Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vol. 20, No 3, 1998

NEURAL NETWORK MODEL OF BINAURAL HEARING BASED ON SPATIAL FEATURE EXTRACTION OF THE HEAD RELATED TRANSFER FUNCTION

Zhenyang Wu, Tao Weng, and Weibin Wang
Department of Radio Engineering
Southeast University, Si Pai Lou #2, Nanjing 210018, P. R. C.
Email: zhenyang@seu.edu.cn

T.F. Lo, Francis H. Y. Chan, and F. K. Lam
Department of Electrical and Electronic Engineering,
University of Hong Kong, Pokfulam Road, Hong Kong
Email: fhychan@hkueee.hku.hk

Abstract - **In spatial hearing, complex valued head-related transfer function (HRTF) can be represented as a real valued head-related impulse response (HRIR). Using Karhunen-Loeve expansion, the spatial features of the normalized HRIRs on measurement space can be extracted as spatial character functions. A neural network model based on Von-Mises function is used to approximate the discrete spatial character function of HRIR. As a result, a time-domain binaural model is established and it fits the measured HRIRs well.**

Keywords: Spatial Hearing, Feature Extraction, Neural Networks

1. Introduction

Extensive physical and behavioral studies have revealed that the external ear plays an important role in spatial hearing. The external ear provides directional amplification of the incident sound pressure level and also modifies the spectrum of the incoming sound according to the incidence angle of that sound. Direction-dependent transformation of the external ear is referred as head-related transfer function (HRTF) or head-related impulse response (HRIR) to acknowledge its primary acoustical importance.

In virtual auditory space (VAS) applications and physiological study, HRTFs in continuous spatial locations are often desired. Kistler and Wightman [1] established a model based on principal component analysis and minimum-phase reconstruction. Principal component analysis (PCA) is applied to the logarithms of the HRTF magnitudes after the removal of direction-independent and subject-dependent spectral features. Chen [2] further proposed a spatial feature extraction and regularization (SFER) model for the HRTFs. In this model the HRTFs are expressed as weighted combinations of a set of complex valued eigentransfer functions. Both the PCA model and the SFER model have focused on the frequency components and involve complex-valued or logarithmic computation, therefore their applicability in real time case is limited.

In a previous work, Wu and Chan et al [3] presented binaural model based on the spatial features extracted from the measured HRIRs of a cat. In this model HRIRs are approximated as weighted combinations of a set of real valued basis functions. A simple linear interpolation algorithm is employed to obtain the modeled binaural HRIRs. The real valued operations and linear interpolation are very effective for speeding up the model computation in real time implementation.

In this paper, we extend this model to the measured HRIRs of a KEMAR and develop a neural network model to establish the continuous human hearing space.

2. Spatial feature extraction of the HRIR[3]

The measured HRIRs come from the Media Laboratory of MIT[4]. The measurement was conducted on a mannequin KEMAR in an anechoic chamber. The spherical space around the KEMAR was sampled at elevations from –40° (40° below the horizontal plane) to 90° (directly overhead) in 10° increments. The azimuth increment sizes were chosen to maintain approximately 5° great-circle increments. In total, 710 locations were sampled. The impulse responses were obtained using a maximum length(ML) sequence measurement technique with a sampling rate of 44.1kHz.

Because these data were contaminated by various types of disturbances such as random noise and acoustic reflections, they must be preprocessed before computation. A denoising algorithm based on singularity detection with wavelet is adopted. The speaker response is also equalized by using the inverse filter of the speaker response. After that, the HRIRs are shortened to 128-point long and shifted along the time axis to remove the ITDs of different locations.

Let $\boldsymbol{h}_j$ denote the HRIR of location (θ_j,φ_j), j=1,2,…P, P=710. The HRIRs are the function of both space and time and through spatial feature extraction

we can separate the spatial features from the temporal features. Using Karhunen-Loeve expansion we derive orthonormal basis functions of $\boldsymbol{h}_j$, that is

$$\begin{aligned}\boldsymbol{h}_j &= \boldsymbol{Q}\boldsymbol{w}_j + \boldsymbol{h}_{av} \\ &= \sum_{i=1}^{M} w_i(\theta_j,\varphi_j)\boldsymbol{q}_j + \boldsymbol{h}_{av} + \varepsilon_j \end{aligned} \quad (1)$$

where the $\boldsymbol{h}_{av}$ is the space sample average and $\mathbf{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots \boldsymbol{q}_M]$ is an orthonormal transformation matrix whose columns are chosen as the eigenvectors of the time auto-covariance matrix $\mathrm{R_h}$.

