OUT-OF-HEAD SOUND LOCALIZATION USING ADAPTIVE INVERSE FILTER

Toshiharu Horiuchi, Haruhide Hokari, and Shoji Shimada

Nagaoka University of Technology,
Nagaoka, Niigata 940-2188, Japan
E-mail: horiuchi@audio.nagaokaut.ac.jp

ABSTRACT

It is well known that the transfer functions for out-of-head sound localization differ with the listener. Our goal is to find a method that easily realizes excellent sound localization without measuring the individual transfer functions. This paper proposes an out-of-head sound localization system that uses adaptive inverse filtering without any measuring preprocesses. This system estimates the inverse Ear Canal Transfer Function (ECTF), and can adaptively obtain the transfer function to fit the listener in real-time. This paper also proposes an adaptive inverse filtering method for out-of-head sound localization that differs from the filtered-x method. In addition, the relationship between convergence time and initial value is assessed. It is clarified that the proposed method is more effective, in terms of convergence, if the initial value is the average of many listeners' impulse responses.

1. INTRODUCTION

With the recent rapid growth of signal processing technologies, binaural techniques have become more attractive for realizing high performance audio-visual systems. Out-of-head sound localization is indispensable for really effective virtual sound systems. Such systems give the illusion that the person or persons you are communicating with are in your room. This kind of technology is very much in demand.

It is well known that the transfer functions for out-of-head sound localization are individually different with each listener[1][2]. Out-of-head sound localization mainly depends on two transfer functions: Head Related Transfer Function (HRTF), and Ear Canal Transfer Function (ECTF). The shapes of the listener's face, ear, and ear canal, strongly influence these two transfer functions, and it is known that the quality of sound localization may be poor if the individual's transfer functions are not accurately reproduced[3]. For that reason, when we try to achieve out-of-head sound localization, it become necessary to obtain the transfer functions of each listener. However, measuring the transfer functions is not user-friendly nor practical. Still more, the conventional method of selecting a pre-measured transfer function is troublesome[4]. Our final goal is to find a method that easily realizes excellent sound localization without measuring the individual impulse responses of the listeners' transfer functions.

This paper proposes an out-of-head sound localization system that uses adaptive inverse filtering without any measuring preprocesses. This system estimates only the inverse ECTF, and can adaptively obtain the transfer function to fit the listener in real-time. This paper also introduces an adaptive inverse filtering method for out-of-head sound localization that is different from the filtered-x method. In addition, the relationship between convergence time and initial value is deduced from the coefficient error vector in order to decide the initial value of the adaptive inverse filter. A computer simulation is carried out using the ECTF measured from many listeners, and the optimum initial value from the viewpoint of convergence time is examined.

2. DEFINITION AND MEASUREMENT OF THE TRANSFER FUNCTIONS

The principle of out-of-head sound localization is to reproduce the sound waveforms of the actual sound field at the listeners' eardrums using stereo headphones or earphones.

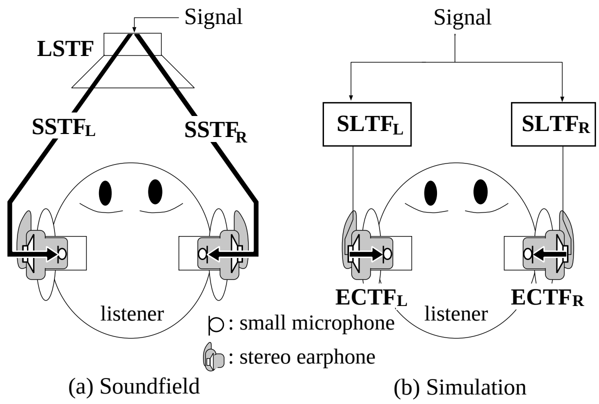


Fig. 1. The relationship of the transfer functions.

Figure 1 shows the relationship of the transfer functions used in this paper. The Spatial Sound Transfer Function (SSTF) models the conversion of the input signal of the loudspeaker into the output signal of the microphone in the ear canal of the listener. The definition of the HRTF described by Blauert[1]. The relationship between SSTF and HRTF is given by

$$\mathrm{HRTF} \cong \mathrm{SSTF} \cdot \mathrm{LSTF}^{-1}. \tag{1}$$

The LoudSpeaker Transfer Function (LSTF) models the conversion of the input signal of the loudspeaker into the output signal of a reference microphone.

