Stacked Autoencoder Based HRTF Synthesis from Sparse Data

Sunil Bharitkar* and Timothy Mauer† Teresa Wells† David Berfanger†
* HP Labs, HP Inc., Palo Alto (CA), USA
E-mail: sunil.bharitkar@hp.com
† HP Inc., Vancouver (WA), USA

***Abstract*—Ipsilateral and contralateral head-related transfer functions (HRTF), $H_{\text{ipsi}}(\omega, r, \phi, \theta)$ and $H_{\text{contra}}(\omega, r, \phi, \theta)$, are used for creating the perception of a virtual sound source at an arbitrary distance r and azimuth-elevation tuple $\underline{\psi} = [\theta, \phi]^T$ relative to the median plane for a given frequency ω. Publicly available databases use a subset of a full-grid of angular directions due to time and complexity to acquire and deconvolve responses. In this paper, we present a subspace-based technique for reconstructing HRTFs at arbitrary directions for the IRCAM-Listen HRTF database, which comprises a sparse set of HRTFs sampled every 15° along the azimuth/elevation direction. The presented technique includes first augmenting the sparse IRCAM dataset using auditory localization blur, then deriving a set of lower-dimensional compressed representation (using an autoencoder) from the augmented HRTFs. The lower dimensional representations are then trained using a fully-connected neural network (FCNN) for the corresponding directions. The reconstruction of HRTF corresponding to an arbitrary direction $\underline{\psi}_p$ is achieved by applying the compressed output from the FCNN, for an arbitrary direction, to a reconstruction system (viz., a decoder of an autoencoder). The results demonstrate the autoencoder approach provides good quality objective and subjective results.**

I. Introduction

Measured Head-related Transfer Functions (HRTF) include binaural cues for localizing a sound-source in a 3D-space. These cues are introduced by the acoustic path, and the reflection and diffraction effects from a listeners anatomical structure. Binaural cues include (a) interaural time difference (ITD) between the two-ears and is a function of the radii of the human head a as well as the direction of arrival $\underline{\psi}$, and (b) interaural spectral differences arising due to the frequency dependent shadowing effect of the head as well as the frequency dependent reflection filtering due to the pinna, shoulders and torso. Typically, these HRTFs are spatially sub-sampled to a point where only a finite set are collected around a mannekin or a human. It becomes critical to synthesize the missing HRTFs for critical interactive applications such as in Virtual Reality (VR)-where asymetrical rendering of high-quality video and spatially inaccurate spatial audio would nonetheless degrade the "presence." There are multiple approaches for synthesizing spatially sub-sampled HRTFs, including (a) parametric models, based on anthropometric features, [5],[2], [6], [7], [11], [10], [8], [9], and (b) using measured HRTFs. In the case of measured HRTFs, among various techniques to synthesize via interpolation, (i) the approach of [12] incorporates a tetrahedral interpolation technique with barycentric weights to synthesize HRTFs, with the technique being extensible for sparse HRTF sets, (ii) functional representation used for deriving HRTFs are obtained by fitting spatial characteristic functions with a thin-plate spline with regularization [13], (iii) state-space approach [14] for HRTF interpolation using the MIT KEMAR data, (iv) wavelet transform [15] based interpolation on the MIT Media Lab KEMAR dataset[1], and (v) manifold learning using ISOMAP [16] on the dense CIPIC HRTF dataset [2]. Additionally, synthesis of distance-based HRTF cues (which is not considered in this paper) can be found, as an example, in [3].

TABLE I
Localization blur for horizontal displacement in the horizontal plane re: median plane ($\Delta(\phi = 0)_{\text{MIN}}$)

Ref.	Stimuli	$\Delta(\theta = 0)_{min}$
Klemm[18]	Impulses (click)	0.75-2°
King et al. [19]	Impulse train	1.6°
Stevens et al.[20]	Sinusoids	4.4°
Schmidt et al.[21]	Sinusoids	>1°
Sandel et al.[22]	Sinusoids	1.1-4°
Mills [23]	Sinusoids	1.1-3.1°
Stiller et al.[24]	Tone bursts	1.4-2.8°
Boerger[25]	Gaussian tone bursts	0.8-3.3°
Gardner[26]	Speech	0.9°
Perrott[27]	Tone bursts	1.8-11.8°
Blauert[28]	Speech	1.5°
Haustein et al.[29]	Broadband noise	3.2°

