Suppression of Late Reverberation Effect on Speech Signal Using Long-Term Multiple-step Linear Prediction

Keisuke Kinoshita, Marc Delcroix, *Senior Member,* Masato Miyoshi

Professor:陳嘉平

Reporter:葉佳璋

Outline

- Introduction
- Signal Model
- Single Channel Algorithm
- Multiple channel Algorithm
- Experiment

Introduction

- A speech signal capture by a distant microphone is generally smeared by reverberation.

- It is desirable to find a reliable way of mitigating the effect of reverberation on ASR.

Introduction

- Reverberant speech is assumed to consist of a direct-path response, early reflection and last reverberations.

- The early reflection may not significantly degrades ASR if they are handled by CMS.

Signal model

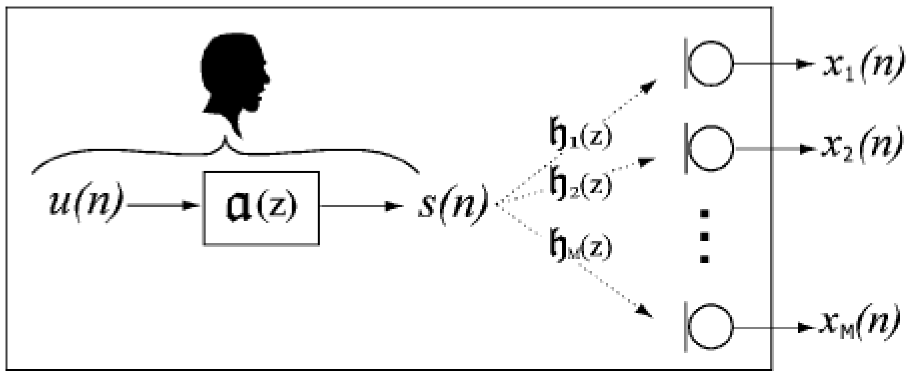


Fig. 1. Acoustic system: $u(n)$ is white noise, $\mathfrak{a}(z)$ is an FIR filter corresponding to vocal tract characteristics, $s(n)$ is a speech signal, $\mathfrak{h}_m(z)$ is the room transfer function between the speaker and the mth microphone, and $x_m(n)$ is an observed signal at the mth microphone.

Signal model

- s(n) is produced through a Pth-order FIR filter

$$s(n) = \sum_{k=0}^{P} a(k)u(n-k)$$

- Where
  - ➢s(n): a source signal(speech signal)
  - ➢a(z): Pth-order FIR filter
  - ➢u(n): white noise

Signal model

- The speech signal recorded with a distant microphone m, $x_m(n)$ can be generally modeled

$$x_m(n) = \sum_i h_m(i) s(n-i),$$

$$= \sum_l g_m(l) u(n-l),$$

$$g_m(l) = \sum_{k=0}^{P} h_m(l-k) a(k)$$

- Where $h_m(n)$ corresponds to the room impulse response between the source signal.

Signal model

- We can reformulate using matrix/vector notation as

$$x_m(n) = G_m u(n)$$

$$u(n) = \left[u(n),\ u(n-1), \ldots,\ u(n-T-N+1)\right]^T$$

$$x_m(n) = \left[x_m(n),\ x_m(n-1), \ldots,\ x_m(n-N)\right]^T$$

$$g_m = \left[g_m(n),\ g_m(n-1), \ldots,\ g_m(T-1)\right]$$

$$G_m = \begin{bmatrix} g_m & 0 & \cdots & 0 \\ 0 & g_m & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & g_m \end{bmatrix}$$

➢N: dimensions of x_m .

➢T: dimension of g_m .

Signal model

- Here let us denote the late reverberation of $g_m(n),\ g_{late,m}$ as

$$g_{late,m} = [g_m(D),\ g_m(D+1),\ \cdots,\ g_m(T-1),\ 0,\ \cdots,\ 0]$$

We consider that late reverberation of g_m corresponding the coefficient of g_m after Dth element.

