MULTICHANNEL TELECONFERENCING SYSTEM WITH MULTI SPATIAL REGION ACOUSTIC ECHO CANCELLATION

Kong-Aik Lee, Woon-Seng Gan, Jun Yang, and Farook Sattar

School of Electrical & Electronic Engineering,
Nanyang Technological University, Singapore 639798

ABSTRACT

In this paper a novel configuration for multichannel teleconferencing with multi spatial region acoustic echo cancellation is developed. In the proposed configuration, the transmission and receiving rooms are divided into equivalent number of spatial regions and adaptive filters are attached to each spatial region for the purpose of echo cancellation. This multi spatial region acoustic echo canceller (AEC) does not suffer from nonuniqueness problem and it involves less computational complexity compared to conventional multichannel AEC. Furthermore, the proposed multi spatial region teleconferencing configuration requires narrower transmission bandwidth. Simulation results show the effectiveness of the proposed configuration.

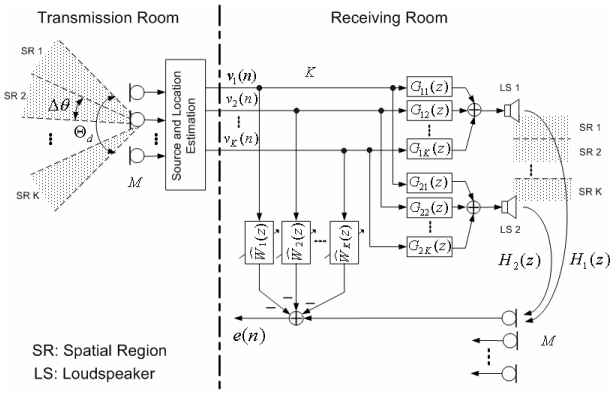

Figure 1: The proposed multichannel teleconferencing system with multi spatial region acoustic echo cancellation.

1. INTRODUCTION

Seamless communication between two remote places is always of great demand. This can be achieved with wider signal bandwidth and increase number of audio channels. With multichannel audio, spatial information is also transmitted. This extra information enables a listener to aurally localize a remote talker in other end. Hence, multichannel teleconferencing provides a more realistic present than monophonic system.

In hands-free communication systems, where acoustic coupling between loudspeaker and microphone is unavoidable, acoustic echo cancellers are necessary. For multichannel system, multichannel acoustic echo cancellers are required. To date, the major problems in multichannel acoustic echo cancellation are high computational complexity of adaptation algorithms and the nonuniqueness problem due to high coherency between channels [1, 2, 3].

A key point in providing a seamless communication is the mapping of sound scene from far-end to near-end. Signal acquisition (in transmission room) and reproduction (in receiving room) methodologies play an important role in providing a realistic and accurate sound scene mapping. In this paper, a novel configuration for multichannel teleconferencing is proposed. In the proposed configuration, the transmission and receiving rooms are divided into equivalent number of spatial regions. Sound images are mapped from transmission to receiving rooms based on their spatial regions. With this multi spatial region configuration, the AEC does not suffer from nonuniqueness problem. Furthermore, it reduces computational complexity and transmission bandwidth as explained in the next section.

2. THE PROPOSED CONFIGURATION

Figure 1 depicts the proposed configuration. In the transmission room, spatial regions are defined based on direction of arrival (DOA), q . Assume that K spatial regions are defined covering all direction of interest, Θ_d . A spatial region is then defined as a range of DOA denoted as

$$\Theta_k = \left[q_k - \frac{\Delta q}{2}, q_k + \frac{\Delta q}{2} \right), \quad k = 1, 2, \ldots, K \tag{1}$$

where q_k is the center of the kth spatial region, and

$$\Delta q = \frac{\Theta_d}{K}, \quad 0 < \Theta_d \le 180^\circ \tag{2}$$

is the range of each spatial region.

