PAPER

# Subjective Assessment of the Desired Echo Return Loss for Subband Acoustic Echo Cancellers

**Sumitaka SAKAUCHI**[†], *Nonmember*, **Yoichi HANEDA**[†], **Shoji MAKINO**[†*], **Masashi TANAKA**[†], *and* **Yutaka KANEDA**[†**], *Regular Members*

**SUMMARY** We investigated the dependence of the desired echo return loss on frequency for various hands-free telecommunication conditions by subjective assessment. The desired echo return loss as a function of frequency ($\mathrm{DERL}_f$) is an important factor in the design and performance evaluation of a subband echo canceller, and it is a measure of what is considered an acceptable echo caused by electrical loss in the transmission line. The $\mathrm{DERL}_f$ during single-talk was obtained as attenuated band-limited echo levels that subjects did not find objectionable when listening to the near-end speech and its band-limited echo under various hands-free telecommunication conditions. When we investigated the $\mathrm{DERL}_f$ during double-talk, subjects also heard the speech in the far-end room from a loudspeaker. The echo was limited to a 250-Hz bandwidth assuming the use of a subband echo canceller. The test results showed that: (1) when the transmission delay was short (30 ms), the echo component around 2 to 3 kHz was the most objectionable to listeners; (2) as the transmission delay rose to 300 ms, the echo component around 1 kHz became the most objectionable; (3) when the room reverberation time was relatively long (about 500 ms), the echo component around 1 kHz was the most objectionable, even if the transmission delay was short; and (4) the $\mathrm{DERL}_f$ during double-talk was about 5 to 10 dB lower than that during single-talk. Use of these $\mathrm{DERL}_f$ values will enable the design of more efficient subband echo cancellers.

***key words:*** *hands-free telecommunication, subband echo canceller, transmission delay, reverberation, auditory characteristics*

## 1. Introduction

Acoustic echo cancellers are widely used for teleconferencing and hands-free telecommunication systems to overcome acoustic feedback, making conversation more comfortable. An acoustic echo canceller is generally composed of an adaptive filter and a nonlinear processor [1], [2].

To design an acoustic echo canceller, it is important to know the desired echo return loss, i.e., how much the returned echo should be suppressed, because the adaptive filter length and loss insertion level depend on the desired echo return loss. Subjective test results for various transmission delays have shown that the desired echo return loss increases significantly when the transmission delay rises above about 50 ms [3], [4]. However, these findings are not necessarily applicable to a subband echo canceller, because the desired echo return loss was investigated assuming the use of a conventional fullband echo canceller.

The configuration of the subband echo canceller is shown in Fig. 1. For subband analysis and synthesis, a polyphase filterbank is used. Subband echo cancellers divide signals into smaller frequency subbands and cancel echoes independently in each subband [5], [6]. Since narrower frequency subbands have a smaller eigenvalue spread than the full band for speech inputs, the convergence speed can be increased. Since downsampling expands the sampling interval and reduces the number of taps needed for the adaptive filter, the subband echo canceller is computationally efficient.

The desired echo return loss as a function of frequency ($\mathrm{DERL}_f$) has been studied theoretically using the threshold of audibility to determine optimal adaptive filter tap allocation tables for a subband echo canceller [7], [8]. These investigations did not, however, sufficiently consider the influence of various conditions such as transmission delay or room reverberation time that affect hands-free telecommunication. Nor did they take into account whether the speech was single-talk or double-talk.

We used subjective assessment to empirically investigate the effect on $\mathrm{DERL}_f$ of transmission delay

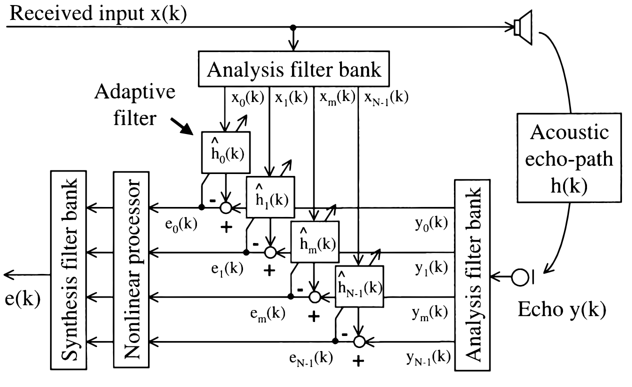


**Fig. 1** Configuration of a subband echo canceller.

Manuscript received March 28, 2000.
Manuscript revised July 7, 2000.

