9

A Stereo Acoustic Echo Canceller Using Cross-Channel Correlation

Shigenobu Minami
Toshiba Corporation
Japan

1. Introduction

Stereo acoustic echo canceller is becoming more and more important as an echo canceller is applied to consumer products like a conversational DTV. However it is well known that if there is strong cross-channel correlation between right and left sounds, it cannot converge well and results in echo path estimation misalignment. This is a serious problem in a conversational DTV because the speaker output sound is combination of a far-end conversational sound, which is essentially monaural, and TV program sound, which has wide variety of sound characteristics, monaural sound, stereo sound or mixture of them. To cope with this problem, many stereo echo cancellation algorithms have been proposed. The methods can be categorized into two approaches. The first one is to de-correlate the stereo sound by introducing independent noise or non-linear post-processing to right and left speaker outputs. This approach is very effective for single source stereo sound case, which covers most of conversational sounds, because the de-correlation prevents rank drop to solve a normal equation in a multi-channel adaptive filtering algorithm. Moreover, it is simple since many traditional adaptation algorithms can be used without any modification. Although the approach has many advantages and therefore widely accepted, it still has an essential problem caused by the de-correlation which results in sound quality change due to insertion of the artificial distortion. Even though the inserted distortion is minimized so as to prevent sound quality degradation, from an entertainment audio equipment view point, such as a conversational DTV, many users do not accept any distortion to the speaker output sound. The second approach is desirable for the entertainment types of equipments because no modification to the speaker outputs is required. In this approach, the algorithms utilize cross-channel correlation change in a stereo sound. This approach is also divided into two approaches, depending on how to utilize the cross-channel correlation change. One widely used approach is affine projection method. If there are small variations in the cross-channel correlation even in a single sound source stereo sound, small un-correlated component appears in each channel. The affine projection method can produce the best direction by excluding the auto-correlation bad effect in each channel and by utilizing the small un-correlated components. This approach has a great advantage since it does not require any modification to the stereo sound, however if the variation in the cross-channel correlation is very small, improvement of the adaptive filter convergence is very small. Since the rank drop problem of the stereo adaptive filter is essentially not solved, we may need slight inserted distortion which reduces merit of this method. Another headache is that the method requires P by P inverse matrix calculation in an each sample. The inverse matrix

operation can be relaxed by choosing P as small number, however small P sometimes cannot attain sufficient convergence speed improvement. To attain better performance even by small P, the affine projection method sometimes realized together with sub-band method. Another method categorized in the second approach is "WARP" method. Unlike to affine projection method which utilizes small change in the cross-channel correlation, the method utilizes large change in the cross-channel correlation. This approach is based on the nature of usual conversations. Even though using stereo sound for conversations, most parts of conversations are single talk monaural sound. The cross-channel correlation is usually very high and it remains almost stable during a single talking. A large change happens when talker change or talker's face movement happens. Therefore, the method applies a monaural adaptive filter to single sound source stereo sound and multi-channel (stereo) adaptive filter to non-single sound source stereo sound. Important feature of the method is two monaural adaptive filter estimation results and one stereo adaptive filter estimation result is transformed to each other by using projection matrixes, called WARP matrixes. Since a monaural adaptive filter is applied when a sound is single source stereo sound, we do not need to suffer from the rank problem.

In this chapter, stereo acoustic echo canceller methods, multi-channel least mean square, affine projection and WARP methods, all of them do not need any modification to the speaker output sounds, are surveyed targeting conversational DTV applications. Then WARP method is explained in detail.

2. Stereo acoustic echo canceller problem

2.1 Conversational DTV

Since conversational DTV should keep smooth speech communication even when it is receiving a regular TV program, it requires following functionalities together with traditional DTV systems as shown in Fig. 1.

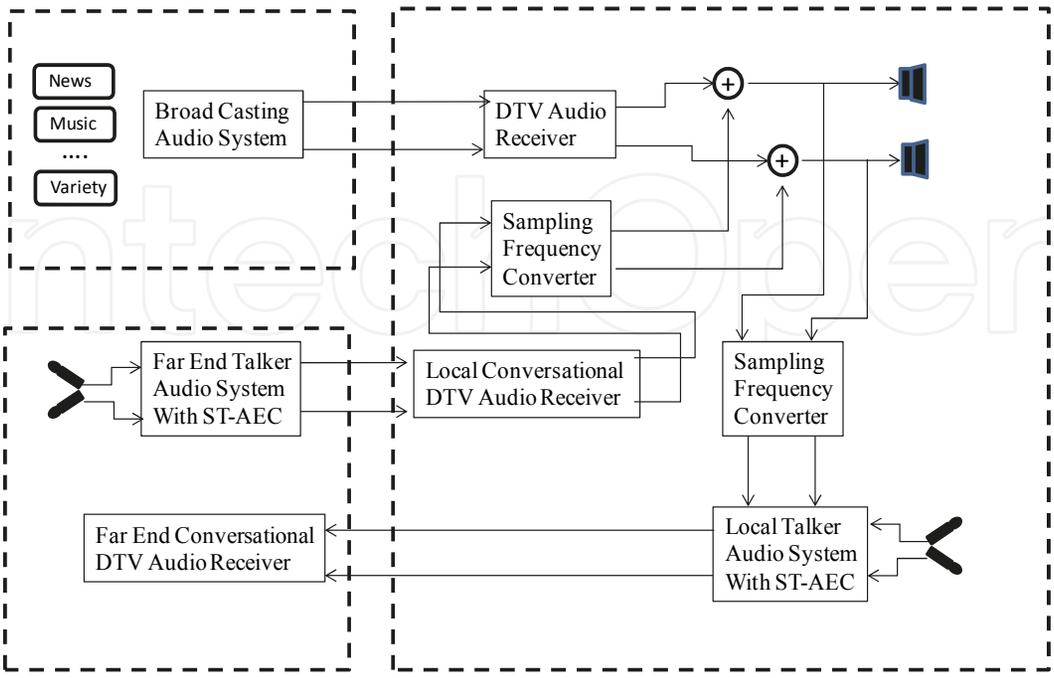


Fig. 1. Audio System Example in a Conversational DTV

1. Mixing of broadcasting sound and communication speech: Two stereo sounds from the DTV audio receiver and local conversational speech decoder are mixed and sent to the stereo speaker system.
2. Sampling frequency conversion: Sampling frequency of DTV sound is usually wider than that of conversational service, such as $f_{SH} = 48kHz$ for DTV sound and $f_S = 16kHz$ for conversational service sound, we need sampling frequency conversion between DTV and conversational service audio parts
3. Stereo acoustic canceller: A stereo acoustic echo canceller is required to prevent howling and speech quality degradation due to acoustic coupling between stereo speaker and microphone.

Among the above special functionalities, the echo canceller for the conversational DTV is technically very challenging because the echo canceller should cancel wide variety of stereo echoes for TV programs as well as stereo speech communications.

2.2 Stereo sound generation model

A stereo acoustic echo canceller system is shown in Fig. 2 with typical stereo sound generation model, where all signals are assumed to be discrete time signals at the kth sampling timing by f_S sampling frequency and the sound generation model is assumed to be linier finite impulse response (FIR) systems which has a sound source signal $x_{Si}(k)$ as an input and stereo sound $x_{Ri}(k)$ and $x_{Li}(k)$ as outputs with additional uncorrelated noises $x'_{URi}(k)$ and $x'_{ULi}(k)$. By using matrix and array notations of the signals as

$$\begin{aligned}
\mathbf{X}_{Si}(k) &= [\mathbf{x}_{Si}(k), \mathbf{x}_{Si}(k-1), \cdots \mathbf{x}_{Si}(k-P+1)] \\
\mathbf{x}_{Si}(k) &= [x_{Si}(k), x_{Si}(k-1), \cdots x_{Si}(k-N+1)]^T \\
\mathbf{x}_{Ri}(k) &= [x_{Ri}(k), x_{Ri}(k-1), \cdots x_{Ri}(k-N+1)]^T \\
\mathbf{x}_{Li}(k) &= [x_{Li}(k), x_{Li}(k-1), \cdots x_{Li}(k-N+1)]^T \\
\mathbf{x}'_{URi}(k) &= [x'_{URi}(k), x'_{URi}(k-1), \cdots x'_{URi}(k-N+1)]^T \\
\mathbf{x}'_{ULi}(k) &= [x'_{ULi}(k), x'_{ULi}(k-1), \cdots x'_{ULi}(k-N+1)]^T
\end{aligned} \tag{1}$$

where P and N are impulse response length of the FIR system and tap length of the adaptive filter for each channel, respectively.

Then the FIR system output $\mathbf{x}_i(k)$ is 2N sample array and is expressed as

$$\mathbf{x}_i(k) = \begin{bmatrix} \mathbf{x}_{Ri}(k) \\ \mathbf{x}_{Li}(k) \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{Si}(k)\mathbf{g}_{Ri}(k) + \mathbf{x}'_{URi}(k) \\ \mathbf{X}_{Si}(k)\mathbf{g}_{Li}(k) + \mathbf{x}'_{ULi}(k) \end{bmatrix}. \tag{2}$$

where $\mathbf{g}_{Ri}(k)$ and $\mathbf{g}_{Li}(k)$ are P sample impulse responses of the FIR system defined as

$$\begin{aligned}
\mathbf{g}_{Ri}(k) &= [g_{Ri,0}(k), g_{Ri1}(k), \cdots, g_{Ri,v}(k), , \cdots, g_{Ri,P-1}(k)]^T \\
\mathbf{g}_{Li}(k) &= [g_{Li,0}(k), g_{Li1}(k), \cdots, g_{Li,v}(k), , \cdots, g_{Li,P-1}(k)]^T
\end{aligned} . \tag{3}$$

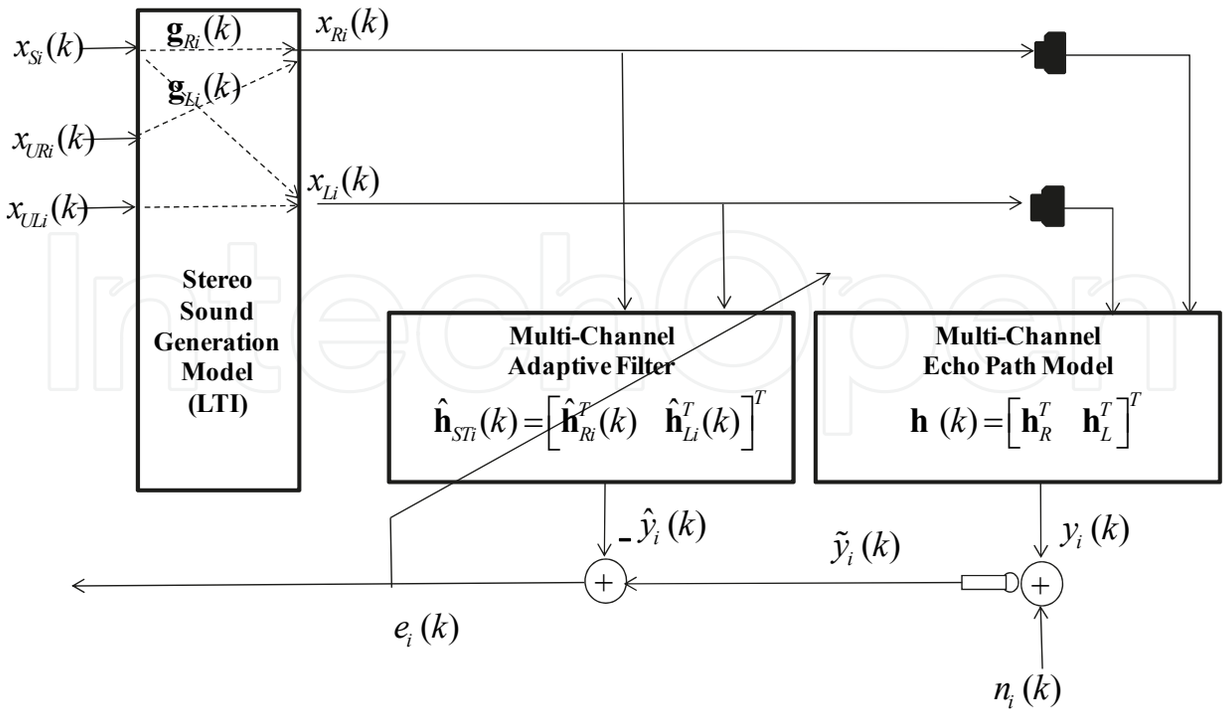


Fig. 2. Traditional Stereo Acoustic Echo Canceller System Configuration with Typical Stereo Sound Generation Model.

In(2), if $\mathbf{g}_{Ri}(k)$ and $\mathbf{g}_{Li}(k)$ are composed of constant array, $\mathbf{g}_{Ri}$ and $\mathbf{g}_{Li}$ during the *ith* period, and small time variant arrays, $\Delta\mathbf{g}_{Ri}(k)$ and $\Delta\mathbf{g}_{Li}(k)$ which are defined as

$$\begin{aligned}
\mathbf{g}_{Ri} &= [g_{Ri,0}, g_{Ri1}, \cdots, g_{Ri,\nu}, \cdots, g_{Ri,P-1}]^T \\
\mathbf{g}_{Li} &= [g_{Li,0}, g_{Li1}, \cdots, g_{Li,\nu}, \cdots, g_{Li,P-1}]^T \\
\Delta\mathbf{g}_{Ri}(k) &= [\Delta g_{Ri,0}(k), \Delta g_{Ri1}(k), \cdots, \Delta g_{Ri,\nu}(k), \cdots, \Delta g_{Ri,P-1}(k)]^T \\
\Delta\mathbf{g}_{Li}(k) &= [\Delta g_{Li,0}(k), \Delta g_{Li1}(k), \cdots, \Delta g_{Li,\nu}(k), \cdots, \Delta g_{Li,P-1}(k)]^T
\end{aligned} . \tag{4}$$

(2) is re-written as

$$\begin{bmatrix} \mathbf{x}_{Ri}(k) \\ \mathbf{x}_{Li}(k) \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{Si}(k)\mathbf{g}_{Ri} + \mathbf{X}_{Si}(k)\Delta\mathbf{g}_{Ri}(k) + \mathbf{x}'_{URi}(k) \\ \mathbf{X}_{Si}(k)\mathbf{g}_{Li} + \mathbf{X}_{Si}(k)\Delta\mathbf{g}_{Li}(k) + \mathbf{x}'_{ULi}(k) \end{bmatrix}. \tag{5}$$

This situation is usual in the case of far-end single talking because transfer functions between talker and right and left microphones vary slightly due to talker's small movement. By assuming the components are also un-correlated noise, (5) can be regarded as a linear time invariant system with independent noise components, $x_{URi}(k)$ and $x_{ULi}(k)$, as

$$\begin{bmatrix} \mathbf{x}_{Ri}(k) \\ \mathbf{x}_{Li}(k) \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{Si}(k)\mathbf{g}_{Ri} + \mathbf{x}_{URi}(k) \\ \mathbf{X}_{Si}(k)\mathbf{g}_{Li} + \mathbf{x}_{ULi}(k) \end{bmatrix}. \tag{6}$$

where

$$\begin{aligned}\mathbf{x}_{URi}(k) &= \mathbf{X}_{Si}(k)\Delta\mathbf{g}_{Ri}(k) + \mathbf{x}'_{URi}(k) \\ \mathbf{x}_{ULi}(k) &= \mathbf{X}_{Si}(k)\Delta\mathbf{g}_{Li}(k) + \mathbf{x}'_{ULi}(k)\end{aligned} . \tag{7}$$

In (6), if there are no un-correlated noises, we call the situation as strict single talking.
In this chapter, sound source signal($x_{Si}(k)$), uncorrelated noises ($x'_{URi}(k)$ and $x'_{ULi}(k)$) are assumed as independent white Gaussian noise with variance σ_{xi} and σ_{Ni} , respectively.

2.3 Stereo acoustic echo canceller problem

For simplification, only one stereo audio echo canceller for the right side microphone's output signal $\tilde{y}_i(k)$, is explained. This is because the echo canceller for left microphone output is apparently treated as the same way as the right microphone case. As shown in Fig.2, the echo canceller cancels the acoustic echo $y_i(k)$ as

$$e_i(k) = y_i(k) - \hat{y}_i(k) + n_i(k) \tag{8}$$

where $e_i(k)$ is acoustic echo canceller's residual error, $n_i(k)$ is a independent background noise, $\hat{y}_i(k)$ is an FIR adaptive filter output in the stereo echo canceller, which is given by

$$\hat{y}_i(k) = \hat{\mathbf{h}}_{Ri}^T(k)\mathbf{x}_{Ri}(k) + \hat{\mathbf{h}}_{Li}^T(k)\mathbf{x}_{Li}(k) \tag{9}$$

where $\hat{\mathbf{h}}_{Ri}(k)$ and $\hat{\mathbf{h}}_{Li}(k)$ are N tap FIR adaptive filter coefficient arrays.
Error power of the echo canceller for the right channel microphone output, $\sigma_{ei}^2(k)$, is given as:

$$\sigma_{ei}^2(k)=(y_{Ri}(k)-\hat{\mathbf{h}}_{STi}^T(k)\mathbf{x}_i(k)+n_i(k))^2 \tag{10}$$

where $\hat{\mathbf{h}}_{STi}(k)$ is a stereo echo path model defined as

$$\hat{\mathbf{h}}_{STi}(k)=\left[\hat{\mathbf{h}}_{Ri}^T(k) \quad \hat{\mathbf{h}}_{Li}^T(k)\right]^T . \tag{11}$$

Optimum echo path estimation $\hat{\mathbf{h}}_{OPT}$ which minimizes the error power $\sigma_e^2(k)$ is given by solving the linier programming problem as

$$Minimize\left[\sum_{k=0}^{N_{LS}-1}\sigma_{ei}^2(k)\right] \tag{12}$$

where N_{LS} is a number of samples used for optimization. Then the optimum echo path estimation for the *ith* LTI period $\hat{\mathbf{h}}_{OPTi}$ is easily obtained by well known normal equation as

$$\hat{\mathbf{h}}_{OPTi}=(\sum_{k=0}^{N_{LS}-1}(\tilde{y}_i(k)\mathbf{x}_i(k)))\mathbf{X}_{NLSi}^{-1} \tag{13}$$

where $\mathbf{X}_{NLSi}$ is an auto-correlation matrix of the adaptive filter input signal and is given by

$$\mathbf{X}_{NLSi} = \sum_{k=0}^{N_{LS}-1} (\mathbf{x}_i(k)\mathbf{x}_i^T(k)) = \begin{bmatrix} \mathbf{A}_i & \mathbf{B}_i \\ \mathbf{C}_i & \mathbf{D}_i \end{bmatrix} = \begin{bmatrix} \sum_{k=0}^{N_{LS}-1} (\mathbf{x}_{Ri}(k)\mathbf{x}_{Ri}^T(k)) & \sum_{k=0}^{N_{LS}-1} (\mathbf{x}_{Ri}(k)\mathbf{x}_{Li}^T(k)) \\ \sum_{k=0}^{N_{LS}-1} (\mathbf{x}_{Li}(k)\mathbf{x}_{Ri}^T(k)) & \sum_{k=0}^{N_{LS}-1} (\mathbf{x}_{Li}(k)\mathbf{x}_{Li}^T(k)) \end{bmatrix}. \quad (14)$$

By (14), determinant of $\mathbf{X}_{NLSi}$ is given by

$$\left|\mathbf{X}_{NLSi}\right| = \left|\mathbf{A}_i\right|\left|\mathbf{D}_i - \mathbf{C}_i\mathbf{A}_i^{-1}\mathbf{B}_i\right| . \quad (15)$$

