A STEREO ECHO CANCELLER WITH SIMULTANEOUS 2-CHANNEL INPUT SLIDES FOR FAST CONVERGENCE AND GOOD SOUND LOCALIZATION

Akihiko Sugiyama, Yuusuke Mizuno[†], *Laurent Kazdaghli*[‡], *Akihiro Hirano*[†], *and Kenji Nakayama*[†]

NEC Common Platform Software Research Laboratories, Kawasaki 211-8666, JAPAN
†Graduate School of Natural Science and Technology, Kanazawa University, Kanazawa 920-1192, JAPAN
‡Eurecom Institute, F-06904 Sophia-Antipolis, FRANCE

ABSTRACT

This paper proposes a stereo echo canceller with simultaneous 2-channel input slides. Periodic delays with different phases in two channels provide an additional condition for echo cancellation, leading to faster convergence. Signal sliding in both channels guarantees symmetrical shifts of a sound image around the original position for better subjective quality of the far-end signal. Simulation results show that the convergence speed of the proposed echo canceller is 20% *faster than that of the conventional echo canceller. Subjective evaluation results demonstrate that the proposed echo canceller achieved a* 0.37 *higher grading difference in the ITU-R five-grade impairment scale than the conventional one with a statistically significant difference.*

1. INTRODUCTION

One of the most successful applications of acoustic echo cancellers is remote conferencing. With growing demands for enhanced presence in conference, use of multiple loudspeakers for presentation of the far-end speech has been spotlighted in these years. With multiple participants at different locations, multichannel presentation is helpful for speaker localization, resulting in easier speaker discrimination [1]. For multichannel conference systems, multichannel acoustic echo cancellers are needed.

In order to model all the echo paths between loudspeakers and microphones, a stereo echo canceller based on linear combination was proposed [2]. Its drawback is a nonuniqueness problem. This structure cannot identify the correct echo paths for multichannel signals with high interchannel correlation [3, 4]. The identified echo path characteristic is dependent on the acoustic characteristics on the remote side. Therefore, a change in the acoustic characteristics on the remote side, such as a talker change or movement, leads to increase in the residual echo [4].

A solution to this problem is a stereo echo canceller based on an input sliding technique that periodically delays one of the input signals [5, 6]. Input sliding theoretically identifies the true echo paths and provides faster convergence than nonlinear processing [7]. However, for signals with strong interchannel correlation, it does not provide sufficiently fast convergence. In addition, periodic delay insertion in only one channel shifts the sound image to the right or the left, making the sound-image sift asymmmetrical and more noticeable.

This paper proposes a stereo echo canceller with simultaneous 2-channel input slides for fast convergence and good sound localization. The following section reviews input sliding in one channel. The new method, simultaneous input sliding in two channels, is presented in Section 3. Finally, in Section 4, convergence speed and subjective quality of the far-end signal with input sliding are evaluated.

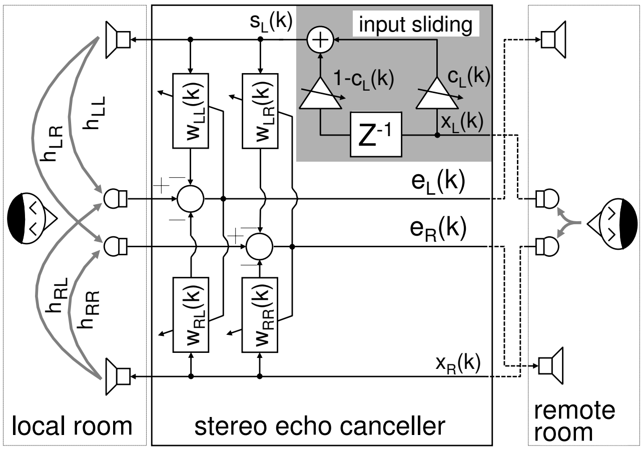

Figure 1: Stereo echo canceller with input sliding in one channel.