Long-Term Muti-step Linear Prediction

- We use the method to identify only the late reverberations.

$$x_1(n) = \sum_{p=0}^{N} w(p) x_1(n - p - D) + e(n)$$

  - w(n): represent the prediction coefficient
  - e(n): prediction error
  - D: step-size(i.e., delay)

Long-Term Muti-step Linear Prediction

- By minimizing the mean square energy of prediction error e(n),

$$(E\{x_1(n-D)x_1^T(n-D)\})w = E\{x_1(n-D)x_1^T(n)\}$$

where

$$w = \left[w(0),\ w(1), \ldots,\ w(N-1)\right]$$

- The prediction coefficient can be obtain as

$$w = (E\{x_1(n-D)x_1^T(n-D)\})^{-1}E\{x_1(n-D)x_1^T(n)\}$$

Long-Term Muti-step Linear Prediction

- Be expanded as

$$E\{x_1(n-D)x_1^T(n-D)\} = G_1 E\{u(n-D)u^T(n-D\}G_1^T$$
$$= \sigma_u^2 G_1 G_1^T$$

  - Where the auto-correlation matrix of white noise u(n) $E\{u(n-D)u^T(n-D\}$ is assume to be $\sigma_u^2 I$.
  - σ_u^2 : scalar that corresponds to the variance of u(n).

Long-Term Muti-step Linear Prediction

- The second term can also be expanded as

$$E\{x_1(n-D)x_1^T(n)\} = G_1 E\{u(n-D)u^T(n)\} g_1^T$$
$$= \sigma_u^2 G_1 g_{late,1}^T$$

- Finally w can be rewrite as

$$w = (G_1 G_1^T)^{-1} G_1 g_{late,1}$$

where

$$g_{late,1} = [g_1(D),\ g_1(D+1),\ \cdots,\ g_1(T-1),\ 0,\ \cdots,\ 0]$$

Long-Term Muti-step Linear Prediction

- Estimate the power of the late reverberation, as follows

$$
\begin{aligned}
&E\{(x_1^T w)^2\} \\
&= \| w^T G_1 E\{u(n-D)u^T(n)\} G_1^T w \| \\
&= \| \sigma_u^2 w^T G_1 G_1^T w \| \\
&= \| \sigma_u^2 g_{late,1}^T G_1^T (G_1 G_1^T)^{-1} G_1 g_{late,1} \| \\
&\leq \| \sigma_u^2 g_{late,1}^T \| \cdot \| G_1^T (G_1 G_1^T)^{-1} G_1 \| \cdot \| g_{late,1} \| \\
&= \| \sigma_u g_{late,1} \|^2
\end{aligned}
$$

Pre-Whitening

- Qth-order prediction filter α(n) was used for pre-whitening to equalize a(z)

$$r_m(c) = E\left[x_m(n)x_m(n+c)\right] \ (c = 0, 1, 2, \cdots)$$

- Then, we take the average of $r_m(c)$ over all channels.

$$\phi(c) = \frac{1}{M}\sum_{m=1}^{M} r_m(c)$$

Pre-Whitening

- As with standard LP using $\phi(c)$, the prediction filter α(n) was calculated based on the following Yule-Walker equation

$$\begin{bmatrix} \alpha(1) \\ \alpha(2) \\ \vdots \\ \alpha(q) \end{bmatrix} = \begin{pmatrix} \phi(0) & \phi(1) & \cdots & \phi(q-1) \\ \phi(1) & \ddots & & \vdots \\ \vdots & & \ddots & \phi(1) \\ \phi(q-1) & & & \phi(0) \end{pmatrix}^{-1} \times \begin{bmatrix} \phi(1) \\ \phi(2) \\ \vdots \\ \phi(q) \end{bmatrix}$$

Spectral Subtraction

- Use of SS to suppress the late reverberations.

$$\left|\hat{S}_m(k\lambda,\omega)\right| = \begin{cases} \sqrt{\left|X_m(k\lambda,\omega)\right|^2 - \left|R_m(k\lambda,\omega)\right|^2} & \\ & (if \ \left|X_m(k\lambda,\omega)\right|^2 - \left|R_m(k\lambda,\omega)\right|^2 > 0) \\ 0 & (\text{otherwise}) \end{cases}$$

- $\hat{S}_m(k\lambda,\omega)$: STFT of the dereverberated signal
- $X_m(k\lambda,\omega)$: STFT of signal at mth microphone
- $R_m(k\lambda,\omega)$: estimated late reverberations

