AN ACOUSTIC ECHO CANCELLATION STRUCTURE FOR SYNTHETIC SURROUND SOUND

Trevor Yensen and Rafik Goubran

Systems and Computer Engineering, Carleton University, Ottawa Ontario, K1S 5B6, Canada

ABSTRACT

This paper proposes an acoustic echo cancellation structure for hands-free synthetic surround sound applications, such as multiple participant conferencing, and virtual reality applications. Voice over Internet protocol (VoIP) and other virtual reality applications can benefit from the addition of 3D spatial audio generated by more than two loudspeakers. When full-duplex audio is present in a system, however, acoustic echo cancellation is required to eliminate the feedback echo path. The acoustic echo cancellation structure proposed by this paper is based on the acoustic echo canceller per spatial region allocation scheme previously introduced by the authors for two channel synthetic stereo. This paper will show that the spatial region allocation scheme is extensible to any number of channels which makes it extremely versatile and flexible, especially for surround sound audio. Microsoft Direct X 7, a commonly used application programmer interface (API), was used in our simulations to generate the 3D spatial audio on a PC.

1. INTRODUCTION

Surround sound spatial imaging is important in virtual reality applications, such as multiple participant voice conferencing over Internet protocol (IP) networks, to enhance the user's ability to distinguish between remote participants or to provide spatial cues for virtual reality. In VoIP applications, it is not always easy to distinguish between participants based on their voice alone, especially if the listener is not familiar with the talker's voice. Surround sound for virtual reality enhances the experience by adding spatial cues, which complement the graphical user environment.

Stereo loudspeakers are able to reproduce limited spatial imaging for multiple participant conferencing. Virtual sources, far-end talkers, can only be convincingly positioned in 3D acoustic space over two loudspeakers through the use of carefully designed head-related transfer functions, HRTF, and loudspeaker decorrelation functions. 3D positioning using this methodology requires complicated head-tracking mechanisms to ensure that the effect does not deteriorate with listener head movement. The addition of surround channels, loudspeakers behind the listener, however, can be used to create realistic spatial audio with a larger sweet spot than two channel methods and is the subject of this paper. A surround sound acoustic echo canceler must deal with the addition of these extra surround channels and the single canceler per spatial region acoustic echo cancellation structure, first proposed by the authors in [1], is well suited for this task.

To facilitate distinguishing among participants in a multiple participant conference, the chosen spatialization method must distribute the participant's audio signals in a manner designed to minimize localization blur and must not sound awkward to the listener. (Localization blur is minimized by ensuring that virtual sources are well separated from one another.)

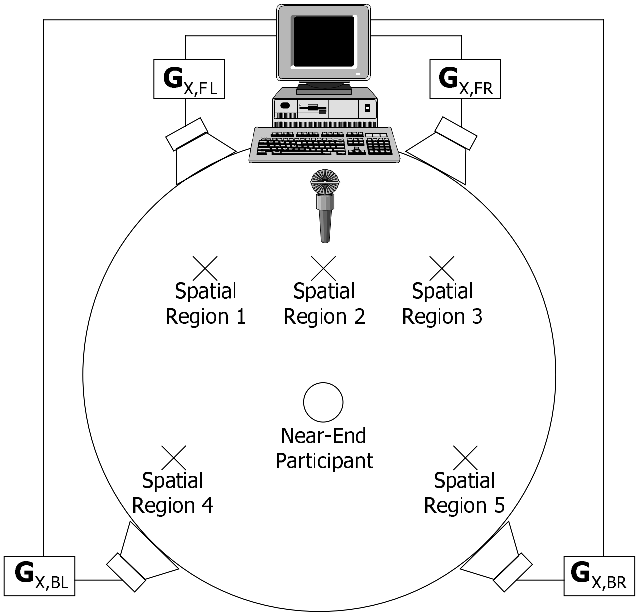

Fig. 1. Near-end talker spatialization

Fig. 1 shows five virtual source positions which can be easily generated by the Microsoft DirectX API.

This paper will explain how the acoustic echo canceller for a surround sound system with three or more loudspeakers operates under a spatial region echo canceller allocation.

