両耳補聴システムにおけるHRTFを利用した画像トラッキング併用型マルチモーダル・ブラインド音声抽出*

☆吉江孝太郎, 猿渡洋, 中村哲 (奈良先端大・情報)

1 Introduction

In the binaural hearing-aid systems, preservation of sound localization cues is required even after noise reduction. However, the most of noise reduction methods are not designed to preserve sound localization cues. Therefore, we should conduct preservation of sound localization with noise reduction.

The basic theory of the minimum mean-square error (MMSE) short-time spectral amplitude (STSA) estimator with dynamic noise estimation based on independent component analysis (ICA) [1] has been proposed by the authors [2]. This method can achieve high-quality noise reduction by estimating the noise signals dynamically using ICA. Then, as an application for hearing-aid system of this method, the sound-localization-preserved MMSE STSA estimator with ICA-based noise estimation has been proposed [3]. However, in the conventional method [3], equi-binaural optimal spectral gain is approximately estimated by using optimal spectral gains independently obtained from single-channel MMSE STSA estimation for each channel. Thus, the conventional method does not use the binaural cue such as a head ralated transfer function (HRTF). This fact often causes serious degradation of noise reduction performance, especially in the case that the target speech source is laterally located in the right- (left-) hand side. In this paper, we propose a new method combining a multichannel MMSE STSA estimator and an image sensor.

2 Previous Works

2.1 MMSE STSA Estimator with ICA-Based Noise Estimation

One of the authors has proposed a blind speech extraction method [2] combining ICA-based dynamic noise estimation and target speech. Hereafter, we introduce a straightforward extension of the method in [2] to binaural hearing-aid system consisting of two inputs and two output for human ears [3]. The noise component estimated by ICA is used to determine *a posteriori* SNR and the optimal spectral gain applied to the L and R channel signal, neglecting phase information; this enables the method to realize error-robust noise reduction.

First, in the reference path, we perform signal separation using an unmixing matrix $\boldsymbol{W}(f)$ obtained by ICA. The output signal $\boldsymbol{y}(f,\tau) = [y_{\mathrm{s}}(f,\tau), y_{\mathrm{n}}(f,\tau)]$ is given by

$$\boldsymbol{y}(f,\tau) = [y_{\mathrm{s}}(f,\tau), y_{\mathrm{n}}(f,\tau)] = \boldsymbol{W}(f)\boldsymbol{x}(f,\tau), \tag{1}$$

$$\boldsymbol{W}^{[p+1]}(f) = \boldsymbol{W}^{[p]}(f) + \mu\left[\boldsymbol{I} - \left\langle \psi(y(f,\tau))y^{\mathrm{H}}(f,\tau)\right\rangle_{\tau}\right]\boldsymbol{W}^{[p]}(f), \tag{2}$$

where $y_{\mathrm{s}}(f,\tau)$ is the estimated target speech signal, $y_{\mathrm{n}}(f,\tau)$ is the estimated noise signal, μ is the step-size parameter, $[p]$ is used to express the value at the pth step of the iteration and $\boldsymbol{I}$ is the identity matrix. In addition, $\langle\cdot\rangle_{\tau}$ denotes a time-averaging operator, $\boldsymbol{M}^{\mathrm{H}}$ denotes the conjugate transpose of matrix $\boldsymbol{M}$, and $\psi(\cdot)$ is an appropriate nonlinear activation function. The estimated target speech signal is discarded as it is not required because we want to estimate only the noise component. Instead, we construct a *noise-only vector*,

$$\boldsymbol{y}^{(\mathrm{noise})}(f,\tau) = [0, y_{\mathrm{n}}(f,\tau)]^{\mathrm{T}}, \tag{3}$$

from the output signal obtained by ICA using (1). Following this, we apply the projection back (PB) operation to remove the ambiguity of amplitude and construct the estimated noise signal at both ears,

$$[\zeta_{\mathrm{L}}(f,\tau), \zeta_{\mathrm{R}}(f,\tau)]^{\mathrm{T}} = \boldsymbol{W}^{-1}(f)\boldsymbol{y}^{(\mathrm{noise})}(f,\tau), \tag{4}$$

where $\zeta_{\mathrm{L}}(f,\tau)$ and $\zeta_{\mathrm{R}}(f,\tau)$ are the noise components estimated at the left and right channels, respectively.