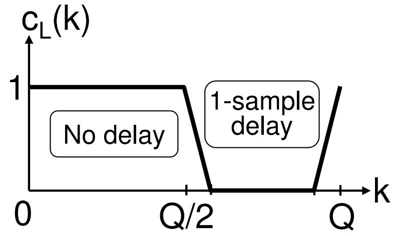

Figure 2: Coefficient $c_L(k)$.

2. INPUT SLIDING IN ONE CHANNEL

Fig. 1 depicts a block diagram of the conventional stereo echo canceller. In addition to the linearly combined echo canceller section that consists of four adaptive filters corresponding to four echo paths, there is an input sliding unit for periodically delaying the left-channel signal. Of course, it is possible to delay the right-channel signal instead of the left. Input sliding unit is realized as a 2-tap FIR filter whose coefficients change the values with a period of Q. Figure 2 illustrates one of the coefficients, $c_L(k)$. The other coefficient $1-c_L(k)$ changes accordingly.

When $c_L(k)=1$, the input sliding unit in Fig. 1 becomes delay-free, allowing the input signal to go through with no change. On the other hand, $c_L(k)=0$ delays the input signal by one sample for the output of the input sliding unit. To avoid audible artifacts, $c_L(k)$ is smoothly changed from 0 to

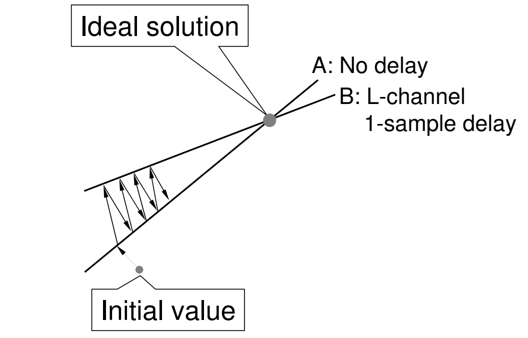

Figure 3: Intuitive illustration of ideal-solution search.

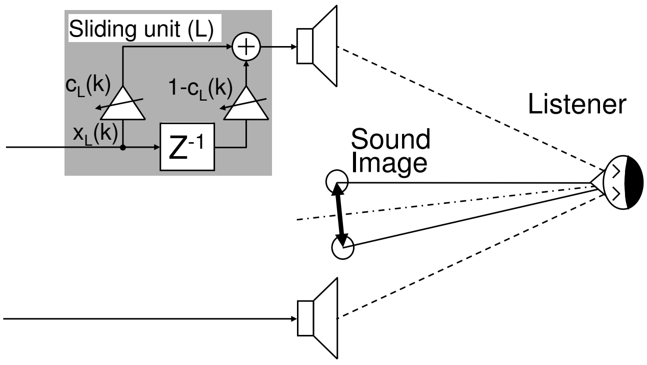

Figure 4: Shift of sound image (1-channel input sliding).

1 or 1 to 0. The input sliding provides two different conditions for echo cancellation with and without the delay. The common solution to the two conditions is the correct impulse response of the echo path that is to be identified.

An intuitive illustration of ideal-solution search by the conventional stereo echo canceller is shown in Fig. 3. Two lines represent two sets of indefinite solutions with and without the delay. Their intersection is the ideal solution. When adaptation by NLMS algorithm starts from an initial condition, the coefficients perpendicularly approach the linear line representing a set of indefinite solutions [8]. They move toward Line A when there is no delay and Line B when there is a delay. Adaptive filter coefficients develop and approach the ideal solution by alternating these two movements.

Input sliding generates shifts of sound image, which were not encountered in the original linear combination structure. It does not move the sound image from its original position. However, input sliding in one channel periodically delays the input signal, leading to a shift of the sound image. Figure 4 shows such a shift of the sound image. Because delay is applied in the left channel in Fig. 4, the right-channel signal arrives at the ear earlier than the left-channel signal. This fact makes a sound image shift toward the left. When there are periodic changes in the input signals with and without a delay, the sound image goes back and forth between the original and the left-shifted positions. The sound image therefore moves around the center between the original and the left-shifted positions as in the dash-dot line in Fig. 4, and becomes unfocused. This is a noticeable artifact.