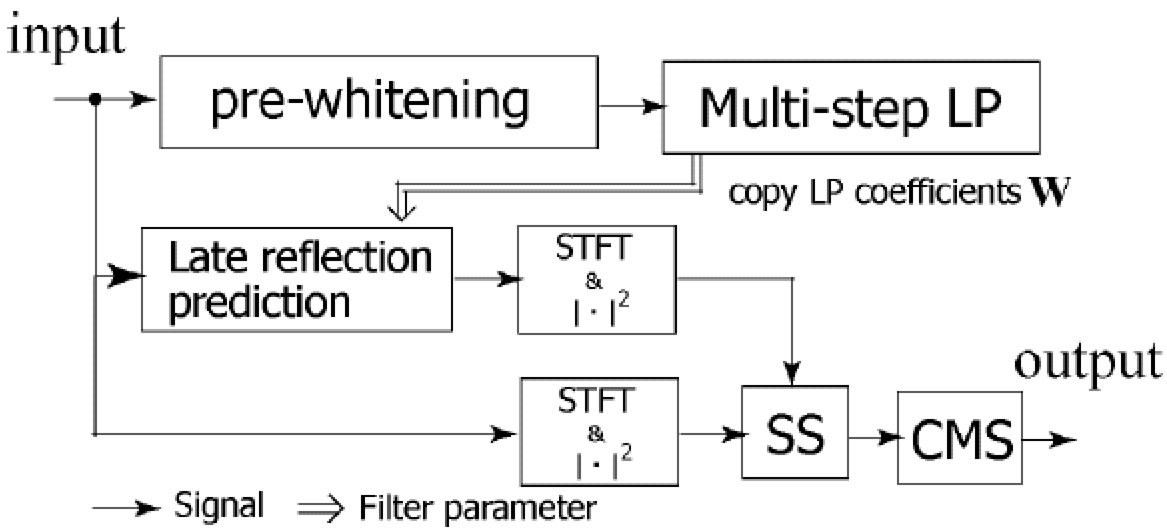


Fig. 2. Schematic diagram of proposed method for single-channel scenario.

Multichannel Long-Term Multi-step Linear Prediction

- We use the method to identify only the late reverberations.

$$x_i(n) = \sum_{m=0}^{M} \sum_{p=0}^{L} w_{m,i}(p) x_m(n - p - D) + e_i(n) \quad (i = 1, 2 \cdots, M)$$

  - $x_m(n)$: signal at the mth microphone
  - $e_i(n)$: prediction error
  - D: step-size(i.e., delay)
  - $w_{m,i}(n)$: represent the prediction coefficient

Multichannel Long-Term Multi-step Linear Prediction

- Single:

$$w = (E\{x_1(n-D)x_1^T(n-D)\})^{-1} E\{x_1(n-D)x_1^T(n)\}$$

$$w = (G_1 G_1^T)^{-1} G_1 g_{late,1}$$

- Multiple:

$$w_i = (E\{x_1(n-D)x_1^T(n-D)\})^{+} E\{x_1(n-D)x_1^T(n)\}$$

$$w_i = (G\ G^T)^{+} G\ g_{late,1}^T$$

$$= (G^T)^{+} g_{late,1}$$

$$where\ \ G = [G_1^T,\ G_2^T,\ \cdots,\ G_M^T]^T$$

Multichannel Long-Term Multi-step Linear Prediction

- We define the observed signal x(n) as

$$x(n) = [x_1^T(n),\ x_2^T(n),\ \cdots,\ x_M^T(n)]^T$$

- Estimated late reverberation can be expressed as follows

$$\begin{aligned} & x^T(n) w_i \\ & = u^T(n) G^T w_i \\ & = u^T(n) G^T (G^T)^+ g_{late,i} \\ & \cong u^T(n) g_{late,i} \end{aligned}$$