Next, using the estimated noise components, we derive the gain function optimized on the basis of the MMSE STSA estimation criterion. Hereinafter, for convenience, the subscript $* \in \{\mathrm{L}, \mathrm{R}\}$ is used to indicate the index of channels. The estimated *a posteriori* signal-to-noise ratio (SNR) $\hat{\gamma}_*(f,\tau)$ is obtained from the observed signal $x_*(f,\tau)$ and the estimated noise signals (4) as

$$\hat{\gamma}_*(f,\tau) = |x_*(f,\tau)|^2 \left\{\mathrm{E}\{|\zeta_*(f,\tau)|^2\}_{(\tau-\tau_{\mathrm{th}})}^{\tau}\right\}^{-1}, \tag{5}$$

where τ_{th} is a smoothing parameter denoting a certain time frame window and $\mathrm{E}\{\cdot\}_A^B$ denotes the expectation operator from A to B. Note that we can momentarily estimate the instantaneous *a posteriori* SNR (5) by utilizing the noise signal estimated by ICA (4), in contrast

* "Multimodal Blind Speech Extraction Using Image Tracking in HRTF-Informed Binaural Hearing-Aid System," By Kotaro Yoshie, Hiroshi Saruwatari, and Satoshi Nakamura (Nara Institute of Science and Technology).

to the case using the original MMSE STSA estimator [4]. Therefore, we consider that our proposed method can suppress nonstationary noise more efficiently than the conventional MMSE STSA estimator.

Next, using (5), the *a priori* SNR estimate $\hat{\xi}_*(f,\tau)$ is given as

$$\hat{\xi}_*(f,\tau) = \alpha\hat{\gamma}_*(f,\tau-1)G_*^2(f,\tau-1) + (1-\alpha)P[\hat{\gamma}_*(f,\tau)-1], \quad (6)$$

where α is the weighting factor of the decision-directed estimation, $G_*(f,\tau)$ is a spectral gain function, and the operator $P[\cdot]$ is a flooring function in which the negative input is floored to zero. Also, $G_*(f,\tau)$ is given as

$$G_*(f,\tau) = \Gamma(1.5)\frac{\sqrt{\nu_*(f,\tau)}}{\hat{\gamma}_*(f,\tau)}\exp\left(\frac{\nu_*(f,\tau)}{2}\right) \cdot\left[\{1+\nu_*(f,\tau)\}I_0\left(\frac{\nu_*(f,\tau)}{2}\right) +\nu_*(f,\tau)I_1\left(\frac{\nu_*(f,\tau)}{2}\right)\right], \quad (7)$$

where $\Gamma(\cdot)$ denotes the gamma function and $I_0(\cdot)$ and $I_1(\cdot)$ denote modified Bessel function of the zeroth and first order, respectively. Moreover, $\nu_*(f,\tau)$ is defined by

$$\nu_*(f,\tau) = \hat{\xi}_*(f,\tau)\hat{\gamma}_*(f,\tau)\{1+\hat{\xi}_*(f,\tau)\}^{-1}. \quad (8)$$

Finally, noise reduction is carried out as follows:

$$z_*(f,\tau) = G_*(f,\tau)x_*(f,\tau), \quad (9)$$

where $z_*(f,\tau)$ is the final output of this method for both ears.

2.2 Equi-Binaural Optimal Spectral Gain Estimation

In [3], it is essential to apply *equivalent* spectral gains for L and R channels to increase the localization accuracy. Therefore, to obtain the optimal spectral gain that maintains the localization accuracy, we introduce spectral gain that minimizes the residual noise power in terms of the MMSE under the condition that the spectral gains are equivalent in both channel. Hereafter, we call this gain the *equi-binaural optimal spectral gain*. The derivation of the equi-binaural optimal spectral gain can be formulated as the minimization problem of the following error e:

