SPATIAL AUDIO FEATURE DISCOVERY WITH CONVOLUTIONAL NEURAL NETWORKS

*Etienne Thuillier**

Dept. of Signal Processing and Acoustics
Aalto University
Espoo, Finland

Hannes Gamper, Ivan J. Tashev

Audio and Acoustics Research Group
Microsoft Research
Redmond, WA, USA

ABSTRACT

The advent of mixed reality consumer products brings about a pressing need to develop and improve spatial sound rendering techniques for a broad user base. Despite a large body of prior work, the precise nature and importance of various sound localization cues and how they should be personalized for an individual user to improve localization performance is still an open research problem. Here we propose training a convolutional neural network (CNN) to classify the elevation angle of spatially rendered sounds and employing Layer-wise Relevance Propagation (LRP) on the trained CNN model. LRP provides saliency maps that can be used to identify spectral features used by the network for classification. These maps, in addition to the convolution filters learned by the CNN, are discussed in the context of listening tests reported in the literature. The proposed approach could potentially provide an avenue for future studies on modeling and personalization of head-related transfer functions (HRTFs).

***Index Terms*—** Spatial sound, virtual reality, HRTF personalization, Deep Taylor Decomposition, acoustic feature discovery

1. INTRODUCTION

With mixed reality entering the mass consumer market, accurate rendering of spatial sound for a large user base is an important problem. Spatial sound rendering engines typically rely on acoustic models encoding the filtering behavior of the human head, torso, and pinnae into sound signals to create the impression of a sound source emanating from a certain location. These acoustic models are referred to as head-related impulse responses (HRIRs) in the time domain or head-related transfer functions (HRTFs) in the frequency domain. When measuring the HRIRs of a human subject, the captured acoustic cues are a direct result of the subject's anthropometric features, and hence highly individual. As the auditory system relies on these cues for localization, deviations of a modelled or generic HRIR set used in the rendering engine from the user's own HRIRs can result in a degraded listening experience. Therefore, identifying the audio cues that should be preserved or modelled for accurate localization is of continued research interest.

A large body of prior work exists on various aspects of spatial audio perception. In this work, we focus on acoustic cues affecting the perception of source elevation. Gardner identifies torso effects below 3.5 kHz and pinna effects above 4 kHz as salient cues for localization on the median plane [1]. Other studies show that the presence of a peak or notch at specific frequencies can be associated with the perceived elevation of a source [2, 3, 4]. Searle et al. derive a mathematical localization model from the results of 40 localization studies that combines the contributions of four types of auditory cues: interaural time delay and head shadow effects, interaural pinna cues, monaural pinna cues, and shoulder reflections [5]. Kulkarni and Colburn find that extreme smoothing of the HRTF can lead to the perception of an elevated sound source [6]. Jin et al. report that both interaural spectral differences and monaural spectral cues are useful for disambiguating source positions within a cone of confusion [7]. Common to these studies is that they rely on listening experiments, which can be time consuming and limited in scope.

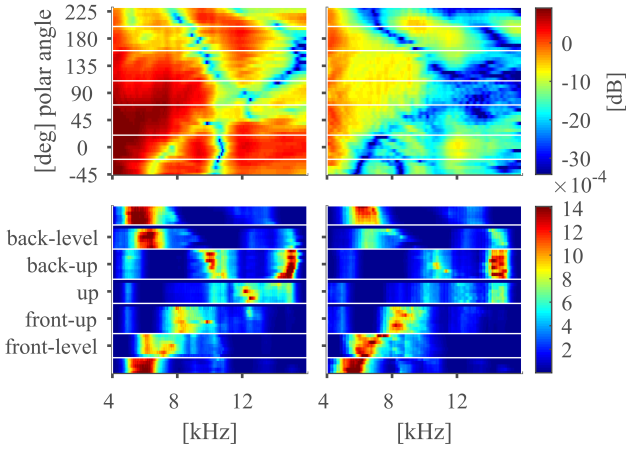

Fig. 1. Ipsilateral and contralateral HRTF magnitude responses (top) at 30 degrees lateral angle for subject 154 from the CIPIC dataset, and the corresponding saliency maps (bottom) for CNN model **HP** produced via Layer-wise Relevance Propagation (LRP).

Jin et al. propose a physiologically inspired localization model consisting of a cochlea model front-end coupled to a time-delay neural network [8]. They show that in a localization test, the model demonstrated qualitatively similar performance to a human subject. A related area of active research aims to personalize generic HRTFs given a user's anthropometric features [9, 10, 11].