In the case of the stereo generation model which is defined by(2), the sub-matrixes in (14) are given by

$$\begin{aligned}
\mathbf{A}_i &= \sum_{k=0}^{N_{LS}-1} (\mathbf{X}_{Si}(k)\mathbf{G}_{RRi}\mathbf{X}_{Si}^T(k) + 2\mathbf{x}_{URi}(k)(\mathbf{X}_{Si}(k)\mathbf{g}_{Ri})^T + \mathbf{x}_{URi}(k)\mathbf{x}_{URi}^T(k)) \\
\mathbf{B}_i &= \sum_{k=0}^{N_{LS}-1} (\mathbf{X}_{Si}(k)\mathbf{G}_{RLi}\mathbf{X}_{Si}^T(k) + \mathbf{x}_{URi}(k)(\mathbf{X}_{Si}(k)\mathbf{g}_{Ri})^T + \mathbf{x}_{ULi}(k)(\mathbf{X}_{Si}(k)\mathbf{g}_{Ri})^T + \mathbf{x}_{URi}(k)\mathbf{x}_{ULi}^T(k)) \\
\mathbf{C}_i &= \sum_{k=0}^{N_{LS}-1} (\mathbf{X}_{Si}(k)\mathbf{G}_{LRi}\mathbf{X}_{Si}^T(k) + \mathbf{x}_{ULi}(k)(\mathbf{X}_{Si}(k)\mathbf{g}_{Ri})^T + \mathbf{x}_{URi}(k)(\mathbf{X}_{Si}(k)\mathbf{g}_{Li})^T + \mathbf{x}_{ULi}(k)\mathbf{x}_{URi}^T(k)) \\
\mathbf{D}_i &= \sum_{k=0}^{N_{LS}-1} (\mathbf{X}_{Si}(k)\mathbf{G}_{LLi}\mathbf{X}_{Si}^T(k) + 2\mathbf{x}_{ULi}(k)(\mathbf{X}_{Si}(k)\mathbf{g}_{Li})^T + \mathbf{x}_{ULi}(k)\mathbf{x}_{ULi}^T(k))
\end{aligned} \quad .(16)$$

where

$$\mathbf{G}_{RRi} = \mathbf{g}_{Ri}\mathbf{g}_{Ri}^T, \mathbf{G}_{RLi} = \mathbf{g}_{Ri}\mathbf{g}_{Li}^T, \mathbf{G}_{LRi} = \mathbf{g}_{Li}\mathbf{g}_{Ri}^T, \mathbf{G}_{LLi} = \mathbf{g}_{Li}\mathbf{g}_{Li}^T . \quad (17)$$

In the cease of strict single talking where $\mathbf{x}_{URi}(k)$ and $\mathbf{x}_{ULi}(k)$ do not exist, (16) becomes very simple as

$$\begin{aligned}
\mathbf{A}_i &= \sum_{k=0}^{N_{LS}-1} (\mathbf{X}_{Si}(k)\mathbf{G}_{RRi}\mathbf{X}_{Si}^T(k)) \\
\mathbf{B}_i &= \sum_{k=0}^{N_{LS}-1} (\mathbf{X}_{Si}(k)\mathbf{G}_{RLi}\mathbf{X}_{Si}^T(k)) \\
\mathbf{C}_i &= \sum_{k=0}^{N_{LS}-1} (\mathbf{X}_{Si}(k)\mathbf{G}_{LRi}\mathbf{X}_{Si}^T(k)) \\
\mathbf{D}_i &= \sum_{k=0}^{N_{LS}-1} (\mathbf{X}_{Si}(k)\mathbf{G}_{LLi}\mathbf{X}_{Si}^T(k))
\end{aligned} . \quad (18)$$

To check the determinant $\left|\mathbf{X}_{NLSi}\right|$, we calculate $\left|\mathbf{X}_{NLSi}\right|\left|\mathbf{C}_i\right|$ considering $\mathbf{B}_i = \mathbf{C}_i^T$ as

$$\begin{aligned}\left|\mathbf{X}_{NLSi}\right|\left|\mathbf{C}_i\right| &= \left|\mathbf{A}_i\right|\left|(\mathbf{D}_i\mathbf{C}_i - \mathbf{C}_i\mathbf{A}_i^{-1}\mathbf{B}_i\mathbf{C}_i\right| \\ &= \left|\mathbf{A}_i\right|\left|(\mathbf{D}_i\mathbf{C}_i - \mathbf{C}_i\mathbf{B}_i\mathbf{C}_i\mathbf{A}_i^{-1}\right|\end{aligned} . \tag{19}$$

Then $\left|\mathbf{D}_i\mathbf{C}_i - \mathbf{C}_i\mathbf{B}_i\mathbf{C}_i\mathbf{A}_i^{-1}\right|$ becomes zero as

$$\begin{aligned}&\left|\mathbf{D}_i\mathbf{C}_i - \mathbf{C}_i\mathbf{A}_i^{-1}\mathbf{B}_i\mathbf{C}_i\right| \\ &= \left|\sum_{k=0}^{N_{LS}-1}(\mathbf{X}_{Si}(k)(\mathbf{G}_{LLi} - \mathbf{G}_{LRi}\mathbf{G}^{-1}{}_{RRi}\mathbf{G}_{RLi})\mathbf{X}_{Si}^T(k))\mathbf{X}_{Si}(k)\mathbf{G}_{LRi}\mathbf{X}_{Si}^T(k))\right| \\ &= N\sigma_{xi}^2\left|\sum_{k=0}^{N_{LS}-1}(\mathbf{X}_{Si}(k)(\mathbf{g}_{Li}^T\mathbf{g}_{Li}(\mathbf{g}_{Li}\mathbf{g}_{Ri}^T - \mathbf{g}_{Li}\mathbf{g}_{Ri}^T(\mathbf{g}_{Ri}\mathbf{g}_{Ri}^T)^{-1}\mathbf{g}_{Ri}\mathbf{g}_{Ri}^T)\mathbf{X}_{Si}^T(k))\right| \\ &= 0\end{aligned} . \tag{20}$$

Hence no unique solution can be found by solving the normal equation in the case of strict single talking where un-correlated components do not exist. This is a well known stereo adaptive filter cross-channel correlation problem.

3. Stereo acoustic echo canceller methods

To improve problems addressed above, many approaches have been proposed. One widely accepted approach is de-correlation of stereo sound. To avoid the rank drop of the normal equation(13), small distortion such as non-linear processing or modification of phase is added to stereo sound. This approach is simple and effective to endorse convergence of the multi-channel adaptive filter, however it may degrade the stereo sound by the distortion. In the case of entertainment applications, such as conversational DTV, the problem may be serious because customer's requirement for sound quality is usually very high and therefore even small modification to the speaker output sound cannot be accepted. From this view point, approaches which do not need to add any modification or artifacts to the speaker output sound are desirable for the entertainment use. In this section, least square (LS), stereo affine projection (AP), stereo normalized least mean square (NLMS) and WARP methods are reviewed as methods which do not need to change stereo sound itself.

3.1 Gradient method

Gradient method is widely used for solving the quadratic problem iteratively. As a generalized gradient method, let denote M sample orthogonalized error array $\boldsymbol{\varepsilon}_{Mi}(\mathrm{k})$ based on original error array $\mathbf{e}_{Mi}(\mathrm{k})$ as

$$\boldsymbol{\varepsilon}_{Mi}(\mathrm{k}) = \mathbf{R}_i(k)\mathbf{e}_{Mi}(\mathrm{k}) \tag{21}$$

where $\mathbf{e}_{Mi}(\mathrm{k})$ is an M sample error array which is defined as

$$\mathbf{e}_{Mi}(\mathrm{k}) = [e_i(k), e_i(k-1), \cdots e_i(k-M+1)]^T \tag{22}$$

and $\mathbf{R}_i(k)$ is a $M \times M$ matrix which orthogonalizes the auto-correlation matrix $\mathbf{e}_{Mi}(\mathrm{k})\mathbf{e}_{Mi}^T(\mathrm{k})$. The orthogonalized error array is expressed using difference between adaptive filter coefficient array $\hat{\mathbf{h}}_{STi}(\mathrm{k})$ and target stereo echo path $2N$ sample response $\mathbf{h}_{ST}$ as

$$\boldsymbol{\varepsilon}_{Mi}(\mathrm{k}) = \mathbf{R}_i(k)\mathbf{X}_{M2Ni}^T(k)(\mathbf{h}_{ST}\text{-}\hat{\mathbf{h}}_{STi}(\mathrm{k})) \tag{23}$$

where $\mathbf{X}_{M2Ni}(k)$ is a Mx2N matrix which is composed of adaptive filter stereo input array as defined by

$$\mathbf{X}_{M2Ni}(k) = [\mathbf{x}_i(k), \mathbf{x}_i(k-1), \cdots \mathbf{x}_i(k-M+1)]. \tag{24}$$

By defining an echo path estimation error array $\mathbf{d}_{STi}(k)$ which is defined as

$$\mathbf{d}_{STi}(k)=\mathbf{h}_{ST}\text{-}\hat{\mathbf{h}}_{STi}(\mathrm{k}) \tag{25}$$

estimation error power $\sigma_{\varepsilon i}^2(\mathrm{k})$ is obtained by

$$\sigma_{\varepsilon i}^2(\mathrm{k}) = \boldsymbol{\varepsilon}_{Mi}^T(\mathrm{k})\boldsymbol{\varepsilon}_{Mi}(\mathrm{k})=\mathbf{d}_{STi}^T(\mathrm{k})\mathbf{Q}_{2N2Ni}(k)\mathbf{d}_{STi}(\mathrm{k}) \tag{26}$$

where

$$\mathbf{Q}_{2N2Ni}(k) = \mathbf{X}_{M2Ni}(k)\mathbf{R}_i^T(k)\mathbf{R}_i(k)\mathbf{X}_{M2Ni}^T(k). \tag{27}$$

Then, (26) is regarded as a quadratic function of $\hat{\mathbf{h}}_{STi}(k)$ as

$$f(\hat{\mathbf{h}}_{STi}(k)) = \frac{1}{2}\hat{\mathbf{h}}_{STi}^T(\mathrm{k})\mathbf{Q}_{2N2Ni}(k)\hat{\mathbf{h}}_{STi}^T(\mathrm{k}) - \hat{\mathbf{h}}_{STi}^T(\mathrm{k})\mathbf{Q}_{2N2Ni}\mathbf{h}_{ST}. \tag{28}$$

For the quadratic function, gradient $\boldsymbol{\Delta}_i(k)$ is given by

$$\boldsymbol{\Delta}_i(k)=\text{-}\mathbf{Q}_{2N2Ni}(k)\mathbf{d}_{STi}(\mathrm{k}). \tag{29}$$

Iteration of $\hat{\mathbf{h}}_{STi}(k)$ which minimizes $\sigma_{\varepsilon i}^2(\mathrm{k})$ is given by

$$\begin{aligned}\hat{\mathbf{h}}_{STi}(k+1)&=\hat{\mathbf{h}}_{STi}(k) - \alpha\boldsymbol{\Delta}_i(k)\\ &= \hat{\mathbf{h}}_{STi}(k) + \alpha\mathbf{Q}_{2N2Ni}(k)\mathbf{d}_{STi}(\mathrm{k})\\ &=\hat{\mathbf{h}}_{STi}(k) + \alpha\mathbf{X}_{M2Ni}(k)\mathbf{R}_i^T(k)\mathbf{R}_i(k)\mathbf{e}_{Mi}(\mathrm{k})\end{aligned} \tag{30}$$

where α is a constant to determine step size.
Above equation is very generic expression of the gradient method and following approaches are regarded as deviations of this iteration.

3.2 Least Square (LS) method (M=2N)

From(30), the estimation error power between estimated adaptive filter coefficients and stereo echo path response , $\mathbf{d}_i^T(k)\mathbf{d}_i(k)$ is given by

$$\mathbf{d}_i^T(k+1)\mathbf{d}_i(k+1)=\mathbf{d}_i^T(k)(\mathbf{I}_{2N}-\alpha\mathbf{Q}_{2N2Ni}(k))(\mathbf{I}_{2N}-\alpha\mathbf{Q}_{2N2Ni}(k))^T\mathbf{d}_i(k) \tag{31}$$

where $\mathbf{I}_{2N}$ is a $2N\times 2N$ identity matrix. Then the fastest convergence is obtained by finding $\mathbf{R}_i(k)$ which orthogonalizes and minimizes eigenvalue variance in $\mathbf{Q}_{2N2Ni}(k)$.

If M=2N, $\mathbf{X}_{2M2Ni}(k)$ is symmetric square matrix as

$$\mathbf{X}_{M2Ni}(k)=\mathbf{X}_{M2Ni}^T(k) \tag{32}$$

and if $\mathbf{X}_{M2Ni}(k)\cdot\mathbf{X}_{M2Ni}^T(k)(=\mathbf{X}_{M2Ni}^T(k)\cdot\mathbf{X}_{M2Ni}(k))$ is a regular matrix so that inverse matrix exists, $\mathbf{R}_i^T(k)\mathbf{R}_i(k)$ which orthogonalizes $\mathbf{Q}_{2N2Ni}(k)$ is given by

$$\mathbf{R}_i^T(k)\mathbf{R}_i(k)=(\mathbf{X}_{2N2Ni}^T(k)\mathbf{X}_{2N2Ni}(k))^{-1} \tag{33}$$

By substituting (33) for (30)

$$\hat{\mathbf{h}}_{STi}(k+1)=\hat{\mathbf{h}}_{STi}(k)+\alpha\mathbf{X}_{2N2Ni}(k)(\mathbf{X}_{2N2Ni}^T(k)\mathbf{X}_{2N2Ni}(k))^{-1}\mathbf{e}_{Ni}(\mathrm{k}) \tag{34}$$

Assuming initial tap coefficient array as zero vector and $\alpha=0$ during 0 to 2N-1th samples and $\alpha=1$ at 2Nth sample , (34) can be re-written as

$$\hat{\mathbf{h}}_{STi}(2N)=\mathbf{X}_{2N2Ni}(\text{2N-1})(\mathbf{X}_{2N2Ni}^T(\text{2N-1})\mathbf{X}_{2N2Ni}(\text{2N-1}))^{-1}\mathbf{y}_i(\text{2N-1}) \tag{35}$$

where $\mathbf{y}_i(\mathrm{k})$ is $2N$ sample echo path output array and is defined as

$$\mathbf{y}_i(\mathrm{k})=[y_i(k),y_i(k-1),\cdots y_i(k-2N+1)]^T \tag{36}$$

This iteration is done only once at $2N-1th$ sample. If $N_{LS}=2N$, inverse matrix term in (35) is written as

$$\mathbf{X}_{2N2Ni}^T(k)\mathbf{X}_{2N2Ni}(k)=\sum_{k=0}^{N_{LS}-1}(\mathbf{x}_i(k)\mathbf{x}_i^T(k))=\mathbf{X}_{NLSi} \tag{37}$$

Comparing (13) and (35) with (37), it is found that LS method is a special case of gradient method when M equals to 2N.

3.3 Stereo Affine Projection (AP) method (M=P ≤ N)

Stereo affine projection method is assumed as a case when M is chosen as FIR response length P in the LTI system. This approach is very effective to reduce 2Nx2N inverse matrix operations in LS method to PxP operations when the stereo generation model is assumed to be LTI system outputs from single WGN signal source with right and left channel independent noises as shown in Fig.2. For the sake of explanation, we define stereo sound signal matrix $\mathbf{X}_{P2Ni}(k)$ which is composed of right and left signal matrix $\mathbf{X}_{Ri}(k)$ and $\mathbf{X}_{Li}(k)$ for P samples as

$$\mathbf{X}_{P2Ni}(k)=\begin{bmatrix}\mathbf{X}_{Ri}^T(k) & \mathbf{X}_{Li}^T(k)\end{bmatrix}^T=\begin{bmatrix}\mathbf{X}_{2Si}(k)\mathbf{G}_{Ri}^T+\mathbf{X}_{URi}(k)\\ \mathbf{X}_{2Si}(k)\mathbf{G}_{Li}^T+\mathbf{X}_{ULi}(k)\end{bmatrix} \tag{38}$$

where

$$\mathbf{X}_{2Si}(k)=[\mathbf{x}_{Si}(k),\mathbf{x}_{Si}(k-1),\cdots\mathbf{x}_{Si}(k-2P+2)] \tag{39}$$

$\mathbf{X}_{URi}(k)$ and $\mathbf{X}_{ULi}(k)$ are un-correlated signal matrix defined as

$$\begin{aligned}\mathbf{X}_{URi}(k)&=[\mathbf{x}_{URi}(k),\mathbf{x}_{URi}(k-1),\cdots\mathbf{x}_{URi}(k-P+1)]\\ \mathbf{X}_{ULi}(k)&=[\mathbf{x}_{ULi}(k),\mathbf{x}_{ULi}(k-1),\cdots\mathbf{x}_{ULi}(k-P+1)]\end{aligned} \tag{40}$$

$\mathbf{G}_{Ri}$ and $\mathbf{G}_{Li}$ are source to microphones response (2P-1)xP matrixes and are defined as

$$\mathbf{G}_{Ri}=\begin{bmatrix}\mathbf{g}_{2R,0,i}^T\\ \mathbf{g}_{2R,1,i}^T\\ \cdots\\ \mathbf{g}_{2R,P-1,i}^T\end{bmatrix}=\begin{bmatrix}\mathbf{g}_{Ri}^T & 0 & \cdots & 0\\ 0 & \mathbf{g}_{Ri}^T & \ddots & 0\\ 0 & 0 & \cdots & 0\\ 0 & 0 & \cdots & \mathbf{g}_{Ri}^T\end{bmatrix},\mathbf{G}_{Li}=\begin{bmatrix}\mathbf{g}_{2RL,0,i}^T\\ \mathbf{g}_{2L,1,i}^T\\ \cdots\\ \mathbf{g}_{2L,P-1,i}^T\end{bmatrix}=\begin{bmatrix}\mathbf{g}_{Li}^T & 0 & \cdots & 0\\ 0 & \mathbf{g}_{Li}^T & \ddots & 0\\ 0 & 0 & \cdots & 0\\ 0 & 0 & \cdots & \mathbf{g}_{Li}^T\end{bmatrix}. \tag{41}$$

As explained by(31), $\mathbf{Q}_{2N2Ni}(k)$ determines convergence speed of the gradient method. In this section, we derive affine projection method by minimizing the max-min eigenvalue variance in $\mathbf{Q}_{2N2Ni}(k)$. Firstly, the auto-correlation matrix is expressed by sub-matrixes for each stereo channel as

$$\mathbf{Q}_{N2Ni}(k)=\begin{bmatrix}\mathbf{Q}_{ANNi}(k) & \mathbf{Q}_{BNNi}(k)\\ \mathbf{Q}_{CNNi}(k) & \mathbf{Q}_{DNNi}(k)\end{bmatrix} \tag{42}$$

where $\mathbf{Q}_{ANNi}(k)$ and $\mathbf{Q}_{DNNi}(k)$ are right and left channel auto-correlation matrixes, $\mathbf{Q}_{BNNi}(k)$ and $\mathbf{Q}_{CNNi}(k)$ are cross channel-correlation matrixes. These sub-matrixes are given by

$$\begin{aligned}\mathbf{Q}_{ANNi}(k)&=\mathbf{X}_{2Si}(k)\mathbf{G}_{Ri}^T\mathbf{R}_i^T(k)\mathbf{R}_i(k)\mathbf{G}_{Ri}\mathbf{X}_{2Si}^T(k)+\mathbf{X}_{URi}(k)\mathbf{R}_i^T(k)\mathbf{R}_i(k)\mathbf{X}_{URi}^T(k)\\ &+2\mathbf{X}_{2Si}(k)\mathbf{R}_i^T(k)\mathbf{R}_i(k)\mathbf{X}_{URi}^T(k)\\ \mathbf{Q}_{BNNi}(k)&=\mathbf{X}_{2Si}(k)\mathbf{G}_{Ri}^T\mathbf{R}_i^T(k)\mathbf{R}_i(k)\mathbf{G}_{Li}\mathbf{X}_{2Si}^T(k)+\mathbf{X}_{URi}(k)\mathbf{R}_i^T(k)\mathbf{R}_i(k)\mathbf{X}_{ULi}^T(k)\\ &+2\mathbf{X}_{URi}(k)\mathbf{R}_i^T(k)\mathbf{R}_i(k)\mathbf{X}_{ULi}^T(k)\\ \mathbf{Q}_{CNNi}(k)&=\mathbf{X}_{2Si}(k)\mathbf{G}_{Li}^T\mathbf{R}_i^T(k)\mathbf{R}_i(k)\mathbf{G}_{Ri}\mathbf{X}_{2Si}^T(k)+\mathbf{X}_{ULi}(k)\mathbf{R}_i^T(k)\mathbf{R}_i(k)\mathbf{X}_{URi}^T(k)\\ &+2\mathbf{X}_{ULi}(k)\mathbf{R}_i^T(k)\mathbf{R}_i(k)\mathbf{X}_{UTi}^T(k)\\ \mathbf{Q}_{DNNi}(k)&=\mathbf{X}_{2Si}(k)\mathbf{G}_{Li}^T\mathbf{R}_i^T(k)\mathbf{R}_i(k)\mathbf{G}_{Li}\mathbf{X}_{2Si}^T(k)+\mathbf{X}_{ULi}(k)\mathbf{R}_i^T(k)\mathbf{R}_i(k)\mathbf{X}_{ULi}^T(k)\\ &+2\mathbf{X}_{2Si}(k)\mathbf{R}_i^T(k)\mathbf{R}_i(k)\mathbf{X}_{ULi}^T(k)\end{aligned} \tag{43}$$