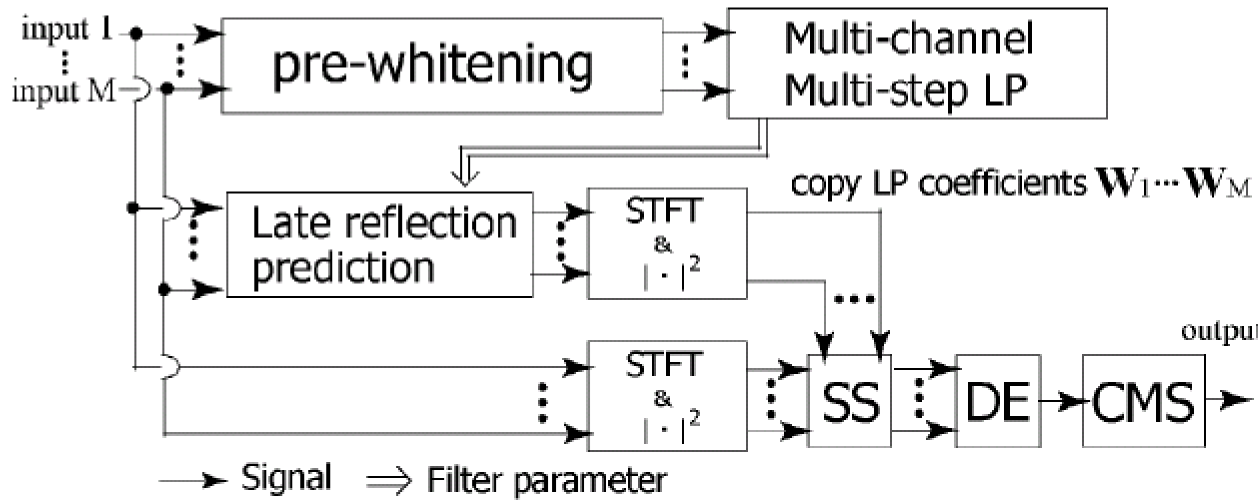


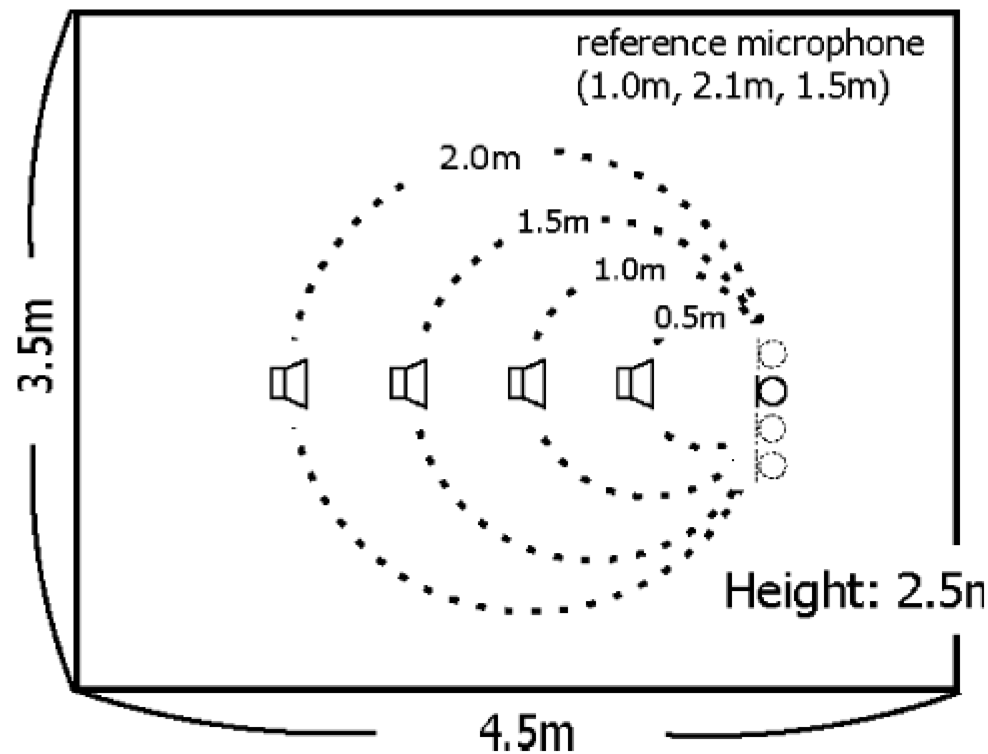
Fig. 3. Schematic diagram of multichannel implementation.