In the approach presented in this paper, we use the IRCAM (Institute for Research and Coordination in Acoustics and Music) Listen HRTF dataset [4]. The IRCAM dataset comprises of HRTFs obtained at an angular spacing of 15°. Given that the human auditory resolution is tuned for discriminating sources with a *localization blur*[1] that is lower bounded on critical test stimuli at 1° intervals in the frontal direction [17], the Listen dataset constitutes a sparse set. Examples of localization blur relative to the median plane are shown in Table 1 [17] with corresponding mean and variance represented in Figure 1. Furthermore, from the compilation of the results in [17], a directional perspective to localization blur is shown in Fig. 1, wherein the auditory system is able to discriminate sources

[1]Localization blur is the amount of displacement of the position of the sound source that is recognized by 50% of experimental subjects as a change in the position of the auditory event

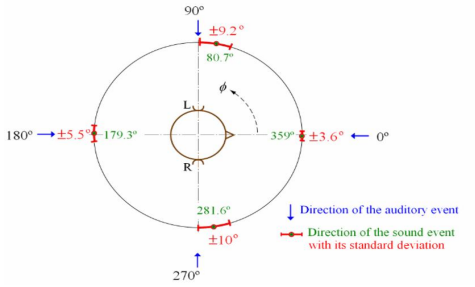

Fig. 1. Localization blur and localization in the azimuth relative to median plane with white-noise pulses [17].

within $\approx \pm 3°$ in the front, while the sensitivity decreases by $\approx \pm 6°$ to the side and it decreases by $\approx \pm 3°$ to the rear. Clearly, the Listen HRTF would benefit from an interpolation scheme that is derived from perceptual cues based on the spatial sensitivity of human hearing (i.e., localization blur).

In the next section, we present an approach to synthesize the missing HRTFs from this sparse set. In this approach, we determine a nonlinear compressed representation, using deep learning architecture, involving a stacked autoencoder. The compressed representation is then used to train an FCNN. The reconstruction is performed by taking the lower-dimensional output, for a given input direction, and applying it to the autoencoder decoder to synthesize the HRTF. In section 3 we present objective and subjective results, while conclusions and future directions are presented in section 4.

II. Localization Blur and Neural Models for HRTF Synthesis

The approach using an autoencoder is shown in Fig. 2. In the first step the original sparse HRTFs are augmented using the percept of localization blur.

A. HRTF Augmentation with Localization Blur

In this step a difference is determined between consecutive HRTF magnitude responses whose envelope is then approximated by a second order discrete time-domain infinite impulse response (IIR) filter, expressed as

$$\begin{aligned} H_{\text{blur}}(z) &= 10^{G/20}\frac{\sum_{k=0}^{2} b_k z^{-k}}{\sum_{k=0}^{2} a_k z^{-k}} \\ b_i &= \gamma_1(f_c, f_s, G); \quad a_0 = 1, a_i = \gamma_2(f_c, f_s) \end{aligned} \quad (1)$$

where f_c is the -3 dB frequency, G controls the gain in dB, and f_s is the sampling frequency, and γ_1 and γ_2 are nonlinear functions. Alternative models for such filters, also referred to as shelf filters, can be found in [30]. An example of this envelope-approximating shelf filter is shown in Fig. 3 for $f_c = 2$ kHz and $G = 3$ dB.

The envelope, between two consecutive HRTF sets, spaced at 15° is interval-stepped in a non-uniform manner predicated on the non-linear spatial auditory resolution (as interpreted by the mean localization blur values, shown in Fig. 1) which is finer in the frontal and rear direction and less-refined towards the sides. The HRTFs from the sparse set are merged with the augmented set to create a system of HRTFs for use in the subsequent ML model. While the finer details (viz., spectral notches width, frequencies, and amplitudes) are not used for the augmented set, in contrast to the envelope, it is expected that an ML model would synthesize these finer representations which are also important for localization.