2. ECHO CANCELER MISCONVERGENCE

A true stereophonic structure has two microphones and two loudspeakers. When two microphones are used in a true stereophonic system the reference signals are cross-correlated [2] which leads to echo canceller misconvergence.

The synthetic stereo and synthetic surround structures have a single microphone and two or more loudspeakers. The far-end synthetic stereo or synthetic surround sound audio signal is generated from the monaural signal using spatialization algorithms appropriate to the system. A recently proposed synthetic stereo structure which allocates an echo canceller structure to each loudspeaker channel [3] is referred to as the synthetic stereo with a single canceler per channel structure. The single canceler per channel structure, however, also suffers from cross-correlation of the reference signals which leads to echo canceller misconvergence. The authors propose adding a small non-linearity to the left and right channels to achieve convergence

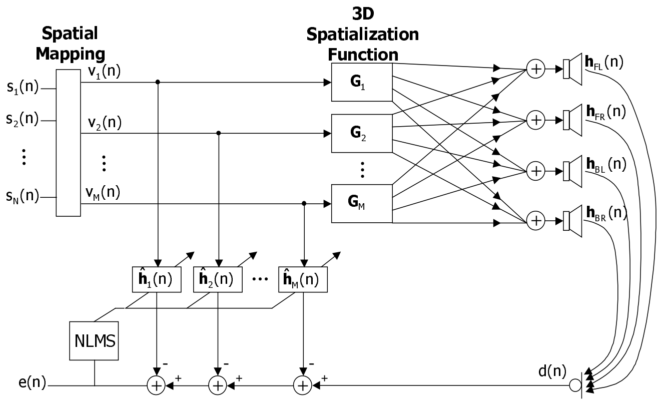

Fig. 2. Synthetic surround acoustic echo cancellation structure with beamformer

and using the fast affine projection (FAP) algorithm or two-channel recursive least squares (RLS).

If the single canceler per channel structure is extended to more than two channels there is an increasing amount of cross-correlation in the reference signal which makes convergence increasingly difficult as the number of channels increases. The number of acoustic echo cancellers required for the per channel structure is equal to the number of channels in the system. This does not apply to the proposed per spatial region structure which has an acoustic echo canceller for each spatial region.

3. SURROUND SOUND ACOUSTIC ECHO CANCELLATION

3.1. Description

This paper proposes performing synthetic surround sound acoustic echo cancellation using the single canceler per spatial region structure. This structure first takes the far-end participant's signal, $s_j(n)$, and maps it into the appropriate spot in the near-end listener's acoustic space, $v_i(n)$, see Fig. 2. The position of the virtual source will likely correspond to far-end participant's location on the video display terminal for VoIP conferencing or their relative 3D position in a virtual reality system. If more than one far-end participant is mapped into the same spatial region, then the signal $v_i(n)$, is a combination of all the far-end participants allocated to the spatial region. The number of far-end participants, N, therefore, does not determine the number of spatial regions, M, since there may be more than one participant assigned to the same spatial region.

The acoustic echo canceller then takes the mapped signal, $v_i(n)$, as its reference before spatialization. This eliminates the fundamental problem of acoustic echo canceller misconvergence [1][2].

The mapped signals are then steered into the appropriate spot in the near-end participant's 3D acoustic space by convolving the mapped signal with the appropriate set of spatialization functions, $\mathbf{h}_{\mathrm{FL}}(n)$, $\mathbf{h}_{\mathrm{FR}}(n)$, $\mathbf{h}_{\mathrm{BL}}(n)$ and $\mathbf{h}_{\mathrm{BR}}(n)$. There are unique spatialization functions associated with each spatial region for each loudspeaker. When four loudspeakers are in use the spatialization function set consists of four impulse responses per spatial region. For example, if two spatial regions are in use then a total of eight impulse responses is required.