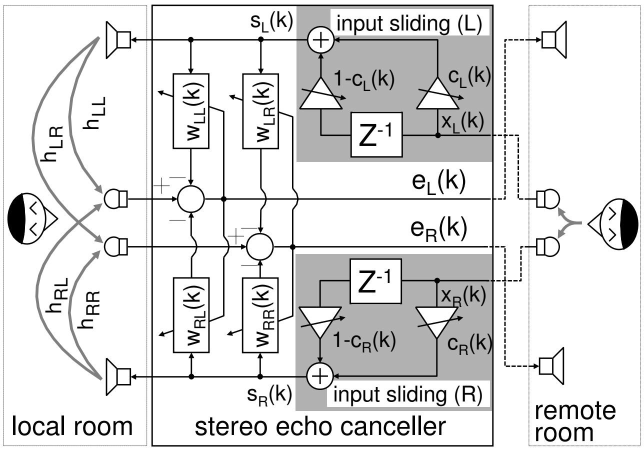

Figure 5: Stereo echo canceller with input sliding in two channels.

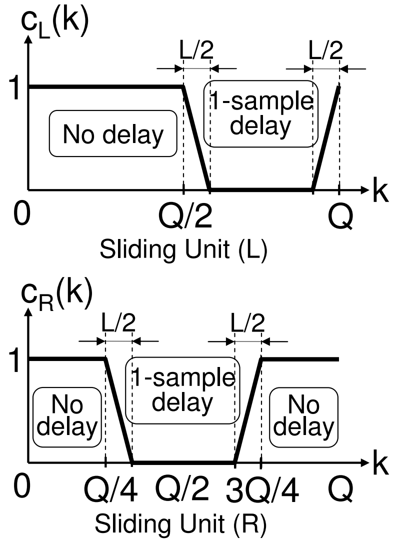

Figure 6: Coefficients $c_L(k)$ and $c_R(k)$.

3. INPUT SLIDING IN TWO CHANNELS

Input sliding in two channels, as shown in Fig. 5, has one sliding unit in both channels. Fig. 6 illustrates coefficients $c_L(k)$ and $c_R(k)$ that are out of phase. Although the left-channel coefficient, $c_L(k)$, is equal to that in Fig. 2, its right-channel counterpart has the same shape with a $Q/4$ advanced phase.

Fig. 6 indicates that there are four states, namely, no delay in both channels, 1-sample delay in the left channel, 1-sample delay in both channels, and 1-sample delay in the right channel. Considering that the first and the third states are equivalent, there are three different states corresponding to three different conditions. Input sliding in one channel provides two different states with and without a delay, leading to two different conditions. Input sliding in two channels achieves fast convergence by three different conditions for echo cancellation, which are generated by simultaneous input slides in two channels.

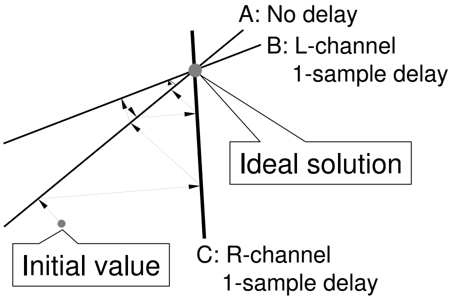

Figure 7: Intuitive illustration of searching for ideal solution.

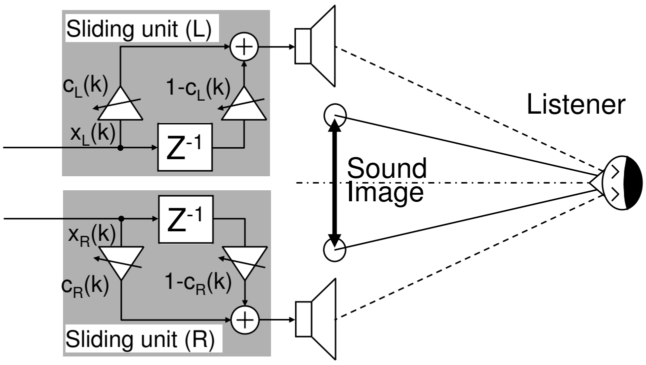

Figure 8: Shift of sound image (2-channel input sliding).