By defining spatial regions, Θ_k , we are actually dividing all the potential sources of interest into separate groups based on their DOAs. Signals acquired from sources located in the same spatial region are combined to form spatial region channel, $v_k(n)$, $k = 1, 2, \ldots, K$. Signals originated from sources located in the kth spatial region can only appear in spatial region channel $v_k(n)$. Hence, spatial region channels carry mutually independent signals with assumption that all sources are statistically independent. This is an important characteristic that

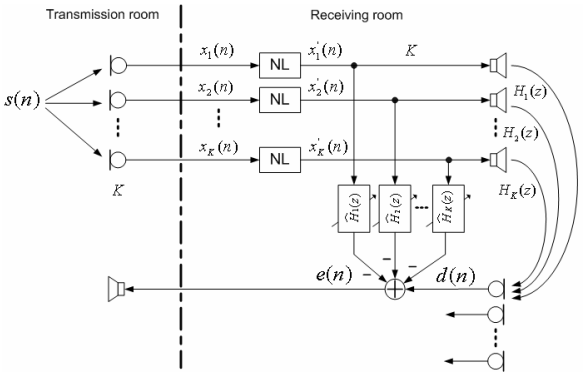

Figure 2: Conventional *K*-channel acoustic echo cancellation configuration.

enables the proposed configuration to avoid the nonuniqueness problem, as explained later.

In the receiving room, sound image of $v_k(n)$ is mapped to its corresponding spatial region. This mapping is performed by a set of spatialization filters, $\mathbf{G}_k(z)$. The overall idea is that, sound images of signals originated from sources located in the kth spatial region in the transmission room are mapped to the kth spatial region in the receiving room, in such a way that the perceived sound image is located at q_k. For reproduction over two loudspeakers, interchannel intensity difference and interchannel time difference can be manipulated to localize the sound images. Sophisticated method as in [4] can be used as well.

2.1. Conventional multichannel acoustic echo cancellation

Figure 2 shows a *K*-channel acoustic echo cancellation configuration as indicated in [1, 2, 3]. The fundamental idea of multichannel acoustic echo cancellation is to model the K acoustic echo paths between K loudspeakers and a microphone with K adaptive FIR filters. These adaptive filters create replica of the echo signal, which can be used to suppress the echo signal picked up by the microphone. The same model applied to other microphones, meaning that $K \times K$ adaptive filters need to be adapted simultaneously.

As indicated in [1, 2, 3], the major problem in this multichannel AEC structure is the high coherency between channels due to the fact that the transmission room signals $x_1(n),\ldots,x_K(n)$ are filtered versions of a common source $s(n)$. High coherency between channels prohibits the adaptive filters to uniquely identify the acoustic echo paths (i.e., the nonuniqueness problem). Consequently, the adaptive filters need to track impulse responses changes in the receiving and the transmission rooms as well. This nonuniqueness problem can be mitigated by adding nonlinear distortion to decorrelate the transmission room signals. This nonlinear preprocessing is indicated as the NL block in Figure 2.

Furthermore, due to the high coherency between channels, common NLMS (Normalized Least-Mean-Square) adaptation algorithm that does not take coherency between channels into account converges slowly. Hence, more sophisticated adaptation algorithms such as APA (Affine Projection) or RLS (Recursive Least-Squares) with higher computational complexity should be employed [3].

2.2. Multi spatial region acoustic echo cancellation

In the proposed configuration, the spatial region channels $v_k(n)$ are mutually independent, which implies zero coherency between channels. Using these mutually independent signals as reference signals to the adaptive filters (as shown in Figure 1), the nonuniqueness problem is essentially eliminated. Hence, nonlinear preprocessing that may cause audible distortion can be avoided. Furthermore, with mutually independent $v_k(n)$, NLMS algorithm with less computational complexity is sufficient to adapt the modeling filters.

With K spatial regions, K adaptive filters are needed, which converge to the weighted sum of the acoustic echo paths $\mathbf{H}(z)$, given by

$$W_k(z) = \mathbf{H}(z)\mathbf{G}_k(z) \qquad (3)$$

where,

$$\mathbf{H}(z) = \begin{bmatrix} H_1(z) & H_2(z) \end{bmatrix}, \quad \mathbf{G}_k(z) = \begin{bmatrix} G_{1k}(z) & G_{2k}(z) \end{bmatrix}^T.$$

As shown in Figure 1, adaptive filter $\widehat{W}_k(z)$ is attached to the kth spatial region and this adaptive filter is activated when there is an active source in the kth spatial region. The same model applied to other microphones. With M microphones defining K spatial regions, where $M > K$, M adaptive filters are attached to each spatial region. Hence, a total of $M \times K$ adaptive filters are required. However, most conversation takes place with one active talker at a time in the transmission room. Hence in most cases, only one spatial region and thus M modeling filters (attached to that spatial region) are activated at a single time. The computational load is therefore greatly reduced. Furthermore, with single active spatial region, only one channel (out of K spatial region channel) needs to be transmitted to the receiving room. Hence, narrower transmission bandwidth is required compared to the conventional configuration.