$$e = \mathrm{E}\Big[\{|h_\mathrm{L}(f)s(f,\tau)| - G(f,\tau)|x_\mathrm{L}(f,\tau)|\}^2 + \{|h_\mathrm{R}(f)s(f,\tau)| - G(f,\tau)|x_\mathrm{R}(f,\tau)|\}^2\Big] \quad (10)$$

where $G(f,\tau)$ is the equi-binaural spectral gain, which is considered as a variable. The optimization problem based on (10) is given by

$$\begin{aligned} G_\mathrm{opt}(f,\tau) &= \underset{G(f,\tau)}{\mathrm{argmin}}\,\mathrm{E}\Big[\{|h_\mathrm{L}(f)s(f,\tau)| - G_\mathrm{L}(f,\tau)|x_\mathrm{L}(f,\tau)|\}^2 \\ &\quad + \{|h_\mathrm{R}(f)s(f,\tau)| - G_\mathrm{R}(f,\tau)|x_\mathrm{R}(f,\tau)|\}^2 \\ &\quad + \{(G(f,\tau) - G_\mathrm{L}(f,\tau))|x_\mathrm{L}(f,\tau)|\}^2 \\ &\quad + \{(G(f,\tau) - G_\mathrm{R}(f,\tau))|x_\mathrm{R}(f,\tau)|\}^2 + 2A\Big], \end{aligned} \quad (11)$$

where $G_\mathrm{opt}(f,\tau)$ is the equi-binaural optimal spectral gain to by estimated, $G_\mathrm{L}(f,\tau)$ and $G_\mathrm{R}(f,\tau)$ are individual spectral gains for the L and R channels, respectively, which are auxiliary parameters for calculating an approximate solution of $G_\mathrm{opt}(f,\tau)$ because the direct Bayesian estimation of $G_\mathrm{opt}(f,\tau)$ is difficult. In addition, A is related to the correlation between the estimation error and the observed signal in each channel when we estimate the target speech signals in the L and R channels using the parameters $G_\mathrm{L}(f,\tau)$ and $G_\mathrm{R}(f,\tau)$, and is defined by

$$\begin{aligned} A &= \{G(f,\tau) - G_\mathrm{L}(f,\tau)\}\{G_\mathrm{L}(f,\tau)|x_\mathrm{L}(f,\tau)| \\ &\qquad -|h_\mathrm{L}(f)s(f,\tau)|\}|x_\mathrm{L}(f,\tau)| \\ &\quad +\{G(f,\tau) - G_\mathrm{R}(f,\tau)\}\{G_\mathrm{R}(f,\tau)|x_\mathrm{R}(f,\tau)| \\ &\qquad -|h_\mathrm{R}(f)s(f,\tau)|\}|x_\mathrm{R}(f,\tau)|. \end{aligned} \quad (12)$$

We discuss the minimization of (11). First, the 1st and 2nd terms of the right-hand side correspond to the problem of target speech estimation in each channel. These terms can be minimized if we obtain the optimal values of $G_\mathrm{L}(f,\tau)$ and $G_\mathrm{R}(f,\tau)$ using the MMSE STSA estimator described in Sect. 2.1. Next, A in the 5th term of the right-hand side can be disregarded if the parameters $G_\mathrm{L}(f,\tau)$ and $G_\mathrm{R}(f,\tau)$ closely estimate the target speech signals by approximately considering A to be negligible. Hence, the residual 3rd and 4th terms, i.e., $\{(G(f,\tau) - G_\mathrm{L}(f,\tau))|x_\mathrm{L}(f,\tau)|\}^2 + \{(G(f,\tau) - G_\mathrm{R}(f,\tau))|x_\mathrm{R}(f,\tau)|\}^2$, are the candidates to be minimized. This problem can be formulated as

$$\begin{aligned} G_\mathrm{opt}(f,\tau) \fallingdotseq \underset{G(f,\tau)}{\mathrm{argmin}}\,\mathrm{E}\Big[&\{(G(f,\tau) - G_{\mathrm{L}_\mathrm{opt}}(f.\tau))|x_\mathrm{L}(f,\tau)|\}^2 \\ &+ \{(G(f,\tau) - G_{\mathrm{R}_\mathrm{opt}}(f.\tau))|x_\mathrm{R}(f,\tau)|\}^2\Big], \end{aligned} \quad (13)$$