[†] The authors are with NTT Cyber Space Laboratories, Musashino-shi, 180-8585 Japan.

[*] Presently, the author is with NTT Communication Science Laboratories, kyoto-fu, 619-0237 Japan.

[**] Presently, the author is with the acoustics signal processing at the Department of Information and Communication Engineering, Tokyo Denki University, chiyoda-ku, 101-8457 Japan.

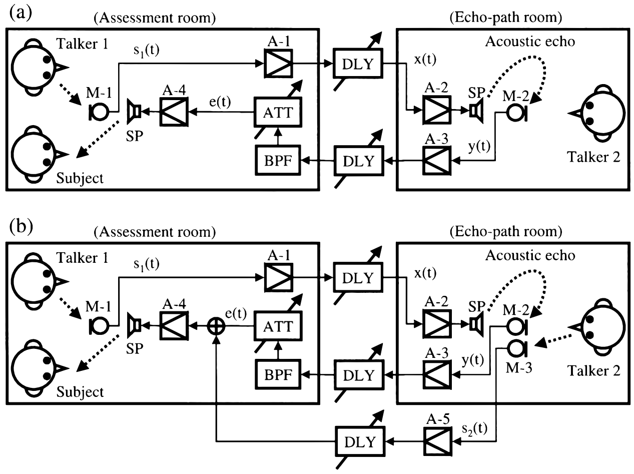


**Fig. 2** Schematic diagram of the subjective assessment system in (a) the single-talk state and (b) the double-talk state.

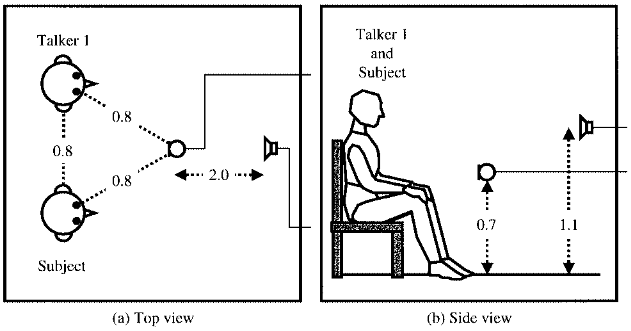
($T_{\mathrm{td}}$), reverberation time ($T_{60}$) in the echo-path room, and conversation state, i.e., single-talk or double-talk. These experiments clarified the perceptual characteristics of the returned echo in the frequency domain, which greatly affect the performance of an acoustic echo canceller. This paper presents our subjective test results and some theoretical discussion of them. We also provide an example of an efficient subband echo canceller design based on our obtained DERL$_f$.

## 2. Test Methods

### 2.1 Assessment System

To investigate DERL$_f$, subjective tests were done with various transmission delays and various room reverberation times under single-talk and double-talk states. Figure 2 shows a schematic diagram of the subjective assessment system which assumes a hands-free teleconferencing system comprising audio facilities, transmission lines, and a pair of conference rooms.

Figure 2(a) shows the single-talk state assessment system. The near-end speech uttered by talker 1 was picked up by microphone M-1 in the assessment (near-end) room. The microphone output signal $s_1(t)$ was amplified by amplifier A-1, delayed by a variable delay unit (DLY) which simulated transmission delay and sent to the echo-path (far-end) room. There, the signal was amplified by amplifier A-2 causing a loudspeaker input signal $x(t)$. The loudspeaker (SP) output signal reached microphone M-2 via the acoustic echo-path in the echo-path room. The microphone's output signal (echo signal) $y(t)$ returned to the assessment room via amplifier A-3 and a variable delay unit (DLY). The returned signal was limited to 250-Hz bandwidth by a band pass filter (BPF) to investigate the frequency dependence of DERL$_f$. The output signal of the BPF was applied to a variable attenuator (ATT), and its output signal $e(t)$ was amplified by amplifier A-4 and applied to a loudspeaker (SP). For the single-talk state assessment, the subject heard both the near-end speech by talker 1 and its band-limited echo $e(t)$ from a loudspeaker. Here, the transmission (round-trip) delay time as the assessment parameter was changed by using the delay units (DLY).