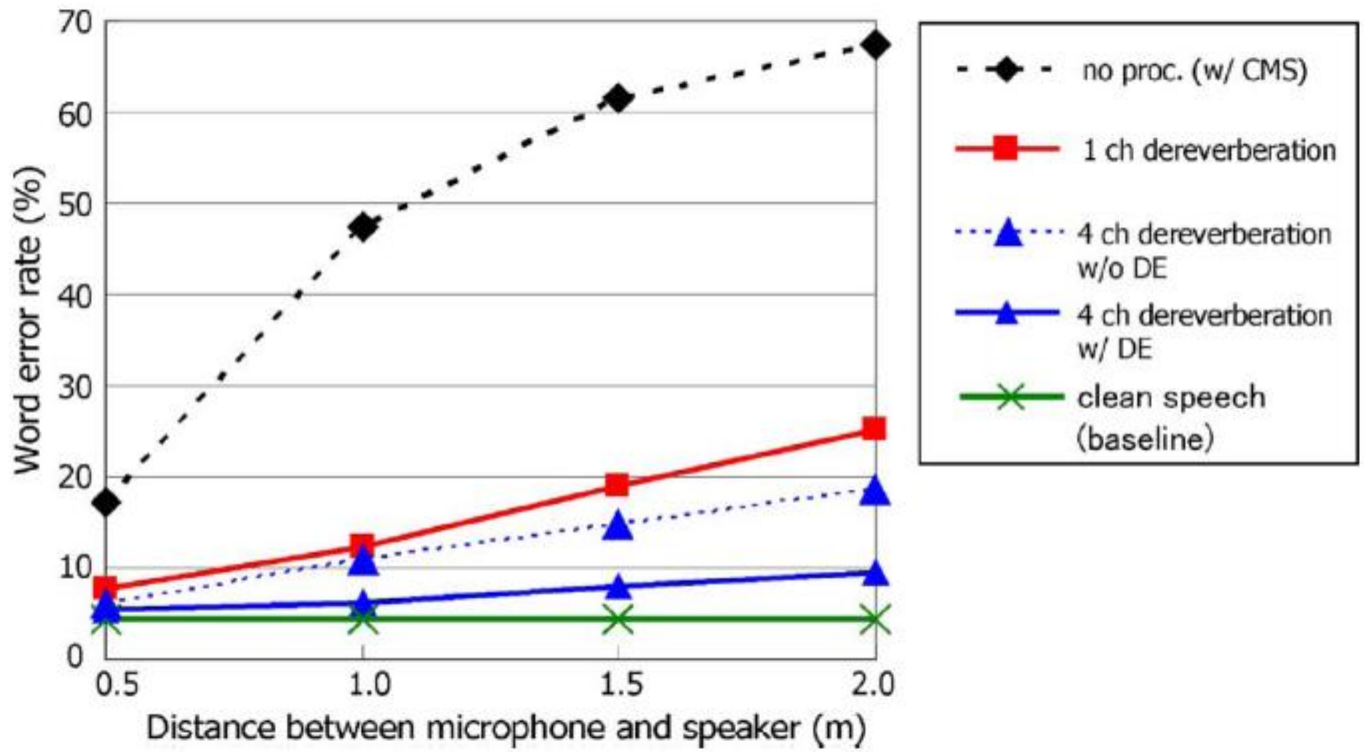
Schematic Processing Diagram

- There are two major modifications
  - ➢ Perform long-term multiple-step LP based on signals captured by multiple microphone
  - ➢ Direct-path Enhancement(DE)
- To enhance the direct-path response in the processed speech we adjust the delays and calculate the sum of the signal from all the channels.

