Evaluation of white noise gain in a binaural system for microphone arrays *

☆ SALVADOR César[1], SAKAMOTO Shuichi[1], TREVIÑO Jorge[1], and SUZUKI Yôiti[1]

[1]Res. Inst. Elect. Comm. and Grad. Sch. Info. Sci., Tohoku University

1 Introduction

Binaural systems aim to convey high-definition listening experiences by synthesizing the sound pressure signals at the ears of the listeners [1]. Microphone array recordings and data sets of head-related transfer functions (HRTFs) are combined with this aim [2–6]. Accurate binaural synthesis requires to capture an auditory scene with high spatial resolution. This implies the use of a large number of microphones, together with the complications that arise when controlling a large number of signals.

Predicting the performance of a microphone array in real conditions is a crucial stage for its design. There is a particular necessity of models for predicting the robustness of an array to the transducer noise, the microphone positioning error, and the effects of space discretization, among other sources of perturbation. These issues have been widely addressed in the theory of beamforming [7], where sensor arrays are used to synthesized spatial patterns. Because binaural synthesis can also be formulated as a beamforming problem, where the spatial patterns are given by the HRTF datasets, beamforming constitutes an adequate framework for investigating the effects of noise in binaural synthesis.

In beamforming, the white noise gain is defined as the output power due to spatially uncorrelated white noise at the sensors [7]. This variable is used as a general measure for robustness to such kind of noise. The kind of arrays examined in conventional beamforming, however, are typically limited to low spatial resolutions [8]. To predict a more precise improvement in the signal-to-noise ratio for higher resolution arrays, additional analyses are required.

In this paper, the propagation of noise through a binaural system is investigated based on the gain in signal-to-noise ratio from the input of the microphone array to the output of the binaural system. For this purpose, a linear model of a system in arbitrary geometries is formulated. The model takes into consideration the contributions of additive white noise, which is assumed spatially uncorrelated and with a uniform distribution of energy around the array.

Spherical arrays and spherical HRTF datasets are of interest in binaural synthesis. The model is therefore tested in spherical geometries. Numerical experiments consider models of high resolution arrays. The results can be used as an objective design recommendation to identify the number of microphones that are necessary to synthesize the spectral information of virtual sound sources with a specified accuracy in binaural systems.

2 Binaural synthesis model

For a single frequency, the microphone array recordings are organized in the vector

$$\mathbf{p} = \begin{bmatrix} p_1 & p_2 & \cdots & p_Q \end{bmatrix}^{\mathsf{T}}. \tag{1}$$

The symbol $^{\mathsf{T}}$ indicates transpose. Each entry p_q of $\mathbf{p}$, where $q = 1, 2, ..., Q$, represents a sample in the frequency domain of a sound pressure signal recorded at an arbitrary position $\vec{r}_q^{\,\mathrm{m}}$, where the superscript 'm' refers to the microphones.

The dataset of HRTFs is organized in the matrix

$$\mathbf{h} = \begin{bmatrix} h_1^{\mathrm{left}} & h_2^{\mathrm{left}} & \cdots & h_L^{\mathrm{left}} \\ h_1^{\mathrm{right}} & h_2^{\mathrm{right}} & \cdots & h_L^{\mathrm{right}} \end{bmatrix}^{\mathsf{T}}. \tag{2}$$

Each entry h_ℓ^{left} or h_ℓ^{right} of $\mathbf{h}$, where $\ell = 1, 2, ..., L$, represents a sample in frequency of a free-field HRTF for the left or right ear, respectively. Each entry is characterized for an arbitrary sound source position $\vec{r}_\ell^{\,\mathrm{v}}$, where the superscript 'v' refers to the sound sources, which are referred to as *virtual loudspeakers* throughout this paper.

The synthesized binaural signals for the left and right ears are organized in the pair

$$\hat{\mathbf{b}} = \begin{bmatrix} \hat{b}^{\mathrm{left}} & \hat{b}^{\mathrm{right}} \end{bmatrix}^{\mathsf{T}}. \tag{3}$$

---

*マイクロホンアレイを用いたバイノーラル合成システムの雑音耐性評価, サルバドル・セザル, 坂本修一, トレビーニョ・ホルヘ, 鈴木陽一 (東北大・通研/大学院情報科学)

Binaural synthesis can be summarized as the following linear combination of $\mathbf{p}$ and $\mathbf{h}$:

$$\hat{\mathbf{b}} = \mathbf{h}^{\dagger}\mathbf{A}\mathbf{p}, \tag{4}$$

where $\mathbf{A} = [a_{\ell q}]$ is a combination matrix of size $L \times Q$ and † indicates conjugate transpose.