Here we propose a machine learning approach to identify salient elevation cues encoded in the HRTFs and shared across a population of subjects. Recently, convolutional neural networks (CNNs) have proven successful for classic speech and audio problems [12, 13], without the need to apply feature extraction to the raw input data [13]. Our approach is based on training a CNN to determine the elevation angle of a virtual sound source and using layer-wise relevance propagation (LRP) to detect the audio features learned by the CNN. An example is shown in Figure 1. The training is per-

*The work was done while Etienne Thuillier was an intern at Microsoft Research Labs in Redmond, WA, USA.

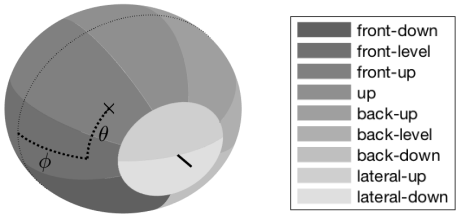

Fig. 2. Elevation classes in horizontal-polar coordinate system [14].

formed on multiple HRTF datasets to account for variability between the measured subjects as well as different measurement setups, thus forcing the CNN to learn common audio features. Experimental results indicate that the proposed network can determine the elevation of a virtual sound source from simulated ear input signals and that the features discovered using LRP seem to be in line with results from the psychoacoustic literature. This indicates that the proposed framework may be a useful tool complementary to listening experiments for studying spatial audio features, with potential applications for HRTF personalization.

2. PROPOSED APPROACH

2.1. Sound source elevation localization using a CNN

The goal of the proposed work is to discover the audio features used by a convolutional neural network (CNN) trained to perform sound source localization. The localization task is posed as a simple classification problem, whereby the CNN determines which elevation class a sound sample belongs to, as illustrated in Figure 2. The hypothesis of our proposed approach is that to perform the classification, the CNN would have to learn audio features specific to each class.

As input data, the CNN is fed with the ear input signals a listener would perceive given a point-source in the far field in anechoic conditions. For the described scenario, the log-magnitude spectrum of the ipsilateral ear input signal $\boldsymbol{E}_{\mathrm{dB,ipsi}}$ is given as

$$\boldsymbol{E}_{\mathrm{dB,ipsi}} = 20\log_{10}|\mathcal{F}(g\boldsymbol{n} \ast \boldsymbol{h}_{\mathrm{ipsi}})| \tag{1}$$

where $\boldsymbol{n}$ is the time-domain source signal, g is a gain factor, $\boldsymbol{h}_{\mathrm{ipsi}}$ is the ipsilateral HRIR corresponding to the source position, $\mathcal{F}$ denotes the Fourier transform and $\ast$ denotes the convolution operator. The log-magnitude spectrum of the contralateral ear input signal $\boldsymbol{E}_{\mathrm{dB,contra}}$ is obtained analogously using the contralateral HRIR $\boldsymbol{h}_{\mathrm{contra}}$.

A CNN training sample $\boldsymbol{S}$ is given as a K $\times$ 2 matrix

$$\boldsymbol{S} = [\boldsymbol{E}_{\mathrm{dB,ipsi}}\ \boldsymbol{E}_{\mathrm{dB,contra}}], \tag{2}$$

where K is the number of frequency bins. Each training sample is obtained via (1) using a random 50 ms long white noise burst as the source signal $\boldsymbol{n}$ and a pair of HRIRs randomly drawn from one of the elevation classes shown in Figure 2. The probability of drawing a specific HRIR pair is determined such that it ensures balancing of all test subjects and classes as well as a uniform spatial representation on the sphere.

	conv. 1	conv. 2	conv. 3	conv. 4	# parameters
WB	25×2	11×1	11×1	10×1	2681
HP	25×1	11×1	11×1	10×2	2741

Table 1. Filter shapes ($\times$ 4 per layer) and number of trainable parameters for each CNN model.

2.2. CNN architectures

Two CNN architectures are considered in this work, one for wideband input features in the range 0.3–16 kHz (**WB**) and one for high-frequency input features in the range 4–16 kHz (**HP**). As illustrated in Figure 3a, for model **WB** the output features of the first convolution layer are generated by (two-dimensional) interaural filters, each combining the ipsilateral and contralateral components of the input features. The hypothesis underlying this choice is that interaural spectral differences contribute to the perception of elevation [7].