subject to

$$G_{\mathrm{L}_\mathrm{opt}}(f,\tau) = \underset{G_\mathrm{L}(f,\tau)}{\mathrm{argmin}}\,\mathrm{E}\Big[\{|h_\mathrm{L}(f)s(f,\tau)| - G_\mathrm{L}(f,\tau)|x_\mathrm{L}(f,\tau)|\}^2\Big], \quad (14)$$

$$G_{\mathrm{R}_\mathrm{opt}}(f,\tau) = \underset{G_\mathrm{R}(f,\tau)}{\mathrm{argmin}}\,\mathrm{E}\Big[\{|h_\mathrm{R}(f)s(f,\tau)| - G_\mathrm{R}(f,\tau)|x_\mathrm{R}(f,\tau)|\}^2\Big], \quad (15)$$

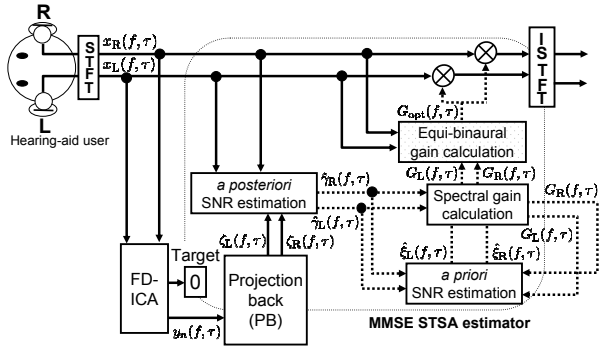

Fig. 1 Block diagram of sound-localization-preserved MMSE STSA estimator [3].

where $G_{\mathrm{L}_{\mathrm{opt}}}(f,\tau)$ and $G_{\mathrm{R}_{\mathrm{opt}}}(f,\tau)$ are the L- and R- channel optimal spectral gains, respectively.

To solve (13) we first obtain $G_{\mathrm{L}_{\mathrm{opt}}}(f,\tau)$ and $G_{\mathrm{R}_{\mathrm{opt}}}(f,\tau)$ from the MMSE STSA estimator in (14) and (15), then by substituting them into (13), we solve the following equation in $G(f,\tau)$:

$$\frac{\partial}{\partial G(f,\tau)}\Big(\{(G(f,\tau)-G_{\mathrm{L}_{\mathrm{opt}}}(f,\tau))|x_{\mathrm{L}}(f,\tau)|\}^2 +\{(G(f,\tau)-G_{\mathrm{R}_{\mathrm{opt}}}(f,\tau))|x_{\mathrm{R}}(f,\tau)|\}^2\Big)=0. \tag{16}$$

The solution of (16) is given by

$$G_{\mathrm{opt}}(f,\tau)=\frac{G_{\mathrm{L}_{\mathrm{opt}}}(f,\tau)|x_{\mathrm{L}}(f,\tau)|^2+G_{\mathrm{R}_{\mathrm{opt}}}(f,\tau)|x_{\mathrm{R}}(f,\tau)|^2}{|x_{\mathrm{L}}(f,\tau)|^2+|x_{\mathrm{R}}(f,\tau)|^2}. \tag{17}$$

Figure 1 shows the configuration of this method based on the equi-binaural optimal spectral gain.

3 Proposed Method

We propose multimodal blind speech extraction using image tracking in hearing-aid system and the multichannel MMSE STSA estimator, which explicitly introduces a HRTF in estimation of the target speech instead of single-channel MMSE STSA estimators for each channel. Figure 2 shows a block diagram of this method.

In the conventional method [3], the equi-binaural optimal spectral gain is approximately estimated using optimal spectral gains independently obtained from single-channel MMSE STSA estimation for each channel. Thus, the conventional method does not use binaural cues such as the HRTF. This fact often causes serious degradation of the noise reduction performance, particularly in the case that the target speech source is laterally located on the right- (left-) hand side. Therefore, in this section, we introduce the *oracle* method using the multichannel MMSE STSA estimator [5], which explicitly uses the HRTF in the estimation of the target speech instead of single-channel MMSE STSA estimators for each channel. This method can give the equi-binaural optimal spectral gain without approximation.