Characterizing $\mathbf{A}$ requires to take into consideration the geometry and physical topology of the microphone array, as well as the distribution of virtual loudspeakers used to obtain the HRTF dataset. Most of the existing methods for obtaining $\mathbf{A}$ can be gathered into two dual approaches depending on weather the products $\mathbf{h}^{\dagger}\mathbf{A}$ or $\mathbf{A}\mathbf{p}$ are optimized.

In this study, the first approach is followed because it is closely related with the theory of filter-and-sum beamforming, which provide a convenient framework for investigating the effects of noise. This leads to the diagram for binaural beamforming shown in Fig. 1. In this context, the binaural synthesis equation in (4) is written as follows:

$$\hat{\mathbf{b}} = \mathbf{w}^{\dagger}\mathbf{p}, \tag{5}$$

where the beamformer matrix

$$\mathbf{w} = \begin{bmatrix} \mathbf{w}^{\text{left}} \\ \mathbf{w}^{\text{right}} \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} w_1^{\text{left}} & w_2^{\text{left}} & \cdots & w_Q^{\text{left}} \\ w_1^{\text{right}} & w_2^{\text{right}} & \cdots & w_Q^{\text{right}} \end{bmatrix}^{\mathsf{T}} \tag{6}$$

contains the weighting coefficients that are applied to $\mathbf{p}$ so as to synthesize the spatial patterns defined by $\mathbf{h}$. The beamformer matrix is defined from (4) and (5) as follows:

$$\mathbf{w} = \mathbf{A}^{\dagger}\mathbf{h}. \tag{7}$$

3 White noise gain

The white noise gain of a beamformer is defined as the output power due to spatially uncorrelated, unit variance white noise of at the sensors. The norm squared of the weight vectors represents the white noise gain [7]:

$$\text{WNG} = \|\mathbf{w}\|^2. \tag{8}$$

For simplicity, $\mathbf{w}$ will represent either $\mathbf{w}^{\text{left}}$ or $\mathbf{w}^{\text{right}}$ in all of what follows. The norm is defined by

$$\|\mathbf{u}\| = \left(\mathbf{u}^{\dagger}\mathbf{u}\right)^{\frac{1}{2}} = \left(\sum_{q=1}^{Q} |u_q|^2\right)^{\frac{1}{2}} < \infty. \tag{9}$$

where $\mathbf{u}$ is an arbitrary vector of size $Q \times 1$. This norm is used throughout this paper.

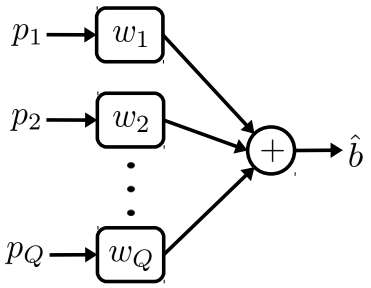

Fig. 1 Binaural beamforming.

If the white noise gain is large, it is expected a poor signal-to-noise ratio at the output of the beamformer due to white noise contributions. The inverse of the white noise gain, WNG^{-1}, should therefore be as high as possible. This variable is used as a general measure for robustness. However, to predict a more precise improvement in the signal-to-noise ratio, an additional analysis is presented below.

4 Gain in signal-to-noise ratio

The propagation of noise through the synthesis model in (4) can be calculated based on the gain in signal-to-noise ratio from the input to the output:

$$G^{\text{SNR}} = \frac{\text{SNR}^{\text{output}}}{\text{SNR}^{\text{input}}}. \tag{10}$$

This quantity is expected to be greater than unit and as high as possible. To calculate G^{SNR}, a signal model with noise is considered below.