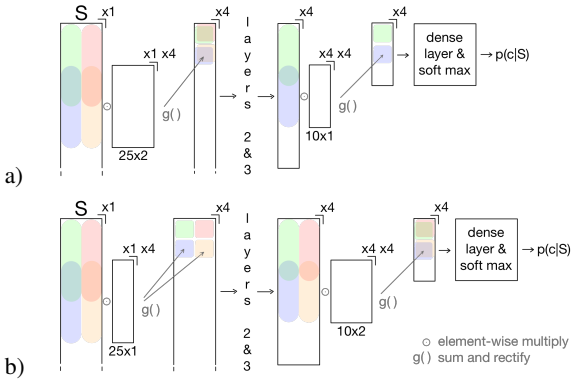

Fig. 3. Schematic diagram of (a) model **WB** and (b) model **HP**. For simplicity, only the first and fourth convolutional layers are shown.

Model **HP** was trained with input features truncated below 4 kHz to force the CNN to learn monaural elevation cues associated with the pinna [1, 15]. In contrast to model **WB**, each of the lower convolution layers extracts monaural features using (single-dimension) monaural filters that are applied to both the ipsilateral and contralateral sides. As shown in Figure 3b, the resulting high-level monaural features from both sides are combined at the top-most convolutional layer.

Models **WB** and **HP** both comprise four convolutional layers with rectified linear units (ReLUs). The filter lengths and strides along the frequency dimension are identical across these models. Specifically, a stride of two samples was used without pooling. Each model further comprises a fully-connected hidden layer and a softmax output layer. A summary of the model parameters is provided in Table 1.

2.3. Feature discovery using LRP

To explain the classification decisions of the CNN, Deep Taylor Decomposition (DTD) [16], a variant of Layer-wise Relevance Propagation (LRP) [17], is performed. DTD performs a weighted redistribution of the network's output activation for the elected class, i.e., its *relevance* R, from network output and backwards, layer-by-layer, to network input. This procedure generates a saliency map that identifies the regions of the input space used by the model to arrive at the

classification decision. Here, these regions are formulated in terms of frequency range and binaural channel. A publicly-available implementation of DTD was used [19].

The relevance R_i of the ith neuron in a lower layer is given as

$$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j, \tag{3}$$

where j and R_j denote the index and relevance of a higher-layer neuron, w_{ij} are the connection weights between the neurons, $^+$ denotes half-wave rectification, and a is the (forward-pass) activation.

Given that the input features are real-valued, the following propagation rule is applied at the model's input layer [18]:

$$R_i = \sum_j \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j. \tag{4}$$

The above expression can be decomposed in terms of the contributions specific to each filter of the model's first layer, allowing to study their respective saliency maps [18].

3. EXPERIMENTAL EVALUATION

3.1. CNN model training

Experiments were carried out using a pool of five HRTF databases, listed in Table 2. All databases except that of Microsoft are publicly available [20] in the Spatially Oriented Format for Acoustics (SOFA) [21]. The resulting pool contained approximately 0.5 million HRIR pairs from 583 subjects and was divided into 80% training data and 20% test data. Training was conducted using the cross-entropy loss function and early stopping regularization under a ten-fold cross-validation scheme. The HRIR pairs from each database were distributed approximately uniformly across the validation folds and test set to ensure robustness against possible database-specific artifacts. Each measured subject was siloed into a single validation fold or the test set to allow performance evaluation on unseen subjects.

The input samples were generated via (2) using randomly generated 50 ms long white noise bursts and raw HRIR data [20] resampled to 32 kHz. The noise gain g in (1) was randomly varied between 0 and -60 dB to provide model robustness to level variations.

For the **WB** model, frequency bins below 300 Hz were discarded. The resulting spectra were weighted with the inverse of the equal-loudness-level contour at 60 phon [22], to approximate human hearing sensitivity. For the **HP** model, frequency bins below 4 kHz were discarded, forcing the CNN to learn high-frequency cues. Both models used 1005 frequency bins per ear up to the Nyquist limit.

Figure 2 shows the boundaries of the nine elevation classes, given in horizontal-polar coordinates [14] as $\pm$ 60 degrees lateral angle ϕ and polar angles θ ranging from -90 to 270 degrees.

3.2. Classification performance

Optimal classification performance was not pursued in this work [8]. Rather, compact models for HRTF-based source localisation which generalise across human subjects were developed. It is worth mentioning that the performance of the trained models are comparable to that of humans, even if achieved on a data representation that is not physiologically accurate, e.g., in terms of the spectral resolution.