(a) Top view

(b) Side view

**Fig. 3** Top and side views of the assessment room layout (unit = meter).

For the double-talk state assessment, we added the far-end speech uttered by talker 2 to the band-limited echo $e(t)$. In the system in Fig. 2(a), however, the far-end speech became limited by BPF and attenuated by ATT. So, we designed the theoretical double-talk state assessment system shown in Fig. 2(b), where the far-end speech and the returned echo were independently picked up by microphones M-2 and M-3. It is, however, impossible to make this system in practice, so the echo-path room was actually set not in a real room but in a simulated room for both state assessments. That is, the impulse responses between a loudspeaker and a microphone and between the talker and a microphone were measured in advance. Then, digital filters that matched the measured impulse responses were used instead of the real acoustic path. Mouth simulators were used instead of real speakers for talkers 1 and 2. These simulations kept the echo-path room and uttered speech conditions fixed so these were not considered to affect the assessment results.

The arrangement of the loudspeaker, microphone, talker 1, and subject in the assessment room is shown in Fig. 3. The arrangement in the echo-path room was almost the same. These setups were based on the conventional work [4]. The microphones had unidirectional acoustic field characteristics.

The output levels of loudspeakers and microphones, and the background noise level in both rooms were set according to ITU-T recommendation P. 34 [9]. The gains of amplifiers A-1, A-3, and A-5 were adjusted to give the delay units an input (electrical) level of −10 dBm when the average speech (acoustic) level was −28.7 dBPa at the point of the microphone. The gains of amplifiers A-2 and A-4 were adjusted to give the loudspeakers an output (acoustic) level of 70 dBspl

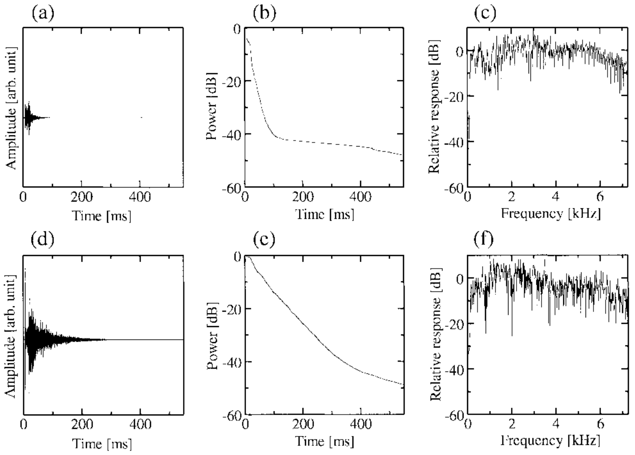


**Fig. 4** Impulse responses and acoustic characteristics of the echo-path room: (a), (b), (c) are impulse response, reverberation curve, and frequency response when $T_{60}$ was 125 ms, (d), (e), (f) are those when $T_{60}$ was 500 ms.

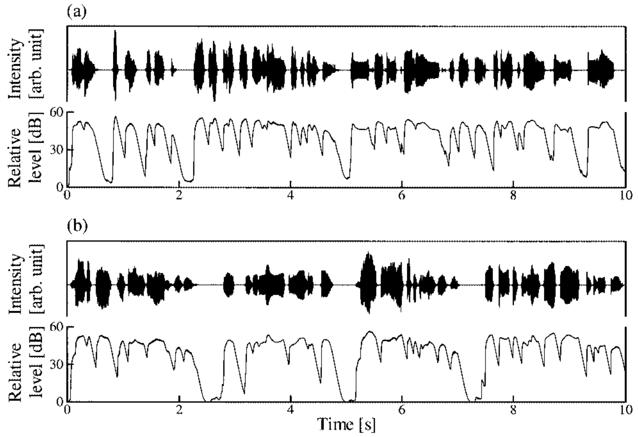


**Fig. 5** Waveforms and their power envelopes presented to the subjects: (a) near-end speech and (b) far-end speech.

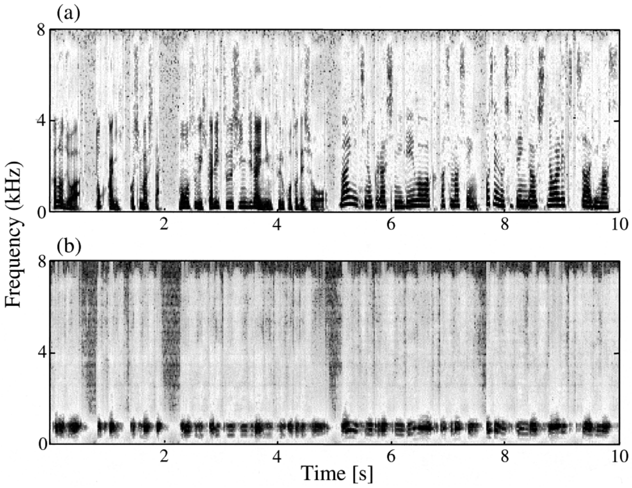


**Fig. 6** Spectrogram examples presented to the subjects: (a) near-end speech and (b) its band-limited echo.

at the subject's ear position when the delay unit output (electrical) level was −10 dBm. The background noise levels in both rooms were kept below 35 dBA.