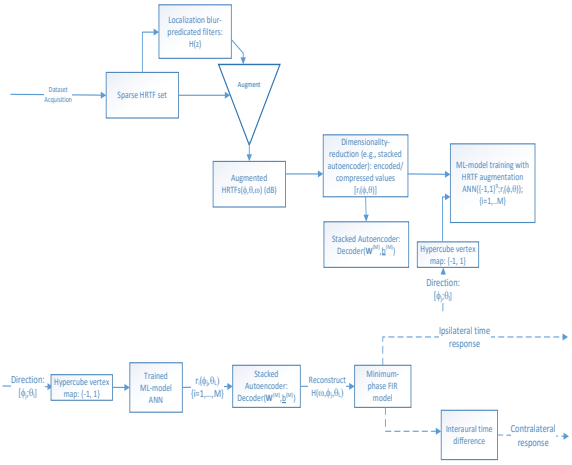

Fig. 2. Stacked autoencoder based approach for HRTF synthesis using blur-based augmentation

B. Stacked Sparse Autoencoders for Signal Compression

In this step the augmented HRTF set is first reduced in dimensionality using stacked sparse autoencoders [31], [32] which are pretrained using a linear weighted combination of (i) a mean-square error term between the input and the estimated input (at the output of the decoder), (b) Kullback-Liebler divergence measure between the activation functions of the hidden layers and a sparsity parameter (ρ) to keep some of the hidden neurons inactive some or most of the time), and (c) with an L_2 regularization on the weights of the autoencoder to keep them constrained in norm. The structure of the stacked autoencoder is shown in Fig. 4. The cost function for optimization of the weights, W, of the sparse and regularized autoencoder is,

$$E = \frac{1}{N}\left(\sum_{k=1}^{N} \|\underline{X}_k - \underline{\hat{X}}_k\|_2| + \alpha\Omega_{\text{KL}}(\rho\|\hat{\rho}_{\text{hidden}})\right) + \beta\|\mathbf{W}\| \quad (2)$$

Adding a term to the cost function that constrains the values of $\hat{\rho}_{\text{hidden}}$ to be low encourages the autoencoder to learn a representation, where each neuron in the hidden layer fires to a small number of training examples. We will explore other autoencoder optimization function involving, for example, restricted Boltzmann machines (RBM) [33].The compressed values, at the output of the deepest encoder layer, are used

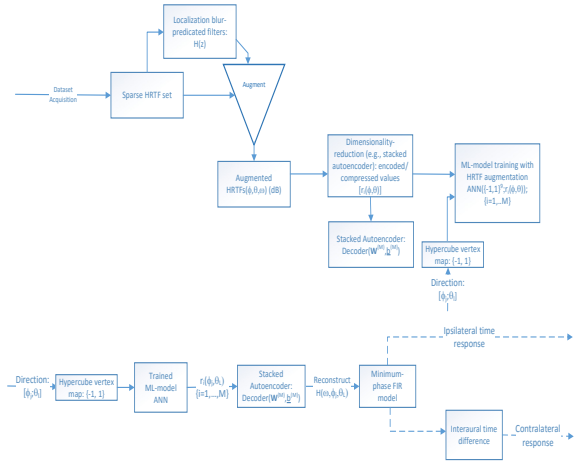

Fig. 3. Stacked autoencoder based approach for HRTF synthesis using blur-based augmentation

for reconstructing the HRTFs at arbitrary directions in the following step. Fig. 2 shows the architecture for synthesizing HRTFs using the sparse autoencoders.