The multichannel MMSE STSA estimator can be viewed as a cascade of minimum variance distortionless response (MVDR) beamforming and the single-channel MMSE STSA estimator [5]. The output of MVDR beamforming is given by

$$Y(f,\tau)=\frac{\boldsymbol{h}(f)^{\mathrm{H}}\boldsymbol{\Sigma}_{\mathrm{N}}(f,\tau)^{-1}\boldsymbol{x}(f,\tau)}{\boldsymbol{h}(f)^{\mathrm{H}}\boldsymbol{\Sigma}_{\mathrm{N}}(f,\tau)^{-1}\boldsymbol{h}(f)}, \tag{18}$$

where $\boldsymbol{\Sigma}_{\mathrm{N}}(f,\tau)$ is a noise covariance matrix momentarily estimated by ICA, which is defined as

$$\boldsymbol{\Sigma}_{\mathrm{N}}(f,\tau) =\mathrm{E}\{[\zeta_{\mathrm{L}}(f,\tau),\zeta_{\mathrm{R}}(f,\tau)]^{\mathrm{H}}[\zeta_{\mathrm{L}}(f,\tau),\zeta_{\mathrm{R}}(f,\tau)]\}^{\tau}_{(\tau-\tau_{\mathrm{th}})}. \tag{19}$$

Then, the a posteriori SNR for $Y(f,\tau)$ is calculated as

$$\hat{\gamma}_{\mathrm{Y}}(f,\tau)=|Y(f,\tau)|^2\boldsymbol{h}(f)^{\mathrm{H}}\boldsymbol{\Sigma}_{\mathrm{N}}(f,\tau)^{-1}\boldsymbol{h}(f). \tag{20}$$

Next, using (20), the a priori SNR estimate $\hat{\xi}_{\mathrm{Y}}(f,\tau)$ is given as

$$\hat{\xi}_{\mathrm{Y}}(f,\tau) =\alpha\hat{\gamma}_{\mathrm{Y}}(f,\tau-1)G_{\mathrm{Y}}(f,\tau-1)^2 +(1-\alpha)P[\hat{\gamma}_{\mathrm{Y}}(f,\tau)-1], \tag{21}$$

where $G_{\mathrm{Y}}(f,\tau)$ is the spectral gain function of the multichannel MMSE STSA estimator and is defined as

$$G_{\mathrm{Y}}(f,\tau) =\Gamma(1.5)\frac{\sqrt{\nu_{\mathrm{Y}}(f,\tau)}}{\hat{\gamma}_{\mathrm{Y}}(f,\tau)}\exp(\nu_{\mathrm{Y}}(f,\tau)/2) \cdot\left[\{1+\nu_{\mathrm{Y}}(f,\tau)\}I_0(\nu_{\mathrm{Y}}(f,\tau)/2) +\nu_{\mathrm{Y}}(f,\tau)I_1(\nu_{\mathrm{Y}}(f,\tau)/2)\right], \tag{22}$$

$$\nu_{\mathrm{Y}}(f,\tau) =\hat{\xi}_{\mathrm{Y}}(f,\tau)\hat{\gamma}_{\mathrm{Y}}(f,\tau)\{1+\hat{\xi}_{\mathrm{Y}}(f,\tau)\}^{-1}. \tag{23}$$

Finally, the resultant equi-binaural optimal spectral gain without approximation that can strictly minimize (10) is given by

$$G_{\mathrm{oracle}}(f,\tau) =\frac{|h_{\mathrm{L}}(f)||x_{\mathrm{L}}(f,\tau)|+|h_{\mathrm{R}}(f)||x_{\mathrm{R}}(f,\tau)|}{|x_{\mathrm{L}}(f,\tau)|^2+|x_{\mathrm{R}}(f,\tau)|^2} \cdot G_{\mathrm{Y}}(f,\tau)Y(f,\tau). \tag{24}$$

We assume that a camera is mounted on the hearing-aid user's head, and prepare the multiple HRTFs for some degrees in advance and switch them according to the camera's image tracking result. However, it is quite difficult to estimate the HRTF blindly in practice because the shape of the hearing-aid users' heads greatly varies. Therefore, we use single "typical" HRTF and some room transfer function (RTF).