ECTF models the conversion of the input signal of the earphone into the output signal of the microphone in the ear canal of the listener.

From these transfer functions, the Sound Localization Transfer Function (SLTF) can be derived using Eq.(2).

$$\mathrm{SLTF} = \mathrm{SSTF} \cdot \mathrm{LSTF}^{-1} \cdot \mathrm{ECTF}^{-1} \tag{2}$$

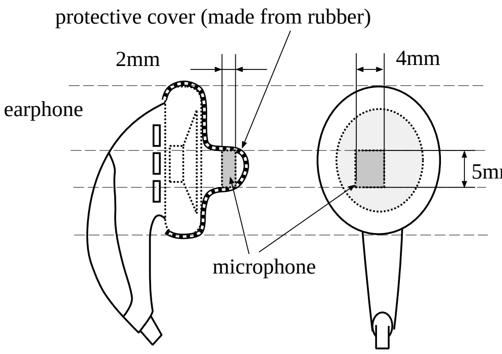


Fig. 2. Stereo earphone-microphone.

Figure 2 shows a sketch of the typical stereo earphone-microphone; it consists of a small microphone and an intra-concha type stereo earphone. The size of electret condenser microphone is 3.6 $\times$ 4.7 $\times$ 2.1mm. Its frequency response is relatively flat from 200Hz to 14kHz. The stereo earphone is a SONY MDR-ED238. Since this earphone is quite permeable to sound, we can easily hear environmental sounds.

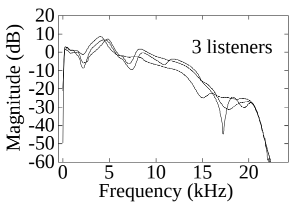


Fig. 3. The frequency magunitude characteristics examples of ECTF.

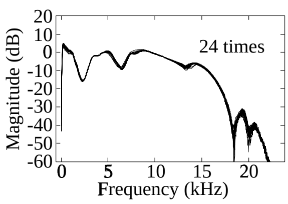


Fig. 4. The characteristics of a certain listener's ECTFs as measured 24 times.

Figure 3 shows, for the same setup and 3 listeners, their right side ECTF. Note that these characteristics were significantly different. Figure 4 shows the characteristics of a listener's ECTFs measured 24 times. In the experiment the earphones were removed after each measurement. There is almost no change in these characteristics, which indicates that the characteristics do not depend on the exact location of the earphone microphone combination in the cavum concha. Therefore, it is necessary to measure the listener's own ECTF.

3. PROPOSED METHOD

Figure 5 shows how the method adaptively obtains the inverse ECTF in real-time by setting an adaptive inverse filter at the ear.

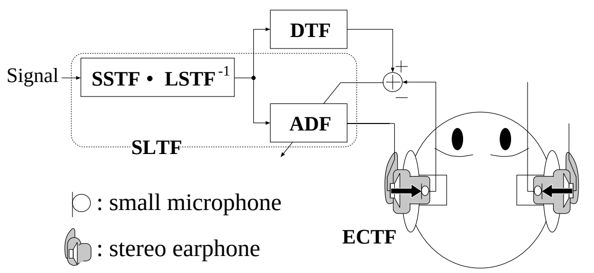


Fig. 5. Out-of-head sound localization using adaptive inverse filtering.

The Desired Transfer Function (DTF) is a BPF with the addition of some delay time to give causality and stability to inverse filtering. The transfer function $\mathrm{SSTF} \cdot \mathrm{LSTF}^{-1}$ is installed as a prefilter. SSTF uses the transfer function common to many listeners. Since ECTF differs more widely with the listener than SSTF[5], ECTF is adapted to each listener's ear. This means that SLTF can be estimated in real-time without requiring measurement-based preprocesses.