Since the iteration process in (30) is an averaging process, the auto-correlation matrix $\mathbf{Q}_{2N2Ni}(k)$ is approximated by using expectation value of it, $\tilde{\mathbf{Q}}_{2N2Ni}(k)=\langle\mathbf{Q}_{2N2Ni}(k)\rangle$. Then expectation values for sub-matrixes in (42) are simplified applying statistical independency between sound source signal and noises and Tlz function defined in Appendix as

$$\begin{aligned}
\tilde{\mathbf{Q}}_{ANNi} &= Tlz(\langle \tilde{\mathbf{X}}_{2Si}(k)\mathbf{G}_{Ri}^T\mathbf{R}_i^T(k)\mathbf{R}_i(k)\mathbf{G}_{Ri}\tilde{\mathbf{X}}_{2Si}^T(k)\rangle) + Tlz(\langle \tilde{\mathbf{X}}_{URi}(k)\mathbf{R}_i^T(k)\mathbf{R}_i(k)\tilde{\mathbf{X}}_{URi}(k)\rangle) \\
\tilde{\mathbf{Q}}_{BNNi} &= Tlz(\langle \tilde{\mathbf{X}}_{2Si}(k)\mathbf{G}_{Ri}^T\mathbf{R}_i^T(k)\mathbf{R}_i(k)\mathbf{G}_{Li}\tilde{\mathbf{X}}_{2Si}^T(k)\rangle) \\
\tilde{\mathbf{Q}}_{CNNi} &= Tlz(\langle \tilde{\mathbf{X}}_{2Si}(k)\mathbf{G}_{Li}^T\mathbf{R}_i^T(k)\mathbf{R}_i(k)\mathbf{G}_{Ri}\tilde{\mathbf{X}}_{2Si}^T(k)\rangle) \\
\tilde{\mathbf{Q}}_{DNNi} &= Tlz(\langle \tilde{\mathbf{X}}_{2Si}(k)\mathbf{G}_{Li}^T\mathbf{R}_i^T(k)\mathbf{R}_i(k)\mathbf{G}_{Li}\tilde{\mathbf{X}}_{2Si}(k)\rangle) + Tlz(\langle \tilde{\mathbf{X}}_{ULi}(k)\mathbf{R}_i^T(k)\mathbf{R}_i(k)\tilde{\mathbf{X}}_{ULi}(k)\rangle)
\end{aligned} \tag{44}$$

where

$$\begin{aligned}
\tilde{\mathbf{X}}_{2Si}(k) &= [\tilde{\mathbf{x}}_{2Si}(k), \tilde{\mathbf{x}}_{2Si}(k-1), \cdots \tilde{\mathbf{x}}_{2Si}(k-P+1)]^T \\
\tilde{\mathbf{X}}_{URi}(k) &= [\tilde{\mathbf{x}}_{URi}(k), \tilde{\mathbf{x}}_{URi}(k-1), \cdots \tilde{\mathbf{x}}_{URi}(k-P+1)] \\
\tilde{\mathbf{X}}_{ULi}(k) &= [\tilde{\mathbf{x}}_{ULi}(k), \tilde{\mathbf{x}}_{ULi}(k-1), \cdots \tilde{\mathbf{x}}_{ULi}(k-P+1)]
\end{aligned} \tag{45}$$

with

$$\begin{aligned}
\tilde{\mathbf{x}}_{2Si}(k) &= [x_{Si}(k), x_{Si}(k-1), \cdots x_{Si}(k-2p+2)]^T \\
\tilde{\mathbf{x}}_{URi}(k) &= [x_{URi}(k), x_{URi}(k-1), \cdots x_{URi}(k-p+1)]^T \\
\tilde{\mathbf{x}}_{ULi}(k) &= [x_{ULi}(k), x_{ULi}(k-1), \cdots x_{ULi}(k-p+1)]^T
\end{aligned} . \tag{46}$$

Applying matrix operations to $\mathbf{Q}_{2N2Ni}$, a new matrix $\mathbf{Q}'_{2N2Ni}$ which has same determinant as $\mathbf{Q}_{2N2Ni}$ is given by

$$\mathbf{Q}'_{2N2Ni}(k) = \begin{bmatrix} \mathbf{Q}'_{ANNi}(k) & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}'_{DNNi}(k) \end{bmatrix} \tag{47}$$

where

$$\mathbf{Q}'_{ANNi} = Tlz(\mathbf{Q}''_{ANNi}), \mathbf{Q}'_{DNNi} = Tlz(\mathbf{Q}''_{DNNi}) . \tag{48}$$

Since both $\tilde{\mathbf{X}}_{2Si}(k)\mathbf{G}_{Ri}^T$ and $\tilde{\mathbf{X}}_{2Si}(k)\mathbf{G}_{Li}^T$ are symmetric PxP square matrixes, $\mathbf{Q}''_{ANNi}$ and $\mathbf{Q}''_{BNNi}$ are re-written as

$$\begin{aligned}
\mathbf{Q}''_{ANNi} &= \langle \tilde{\mathbf{X}}_{2Si}(k)\mathbf{G}_{Ri}^T\mathbf{R}_i^T(k)\mathbf{R}_i(k)\mathbf{G}_{Ri}\tilde{\mathbf{X}}_{2Si}^T(k) + \tilde{\mathbf{X}}_{2Si}(k)\mathbf{G}_{Li}^T\mathbf{R}_i^T(k)\mathbf{R}_i(k)\mathbf{G}_{Li}\tilde{\mathbf{X}}_{2Si}^T(k)\rangle \\
&+ \langle \tilde{\mathbf{X}}_{URi}(k)\mathbf{R}_i^T(k)\mathbf{R}_i(k)\tilde{\mathbf{X}}_{URi}^T(k)\rangle \\
&= \langle \tilde{\mathbf{X}}_{2Si}(k)(\mathbf{G}_{Ri}^T\mathbf{G}_{Ri} + \mathbf{G}_{Li}^T\mathbf{G}_{Li})\tilde{\mathbf{X}}_{2Si}^T(k)\rangle \mathbf{R}_i^T(k)\mathbf{R}_i(k) + \langle \tilde{\mathbf{X}}_{URi}(k)\tilde{\mathbf{X}}_{URi}^T(k)\rangle \mathbf{R}_i^T(k)\mathbf{R}_i(k) . \\
&= (N\sigma_{Xi}^2(\mathbf{G}_{Ri}\mathbf{G}_{Ri}^T + \mathbf{G}_{Li}\mathbf{G}_{Li}^T) + N\sigma_{Ni}^2\mathbf{I}_P)\mathbf{R}_i^T(k)\mathbf{R}_i(k) \\
\mathbf{Q}''_{DNNi} &= N\sigma_{Ni}^2\mathbf{I}_P\mathbf{R}_i^T(k)\mathbf{R}_i(k)
\end{aligned} \tag{49}$$

As evident by(47), (48) and(49), $\mathbf{Q}'_{2N2Ni}(k)$ is composed of major matrix $\mathbf{Q}'_{ANNi}(k)$ and noise matrix $\mathbf{Q}'_{DNNi}(k)$. In the case of single talking where sound source signal power σ_X^2 is much

larger than un-correlated signal power σ_{Ni}^2, $\mathbf{R}_i^T(k)\mathbf{R}_i(k)$ which minimizes eigenvalue spread in $\mathbf{Q}_{2N2Ni}(k)$ so as to attain the fastest convergence is given by making $\mathbf{Q}''_{ANNi}$ as a identity matrix by setting $\mathbf{R}_i^T(k)\mathbf{R}_i(k)$ as

$$\mathbf{R}_i^T(k)\mathbf{R}_i(k) \simeq (N\sigma_{Xi}^2(\mathbf{G}_{Ri}\mathbf{G}_{Ri}^T + \mathbf{G}_{Li}\mathbf{G}_{Li}^T))^{-1} \tag{50}$$

In other cases such as double talking or no talk situations, where we assume σ_X^2 is almost zero, $\mathbf{R}_i^T(k)\mathbf{R}_i(k)$ which orthogonalizes $\mathbf{Q}''_{ANNi}$ is given by

$$\mathbf{R}_i^T(k)\mathbf{R}_i(k) \simeq (N\sigma_{Ni}^2\mathbf{I}_P)^{-1} \tag{51}$$

Summarizing the above discussions, the fastest convergence is attained by setting $\mathbf{R}_i^T(k)\mathbf{R}_i(k)$ as

$$\mathbf{R}_i^T(k)\mathbf{R}_i(k) = \left(\mathbf{X}_{P2Ni}^T(k)\mathbf{X}_{P2Ni}(k)\right)^{-1}. \tag{52}$$

Since

$$\begin{aligned}
&\left\langle \mathbf{X}_{P2Ni}^T(k)\mathbf{X}_{P2Ni}(k)\right\rangle = \\
&\left\langle \begin{bmatrix} \mathbf{G}_{Ri}\mathbf{X}_{2Si}^T(k) + \mathbf{X}_{URi}^T(k) & \mathbf{G}_{Li}\mathbf{X}_{2Si}^T(k) + \mathbf{X}_{ULi}^T(k)\end{bmatrix} \begin{bmatrix} \mathbf{X}_{2Si}(k)\mathbf{G}_{Ri}^T + \mathbf{X}_{URi}(k) \\ \mathbf{X}_{2Si}(k)\mathbf{G}_{Li}^T + \mathbf{X}_{ULi}(k)\end{bmatrix} \right\rangle \\
&= \left\langle \mathbf{G}_{Ri}\mathbf{X}_{2Si}^T(k)\mathbf{X}_{2Si}(k)\mathbf{G}_{Ri}^T + \mathbf{G}_{Li}\mathbf{X}_{2Si}^T(k)\mathbf{X}_{2Si}(k)\mathbf{G}_{Li}^T + \mathbf{X}_{URi}^T(k)\mathbf{X}_{URi}(k) + \mathbf{X}_{ULi}^T(k)\mathbf{X}_{ULi}(k)\right\rangle \\
&\simeq N\sigma_{Xi}^2(\mathbf{G}_{Ri}\mathbf{G}_{Ri}^T + \mathbf{G}_{Li}\mathbf{G}_{Li}^T) + 2N\sigma_{Ni}^2\mathbf{I}_P
\end{aligned} \tag{53}$$

By substituting (52) for (30), we obtain following affine projection iteration :

$$\hat{\mathbf{h}}_{STi}(k+1) = \hat{\mathbf{h}}_{STi}(k) + \alpha\mathbf{X}_i(k)(\mathbf{X}_{P2Ni}^T(k)\mathbf{X}_{P2Ni}(k))^{-1}\mathbf{e}_{Pi}(\mathrm{k}). \tag{54}$$

In an actual implementation α is replaced by $\boldsymbol{\mu}$ for forgetting factor and $\delta\mathrm{I}$ is added to the inverse matrix to avoid zero division as shown bellow.

$$\hat{\mathbf{h}}_{ST}(k+1) = \hat{\mathbf{h}}_{ST}(k) + \alpha\mathbf{X}_{P2Ni}(\mathrm{k})[\mathbf{X}_{P2Ni}^T(\mathrm{k})\mathbf{X}_{P2Ni}(\mathrm{k}) + \delta\mathrm{I}]^{-1}\boldsymbol{\mu}\mathbf{e}_{Pi}(k) \tag{55}$$

where $\delta(\ll 1)$ is very small positive value and

$$\boldsymbol{\mu} = diag[1,(1-\mu),\cdots,(1-\mu)^{p-1}]. \tag{56}$$

The method can be intuitively understood using geometrical explanation in Fig. 3. As seen here, from a estimated coefficients in a k-1th plane a new direction is created by finding the nearest point on the i th plane in the case of traditional NLMS approach. On the other hand, affine projection creates the best direction which targets a location included in the both i-1 and i th plane.

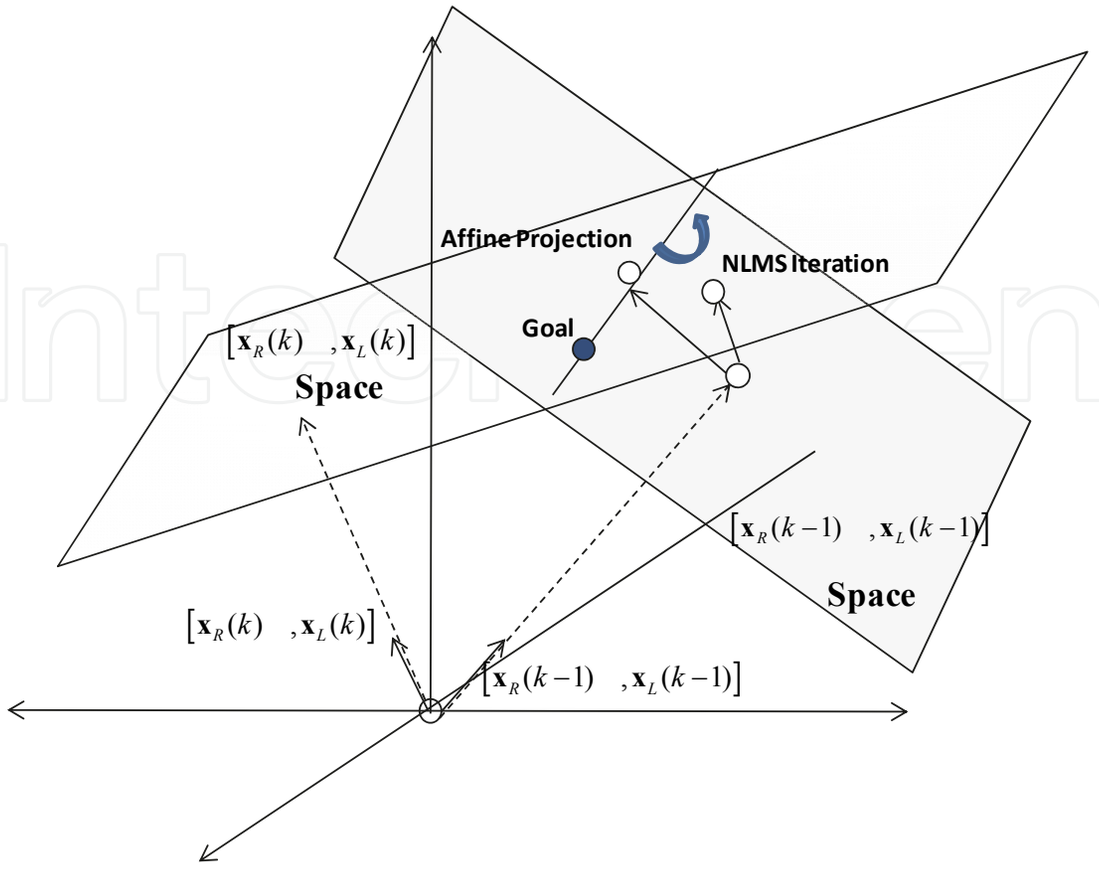


Fig. 3. Very Simple Example for Affine Method

3.4 Stereo Normalized Least Mean Square (NLMS) method (M=1)

Stereo NLMS method is a case when M=1 of the gradient method.
Equation (54) is re-written when M =1 as

$$\hat{\mathbf{h}}_{STi}(k+1)=\hat{\mathbf{h}}_{STi}(k)+\alpha\mathbf{x}_i(k)(\mathbf{x}_{Ri}^T(k)\mathbf{x}_{Ri}(k)+\mathbf{x}_{Li}^T(k)\mathbf{x}_{Li}(k))^{-1}e_i(\mathrm{k}) \tag{57}$$

It is well known that convergence speed of (57) depends on the smallest and largest eigenvalue of the matrix $\mathbf{Q}_{2N2Ni}$. In the case of the stereo generation model in Fig.2 for single talking with small right and left noises, we obtain following determinant of $\mathbf{Q}_{2N2Ni}$ for M=1 as

$$\begin{aligned}\left|\mathbf{Q}_{2N2Ni}(k)\right| &= \left|\mathbf{x}_i(k)(\mathbf{x}_i^T(k)\mathbf{x}_i(k))^{-1}\mathbf{x}_i^T(k)\right| \\ &\simeq (\mathbf{g}_{Ri}^T\mathbf{g}_{Ri}+\mathbf{g}_{Li}^T\mathbf{g}_{Li})^{-1}\left|(\mathbf{g}_{Ri}\mathbf{g}_{Ri}^T+\mathbf{g}_{Li}\mathbf{g}_{Li}^T)\right|\left|\sigma_N^2\mathbf{I}_N\right|\end{aligned} \tag{58}$$

If eigenvalue of $\mathbf{g}_{Ri}\mathbf{g}_{Ri}^T+\mathbf{g}_{Li}\mathbf{g}_{Li}^T$ are given as

$$\left|(\mathbf{g}_{Ri}\mathbf{g}_{Ri}^T+\mathbf{g}_{Li}\mathbf{g}_{Li}^T)\right|=\lambda_{\min i}^2\cdots\lambda_{\max i}^2 \tag{59}$$

where $\lambda_{\min i}^2$ and $\lambda_{\max i}^2$ are the smallest and largest eigenvalues, respectively.

$\left|\mathbf{Q}_{2N2Ni}(k)\right|$ is given by assuming un-correlated noise power σ_{Ni}^2 is very small ($\sigma_{Ni}^2 \ll \lambda_{\min i}^2$) as

$$\left|\mathbf{Q}_{2N2Ni}(k)\right| = (\mathbf{g}_{Ri}^T\mathbf{g}_{Ri} + \mathbf{g}_{Li}^T\mathbf{g}_{Li})^{-1} \cdot \sigma_{Ni}^2 \cdots \sigma_{Ni}^2 \cdot \lambda_{\min i}^2 \cdots \lambda_{\max i}^2 \tag{60}$$

Hence, it is shown that stereo NLMS echo-canceller's convergence speed is largely affected by the ratio between the largest eigenvalue of $\mathbf{g}_{Ri}\mathbf{g}_{Ri}^T + \mathbf{g}_{Li}\mathbf{g}_{Li}^T$ and non-correlated signal power σ_{Ni}^2. If the un-correlated sound power is very small in single talking, the stereo NLMS echo canceller's convergence speed becomes very slow.

3.5 Double adaptive filters for Rapid Projection (WARP) method

Naming of the WARP is that this algorithm projects the optimum solution between monaural space and stereo space. Since this algorithm dynamically changes the types of adaptive filters between monaural and stereo observing sound source characteristics, we do not need to suffer from rank drop problem caused by strong cross-channel correlation in stereo sound. The algorithm was originally developed for the acoustic echo canceller in a pseudo-stereo system which creates artificial stereo effect by adding delay and/or loss to a monaural sound. The algorithm has been extended to real stereo sound by introducing residual signal after removing the cross-channel correlation.

In this section, it is shown that WARP method is derived as an extension of affine projection which has been shown in 3.3.