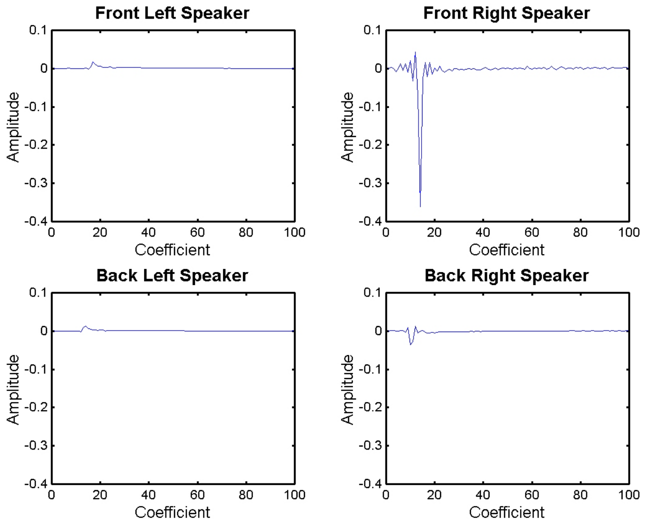

Fig. 3. Front right position loudspeaker spatialization impulse responses

When the synthetic surround sound acoustic echo cancellation structure is being used for VoIP conferencing the position of the 3D virtual source does not change throughout the duration of the conference. Not allowing a virtual source to change positions during a conference makes it easier for the listener to distinguish between participants. When the synthetic surround sound structure is used for virtual reality systems, the virtual sources can move in the environment, changing the loudspeaker spatialization function set. The acoustic echo canceller is able to track the changes in the spatialization functions and the impulse response of the reception room.

Each spatial region is generated by implementing a set of spatialization functions for each loudspeaker. The virtual sources used in this paper were created using the Microsoft DirectX 7 API. Fig. 3 shows the loudspeaker spatialization impulse responses used to create the impression that the talker appears 45 degrees to the front-right, spatial region 3 in Fig. 1. The DirectX API allows the user to choose several options for positioning a virtual source in acoustic space. The spatial audio is generated in 3D buffers with parameters relative to a 3D listener.

3.2. Analysis

The synthetic surround acoustic echo cancellation structure relies on a separate acoustic echo canceller for each spatial region. If there are M spatial regions present then there must be M acoustic echo cancellers, $\hat{\mathbf{h}}_1(n)$, $\hat{\mathbf{h}}_2(n)$, $\cdots$ $\hat{\mathbf{h}}_M(n)$. This assignment is scalable since there are only a limited number of spatial positions that the listener can distinguish between due to localization blur. There establishes a practical limit of the number of acoustic echo cancellers present in the system.

The number of echo canceller coefficients and taps in the system is dependent only on the number of spatial regions being used. The number of coefficients and taps is not dependent on the number of channels being used in the surround sound system. If five spatial regions are being used, M=5, then a total of $5T_{\mathrm{F}}$

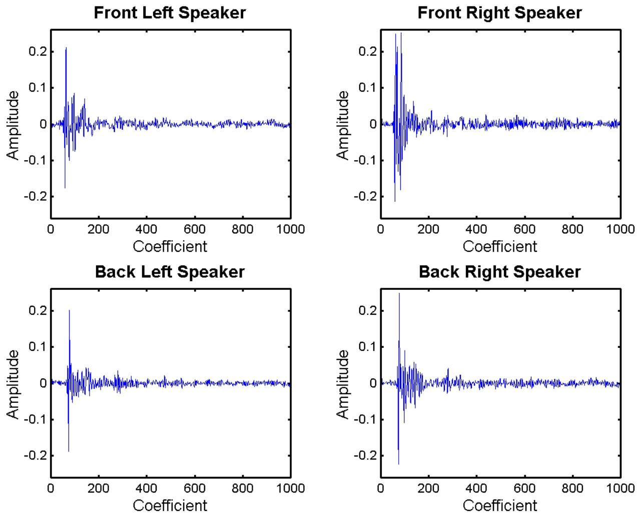

Fig. 4. Reception room impulse responses

coefficients and taps will be required regardless of the number of surround sound channels, where T_F is the length of the adaptive echo path impulse response.