The system's frequency range was set from 0.1 to 7 kHz. Both rooms were variable reverberation rooms, 7 × 6 × 4 m. The assessment room had a reverberation time of about 270 ms, which corresponds to an acoustically treated meeting room. The echo-path room was given two different sets of acoustic characteristics, distinguished by reverberation times of about 125 and 500 ms as an assessment parameter. Figure 4 shows the impulse response waveform, reverberation curve, and frequency response from the loudspeaker to the microphone for each reverberation time in the echo-path room. The acoustic coupling from the loudspeaker to the microphone in the echo-path room was about −3 dB in both cases.

### 2.2 Test Procedures

The test procedures were as follows: The subject was seated in the assessment room, and listened to the band-limited echo. The subject increased the channel loss with an attenuator (ATT) until the band-limited echo was no longer objectionable, and the attenuation level at that point was regarded as $\mathrm{DERL}_f$. The frequency bands were those with 250-Hz bandwidth centered at 250, 500, 750, ⋯, 6750, 7000 Hz. The subjective tests were repeated in a random pattern with each band-limited echo.

The near-end and far-end speech presented to the subject were different sets of several short Japanese sentences (about 10 s in total) spoken by either a man or a woman. Figure 5 shows the waveforms and their power envelopes of (a) the near-end speech uttered by talker 1 and (b) the far-end speech uttered by talker 2. Figure 6 shows the spectrograms of (a) the near-end speech and (b) an example of its band-limited echo, whose frequency range was from 625 to 875 Hz (the third subband of 250-Hz bandwidth).

There were 15 subjects whose ages ranged from 25 to 35. Here, all experimental data was evaluated by a t-test where the significance level was 0.01.

## 3. Test Results and Analysis

### 3.1 Transmission Delay

Figure 7 shows subjective test results of $\mathrm{DERL}_f$ for transmission delays of 30 ms (circles), 50 ms (squares), 100 ms (closed triangles), and 300 ms (open triangles). The reverberation time of the echo-path room was set to 125 ms.

The longer the transmission delay was, the larger the $\mathrm{DERL}_f$ values at any frequency were. The frequency bands of maximum $\mathrm{DERL}_f$, however, changed with the transmission delay. When the transmission de-

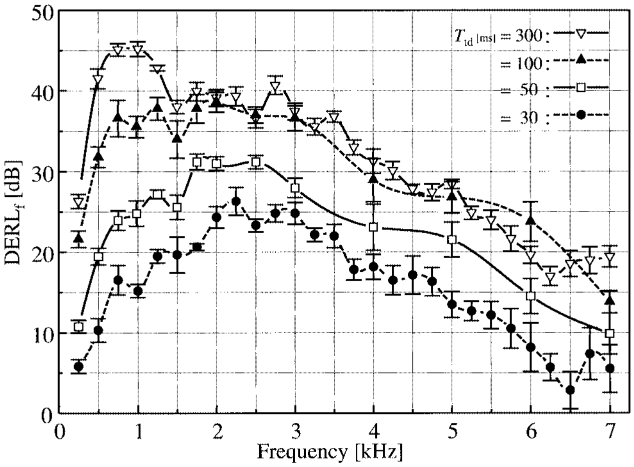


**Fig. 7** Subjective DERL$_f$ test results for transmission delays from 30 to 300 ms.

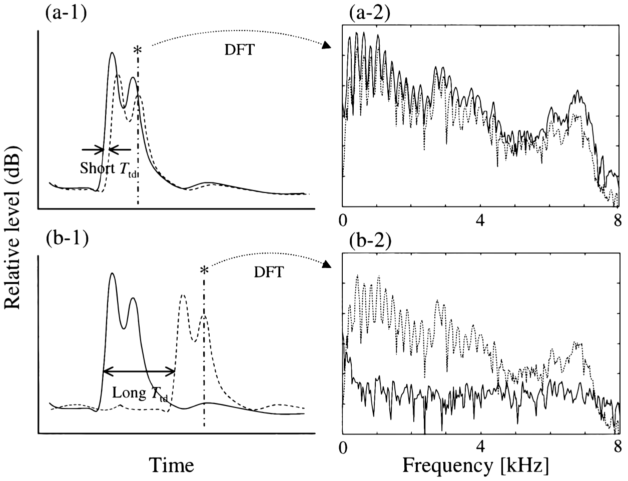


**Fig. 8** Signal power envelopes and short-time spectra of the near-end speech (solid line) and its echo (dashed line) for different transmission delays.