By introducing error matrix $\mathbf{E}_i(k)$ which is defined by

$$\mathbf{E}_i(k) = \begin{bmatrix} \mathbf{e}_{Pi}(k) & \mathbf{e}_{Pi}(k-1) & \cdots & \mathbf{e}_{Pi}(k-p+1) \end{bmatrix} \tag{61}$$

iteration of the stereo affine projection method in (54) is re-written as

$$\hat{\mathbf{H}}_{STi}(k+1) = \hat{\mathbf{H}}_{STi}(k) + \alpha\mathbf{X}_{P2Ni}(k)(\mathbf{X}_{P2Ni}^T(k)\mathbf{X}_{P2Ni}(k))^{-1}\mathbf{E}_i(k) \tag{62}$$

where

$$\hat{\mathbf{H}}_{STi}(k) = \begin{bmatrix} \hat{\mathbf{h}}_{STi}(k) & \hat{\mathbf{h}}_{STi}(k-1) & \cdots & \hat{\mathbf{h}}_{STi}(k-p+1) \end{bmatrix} \tag{63}$$

In the case of strict single talking, following assumption is possible in the ith LTI period by (53)

$$\left\langle \mathbf{X}_{PNi}^T(k)\mathbf{X}_{PNi}(k) \right\rangle \cong \mathbf{G}_{RRLLi} \tag{64}$$

where $\mathbf{G}_{RRLLi}$ is a PxP symmetric matrix as

$$\mathbf{G}_{RRLLi} = N\sigma_{Xi}^2(\mathbf{G}_{Ri}\mathbf{G}_{Ri}^T + \mathbf{G}_{Li}\mathbf{G}_{Li}^T) \tag{65}$$

By assuming $\mathbf{G}_{RRLLi}$ as a regular matrix, (62) can be re-written as

$$\hat{\mathbf{H}}_{STi}(k+1)\mathbf{G}_{RRLLi} = \hat{\mathbf{H}}_{STi}(k)\mathbf{G}_{RRLLi} + \alpha\mathbf{X}_{P2Ni}(k)\mathbf{E}_i(k) \tag{66}$$

Re-defining echo path estimation matrix $\hat{\mathbf{H}}_{STi}(k)$ by a new matrix $\hat{\mathbf{H}}'_{STi}(k)$ which is defined by

$$\hat{\mathbf{H}}'_{STi}(k)=\hat{\mathbf{H}}_{STi}(k)\mathbf{G}_{RRLLi} \tag{67}$$

(66) is re-written as

$$\hat{\mathbf{H}}'_{STi}(k+1)=\hat{\mathbf{H}}'_{STi}(k)+\alpha\mathbf{X}_{P2Ni}(k)\mathbf{E}_i(k) \tag{68}$$

Then the iteration is expressed using signal matrix $\mathbf{X}_{2Si}(k)$ as

$$\hat{\mathbf{H}}'_{STi}(k+1)=\hat{\mathbf{H}}'_{STi}(k)+\alpha\begin{bmatrix}\mathbf{X}_{2Si}(k)\mathbf{G}_{Ri}^T+\mathbf{X}_{URi}(k)\\ \mathbf{X}_{2Si}(k)\mathbf{G}_{Li}^T+\mathbf{X}_{ULi}(k)\end{bmatrix}\mathbf{E}_i(k) \tag{69}$$

In the case of strict single talking where no un-correlated signals exist, and if we can assume $\mathbf{G}_{Li}$ is assumed to be an output of a LTI system $\mathbf{G}_{RLi}$ which is PxP symmetric regular matrix with input $\mathbf{G}_{Ri}$, then (69) is given by

$$\begin{bmatrix}\hat{\mathbf{H}}'_{STRi}(k+1)\\ \hat{\mathbf{H}}'_{STLi}(k+1)\end{bmatrix}=\begin{bmatrix}\hat{\mathbf{H}}'_{STRi}(k)\\ \hat{\mathbf{H}}'_{STLi}(k)\end{bmatrix}+\alpha\begin{bmatrix}\mathbf{X}_{2Si}(k)\mathbf{G}_{Ri}\mathbf{E}_i(k)\\ \mathbf{X}_{2Si}(k)\mathbf{G}_{Ri}\mathbf{G}_i\mathbf{E}_i(k)\end{bmatrix}$$
$$\begin{bmatrix}\hat{\mathbf{H}}'_{STRi}(k+1)\\ \hat{\mathbf{H}}'_{STLi}(k+1)\mathbf{G}_{RLi}^{-1}\end{bmatrix}=\begin{bmatrix}\hat{\mathbf{H}}'_{STRi}(k)\\ \hat{\mathbf{H}}'_{STLi}(k)\mathbf{G}_{RLi}^{-1}\end{bmatrix}+\alpha\begin{bmatrix}\mathbf{X}_{2Si}(k)\mathbf{G}_{Ri}\mathbf{E}_i(k)\\ \mathbf{X}_{2Si}(k)\mathbf{G}_{Ri}\mathbf{E}_i(k)\end{bmatrix} \tag{70}$$

It is evident that rank of the equation in (70) is N not 2N, therefore the equation becomes monaural one by subtracting the first law after multiplying $(\mathbf{G}_{RLi})^{-1}$ from the second low as

$$\hat{\mathbf{H}}_{MONRLi}(k+1)=\hat{\mathbf{H}}_{MONRLi}(k)+2\alpha\mathbf{X}_{Ri}(k)\mathbf{E}_i(k) \tag{71}$$

where

$$\hat{\mathbf{H}}_{MONRLi}(k)=\hat{\mathbf{H}}'_{STRi}(k)+\hat{\mathbf{H}}'_{STLi}(k)\mathbf{G}_{RLi}^{-1} \tag{72}$$

or assuming $\mathbf{G}_{Ri}=\mathbf{G}_{Li}\mathbf{G}_{LRi}$

$$\hat{\mathbf{H}}_{MONLRi}(k+1)=\hat{\mathbf{H}}_{MONLRi}(k)+2\alpha\mathbf{X}_{Li}(k)\mathbf{E}_i(k) \tag{73}$$

where

$$\hat{\mathbf{H}}_{MONRLi}(k)=\hat{\mathbf{H}}'_{STLi}(k)+\hat{\mathbf{H}}'_{STRi}(k)\mathbf{G}_{LRi}^{-1} \tag{74}$$

Selection of the iteration depends on existence of the inverse matrix $\mathbf{G}_{RLi}^{-1}$ or $\mathbf{G}_{LRi}^{-1}$ and the detail is explained in the next section.
By substituting (67) to (72) and (74), we obtain following equations;

$$\hat{\mathbf{H}}_{MONRLi}(k)=\hat{\mathbf{H}}_{STRi}(k)\mathbf{G}_{RRLLi}+\hat{\mathbf{H}}_{STLi}(k)\mathbf{G}_{RRLLi}\mathbf{G}_{RLi}^{-1} \tag{75}$$

or

$$\hat{\mathbf{H}}_{MONLRi}(k) = \hat{\mathbf{H}}_{STRi}(k)\mathbf{G}_{RRLLi}\mathbf{G}_{LRi}^{-1} + \hat{\mathbf{H}}_{STLi}(k)\mathbf{G}_{RRLLi} \tag{76}$$

From the stereo echo path estimation view point, we can obtain $\hat{\mathbf{H}}_{MONRLi}(k)$ or $\hat{\mathbf{H}}_{MONLRi}(k)$, however we can't identify right and left echo path estimation from the monaural one. To cope with this problem, we use two LTI periods for separating the right and left estimation results as

$$\begin{bmatrix} \hat{\mathbf{H}}_{MONLRi}^T \\ \hat{\mathbf{H}}_{MONLRi-1}^T \end{bmatrix} = \begin{bmatrix} \mathbf{G}_{RRLLi}^T & \mathbf{G}_{RRLLi}\mathbf{G}_{RLi}^{-1} \\ \mathbf{G}_{RRLLii-1}^T & \mathbf{G}_{RRLLi-1}\mathbf{G}_{RLi-1}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{H}}_{STRi}^T \\ \hat{\mathbf{H}}_{STLi} \end{bmatrix} \cdots \mathbf{G}_{RLi} \text{ and } \mathbf{G}_{RLi-1}$$

are regular matrix

$$\begin{bmatrix} \hat{\mathbf{H}}_{MONLRi}^T \\ \hat{\mathbf{H}}_{MONLRi-1}^T \end{bmatrix} = \begin{bmatrix} \mathbf{G}_{RLRLi}^T\mathbf{G}_{LRi}^{-1} & \mathbf{G}_{RRLLi} \\ \mathbf{G}_{RRLLi-1}^T\mathbf{G}_{LRi-1}^{-1} & \mathbf{G}_{RRLLi-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{H}}_{STRi}^T \\ \hat{\mathbf{H}}_{STLi} \end{bmatrix} \cdots \mathbf{G}_{LRi} \text{ and } \mathbf{G}_{LRi-1}$$

are regular matrix

$$\begin{bmatrix} \hat{\mathbf{H}}_{MONLRi}^T \\ \hat{\mathbf{H}}_{MONLRi-1}^T \end{bmatrix} = \begin{bmatrix} \mathbf{G}_{RRLLi}^T & \mathbf{G}_{RRLLi}\mathbf{G}_{RLi}^{-1} \\ \mathbf{G}_{RRLLi-1}^T\mathbf{G}_{LRi-1}^{-1} & \mathbf{G}_{RRLLi-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{H}}_{STRi}^T \\ \hat{\mathbf{H}}_{STLi} \end{bmatrix} \cdots \mathbf{G}_{RLi} \text{ and } \mathbf{G}_{LRi-1} \tag{77}$$

are regular matrix

$$\begin{bmatrix} \hat{\mathbf{H}}_{MONLRi}^T \\ \hat{\mathbf{H}}_{MONLRi-1}^T \end{bmatrix} = \begin{bmatrix} \mathbf{G}_{RRLLi}^T\mathbf{G}_{LRi}^{-1} & \mathbf{G}_{RRLLi} \\ \mathbf{G}_{RRLLi-1}^T & \mathbf{G}_{RRLLi-1}\mathbf{G}_{RLi-1}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{H}}_{STRi}^T \\ \hat{\mathbf{H}}_{STLi} \end{bmatrix} \cdots \mathbf{G}_{LRi} \text{ and } \mathbf{G}_{RLi-1}$$

are regular matrix

where $\hat{\mathbf{H}}_{MONLRi}$ and $\hat{\mathbf{H}}_{MONLRi-1}$ are monaural echo canceller estimation results at the end of each LTI period, $\hat{\mathbf{H}}_{STRi}$ and $\hat{\mathbf{H}}_{STLi}$ are right and left estimated stereo echo paths based on the $i-1th$ and ith LTI period's estimation results.
Equation (77) is written simply as

$$\hat{\mathbf{H}}_{MONi,i-1} = \mathbf{W}_i^{-1}\hat{\mathbf{H}}_{STi} \tag{78}$$

where $\hat{\mathbf{H}}_{MONRLij}^T$ is estimation result matrix for the $i-1th$ and ith LTI period's as

$$\hat{\mathbf{H}}_{MONi,i-1} = \begin{bmatrix} \hat{\mathbf{H}}_{MONRLi}^T \\ \hat{\mathbf{H}}_{MONRLi-1}^T \end{bmatrix} \tag{79}$$

$\hat{\mathbf{H}}_{STi}^T$ is stereo echo path estimation result as

$$\hat{\mathbf{H}}_{STi} = \begin{bmatrix} \hat{\mathbf{H}}_{STRi}^T \\ \hat{\mathbf{H}}_{STLi}^T \end{bmatrix} \tag{80}$$

$\mathbf{W}_i^{-1}$ is a matrix which projects stereo estimation results to two monaural estimation results and is defined by

$$\mathbf{W}_i^{-1} = \begin{cases} \begin{bmatrix} \mathbf{G}_{RRLLi}^T & \mathbf{G}_{RRLLi}\mathbf{G}_{RLi}^{-1} \\ \mathbf{G}_{RRLLi-1}^T & \mathbf{G}_{RRLLi-1}\mathbf{G}_{RLi-1}^{-1} \end{bmatrix} \cdots \mathbf{G}_{RLi} \text{ and } \mathbf{G}_{RLi-1} \text{ are regular matrix} \\ \begin{bmatrix} \mathbf{G}_{RRLLi}^T\mathbf{G}_{LRi}^{-1} & \mathbf{G}_{RRLLi} \\ \mathbf{G}_{RRLLi-1}^T\mathbf{G}_{RLRi-1}^{-1} & \mathbf{G}_{RRLLi-1} \end{bmatrix} \cdots \mathbf{G}_{LRi} \text{ and } \mathbf{G}_{LRi-1} \text{ are regular matrix} \\ \begin{bmatrix} \mathbf{G}_{RRLLi}^T & \mathbf{G}_{RRLLi}\mathbf{G}_{RLi}^{-1} \\ \mathbf{G}_{RLi-1}^T\mathbf{G}_{LRi-1}^{-1} & \mathbf{G}_{RRLLi-1} \end{bmatrix} \cdots \mathbf{G}_{RLi} \text{ and } \mathbf{G}_{LRi-1} \text{ are regular matrix} \\ \begin{bmatrix} \mathbf{G}_{RRLLi}^T\mathbf{G}_{LRi}^{-1} & \mathbf{G}_{RRLLi} \\ \mathbf{G}_{RRLLi-1}^T & \mathbf{G}_{RRLLi-1}\mathbf{G}_{RLi-1}^{-1} \end{bmatrix} \cdots \mathbf{G}_{LRi} \text{ and } \mathbf{G}_{RLi-1} \text{ are regular matrix} \end{cases} \tag{81}$$

By swapping right side hand and left side hand in(78), we obtain right and left stereo echo path estimation using two monaural echo path estimation results as

$$\hat{\mathbf{H}}_{STi} = \mathbf{W}_i \hat{\mathbf{H}}_{MONi,i-1} \,. \tag{82}$$

Since $\mathbf{W}_i^{-1}$ and $\mathbf{W}_i$ are used to project optimum solutions in two monaural spaces to corresponding optimum solution in a stereo space and vice-versa, we call the matrixes as WARP functions. Above procedure is depicted in Fig. 4. As shown here, the WARP system is regarded as an acoustic echo canceller which transforms stereo signal to correlated component and un-correlated component and monaural acoustic echo canceller is applied to the correlated signal. To re-construct stereo signal, cross-channel correlation recovery matrix is inserted to echo path side. Therefore, WARP operation is needed at a LTI system change.

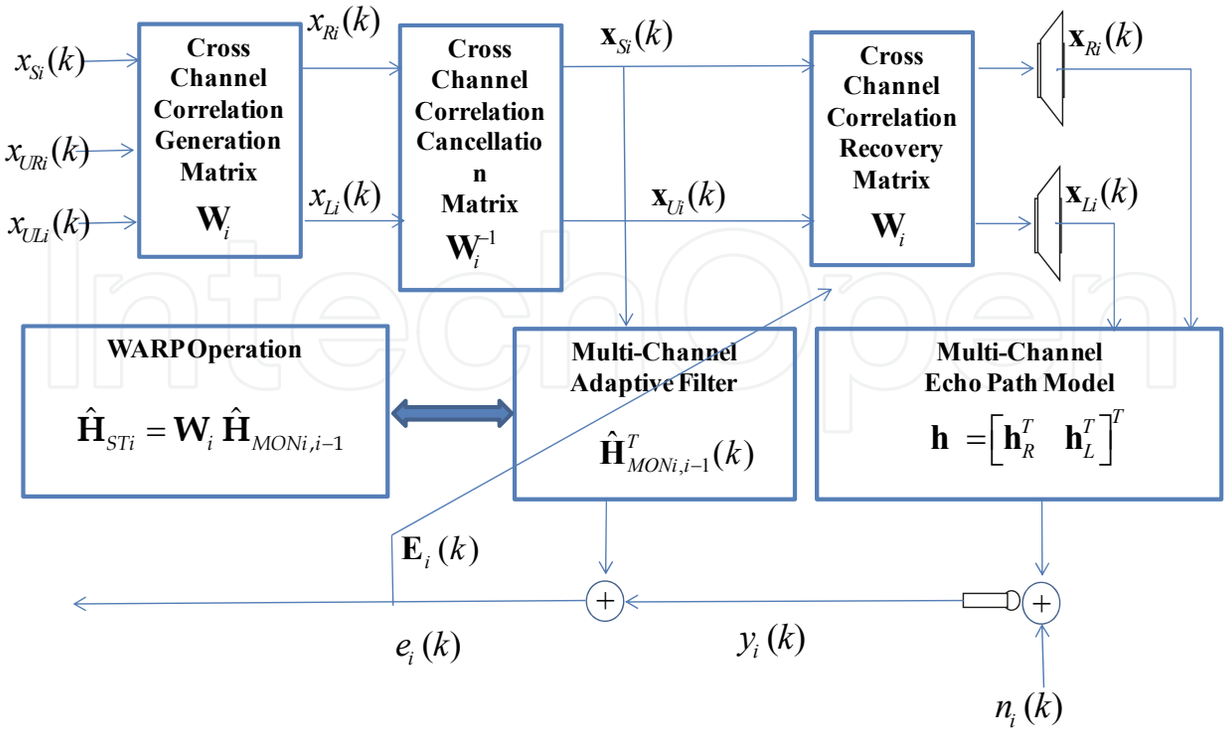


Fig. 4. Basic Principle for WARP method

In an actual application such as speech communication, the auto-correlation characteristics $\mathbf{G}_{RRLLi}$ varies frequently corresponding speech characteristics change, on the other hand the cross-channel characteristics $\mathbf{G}_{RLi}$ or $\mathbf{G}_{LRi}$ changes mainly at a far-end talker change. So, in the following discussions, we apply NLMS method as the simplest affine projection (P=1).
The mechanism is also intuitively understood by using simple vector planes depicted in Fig. 5.

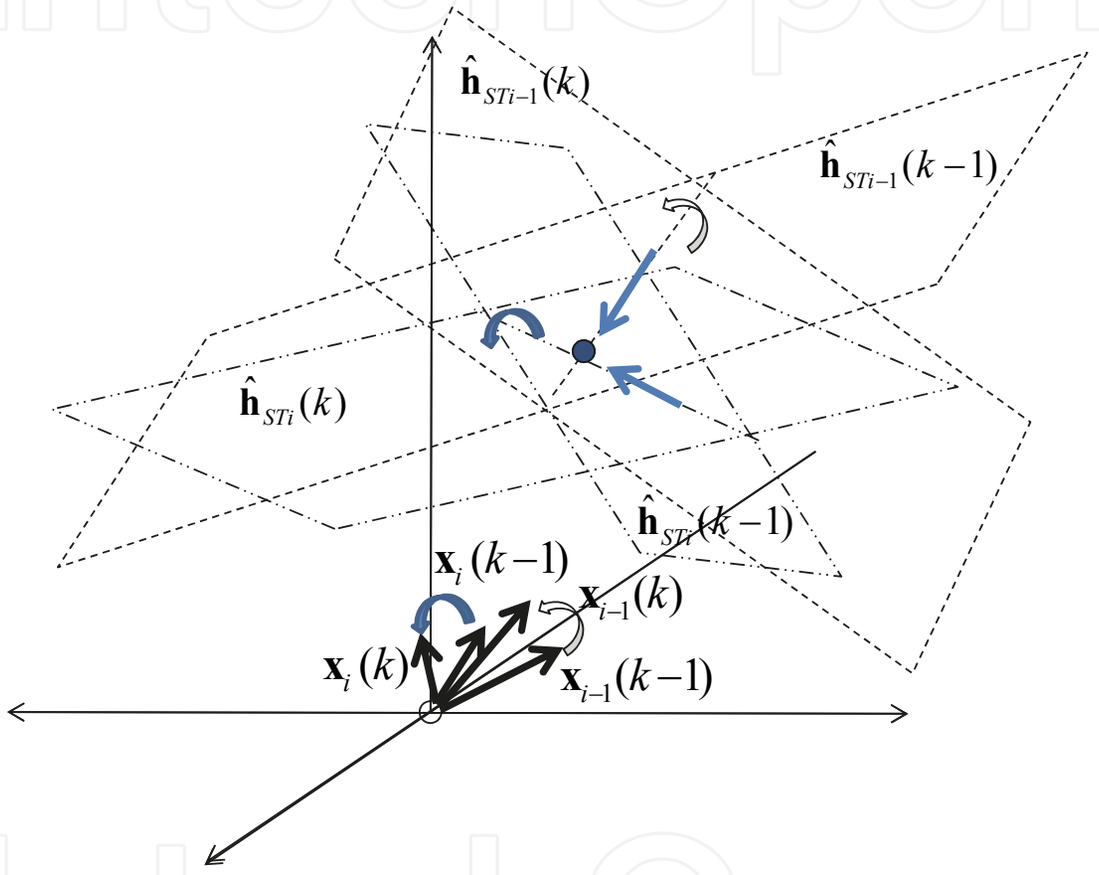


Fig. 5. Very Simple Example for WARP method

As shown here, using two optimum solutions in monaural spaces (in this case on the lines) the optimum solution located in the two dimensional (stereo) space is calculated directly.

4. Realization of WARP

4.1 Simplification by assuming direct-wave stereo sound

Both stereo affine projection and WARP methods require P x P inverse matrix operation which needs to consider its high computation load and stability problem. Even though the WARP operation is required only when the LTI system changes such as far-end talker change and it is much smaller computation than inverse matrix operations for affine projection which requires calculations in each sample, simplification of the WARP operation

is still important. This is possible by assuming that target stereo sound is composed of only direct wave sound from a talker (single talker) as shown in Fig. 6.