4. COMPARISON OF ADAPTIVE INVERSE FILTER METHODS

4.1. Method 1

One of the most widely used adaptive filter algorithms for active noise control (ANC)[6], is the filtered-x type[7]. A block diagram of the out-of-head sound localization system with the filtered-x type algorithm is shown in Figure 6.

The prefilter $\mathrm{SSTF} \cdot \mathrm{LSTF}^{-1}$ was omitted in Figure 6, so input signal $x(n)$ is the output signal of the prefilter. The unknown system $\mathbf{c}$ is the listener's impulse response (ECTF). The reference system $\widehat{\mathbf{c}}$ is an approximation of the listener's impulse response. The desired system $\mathbf{b}$ is the impulse response of BPF with additional delay time. $x(n)$, $y(n)$, $z'(n)$, $u(n)$, $d(n)$, and $e(n)$ in Figure 6 are the input signals, the adaptive filter output, the output signal from the unknown system, the filtered reference signals, the desired system output, and the error signal, respectively. $v(n)$ is the noise signals, here $v(n) = 0$. n is the discrete time index.

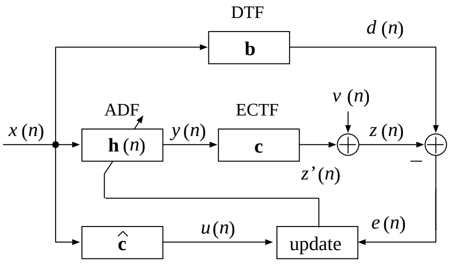


Fig. 6. Block diagram of filtered-x type algorithm implementation (Method 1).

The adaptive inverse filter $\mathbf{h}(n)$ is adjusted to make $z(n)$ approach $d(n)$ by adding adjustment vector $\Delta\mathbf{h}(n)$

$$\mathbf{h}(n+1) = \mathbf{h}(n) + \mu\Delta\mathbf{h}(n), \tag{3}$$

where μ is a scaling factor called the step size, which controls the convergence rate and the amount of residual error.

However, there is a problem with the adaptive filter in that it often fails to converge due to the modeling error of the filtered-x type algorithm.

4.2. Method 2

This method is the one proposed in this paper. The initial value is given as $\mathbf{h}'(n)$ and $\mathbf{h}(n)$ at $n = 1$.

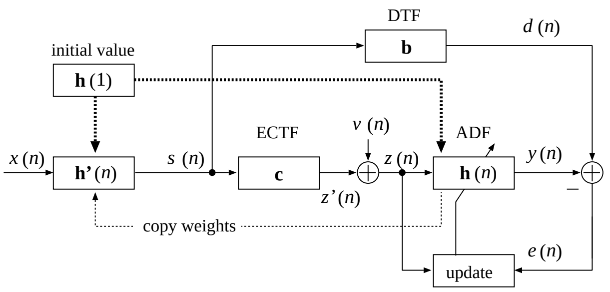


Fig. 7. Block diagram of the adaptive inverse filter with initial value (Method 2).

$x(n)$, $s(n)$, $z'(n)$, $y(n)$, $d(n)$, and $e(n)$ in Figure 7 are the input signal, the input signal to the unknown system, the output signal from the unknown system, the adaptive filter output, the desired system output, and the error signal, respectively. $v(n)$ represents the noise signals.

The advantages of this method is that it is simpler to implement and contains no modeling error.

5. INITIAL VALUE OF ADAPTIVE INVERSE FILTER

With method 2 the filter must be given an initial value. This section describes the relationship between initial value and convergence time and examines the optimum initial value from the viewpoint of convergence time.

The coefficient error vector is defined by

$$\mathbf{E}(n+1) = \mathbf{h}(n+1) - \mathbf{w}. \tag{4}$$

Here, $\mathbf{w}$ is the optimal value. When the input signal of adaptive filter is assumed to be white noise, $\mathbf{E}(n+1)$ is given by

$$\mathbf{E}(n+1) = (1-\mu)^n\,\mathbf{E}(1). \tag{5}$$

The coefficient error vector norm is given as

$$\|\mathbf{E}(n+1)\|^2 = (1-\mu)^{2n}\,\|\mathbf{E}(1)\|^2. \tag{6}$$

$\mathbf{h}^{(j)}(1)$ is the initial value j, $\mathbf{w}_i$ is the optimal value i while $\mathbf{h}_i^{(j)}(n+1)$ is the adaptive inverse filter coefficient after time n,

$$\|\mathbf{E}_i^{(j)}(n+1)\|^2 = (1-\mu)^{2n}\|\mathbf{h}^{(j)}(1) - \mathbf{w}_i\|^2. \tag{7}$$