lay was short ($T_{\rm td}$ = 30 or 50 ms), the DERL$_f$ reached a maximum at 2 to 3 kHz. When the transmission delay was 300 ms, the DERL$_f$ in the low-frequency bands around 1 kHz was significantly higher than when the transmission delay was short. Put another way, when the transmission delay was short, the echo component around 2 to 3 kHz was the most objectionable to subjects (listeners). As the transmission delay rose to 300 ms, the echo component around 1 kHz became the most objectionable. The maximum values of DERL$_f$ for the various transmission delays are consistent with previous desired echo return loss results for a fullband acoustic echo canceller ([4]), and they also fit the required echo return loss in the ITU-T recommendation G. 165 [10].

To understand the dependency of the most objectionable frequency component on the transmission delay, we investigated the masking effect of the echo. Figure 8 shows examples of signal power envelopes (a-1, b-1) and short-time spectra (a-2, b-2) of the near-end speech (solid line) and its echo (dashed line) at the point (*). The upper figures show the relationship between the two signals in the case of a short transmission delay and the lower figures show that for a long one.

When the transmission delay was short, the subject heard both the near-end speech and its echo at almost the same time, as shown in Fig. 8(a-1). In this case, thier short-time spectra were very similar, as shown in Fig. 8(a-2). Here, it was supposed that the DERL$_f$ was equal to the power differences of the echo and the simultaneous masking threshold by the near-end speech. The simultaneous masking threshold was proportional to the power densities of the near-end speech in the frequency domain and the threshold at low frequencies was larger than that at high frequencies. Therefore, it was estimated that the DERL$_f$, which corresponds to the power differences, was almost the same over all frequencies. However, the DERL$_f$ peaked at about 2 or 3 kHz, where the human ear is perceptually the most sensitive [11]. When the transmission delay was long (e.g., $T_{\rm td}$ = 300 ms), the waveforms of the near-end speech and its echo had a long time lag, as shown in Fig. 8(b-1). So the echo sometimes appeared when the near-end speech was not present. In this case, the subjects heard only the echo that had high energy at low frequencies, as shown in Fig. 8(b-2). Therefore, DERL$_f$ peaked in the low-frequency bands around 1 kHz.

## 3.2 Reverberation Time

Figure 9 shows subjective DERL$_f$ test results for reverberation times in the echo-path rooms of 125 and 500 ms for short (30 ms) and long (300 ms) transmission delays. The combinations of $T_{\rm td}$ and $T_{60}$ conditions were 300 and 125 ms (open triangles), 300 and 500 ms (closed triangles), 30 and 500 ms (squares), and 30 and 125 ms (circles).

When the transmission delay was short ($T_{\rm td}$ = 30 ms), DERL$_f$ in the low-frequency bands around 1 kHz for a long reverberation time (squares) was higher than that for a short reverberation time (circles). That is to say, when the room reverberation time was relatively long, the echo component around 1 kHz became the most objectionable, even when the transmission delay was short. One possible reason for this phenomenon is that the latter reverberations affected DERL$_f$ in the same way that the long-transmission-delayed echo did, although the transmission delay was short. The DERL$_f$ value for the long reverberation time was, however, lower than that for the long transmission delay, because the latter reverberations had less energy than the direct sound. On the other hand, when the transmission delay was long ($T_{\rm td}$ = 300 ms), changes in the reverberation time hardly affected DERL$_f$. The DERL$_f$ seemed to