The vector of microphone array recordings affected by additive, spatially uncorrelated white noise is modeled as follows:

$$\mathbf{p} = \mathbf{s} + \boldsymbol{\nu}, \tag{11}$$

where the vector $\mathbf{s} = \begin{bmatrix} s_1 & s_2 & \cdots & s_Q \end{bmatrix}^{\mathsf{T}}$ contains the microphone signals without noise, and the vector $\boldsymbol{\nu} = \begin{bmatrix} \nu_1 & \nu_2 & \cdots & \nu_Q \end{bmatrix}^{\mathsf{T}}$ contains the noise only. For this model, binaural synthesis results in

$$\begin{aligned} \hat{\mathbf{b}} &= \mathbf{w}^{\dagger}\left[\mathbf{s} + \boldsymbol{\nu}\right], \\ &= \underbrace{\mathbf{w}^{\dagger}\mathbf{s}}_{\hat{\mathbf{b}}^{(s)}} + \underbrace{\mathbf{w}^{\dagger}\boldsymbol{\nu}}_{\hat{\mathbf{b}}^{(\nu)}}, \end{aligned} \tag{12}$$

where $\hat{\mathbf{b}}^{(s)}$ denote the binaural signals from recordings without noise, and $\hat{\mathbf{b}}^{(\nu)}$ represents the binaural signals due to noise only.

The multichannel signal-to-noise ratio at the input can be defined by

$$\text{SNR}^{\text{input}} := \frac{\|\mathbf{s}\|^2}{\|\boldsymbol{\nu}\|^2} \tag{13}$$

with the norm in (9). The signal-to-noise ratio at the output, for the left or right ear, can also be defined in consistency with (9) but for the case of a single-channel signal:

$$\mathrm{SNR}^{\mathrm{output}} := \frac{|\hat{b}^{(s)}|^2}{|\hat{b}^{(\nu)}|^2} = \frac{|\mathbf{w}^\dagger \mathbf{s}|^2}{|\mathbf{w}^\dagger \boldsymbol{\nu}|^2}. \tag{14}$$

When $\boldsymbol{\nu}$ has a uniform distribution of energy,

$$\boldsymbol{\nu} = \nu_0 \cdot \boldsymbol{\varphi}, \tag{15}$$

where $\boldsymbol{\varphi} = \begin{bmatrix} e^{j\varphi_1} & e^{j\varphi_2} & \cdots & e^{j\varphi_Q} \end{bmatrix}^\mathsf{T}$ is a vector with random phases φ_q, (10) results in

$$G^{\mathrm{SNR}} = \frac{\frac{|\mathbf{w}^\dagger \mathbf{s}|^2}{|\nu_0|^2 |\mathbf{w}^\dagger \boldsymbol{\varphi}|^2}}{\frac{\|\mathbf{s}\|^2}{|\nu_0|^2 \|\boldsymbol{\varphi}\|^2}} = \frac{|\mathbf{w}^\dagger \mathbf{s}|^2}{|\mathbf{w}^\dagger \boldsymbol{\varphi}|^2 \frac{\|\mathbf{s}\|^2}{\|\boldsymbol{\varphi}\|^2}}. \tag{16}$$

By virtue of the Cauchy-Schwarz inequality for the norm in (9), the left factor in the denominator of (16) has the following upper bound:

$$|\mathbf{w}^\dagger \boldsymbol{\varphi}|^2 \leq \|\mathbf{w}\|^2 \|\boldsymbol{\varphi}\|^2. \tag{17}$$

Equating (17) and (16), it is shown that

$$\begin{aligned} G^{\mathrm{SNR}} &\geq \frac{|\mathbf{w}^\dagger \mathbf{s}|^2}{\|\mathbf{w}\|^2 \|\mathbf{s}\|^2}, \\ &\geq \frac{|\sum_{\ell,q} h_\ell^* a_{\ell q} s_q|^2}{\left(\sum_q |\sum_\ell h_\ell^* a_{\ell q}|^2\right)\left(\sum_q |s_q|^2\right)}, \end{aligned} \tag{18}$$

where * denotes complex conjugate. When the recording signals are affected by additive noise that is spatially uncorrelated and has a uniform distribution of energy, the right side of (18) provides a theoretical lower bound for the propagation of this kind of noise through the synthesis model in (4).