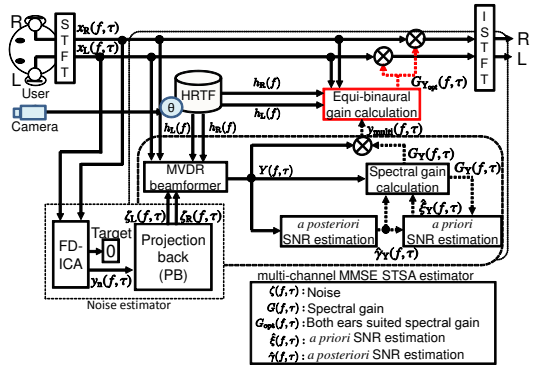

Fig. 2 Block diagram of proposed method.

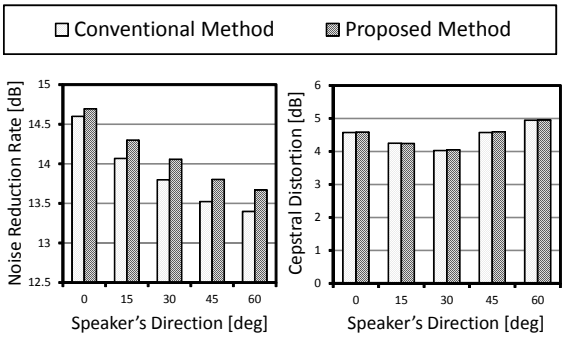

Fig. 3 Results of noise reduction rate and cepstral distortion.

4 Experiment and Results

To confirm the effectiveness of the proposed method, we conducted an experiment. In this experiment, we compared the noise reduction performance of the conventional and proposed methods by objective evaluations.

We used 20 utterances (10 males and 10 females from the Japanese newspaper dictation database) as the target speech signals and white Gaussian noise signal. Furthermore, the binaural speech signals from five directions, 0, 15, 30, 45 or 60 degrees, were obtained by convolution of the target speech signals and HRTF of each direction. We used 10 patterns of RTFs (for 10 persons) to confirm that the proposed method achieves a larger amount of noise reduction than the conventional method.

The test data were obtained by combining the binaural speech signals and noise signals. The input SNR was set to 0 dB. The weighting factor of the decision-directed estimation is 0.97 in MMSE STSA estimation. To compare the amount of noise reduction and sound quality, we calculated the noise reduction rate (NRR) (output SNR - input SNR in dB) and cepstral distortion (CD) (a measure of the degree of spectral envelope distortion) of the processed signals.

Figure 3 shows the results for the average NRR and CD of all the target speakers and all RTFs patterns at each direction. In Fig. 3, we can confirm that NRR of the proposed method is superior to that of the conventional method, especially in the case of laterally located sources. Also, CDs of both methods are the same.

5 Conclusions

We presented the Multimodal blind speech extraction using image tracking in hearing-aids system with some degrees HRTF filter bank.

From the evaluation, we confirmed that to apply different speech priors for each signals of the left and right channels is effective for improvement of noise reduction performance. The proposed method is advantageous to the application of binaural hearing-aid systems.

Reference

[1] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol.36, pp.287–314, 1994.

[2] R. Okamoto, Y. Takahashi, H. Saruwatari, K. Shikano, "MMSE STSA estimator with nonstationary noise estimation based on ICA for high-quality speech enhancement," *Proc. ICASSP2010*, pp.4778–4781, 2010.

[3] H. Saruwatari, M. Go, R. Okamoto, K. Shikano, "Binaural hearing aid using sound-localization-preserved MMSE STSA estimator with ICA-based noise estimation," *Proc. IWAENC*, 2010.

[4] Y. Ephraim, D. Malah. "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol.ASSP-32, no.6 pp.1109–1121, 1984.

[5] R. C. Hnedriks, R. Heusdens, U. Kjems, J. Jensen, "On optimal multichannel mean-square error estimators for speech enhancement," *IEEE Signal Processing Letters*, vol.16, no.10, pp.885–888, 2009.