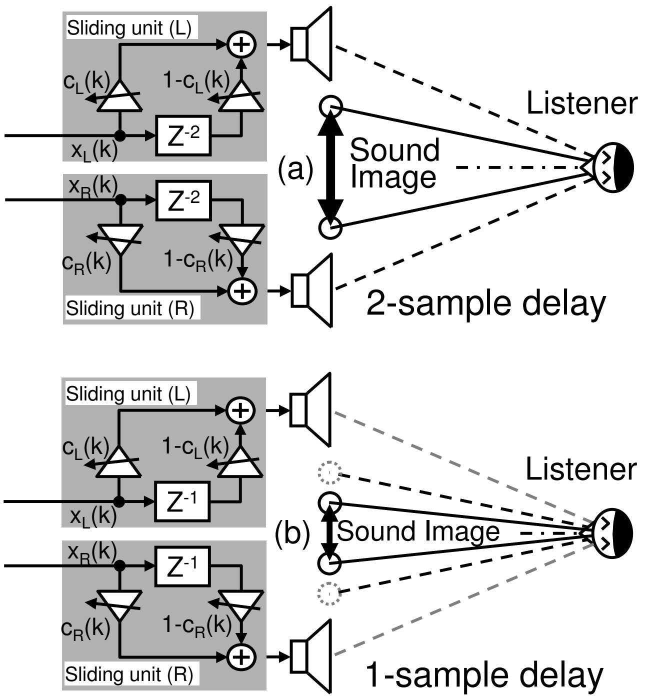

Figure 9: Delay and angle of sound-image shift.

Table 1: Parameters

Input Sliding	LC	LC-1IS1	LC-1IS2	LC-2IS
Number of taps	64 or 1000			
Stepsize	0.5			
Q	–	4000		
L	–	400		
Delay in samples	0	1	2	1
Ambient noise	$-40dB$			

Ideal-solution search in the case of input sliding in two channels is illustrated in Fig. 7. An additional channel with input sliding generates another set of indefinite solutions represented by a new linear line. The coefficients approach the correct solution by going back and forth among three lines corresponding to A: no delay, B: 1-sample delay in the left channel, and C: 1-sample delay in the right channel. Compared to Fig. 3, the coefficients converge to the correct solution in a shorter time in the case of Fig. 7.

Figure 8 shows sound-image shifts in case of two-channel input sliding. Because there are input sliding units in two channels, the sound-image moves toward both the left and the right. The sound image at the center moves around the center, making the artifacts less noticeable than input sliding in one channel. However, the angle of the sound-image shift is twice as wide as that in the other case. It means that the artifacts are more serious. For this reason, the amount of delay in the case of two-channel input sliding should be set to the half of the delay in the single-channel input sliding.

Fig. 9 depicts shift of the sound image originally located at the center in the case of two-channel input sliding. Sound image locations (a) and (b) correspond to 2- and 1-sample delays, respectively. Because the latter results in half the shift of the former, 2-sample delay in one channel and 1-sample delay in two channels have the same shift angle with different center positions.

4. EVALUATIONS

4.1 Convergence Speed

Linear Combination structure (LC), LC structure with 1-channel input sliding (LC-1IS), LC structure with 2-channel input sliding (LC-2IS) were evaluated from viewpoints of convergence characteristics and ERLE (echo return-loss enhancement) by simulations. Q and L are parameters for determining the period of input slides. By setting a large value, artifacts by input slides become less noticeable. Parameter values are summarized in Tab. 1. 10 speech signals (5 male and 5 female voice signals) sampled at 16kHz were used as the input signal. Coefficients of the adaptive filters were updated by the NLMS algorithm. As an ambient noise, a white noise was added to the echo with an echo-to-noise ratio (ENR) of -40 dB. LC-1IS was evaluated with two different values of the delay. LC-1IS1 and LC-1IS2 have 1- and 2-sample delays for one of the far-end signals, respectively. LC-1IS2 has a 2-sample delay, which has the same angle of sound-image shift as that of LC-2IS.

