traitées dans le domaine cepstral, où l'on effectue un moyennage spatial suivi d'un filtrage passe-bas. Les composantes passe-tout, qui contiennent l'information de position de la source, sont traitées dans le domaine fréquentiel en effectuant une formation de voie suivie de l'extraction d'une composante en phase-minimale. Puisqu'elle repose sur le traitement spatio-temporel d'un ensemble de trames synchronisées provenant de plusieurs microphones, cette technique peut être utilisée dans des environnements acoustiques variants tels que l'on rencontre en pratique. Des réponses impulsionnelles de salles synthétisées au moyen d'un ordinateur sont utilisées afin d'évaluer la nouvelle technique et de la comparer à une formation de voie conventionnelle sous des conditions contrôlées. Les résultats indiquent une augmentation significative du gain d'antenne et un effet de déréverbération marqué.

Keywords: Microphone array; Room dereverberation; Speech enhancement

1. Introduction

In many applications of speech communications such as hands-free telephony and audio-conferencing in small rooms, dereverberation techniques are required for enhancing the intelligibility of speech degraded through the addition of multiple echoes. For single microphone acquisition, a direct solution to this problem is provided by conventional inverse filtering techniques. If the room impulse response between the speaker and the microphone is known from calculations or measurements, the reverberation can be removed by the use of an inverse filter or by minimum mean-square error deconvolution. However, since the impulse responses of typical rooms are non-minimum-phase and have therefore unstable inverses (Neely and Allen, 1979; Mourjopoulos, 1985; Miyoshi and Kaneda, 1988), inverse filtering-based methods have a limited scope in practice (Walsh, 1985). The situation is further complicated by the difficulty of measuring and tracking the room impulse response in real-time applications.

An alternative approach for the enhancement of reverberant speech with a single microphone is provided by cepstrum filtering techniques (Oppenheim and Schafer, 1975). The underlying motivation is the fact that deconvolution in the time domain corresponds to subtraction in the cepstrum (i.e. quefrency) domain. Since the complex cepstrum of a speech signal is usually concentrated around the cepstral origin, while that of the echoes is composed of pulses extending far away from the origin, it follows that low-time filtering or peak-picking in the quefrency domain can be used to remove the echo's cepstrum.

While cepstrum filtering has been applied successfully to the enhancement of speech degraded by simple echoes, its use for the enhancement of single microphone speech affected by room reverberation poses several practical problems. These are due mainly to the effect of segmentation errors on the evaluation of complex cepstra (Bees et al., 1991) and to certain numerical errors associated with the use of exponential weighting. The use of various windowing and segmentation schemes to reduce these types of errors was investigated by Bees et al. (1991). They also proposed to use temporal averaging of the echo cepstrum over successive frames to achieve significant enhancement of the reverberant speech. However, this approach implicitly assumes that the room impulse response is invariant over a long period of time, which is not appropriate for real-time processing of speech under time-varying conditions. For example, in practical environments, the impulse response usually changes from frame to frame due to the variation of the speaker's position and even the position changes of the physical objects in the room (e.g., the opening doors, people moving about, etc.).

Microphone array techniques have long been proposed for the removal of room reverberation. Compared to single microphone techniques which are limited to temporal processing, array processing offers the additional advantage of spatial processing. In an early paper, Allen et al. (1977) proposed a two-

microphone technique to remove room reverberation from speech signals. This is accomplished by compensating for the phase and amplitude differences between the two microphone channels and by summing them coherently. This approach, which is a form of delay-and-sum beam-forming, takes advantage of the uncorrelated nature of reverberant speech tails at different locations. Two-dimensional microphone array systems based on delay-and-sum beamforming that can be used for dereverberation of speech are described in (Flanagan et al., 1985, 1991). Because of the wide-band nature of speech signals, several studies (Pirz, 1979; Goodwin and Elko, 1993; Sydow, 1994) have focussed on the design of wide-band microphones arrays with constant beamwidth. Adaptive beamforming algorithms (Van Compernolle et al., 1990; Dowling et al., 1992) have also been considered for the suppression of directional interference, but they fail to reduce speech reverberation because of the correlation that exists between the direct-path speech signal and its echoes.

In this paper, we present a new technique to remove room reverberation from speech signals which is based on the joint use of a microphone array combined with cepstrum-based processing. In the proposed technique, the signal received at each microphone is factored into a minimum-phase and an all-pass component. These components are processed as follows:

1. The minimum-phase component is related to the real cepstrum which needs neither phase unwrapping nor exponential windowing. This component was found experimentally to be affected less by reverberation than the all-pass component. To recover the minimum-phase component of the original speech, spatial averaging followed by low-time filtering in the quefrency domain is applied to the minimum-phase components of the individual microphone signals.
2. The phase information of the microphone signals is preserved in their all-pass components. The all-pass component of the original speech is recovered from the all-pass component of a conventional beamformer applied to the all-pass components of the microphone signals.

The final dereverberated speech is obtained from the synthesis of the recovered minimum-phase and all-pass components. Simulation results and listening tests of the new processor indicate an improvement in dereverberation performance as compared to conventional beamforming techniques.

The paper is organized as follows. Basic microphone array concepts are introduced in Section 2. The effect of reverberation on the minimum-phase and the all-pass components of a room impulse response is investigated in Section 3. Separate processing of the minimum-phase and the all-pass components of the microphone array signals is described in Section 4 and the software implementation of the new processor is described in Section 5. Experimental results are provided in Section 6. Finally, a discussion is presented in Section 7.

2. Beamforming and microphone arrays

Consider an array of M omni-directional microphones in a reverberant acoustical enclosure. A conventional (delay-and-sum) beamformer structure for this array is illustrated in Fig. 1. The sampled output of the ith microphone, denoted $x_i(n)$ $(i = 1, \ldots, M)$, is first shifted by a time-delay τ_i and then scaled by a corresponding weight w_i. The resulting delayed and scaled signals from all microphones are then summed to produce the beamformer output $y(n)$. Assuming that the background noise is negligible, the ith microphone output can be expressed as

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

where $s(n)$ represents the anechoic speech signal, $h_i(n)$ denotes the impulse response between the speech source and the ith microphone in the room and $*$ denotes convolution. Thus, according to Fig. 1, the output of the conventional beamformer is

$$y(n) = s(n) * b_0(n), \tag{2}$$

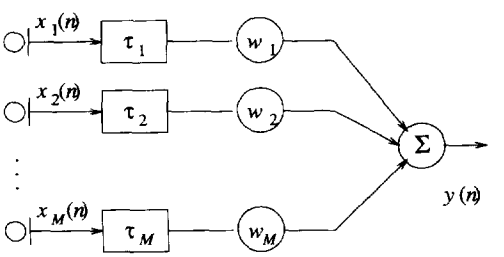

Fig. 1. Delay and sum beamformer.

where

$$b_0(n) = \sum_{i=1}^{M} w_i h_i(n - \tau_i). \quad (3)$$

In Eq. (3), the purpose of the delays τ_i is to time-align the direct-path components of the impulse responses $h_i(n)$ so as to steer the beamformer in the direction of the desired speech source. This way the direct-path signals are reinforced while echoes apart from the steering direction are attenuated. One basic requirement associated to the beamformer in Fig. 1 is the determination of the steering delays τ_i. This can be achieved through the use of time-delay estimation (Carter, 1993) or direction finding techniques (Dowling et al., 1992; Silverman and Kirtman, 1992; Tanaka and Kaneda, 1993;). In the sequel, the time delays τ_i are assumed to be known. The weights w_i in Eq. (3) are used to shape the spatial directivity pattern of the beamformer. We note that the beampattern (and in particular the beamwidth) is dependent on the signal frequency and the array configuration (Flanagan, 1985). For a uniformly spaced linear array with a fixed aperture size, the beamwidth is inversely proportional to the signal frequency. At low frequency, the spatial resolution is poor and the echo power at the output of the beamformer is larger. A more general wideband beamforming structure would involve replacing the weights w_i in Fig. 1 by time-domain filters whose purpose is to control the frequency properties of the desired response.

The design of wideband microphone arrays with constant beamwidth has been addressed by several authors (Pirz, 1979; Goodwin and Elko, 1993; Sydow, 1994). The design parameters can be specified in both spatial and temporal domains. Typical parameters include the positions of the microphones, the channel weights w_i or even the coefficients of time-domain filters in a more general wideband structure. This design usually leads to complex nonlinear optimization problems. A simple and intuitively pleasing approach for the design of microphone array configurations with uniform directivity patterns over several octaves is based on the concept of harmonic nesting (Pirz, 1979; Flanagan et al., 1991; Grenier, 1993; Kellermann, 1991). The basic principle is the following: if d is the optimal element spacing in a uniformly spaced linear array at some frequency f, then $2d$ will be the optimal spacing at frequency $f/2$, and so on. An harmonically nested array with identical beampattern over three octaves is shown in Fig. 2. By nesting the subarrays, the total number of microphones needed can be reduced.