3. REALIZATION OF THE PROPOSED CONFIGURATION

In the previous section, an ideal configuration for multi spatial region acoustic echo cancellation has been proposed. The difficult part in realizing the proposed configuration is the implementation of the *source and location estimation* block, which defines the spatial regions, as shown in Figure 1. Blind signal separation (BSS) with DOA estimation, as well as beamforming techniques are possible candidates. In this section, a realization of the proposed configuration with multiple fixed beamformers is demonstrated.

As shown in Figure 3, three spatial regions ($K = 3$) are defined. These spatial regions are formed with fixed beamformers looking at the center of each spatial region, q_k, $k = 1,2,3$. Ideally, the spatial response of the kth beamformer should be given as:

$$b_k(f,q) = \begin{cases} 1, & q = q_k \pm \Delta q/2 \quad \forall f \\ 0, & \text{elsewhere} \end{cases} \qquad (4)$$

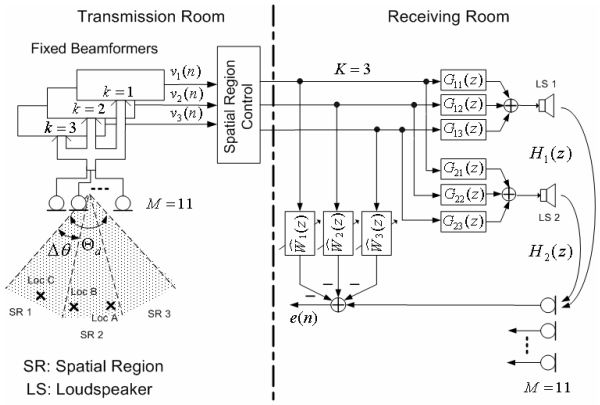

Figure 3: Realization of the proposed configuration with beamforming technique.

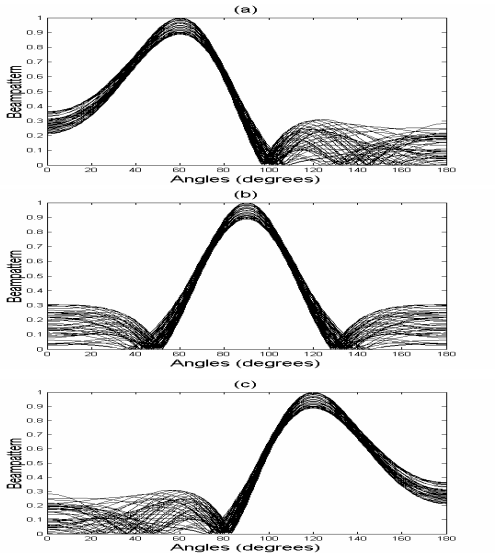

Figure 4: Beampattern for 44 frequency bins within the design bandwidth ranging from 300 to 3000 Hz: (a) $q_1 = 60^\circ$, (b) $q_2 = 90^\circ$ and (c) $q_3 = 120^\circ$.

The implementation will be straight forward if this ideal spatial response is obtainable. However, this ideal response with constant gain over the pass band and infinite attenuation in the stop band is not achievable in practical with a discrete and finite aperture.

In a realistic case where the array aperture is constrained by the enclosure's dimensions, a *sinc* like spatial response can be achieved as shown in Figure 4. Clearly, with these spatial responses, signal originated from a source located in one of the spatial regions will present in all beams. This is undesirable as explained in Section 2. To cope with this problem, a *spatial region control* block is added as shown in Figure 3. It performs beam selection base on the signal level and only the beam with highest level is activated. The rationale is that beam defining the spatial region where the active source located will have a strongest output compared to other beams. The beam selection algorithm is as outlined in [5] with a slight modification. Beam selection is made in every 120 ms interval to cope with multi-talker situation.