As $(1-\mu)^2 = \exp(-a)$, it converts to the sampling time t,

$$\|\mathbf{E}_i^{(j)}(1)\|^2 = \|\mathbf{h}^{(j)}(1) - \mathbf{w}_i\|^2 = \exp(-at_1^{(j)})\|\mathbf{w}_i\|^2, \tag{8}$$

$$\begin{aligned}\|\mathbf{E}_i^{(j)}(n+1)\|^2 &= \|\mathbf{h}_i^{(j)}(n+1) - \mathbf{w}_i\|^2 \\ &= \exp(-at_{n+1}^{(j)})\|\mathbf{w}_i\|^2.\end{aligned} \tag{9}$$

Therefore, convergence time $T_i^{(j)} (= t_{n+1}^{(j)} - t_1^{(j)})$

$$T_i^{(j)} = \frac{1}{a}\ln\left(\frac{\|\mathbf{h}^{(j)}(1) - \mathbf{w}_i\|^2}{\|\mathbf{h}_i^{(j)}(n+1) - \mathbf{w}_i\|^2}\right) \tag{10}$$

is given. The average convergence time is deduced from the coefficient error vector, $T_{ave}^{(j)}$ of the initial value $\mathbf{h}^{(j)}(1)$, is given by

$$T_{ave}^{(j)} = \frac{1}{N}\sum_{i=1}^{N} T_i^{(j)}. \tag{11}$$

It is difficult to acquire the initial value j in which $T_{ave}^{(j)}$ becomes minimum value by arithmetic analysis method, from Eq.(11). Therefore, we used an alternative method, which is computer simulation.

6. COMPUTER SIMULATION

The input signal $x(n)$, we used white noise. In determining the unknown impulse response $\mathbf{c}$, we used 118 ECTF, which was measured from both ears of 59 listeners. To determine the desired impulse response $\mathbf{b}$, we used the BPF (20Hz-15kHz) with additional delay time of 5.33ms. In determining the reference impulse response $\widehat{\mathbf{c}}$ required in method 1, the average of 118 ECTF was used.

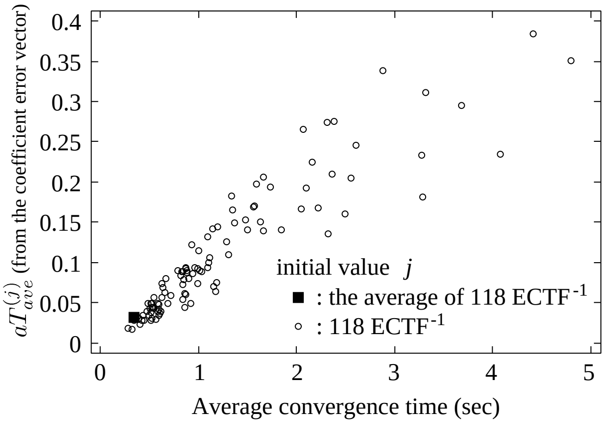
The estimation error is defined by

$$\varepsilon(n) = 10\log_{10}\frac{\sum_f(|B(f)| - |C(f)H(f,n)|)^2}{\sum_f|B(f)|^2} \quad \text{(dB)}, \tag{12}$$

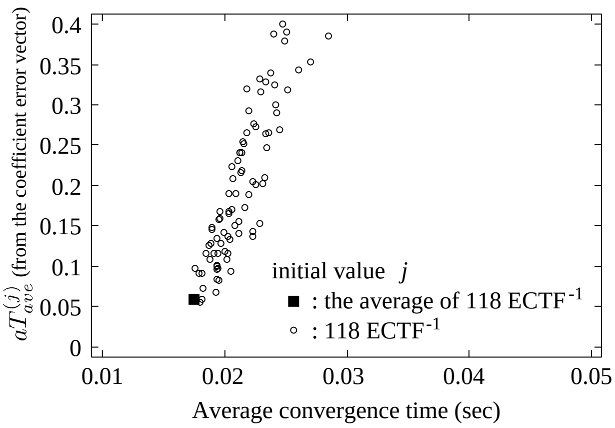
where $B(f)$, $C(f)$, and $H(f,n)$ are converted into the frequency domain equivalents of $\mathbf{b}$, $\mathbf{c}$, and $\mathbf{h}(n)$, respectively.