C. Function Approximation Model

While several models are available that perform function approximation, we selected a multilayer fully-connected neural network (FCNN) for developing the subspace synthesis model due to its universal approximation properties (e.g., single hidden-layer [34], multi-hidden layer [35]). The input to the neural network is the direction of the HRTF $\underline{\psi}_p$ and the output vector corresponds to the M-dimension compressed representation. The direction input transformed initially to binary form with the actual values mapped to the vertices of a q-dimensional hypercube (viz., $V = \{(v_1, \ldots, v_q) | v_i \in \{-1, 1\} (i = 1, \ldots, q)\}$) in order to normalize the input to the first hidden layer of the ANN. For example, with $\phi_i, \theta_i \in [0, 360]$ taking extremal values, the input to the hidden layer with sigmoidal activation functions would result in saturation and degrade convergence behavior during training[2]. Accordingly, the input space is transformed to a binary representation having 9-element input layer for the horizontal and elevation directions[3]. Among the various training approaches, we found the gradient descent with momentum term and adaptive learning rate providing an acceptable balance in terms of convergence time and approximation error on the training data.

[2]The importance of normalization was studied, for example, in [36]

[3]The range 0-360° can be mapped to ±180° with positive angles representing the hemisphere to the right of the median plane and negative ψ angles the left hemisphere, and for which the binary representation is expressed using 9 bits including an MSB as a sign bit corresponding to the left or right hemispheres relative to the median plane

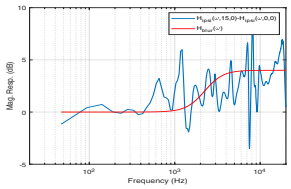

Fig. 4. Example of adjacent ipsilateral HRTF difference and blur-based envelope augmentation model

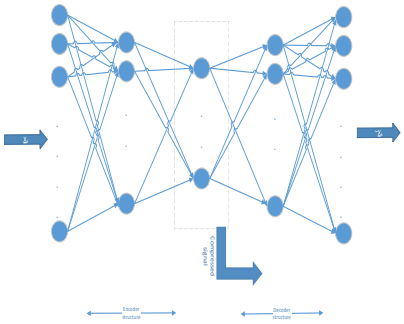

Fig. 5. Stacked Autoencoder

III. Results

Due to approximate symmetry between the left and right half of the median plane, with respect to spatial hearing, the synthesis for ipsilateral and contralateral HRTFs are demonstrated for only the left half of the median plane and on the *horizontal plane*. In the first step the envelope filter is estimated between two consecutive Listen HRTFs. An example of the difference in the ipsilateral magnitude responses between 0° and 15° is shown in Fig. 3, for subject 1008, along with the plot of the envelope estimating filter. The augmented HRTFs, as an example in this case, are derived at $\{3, 6, 9, 12\}°$ with the angular spacing being $\approx 3°$ by first interpolating $H_{\text{blur}(\omega)}$ linearly along the G (gain)-axis and applying these filters to the $H_{\text{ipsi}}(\omega, 0, 0)$ HRTF. The angular spacing in the presented approach successively becomes wider towards the side (e.g., the augmented HRTFs based on blur are derived at $\{80, 85\}°$ towards the side in between the $H_{\text{ipsi}}(\omega, 75, 0)$ and $H_{\text{ipsi}}(\omega, 90, 0)$ yielding $[\mathbf{R}_{\text{ipsi}}(\omega, \phi, \theta), \mathbf{R}_{\text{contra}}(\omega, \phi, \theta)] \in \Re^{1024 \times 44}$.

A. Sparse Stacked Autoencoders

The number of stacked autoencoders used was set to two for first achieving a compression from 1024 fft bins to 64 values and then from 64 dimension-representation down to 6-dimensional representation in the encoder part (this allows

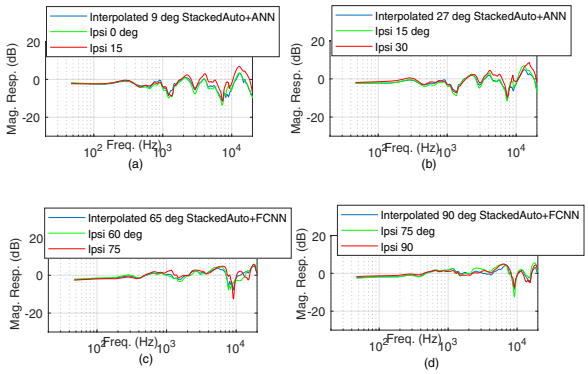

Fig. 6. Sample results: (a)-(d) Autoencoder+FCNN synthesis for ipsilateral responses