The N far-end talkers in an $N+1$ participant conference or virtual reality system are denoted by $s_j(n)$:

$$s_j(n), \qquad j = 1, 2, \ldots, N \qquad (1)$$

These far-end participants are mapped into one of M near-end spatial regions, to form the acoustic echo canceller reference for spatial region *i*,

$$\mathbf{v}_i(n) = [v_i(n) \quad v_i(n-1) \quad \cdots \quad v_i(n-T_F+1)]^T, \; i = 1, \ldots, M \quad (2)$$

where $v_i(n)$ is the sum of all the far-end signals mapped to that spatial region.

The adaptive echo path model for spatial region, *i*, is denoted by, $\hat{\mathbf{h}}_i(n)$:

$$\hat{\mathbf{h}}_i(n) = [\hat{h}_{i,0}(n) \quad \hat{h}_{i,1}(n) \quad \cdots \quad \hat{h}_{i,T_F-1}(n)]^T, \quad i = 1, \ldots, M \quad (3)$$

If four surround sound channels are being used in the system then the spatial regions are realized through the front left, front right, back left and back right spatialization functions, $\mathbf{g}_{FLi}$, $\mathbf{g}_{FRi}$, $\mathbf{g}_{BLi}$ and $\mathbf{g}_{BRi}$. For example, the spatialization function for the front left channel is

$$\mathbf{g}_{FLi} = [g_{FLi,0} \quad g_{FLi,1} \quad \cdots \quad g_{FLi,T_S-1} \quad \mathbf{0}_{T_F-T_S}]^T, \; i = 1, \ldots, M \, (4)$$

where T_S is the length of the spatialization functions and $\mathbf{0}_{T_F-T_S}$ is a zero vector of length $T_F - T_S$.

The loudspeaker outputs, $\mathbf{x}_{FL}(n)$, $\mathbf{x}_{FR}(n)$, $\mathbf{x}_{BL}(n)$ and $\mathbf{x}_{BR}(n)$ are formed by convolving the mapped far-end signals, $v_i(n)$, by the appropriate spatialization function for the channel and spatial region. For example the front left loudspeaker output, $\mathbf{x}_{FL}(n)$, is formed by convolving it with the front left spatialization function, $\mathbf{g}_{FLi}$, for the appropriate spatial region *i*

$$\mathbf{x}_{FL}(n) = [x_{FL}(n) \quad x_{FL}(n-1) \quad \cdots \quad x_{FL}(n-T_R+1)]^T \quad (5)$$

$$x_{FL}(n) = \sum_{i=1}^{M} \mathbf{v}_i^T(n)\mathbf{g}_{FLi} \qquad (6)$$

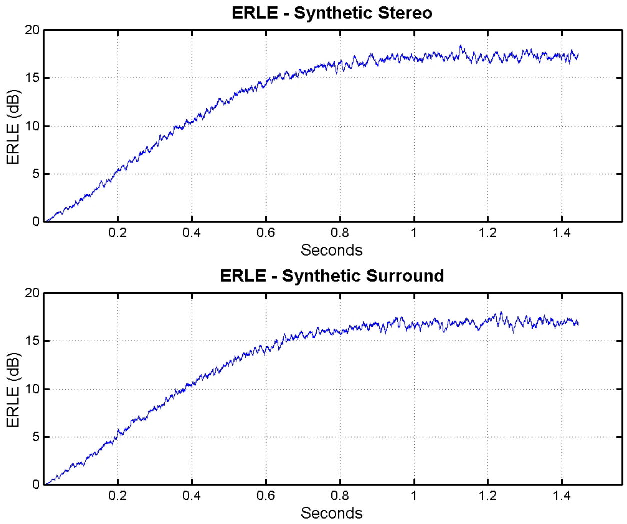

Fig. 5. Synthetic stereo versus synthetic surround

where T_R is the length of the reception room impulse response. The loudspeaker outputs for the remaining channels are formed in a similar fashion.