3.1. Wideband beamforming

As shown in Figure 3, direction of interest is $\Theta_d = 90^\circ = \left[45^\circ, 135^\circ\right)$. With $K = 3$ spatial regions, range of each spatial region is $\Delta q = 30^\circ$. These spatial regions are defined with three wideband beamformers looking at the center of each spatial region, $q_1 = 60^\circ$, $q_2 = 90^\circ$ and $q_3 = 120^\circ$. Wideband beamformers are needed due to the broadband nature of audio signal in teleconferencing system. Constant directivity beamformers for wideband signals have been discussed extensively, for example in [6, 7].

Following the indications in [6], a linear array is designed to provide frequency invariant spatial response covering a bandwidth of 300-3000 Hz with an aperture size of $Q = 2$ wavelength. With logarithmically spaced array geometry, $M = 11$ elements are needed. Using an FFT size of 128 resulted in 44 bins within the design band, with each bin having a width of 62.5 Hz. An 11 taps reference FIR filter with low pass characteristic is used in generating the complex weights for each frequency bin. Figure 4 shows beampattern for all the 44 frequency bins within design band for $q_1 = 60^\circ$, $q_2 = 90^\circ$, and $q_3 = 120^\circ$ respectively. Clearly, main lobes of these beamformers are overlapping.

4. SIMULATION RESULTS

For the proposed configuration, there are $M = 11$ microphones and 2 loudspeakers in the transmission and receiving rooms. With $K = 3$, there are three spatial regions channels $v_k(n)$ as shown in Figure 3. Equivalently, for a 3-channel system, the conventional configuration consists of $M = 3$ microphones and three loudspeakers in both rooms.

Transmission and receiving rooms are modeled as $5 \times 5 \times 3 = 75 m^3$ rectangular enclosures. The reflection coefficients of the six inner surfaces are chosen so that the reverberation time is about 300 *ms*. Outputs of the microphones are generated by convolving a Gaussian noise with M impulse responses. Three sets of impulse responses are derived for different source locations at *Loc A*, *Loc B* and *Loc C* as shown in Figure 3. These impulse responses are generated with source image method [8]. Length of these impulse responses is 2048 taps. In the receiving room, the acoustic echo paths to identify are modeled with 2048 taps impulse responses generated with the same setting as the transmission room model. The adaptive filters to model these acoustic echo paths are of 768 taps and are adapted with NLMS algorithm. The overall sampling frequency used in the simulations is 8 kHz.

In the simulations, location of the active source changes at $t = 10\ s$ and $t = 20\ s$. In the first simulation, changes of location occur within SR 2 (SR denotes spatial region). As discussed earlier, conventional configuration suffers from nonuniqueness problem. Consequently, its adaptive filters need to reconverge

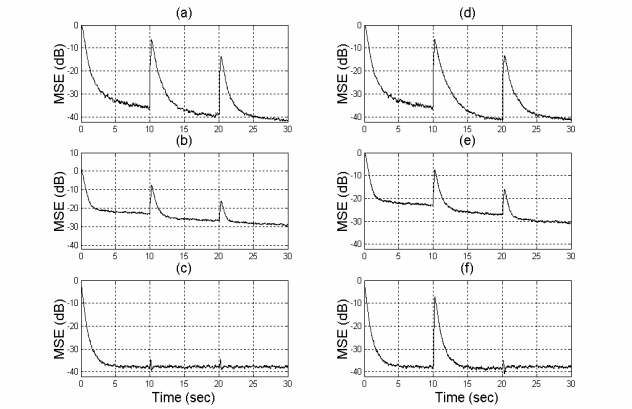

Figure 5: Behavior of the MSE (mean square error) when location of an active source changes at $t = 10\ s$ and $t = 20\ s$ for conventional configuration without nonlinear preprocessing (a, d), conventional configuration with nonlinear preprocessing (b, e), and the proposed configuration (c, f).