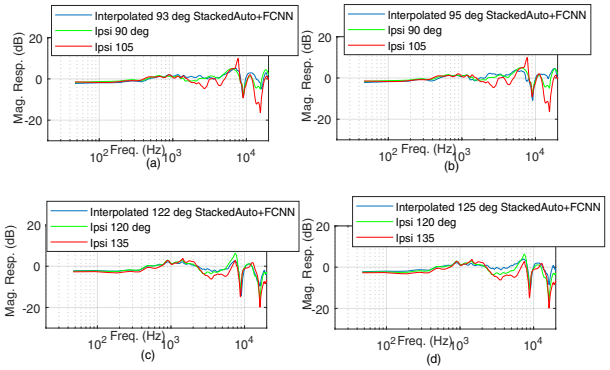

Fig. 7. Sample results: (a)-(d) Autoencoder+FCNN synthesis for ipsilateral responses

comparison against the PCA-based approach described earlier which used $M = 6$ principal components) with the sparsity proportion set to 0.8 for the first encoder and 0.7 for the second encoder. The multilayer neural network used in this paper two hidden layers involving 29 and 15 neurons in the first and second hidden layer, respectively, to perform function approximation over the training set comprising the input direction $\in \Re^{9\times 44}$ (with 9 input neurons for the "8-bit+MSB sign bit" binary directional representation and 44 horizontal directions) and output comprising the $M = 6$. Each of the hidden and output neurons use the $\tanh.()$ function since the maximum of each of the PC over all directions is $\in [0, 1]$ and minimum is $\in [-1, 0]$.

The autoencoder+FCNN results are shown, as examples, in Figs. 5-8 for directions not in the training set. The blue curve shows the synthesized HRTF (ipsilateral or contralateral), whereas the red and green curve correspond to the nearest HRTFs from the Listen dataset. The results show a good objective performance where the synthesized HRTFs fall within the limits established on the quantized (viz., original) Listen/IRCAM HRTFs.

B. Time-domain Synthesis using ITD

The estimated magnitude response (for PCA and stacked autoencoder approach) is then converted to a 512-point finite-impulse-response (FIR) using the frequency resampling approach [37], and the appropriate interaural time-delay (ITD) is inserted into the contralateral response based on the direction ψ_p. The delay is modeled using the Woodworth-Schlossberg formula [17] $(a/c)(\phi_p + \sin(\phi_p))$ in seconds where a is the average head-radii of 0.0875 m. In this paper we estimated the response by cross-correlating the ipsilateral and contralateral

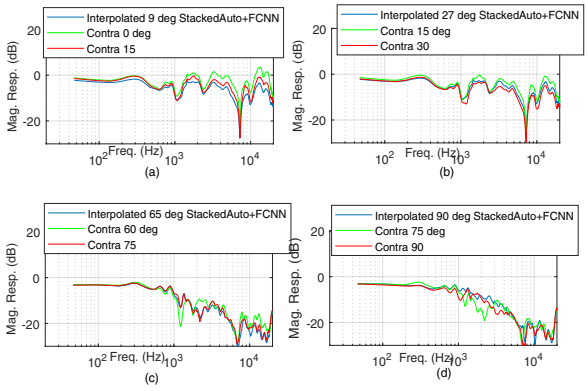

Fig. 8. Sample results: (a)-(d) Autoencoder+FCNN synthesis for contralateral responses

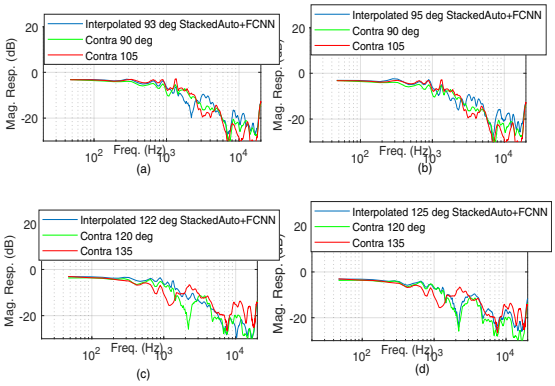

Fig. 9. Sample results: (a)-(d) Autoencoder+FCNN synthesis for contralateral responses

responses over the 15° separation Listen HRTF data. The ITDs for arbitrary angles can be determined using polynomial interpolation technique.