To analyze the performance of (8) and (18) on predicting improvement in signal-to-noise ratio, a physical model for $\mathbf{A}$ would be useful. We describe a model for binaural synthesis in spherical geometries.

5 Model for spherical geometries

In the spherical coordinate system shown in Fig. 2, a point in space $\vec{r} = (r, \theta, \phi)$ is specified by its radial distance r, azimuth angle $\theta \in [-\pi, \pi]$ and elevation angle $\phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. Angles can be merged into a variable $\Omega = (\theta, \phi)$ in such a way that a point in space is also represented by $\vec{r} = (r, \Omega)$.

When a rigid spherical microphone array of radius r_m is used for recording, each entry p_q of $\mathbf{p}$ in (1) correspond to a microphone signal recorded at $\vec{r}_q^{\,\mathrm{m}} = (r_\mathrm{m}, \Omega_q^\mathrm{m})$. Similarly, when a spherical virtual loudspeaker array of radius r_v is used to obtain the HRTF dataset $\mathbf{h}$ in (2), each entry h_ℓ^{left} or h_ℓ^{right} corresponds to a virtual loudspeaker position $\vec{r}_\ell^{\,\mathrm{v}} = (r_\mathrm{v}, \Omega_\ell^\mathrm{v})$.

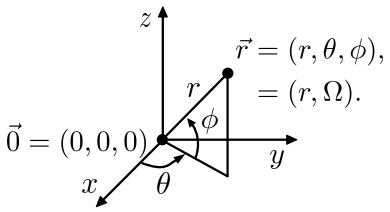

Fig. 2 Spherical coordinates. The origin $\vec{0}$ coincides with the centers of the array and head.

In such spherical geometries, the entries $a_{\ell q}$ of $\mathbf{A}$ in (4) can be modeled by [4,9]:

$$\begin{aligned} a_{\ell q} = \frac{1}{4\pi(N+1)^2} \cdot \frac{\exp(jkr_\mathrm{v})}{r_\mathrm{v}} \cdot \\ \sum_{n=0}^{(Q+1)^2} (2n+1) R_n^{\mathrm{reg}}(r_\mathrm{m}, r_\mathrm{v}, k) P_n(\cos\Theta_{\ell q}). \end{aligned} \tag{19}$$

The angular part of the sum in (19) is defined by the Legendre polynomial P_n of order n evaluated at the cosine of the angle $\Theta_{\ell q}$ between $\vec{r}_\ell^{\,\mathrm{v}}$ and $\vec{r}_q^{\,\mathrm{m}}$. The radial part of the sum in (19) is defined by the regularized radial filter

$$R_n^{\mathrm{reg}} = \frac{R_n}{1 + \lambda^2 |R_n|^2}, \quad R_n = -\frac{kr_\mathrm{m}^2 h_n'(kr_\mathrm{m})}{h_n(kr_\mathrm{v})}. \tag{20}$$

Here, λ is the regularization parameter, h_n represents the spherical Hankel function of second kind and order n, and $'$ indicates derivative with respect to the argument.

6 Results of the experiments

Figure 3 shows the predictions of robustness obtained with WNG^{-1} from (8), and with the proposed lower bound for G^{SNR} in (18). The binaural beamformer was modeled by using (19) with $\lambda = 1 \times 10^{-3}$. The minimum of G^{SNR} was obtained from the simulation of 5000 sources randomly distributed around the array at 1.5 m distance. The microphones and virtual loudspeakers were distributed using spherical grids based on the geometry of an icosahedron.

The top panel shows the results for WNG^{-1}, where it can be observed that increasing the number of microphones improved the expected robustness at