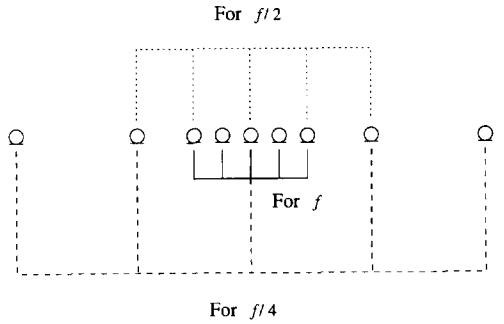

Fig. 2. Harmonically nested linear array.

Array gain is frequently used to measure the improvement in signal-to-noise ratio at the output of a beamformer. For the purpose of dereverberation, the desired signal is the direct-path signal from the speech source, while the unwanted signal consists of all the echoes and possibly some additive background noise. In this paper, we therefore define the array gain as follows:

$$\mathrm{AG} = \frac{\mathrm{SENR}_0}{\frac{1}{M}\left(\sum_{i=1}^{M} \mathrm{SENR}_i\right)}, \quad (4)$$

where SENR_0 is the signal to echo-plus-noise power ratio (SENR) at the output of the array processor and SENR_i $(i = 1, \ldots, M)$ is the SENR at the output of the ith microphone. Hence, the denominator in Eq. (4) represents the average input SENR over all microphones.

Under the assumption of negligible background noise, the microphone outputs are given by Eq. (1). If we further assume that the array processor under study acts as a linear time-invariant systems on its inputs $s(n)$, then the array output can be written as

$$y(n) = s(n) * h_0(n), \quad (5)$$

where $h_0(n)$ is the impulse response between the source and the array output. A simple yet very useful expression for SENR_i $(i = 0, 1, \ldots, M)$ can be obtained as follows. First, let us write $h_i(n)$ in the form

$$h_i(n) = h_{i,\mathrm{d}}(n) + h_{i,\mathrm{e}}(n), \quad (6)$$

where $h_{i,d}(n)$ and $h_{i,e}(n)$ represent the components of $h_i(n)$ corresponding to the direct-path and the echoes, respectively. Notice that due to the sampling function and the transfer functions of the sensors, the direct-path response $h_{i,d}(n)$ is not assumed to be a pure pulse. It can be shown easily that for either an impulsive or a white noise excitation $s(n)$, we have

$$\mathrm{SENR}_i = \frac{\sum_{n=0}^{\infty} |h_{i,d}(n)|^2}{\sum_{n=0}^{\infty} |h_{i,e}(n)|^2}, \quad i = 0, \ldots, M, \tag{7}$$

where the index $i = 0$ refers to the array processor's response. The above measure will be used to evaluate the performance of the proposed array processor in Section 6.

3. Minimum-phase and all-pass components of room impulse responses

Let $H(\omega)$ denote the Fourier transform of the room impulse response $h(n)$ between a source and a receiver in a reverberant room. In this section, to simplify the notations, we will omit the microphone index i. The most commonly used representation of $H(\omega)$ is in terms of its magnitude $|H(\omega)|$ and phase $\phi(\omega)$, that is,

$$H(\omega) = |H(\omega)| \exp[\mathrm{j}\phi(\omega)]. \tag{8}$$

Another useful representation of $H(\omega)$ is given by the factorization (Neely and Allen, 1979; Oppenheim and Schafer, 1975)

$$H(\omega) = H_{\mathrm{Min}}(\omega) \cdot H_{\mathrm{All}}(\omega), \tag{9}$$

or equivalently, in the time domain,

$$h(n) = h_{\mathrm{Min}}(n) * h_{\mathrm{All}}(n), \tag{10}$$

where $H_{\mathrm{Min}}(\omega)$ and $H_{\mathrm{All}}(\omega)$ are the minimum-phase and all-pass components of $H(\omega)$ and $h_{\mathrm{Min}}(n)$ and $h_{\mathrm{All}}(n)$ are the corresponding inverse Fourier transforms. A signal is said to be minimum-phase if its z-transform contains no poles or zeros outside the unit circle in the z-domain. Minimum-phase signals are of particular interest here because they have stable and causal inverses. Unfortunately, typical room impulse responses are generally nonminimum-phase (Neely and Allen, 1979); their z-transforms have zeros outside the unit circle. These zeros are represented by the all-pass component of $H(\omega)$.

The minimum-phase component $H_{\mathrm{Min}}(\omega)$ can be expressed as (Neely and Allen, 1979; Oppenheim and Schafer, 1975)

$$H_{\mathrm{Min}}(\omega) = |H(\omega)| \exp[\mathrm{j}\phi_{\mathrm{Min}}(\omega)], \tag{11}$$

where $\phi_{\mathrm{Min}}(\omega)$ is the Hilbert transform of $\log|H(\omega)|$. Thus, $H_{\mathrm{Min}}(\omega)$ depends only on the magnitude of $H(\omega)$ and not on its phase. The phase information is entirely contained in the all-pass component $H_{\mathrm{All}}(\omega)$ which can be obtained by dividing $H(\omega)$ in Eq. (8) by $H_{\mathrm{Min}}(\omega)$ in Eq. (11):

$$H_{\mathrm{All}}(\omega) = \exp\{\mathrm{j}[\phi(\omega) - \phi_{\mathrm{Min}}(\omega)]\}. \tag{12}$$

The all-pass component $H_{\mathrm{All}}(\omega)$ contains only a phase term and has a unit magnitude.

The decomposition of a signal into a minimum-phase and an all-pass component can also be carried out in the cepstrum domain. Let F and F^{-1} denote the Fourier transform and its inverse, respectively. By definition, the complex cepstrum of a signal $h(n)$ is given by

$$\hat{h}(n) = C\{h(n)\} = F^{-1}\{\log[H(\omega)]\}, \tag{13}$$

where the symbol $\hat{}$ is used to indicate a cepstral representation, C denotes the complex cepstrum operator and log is the complex logarithm (Oppenheim and Schafer, 1975). Let $\hat{h}_{\mathrm{Min}}(n) = C\{h_{\mathrm{Min}}(n)\}$ denote the complex cepstrum of the minimum-phase component of $h(n)$. It can be shown that

$$\hat{h}_{\mathrm{Min}}(n) = \begin{cases} \hat{h}_r(n), & n = 0, \\ 2\hat{h}_r(n), & n > 0, \\ 0, & n < 0, \end{cases} \tag{14}$$

where

$$\hat{h}_r(n) = F^{-1}\{\log|H(\omega)|\} \tag{15}$$

is also known as the real cepstrum of $h(n)$. Eqs. (14) and (15) provide an attractive way of performing the decomposition $h(n) = h_{\mathrm{Min}}(n) * h_{\mathrm{All}}(n)$. Indeed, once $\hat{h}_{\mathrm{Min}}(n)$ (Eq. (14)) is available, $h_{\mathrm{Min}}(n)$ can be recovered as $C^{-1}\{\hat{h}_{\mathrm{Min}}(n)\}$, where C^{-1} denotes the inverse of the complex cepstrum operator, and $h_{\mathrm{All}}(n)$ can then be obtained easily from Eq. (9) or Eq. (10). A computational realization of this procedure will be

used in Section 5 to compute the minimum-phase and all-pass components of the microphone signals.

When there is no reverberation, it can easily be verified that $H_{\mathrm{Min}}(\omega) = 1$ and $H_{\mathrm{All}}(\omega) = \exp(-j\omega n_0)$, where n_0 represents the propagation delay between the source and the receiver. Consider a simple propagation scenario in which the direct-path signal is received in the presence of a single echo, delayed by one sample. In this case, we can write

$$h(n) = \delta(n - n_0) + a\,\delta(n - n_0 - 1), \tag{16}$$

where a is a positive scaling factor representing the amplitude of the echo (without loss of generality, it is assumed that n_0 is a positive integer). The z-transform of $h(n)$ (Eq. (16)), which is given by

$$H(z) = z^{-n_0}(1 + az^{-1}), \tag{17}$$

has only one zero at $z = -a$. When $a < 1$, i.e. the echo is weaker than the direct-path signal, the minimum-phase component of $H(z)$ is $(1 + az^{-1})$ and the all-pass component is z^{-n_0}. In this case the minimum-phase response $h_{\mathrm{Min}}(n) \equiv Z^{-1}\{H_{\mathrm{Min}}(z)\}$, where Z^{-1} denotes the inverse z-transform, is identical to the impulse response except for the absence of the direct-path delay n_0, which contains the location information of the source. The latter is contained in the all-pass response $h_{\mathrm{All}}(n) \equiv Z^{-1}\{H_{\mathrm{All}}(z)\}$.