In particular, the classification error (CE) rates of the **WB** and **HP** models on unseen test data are 27.19% and 45.05% respectively.

	year	# subjects	# meas.	# pairs
ARI* [23]	2010	135	1150	138000
CIPIC [24]	2001	45	1250	56250
ITA** [25]	2016	46	2304	110592
Microsoft [9]	2015	252	400	100800
RIEC [26]	2014	105	865	90825

* Subjects 10 and 22 as well as all subjects not measured in-ear were removed.
** Subjects 02 and 14 were removed due to SOFA meta-data inconsistencies.

Table 2. Curated HRTF databases used for training.

	CE [%]	RMSE [deg]	MAE [deg]	r
random	91.3	74.5	59.5	0.65
[15]	-	25.2	-	0.85
[27]	-	-	22.3	0.82
[28]	-	-	≈25	-
WB	45.1	43.2	16.5	0.90

Table 3. Comparison of **WB** model to human localization performance.

To put the CE rates into context, performance metrics can be derived from the corresponding angular error rates after removing the lateral-up and lateral-down classes and accounting for front–back confusions [15]. Table 3 compares the root-mean-squared error (RMSE), mean absolute error (MAE), and correlation coefficient (r) to human localization performance reported in the literature. As can be seen, the **WB** model performs comparably to human subjects.

3.3. Subject-specific saliency map

To analyze the cues learned by the CNN the saliency map of a specific subject is computed. Figure 4 shows the confusion matrices for subject 154 of the CIPIC database. Given the high classification performance for this subject, the HRIRs are expected to present structures representative of the elevation cues learned by the models. Given input samples generated using randomly-drawn HRIR pairs via (2), 1-D saliency maps can be obtained using DTD. Averaging and stacking the 1-D maps of successful classifications according to their polar angle produces the 2-D saliency map shown in Figure 1 for model **HP**.

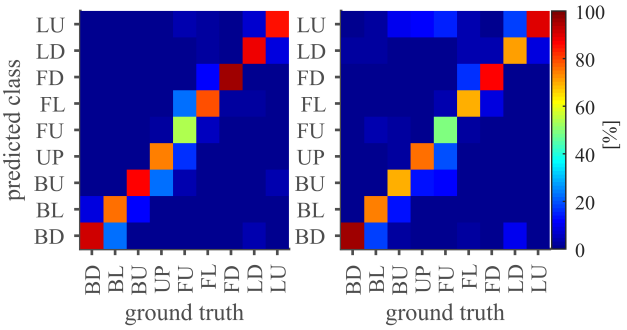

Fig. 4. Confusion matrix for subject 154 from the CIPIC dataset for model **WB** (left) and **HP** (right).