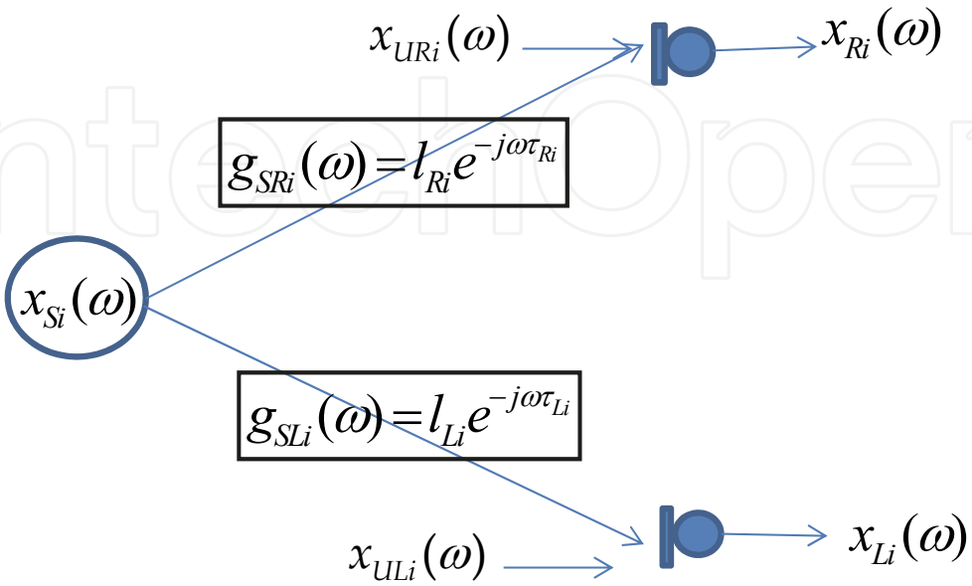


Fig. 6. Stereo Sound Generation System for Single Talking

In figure 6, a single sound source signal at an angular frequency ω in the ith LTI period, $x_{Si}(\omega)$, becomes a stereo sound composed of right and left signals, $x_{Ri}(\omega)$ and $x_{Li}(\omega)$, through out right and left LTI systems, $g_{SRi}(\omega)$ and $g_{SLi}(\omega)$ with additional un-correlated noise $x_{URi}(\omega)$ and $x_{ULi}(\omega)$ as

$$\begin{aligned} x_{Ri}(\omega) &= g_{SRi}(\omega)x_{Si}(\omega) + x_{URi}(\omega) \\ x_{Li}(\omega) &= g_{SLi}(\omega)x_{Si}(\omega) + x_{ULi}(\omega) \end{aligned}. \tag{83}$$

In the case of simple direct-wave systems, (83) can be re-written as

$$\begin{aligned} x_{Ri}(\omega) &= l_{Ri}e^{-j\omega\tau_{Ri}}x_{Si}(\omega) + x_{URi}(\omega) \\ x_{Li}(\omega) &= l_{Li}e^{-j\omega\tau_{Li}}x_{Si}(\omega) + x_{ULi}(\omega) \end{aligned} \tag{84}$$

where l_{Ri} and l_{Li} are attenuation of the transfer functions and τ_{Ri} and τ_{Li} are analog delay values.

Since the right and left sounds are sampled by $f_S(=\omega_S/2\pi)$ Hz and treated as digital signals, we use z- domain notation instead of ω-domain as

$$z = \exp[2\pi\omega j/\omega_s]. \tag{85}$$

In z-domain, the system in Fig.4 is expressed as shown in Fig. 7.

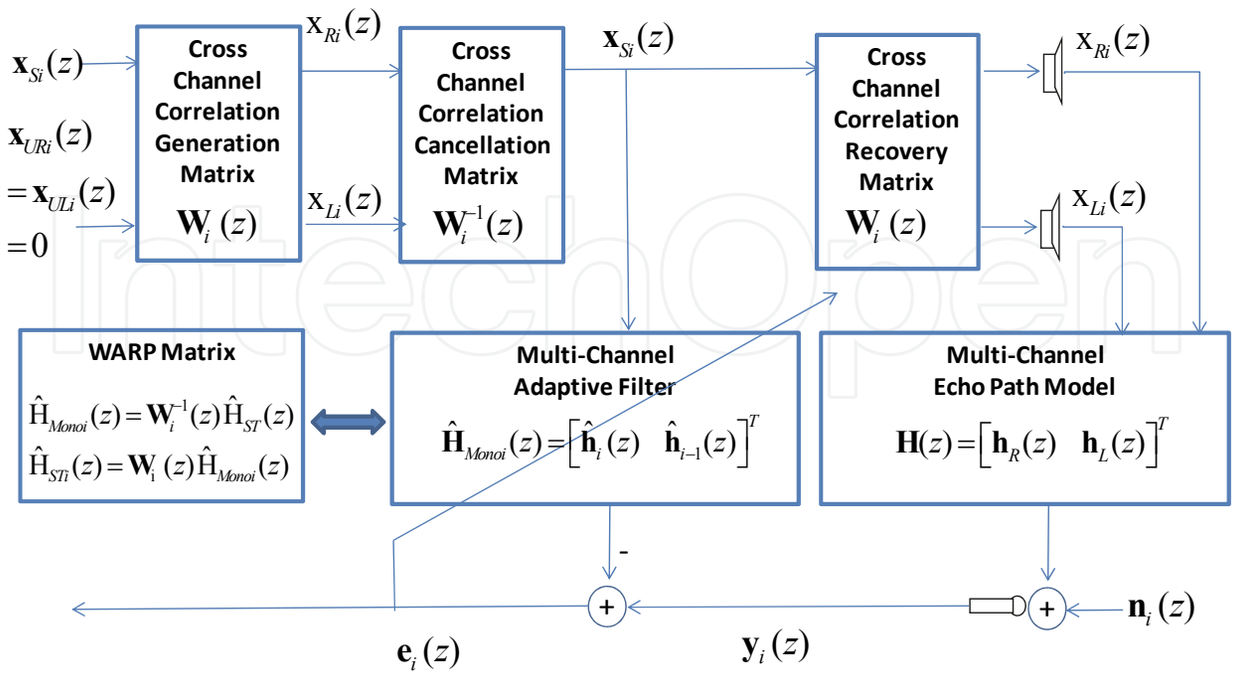


Fig. 7. WARP Method using Z-Function

As shown in Fig.7, the stereo sound generation model for $\mathbf{x}_i(z)$ is expressed as

$$\mathbf{x}_i(z)=\begin{bmatrix}\mathbf{x}_{Ri}(z)\\ \mathbf{x}_{Li}(z)\end{bmatrix}=\begin{bmatrix}\mathbf{g}_{SRi}(z)\mathbf{x}_{Si}(z)+\mathbf{x}_{URi}(z)\\ \mathbf{g}_{SLi}(z)\mathbf{x}_{Si}(z)+\mathbf{x}_{ULi}(z)\end{bmatrix} \tag{86}$$

where $\mathbf{x}_{Ri}(z)$, $\mathbf{x}_{Li}(z)$, $\mathbf{g}_{SRi}(z)$, $\mathbf{g}_{SLi}(z)$, $\mathbf{x}_{Si}(z)$, $\mathbf{x}_{URi}(z)$ and $\mathbf{x}_{ULi}(z)$ are z-domain expression of the band-limited sampled signals corresponding to $x_{Ri}(\omega)$, $x_{Li}(\omega)$, $g_{SRi}(\omega)$, $g_{SLi}(\omega)$, $x_{URi}(\omega)$ and $x_{ULi}(\omega)$, respectively. Adaptive filer output $\hat{\mathbf{y}}_i(z)$ and microphone output $\mathbf{y}_i(z)$ at the end of *ith* LTI period is defined as

$$\begin{aligned}\hat{\mathbf{y}}_i(z)&=\hat{\mathbf{h}}_i^T(z)\mathbf{x}_i(z)\\ \mathbf{y}_i(z)&=\mathbf{h}^T(z)\mathbf{x}_i(z)+\mathbf{n}_i(z)\end{aligned} \tag{87}$$

where $\mathbf{n}_i(z)$ is a room noise, $\hat{\mathbf{h}}_i(z)$ and $\hat{\mathbf{h}}_i(z)$ are stereo adaptive filter and stereo echo path characteristics at the end of *ith* LTI period respectively and which are defined as

$$\hat{\mathbf{H}}_{STi}(z)=\begin{bmatrix}\hat{\mathbf{h}}_{Ri}(z)\\ \hat{\mathbf{h}}_{Li}(z)\end{bmatrix},\mathbf{H}_{ST}(z)=\begin{bmatrix}\mathbf{h}_R(z)\\ \mathbf{h}_L(z)\end{bmatrix}. \tag{88}$$

Then cancellation error is given neglecting near end noise by

$$\mathbf{e}_i(z)=\mathbf{y}_i(z)-\hat{\mathbf{H}}_{STi}^T(z)\mathbf{x}_i(z) \tag{89}$$

In the case of single talking, we can assume both $\mathbf{x}_{URi}(z)$ and $\mathbf{x}_{ULi}(z)$ are almost zero, and (89) can be re-written as

$$\mathbf{e}_i(z) = \mathbf{y}_i(z) - (\mathbf{g}_{SRi}(z)\hat{\mathbf{h}}_{Ri}(z) + \mathbf{g}_{SLi}(z)\hat{\mathbf{h}}_{Li}(z))\mathbf{x}_{Si}(z) \tag{90}$$

Since the acoustic echo can also be assumed to be driven by single sound source $\mathbf{x}_{Si}(z)$, we can assume a monaural echo path $\mathbf{h}_{Monoi}(z)$ as

$$\mathbf{h}_{Monoi}(z) = \mathbf{g}_{SRi}(z)\mathbf{h}_R(z) + \mathbf{g}_{SLi}(z)\mathbf{h}_L(z)\,. \tag{91}$$

Then (90) is re-written as

$$\mathbf{e}_i(z) = (\mathbf{h}_{Monoi}(z) - (\mathbf{g}_{SRi}(z)\hat{\mathbf{h}}_{Ri}(z) + \mathbf{g}_{SLi}(z)\hat{\mathbf{h}}_{Li}(z)))\mathbf{x}_{Si}(z)\,. \tag{92}$$

This equation implies we can adopt monaural adaptive filter by using a new monaural quasi-echo path $\hat{\mathbf{h}}_{Monoi}(z)$ as

$$\hat{\mathbf{h}}_{Monoi}(z) = \mathbf{g}_{SRi}(z)\hat{\mathbf{h}}_{Ri}(z) + \mathbf{g}_{SLi}(z)\hat{\mathbf{h}}_{Li}(z)\,. \tag{93}$$

However, it is also evident that if LTI system changes both echo and quasi-echo paths should be up-dated to meet new LTI system. This is the same reason for the stereo echo canceller in the case of pure single talk stereo sound input. If we can assume the acoustic echo paths is time invariant for two adjacent LTI periods, this problem is easily solved by satisfying require rank for solving the equation as

$$\begin{bmatrix} \hat{\mathbf{h}}_{Monoi}(z) \\ \hat{\mathbf{h}}_{Monoi-1}(z) \end{bmatrix} = \mathbf{W}_i^{-1}(z) \begin{bmatrix} \hat{\mathbf{h}}_{Ri}(z) \\ \hat{\mathbf{h}}_{Li}(z) \end{bmatrix}. \tag{94}$$

where

$$\mathbf{W}_i^{-1}(z) = \begin{bmatrix} \mathbf{g}_{SRi}(z) & \mathbf{g}_{SLi}(z) \\ \mathbf{g}_{SRi-1}(z) & \mathbf{g}_{SLi-1}(z) \end{bmatrix} \tag{95}$$

In other words, using two echo path estimation results for corresponding two LTI periods, we can project monaural domain quasi-echo path to stereo domain quasi echo path or vice - versa using WARP operations as

$$\begin{aligned} \hat{\mathbf{H}}_{STi}(z) &= \mathbf{W}_i\,(z)\hat{\mathbf{H}}_{Monoi}(z) \\ \hat{\mathbf{H}}_{Monoi}(z) &= \mathbf{W}_i^{-1}(z)\hat{\mathbf{H}}_{STi}(z) \end{aligned}\,. \tag{96}$$

where

$$\hat{\mathbf{H}}_{Monoi}(z) = \begin{bmatrix} \hat{\mathbf{h}}_{Monoi}(z) \\ \hat{\mathbf{h}}_{Monoi-1}(z) \end{bmatrix}, \hat{\mathbf{H}}_{STi}(z) = \begin{bmatrix} \hat{\mathbf{h}}_{Ri}(z) \\ \hat{\mathbf{h}}_{Li}(z) \end{bmatrix}. \tag{97}$$

In actual implementation, it is impossible to obtain real $W_i(z)$, which is composed of unknown transfer functions between a sound source and right and left microphones, so use one of the stereo sounds as a single talk sound source instead of a sound source. Usually, higher level sound is chosen as a pseudo-sound source because higher level sound is usually closer to one of the microphones. Then, the approximated WARP function $\tilde{\mathbf{W}}_i(z)$ is defined as

$$\tilde{\mathbf{W}}_i(z) = \begin{cases} \begin{bmatrix} 1 & \mathbf{g_{RLi}}(z) \\ 1 & \mathbf{g_{RLi\text{-}1}}(z) \end{bmatrix} \cdots RR-Transition \\ \begin{bmatrix} 1 & \mathbf{g_{RLi}}(z) \\ \mathbf{g_{LRi\text{-}1}}(z) & 1 \end{bmatrix} \cdots RL-Transition \\ \begin{bmatrix} \mathbf{g_{LRi}}(z) & 1 \\ 1 & \mathbf{g_{RLi\text{-}1}}(z) \end{bmatrix} \cdots LR-Transition \\ \begin{bmatrix} \mathbf{g_{LRi}}(z) & 1 \\ \mathbf{g_{LRi\text{-}1}}(z) & 1 \end{bmatrix} \cdots LL-Transition \end{cases} \tag{98}$$

where $\mathbf{g}_{RLi}(z)$ and $\mathbf{g}_{LRi}(z)$ are cross-channel transfer functions between right and left stereo sounds and are defined as

$$\mathbf{g}_{RLi}(z) = \mathbf{g}_{SLi}(z) / \mathbf{g}_{SRi}(z), \mathbf{g}_{LRi}(z) = \mathbf{g}_{SRi}(z) / \mathbf{g}_{SLi}(z). \tag{99}$$

The RR, RL, LR and LL transitions in (98) mean a single talker's location changes. If a talker' location change is within right microphone side (right microphone is the closest microphone) we call RR-transition and if it is within left-microphone side (left microphone is the closest microphone) we call LL-transition. If the location change is from right-microphone side to left microphone side, we call RL-transition and if the change is opposite we call LR-transition. Let's assume ideal direct-wave single talk case. Then the ω domain transfer functions, $g_{RLi}(\omega)$ and $g_{LRi}(\omega)$ are expressed in z-domain as

$$\mathbf{g}_{RLi}(z) = l_{RLi}\varphi(\delta_{RLi,}, z)z^{-d_{RLi}}, \mathbf{g}_{LRi}(z) = l_{LRi}\varphi(\delta_{LRi,}, z)z^{-d_{LRi}} \tag{100}$$

where $\delta_{RLi,}$, and $\delta_{LRi,}$ are fractional delays and d_{RLi} and d_{LRi} are integer delays for the direct-wave to realize analog delays τ_{RLi} and τ_{LRi}, these parameters are defined as

$$\begin{aligned} &d_{RLi} = INT[\tau_{RLi} f_S]. d_{LRi} = INT[\tau_{LRi} f_S], \\ &\delta_{RLi,} = Mod[\tau_{RLi} f_S], \delta_{LRi,} = Mod[\tau_{LRi} f_S] \end{aligned} \tag{101}$$

$\varphi(\delta, z)$ is a "Sinc Interpolation" function to interpolate a value at a timing between adjacent two samples and is given by

$$\boldsymbol{\varphi}(\delta, z) = \sum_{\nu=-\infty}^{\infty} \frac{\sin(\pi\nu - \delta)}{(\pi\nu - \delta)} z^{-\nu}. \tag{102}$$

4.2 Digital filter realization of WARP functions

Since LL-transition and LR transition are symmetrical to RR-transition and RL-transition respectively, Only RR and RL transition cases are explained in the following discussions. By solving (96) applying WARP function in(98), we obtain right and left stereo echo path estimation functions as

$$\begin{aligned}\hat{\mathbf{h}}_{Ri}(z) &= \frac{\hat{\mathbf{h}}_{Monoi}(z)-\hat{\mathbf{h}}_{Monoi-1}(z)}{\mathbf{g}_{\ RLi-1}(z)-\mathbf{g}_{\ RLi}(z)} \\ \hat{\mathbf{h}}_{Li}(z) &= \frac{\mathbf{g}_{\ RLi-1}(z)\hat{\mathbf{h}}_{Monoi}(z)-\mathbf{g}_{\ RLi}(z)\hat{\mathbf{h}}_{Monoi-1}(z)}{\mathbf{g}_{\ RLi-1}(z)-\mathbf{g}_{\ RLi}(z)}\end{aligned} \quad \cdots RR-Transition \tag{103}$$

or

$$\begin{aligned}\hat{\mathbf{h}}_{Ri}(z) &= \frac{\hat{\mathbf{h}}_{Monoi}(z)-\mathbf{g}_{\ RLi-1}(z)\hat{\mathbf{h}}_{Monoi-1}(z)}{1-\mathbf{g}_{\ LRi}(z)\mathbf{g}_{\ RLi-1}(z)} \\ \hat{\mathbf{h}}_{Li}(z) &= \frac{\hat{\mathbf{h}}_{Monoi-1}(z)-\mathbf{g}_{\ LRi}(z)\hat{\mathbf{h}}_{Monoi}(z)}{1-\mathbf{g}_{\ LRi}(z)\mathbf{g}_{\ RLi-1}(z)}\end{aligned} \quad \cdots RL-Transition \tag{104}$$

By substituting (100) for (104), we obtain

$$\begin{aligned}\hat{\mathbf{h}}_{Ri}(z) &= \frac{\hat{\mathbf{h}}_{Monoi}(z)-\hat{\mathbf{h}}_{Monoi-1}(z)}{l_{RLi-1}\boldsymbol{\varphi}(\delta_{RLi-1,}z)z^{-d_{RLi-1}}-l_{RLi}\boldsymbol{\varphi}(\delta_{RLi,}z)z^{-d_{RLi}}} \\ \hat{\mathbf{h}}_{Li}(z) &= \frac{l_{RLi-1}\boldsymbol{\varphi}(\delta_{RLi-1,}z)z^{-d_{RLi-1}}\hat{\mathrm{h}}_{Monoi}(z)-l_{RLi}\boldsymbol{\varphi}(\delta_{RLi,}z)z^{-d_{RLi}}\hat{\mathrm{h}}_{Monoi-1}(z)}{l_{RLi-1}\boldsymbol{\varphi}(\delta_{RLi-1,}z)z^{-d_{RLi-1}}-l_{RLi}\boldsymbol{\varphi}(\delta_{RLi,}z)z^{-d_{RLi}}}\end{aligned} \quad \cdots RR-Transition \tag{105}$$

and

$$\begin{aligned}\hat{\mathbf{h}}_{Ri}(z) &= \frac{\hat{\mathbf{h}}_{Monoi}(z)-l_{RLi-1}\varphi(\delta_{RLi-1,}z)z^{-d_{RLi-1}}\hat{\mathbf{h}}_{Monoi-1}(z)}{1-l_{LRi}\varphi(\delta_{LRi,}z)l_{RLi-1}\varphi(\delta_{RLi-1,}z)z^{-(d_{RLi-1}+d_{LRi})}} \\ \hat{\mathbf{h}}_{Li}(z) &= \frac{\hat{\mathbf{h}}_{Monoi-1}(z)-l_{LRi}\varphi(\delta_{LRi,}z)z^{-d_{LRi}}\hat{\mathbf{h}}_{Monoi}(z)}{1-l_{LRi}\varphi(\delta_{LRi,}z)l_{RLi-1}\varphi(\delta_{RLi-1,}z)z^{-(d_{RLi-1}+d_{LRi})}}\end{aligned} \quad \cdots RL-Transition \tag{106}$$

Since $\boldsymbol{\varphi}(\delta_{,}z)$ is an interpolation function for a delay δ, the delay is compensated by $\boldsymbol{\varphi}(-\delta_{,}z)$ as

$$\boldsymbol{\varphi}(-\delta_{,}z)\cdot\boldsymbol{\varphi}(\delta_{,}z)=1. \tag{107}$$

From(107), (105) is re-written as

$$\begin{aligned}\hat{\mathbf{h}}_{Ri}(z) &= \frac{(\hat{\mathbf{h}}_{Monoi}(z)-\hat{\mathbf{h}}_{Monoi-1}(z))l_{RLi-1}^{-1}\boldsymbol{\varphi}(-\delta_{RLi-1,}z)z^{d_{RLi-1}}}{1-(l_{RLi}l_{RLi-1}^{-1})\boldsymbol{\varphi}(-\delta_{RLi-1,}z)\boldsymbol{\varphi}(\delta_{RLi,}z)z^{-(d_{RLi}-d_{RLi-1})}} \\ \hat{\mathbf{h}}_{Li}(z) &= \frac{\hat{\mathbf{h}}_{Monoi}(z)-l_{RLi}l_{RLi-1}^{-1}\boldsymbol{\varphi}(\delta_{RLi,}z)\boldsymbol{\varphi}(-\delta_{RLi-1,}z)z^{-d_{RLi}+d_{RLi-1}}\hat{\mathbf{h}}_{Monoi-1}(z)}{1-(l_{RLi}l_{RLi-1}^{-1})\boldsymbol{\varphi}(-\delta_{RLi-1,}z)\boldsymbol{\varphi}(\delta_{RLi,}z)z^{-(d_{RLi}-d_{RLi-1})}}\end{aligned} \quad \cdots RR-Transition \tag{108}$$

These functions are assumed to be digital filters for the echo path estimation results as shown in Fig.8.