Norm of the coefficient-error vector (NCEV) was used as a measure for correct echo-path identification. NCEV is determined as an error norm between the echo-path impulse response and the filter coefficient vectors. NCEV is calculated by

$$NCEV(k) = 10\log_{10}\frac{||\mathbf{h}-\mathbf{w}(k)||^2}{||\mathbf{h}||^2} \tag{1}$$

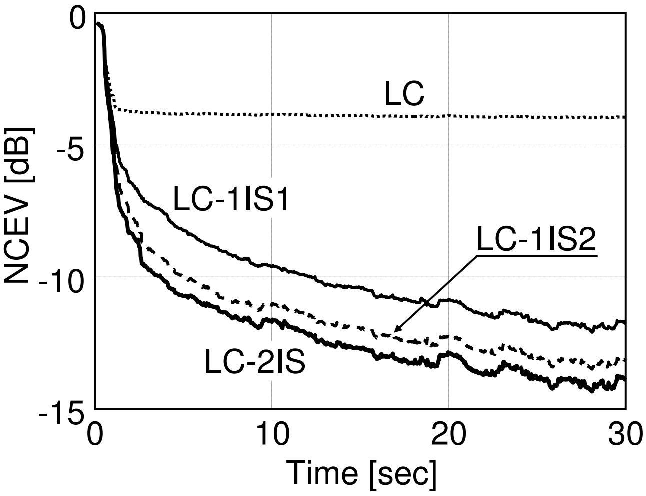

Figure 10: Normalized coefficient-error vector (NCEV) for 64 taps.

where $\mathbf{h}$ and $\mathbf{w}(k)$ are the echo path and the filter coefficient vectors, respectively.

NCEV for a 64-tap echo path is shown in Fig. 10. Each curve is an ensemble average of the results for the 10 speech signals. NCEV goes as low as -4.0dB and is saturated in the case of the LC structure. This indicates that the coefficients have not converged to the correct values. When there is a talker change, ERLE will be degraded. The coefficient vector of LC-1IS1 is heading for the true echo-path impulse response. NCEV is reduced to as low as -12dB in 30 sec. LC-1IS2 achieves even better convergence with an NCEV of -13.5dB in 30 sec. It confirms that a larger number of delay samples achieves faster convergence. Speed-up is due to lower interchannel correlation by a larger number of interchannel time lags. However, fast convergence of LC-1IS2 is obtained at the price of twice as wide an angle of sound-image shift as that for LC-1IS1. It should be noted that LC-2IS provides faster convergence with an NCEV of -14.3dB in 30 sec. with exactly the same angle of sound-image shift as that for LC-1IS1.

Shown in Fig. 11 is the NCEV for a 1000-tap echo path. Similar to the 64-tap case, NCEV by LC is saturated at -2.5dB, indicating a failure of correct echo-path identification. LC-1IS2 achieves -8.4dB NCEV in 30 sec. LC-2IS exhibits even faster convergence with a 30-sec NCEV of -9.0dB. When the time for NCEV convergence to -8.0dB is compared, LC-2IS needs 19 sec. with 20% speed-up over LC-1IS2 with 24 sec.

Evolution of ERLE for 1000 tap is depicted in Fig. 12. LC-1IS1 achieves 1dB lower ERLE than those for LC-1IS2 and LC-2IS after 10 sec. This is due to a smaller number of delay samples compared to the other two structures, leading to coefficients that are not sufficiently close to the true solution.