We call q_j (j=1,2,...M) the eigentranfer functions (EF). EFs only contain the temporal information of the HRIRs while the weight vector $\mathbf{w}_j = [w_1(\theta_j,\varphi_j), w_2(\theta_j,\varphi_j), \cdots, w_M(\theta_j,\varphi_j)]$ only contain the spatial information of the HRIRs. We call $\mathbf{w}_j$ the spatial character functions (SCF) and $\mathbf{w}_j$ is calculated by

$$\mathbf{w}_j = \mathbf{Q}^T(\mathrm{h}_j - \mathrm{h}_{av}) \quad (2)$$

In our model, the first 10 components can represent more than 95% of the variation in the normalized HRIRs.

3. The neural network model of HRIR in continuous hearing space

Because the SCF we obtained from the spatial feature extraction exists only in several discrete space locations, we hope to find the SCF in an arbitrary space location.

In practice, we can use several basis functions to approach the SCF in the desired position, that is

$$\begin{aligned}\hat{w}(\theta,\varphi) &= c_1 f_1(\theta,\varphi) + c_2 f_2(\theta,\varphi) \\ &+ \cdots + c_n f_n(\theta,\varphi) \end{aligned} \quad (3)$$

In this case, a two-layer BP (Back-Propagation) network is developed to approach the SCF. The inputs to the network are the location in the space. The hidden layer unit gives the corresponding output according to its input θ and φ. Here the Von Mises Function is used as the basis function[5]:

$$\begin{aligned}VM(\theta,\varphi) &= \exp\{k[\sin\varphi \sin\beta \cos(\theta-\alpha) \\ &+ \cos\varphi \cos\beta - 1]\} \end{aligned} \quad (4)$$

where k,α,β are the parameters and $\alpha,\theta \in [0,2\pi]$, $\beta,\varphi \in [0,\pi]$. The output of the network is SCF. The connectivity of the network is indicated schematically in figure 1.

The learning algorithm used here is as follows:

$$\begin{aligned}\Omega(n+1) &= \Omega(n) + \eta(n)\Delta\Omega \\ &+ \mu[\Omega(n) - \Omega(n-1)] \end{aligned} \quad (5)$$

where $\Omega = [\mathrm{wt}^T, \alpha^T, \beta^T, \mathrm{k}^T]$,

$\Delta\Omega = [\Delta\mathrm{wt}^T, \Delta\alpha^T, \Delta\beta^T, \Delta\mathrm{k}^T]$.

$$\Delta wt_{i,j} = [t_i - y_i(\theta,\varphi)]VM(\theta,\varphi,\alpha_j,\beta_j,k_j) \quad (6)$$

$$\begin{aligned}\Delta\alpha_j &= k_j[\sin\varphi \sin\beta_j \sin(\theta-\alpha_j)] \\ &\times \sum_i^N \{[t_i - y_i(\theta,\varphi)]wt_{i,j}\} \\ &\times VM(\theta,\varphi,\alpha_j,\beta_j,k_j) \end{aligned} \quad (7)$$

$$\begin{aligned}\Delta\beta_j &= k_j[\sin\varphi \cos\beta_j \cos(\theta-\alpha_j) \\ &- \cos\varphi \sin\beta_j] \times \sum_i^N \{[t_i - y_i(\theta,\varphi)] \\ &\times wt_{i,j}\} \times VM(\theta,\varphi,\alpha_j,\beta_j,k_j) \end{aligned} \quad (8)$$

$$\begin{aligned}\Delta k_j &= k_j[\sin\varphi \cos\beta_j \cos(\theta-\alpha_j) - \cos\varphi \cos\beta_j \\ &- 1] \times \sum_i^N \{[t_i - y_i(\theta,\varphi)]wt_{i,j}\} \\ &\times VM(\theta,\varphi,\alpha_j,\beta_j,k_j) \end{aligned} \quad (9)$$

where t_j is teacher and y_j is the output. Among the total 710 (θ_j,φ_j) data sets, 600 are chosen for training the net and the rest are used to test the generalization performance of the model.

4. Results

Figure 2 gives the 3D plot of the first component of SCF . Above is the SCF in discrete space locations and below is the network output of the SCF in different space locations.