When $a > 1$, i.e. the echo is stronger than the direct-path signal, the minimum-phase and all-pass components of $H(z)$ are given by

$$H_{\mathrm{Min}}(z) = a + z^{-1}, \tag{18}$$

$$H_{\mathrm{All}}(z) = z^{-n_0}\frac{a^{-1} + z^{-1}}{1 + a^{-1}z^{-1}}, \tag{19}$$

and the corresponding minimum-phase and all-pass responses are given by

$$h_{\mathrm{Min}}(n) = a\,\delta(n) + \delta(n - 1), \tag{20}$$

$$h_{\mathrm{All}}(n) = a^{-1}\delta(n - n_0) + (a^{-1} - a)\sum_{m=1}^{\infty}(-a)^{-m}\delta(n - n_0 - m). \tag{21}$$

These responses are illustrated in Fig. 3 for $n_0 = 2$ and $a = 1.5$. From Fig. 3 and Eq. (20), we see that in the minimum-phase response, the direct-path component is still stronger than the echo. Thus, it seems that the minimum-phase response is less severely affected by reverberation than is the original impulse response (Eq. (16)). The situation is quite different for the all-pass response. Indeed, as can be seen from Fig. 3 and Eq. (21), $h_{\mathrm{All}}(n)$ has now an echo tail of infinite duration, in agreement with the pole-zero nature of Eq. (19). Thus, we conclude that the all-pass response is more severely affected by reverberation than is the minimum-phase response. However, and more importantly, we note that reverberation has no effect on the direct-path delay information which is contained in the all-pass response. This location information plays a fundamental role in array processing applications.

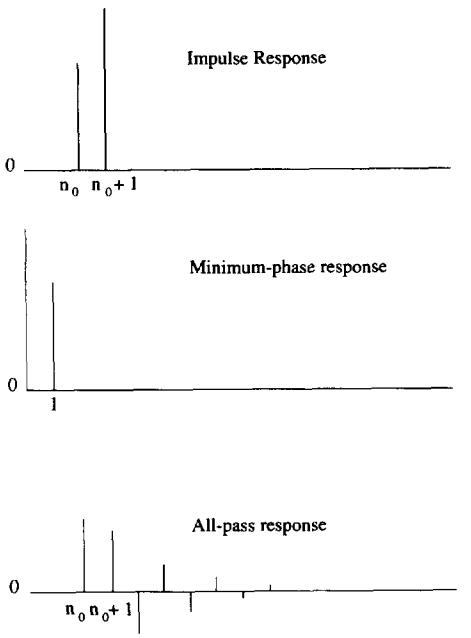

Fig. 3. Impulse response $h_i(n) = \delta(n - n_0) + 1.5\delta(n - n_0 - 1)$ and its minimum-phase and all-pass components.

The above discussion applies only to the simple echo model in Eq. (16). Typical room impulse responses are considerably more complex and contain a very large number of echoes distributed in a random-like fashion on the time axis. As a result, it is difficult to derive closed-form expressions for the minimum-phase and all-pass components of a room impulse response (even for a simple room model) and to determine whether or not the above observations can be generalized. For this purpose, we had recourse to an experimental approach. Following a standard procedure, we generated several synthetic

room impulse responses on a computer using the well-known image model technique (Allen and Berkley, 1979; Peterson, 1986). These responses were then decomposed into minimum-phase and all-pass components. It was found that the conclusions made above regarding the effects of reverberation on the minimum-phase and all-pass responses can also be applied to these synthetic room impulse responses.

To illustrate this point, Fig. 4 shows two synthetic room impulse responses between a common source position and two distinct microphone locations in a reverberant room. The corresponding minimum-phase and all-pass components are also illustrated. As can be seen from Fig. 4(b), each minimum-phase response consists of a main positive peak at the origin followed by several secondary peaks of smaller amplitudes whose envelope decays quite rapidly (i.e. weak echo tail). This is in agreement with the fact that the energy of a minimum-phase sequence is concentrated around the time origin (Oppenheim and Schafer, 1975). As can be seen from Fig. 4(c), the effects of reverberation on the all-pass response are considerably more severe. In particular, the all-pass response is noisier than the minimum-phase one and it contains several echoes whose amplitudes are comparable to (or even larger than) the direct-path component. Yet, as in the simple echo model discussed previously, the direct path delay information is not affected by the reverberation. That is, the location of the first dominant positive peak of $H_{\mathrm{All}}(\omega)$ on the time axis still corresponds to the correct value of direct path propagation delay between the source and the corresponding microphone location.

We have seen that the effects of reverberation on the minimum-phase and the all-pass components of the room impulse response are fundamentally different. To complete this section, let us discuss how these differences should affect our processing philosophy for the microphone signals. For simplicity, assume that there are no common zeros and poles between the z-transforms of the room impulse response and that of the signal. Then, the Fourier transform of the microphone signal, $x(n)$, can be decomposed as

$$X(\omega) = X_{\mathrm{Min}}(\omega) \cdot X_{\mathrm{All}}(\omega), \tag{22}$$

where $X_{\mathrm{Min}}(\omega)$ and $X_{\mathrm{All}}(\omega)$ are the minimum-phase and all-pass components of $X(\omega)$, respectively. These can be further decomposed into

$$X_{\mathrm{Min}}(\omega) = S_{\mathrm{Min}}(\omega) \cdot H_{\mathrm{Min}}(\omega), \tag{23}$$

$$X_{\mathrm{All}}(\omega) = S_{\mathrm{All}}(\omega) \cdot H_{\mathrm{All}}(\omega), \tag{24}$$

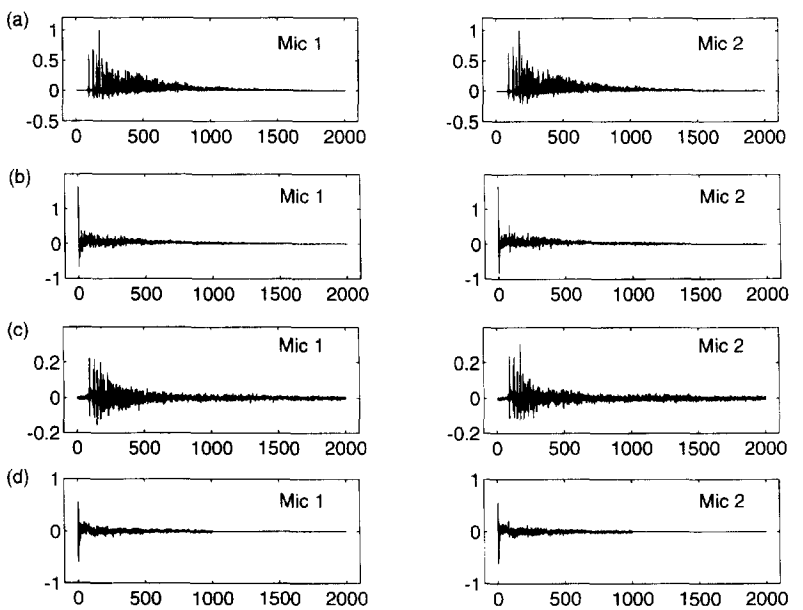

Fig. 4. Minimum-phase and all-pass decompositions of room impulse responses for two spatially separated microphones: (a) impulse response, (b) minimum-phase response, (c) all-pass response and (d) minimum-phase cepstrum.

where $S_{\mathrm{Min}}(\omega)$ and $S_{\mathrm{All}}(\omega)$ respectively denote the minimum-phase and the all-pass components of the speech signal. Thus the minimum-phase and the all-pass components of the speech signal are separately affected by the minimum-phase and the all-pass components of the room impulse response. This suggests that they could be processed in different ways for the purpose of dereverberation.

4. Separate processing of minimum-phase and all-pass components

In the new dereverberation technique that we propose in this paper, the microphone signals are decomposed into minimum-phase and all-pass components. Based on the above discussion, these components are processed separately and differently. In Sections 4.1, 4.2 and 4.3, we describe the processing operations that we propose for each components and how they should be recombined. The implementation details of the decomposition, processing and synthesis will be described in Section 5.