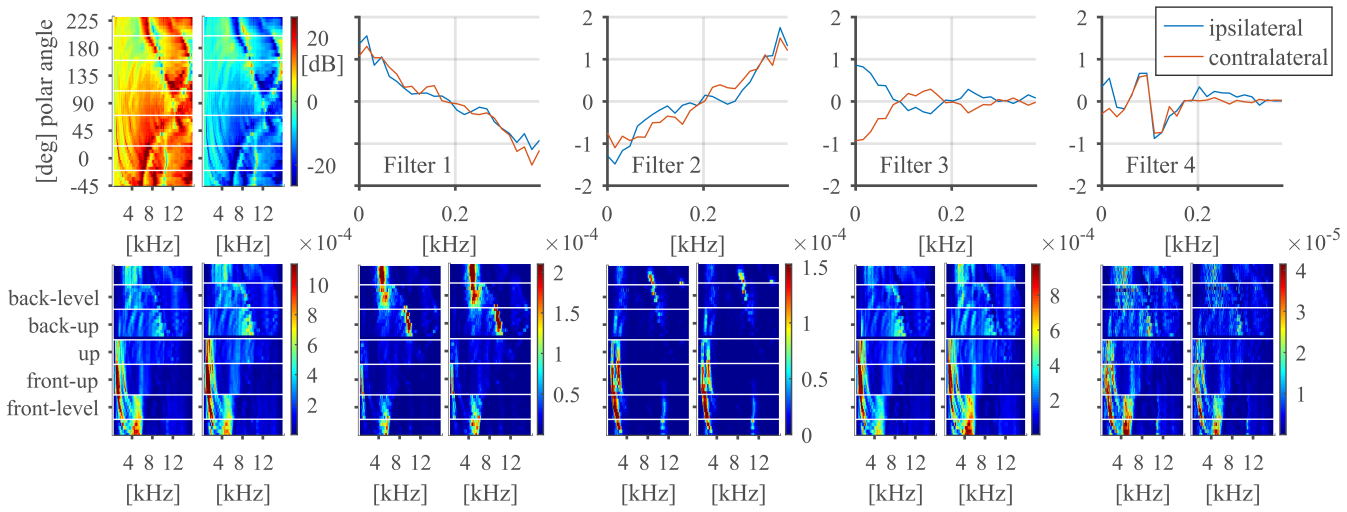

Fig. 5. Interaural transfer function [4] at 30 degrees lateral angle for subject 154 (top left); filters of the first convolution layer from model **WB** (top row) and their corresponding saliency contributions (bottom row); combined saliency map (bottom left).

3.3.1. Model **WB**

As shown in Figure 5, the filters of the first convolution layer in model **WB** are readily interpretable. Filters 1 and 2 form a complementary pair of falling and rising spectral edge detectors. Filter-specific saliency maps are shown in Figure 5. These maps indicate that model **WB** uses filter 2 to extract ripples in the range from 0.3 to 4 kHz caused by shoulder and torso reflections [15]. One limitation of the datasets used in this study is that the dependence of shoulder reflections on head orientation [29] is not accounted for. Training the model on variable shoulder reflections might potentially lower the contribution of these cues.

Filters 1 and 3 appear to contribute to the classification especially at low elevations. Filter 3 implements interaural differentiation and thus provides robust features to the upper layers of the network that are invariant to changes in the frequency composition of the sound source. Interaural cues are shown to enhance localization of sound sources in elevation [7]. At low elevations, these might be due to torso shadowing [30].

3.3.2. Model **HP**

Figure 1 illustrates that model **HP** relies on spectral notches as a primary cue for detecting sound sources located in frontal directions (i.e. 'front-down', 'front-level' and 'front-up'). Spectral notches varying as a function of elevation have been identified as important localization cues for humans [7]. As can be seen, the center frequency of the notch varies progressively from 6 kHz to 9 kHz as the polar angle increases, which is consistent with pinna models from the literature [31, 32]. Human pinnae typically produce several spectral notches resulting from reflections off various pinna features. In the example shown in Figure 1, the model seems to rely on the lowest-frequency notch, presumably stemming from the largest pinna feature, which might indicate that this feature is more consistent across the population than finer pinna details.

Other features visible in Figure 1 include:

- a relatively extended low-magnitude region above 10 KHz that seems to be indicative of class 'up';
- a sharp spectral notch in the the 15 kHz region that seems to be indicative of class 'back-up'; and
- a shadowing of the ipsilateral ear in the 4-7 kHz range that seems to be indicative of classes 'back-level' and 'back-down' [33].

Further work is required to determine exactly what type of feature was used by the model, and if these are relevant in a psycho-acoustic sense. In particular, it is doubtful that an adult subject would rely on features lying at the upper frequency limit of the hearing range, as in the example of the 15 kHz notch.

4. CONCLUSIONS

Experimental results indicate that a convolutional neural network (CNN) can be trained to achieve a classification performance comparable to that of humans in a simple sound localization task while being robust to inter-subject and measurement variability. The model seems to learn features from the input data, consisting of noise bursts convolved with measured head-related impulse responses (HRIRs), that are common to the tested population. Applying Deep Taylor Decomposition (DTD), a variant of Layer-wise Relevance Propagation (LRP), to the output of the trained model and stacking the resulting saliency maps as a function of polar angle provides an intuitive visualization of the features the CNN relies on for classification. The features illustrated by the saliency maps, as well as the convolution filters learned by the network, seem to be in line with results from the psychoacoustic literature. This indicates that the proposed approach may be useful for discovering or verifying spatial audio features shared across a population and possibly open avenues for better modeling and personalization of HRIRs. Future work includes training the network using non-white sound samples [8].

5. REFERENCES

[1] M. B. Gardner, "Some monaural and binaural facets of median plane localization," *J. Acoust. Soc. Am.*, vol. 54, no. 6, pp. 1489–1495, 1973.

[2] J. Hebrank and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," *J. Acoust. Soc. Am.*, vol. 56, no. 6, pp. 1829–1834, 1974.

[3] P. J. Bloom, "Creating source elevation illusions by spectral manipulation," *J. Audio Eng. Soc.*, vol. 25, no. 9, pp. 560–565, 1977.

[4] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*, MIT press, 1997.