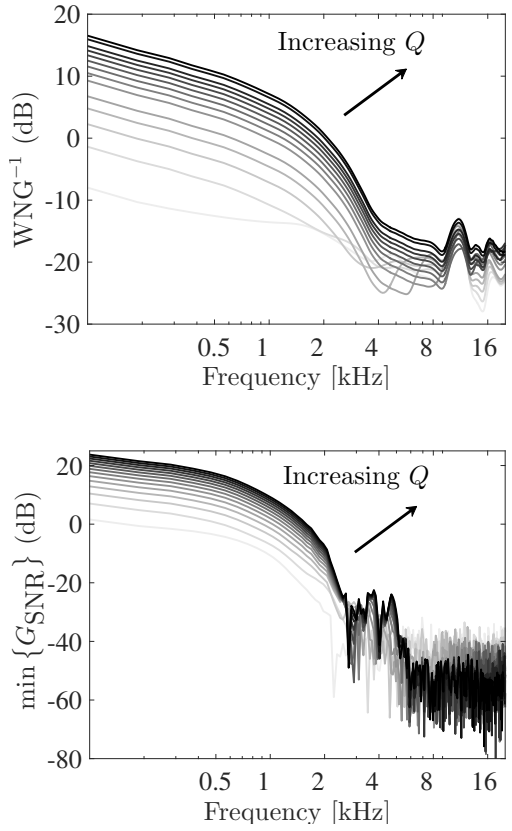

Fig. 3 Evaluation of inverse of white noise gain (WNG^{-1}) and gain in signal-to-noise ratio (G^{SNR}) for a binaural beamformer in spherical geometries. Results where obtained for $r_{\mathrm{m}} = 8.5$ cm, $r_{\mathrm{v}} = 1.5$ m, $L = 1962$ virtual loudspeakers, and different numbers of microphones (Q =12, 42, 92, 162, 252, 362, 492, 642, 812, 1002, 1212, 1442, 1692, 1962.).

lower and middle frequencies. However, when the number of microphones increased, its contribution to robustness only increased slightly. This was specially true at higher frequencies.

The bottom panel shows the results for G^{SNR}. At lower frequencies up to around 2 kHz, the results confirmed the tendency predicted by WNG^{-1}. Nevertheless, the results showed that adding more microphones to the system does not necessarily imply an improvement in robustness at higher frequencies.

7 Conclusion

A linear model for binaural systems in arbitrary geometries was formulated. The model takes into consideration the contributions of additive white noise, which is assumed spatially uncorrelated and with a uniform distribution of energy around the array. The propagation of such kind of noise through the model was investigated based on two predictors of robustness: 1) the inverse of white noise gain used in beamforming, and 2) a proposed lower bound for the gain in signal-to-noise ratio.

Numerical experiments considering a binaural system in spherical geometries showed that similar predictions at lower frequencies can be obtained with the white noise gain and the gain in signal-to-noise ratio. However, results at higher frequencies showed that the white noise gain might not be sufficient to predict robustness in this region. In this regard, the estimates obtained with the lower bound to the gain in signal to noise ratio predicted much lower levels of robustness at higher frequencies.

Additional experiments based on a more precise model of noise, accompanied with objective validations in real world conditions, could give more insight into the findings reported in this work.

Acknowledgements A part of this study was supported by a JSPS Grant-in-Aid for Scientific Research (no. 24240016 and 16H01736) and the A3 Foresight Program for "Ultra-realistic acoustic interactive communication on next-generation Internet". The authors thank M. Otani for admitting us to use his BEM solver [10] to calculate HRTFs.

References

[1] M. Morimoto *et al., Proc. Congress Acoust. Soc. Jpn.*, 1975.
[2] V. Algazi *et al., J. Audio Eng. Soc.*, Vol. 23, No. 11, pp. 1142–1156, 2004.
[3] R. Duraiswami *et al., Proc. Audio Eng. Soc. Conv.*, 2005.
[4] C. Salvador *et al., Proc. Mtgs. Acoust.*, Vol. 19, No. 1, 2013.
[5] E. Rasumow *et al., IEEE/ACM Trans. Audio, Speech, Language Process.*, Vol. 24, No. 2, pp. 215-225, 2016.
[6] S. Sakamoto *et al., Acoust. Sci. Technol.*, Vol. 36, No. 6, pp. 516-526, 2015.
[7] B. Van Veen and K. Buckley, *IEEE ASSP*, Vol. 5, No. 2, pp. 4-24, 1988.
[8] B. Rafaely *et al., IEEE Trans. Speech and Audio Process.*, Vol. 13, No. 1, pp. 135-143, 2005.
[9] E.G. Williams, *Fourier Acoustics*, 1999.
[10] M. Otani, S. Ise, J. Acoust. Soc. Am., Vol. 119, pp. 2589–2598, May 2006.