Experiment in Simulated Reverberant Environment

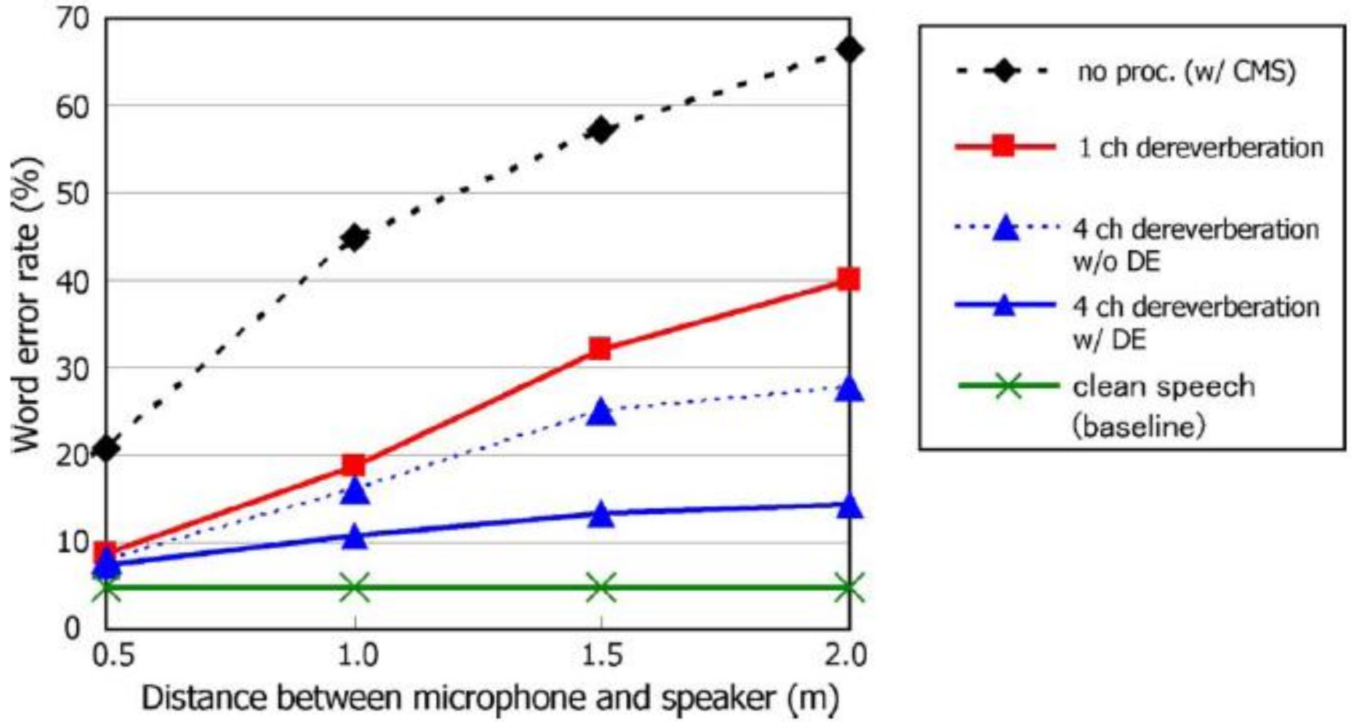
- The Japanese Newspaper Article Sentence(JANS) corpus was used.
- 12 order MFCC+ energy, delta and delta-delta.
- Simulated reverberant environment, where our noise-free assumption holds.


reference microphone
(1.0m, 2.1m, 1.5m)
2.0m
1.5m
1.0m
0.5m
3.5m
Height: 2.5m
4.5m


70
60
50
40
30
20
10
0
Word error rate (%)
0.5
1.0
1.5
2.0
Distance between microphone and speaker (m)
no proc. (w/ CMS)
1 ch dereverberation
4 ch dereverberation w/o DE
4 ch dereverberation w/ DE
clean speech (baseline)