For four channel surround sound the primary signal, $d(n)$, received by the microphone is

$$d(n) = \mathbf{x}_{FL}^T(n)\mathbf{h}_{FL}(n) + \mathbf{x}_{FR}^T(n)\mathbf{h}_{FR}(n) + \mathbf{x}_{BL}^T(n)\mathbf{h}_{BL}(n) + \mathbf{x}_{BR}^T(n)\mathbf{h}_{BR}(n) \qquad (7)$$

where the reception room impulse responses are of length T_R and denoted by $\mathbf{h}_{FL}(n)$, $\mathbf{h}_{FR}(n)$, $\mathbf{h}_{BL}(n)$ and $\mathbf{h}_{BR}(n)$. For example, the front left reception room impulse response is given by:

$$\mathbf{h}_{FL}(n) = [h_{FL}(n) \quad h_{FL}(n-1) \quad \cdots \quad h_{FL}(n-T_R+1)]^T \qquad (8)$$

The adaptive echo path model for spatial region, *i*, is therefore a sum of the convolutions of the loudspeaker spatialization function for the given spatial region and the respective reception room impulse response:

$$\hat{\mathbf{h}}_i(\mathrm{n}) = \mathbf{g}_{FLi} * \mathbf{h}_{FL}(n) + \mathbf{g}_{FRi} * \mathbf{h}_{FRi}(n) + \mathbf{g}_{BLi} * \mathbf{h}_{BL}(n) + \mathbf{g}_{BRi} * \mathbf{h}_{BRi}(n) \qquad (9)$$

where * is the symbol used for convolution.

3.3. Comparing Single Channel to Synthetic Multi-Channel Acoustic Echo Cancellation

When the single canceler per spatial region structure is used in a monaural, stereo or surround sound scenario with only one active spatial region at a time, the acoustic echo canceller performs identically with only a change in the echo path model, $\hat{\mathbf{h}}(n)$. When the proposed structure is used with multiple channels the echo path model changes from including only the reception room impulse response, in the monaural case with no spatialization, to having spatialization information included.

For single channel applications the acoustic echo canceller simply converges to the reception room impulse response:

$$\hat{\mathbf{h}}(n) = \mathbf{h}(n) \qquad (10)$$

For synthetic stereo applications with the single canceler per spatial region structure the acoustic echo canceller converges to the sum of the left loudspeaker spatialization function for region *i*

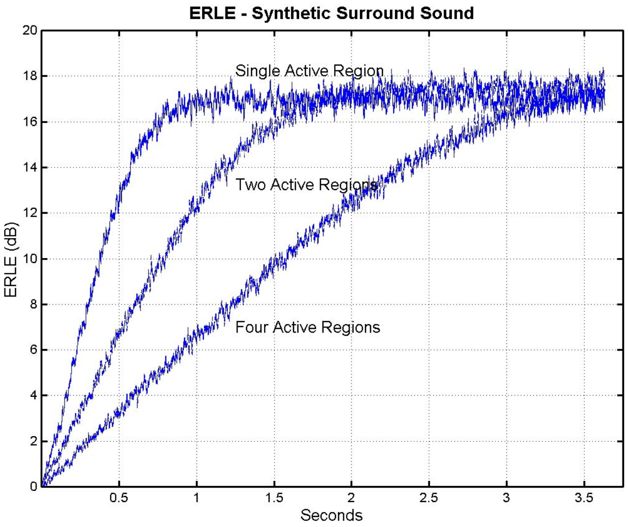

Fig. 6. Multiple active spatial regions

convolved with the left loudspeaker reception room impulse response and the right loudspeaker spatialization function for region *i* convolved with the right loudspeaker reception room impulse response:

$$\hat{\mathbf{h}}_i(n) = \mathbf{g}_{\mathrm{L}i}(n) * \mathbf{h}_{\mathrm{L}}(n) + \mathbf{g}_{\mathrm{R}i}(n) * \mathbf{h}_{\mathrm{R}}(n) \qquad (11)$$

For synthetic surround applications with the single canceler per spatial region structure and four loudspeakers, the acoustic echo canceller converges to the sum of all four loudspeaker spatialization functions for region *i* convolved with their respective reception room impulse responses (9).

It is easy to see that acoustic echo canceler allocation by spatial region works well for any number of loudspeakers and the convergence rate and tracking performance is equivalent as long as only one spatial region is allowed to be active at one time, see section 4.