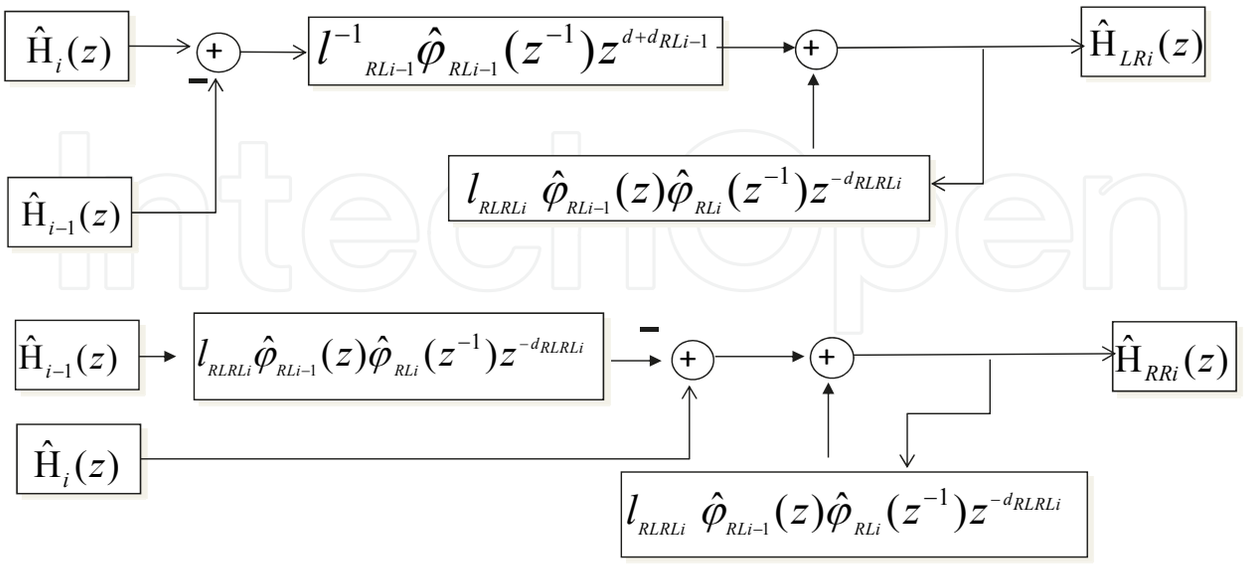


Fig. 8. Digital Filter Realization for WARP Functions

4.3 Causality and stability of WARP functions

Stability conditions are obtained by checking denominator of (108) and(106) $\mathbf{D}_{RRi}(z)$ and $\mathbf{D}_{RLi}(z)$ which are defined as

$$\begin{aligned}\left|\mathbf{D}_{RRi}(z)\right| &< 1\cdots RR-Transition\\ \left|\mathbf{D}_{RLi}(z)\right| &< 1\cdots RL-Transition\end{aligned} \tag{109}$$

where

$$\begin{aligned}\mathbf{D}_{RRi}(z) &= l_{RLi}l_{RLi-1}^{-1}\boldsymbol{\varphi}(-\delta_{RLi-1,}z)\boldsymbol{\varphi}(\delta_{RLi,}z)z^{-(d_{RLi}-d_{RLi-1})}\cdots RR-Transition\\ \mathbf{D}_{RLi}(z) &= l_{LRi}l_{RLi-1}\boldsymbol{\varphi}(\delta_{LRi,}z)\boldsymbol{\varphi}(\delta_{RLi-1,}z)z^{-(d_{RLi-1}+d_{LRi})}\cdots RL-Transition\end{aligned} \tag{110}$$

From(109),

$$\begin{aligned}\left|\boldsymbol{\varphi}(-\delta_{RLi-1,}z)\boldsymbol{\varphi}(\delta_{RLi,}z)\right| &\le \left|\boldsymbol{\varphi}(-\delta_{RLi-1,}z)\right|\left|\boldsymbol{\varphi}(\delta_{RLi,}z)\right|\cdots RR-Transition\\ \left|\boldsymbol{\varphi}(\delta_{LRi,}z)\boldsymbol{\varphi}(\delta_{RLi-1,}z)\right| &\le \left|\boldsymbol{\varphi}(\delta_{LRi,}z)\right|\left|\boldsymbol{\varphi}(\delta_{RLi-1,}z)\right|\cdots RR-Transition\end{aligned} \tag{111}$$

By using numerical calculations,

$$\left|\boldsymbol{\varphi}(\delta,z)\right| < 1.2 \tag{112}$$

Substituting (112) for (109),

$$\begin{aligned}l_{RLi}l_{RLi-1}^{-1} &< 1/1.44\cdots RR-Transition\\ l_{LRi}l_{RLi-1} &< 1/1.44\cdots RL-Transition\end{aligned} \tag{113}$$

Secondly, conditions for causality are given by checking the delay of the feedback component of the denominators $\mathbf{D}_{RRi}(z)$ and $\mathbf{D}_{RLi}(z)$. Since convolution of a "Sinc Interpolation" function is also a "Sinc Interpolation" function as

$$\boldsymbol{\varphi}(\delta_{A,}z)\cdot\boldsymbol{\varphi}(\delta_{B,}z)=\boldsymbol{\varphi}(\delta_A+\delta_{B,}z). \tag{114}$$

Equation (110) is re-written as

$$\begin{aligned}\mathbf{D}_{RRi}(z)&=l_{RLi}l_{RLi-1}^{-1}\boldsymbol{\varphi}(\delta_{RLi,}-\delta_{RLi-1,}z)z^{-(d_{RLi}-d_{RLi-1})}\cdots RR-Transition\\ \mathbf{D}_{RLi}(z)&=l_{LRi}l_{RLi-1}\boldsymbol{\varphi}(\delta_{LRi}+\delta_{RLi-1,}z)z^{-(d_{RLi-1}+d_{LRi})}\cdots RL-Transition\end{aligned} \tag{115}$$

The "Sinc Interpolation" function is an infinite sum toward both positive and negative delays. Therefore it is essentially impossible to endorse causality. However, by permitting some errors, we can find conditions to maintain causality with errors. To do so, we use a "Quasi-Sinc Interpolation" function which is defined as

$$\tilde{\boldsymbol{\varphi}}(\delta,z)=\sum_{\nu=-N_F+1}^{N_F}\frac{\sin(\pi\nu-\delta)}{(\pi\nu-\delta)}z^{-\nu}. \tag{116}$$

where $2N_F$ is a finite impulse response range of the "Quasi-Sinc Interpolation" $\tilde{\varphi}(\delta,z)$. Then the error power by the approximation is given as

$$\oint\tilde{\boldsymbol{\varphi}}(\delta,z)\tilde{\boldsymbol{\varphi}}^{*}(\delta,z)dz=\sum_{\nu=-\infty}^{-N_F}\frac{\sin^2(\pi\nu-\delta)}{(\pi\nu-\delta)^2}z^{-\nu}+\sum_{\nu=N_F+1}^{\infty}\frac{\sin^2(\pi\nu-\delta)}{(\pi\nu-\delta)^2}z^{-\nu}. \tag{117}$$

Equation (116) is re-written as

$$\tilde{\boldsymbol{\varphi}}(\delta,z)=\sum_{\nu=0}^{2N_F-1}\frac{\sin(\pi\nu-\delta)}{(\pi\nu-\delta)}z^{-\nu-N_F+1}. \tag{118}$$

By substituting (118) for (115),

$$\begin{aligned}\mathbf{D}_{RRi}(z)&\simeq l_{RLi}l_{RLi-1}^{-1}\tilde{\boldsymbol{\varphi}}(\delta_{RLi,}-\delta_{RLi-1,}z)z^{-(d_{RLi}-d_{RLi-1}-N_F+1)}\cdots RR-Transition\\ \mathbf{D}_{RLi}(z)&\simeq l_{LRi}l_{RLi-1}\tilde{\boldsymbol{\varphi}}(\delta_{LRi}+\delta_{RLi-1,}z)z^{-(d_{RLi-1}+d_{LRi}-N_F+1)}\cdots RL-Transition\end{aligned} \tag{119}$$

Then conditions for causality are

$$\begin{aligned}d_{RLi}-d_{RLi-1}&\geq N_F-1\cdots RR-Transition\\ d_{RLi-1}+d_{LRi}&\geq N_F-1\cdots RL-Transition\end{aligned} \tag{120}$$

The physical meaning of the conditions are the delay difference due to talker's location change should be equal or less than cover range of the "Quasi-Sinc Interpolation" $\tilde{\boldsymbol{\varphi}}(\delta,z)$ in the case of staying in the same microphone zone and the delay sun due to talker's location change should be equal or less than cover range of the "Quasi-Sinc Interpolation" $\tilde{\boldsymbol{\varphi}}(\delta,z)$ in the case of changing the microphone zone.

4.4 Stereo echo canceller using WARP

Total system using WARP method is presented in Fig. 9, where the system composed of five components, far-end stereo sound generation model, cross-channel transfer function (CCTF) estimation block, stereo echo path model, monaural acoustic echo canceller (AEC-I) block, stereo acoustic echo canceller (AEC-II) block and WARP block.

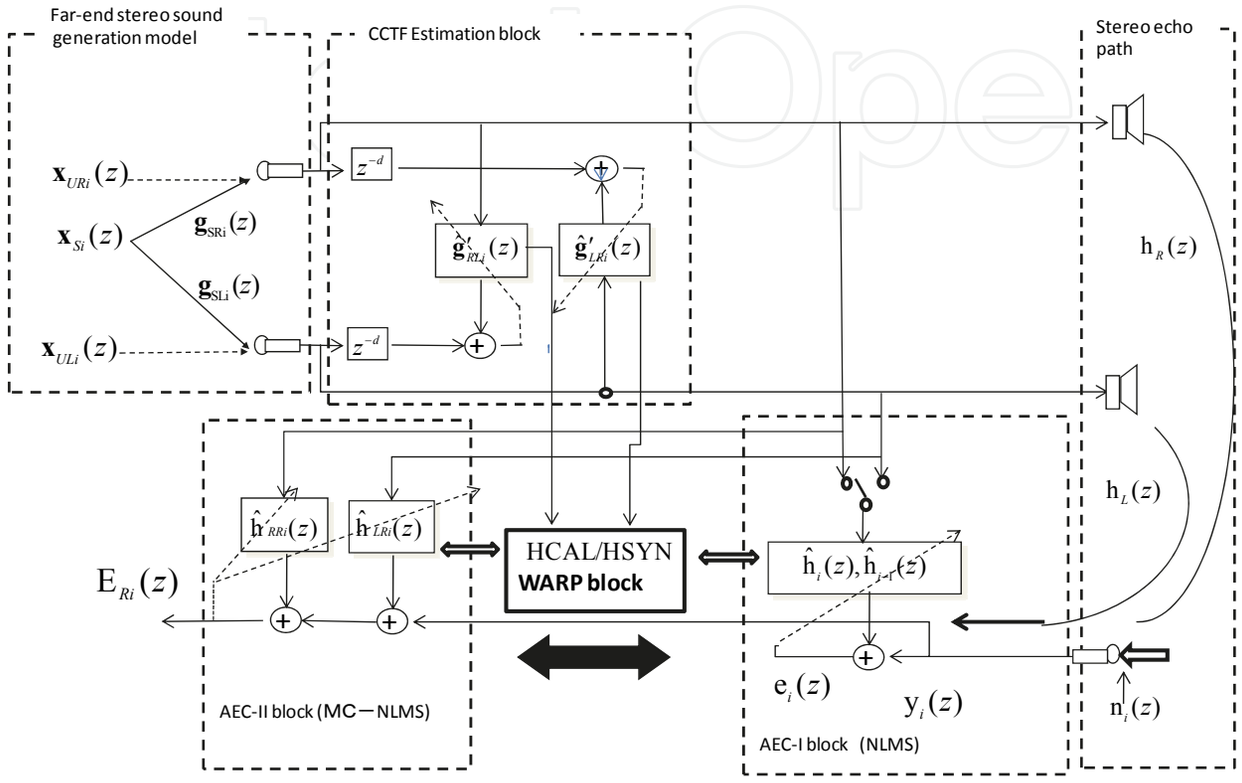


Fig. 9. System Configuration for WARP based Stereo Acoustic Echo Canceller

As shown in Fig.9, actual echo cancellation is done by stereo acoustic echo canceller (AEC-II), however, a monaural acoustic echo canceller (AEC-I) is used for the far-end single talking. The WARP block is active only when the cross-channel transfer function changes and it projects monaural echo chancellor echo path estimation results for two LTI periods to one stereo echo path estimation or vice-versa.

5. Computer simulations

5.1 Stereo sound generation model

Computer simulations are carried out using the stereo generation model shown in Fig.10 for both white Gaussian noise (WGN) and an actual voice. The system is composed of cross-channel transfer function estimation blocks (CCTF), where all signals are assumed to be sampled at $f_S = 8KHz$ after 3.4kHz cut-off low-pass filtering. Frame length is set to 100 samples. Since the stereo sound generation model is essentially a continuous time signal system, over-sampling (x6, $f_A = 48KHz$) is applied to simulate it. In the stereo sound

generation model, three far-end talker's locations, A Loc(1)=(-0.8,1.0), B Loc(2)=(-0.8,0.5), C Loc(3)=(-0.8,0.0), D Loc(4)=(-0.8,-0.5) and D Loc(5)=(-0.8,-1.0) are used and R/L microphone locations are set to R-Mic=(0,0.5) and L-Mic=(0,-0.5), respectively. Delay is calculated assuming voice wave speed as 300m/sec. In this set-up, talker's position change for WGN is assumed to be from location A to location B and finally to location D, in which each talker stable period is set to 80 frames. The position change for voice is from C->A and the period is set to 133 frames. Both room noise and reverberation components in the far-end terminals is assumed, the S/N is set to 20dB ~ 40dB.

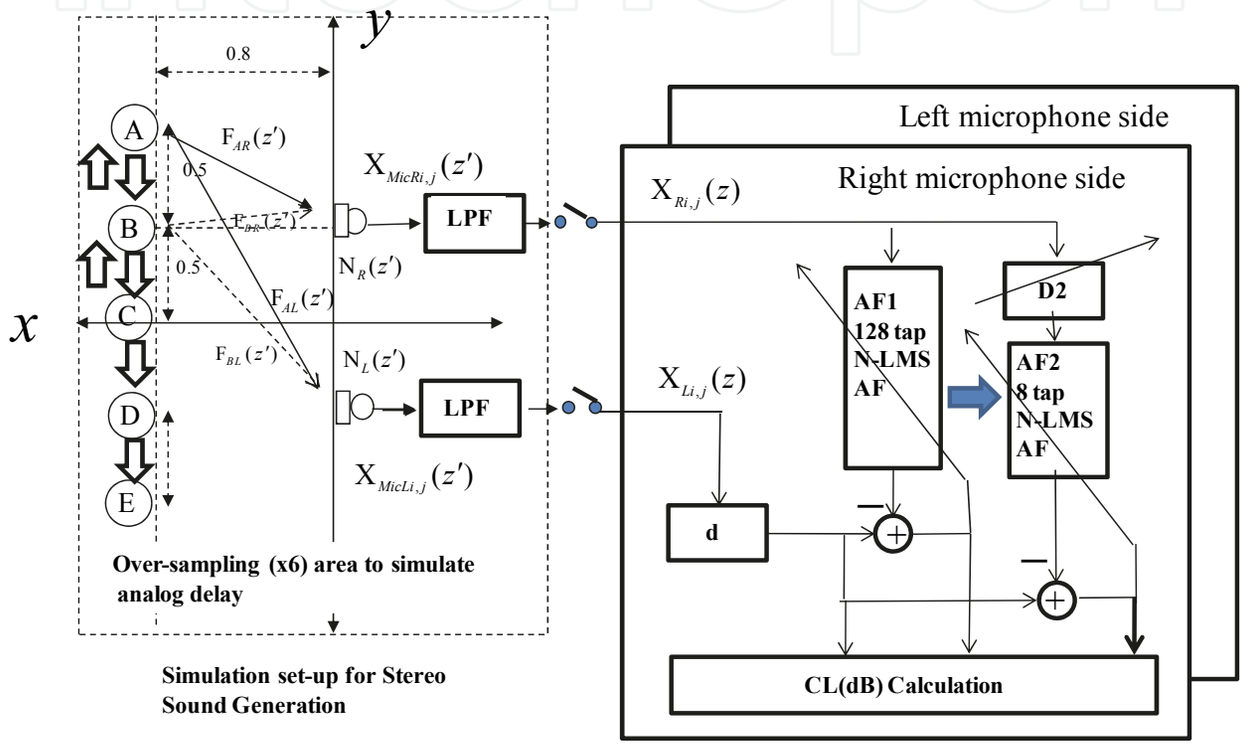


Fig. 10. Stereo Sound Generation Model and Cross-Channel Transfer Function Detector

5.2 Cross-channel transfer function estimation

In WARP method, it is easily imagine that the estimation performance of the cross-channel transfer function largely affects the echo canceller cancellation performances. To clarify the transfer function estimation performance, simulations are carried out using the cross-channel transfer function estimators (CCTF). The estimators are prepared for right microphone side sound source case and left microphone side sound source case, respectively. Each estimator has two NLMS adaptive filters, longer (128) tap one and shorter (8) tap one. The longer tap adaptive filter (AF1) is used to find a main tap and shorter one (AF2) is used to estimate the transfer function precisely as an impulse response.

Figure 11 shows CCTF estimation results as the AF1 tap coefficients after convergence setting single male voice sound source to the locations C, B and A in Fig. 11. Detail responses obtained by AF2 are shown in Fig. 12.As shown the results, the CCTF estimation works correctly in the simulations.