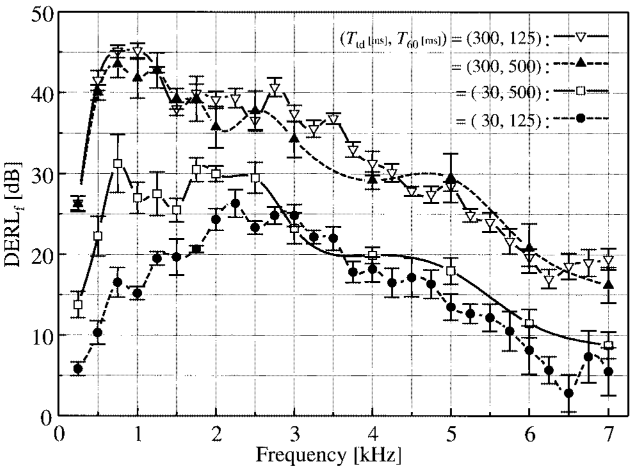


**Fig. 9** Subjective DERL$_f$ test results for different reverberation times in the echo-path room.

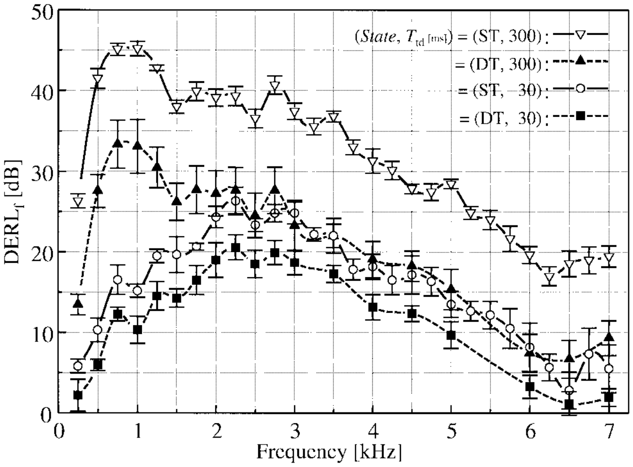


**Fig. 10** Subjective DERL$_f$ test results for the differences between single-talk (ST) and double-talk (DT).

be almost saturated.

### 3.3 Single- and Double-Talk States

Figure 10 shows subjective DERL$_f$ test results for both single-talk (ST) and double-talk (DT). The combinations of the state and $T_{\mathrm{td}}$ conditions were single-talk state and 300 ms (open triangles), double-talk state and 300 ms (closed triangles), single-talk state and 30 ms (circles), and double-talk state and 30 ms (squares). When the transmission delay was short ($T_{\mathrm{td}} = 30$ ms), DERL$_f$ during double-talk was about 5 dB lower than that for single-talk at all frequencies. When the transmission delay was long ($T_{\mathrm{td}} = 300$ ms), DERL$_f$ during double-talk was about 10 dB lower.

When the transmission delay was short, the subject heard both the near-end speech and its echo at almost the same time. In this case, thier short-time spectra were very similar. Thus, most of the echo was simultaneously masked by the near-end speech, and there were few non-masked parts of the echo even during the single-talk state. During the double-talk state, the far-end speech slightly masked the parts of the echo that were not masked by the near-end speech. When the transmission delay was short, the masking effect by the near-end speech was dominant and that by the far-end speech was small. Therefore, DERL$_f$ during double-talk was slightly lower than that during single-talk.

When the transmission delay was long, the subject heard both the near-end speech and its echo separately because both had a long time lag. So a large simultaneous masking effect by the near-end speech like that in the short transmission delay did not occur. The only effect of the near-end speech was partial masking and the subjects heard the echo when the near-end speech was not present. During the double-talk state, the partial masking effect by the far-end speech was almost the same as that by the near-end speech, because both the near-end and far-end speech had almost the same averaged power in the time domain. In particular, the far-end speech partially masked the echo when the near-end speech was not present. When the transmission delay was long, therefore, the masking effect of the far-end speech was larger than that in the short transmission delay so there was a larger difference in DERL$_f$ between double-talk and single-talk.

These experimental results show that the difference in conversation state (single-talk or double-talk) has as much influence on the DERL$_f$ as the difference in the transmission delay. If the system can detect the conversation state, then the loss insertion level can probably be reduced during double-talk to improve the speech quality.

## 4. Example of Efficient Subband Echo Cancellers Design

These subjective test results can be used to design subband echo cancellers. Figure 11 shows a subband echo canceller performance profile that was calculated based on our obtained DERL$_f$ for transmission delay of 300 ms and reverberation time of 500 ms during single-talk in accordance with ITU-T recommendation G. 167 [12]. The vertical axis is the total echo suppression level, corresponding to the DERL$_f$, composed of desired echo return loss enhancement due to the adaptive digital filters (*ERLE*-ADF) and insertion loss levels using the nonlinear processor (NLP).