4.2 Far-End Signal Quality after Input Sliding

A subjective evaluation was performed to assess the far-end signal quality with simultaneous input sliding in two channels. The triple stimulus/hidden reference/double blind approach [9] was used in the evaluation. Subjects first listen to the reference, the hidden reference, and the far-end signal with slides. The order of presentation for the last two were randomly selected. They were then asked to score the latter two with respect to their difference from the reference according to the ITU-R 5-grade impairment scale [10] as shown in Tab. 2. Eight subjects participated in the evaluation. The room layout is illustrated in Fig. 13. Speech samples were 10-second excerpts from those used in simulations. The number of delay samples for LC-1IS was set to 2, *i.e.* LC-1IS2 was evaluated as the conventional structure.

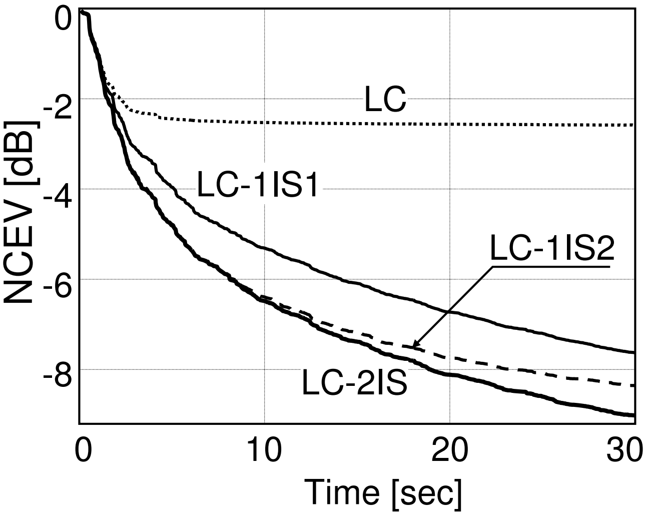

Figure 11: Normalized coefficient-error vector (NCEV) for 1000 taps.

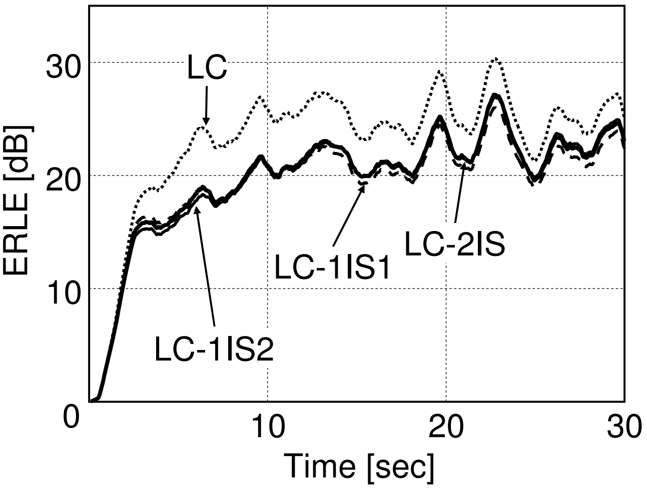

Figure 12: ERLE for 1000 taps.

Figure 14 presents the evaluation result. The length of each bar shows an average score for all subjects and speech signals. The 95% confidence interval is represented by the horizontal line at the right end of the bar. LC-1IS and LC-2IS both have lower scores than that of the reference, indicating that input-sliding introduces some degradation in subjective quality. However, LC-2IS with an average score of 4.65 is closer to the reference than LC-1IS with an average score of 4.15.

Table 2: Grading Scale.

Score	Description
5	Imperceptible
4	Perceptible, but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

A grading difference, which is calculated by subtracting the score of the signal for evaluation from that of the reference, was also calculated. A larger negative grading difference means more serious degradation in the subjective quality. As can be seen in Fig. 15, the grading difference for LC-2IS is -0.24 and 0.37 higher than that for LC-1IS with -0.61. In addition, there is a statistically significant difference between the scores of LC-1IS and LC-2IS, suggesting smaller subjective-quality degradation of the latter. This superior subjective quality is mainly due to symmetrical shifts of the sound image. Such symmetrical shifts, resulting from input sliding in two channels, are less noticeable than asymmetrical shifts.