4. SIMULATION

The simulations performed for this paper required measuring the reception room impulse responses for a synthetic stereo configuration with two loudspeakers, $\mathbf{h}_{\mathrm{L}}(n)$ and $\mathbf{h}_{\mathrm{R}}(n)$, and a synthetic surround configuration with four loudspeakers, $\mathbf{h}_{\mathrm{FL}}(n)$, $\mathbf{h}_{\mathrm{FR}}(n)$, $\mathbf{h}_{\mathrm{BL}}(n)$ and $\mathbf{h}_{\mathrm{BR}}(n)$, see Fig. 4. Spatialization functions for two channel stereo, $\mathbf{g}_{\mathrm{L}i}$ and $\mathbf{g}_{\mathrm{R}i}$, and four channel surround sound, $\mathbf{g}_{\mathrm{FL}i}$, $\mathbf{g}_{\mathrm{FR}i}$, $\mathbf{g}_{\mathrm{BL}i}$ and $\mathbf{g}_{\mathrm{BR}i}$, were generated using the Microsoft DirectX API. Fig. 3 shows the spatialization impulse responses for the front right spatial region generated by the DirectX API.

Fig. 5 compares the convergence rate for a synthetic stereo and a four channel synthetic surround sound acoustic echo canceller configuration with only one active spatial region at a time. The input signal to the single spatial region, $\nu_1(\mathrm{n})$, was independent Gaussian white noise. It can be seen that both the convergence rate and the maximum echo return loss enhancement (ERLE) [5] are equivalent for each configuration. This simulation can be extended to the single channel case with similar results. The simulation results for Fig. 6 are valid for 25000 iterations at a sampling rate of 16kHz using the normalized least mean square (NLMS) algorithm with a step-size of 0.5.

If more than one spatial region is allowed to be active at a time then convergence is still possible but the rate of convergence decreases. Fig. 6 shows the synthetic surround sound acoustic echo canceller configuration with one active spatial region, two active spatial regions and four active spatial regions. This simulation shows that the rate of convergence decreases when more than echo canceller is allowed to be active at any one time, but the maximum ERLE value is still attainable. To maximize convergence rate and tracking it may be desirable to only allow one acoustic echo canceler to converge at a time based on the spatial region with the largest amount of energy although convergence is possible for multiple active spatial regions. The simulation results for Fig. 6 are valid for 60000 iterations at a sampling rate of 16kHz using the normalized least mean square (NLMS) algorithm with a step-size of 0.5 with independent Gaussian white noise input signal, $\nu_1(\mathrm{n})$.

5. CONCLUSION

This paper proposes a new structure for supporting acoustic echo cancellation with synthetic surround sound for multiple participant VoIP conferencing and virtual reality systems. The structure proposed in this paper allocates a single acoustic echo canceler per spatial region, which eliminates the problem of reference signal correlation and echo canceler misconvergence. It was shown that the number of loudspeaker channels does not affect the convergence rate or tracking performance of this structure and that convergence is possible with multiple active spatial regions.

6. ACKNOWLEDGEMENTS

This work was partially funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) and Communications and Information Technology Ontario (CITO).

7. REFERENCES

[1] T. Yensen, R. Goubran and I. Lambadaris, “A multiple participant CTI acoustic echo cancellation structure,” in *Proc. IEEE Nordic Signal Processing Symp.*, 1998, pp. 193-196.

[2] M. Sondhi, D. Morgan and J. Hall, “Stereophonic acoustic echo cancellation – An overview of the fundamental problem,” *IEEE Signal Processing Lett.*, Vol. 2, No. 8, pp. 148-151, Aug. 1995.

[3] J. Benesty, D. Morgan, J. Hall and M. Sondhi, “Synthesized stereo combined with acoustic echo cancellation for desktop conferencing,” in *Proc.IEEE Int’l Conf. Acoustics, Speech and Signal Processing*, 1999, pp. 853-856.

[4] R. Dunlop, D. Shepherd, M. Martin, et al, Sams Teach Yourself DirectX 7 in 24 Hours, Indianapolis, IN: Sams Publishing, 1999.

[5] M. Knappe and R. Goubran, “Steady state performance limitations of full-band acoustic echo cancellers,” in *Proc. IEEE Int’l Conf. Acoustics Speech and Signal Processing*, 1994, pp. 2073-2077.