Figure 12 shows a conventional fullband echo canceller performance profile transformed into the frequency domain. Both performance profiles are designed for the same amount of hardware. In our proposed design, the echo suppression performance is weighted at low frequencies around 1 kHz, because DERL$_f$ in the low-frequency bands is maximum for the transmission delay of 300 ms. The echo suppression per-

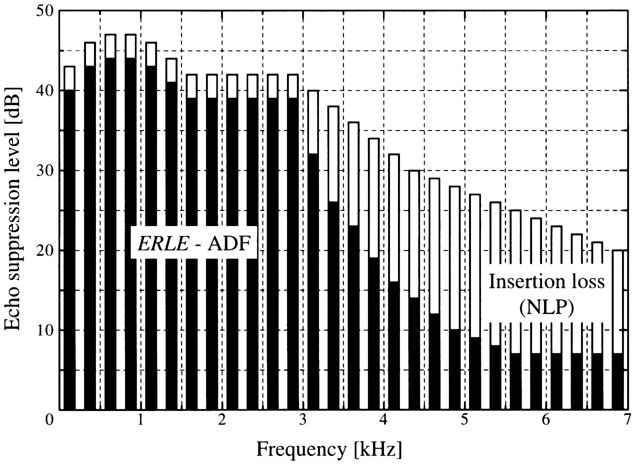


**Fig. 11** Efficient subband echo canceller performance profile.

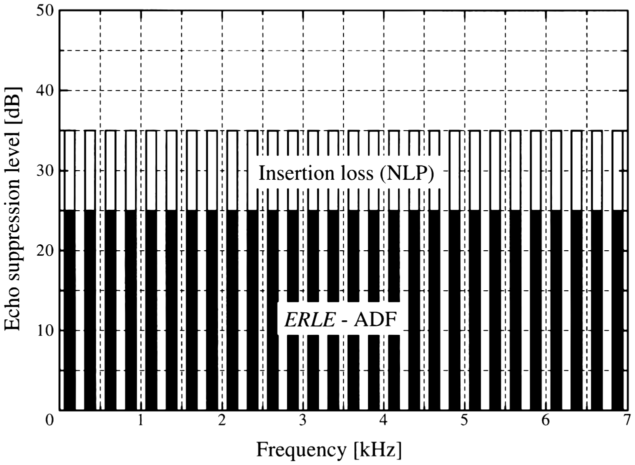


**Fig. 12** Conventional fullband echo canceller performance profile.

formance in the high-frequency bands (4 kHz or more) is reduced, because the echoes at high frequency are not very annoying. Although the hardware volumes in Figs. 11 and 12 are the same, our design is perceptually suited to DERL$_f$.

Moreover the performance of the adaptive filters is dominant when a long filter tap length in the low-frequency bands around 1 kHz is set while the performance of the nonlinear processor is dominant in the high-frequency bands (4 kHz or more). In the low-frequency bands, there exist several formants of speech. Thus, the far-end speech is not choppy even during the double-talk state.

## 5. Conclusion

We investigated the desired echo return loss as a function of frequency (DERL$_f$) for various hands-free telecommunication conditions by subjective assessment. Subjective test results showed that DERL$_f$ in the low-frequency bands around 1 kHz increased significantly when the transmission delay was long. The effect of the reverberation time on DERL$_f$ was weaker than that of the transmission delay. When the reverberation time was long, however, DERL$_f$ in the low-frequency bands increased, even if the transmission delay was short. The DERL$_f$ during double-talk was lower than that during single-talk, especially when the transmission delay was long. Using these results, we obtained guidelines for efficient subband echo canceller design.

## Acknowledgments

We thank Dr. N. Kitawaki and Mr. K. Yamamori for their advice on our approach to analyzing the desired echo return loss and its impact on echo canceller performance.