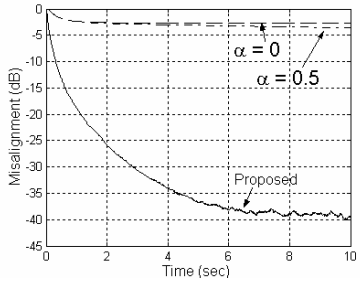

Figure 6: Misalignment plots for the conventional ($a = 0, a = 0.5$) and the proposed configurations.

whenever there are any changes of acoustic path in the transmission room. This can be seen from a severe increase of mean square error (MSE) at $t = 10\ s$ and $t = 20\ s$, where source position changes instantaneously from *Loc A* to *Loc B*, and from *Loc B* back to *Loc A* respectively, as shown in Figure 5(a). This problem can be mitigated by adding nonlinear distortion (half-wave rectifier, level of nonlinearity $a = 0.5$), as shown in Figure 5(b). On the other hand, for the proposed configuration, nonuniqueness problem does not occur. This can be seen from Figure 5(c) where the MSE is unaffected by source position changes in the transmission room.

In the second simulation, location of the active source changes from SR 2 (*Loc A*) to SR 1 (*Loc C*), and from SR 1 (*Loc C*) back to SR 2 (*Loc A*) at $t = 10\ s$ and $t = 20\ s$ respectively. Increase of MSE can be observed for the conventional configuration with and without NL preprocessing as shown in Figure 5(e) and (d) respectively. For the proposed configuration, MSE plot is as shown in Figure 5(f). In the first $10\ s$, SR 2 is active and the adaptive filter attached to SR 2 is converging. In the following $10\ s$, SR 1 is active instead, and its adaptive filter starts to converge at $t = 10\ s$ causing an increase in MSE. This should not be considered as the effect of nonuniqueness problem because the adaptive filters have indeed converged to their desired echo path solutions during the first and second $10\ s$ intervals. This can be seen at $t = 20\ s$ where MSE is unaffected when location of the active source changes instantaneously from SR 1 (*Loc C*) back to SR 2 (*Loc A*).

Figure 6 shows (from uppermost to lowest lines) the misalignment for the conventional configuration without NL processing ($a = 0$), with NL processing ($a = 0.5$) and the proposed configuration. For the conventional configuration, due to high coherency between channels, NLMS algorithm converges slowly to the echo path solution. Whereas, for the proposed configuration, it is clear that, NLMS is sufficiently fast to drive the modeling filters to its echo path solution with a misalignment of $-38\ dB$ after $6\ s$.

5. CONCLUSIONS

A novel configuration for multichannel teleconferencing has been proposed. In the proposed configuration, the transmission and receiving rooms are divided into equivalent number of spatial regions. Simulation results confirm that, with this multi spatial region configuration the AEC does not suffers from nonuniqueness problem and low complexity NLMS algorithm is sufficient to adapt the modeling filters. Furthermore, compared to the conventional configuration, the proposed configuration requires narrower transmission bandwidth. Finally, we have also shown that the proposed configuration can be realized using beamforming technique.

6. REFERENCES

[1] M. M. Sondhi and D. R. Morgan, "Stereophonic acoustic echo cancellation – an overview of the fundamental problem," *IEEE Signal Processing Letters*, vol.2, pp.148-151, Aug. 1995.

[2] T. Gänsler and J. Benesty, "Stereophonic acoustic echo cancellation and two-channel adaptive filtering: an overview," *International Journal of Adaptive Control and Signal Processing*, vol. 14, pp. 565-586, Sept. 2000.

[3] T. Gänsler and J. Benesty, "Multichannel acoustic echo cancellation: what's new?" in *Proc. IWAENC*, 2001.

[4] H. Buchner, S. Spors, W. Kellermann and R. Robenstein, "Full-duplex communication systems using loudspeaker arrays and microphone arrays," in *Proc. ICASSP*, pp. 509-512, 2002.

[5] P. L. Chu, "Desktop mic array for teleconferencing," in *Proc. ICASSP*, pp. 2999-3002, May 1995.

[6] D. B. Ward, R. A. Kennedy and R. C. Williamson, "Constant directivity beamforming," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. B. Ward, Eds., Berlin, Germany: Springer-Verlag, 2001.

[7] T. Chou, "Frequency-independent beamformer with low response error," in *Proc. ICASSP*, pp. 2995-2998, May 1995.

[8] J. Garas, *Room Impulse Response v2.5 software and documentation*, 2002. http://www.dspalgorithms.com.