5. CONCLUSION

A stereo echo canceller with simultaneous 2-channel input slides has been proposed. An additional input-sliding unit has been introduced to the channel without such a unit for an additional condition for echo cancellation. This unit has been shown effective for fast convergence and better sound quality of the far-end signal. Simulation results have shown that the proposed echo canceller achieves 20% faster convergence than the conventional one. Subjective evaluation results have demonstrated that the far-end signal with simultaneous input slides in two channels obtained a 0.37 higher score in the ITU-T 5-grade impairment scale with a statistically significant difference.

REFERENCES

[1] R. Botros, O. Abdel-Alim, and P. Damaske, “Stereophonic Speech Teleconferencing,” Proc. ICASSP’86, pp.1321–1324, Apr. 1986.

[2] T. Fujii and S. Shimada, “A note on Multi-Channel Echo cancellers,” Technical Report of IEICE, CS84-178, pp. 7–14, Jan. 1985 (in Japanese).

[3] A. Hirano and A. Sugiyama, “A New Multi-Channel Echo canceller with a Single Adaptive Filter per Channel,” Proc. Nat. Conv. of IEICEJ, A–202, Mar 1991. (in Japanese)

[4] A. Hirano and A. Sugiyama, “Convergence Characteristics of a Multichannel Echo canceller with Strongly Cross-Correlated Input Signals –Analytical Results–,” Proc. of 6th DSP Symposium of IEICEJ, pp. 144–149, Nov. 1991.

[5] Y. Joncour and A. Sugiyama, “A Stereo Echo Canceler with Correct Echo-Path Identification, ” Proc. of ICASSP’98, pp.3677–3680, May 1998.

[6] A. Sugiyama, Y. Joncour, and A. Hirano, “A Stereo Echo Canceler with Correct Echo-Path Identification Based on an Input Sliding Technique, ” IEEE Trans. SP, pp.2577–2587, Nov. 2001.

[7] J. Benesty, D. R. Morgan and M. M. Sondhi, “A Better Understanding and an Improved Solution to the Problems of Stereophonic Acoustic Echo Cancellation,” IEEE Trans. SAP, Vol. 6, No. 2, pp.156–165, Mar. 1998.

[8] S. Haykin, “Adaptive Filter Theory, Third Edition, ” Prentice-Hall, 1996

[9] S. Bergman, C. Grewin and T. Rydén, “The SR Report on The MPEG/Audio Subjective Listening Test Stockholm April/May 1991, ” ISO/IEC JTC1/SC2/WG11 MPEG 91/010, May 1991.

[10] CCIR Recommendation 562, 1990.

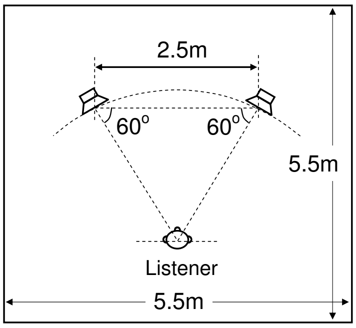

Figure 13: Room layout.

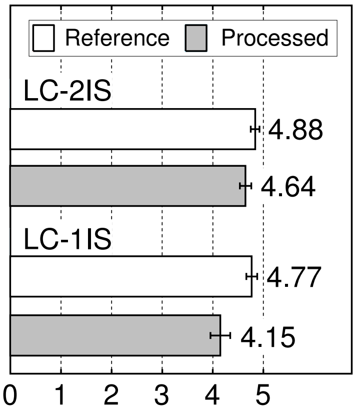

Figure 14: Subjective evaluation result.

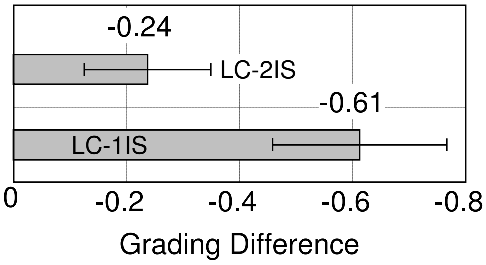

Figure 15: Grading difference.