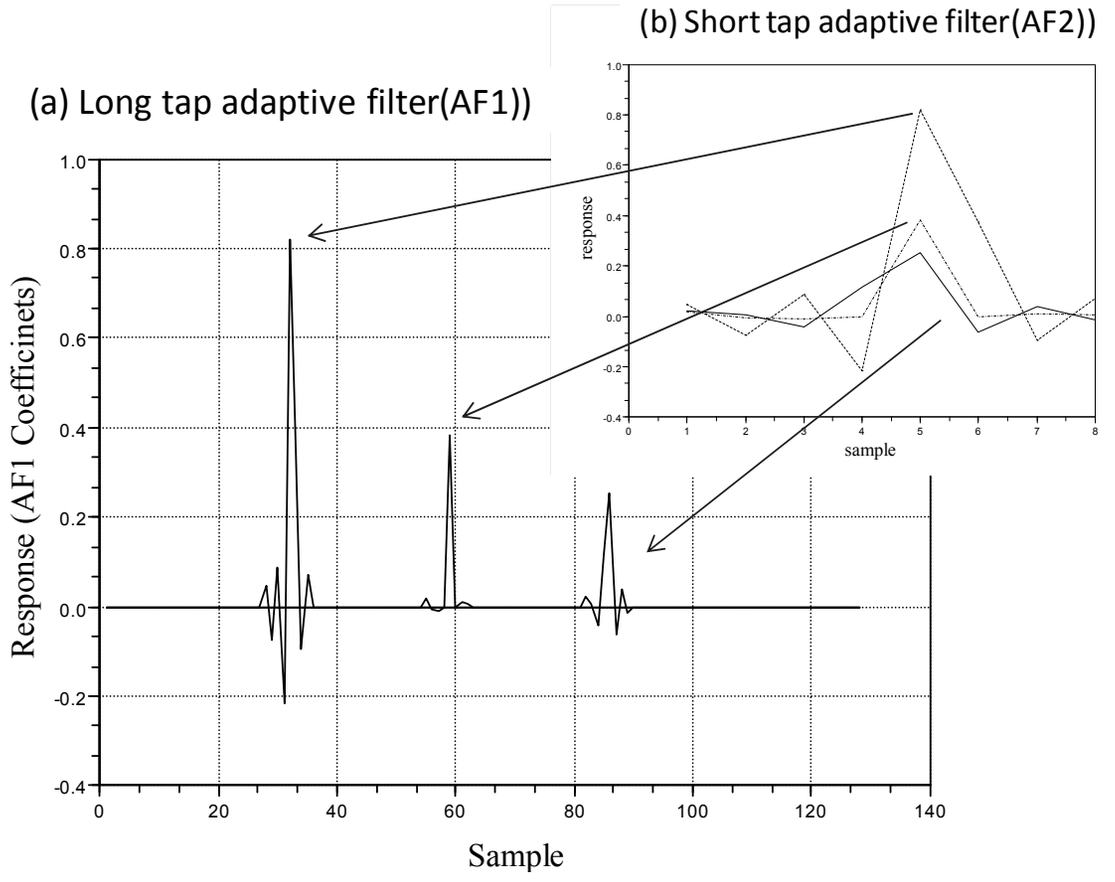


Fig. 11. Impulse Response Estimation Results in CCTF Block

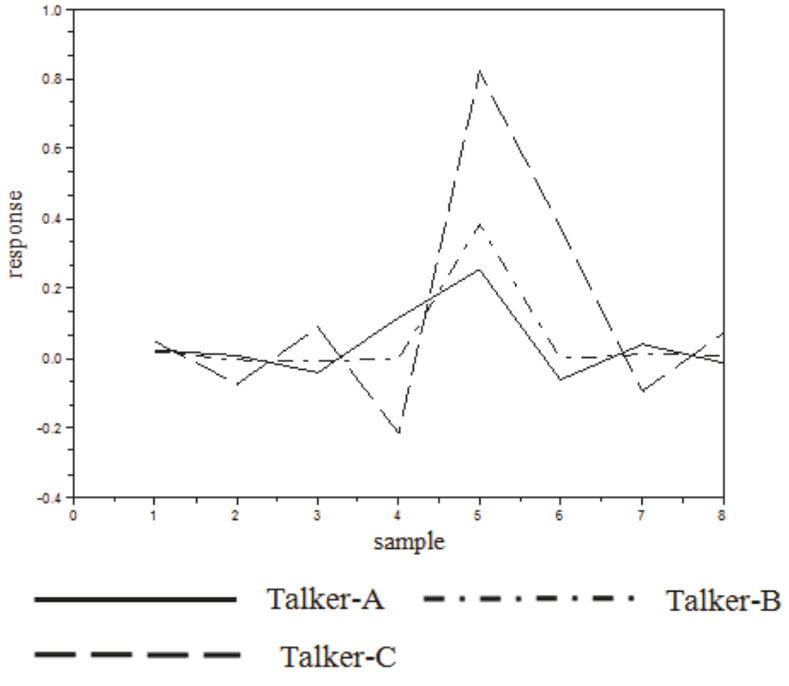


Fig. 12. Estimated Tap Coefficients by Short Tap Adaptive Filter in CCTF Estimation Block

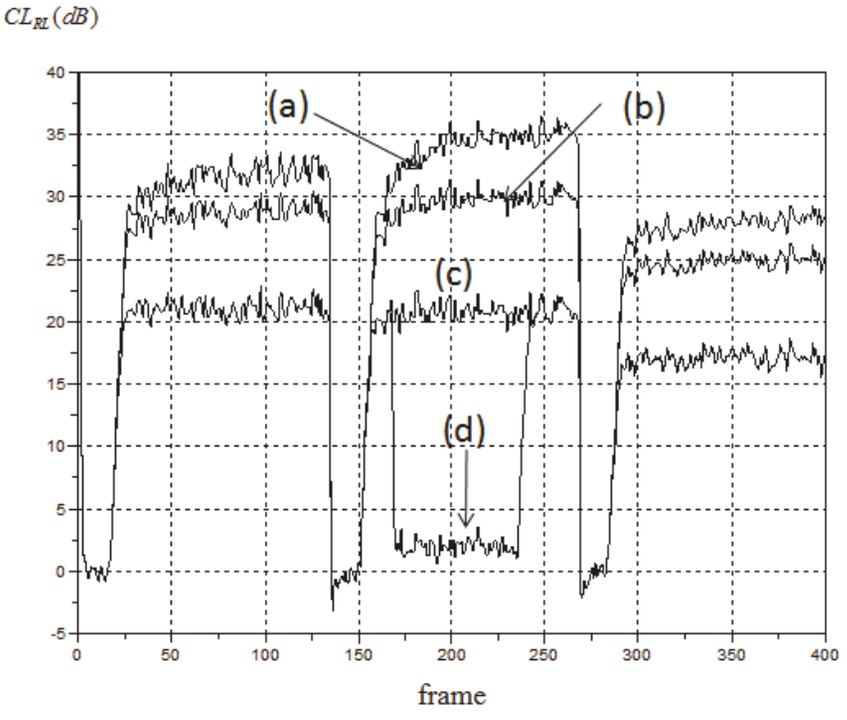


Fig. 13. Cross-Channel Correlation Cancellation Performances

Cancellation performances of the cross-channel correlation under room noise (WGN) are obtained using the adaptive filter (AF2) and are shown is Fig. 13, where S/N is assumed to be 20dB, 30dB and 40dB. In the figure $CL_{RL}(dB)$ is power reduction in dB which is observed by the signal power before and after cancellation of the cross-channel correlation by AF2. As shown here, more than 17dB cross-channel correlation cancellation is attained.

5.3 Echo canceller performances

To evaluate echo cancellation performances of the WARP acoustic echo canceller which system is shown in Fig. 10, computer simulations are carried out assuming 1000tap NLMS adaptive filters for both stereo and monaural echo cancellers. The performances of the acoustic echo canceller are evaluated by two measurements. The first one is the echo return loss enhancement $ERLE_{ij}(dB)$, which is applied to the WGN source case and is defined as

$$\mathrm{ERLE}_{L\cdot(i-1)+j-1} = \begin{cases} 10\log_{10}\left(\sum_{k=0}^{N_F-1} y^2{}_{i,j,k} \Big/ \sum_{k=0}^{N_F-1} e^2_{MONi,j,k}\right)\cdots MonauralEchoCanceller \\ 10\log_{10}\left(\sum_{k=0}^{N_F-1} y^2{}_{i,j,k} \Big/ \sum_{k=0}^{N_F-1} e^2_{STi,j,k}\right)\cdots StereoEchoCanceller \end{cases} \quad (121)$$

where $e_{MONi,j,k}$ and $e_{MONi,j,k}$ are residual echo for the monaural echo canceller (AEC-I) and stereo echo canceller (AEC-II) for the kth sample in the jth frame in the ith LTI period, respectively. The second measurement is normalized misalignment of the estimated echo paths and are defined as

$$\mathrm{NORM}_{L\cdot(i-1)+j-1} = 10\log_{10}\left(\frac{(\mathbf{h}_R)^T(\mathbf{h}_R)+(\mathbf{h}_L)^T(\mathbf{h}_L)}{(\mathbf{h}_R-\hat{\mathbf{h}}_{Ri,j})^T(\mathbf{h}_L-\hat{\mathbf{h}}_{Li,j})}\right) \tag{122}$$

where $\hat{\mathrm{h}}_{Ri,j}$ and $\hat{\mathrm{h}}_{Li,j}$ are stereo echo canceller estimated coefficient arrays at the end of $(i,j)th$ frame, respectively. h_R and h_L are target stereo echo path impulse response arrays, respectively.

5.3.1 WARP echo canceller basic performances for WGN

The simulation results for WARP echo canceller in the case of WGN sound source, no far-end double talking and no local noise, are shown in Fig. 14. In the simulations, talker is assumed to move from A to E every 80 frames (1sec). In Fig.14, the results (a) and (b) show ERLEs for monaural and stereo acoustic echo cancellers (AEC-I and AEC-II), respectively.

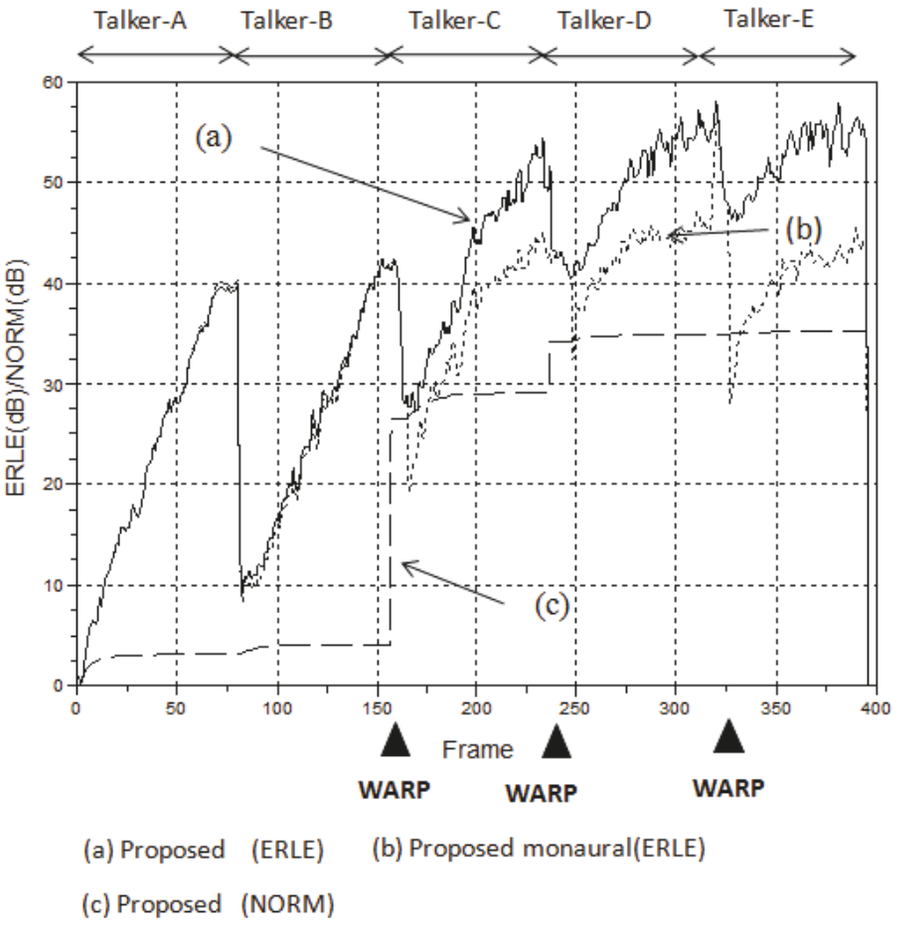


Fig. 14. WARP Echo Cancellation Performances for WGN Source

The WARP operations are applied at the boundaries of the three LTI periods for the talkers C, D and E. NORM for the stereo echo canceller (AEC-II). As shown here, after two LTI periods (A, B periods), NORM and ERLE improves quickly by WARP projection at WARP timings in the Fig. 16. As for ERLE, stereo acoustic echo canceller shows better performance than monaural echo canceller. This is because the monaural echo canceller estimates an echo path model which is combination of CCTF and real stereo echo path and therefore the performance is affected by the CCTF estimation error. On the other hand, the echo path model for the stereo echo canceller is purely the stereo echo path model which does not include CCTF.

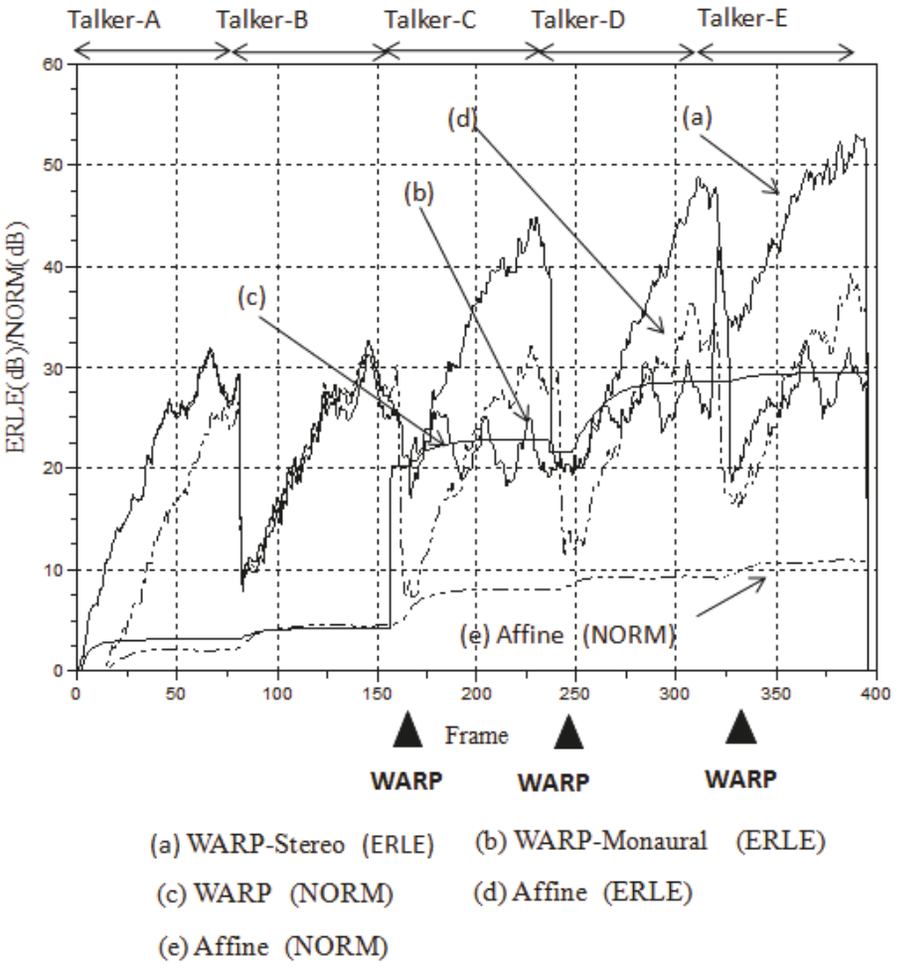


Fig. 15. Echo Cancellation Performance Comparison for WGN Source

Secondary, the WARP acoustic echo canceller is compared with a stereo echo canceller based on an affine projection method. In this case, the right and left sounds at kth sample in the $(i,j)th$ frame, x'_{Rijk} and x'_{Lijk}, are assumed to have independent level shift to the original right and left sounds, x_{Rijk} and x_{Lijk}, for simulating small movement of talker's face as

$$\begin{aligned} x'_{Rijk} &= (1+\alpha_{Level}\cdot\sin(2\pi k/(f_s\cdot T_X)))x_{Rijk} \\ x'_{Lijk} &= (1+\alpha_{Level}\cdot\cos(2\pi k/(f_s\cdot T_X)))x_{Lijk} \end{aligned} \tag{123}$$

where α_{Level} and T_X are constants which determine the level shift ratio and cycle. Figure 15 shows the cancellation performances when α_{Level} and T_X are 10% and 500msec, respectively.

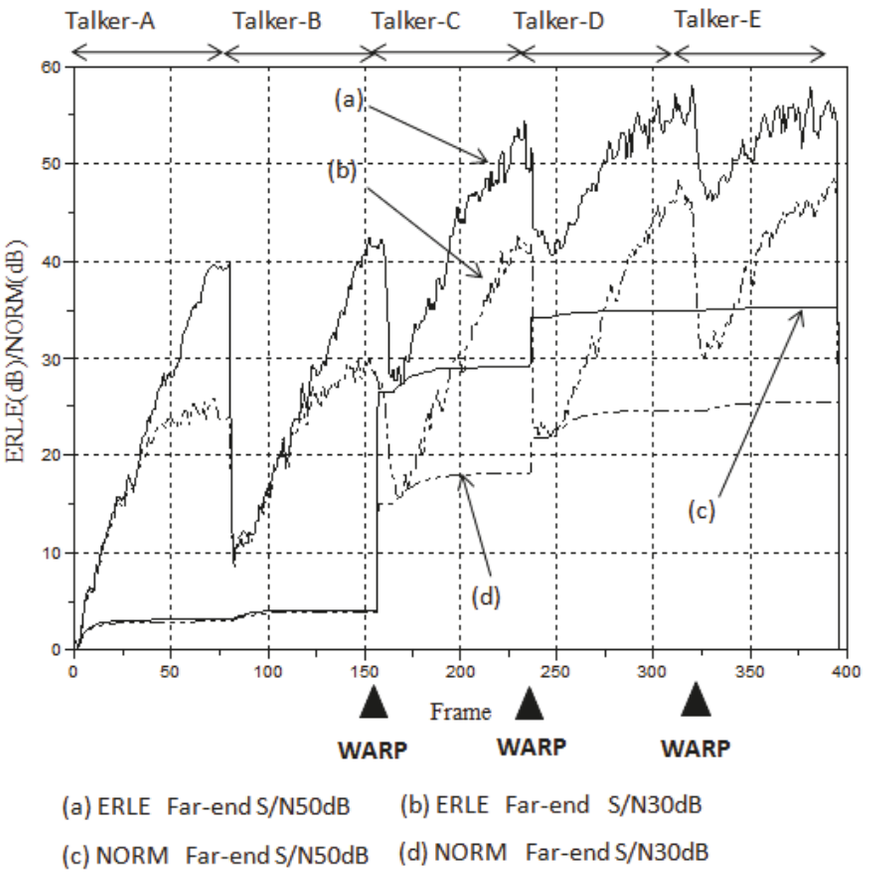


Fig. 16. WARP Echo Canceller Performances Affected by Far-End Back Ground Noise

In Fig. 15, the WARP method shows more than 10dB better stereo echo path estimation performance, NORM, than that of affine projection (P=3). ERLE by stereo echo canceller base on WARP method is also better than affine projection (P=3). ERLE by monaural acoustic echo canceller based on WARP method is somehow similar cancellation performance as affine method (P=3), however ERLE improvement after two LTI periods by the WARP based monaural echo canceller is better than affine based stereo echo canceller.

Figure 16 shows the echo canceller performances in the case of CCTF estimation is degraded by room noise in the far-end terminal. S/N in the far-end terminal is assumed to be 30dB or 50dB. Although the results clearly show that lower S/N degrade ERLR or NORM, more than 15dB ERLE or NORM is attained after two LTI periods.

Figure 17 shows the echo canceller performances in the case of echo path change happens. In this simulation, echo path change is inserted at 100frame. The echo path change is chosen 20dB, 30dB and 40dB. It is observed that echo path change affects the WARP calculation and therefore WARP effect degrades at 2nd and third LTI period boundary.