C. Subjective Testing

The testing was done using an interface shown in Figs. 9 (a)-(b), with 18 subjects, where each subject annotated the angle they perceived the stimuli (results in Fig. 10) at 30 degrees and 135 degrees (where 0 degrees is in the front and 90 degrees is to the left). The testing was done with stimuli (pink noise, speech, bandlimited noise, tone-burst) ordered randomly and using Etymotic Research ER4SR reference in-ear headphones with foam inserts, whereas the stimuli selected were pink noise (spectrally 3 dB/octave roll-off but perceptually neutral), speech (female voice), bandlimited pink noise centered at 2 kHz (corresponding to a high-sensitivity region as predicated on the equal-loudness contours) with 1/3-octave wide bandwidth, and tone-burst centered at 2 kHz with 1/3-octave wide bandwidth. The results for the quality of the synthesized stimuli was also tested and the results are shown in Fig. 11. Each of the stimuli was $\approx$ 2 seconds long and was normalized to a loudness of -26 LKFS using ITU-R BS. 1116-4 [38]. As is evident for the frontal perception, the stacked autoencoder mean localized angle is closer to the desired 30 degrees for speech and tone bursts compared to the PCA approach. The reference HRTF dataset, corresponding to subject 1008, was perceived for each stimuli, having a mean localization angle different than 30 degrees. In addition, there wasn't a significant difference between the mean localized values for the raw IRCAM HRTF set, over the test stimuli. Fig. 11 shows the

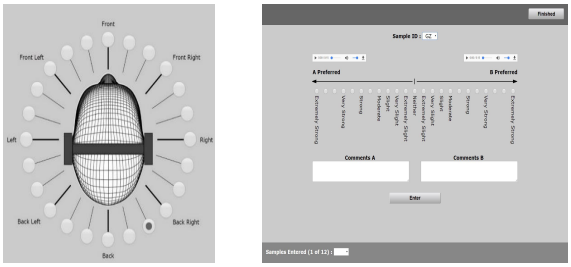

Fig. 10. Subjective testing: (a) Directional interface, (c) Qualitative interface.

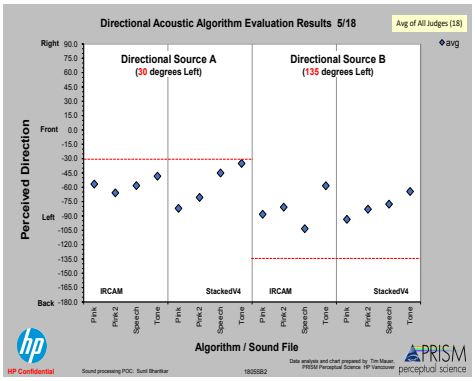

Fig. 11. Comparison of localization performance between Original HRTFs (IRCAM) and Autoencoder approach for 30 and 135 degrees.

qualitative performance (as judged by naturalness, audibility of artifacts such as excess energy in certain frequency region, temporal artifacts, and spatial imaging artifacts such as split image) between the raw HRTFs and the synthesized HRTFs. For the 30 degrees case, the stacked autoencoder approach was judged on an average better than the raw IRCAM HRTFs for broadband pink noise. For the side-rear localization perception, the mean results for the autoencoder show a close match between the 135 degrees IRCAM and 135 degrees autoencoder in the mean values for all stimuli (except speech).