Table 1 gives the final error of the SCF and the reconstruction error according to these SCFs.

Acknowlegement

The present work was supported in part by University of Hong Kong Research Grants and National Natural Science Foundation of China. The authors would like to thank Mr. Bill Gardener and Kieth Martin of MIT lab for providing the KEMAR data.

REFERENCES

[1] D. J. Kistler, and F. L. Wightman "A Model of Head-related Transfer Functions Based on Principal Components Analysis and Minimum-phase Reconstruction," J. Acoust. Soc. Am. 91, 1637-1647.

[2] J. Chen, B. D.Van Veen, and K. E. Hecox "A Spatial Feature Extraction and Regularization Model for the Head-related Transfer Function," J. Acoust. Soc. Am. 97, 439-452.

[3] Z. Wu, F. H. Y. Chan, F. K. Lam and J. Chan "A time domain binaural model based on spatial feature extraction for the head-related transfer function" J. Acoust. Soc. Am. 102, 2211-2218.

[4] W. G. Gardner and K. D. Martin " HRTF measurements of a KEMAR," J. Acoust. Soc. Am. 97, 3907-3908.

[5] R. L. Jenison, et al. "A spherical basis function neural network for modeling auditory space," Neural Computation, 8, 115-128.

Table 1

Components of SCF	Rms error of the SCF		Rms error of the HRIR using the output SCF		Rms error of the HRIR using original SCF
	Training set	Test set	Training set	Test set	Total space positions
1	0.0521	0.0519	0.3001	0.3041	0.2978
2	0.0829	0.0856	0.2483	0.2484	0.2418
3	0.1148	0.1064	0.1792	0.1870	0.1640
4	0.1024	0.1269	0.1352	0.1379	0.1127
5	0.1118	0.1269	0.1352	0.1379	0.1127
6	0.0822	0.0821	0.0991	0.1009	0.0728
7	0.0661	0.0722	0.0816	0.0848	0.0533
8	0.0402	0.0521	0.0758	0.0789	0.0464
9	0.0393	0.0431	0.0673	0.0675	0.0368
10	0.0337	0.0408	0.0503	0.0539	0.0292

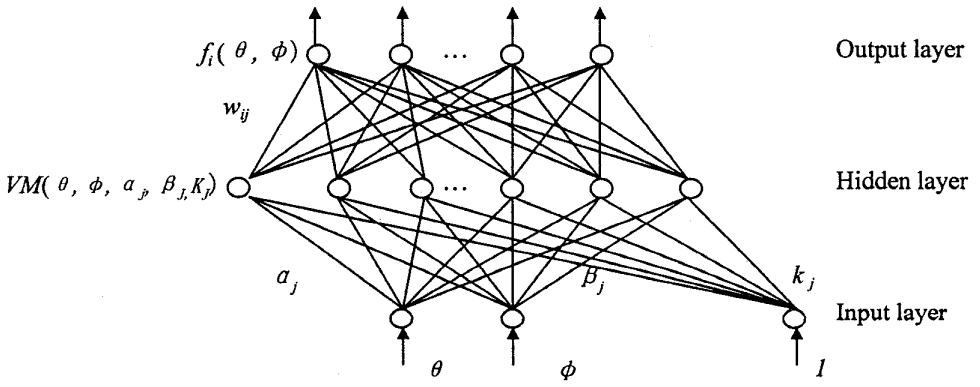

Figure 1 The neural network model

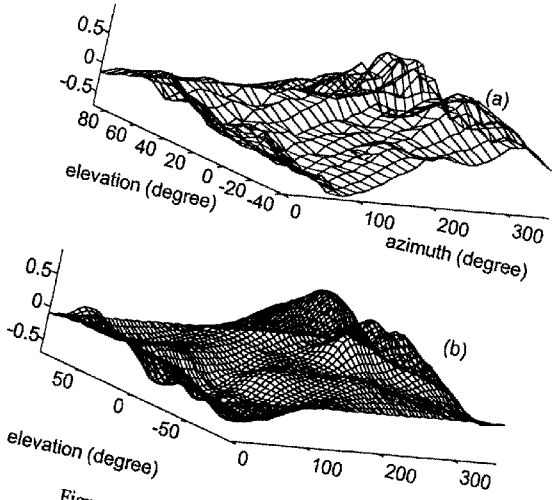

Figure 2 Mesh plot of first component of SCF