[5] C. L. Searle, L. D. Braida, M. F. Davis, and H. S. Colburn, "Model for auditory localization," *J. Acoust. Soc. Am.*, vol. 60, no. 5, pp. 1164–1175, 1976.

[6] A. Kulkarni and H. S. Colburn, "Role of spectral detail in sound-source localization.," *Nature*, vol. 396, pp. 747–9, 1998.

[7] C. Jin, A. Corderoy, S. Carlile, and A. van Schaik, "Contrasting monaural and interaural spectral cues for human sound localization," *J. Acoust. Soc. Am.*, vol. 115, no. 6, pp. 3124–3141, 2004.

[8] C. Jin, M. Schenkel, and S. Carlile, "Neural system identification model of human sound localization," *J. Acoust. Soc. Am.*, vol. 108, no. 3, pp. 1215–1235, 2000.

[9] P. Bilinski, J. Ahrens, M. R. P. Thomas, I. J. Tashev, and J. C. Platt, "HRTF magnitude synthesis via sparse representation of anthropometric features," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4468–4472.

[10] A. Politis, M. R. P. Thomas, H. Gamper, and I. J. Tashev, "Applications of 3D spherical transforms to personalization of head-related transfer functions," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, Brisbane, Australia, Mar 2016, pp. 306–310.

[11] R. Sridhar and E. Choueiri, "A method for efficiently calculating head-related transfer functions directly from head scan point clouds," in *Audio Engineering Society Convention 143*, Oct 2017.

[12] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[13] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. IEEE Workshop Appl. Signal Process. to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct 2017.

[14] J. C. Middlebrooks, "Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1493–1510, 1999.

[15] V. R. Algazi, C. Avendano, and R. O. Duda, "Elevation localization and head-related transfer function analysis at low frequencies," *The Journal of the Acoustical Society of America*, vol. 109, no. 3, pp. 1110–1122, 2001.

[16] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

[17] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, pp. e0130140, 2015.

[18] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *arXiv preprint arXiv:1706.07979*, 2017.

[19] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, and W. Samek, "The LRP toolbox for artificial neural networks," *J. Machine Learning Research*, vol. 17, no. 1, pp. 3938–3942, 2016.

[20] "SOFA general purpose database," `https://www.sofaconventions.org/mediawiki/index.php/Files`, Online; accessed 25-Oct-2017.

[21] Inc. Audio Engineering Society, "AES69-2015 - AES standard for file exchange - spatial acoustic data file format," 2015.

[22] "ISO 226: 2003(e): Acoustics-normal equal-loudness-level contours," 2003.

[23] P. Majdak, M. J. Goupell, and B. Laback, "3-D localization of virtual sound sources: effects of visual environment, pointing method, and training," *Attention, perception, & psychophysics*, vol. 72, no. 2, pp. 454–469, 2010.

[24] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2001, pp. 99–102.

[25] R. Bomhardt, M. de la Fuente Klein, and J. Fels, "A high-resolution head-related transfer function and three-dimensional ear model database," in *Proceedings of Meetings on Acoustics 172*. ASA, 2016, vol. 29, p. 050002.

[26] K. Watanabe, Y. Iwaya, Y. Suzuki, S. Takane, and S. Sato, "Dataset of head-related transfer functions measured with a circular loudspeaker array," *Acoustical science and technology*, vol. 35, no. 3, pp. 159–165, 2014.

[27] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. II: Psychophysical validation," *J. Acoust. Soc. Am.*, vol. 85, no. 2, pp. 868–878, 1989.

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

[29] M. Guldenschuh, A. Sontacchi, F. Zotter, and R. Höldrich, "HRTF modeling in due consideration variable torso reflections," *J. Acoust. Soc. Am.*, vol. 123, pp. 3080, 2008.

[30] V. R. Algazi, R. O. Duda, R. 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. Am.*, vol. 112, no. 5, pp. 2053–2064, 2002.

[31] V. C. Raykar, R. Duraiswami, and B. Yegnanarayana, "Extracting the frequencies of the pinna spectral notches in measured head related impulse responses," *J. Acoust. Soc. Am.*, vol. 118, no. 1, pp. 364–374, 2005.

[32] S. Spagnol, M. Geronazzo, and F. Avanzini, "On the relation between pinna reflection patterns and head-related transfer function features.," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 3, pp. 508–519, 2013.

[33] E.A.G. Shaw and R. Teranishi, "Sound pressure generated in an external-ear replica and real human ears by a nearby point source," *J. Acoust. Soc. Am.*, vol. 44, no. 1, pp. 240–249, 1968.