IV. Conclusions & Future Directions

The modeling and synthesis of HRTFs from a sparse dataset is an important problem in order to recreate the perception of sound sources at arbitrary positions in 3D-space for interactive applications such as VR. Developing a low-complexity approach for synthesis, additionally, is important for such interactive applications. In this paper we presented an

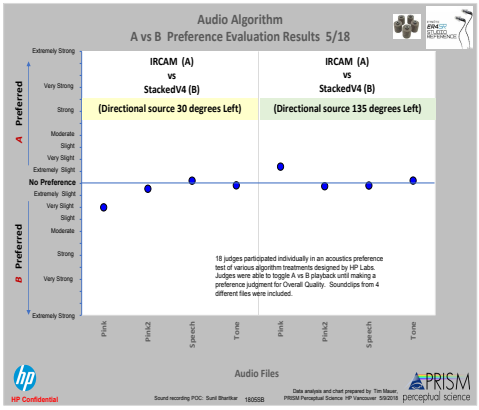

Fig. 12. Comparison of qualitative performance between Original HRTFs (IRCAM) and Autoencoder approach for 30 and 135 degrees.

approach involving auditory perception (viz., localization blur) to create an augmented HRTF dataset, followed by a subspace decomposition involving a stacked autoencoder to train a low-complexity ML-model. The objective and subjective results on an average that the stacked autoencoder approach performs well in synthesizing HRTFs that were not in the original HRTF sparse dataset. Future directions include exploring improvements by hyper-parameter tuning of the ML-model, exploring new architectures from deep learning as well as alternative training models, synthesizing vertical responses, generalizing over more test subject HRTFs from Listen listeners, and finally generalizing to diverse datasets (including MIT, ITA, CIPIC, etc.). Test stimuli as .wav files, and models (in .mat Matlab format), Listen HRTFs in .mat, and a readme are located at: https://github.com/bharitka/APSIPA2018.

REFERENCES

[1] W. Gardner,and K. Martin, "HRTF measurements of a KEMAR," *J. Acoust. Soc. Am.*, vol. 97(6), pp.3907-3908, 1995.

[2] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The CIPIC hrtf database," *2001 IEEE ASSP Wkshp. Appl. Sig. Proc. to Audio and Acoust. (IEEE-WASPAA)*, Oct. 2001.

[3] A. Kan, C. Jin, and A. Schaik, "A psychophysical evaluation of near-field head-related transfer functions synthesized using a distance variation function," *J. Acoust. Soc. Amer.*, vol. 125, pp. 2233, April 2009.

[4] O. Warusfel, "http://recherche.ircam.fr/equipes/salles/listen/," *IRCAM Listen HRTF Database*, 2002.

[5] C. Brown, and R. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. on Speech and Audio Proc.*, vol. 6(5), pp. 476488, 1998.

[6] G. Ramos, and M. Cobos, "Parametric head-related transfer function modeling and interpolation for cost-efficient binaural sound applications," *J. Acoust. Soc. Am.*, vol. 134, pp. 17351738, 2013.

[7] G. Ramos, M. Cobos, B. Bank, ad J. Belloch, "A Parallel Approach to HRTF Approximation and Interpolation Based on a Parametric Filter Model," *IEEE Sig. Proc. Lett.*, 24(10), pp. 15071511, 2017.

[8] V. R. Algazi, R. Duda, R. O. Duraiswami, N. A. Gumerov, and Z. Tang, "Approximating the head-related transfer function using simple geometric models of the head and torso," *J. Acoust. Soc. Amer.*, 112, pp. 2053-2064, 2002.

[9] J. Mackenzie, J. Huopaniemi, V. Valimaki, and I. Kale, "Low-order modeling of head-related transfer functions using balanced model truncation," *IEEE Sig. Proc. Lett.*, 4(2), pp. 39-41, Feb. 1997.

[10] A. Kulkarni, and H. S. Colburn, "Infinite-impulse-response models of the head-related transfer function," *J. Acoust. Soc. Amer.*, 115(4), pp. 17141728, 2004.

[11] R. Duraiswami, D. N. Zotkin, and N. A. Gumerov, "Interpolation and range extrapolation of HRTFs," *Proc. IEEE Int. Conf. Acoust. Speech & Sig. Proc.*, Canada 2004.

[12] H. Gamper, "Head-related transfer function interpolation in azimuth, elevation, and distance," *J. Acoust. Soc. Amer.*, 134(6), Dec. 2013.

[13] J. Chen, B. Van Veen, and K. Hecox, "A spatial feature extraction and regularization model for the head-related transfer function," *J. Acoust. Soc. Amer.*, 97(1), pp. 439-452, Jan. 1995.