4.1. Minimum-phase components

Let $s_{\mathrm{Min}}(n)$, $h_{i,\mathrm{Min}}(n)$ and $x_{i,\mathrm{Min}}(n)$ respectively denote the minimum-phase components of $s(n)$, $h_i(n)$ and $x_i(n)$ (as defined in Section 2) and let $\hat{s}_{\mathrm{Min}}(n)$, $\hat{h}_{\mathrm{i,Min}}(n)$ and $\hat{x}_{\mathrm{i,Min}}(n)$ be the corresponding complex cepstra. From Eq. (23) and the properties of the complex cepstrum, it follows that

$$\begin{aligned}\hat{x}_{i,\mathrm{Min}}(n) &= C\left[s_{\mathrm{Min}}(n) * h_{i,\mathrm{Min}}(n)\right] \\ &= \hat{s}_{\mathrm{Min}}(n) + \hat{h}_{i,\mathrm{Min}}(n). \end{aligned} \tag{25}$$

As shown in Eq. (25), the minimum-phase signal cepstrum $\hat{s}_{\mathrm{Min}}(n)$ is kept invariant for different channels, while the minimum-phase channel cepstrum $\hat{h}_{\mathrm{i,Min}}(n)$ changes from channel to channel. In Fig. 4(d), we give two examples for the minimum-phase channel cepstra at two different microphones. Each of them consists of a main part around the origin in the quefrency domain followed by an echo part (recall that an ideal impulse function at the origin in the quefrency domain corresponds to an ideal impulse function in the time domain). For different microphone positions, the main parts have some correlation, but the echo parts are found experimentally to have weak spatial correlation. Similar phenomena was observed from the real room impulse responses (Tohyama et al., 1993). Therefore, we propose first to average the minimum-phase cepstra of the individual microphone signals (Eq. (25)) to enhance the signal cepstrum $\hat{s}_{\mathrm{Min}}(n)$, which yields

$$\hat{x}_{0,\mathrm{Min}}(n) = \frac{1}{M}\sum_{i=1}^{M} \hat{x}_{i,\mathrm{Min}}(n). \tag{26}$$

This spatial averaging operation in the quefrency domain is actually equivalent to a geometrical averaging in the z-domain. This observation has a rather important consequence, namely: Eq. (26) preserves the minimum-phase property of the input $\hat{x}_{\mathrm{i,Min}}(n)$.

As mentioned earlier, the speech cepstrum $\hat{s}_{\mathrm{Min}}(n)$ is concentrated mostly in the low-quefrency region. Based on this observation, a low-quefrency cepstrum window is further used to cut off reverberant components of $\hat{h}_{\mathrm{i,Min}}(n)$ remaining in the high-quefrency region. More specifically, let

$$\hat{w}_{\mathrm{Low}}(n) = \begin{cases} 1, & 0 \le n \le n_{\mathrm{c}}, \\ 0, & \text{otherwise}, \end{cases} \tag{27}$$

where the positive quefrency index n_{c} is the cutoff quefrency of the window. In our software implementation of the method, the value of n_{c} is chosen as one quarter of the length of the analysis segment. Applying this window to $\hat{x}_{0,\mathrm{Min}}(n)$ in Eq. (26), we have

$$\hat{y}_{\mathrm{Min}}(n) = \hat{w}_{\mathrm{Low}}(n)\,\hat{x}_{0,\mathrm{Min}}(n). \tag{28}$$

Note that Eq. (28) preserves the minimum-phase nature of the microphone signals, that is, $\hat{y}_{\mathrm{Min}}(n)$ is also minimum-phase.

The final processing step consists of recovering a time-domain signal from $\hat{y}_{i,\mathrm{Min}}(n)$, that is,

$$y_{\mathrm{Min}}(n) = C^{-1}\left[\hat{y}_{\mathrm{Min}}(n)\right]. \tag{29}$$

In the above dereverberation scheme for the minimum-phase signal, both quefrency and spatial processing are applied. The spatial averaging can attenuate the reverberant components in the whole quefrency region. The low-quefrency filtering then removes the remaining echoes in the high quefrency region. The advantage of spatial processing is evident: if only low-quefrency filtering is used

(Tohyama et al., 1993), the reverberant components in the low-quefrency can only be decreased by using a smaller cutoff time, but signal distortion will be noticeable.

Now consider how the source speech is affected by the nonlinear operations in the above procedure. Based on previous Eqs. (25)–(29), we can express $y_{\mathrm{Min}}(n)$ in the form

$$y_{\mathrm{Min}}(n) = C^{-1}\left[\hat{w}_{\mathrm{Low}}(n)\,\hat{s}_{\mathrm{Min}}(n) + \hat{h}_{0,\mathrm{Min}}(n)\right], \tag{30}$$

where

$$\hat{h}_{0,\mathrm{Min}}(n) = \hat{w}_{\mathrm{Low}} \cdot \frac{1}{M}\sum_{i=1}^{M} \hat{h}_{i,\mathrm{Min}}(n). \tag{31}$$

Assume that the loss of source speech through low-quefrency filtering is negligible, i.e., $\hat{w}_{\mathrm{Low}}(n)\hat{s}_{\mathrm{Min}}(n) \approx \hat{s}_{\mathrm{Min}}(n)$. Then Eq. (30) becomes

$$\begin{aligned} y_{\mathrm{Min}}(n) &= C^{-1}\left[\hat{s}_{\mathrm{Min}}(n) + \hat{h}_{0,\mathrm{Min}}(n)\right] \\ &= s_{\mathrm{Min}}(n) * h_{0,\mathrm{Min}}(n), \end{aligned} \tag{32}$$

where $h_{0,\mathrm{Min}}(n) = C^{-1}[\hat{h}_{0,\mathrm{Min}}(n)]$. Thus no nonlinear distortion is introduced for the minimum-phase recovery of the source speech although nonlinear processing is applied.

4.2. All-pass components

As exemplified in Section 3, the all-pass component of a typical room impulse response preserves the position of the direct-path pulse; however, it contains strong echoes with both positive and negative amplitudes that seem to be distributed randomly along the time axis. In fact, we have found that the contributions of these echoes on the all-pass components of different microphone channels are comparable to spatially uncorrelated additive noise. Thus spatial filtering or beamforming can be applied to the all-pass responses $H_{i,\mathrm{All}}(\omega)$, $i = 1, \ldots, M$, in an attempt to attenuate the echo pulses. According to Eq. (24), this is equivalent to applying beamforming to the all-pass components of the microphone signals $X_i(\omega)$.

Assuming that the microphone array has been pre-steered in the direction of the desired source, delay-and-sum beamforming of the all-pass components can be simply expressed in the frequency-domain as

$$Y_{\mathrm{Beam}}(\omega) = \frac{1}{M}\sum_{i=1}^{M} X_{i,\mathrm{All}}(\omega). \tag{33}$$

Substituting Eq. (24) into Eq. (33), we obtain

$$Y_{\mathrm{Beam}}(\omega) = S_{\mathrm{All}}(\omega) \cdot H_{\mathrm{Beam}}(\omega), \tag{34}$$

where

$$H_{\mathrm{Beam}}(\omega) = \frac{1}{M}\sum_{i=1}^{M} H_{i,\mathrm{All}}(\omega). \tag{35}$$

In Eq. (34), $S_{\mathrm{All}}(\omega)$ is the all-pass component of the source speech and so has unit magnitude. However, it is not true in general that the output $H_{\mathrm{Beam}}(\omega)$ of the beamforming operation in Eq. (35) has unit magnitude. Hence, in general $|Y_{\mathrm{Beam}}(\omega)| \neq 1$ so that $Y_{\mathrm{Beam}}(\omega)$ is not an all-pass component. Since the purpose of the processing of the all-pass components of the microphone signals is to obtain an estimate of the all-pass component of the original speech signal, further processing is required.

For this purpose, we propose removing the minimum-phase component of $Y_{\mathrm{Beam}}(\omega)$. For an arbitrary Fourier transform $X(\omega)$, this can be achieved by dividing $X(\omega)$ by its minimum-phase component, which in turn can be evaluated using the procedure described in Section 3. Let $\mathscr{A}\{X(\omega)\}$ denote the operation which assigns to $X(\omega)$ the corresponding all-pass component obtained in this manner. Applying the operator $\mathscr{A}\{\cdot\}$ to Eq. (34) yields

$$Y_{\mathrm{All}}(\omega) \equiv \mathscr{A}\{Y_{\mathrm{Beam}}(\omega)\} = S_{\mathrm{All}}(\omega) \cdot H_{0,\mathrm{All}}(\omega), \tag{36}$$

where $H_{0,\mathrm{All}}(\omega) = \mathscr{A}\{H_{\mathrm{Beam}}(\omega)\}$ is the all-pass component of $H_{\mathrm{Beam}}(\omega)$. In the time domain, Eq. (36) becomes

$$y_{\mathrm{All}}(n) = s_{\mathrm{All}}(n) * h_{0,\mathrm{All}}(n), \tag{37}$$

where $h_{0,\mathrm{All}}(n)$ is the inverse Fourier transform of $H_{0,\mathrm{All}}(\omega)$.

Two other approaches were also tried for the all-pass recovery of the original speech. The first one consists of normalizing $Y_{\mathrm{Beam}}(\omega)$ in Eq. (34) by its magnitude. The second one consists of using the all-pass component of the output of a conventional beamformer applied to the microphone signals as the

all-pass recovery $Y_{All}(\omega)$. However, simulation results show some loss in performance (particularly array gain) when these approaches are used. Thus, we prefer to use Eq. (36) for the all-pass recovery of the original speech.

4.3. Combination of the minimum-phase and all-pass components

In the proposed dereverberation technique, the minimum-phase component $y_{Min}(n)$ (Eq. (32)) and the all-pass component $y_{All}(n)$ (Eq. (37)) are convolved to produce the final output of the system. Thus the recovered speech can be expressed as

$$y(n) = y_{Min}(n) * y_{All}(n). \tag{38}$$

Using Eq. (32) and Eq. (37), we can also write

$$y(n) = s(n) * h_0(n), \tag{39}$$

where

$$h_0(n) = h_{0,Min}(n) * h_{0,All}(n). \tag{40}$$

According to Eq. (39), the final output $y(n)$ is a linear convolution of the speech signal $s(n)$ with $h_0(n)$, which is independent of $s(n)$ and hence can be viewed as an equivalent impulse response for the processor. Thus, in principle, the dereveberation performance can be evaluated independently of the source speech by examining the impulse response $h_0(n)$.

Experimental results showed that considerable reduction of reverberation could be obtained from the equivalent processor impulse response $h_0(n)$ by using the technique described above. However, it was also observed that a large negative peak could be created following the direct-path peak in $h_0(n)$ when the room reflectivity was large. This negative peak was found to result from the non-linear operator $\mathscr{A}$ in Eq. (36), which is used to remove the minimum-phase component of $Y_{Beam}(\omega)$. The operator $\mathscr{A}$ seems to introduce some direct-path signal distortion during the recovery of the all-pass component. To offset this effect, the following fixed filter $D(z)$ was found effective:

$$D(z) = \frac{1}{1 - \alpha z^{-1}}, \tag{41}$$

where $0 < \alpha < 1$. The specific value of α is dependent on the room reflection coefficients. Typical values of α in our applications range from 0.3 to 0.5 (room reflection coefficient 0.7–0.9).

5. Implementation of the new processor

Short-term discrete Fourier transform (DFT) analysis and synthesis techniques are required in a practical implementation of the new processor described above. These techniques consists of two fundamental steps, namely segmentation and reconstruction (Allen, 1977; Portnoff, 1976). In the segmentation step, segments of the reverberant speech are obtained by applying a finite length window to the microphone signals. In essence, dereverberation processing would be applied to these segments. In the reconstruction step, processed speech segments are recombined to form the final output signal. The latter is usually dependent on the type of window used in the implementation. In particular, the output may deviate from the desired response due to the effect of the window. This effect is difficult to remove in the synthesis output if a time-varying spectral modification (i.e., changing from segment to segment) is applied during the dereverberation processing (Allen, 1977), as would be the case with the new technique. In our implementation of the new processor, Allen's short-term DFT analysis and synthesis technique (Allen, 1977) is employed. The overall processor implementation is outlined in block diagram form in Fig. 5. Related computational details are described below.

5.1. Window and segmentation

Following delay alignment (as indicated by the variables τ_i in Fig. 5), a low-pass window function of length L is applied synchronously to each microphone signal $x_i(n)$ $(i = 1, \ldots, M)$. We used a Hamming window. The windowed signal of each channel is padded with zeros to form a segment of length $N > L$. Since modifications will be made to the spectrum of each segment, the values of L and N should be carefully chosen so as to avoid serious time aliasing during synthesis (Allen, 1977; Allen et al., 1977). A typical value of N is $2L$. The array of synchronous segments taken from the M microphone channels are processed as described below to produce a single segment of enhanced speech at the

processor output. Following this processing, the window is shifted along the time axis and the procedure is repeated. To properly reconstruct the successive output segments, overlapping of the analysis window is needed. For a Hamming window, a reasonable length for the window overlap is $3L/4$ (Allen, 1977; Allen et al., 1977).

5.2. Decomposition into minimum-phase and all-pass components

The decomposition of the segments into minimum-phase and all-pass components is performed in the frequency domain. Let $x_i(n, k)$ denote the kth segment of the ith microphone signal, where $i = 1, \ldots, M$, $n = 0, \ldots, N-1$ and $k = 1, \ldots, \infty$. Its discrete Fourier transform is first computed by using the fast Fourier transform (FFT) algorithm, as indicated below:

$$X_i(\omega_l, k) = \mathrm{FFT}[x_i(n, k)], \quad i = 1, \ldots, M, \quad (42)$$

where $\omega_l = 2\pi l/N$ $(l = 0, \ldots, N-1)$ is the discrete frequency. The real cepstrum of the segment $x_i(n, k)$ is then calculated as

$$\hat{x}_i(n, k) = \mathrm{FFT}^{-1}[\log|X_i(\omega_l, k)|], \quad i = 1, \ldots, M, \quad (43)$$

and the cepstrum of the minimum-phase component of $x_i(n, k)$ is

$$\hat{x}_{i,\mathrm{Min}}(n,k) = \hat{r}(n)\, \hat{x}_i(n, k), \quad (44)$$

where $\hat{r}(n)$ is a window function whose purpose is to zero the cepstrum for negative quefrencies:

$$\hat{r}(n) = \begin{cases} 1, & n = 0, N/2, \\ 2, & 1 \le n < N/2, \\ 0, & N/2 < n < N-1. \end{cases} \quad (45)$$

These M channel cepstra are fed into the subprocessor for the minimum-phase recovery while the mini-

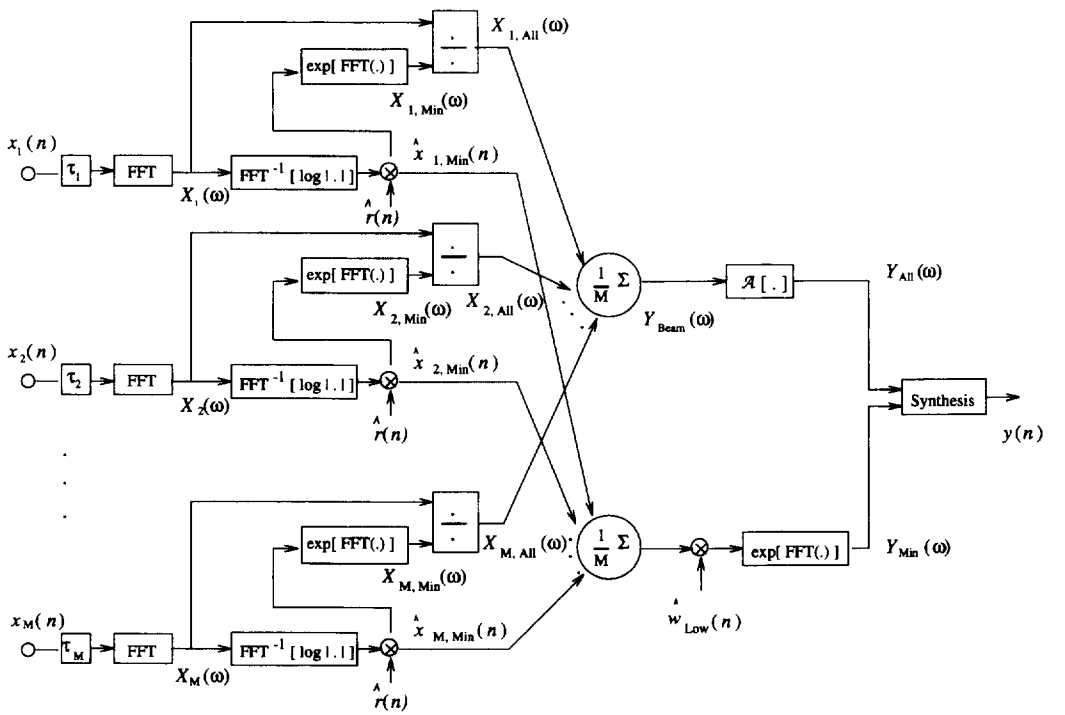

Fig. 5. Block diagram of the new dereverberation processor. In this figure, $\mathcal{A}[\cdot]$ denotes the operation which assigns to its input the corresponding all-pass component.

mum-phase components of $X_i(\omega_l, k)$, $i = 1, \ldots, M$, are obtained as

$$X_{i,\mathrm{Min}}(\omega_l, k) = \exp\left[\mathrm{FFT}\left[\hat{x}_{i,\mathrm{Min}}(n, k)\right]\right]. \quad (46)$$

Finally, the all-pass component $X_{i,\mathrm{All}}(\omega_l, k)$ is obtained as

$$X_{i,\mathrm{All}}(\omega_l, k) = X_i(\omega_l, k)/X_{i,\mathrm{Min}}(\omega_l, k). \quad (47)$$

5.3. Processor kernel

The two subprocessors for the minimum-phase and all-phase components have been described in Section 4. For the kth segment, the minimum-phase and all-pass recoveries of the desired speech signal are implemented according to the following expressions, respectively:

$$Y_{\mathrm{Min}}(\omega_l, k) = \exp\left\{\mathrm{FFT}\left[\hat{w}_{\mathrm{Low}}(n) \cdot \frac{1}{M}\sum_{i=1}^{M} \hat{x}_{i,\mathrm{Min}}(n, k)\right]\right\}, \quad (48)$$

$$Y_{\mathrm{All}}(\omega_l, k) = \mathscr{A}\left[\frac{1}{M}\sum_{i=1}^{M} X_{i,\mathrm{All}}(\omega_l, k)\right], \quad (49)$$

where $\hat{w}_{\mathrm{Low}}(n)$ is the periodic version (with period N) of the window of Eq. (27) and the operator $\mathscr{A}[\cdot]$ is as defined in Section 4.2. Once the all-pass and minimum-phase components are available, the composite signal recovery is obtained as

$$y(n, k) = \mathrm{FFT}^{-1}\left[Y_{\mathrm{Min}}(\omega_l, k) \cdot Y_{\mathrm{All}}(\omega_l, k)\right] \quad (50)$$

for $n = 0, \ldots, N-1$ and $k = 1, \ldots, \infty$.

5.4. Synthesis

Before the synthesis, a final modification is made to each output segment $y(n, k)$. According to Eq. (39), the output segment $y(n, k)$ is (approximately) the linear convolution of the source speech with the impulse response $h_0(n)$. As explained earlier, the position of the direct-path peak in $h_0(n)$ is left invariant by the processing and is the same as that of the individual impulse responses $h_i(n)$ (after delay alignment). So in each output segment $y(n, k)$, the direct-path signal has approximately no time shift relative to that in the input segments $x_i(n, k)$. Because $x_i(n, k)$ was obtained by padding $N - L$ zeros to an L-point microphone signal, it follows from the above observation that in each output segment of length N, the desired signal occupies only the first L samples. Thus, in the processor, the last $N - L$ samples of each output segment are set to zero before the final synthesis, without introducing any signal distortion.

The segment synthesis technique (Allen, 1977) is finally applied to the modified output segments. In essence, it consists of summing up all output segments in the time domain while maintaining the proper phase relationship between the successive time windows used in the segmentation.

6. Results

This section presents the results of simulations and audio tests conducted to evaluate the new dereverberation technique described above. Results for a conventional beamformer are also provided for the purpose of comparison.

A computer implementation of the image method as described in (Allen and Berkley, 1979; Peterson, 1986) is used to generate synthetic room impulse responses for the microphones. The sampling frequency used for the synthesis of the impulse responses is 10 kHz. The room size is assumed to be 5 m (length) × 4 m (width) × 3 m (height). The six walls of the room have the same reflection coefficient. Two different array configurations are used for the evaluation of the new processor, namely: a uniform linear array and a harmonically nested linear array. The simulation scenario for the case of the uniform linear array is illustrated in Fig. 6, where the number of microphones is $M = 17$ and the microphone spacing is 4 cm. The array and the source speaker lie in an horizontal plane at an elevation of 1.5 m. The nested array is obtained by repositioning the last 4 microphones of the two ends of the above linear array. In effect, it can be viewed as the superposition of three uniform linear subarrays with element spacing 4 cm, 8 cm and 16 cm, respectively. Each subarray contains 9 microphones, some of them being shared by the subarrays, so that the total number of microphones is also 17. These subarrays would be applied to speech signals which cover three

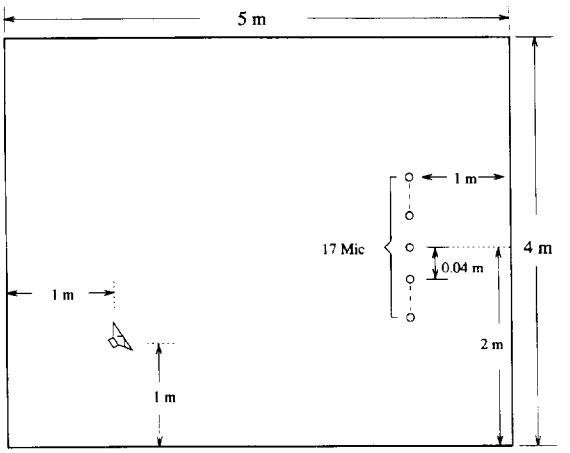

Fig. 6. Simulation scenario: the room size is 5 m × 4 m × 3 m; the speaker and the microphone array are positioned on an horizontal plane at elevation of 1.5 m.

different octaves respectively. However, for simplicity, the whole nested array is used for the full-band speech signals in the following simulations.

Note that a pre-steering of the array is needed in our algorithm. To do this, digital interpolation techniques could be used in practice to obtain a high spatial resolution. In our simulations, for simplicity, we performed the pre-steering before the time sampling in the room response synthesis program. As a result, exact time-alignment of the direct-path microphone signals could be achieved. Higher sampling frequency could also be used to simulate the behavior of an array correctly with the image method.

6.1. Evaluation of processor's impulse response

In this subsection, we first investigate the associated impulse response of the new processor. For each array configuration, the room impulse responses $h_i(n)$ of the M microphones are passed through the new processor and a conventional beamformer, resulting in an equivalent impulse response $h_0(n)$ at the output. The array gain defined in Eq. (4) and Eq. (7) is then used to evaluate the dereverberation performance. This procedure is repeated for several values of the wall reflection coefficient

Array gain versus room reflection coefficient for the new processor and the conventional beamformer is shown in Fig. 7. As can be seen from this figure, for small reflection coefficients, the array gains of the processors are almost identical. But for a typical environment where the wall reflectivity $\beta > 0.7$, the new processor shows a significant array gain improvement over the conventional beamformer. For the uniform linear array, this improvement is between 3 to 6 dB, while for the nested array, the improvement is between 4 to 8 dB. Note that above a certain reflectivity threshold, the array gain of the conventional beamformer decreases when the reflection coefficient increases. This may be due to the fact that several of the echoes in different room impulse responses $h_i(n)$ are 'co-phased' or time-aligned. Thus, as the wall reflectivity increases, conventional beamforming becomes inefficient in smoothing out such spatially correlated echo pulses. Fig. 7 shows that the two array configurations yield different results. In particular, the nested array exhibits an obvious advantage over the uniform linear array (this is also true for the audio tests described below). This result is consistent with other observations found in the literature (Grenier, 1993; Kellermann, 1991) and may be attributed to the better spatial resolution properties of the nested array, as explained in Section 2. In the sequel, only the results for the nested array are presented.

Fig. 8 shows the impulse responses at the outputs of a single microphone, the conventional beamformer and the new processor for the nested array. The wall reflection coefficient used to obtain these

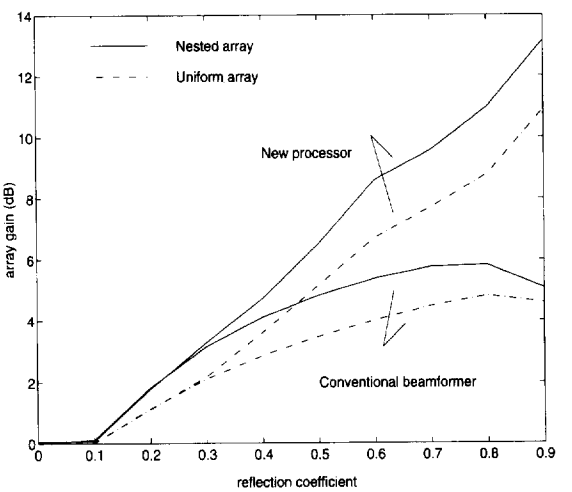

Fig. 7. Array gain versus wall reflection coefficient.

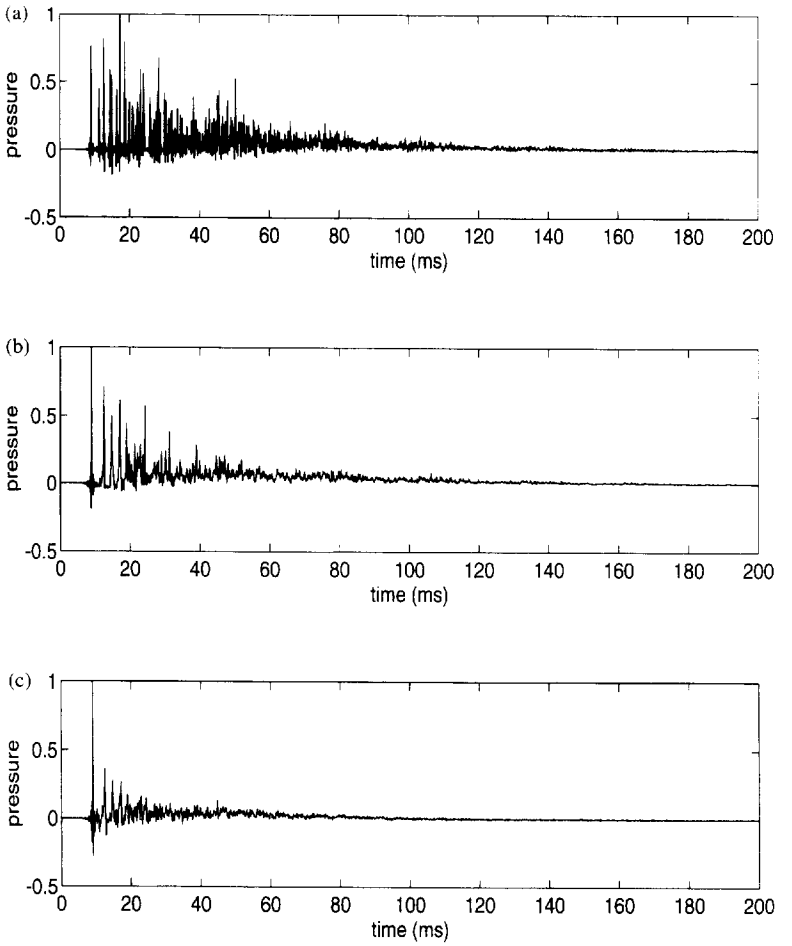

Fig. 8. Impulse responses when the reflection coefficient is 0.8 (reverberation time 0.3 sec): (a) single microphone, (b) conventional beamformer and (c) new processor.

results is $\beta = 0.8$ (the corresponding reverberation time $T_R = 0.3$ sec). It can be seen that the conventional beamformer is effective in suppressing the echoes, while the new processor makes a further significant improvement.

6.2. Dereverberation performance for speech signals

Another set of experiments were performed to evaluate and compare the dereverberation performance of the new processor and the conventional beamformer on speech signals. In the experiment, clean speech with 10 kHz sampling rate was convolved with the room impulse responses $h_i(n)$ to produce the 17 microphone outputs of the nested array. The following parameter values were used for the implementation of the new processor:

Hamming window: $L = 2048$;
FFT length: $N = 2L = 4096$;
Frame overlapping: $3L/4$;
Cutoff time for $\hat{w}_{\text{Low}}(n)$: $n_c = 512$.

Following the dereverberation processing, speech from the output of a selected microphone as well as the enhanced speech at the output of the conventional beamformer and the new processor was sent to a 16-bit digital audio card for listening tests. When the reflection coefficient is large, i.e. $\beta > 0.9$ (i.e. $T_R > 0.54$ sec), strong reveberation is audible for the single microphone speech. An evident reduction of reveberation can be heard from the output of the

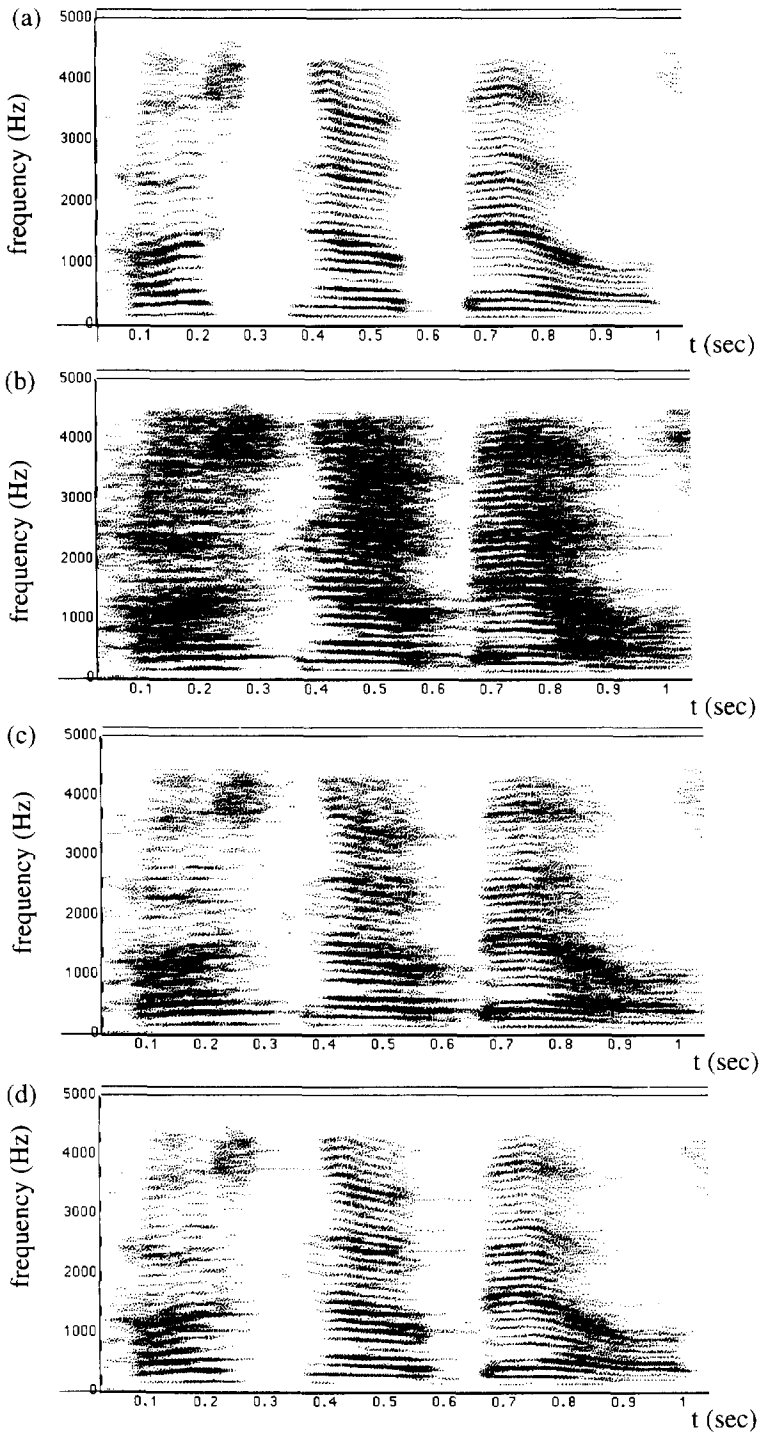

Fig. 9. Speech spectrograms when the reflection coefficient is 0.8 (reverberation time 0.3 sec): (a) anechoic speech, (b) reverberant speech at one single microphone, (c) speech processed by the conventional beamformer and (d) speech processed by the new processor. (The speech shown is the first part of the sentence "post no bills on this office wall".)

conventional beamformer; but there still remain some audible reverberation especially in low frequencies. For the new processor, a further reduction of reverberation was obtained as compared to the conventional beamformer. Both male and female speeches were tested and the results were similar.

Fig. 9 shows the narrow-band spectrograms of the original (anechoic), reverberant (single microphone) and the dereverberated (conventional beamformer and new cepstral processor) speech signals when $\beta = 0.8$ ($T_R = 0.3$ sec). These spectrograms are obtained from the convolutions of an anechoic speech with the corresponding impulse responses. Comparing Fig. 9(d) with (c), we see that the new processor shows a cleaner harmonic structure in the spectrogram. Furthermore, a significant reduction of reverberation, especially in the low frequencies, is observed with the new processor. These observations are consistent with the results of the audio tests. [2]

The new technique presented in this paper is also applicable in the presence of low-level uncorrelated background noise. In particular, for signal-plus-reverberation to noise ratio on the order of 20 to 30 dB, the background noise has no audible effects on the dereverberated speech. An interesting avenue for future research would be to investigate more extensively the effects of background noise, directional interferences and other modeling errors on the performance of the new dereverberation technique.

7. Discussion

In this paper, a new multi-microphone dereverberation technique was proposed which is well suited to acoustic environments in which the impulse responses are time-variant. The new technique combines spatial and cepstral processing of the delay-steered microphone signals and is motivated by the observation that the minimum-phase and the all-pass components of the microphone signals are affected differently by the room acoustics. When compared to a conventional beamformer, the new processor resulted in a 4–8 dB array gain improvement. The dereverberation effects of the new processor and its advantages over the conventional beamformer were also verified in listening tests. The simulations and tests for the new processor also raised several interesting issues, some of which are discussed below.

In general, the process of minimum-phase and all-pass decomposition will introduce extra zeros in the minimum-phase component and extra poles in the all-pass components. These zeros and poles will cancel each other when the two components are recombined. Even if a mixed-phase signal has finite duration, its all-pass component will be infinite due to these extra poles. Thus, time-aliasing will occur in the discrete expressions of the all-pass components as shown in Fig. 4(c). When the minimum-phase and all-pass components are processed differently, their combination may fail to eliminate the extra poles. As a result, time-aliasing will also appear in the final output. This explains why some time-aliasing occurs in the impulse response of the new processor, as shown in Fig. 8(c). As the length of the impulse responses $h_i(n)$ increases, more poles are introduced in their all-pass components and the time-aliasing becomes more serious. Padding zeros to the analysis window can reduce the extent of time-aliasing, but a too long analysis window may be impractical for implementation.

Exponential weighting can be used to convert some mixed-phase components into minimum-phase. In effect, exponential weighting emphasizes the minimum-phase processing by reducing the number of poles in the all-pass component. As a result, the time-aliasing phenomenon discussed above will become less significant when exponential weighting is used. However, if the analysis window is very long, which is often required in the case of cepstrum-based processing, serious numerical errors may be produced in the recovery process due to exponential deweighting.

One possible solution to this problem is to deweight only the minimum-phase component which decays rapidly, but not the all-pass component which is more severely affected by deweighting. This process is further motivated by the fact that the phase information, which is contained in the all-pass component, is not too important for the perception of speech signals. Indeed, several simulations of this approach have shown better results in terms of the

[2] The corresponding audiofiles are available at http://www.elsevier.nl/locate/specome .

processor's impulse response. However, when it is used in connection with the short-time DFT analysis/synthesis technique described previously for the dereverberation of continuous speech signals, there still exist some problems. If only the minimum-phase component is deweighted before the combination, discontinuity between consecutive segments may occur due to a loss of phase information. When the length of the segment is small (< 25 msec in our applications), no speech distortion can be perceived as a result of this effect. However, if the length of the segment is increased for the purpose of efficient dereverberation, phase discontinuity can be heard.

These preliminary results indicate that the parameter of the exponential weighting, the length of the analysis window and the type of segmentation/reconstruction scheme are important factors for the efficient application of exponential weighting. Further research is needed in this area.

As for the case of continuous speech dereverberation, there also exist some implementation errors in the new processor. We showed in (39) that the output of the processor is approximately a convolution of the source speech with the impulse response of the processor. But when continuous speech signals are processed, this result may not be true for each output segment. This is due to the fact that each segment of reverberant speech cannot be expressed exactly as the convolution of a segment of clean speech with the room impulse response (Bees et al., 1991). These segmentation errors, which are common in many segment-based processing systems, become more significant as the length of the room impulse response increases or, equivalently, as the analysis window is made shorter. Therefore, the use of an analysis/synthesis scheme for the processing of continuous speech usually does not produce exactly the same result as the direct convolution of the processor equivalent impulse response with the original speech. In our experiments, this type of implementation errors sometimes introduced some slight performance reduction for dereverberation when the analysis/synthesis technique (Allen, 1977) was employed. Thus, more robust analysis/synthesis techniques could be considered for use in continuous speech dereverberation.

Finally, we need to point out that most of the observations and the conclusions made in this paper are based on the use of synthetic room impulse responses generated with the image method. This type of approach has been widely used in basic studies of room acoustics because of its controllability and reproducibility. Yet, an interesting avenue for a future work would be to study the properties of the minimum phase and all-pass components of real room impulse responses and to evaluate the proposed algorithm in a practical environment.

Acknowledgements

The authors wish to thank the two anonymous reviewers for their many helpful comments and suggestions for improving the early versions of the paper.

This work was supported in part by a grant from Fonds pour la Formation de Chercheurs et l'Aide à la Recherche (FCAR), Government of Quebec.

References

J.B. Allen (1977), "Short term spectral analysis, synthesis, and modification by discrete Fourier transform", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 25, pp. 235–238.

J.B. Allen and D.A. Berkley (1979), "Image method for efficiently simulating small-room acoustics", *J. Acoust. Soc. Amer.*, Vol. 65, No. 4, pp. 943–950.

J.B. Allen, D.A. Berkley and J. Blauert (1977), "Multimicrophone signal-processing technique to remove room reverberation from speech signal", *J. Acoust. Soc. Amer.*, Vol. 62, No. 4, pp. 912–915.

D. Bees, M. Blostein and P. Kabal (1991), "Reverberant speech enhancement using cepstral processing", *Proc. Internat. Conf. Acoust. Speech Signal Process.'91*, Toronto, Canada, pp. 977–980.

G. Clifford Carter, Ed. (1993), *Coherence and Time Delay Estimation* (IEEE Press, New York).

E.M. Dowling, D.A. Linebarger, Y. Tong and M. Munoz (1992), "An adaptive microphone array processing system", *Microprocessor and Microsystems*, Vol. 16, No. 10, pp. 507–516.

J.L. Flanagan (1985), "Beamwidth and usable bandwidth of delay-steered microphone arrays", *AT&T Tech. J.*, Vol. 64, No. 4, pp. 983–995.

J.L. Flanagan, J.D. Johnston, R. Zahn and G.W. Elko (1985), "Computer-steered microphone arrays for sound transduction in large rooms", *J. Acoust. Soc. Amer.*, Vol. 78, No. 5, pp. 1508–1518.

J.L. Flanagan, D.A. Berkley, G.W. Elko and M.M. Sondhi (1991), "Autodirective microphone arrays systems", *Acustica*, Vol. 73, pp. 58–71.

M.M. Goodwin and G.W. Elko (1993), "Constant beamwidth beamforming", *Proc. Internat. Conf. Acoust. Speech Signal Process.'93*, Minneapolis, MN, pp. I.169–I.172.

Y. Grenier (1993), "A microphone array for car environments", *Speech Communication*, Vol. 12, No. 1, pp. 25–39.

W. Kellermann (1991), "A self-steering digital microphone array", *Proc. Internat. Conf. Acoust. Speeech Signal Process.'91*, Toronto, Canada, pp. 3581–3584.

M. Miyoshi and Y. Kaneda (1988), "Inverse filtering of room acoustics", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 36, pp. 145–152.

J. Mourjopoulos (1985), "On the variation and invertibility of room impulse response function", *J. Sound and Vibration*, Vol. 102, No. 2, pp. 217–228.

S.T. Neely and J.B. Allen (1979), "Invertibility of a room impulse response", *J. Acoust. Soc. Amer.*, Vol. 66, pp. 165–169.

A.V. Oppenheim and R.W. Schafer (1975), *Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ).

P.M. Peterson (1986), "Simulating the response of multiple microphones to a single acoustic source in a reverberant room", *J. Acoust. Soc. Amer.*, Vol. 80, No. 5, pp. 1527–1529.

F. Pirz (1979), "Design of a wideband, constant beamwidth, array microphone for use in the near field", *AT&T Bell Syst. Tech. J.*, Vol. 58, No. 8, pp. 1839–1850.

M.R. Portnoff (1976), "Implementation of the digital phase vocoder using the fast Fourier transform", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 24, pp. 243–248.

H.F. Silverman and S.E. Kirtman (1992), "A two-stage algorithm for determining talking location from linear microphone array data", *Computer Speech and Language*, Vol. 6, pp. 129–192.

C. Sydow (1994), "Broadband beamforming for a microphone array", *J. Acoust. Soc. Amer.*, Vol. 96, No. 2, pp. 845–849.

M. Tanaka and Y. Kaneda (1993), "Performance of sound source direction estimation methods under reverberant conditions", *J. Acoust. Soc. Japan (E)*, Vol. 14, No. 4, pp. 291–292.

M. Tohyama, R.H. Lyon and T. Koike (1993), "Source waveform recovery in a reverberant space by cepstrum dereverberation", *Proc. Internat. Conf. Acoust. Speech Signal Process.'93*, Minneapolis, MN, pp. I.157–I.160.

D. van Compernolle, W. Ma, F. Xie and M. van Diest (1990), "Speech recognition in noisy environments with the aid of microphone arrays", *Speech Communication*, Vol. 9, Nos. 5/6, pp. 433–442.

J.P. Walsh (1985), "On limitations of minimum mean-square error deconvolution in deriving impulse response of rooms", *J. Acoust. Soc. Amer.*, Vol. 77, No. 2, pp. 547–556.