### References

[1] E. Hänsler, "The hands-free telephone problem—An annotated bibliography," Signal Processing, vol.27, pp.259–271, 1992.
[2] E. Hänsler, "The hands-free telephone problem: An annotated bibliography update," Annales des Télécommunications, vol.49, no.7–8, pp.360–367, July-Aug. 1994.
[3] N. Kishimoto, K. Ishimaru, and K. Takahashi, "Transmission quality of hand-free audio teleconference services," IEEE ICC'88, no.8.4, 1988.
[4] H. Yasukawa, M. Ogawa, and M. Nishino, "Echo return loss required for acoustic echo controller based on subjective assessment," IEICE Trans., vol.E-74, no.4, pp.692–705, April 1991.
[5] W. Kellermann, "Analysis and design of multirate systems for cancellation of acoustical echoes," Proc. ICASSP'88, pp.2570–2573, April 1988.
[6] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: Analysis, experiments, and applications to acoustic echo cancellation," IEEE Trans. Signal Processing, vol.40, pp.1862–1875, Aug. 1992.
[7] E.J. Diethorn, "Perceptually optimum adaptive filter tap profiles for subband acoustic echo cancellers," Proc. ICSPAT'95, vol.I, pp.290–293, Boston, 1995.
[8] M. Vukadinovic and T. Aboulnasr, "A study of adaptive intersubband tap assignment algorithms from a psychoacoustic point of view," Proc. ISCAS'96, vol.2, pp.65–68, 1992.
[9] International Telecommunication Union, "Transmission performance of hands-free telephones," Recommendation P. 34 CCITT Blue Book, 5, 1988.
[10] International Telecommunication Union, "General characteristics of international telephone connections and international telephone circuits—Echo cancellers," ITU-T Recommendation G. 165, 1993.
[11] E. Zwicker, Psychoakustik, Springer Verlag, 1982.
[12] International Telecommunication Union, "General characteristics of international telephone connections and international telephone circuits—Acoustic echo controllers," ITU-T Recommendation G. 167, 1993.

**Sumitaka Sakauchi** received the B.S. degree from Yamagata University in 1993 and the M.S. degree from Tohoku University in 1995. He joined NTT Human Interface Laboratories in 1995. He is now a research engineer in NTT Cyber Space Laboratories. His research interests include acoustic signal processing and acoustic echo cancellation. He is a member of the Acoustical Society of Japan.

**Yoichi Haneda** received the B.S., M.S., and Ph.D. degrees from Tohoku University in Sendai, in 1987, 1989, and 1999, respectively. Since joining Nippon Telegraph and Telephone Corporation (NTT) in 1989, he has been investigating the modeling of acoustic transfer functions, acoustic signal processing, and acoustic echo cancellers. He is now a senior researcher in NTT Cyber Space Laboratories. Dr. Haneda is a member of the Acoustical Society of Japan and the Institute of Electronics, Information, and Communication Engineers of Japan.

**Shoji Makino** received the B.E., M.E., and Ph.D. degrees from Tohoku University, Sendai, Japan, in 1979, 1981, and 1993, respectively. He joined the Electrical Communication Laboratory of Nippon Telegraph and Telephone Corporation (NTT) in 1981. Since then, he has been engaged in research on electroacoustic transducers and acoustic echo cancellers. His research interests include acoustic signal processing, and adaptive filtering and its applications. Dr. Makino received the Outstanding Technological Development Award of the Acoustical Society of Japan in 1995, and the Achievement Award of the Institute of Electronics, Information, and Communication Engineers of Japan in 1997. He is the author or co-author of more than 100 articles in journals and conference proceedings, and more than 130 patents. He is a member of the Audio and Electroacoustics Technical Committee of the IEEE Signal Processing Society. He served on the Technical Committee of the 1999 IEEE Workshop on Acoustic Echo and Noise Control. He is a Senior Member of the IEEE and a member of the Acoustical Society of Japan and the Institute of Electronics, Information, and Communication Engineers of Japan.

**Masashi Tanaka** received B.E. and M.E. from Hokkaido University in 1988 and 1990. He joined Nippon Telegraph and Telephone Corporation (NTT) in 1990. Since then, he has been engaged in signal processing in microphone arrays and acoustic echo cancellation. He is a member of the Acoustical Society of Japan and the Institute of Electronics, Information, and Communication Engineers of Japan.

**Yutaka Kaneda** received the B.E., M.E. and Doctor of Engineering degrees from Nagoya University, Nagoya, Japan, in 1975, 1977, and 1990, respectively. From 1977 to 2000, he was with the Electrical Communication Laboratory of Nippon Telegraph and Telephone Corporation (NTT), Musashino, Tokyo, Japan, where his work included microphone array signal processing, sound field control, and acoustic echo cancellation. He is now professor in acoustic signal processing at the Department of Information and Communication Engineering, Tokyo Denki University, Tokyo, Japan. Dr. Kaneda is a member of the ASJ, ASA, IEICE, and IEEE.