The convergence time is defined as the sampling time t taken to achieve the estimation error $\varepsilon = -10$dB down.

The other simulation parameters are shown below: Step size $\mu = 0.5$, Affine Projection (AP) algorithm's dimension $r = 2$, and Block length $L = 257$ for Method 1. It is necessary to use the block processing in order to apply AP algorithm to the filtered-x type, because AP algorithm needs error signal vector of dimension r.



(a) Method 1



(b) Method 2

Fig. 8. The average convergence time.

The simulation results are shown in Figure 8 and Table 1. Figure 8 shows the relationship between average convergence time and Eq.(11). The result (a) and (b) in Figure 8 indicate that (1) the average convergence time becomes short when the average of many listeners' ECTF^{-1} is used as the initial value, (2) the convergence time of method 2 is also generally shorter than of that of method 1, and (3) the dispersion of method 2 is smaller than of method 1.

Table 1 shows the 1st order autocorrelation coefficient of the input signal of adaptive filter. σ in Table 1 is the standard deviation. In this case, initial value $\mathbf{h}'(1)$, the average of 118 ECTF^{-1} was used for method 2. The 1st order

Table 1. The 1st order autocorrelation coefficient of the input signal of adaptive filter

$\phi_{uu}(1)$ (Method 1)	$\phi_{zz}(1)$ (Method 2)
0.66	0.43 ($\sigma = 0.13$)

autocorrelation coefficient of method 2 is smaller than of method 1.

Therefore, these results indicate that method 2 is better than method 1 for an out-of-head sound localization system.

7. CONCLUSION

This paper proposes an out-of-head sound localization system that uses an adaptive inverse filter that needs no preprocessing measurements. This system estimates the inverse ECTF, and can adaptively obtain the transfer function to fit the listener in real-time. This paper also proposed an adaptive inverse filtering method for out-of-head sound localization that is different from the filtered-x method. In addition, the relationship between convergence time and initial value is deduced from the coefficient error vector in order to decide the initial value given to the adaptive inverse filter. A computer simulation was carried out using the ECTF measured from many listeners, and the optimum initial value from the viewpoint of convergence time was examined. As a result, we clarified that the proposed method is more effective, in terms of convergence, if the initial value of the adaptive inverse filter is set as the average of many listeners' impulse responses. This is extremely effective since it makes the initial value easy to set.

8. REFERENCES

[1] J. Blauert, *Spatial Hearing*, pp. 78-93, The MIT Press, 1997.

[2] H. Møller, M.F. Sørensen, C.B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?," *J. Audio Eng. Soc.*, vol. 44, no. 6, pp. 451-469, 1996.

[3] E.M. Wenzel, M. Arruda, D.J. Kistler, and F.L. Wightman, "Localization using non-individualized head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 94, no. 1, pp. 111-123, 1993.

[4] S. Shimada, N. Hayashi, and S. Hayashi, "A clustering method for sound localization transfer function," *J. Audio Eng. Soc.*, vol. 42, no. 7/8, pp. 577-584, 1994.

[5] S. Yano, H. Hokari, and S. Shimada, "A Study on Personal Difference in the Transfer Functions of Sound Localization Using Stereo Earphones," *IEICE Trans. Fundamentals*, vol. E83-A, no. 5, pp. 877-887, May 2000.

[6] P.A. Nelson and S.J. Eliott, *Active Control of Sound*, Academic Press, 1992.

[7] B. Widrow and S.D. Stearns, *Adaptive Signal Processing*, pp. 288-297, Prentice-Hall, 1985.