Experiment in Real Reverberant Environment

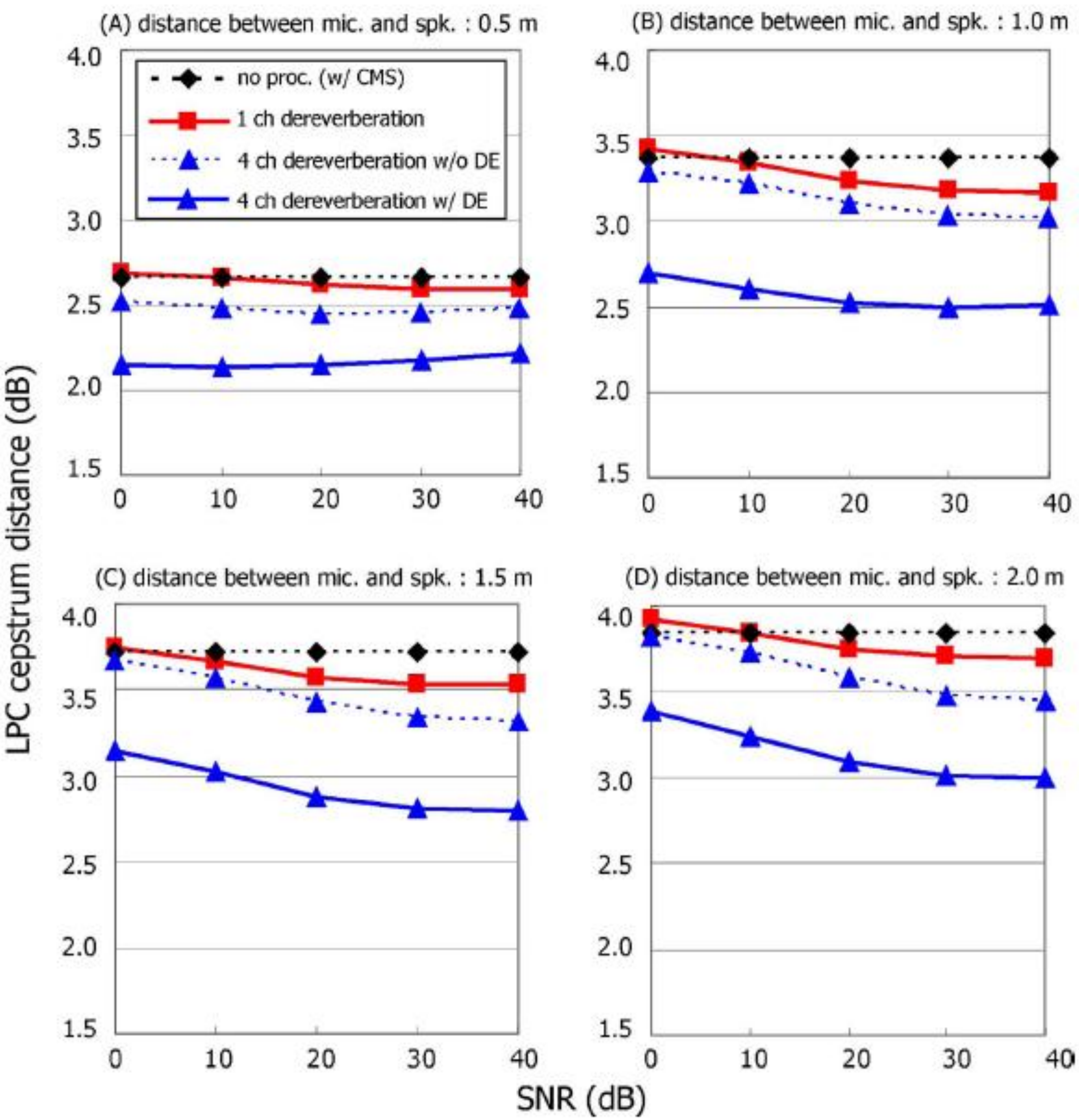
- The recordings were made in a reverberant chamber with same dimension as the simulated room.
- JANS database were played through a BOSE101VM loudspeaker.
- SNRs of the recordings were about 15 to 20 dB.
- After a high pass-filtering, the SNRs about 30 dB.



70
60
50
40
30
20
10
0
Word error rate (%)
0.5
1.0
1.5
2.0
Distance between microphone and speaker (m)
no proc. (w/ CMS)
1 ch dereverberation
4 ch dereverberation w/o DE
4 ch dereverberation w/ DE
clean speech (baseline)

Robustness of Proposed Dereverberation Method to Diffusive Noise

- White noise was artificially generated and added to reverberant speech with SNRs of 0, 10, 20, 30, 40 dB.
- Calculated the LPC cestrum distance between clean speech processed with CMS and the target speech.


(A) distance between mic. and spk. : 0.5 m
(B) distance between mic. and spk. : 1.0 m
(C) distance between mic. and spk. : 1.5 m
(D) distance between mic. and spk. : 2.0 m
no proc. (w/ CMS)
1 ch dereverberation
4 ch dereverberation w/o DE
4 ch dereverberation w/ DE
LPC cepstrum distance (dB)
SNR (dB)