[14] F. Keyrouz, and K. Diepold, "A New HRTF Interpolation Approach for Fast Synthesis of Dynamic Environmental Interaction," *J. Audio Eng. Soc.*, 56(1/2), pp. 28-35,Jan./Feb. 2008 2004.

[15] R. Duraiswami, D. N. Zotkin, and N. A. Gumerov, "HRTF Interpolation in the wavelet transform domain," *Proc. IEEE Wkshp. Appl. Sig. Proc. Audio & Acoust.*, Oct. 2009.

[16] F. Grijalva, L. Martini, D. Florencio, and S. Goldenstein, "Interpolation of Head-Related Transfer Functions Using Manifold Learning," *IEEE Sig. Proc. Lett.*, 24(2), pp. 221-225, Feb. 2017.

[17] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT Press (Cambridge), 1999.

[18] O. Klemm, "Investigations of the localization of sound stimuli IV: On the influence of binaural time difference on localization," *Arch. Ges. Psychol.*, 40, pp. 117-145, 1920.

[19] W. King, and D. Laird, "The effect of noise intensity and pattern on locating sounds," *J. Acoust. Soc. Am.*, 2, pp. 99-102, 1930.

[20] S. Stevens, and E. Newman, "The localization of actual sources of sound," *Amer. J. Psychol.*, 48, pp. 297-306, 1936.

[21] P. Schmidt, A. Gemmert, R. Fries, J. Duyff, "Binaural threshold for azimuth difference," *Acta Physio. Pharmacol. Nederl.*, 3, pp. 2-18, 1953.

[22] T. Sandel, D. Teas, W. Feddersen, and L. Jeffress, "Localization of sound from single and paired sources," *J. Acoust. Soc. Amer.*, 27, pp. 842-852, 1955.

[23] A. Mills, "On the minimum audible angle," *J. Acoust. Soc. Amer.*, 30, pp. 237-246, 1958.

[24] D. Stiller, "The faculty of localization," *Elektroakustik*, 71, pp. 76, 1960.

[25] G. Boerger, "The localization of Gaussian tones," *Dissertation*, Technische Universitat, Berlin, 1965.

[26] M. Gardner, "Lateral localization of 0° or near-0°-oriented speech signals in anechoic space," *J. Acoust. Soc. Amer.*, 44, pp. 797-803, 1968.

[27] D. Perrott, "Role of signal onset in sound localization," *J. Acoust. Soc. Amer.*, 45, pp. 436-445, 1969.

[28] J. Blauert, "An experiment in directional hearing with simultaneous optical stimulation," *Acustica*, 23, pp. 118-119, 1970.

[29] B. Haustein, and W. Schirmer, "A measuring apparatus for the investigation of the faculty of directional localization," *Elektroakustik*, 79, pp. 96-101, 1970.

[30] S. Bharitkar and C. Kyriakakis, *Immersive Audio Signal Processing*, Springer-Verlag (NY), 2006.

[31] M. F. Moller, "A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning," *Neural Networks*, Vol. 6, 1993, pp. 525533.

[32] B. A. Olshausen and D. J. Field. "Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1," *Vision Research*, Vol.37, 1997, pp. 33113325.

[33] G. E. Hinton, and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, 313, pp. 504-507, July 2006.

[34] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Infor. Theory*, 39(3), pp. 930-945, 1993.

[35] H. Lee, R. Ge, T. Ma, A. Risteki, S. Arora, "On the ability of neural nets to express distributions," *Proc. of Mach. Learn. Res.*, 65, pp. 1-26, 2017.

[36] J. Sola, and J. Sevilla, "Importance of input data normalization for the application of neural networks to complex industrial problems," *IEEE Trans. Nucl. Sci.*, 44(3), pp. 1464-1468, 1997.

[37] S.Mitra, *Digital Signal Processing: A Computer Based Approach*, McGraw-Hill (NY), 1998.

[38] ITU-R BS.1770 International Telecommunication Union-Radiocommunication, *Algorithms to measure audio programme loudness and true-peak audio level*, Geneva, 2015