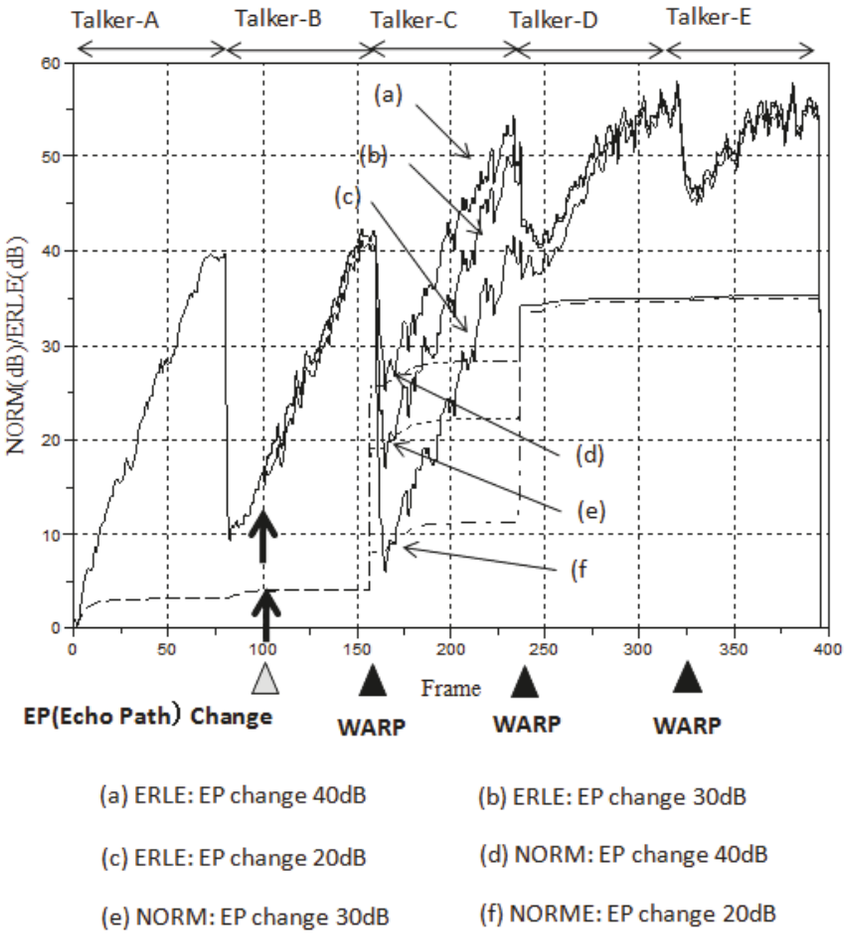


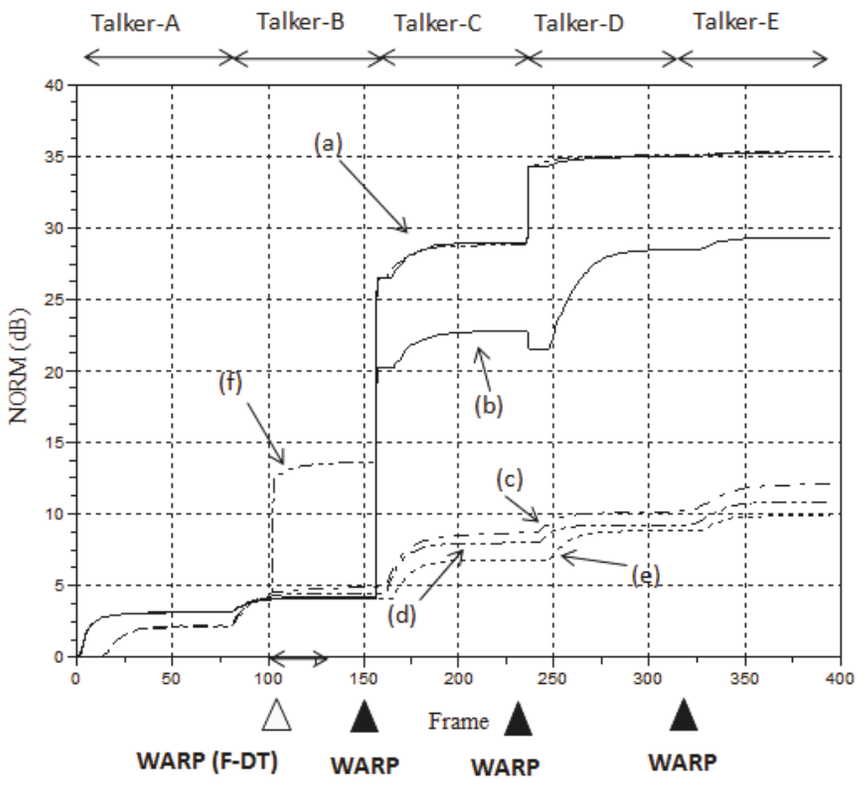
Fig. 17. WARP Echo Canceller Cancellation Performance Drops Due to Echo Path Chance

Figure 18 summarizes NORM results for stereo NLMS method, affine projection method as WARP method. In this simulation, as a non-linear function for affine projection, independent absolute values of the right and left sounds are added by

$$x'_{Rijk} = x_{Rijk} + 0.5 \cdot \alpha_{ABS} \cdot (x_{Rijk} + |x_{Rijk}|)$$
$$x'_{Lijk} = x_{Lijk} + 0.5 \cdot \alpha_{ABS} \cdot (x_{Lijk} - |x_{Lijk}|) \quad (124)$$

where α_{ABS} is a constant to determine non-liner level of the stereo sound and is set to 10%. In this simulation, an experiment is carried out assuming far-end double talking, where WGN which power is same as far-end single talking is added between 100 and 130 frames.

As evident from the results in Fig. 18, WARP method shows better performances for the stereo echo path estimation regardless far-end double talking existence. Even in the case 10% far end signal level shit, WARP method attains more than 20% NORM comparing affine method (P=3) with 10% absolute non-linear result.



(a) Proposed (No level shift with/without F-DT)

(b) Proposed 10% level shift (c) Affine with NL (ABS)

(d) Affine 10% level shift (e) MC-NLMS

(f) WARP (No level shift with F-DT)

F-DT(Far-end double talk (100-130frame)

Fig. 18. Echo Path Estimation Performance Comparison for NLMS, Affine and WARP Methods

5.3.2 WARP echo canceller basic performances for voice

Figure 19 shows NORM and residual echo level (Lres) for actual male voice sound source. Since voice sound level changes frequently, we calculate residual echo level Lres (dB) instead of ERLE(dB) for white Gaussian noise case. Although slower NORM and Lres convergence than white Gaussian is observed, quick improvement for the both metrics is observed at the talker B and A border. In this simulation, we applied 500 tap NLMS adaptive filter. Affine projection may give better convergence speed by eliminating auto-correlation in the voice, however it is independent effect from WARP effect. WARP and affine projection can be used together and may contribute to convergence speed up independently.

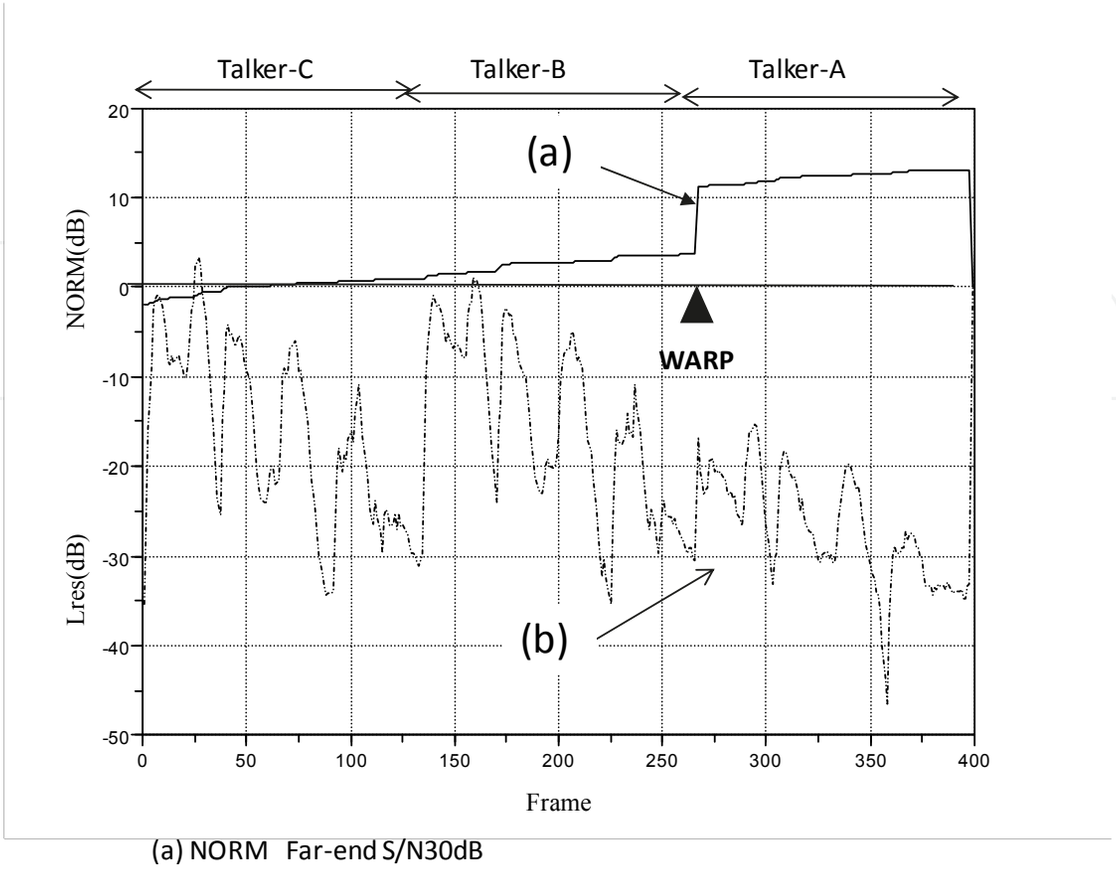


Fig. 19. Residual Echo (Lres (dB) Level and Normalized Estimated Echo Misalignment. (NORM) for the voice Source at Far-end Terminal S/N=30dB.
(Level shift 0, 500tap Step gain=1.0)

6. Conclusions

In this chapter stereo acoustic echo canceller methods are studied from cross-channel correlation view point aiming at conversational DTV use. Among many stereo acoustic echo cancellers, we focused on AP (including LS and stereo NLMS methods) and WARP methods, since these approaches do not cause any modification nor artifacts to speaker output stereo sound which is not desirable consumer audio-visual products such as DTV. In this study, stereo sound generation system is modeled by using right and left Pth order LTI systems with independent noises. Stereo LS method (M=2P) and stereo NLMS method (M=P=1) are two extreme cases of general AP method which requires MxM inverse matrix operation in each sample. Stereo AP method (M=P) can produce the best iteration direction fully adopting un-correlated component produced by small fluctuation in the stereo cross-channel correlation by calculating PxP inverse matrix operations in each sample. Major problem of the method is that it cannot cope with strict single talking where no un-correlated signals exist in right and left channels and therefore rank drop problem happens. Contrary to AP method, WARP method creates a stereo echo path estimation model applying a monaural adaptive filter for two LTI periods at a chance of far-end talker change. Since it creates stereo echo path estimation using two monaural echo path models for two

LTI periods, we do not suffer from any rank drop problem even in a strict single talking. Moreover, using WARP method, computational complexity can be reduced drastically because WARP method requires PxP inverse matrix operations only at LTI characteristics change such as far-end talker change. However, contrary to AP method, it is clear that performance of WARP method may drop if fluctuation in cross-channel correlation becomes high. Considering above pros-cons in affine projection and WARP methods, it looks desirable to apply affine method and WARP method dynamically depending on the nature of stereo sound. In this chapter, an acoustic echo canceller based on WARP method which equips both monaural and stereo adaptive filters is discussed together with other gradient base stereo adaptive filter methods. The WARP method observes cross-channel correlation characteristics in stereo sound using short tap pre-adaptive filters. Pre-adaptive filter coefficients are used to calculate WARP functions which project monaural adaptive filter estimation results to stereo adaptive filter initial coefficients or vice-versa.

To clarify effectiveness WARP method, simple computer simulations are carried out using white Gaussian noise source and male voice, using 128tap NLMS cross-channel correlation estimator, 1000tap monaural NLMS adaptive filter for monaural echo canceller and 2x1000tap (2x500tap for voice) multi-channel NLMS adaptive filter for stereo echo canceller. Followings are summary of the results:

1. Considering sampling effect for analog delay, x6 over sampling system is assumed for stereo generation model. 5 far-end talker positions are assumed and direct wave sound from each talker is assumed to be picked up by far-end stereo microphone with far-end room background noise. The simulation results show we can attain good cross-channel transfer function estimation rapidly using 128tap adaptive filter if far-end noise S/N is reasonable (such as 20-40dB).
2. Using the far-end stereo generation model and cross-channel correlation estimation results, 1000tap NLMS monaural NLMS adaptive filter and 2-1000 tap stereo NLMS adaptive filters are used to clarify effectiveness of WARP method. In the simulation far-end talker changes are assumed to happen at every 80frames (1frame=100sample). Echo return loss Enhancement (ERLE) MORMalized estimation error power (NORM) are used as measurements. It is clarified that both ERLE and NORM are drastically improved at the far-end talker change by applying WARP operation.
3. Far-end S/N affects WARP performance, however, we can still attain around SN-5dB ERLE or NORM.
4. We find slight convergence improvement in the case of AP method (P=3) with non-linear operation. However, the improvement is much smaller than WARP at the far-end talker change. This is because sound source is white Gaussian noise in this simulation and therefore merit of AP method is not archived well.
5. Since WARP method assumes stereo echo path characteristics remain stable, stereo echo path characteristics change degrade WARP effectiveness. The simulation results show the degradation depends on how much stereo echo path moved and the degradation appears just after WARP projection.
6. WARP method works correctly actual voice sound too. Collaboration with AP method may improve total convergence speed further more because AP method improves convergence speed for voice independent from WARP effect.

As for further studies, more experiments in actual environments are necessary. The author would like to continue further researches to realize smooth and natural conversations in the future conversational DTV.

7. Appendix

If $N \times N$ matrix $\mathbf{Q}$ is defined as

$$\mathbf{Q} = \mathbf{X}_{2S}^{T}(k)\mathbf{G}^{T}\mathbf{G}\ \mathbf{X}_{2S}(k) \tag{A-1}$$

where $\mathbf{X}_{2S}(k)$ is a $(2P-1)$ sample array composed of white Gaussian noise sample $x\ (k)$ as

$$\begin{aligned}\mathbf{X}_{2S}(k) &= [\mathbf{x}\ (k), \mathbf{x}\ (k-1), \cdots \mathbf{x}\ (k-N+1)] \\ \mathbf{x}\ (k) &= [x\ (k), x\ (k-1), \cdots x\ (k-2p+2)]^{T}\end{aligned} \tag{A-2}$$

$\mathbf{G}$ is defined as a $(2P-1)\times P$ matrix as

$$\mathbf{G} = \begin{bmatrix} \mathbf{g}^{T} & 0 & \cdots & 0 \\ 0 & \mathbf{g}^{T} & \ddots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \mathbf{g}^{T} \end{bmatrix} \tag{A-3}$$

where $\mathbf{g}$ is P sample array defined as

$$\mathbf{g} = [g_0, g_1, \cdots, g_v \cdots g_{P-1}]^{T}. \tag{A-4}$$

Then $\langle \mathbf{Q} \rangle$ is a Toepliz matrix and is expressed using $P \times P$ ($P \le N$) Toepliz matrix $\langle \mathbf{Q}' \rangle$ as

$$\langle \mathbf{Q} \rangle = Tlz(\langle \mathbf{Q}' \rangle) \tag{A-5}$$

This is because $(u,v)th$ element of the matrix $\langle \mathbf{Q} \rangle$, $a_{TlZ}(u,v)$ is defined as

$$a_{TlZ}(u,v) = \left\langle \mathbf{x}^{T}(k-u)\mathbf{G}^{T}\mathbf{G}\ \mathbf{x}(k-v) \right\rangle. \tag{A-6}$$

Considering

$$\left\langle \mathbf{x}^{T}(k-u)\mathbf{G}^{T}\mathbf{G}\mathbf{x}\ (k-v) \right\rangle = 0 \cdots \text{for all } |u-v| \ge P \tag{A-7}$$

the element $a_{TlZ}(u,v)$ is given as

$$a_{TlZ}(u,v) = \begin{cases} a(u-v,0) \cdots P-1 \ge u-v \ge 0 \\ a(0,v-u) \cdots P-1 \ge v-u > 0 \\ 0 \cdots |u-v| \ge P \end{cases}. \tag{A-8}$$

By setting the (u,v) th element of $P \times P$ ($P \le N$) Toepliz matrix $\langle \mathbf{Q}' \rangle$ as $a_{TlZ}(u,v)$ ($(0 \le u < P, 0 \le v < P)$), we define a function $Tlz(\langle \mathbf{Q}' \rangle)$ which determines $N \times N$ Toepliz matrix $\mathbf{Q}$.

It is noted that if $\mathbf{Q}'$ is a identity matrix $\mathbf{Q}$ is also identity matrix.

8. References

J. Nagumo, “A Learning Identification Method for System Identification”, IEEE Trans. AC. 12 No.3 Jun 1967 p282

M.M.Sondhi et.al. "Acoustic Echo Cancellation for Stereophonic Teleconferencing", Workshop on Applications of Signal Processing to Audio and Acoustics, May 1991.

Benesty. J, Amand. F, Gillorie A, Grenier Y, “adaptive filtering algorithm for a stereophonic echo cancellation” Proc. Of ICASSP-96, Vol.5, May 1996, 3099-3012..

J. Benesty, D.R. Morgan and M.M. Sondhi, “A better understanding and an improved solution to the specific problems of stereophonic acoustic echo canceller”, IEEE Trans. Speech Audio Processing, vol. 6, No. 2 pp156-165, Mar 1998.

Bershad NJ, “Behavior of the ε -normalized LMS algorithm with Gaussian inputs”, IEEE Transaction on Acoustic, Speech and Signal Processing 1987, ASSP-35(5): 636-644.

T. Fujii and S.Shimada, "A Note on Multi-Cannel Echo Cancelers," technical report of ICICE on CS, pp 7-14, Jan. 1984

A. Sugiyama, Y. Joncour and A. Hirano, “A stereo echo canceller with correct echo-path identification based on an input-sliding technique”, IEEE Trans. On Signal Processing, vol. 49, No. 11, pp2577-2587 2001.

Jun-Mei Yang;Sakai,”Stereo acoustic echo cancellation using independent component analysis” IEEE, Proceedings of 2007

International Symposium on Intelligent Signal Processing and Communication Systems (USA) P.P.121-4

Jacob Benesty, R.Morgan, M. M. Sondhi, “A hybrid Momo/Stereo Acoustic Echo Canceller”, IEEE Transactions on Speech and Audio Processing, Vol. 6. No. 5, September 1998.

S. Shimauchi, S.;Makino, S., “Stereo projection echo canceller with true echo path estimation”, IEEE Proc. of ICASSP95, vol. 3662 P.P.3059-62 vol.5 PD:1995

S. Makino, K. Strauss, S. Shimauchi, Y. Haneda, and A.Nakagawa,"Subband Stereo Echo Canceller using the projection algorithm with fast convergence to the true echo path”, IEEE Proc. of ICASSP 97, pp299-302, 1997

S. Shimauchi, S. Makino, Y. Haneda, and Y.Kaneda, "New configuration for a stereo echo canceller with nonlinier pre-processing”, IEEE Proc. of ICASSP 98, pp3685-3688, 1998

S. Shimauchi, S. Makino, Y. Haneda, A. Nakagawa, S. Sakauchi, "A stereo echo canceller implemented using a stereo shaker and a duo-filter control system”, IEEE ICASSP99 Vo. 2 pp857-60, 1999

Akira Nakagawa and Youichi Haneda, " A study of an adaptive algorithm for stereo signals with a power difference”, IEEE ICASSP2002,Vol. 2, II-1913-16, 2002

S. Minami, “An Echo Canceller with Comp. & Decomposition of Estimated Echo Path Characteristics for TV Conference & Multi-Media Terminals”, The 6th Karuizawa Workshop on Circuits and Sytstems, April 19-20 1993 pp 333-337

S.Minami,"An Acoustic Echo Canceler for Pseudo Stereophonic Voice", IEEE GLOBCOM'87 35.1 Nov. 1987

S.Minami, " A stereophonic Voice Coding Method for teleconferencing", IEEE ICCC. 86, 46.6, June 1986

Multi-Channel Acoustic Echo Canceller with Microphone/Speaker Array ITC-CSCC’09 pp 397-400 (2009)

WARP-AEC: A Stereo Acoustic Echo Canceller based on W-Adaptive filters for Rapid Projection IEEE ISPACS’09

Adaptive Filtering

Edited by Dr Lino Garcia

ISBN 978-953-307-158-9
Hard cover, 398 pages
Publisher InTech
Published online 06, September, 2011
Published in print edition September, 2011

Adaptive filtering is useful in any application where the signals or the modeled system vary over time. The configuration of the system and, in particular, the position where the adaptive processor is placed generate different areas or application fields such as prediction, system identification and modeling, equalization, cancellation of interference, etc., which are very important in many disciplines such as control systems, communications, signal processing, acoustics, voice, sound and image, etc. The book consists of noise and echo cancellation, medical applications, communications systems and others hardly joined by their heterogeneity. Each application is a case study with rigor that shows weakness/strength of the method used, assesses its suitability and suggests new forms and areas of use. The problems are becoming increasingly complex and applications must be adapted to solve them. The adaptive filters have proven to be useful in these environments of multiple input/output, variant-time behaviors, and long and complex transfer functions effectively, but fundamentally they still have to evolve. This book is a demonstration of this and a small illustration of everything that is to come.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Shigenobu Minami (2011). A Stereo Acoustic Echo Canceller Using Cross-Channel Correlation, Adaptive Filtering, Dr Lino Garcia (Ed.), ISBN: 978-953-307-158-9, InTech, Available from: http://www.intechopen.com/books/adaptive-filtering/a-stereo-acoustic-echo-canceller-using-cross-channel-correlation

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen