Proc. IEEE Pacific Rim Conf. Commun., Computers, Signal Processing (Victoria, BC), pp. 571-574, May 1995

Room Speech Dereverberation via Minimum-phase and All-pass Component Processing of Multi-microphone Signals

Qing-Guang Liu [1], Benoît Champagne[1] and Peter Kabal [1,2]

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

[2] Dept. of Electrical Engineering, McGill University, 3480 University Street
Montreal, Québec, Canada H3A 2A7

Abstract - In this paper, a new microphone array processing technique is proposed for blind dereverberation of speech signals effected by room acoustics. It is based on the separate processing of the minimum-phase and all-pass components of multi-microphone signals. 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. A significant improvement in array gain and a reduction of reverberation in listening tests are observed.

I. INTRODUCTION

In many applications of speech communications such as hands-free telephony and audio-conferencing, dereverberation techniques are required for enhancing the intelligibility of speech degraded through the addition of multiple echoes. Since the impulse responses of typical rooms are non-minimum-phase and have therefore unstable inverses [1], inverse filtering-based deconvolution methods have a limited scope in practice. 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 is provided by cepstrum filtering techniques [2], where low-time filtering or peak-picking in the quefrency domain is 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 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 [3] and to certain numerical errors associated with the use of exponential weighting.

Microphone array techniques have long been proposed for the removal of room reverberation. In [4], Allen *et al.* describe a two-microphone technique to remove room reverberation from speech signals. This approach, which is a form of delay-and-sum beamforming, takes advantage of the uncorrelated nature of reverberant speech tails at different locations. Other multi-microphone techniques which can be used for room dereveberation have also been presented in the literature (e.g., [5]-[6]).

In this paper, we present a new technique to remove room reverberation which is based on the joint use of microphone array and cepstrum processing. In this technique, the microphone signals are first delay-steered and then decomposed into minimum-phase and all-pass components. The former are processed in the cepstrum-domain, where spatial averaging followed by low-time filtering is applied. The later are processed in the frequency-domain by performing spatial averaging and by retaining only the all-pass component of the resulting output. Simulation results and listening tests of the new processor indicate a significant improvement in dereverberation performance.

II. BEAMFORMING AND MICROPHONE ARRAYS

Consider an array of M microphones in a reverberant acoustical space. 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 and * denotes convolution. In a conventional (delay-and-sum) beamformer, $x_i(n)$ $(i = 1,..., M)$ 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 $y(n)$:

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

where

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

In (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 phase-aligned and reinforced while echoes apart from the steering direction are attenuated. In the sequel, the time delays τ_i are assumed to be known. The weights w_i in (3) are used to shape the spatial

Support for this work has been provided by FCAR grant 93-95-ER-1577.

0-7803-2553-2/95/$4.00

directivity pattern of the beamformer. We note that the beampattern is dependent on the signal frequency and the array configuration. 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 (e.g., [6]). An harmonically nested array with identical beampattern over three octaves will be used in the simulation section.

III. 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. A useful representation of $H(\omega)$ is given by the factorization [1]-[2]:

$$H(\omega) = H_{Min}(\omega) \cdot H_{All}(\omega) \qquad (4)$$

where $H_{Min}(\omega)$ and $H_{All}(\omega)$ are respectively the minimum-phase and all-pass components of $H(\omega)$. A signal is said to be minimum-phase if its z-transform contains no poles or zeros outside the unit circle. The minimum-phase component $H_{Min}(\omega)$ depends only on the magnitude of $H(\omega)$ and not on its phase. The phase information is contained in the all-pass component $H_{All}(\omega)$, which has a unit magnitude. The decomposition of a signal into a minimum-phase and an all-pass component is usually carried out in the cepstrum domain [2].

Now consider the effect of reverberation on the minimum-phase and all-pass components of room impulse responses. Fig.2 (a) shows two synthetic room impulse responses [7] between a common source position and two distinct microphone locations in a reverberant room. The corresponding minimum-phase and all-pass components are illustrated in Fig.2 (b) and (c), respectively. As can be seen, 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. This is in agreement with the fact that the energy of a minimum-phase sequence is concentrated around the time origin [2]. From Fig.2 (c), we see that the effects of reverberation on the all-pass response are considerably more severe. In particular, the all-pass response is more noisy than the minimum-phase one. However, and more importantly, we note that the direct path delay information in the all-pass components is not affected by the reverberation. That is, the location of the first dominant positive peak of $H_{All}(\omega)$ on the time axis still corresponds to the correct delay of direct path propagation between the source and the corresponding microphone. This location information plays a fundamental role in array processing applications.

In summary, the effects of reveberation on the minimum-phase and all-pass components of room impulse responses are fundamentally different. As for the microphone signal, its minimum-phase and all-pass components are separately affected by the corresponding components of the room impulse response. This suggests that they should be processed in different ways for the purpose of dereverberation.

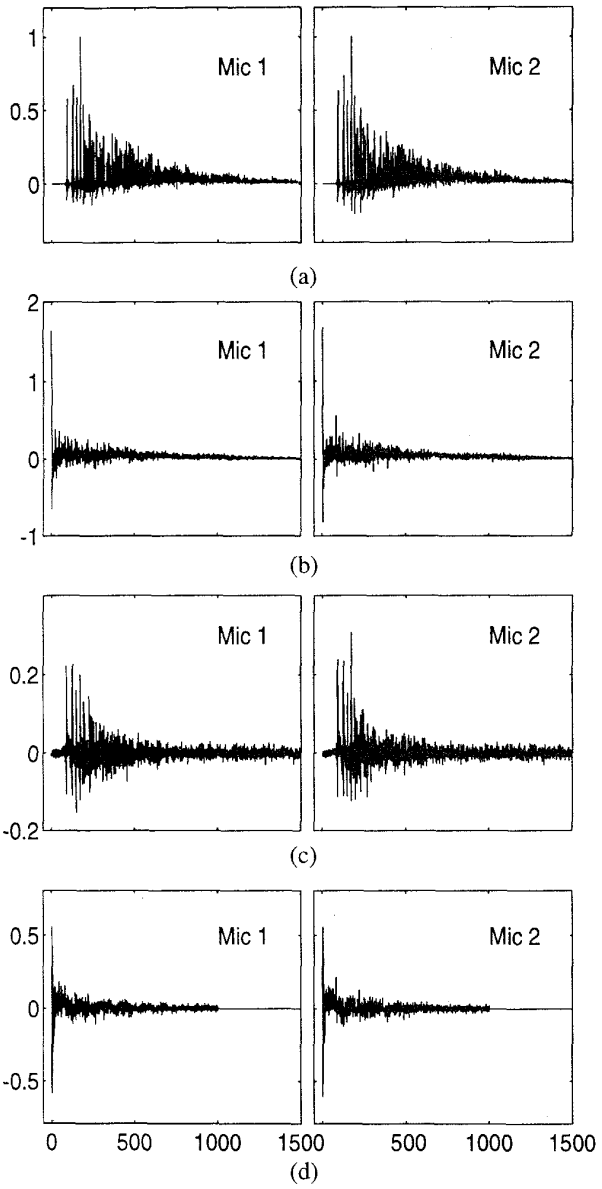

Fig.1. Minimum-phase and all-pass decomposition of room impulse responses for two spatially separated microphones: (a) impulse response, (b) minimum-phase response, (c) all-pass response and (d) minimum-phase cepstrum.

IV. SEPARATE PROCESSING

A. Minimum-phase recovery

Let $s_{Min}(n)$, $h_{i,Min}(n)$ and $x_{i,Min}(n)$ denote the minimum-phase components of $s(n)$, $h_i(n)$ and $x_i(n)$ and let $\hat{s}_{Min}(n)$, $\hat{h}_{i,Min}(n)$ and $\hat{x}_{i,Min}(n)$ be the corresponding complex cepstra. From the properties of the cepstrum [2], we have

$$\hat{x}_{i,Min}(n) = \hat{s}_{Min}(n) + \hat{h}_{i,Min}(n) \quad (5)$$

As shown in (5), the minimum-phase signal cepstrum $\hat{s}_{Min}(n)$ is kept invariant for different channels while the minimum-phase channel cepstrum $\hat{h}_{i,Min}(n)$ changes from channel to channel. In Fig.2 (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. For different microphone positions, the main parts have some correlation, but the echo partts are found experimentally to have a weak spatial correlation. Therefore, we propose first to average the minimum-phase cepstra of the individual microphone signals (5) to enhance the signal cepstrum $\hat{s}_{Min}(n)$, which yields

$$\hat{x}_{0,Min}(n) = \frac{1}{M}\sum_{i=1}^{M}\hat{x}_{i,Min}(n) \quad (6)$$

It is known [2] that the speech cepstrum $\hat{s}_{Min}(n)$ is concentrated mostly in the low-quefrency region. Thus, a low-quefrency cepstrum window $\hat{w}_{Low}(n)$ is further applied to $\hat{x}_{0,Min}(n)$ to cut off reverberant components which remain in the high-quefrency region. As a result, we have:

$$\hat{y}_{Min}(n) = \hat{w}_{Low}(n)\,\hat{x}_{0,Min}(n) \quad (7)$$

The final processing step consists of recovering a time-domain signal $y_{Min}(n)$ from $\hat{y}_{Min}(n)$. In the above dereverberation scheme for the minimum-phase component, both quefrency and spacial 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.

B. All-pass recovery

As we discussed in the previous section, 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 quefrency axis. Thus spatial filtering can be applied to the all-pass responses $H_{i,All}(\omega)$ (i =1,...,M) in an attempt to attenuate the echo pulses. 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 in the frequency-domain can be expressed as

$$Y_{Beam}(\omega) = \frac{1}{M}\sum_{i=1}^{M}X_{i,All}(\omega) \quad (8)$$

It is not true in general that the output $Y_{Beam}(\omega)$ of the beamforming operation in (8) has unit magnitude. Hence, $Y_{Beam}(\omega)$ is not an all-pass component. To recover an estimate of the all-pass component of the original speech signal, we propose to remove the minimum-phase component of $Y_{Beam}(\omega)$. Let $\mathcal{A}\{X(\omega)\}$ denote the operation which assigns to an arbitrary Fourier transform $X(\omega)$ the corresponding all-pass component. The final all-pass recovery will be:

$$Y_{All}(\omega) = \mathcal{A}\{Y_{Beam}(\omega)\} \quad (9)$$

The above recovery is then combined with the minimum-phase recovery (7) to produce the final output of the system. It can be proved that the final output y(n) can be approximately expressed as a linear convolution of the speech signal $s(n)$ with some signal $h_0(n)$, i.e.,

$$y(n) = s(n) * h_0(n) \quad (10)$$

where $h_0(n)$ is independent of the input signal $s(n)$ and can be viewed as an equivalent impulse response for the processor. Thus, the dereverberation performance can be evaluated independently of the source speech by examining the impulse response $h_0(n)$.

V. RESULTS

A computer implementation of the image method as described in [7] 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 5m(length) × 4m(width) × 3m(height). A harmonically nested linear array is used in simulations and tests. It can be viewed as the superposition of three uniform linear subarrays with element spacing 4cm, 8cm and 16cm respectively. Each subarray contains 9 microphones, some of them being shared by the subarrays, so that the total number of microphones is 17.

We first investigate the associated impulse response of the new processor and the conventional beamformer. Array gain is employed to measure the improvement of the Signal-to-Echo Ratio (SER). In this paper, the array gain is defined as

$$AG = M\frac{\mathrm{SER}_0}{\sum_{i=1}^{M}\mathrm{SER}_i} \quad (11)$$

where SER_0 is the SER at the output of the array processor and SER_i (i =1,...,M) is the SER at the output of the ith microphone. Array gains versus reflection coefficient for the new processor and the conventional beamformer are shown in Fig.2. As can be seen from this figure, for a typical environment where the wall reflectivity β is greater than 0.7, the new processor shows a significant improvement over the conventional beamformer. Fig.3 shows the impulse responses at the outputs of a single microphone, the conventional beamformer and the new processor for β = 0.8 (the corresponding reverberation time T_R is approximately 0.3 sec).

Another set of experiments were performed to evaluate the dereverberation performance of the new processor on speech

signals. In the dereverberation process for continuous speech signals, short-term DFT analysis/synthesis techniques are required. These techniques consists of two fundamental steps, namely: segmentation and reconstruction. In the segmentation step, segments of the reverberant speech are obtained by applying a finite length window to the microphone signals. To properly reconstruct the successive output segments, overlapping of the analysis window is needed. In the reconstruction step, processed speech segments are recombined to form the final output signal.

In the implementation of the new processor, Allen's short-term DFT analysis and synthesis technique [4] is employed. In the experiment, clean speech with 10 kHz sampling rate is convolved with the room impulse responses $h_i(n)$ to produce the 17 microphone outputs of the nested array. In the implementation of the new processor, a Hamming window of length 2048 is used for each frame of speech and an 4096-point FFT provides analysis bins. The cutoff time for the low-time cepstral filter $\hat{w}_{Low}(n)$ in (7) is 512 points. Following the dereverberation processing, speech from the output of a selected microphone as well as the enhanced speech at the output of the new processor is sent to a 16-bit digital audio card for listening tests. When the reflection coefficient is large ($\beta > 0.9$ or equivalently $T_R > 0.54$s), strong echoes are audible for the single microphone speech. An evident reduction of echoes can be heard from the output of the new processor.

VI. CONCLUSIONS AND DISCUSSION

In this paper, a new segment-based multi-microphone dereverberation technique was proposed which is particularly well suited to acoustic environments in which the impulse responses are time-variant. The new technique combines spatial and cepstral processing of the 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.

By examining the impulse response in simulations, significant array gain improvement is obtained for the proposed technique. Listening test for continuous speech signal also show evident effect of dereverberation. We note however that the use of an analysis/synthesis technique for the processing of continuous speech does not produce the same result as the direct convolution of the processor equivalent impulse response with the original speech. This is caused by errors in the convolution model as a result of segmentation [3]. Thus, more robust analysis/synthesis techniques are needed for the application of the new dereverberation technique presented here to continuous speech.

REFERENCES

[1] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Am.*, vol.66, pp. 165-169, 1979.

[2] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, 1975.

[3] D. Bees, M. Blostein and P. Kabal, "Reverberant speech enhancement using cepstral processing," *Proc. ICASSP'91*, Toronto, Canada, May 1991, pp. 977-980.

[4] J. B. Allen, D. A. Berkley and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signal," *J. Acoust. Soc. Am.*, vol. 62, No.4, pp. 912-915, 1977.

[5] J. L. Flanagan, J. D. Johnson, R. Zahn and G. W. Elko, "Computer steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.*, vol. 78, No. 5, pp. 1508-1518, 1985.

[6] F. Pirz, "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, 1979.

[7] P. M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Am.* vol. 80, No. 5, pp. 1527-1529, 1986.

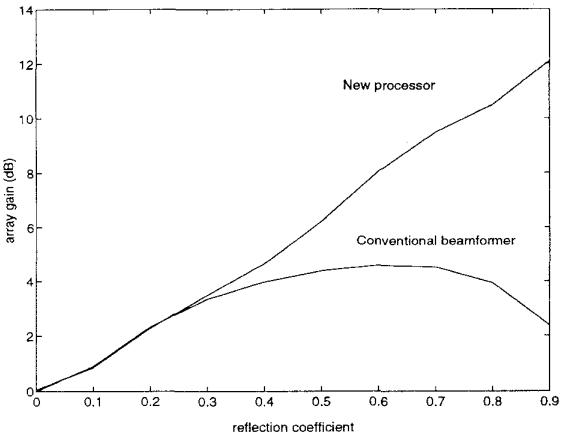

Fig.2. Array gain versus wall reflection coefficient.

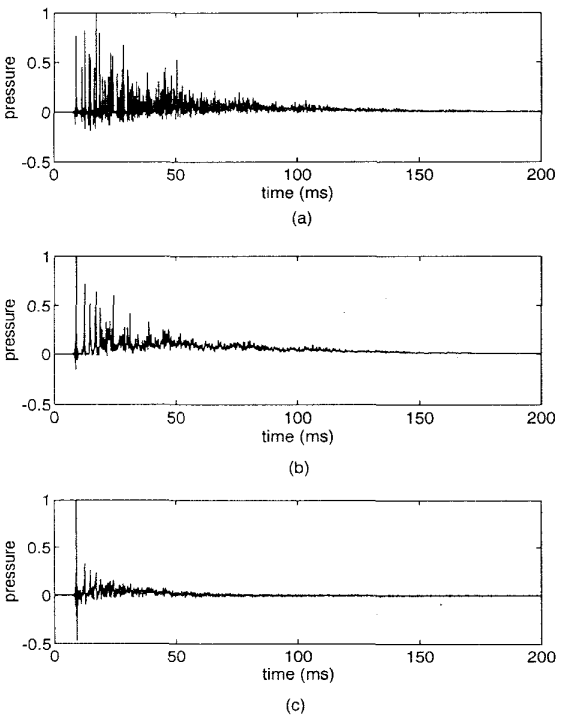

Fig.3. Impulse responses: (a) single microphone, (b) conventional beamformer and (c) new processor.