A framework for designing head-related transfer function distance metrics that capture localization perception

Cite as: JASA Express Lett. **1**, 044401 (2021); https://doi.org/10.1121/10.0003983
Submitted: 30 October 2020 . Accepted: 11 March 2021 . Published Online: 02 April 2021

Ishwarya Ananthabhotla, Vamsi Krishna Ithapu, and W. Owen Brimijoin

A framework for designing head-related transfer function distance metrics that capture localization perception

Ishwarya Ananthabhotla,[1,a)] Vamsi Krishna Ithapu,[2] and W. Owen Brimijoin[2,b)]
[1]*MIT Media Lab, 75 Amherst Street, Cambridge, Massachusetts, USA*
[2]*Facebook Reality Labs, 9845 Willows Road, Redmond, Washington 98052, USA*
ishwarya@mit.edu, *ithapu@fb.com*, *owen.brimijoin@fb.com*

Abstract: Linear comparisons can fail to describe perceptual differences between head-related transfer functions (HRTFs), reducing their utility for perceptual tests, HRTF selection methods, and prediction algorithms. This work introduces a machine learning framework for constructing a perceptual error metric that is aligned with performance in human sound localization. A neural network is first trained to predict measurement locations from a large database of HRTFs and then fine-tuned with perceptual data. It demonstrates robust model performance over a standard spectral difference error metric. A statistical test is employed to quantify the information gain from the perceptual observations as a function of space.

[Editor: Christopher Bergevin] https://doi.org/10.1121/10.0003983

Received: 30 October 2020 **Accepted:** 11 March 2021 **Published Online:** 2 April 2021

1. Introduction

Individual head-related transfer functions (HRTFs) are complex and vary substantially from person to person because of anthropometric differences like ear shape, head circumference, and torso size.[1] Their high dimensionality and idiosyncratic nature pose problems for binaural rendering because the apparent realism of any virtual acoustics depends heavily on how closely matched the HRTF is to that of the individual listener. Indeed, using HRTFs that are different from a listener's individual HRTF can result in unacceptable distortions in the perception of sound source direction,[2,3] coloration,[4] and width, as well as a potential collapse of externalization.[5] But despite understanding these perceptual impacts, it remains unknown what constitutes a perceptually meaningful acoustic difference between two HRTFs.

It would be of great value to have an HRTF distance metric that varied along perceptually meaningful dimensions. From a basic research perspective such a metric would help us to understand what HRTF cues the brain uses to construct our internal representation of the spatial acoustic world, and how to design perceptual experiments and select spatial regions that would reveal the most about our perception of sound source location. From an applied perspective, such a metric could serve as a loss function for training HRTF prediction methods, a means for validating the reliability of acoustically captured HRTFs, or a system by which we could select an HRTF which is most appropriate for a given person from a set of HRTFs.

The most frequently used HRTF distance metric is essentially a spectral subtraction, computed as a log-scaled l_p norm directly on the filter representations in the magnitude frequency domain, known as a spectral difference error (SDE).[2] This class of error metrics is robust, easy to implement, and simple to understand, but it can fail to correlate well with, e.g., errors in a sound localization task. Simple additions such as using a spectral weighting on the residual between the two HRTFs can reduce the impact of errors at frequencies well outside the range where measurements are reliable, but cannot solve what is in essence a high dimensional problem.[2,6,7] Given the perceptual impacts of coloration shifts, individual notch position and slope, and cross-frequency interactions to name but a few, it is highly unlikely that a generalizable perceptual error metric would consist of a simple weighted spectral subtraction.

In this work, we turned instead to a statistical model to develop a more effective error metric for HRTF comparison that is better aligned with perception, focusing here on localization perception. A naive approach to this problem would entail collecting data in an experimental setting wherein (1) high-fidelity HRTFs are acoustically measured for each participant, (2) participants are presented with a series of sound sources rendered at different spatial locations using these

[a)] Author to whom correspondence should be addressed, ORCID: 0000-0002-4624-0208.
[b)] ORCID: 0000-0001-6124-3987.

measurements, and (3) are asked to relay their perception of the location of the sources. This data would then be used to create a model to describe the relationship between a given HRTF and a perceived spatial location. In practice, however, collecting the comprehensive amounts of data that this would require is costly, time-consuming, and inefficient. Since no such appropriate datasets exist, learning models with this approach would have to be derived solely from those small, sparse, noisy datasets that do exist, and the resulting models are likely to be unstable and generalize poorly—a well-known challenge in classical data learning contexts.[8,9]

To this end, we propose and demonstrate a more flexible framework for constructing an HRTF perceptual error metric. We suggest (1) constructing a model that is first built on large amounts of informative, non-perceptual data, that constitutes a "prior" on the relationship between an HRTF spectrum and its corresponding spatial location; (2) fine-tuning this model, taking into account sparse, noisy perceptual observations from existing small-scale datasets collected from procedures like the hypothetical setting described above, which we consider the "posterior" model; and (3) computing measures of statistical significance as a function of spatial location between the prior and posterior model. This has the benefit of both validating the model itself, but also informing subsequent collection of perceptual data to further improve the model in an iterative fashion.

We demonstrate this idea in practice by constructing a neural network model that is designed to predict a spatial location from a left-right pair of HRTF magnitude frequency responses. The model is first trained on a large database of acoustic HRTF measurements, and then is fine-tuned using a transfer learning approach incorporating a small set of observations from a spatial localization experiment. We illustrate the utility of this approach by showing that:

(1) The metric obtained from the "priors" is monotonic and linear with subtended angles between sound sources, in contrast to SDE which often lacks linearity.
(2) The performance of the fine-tuned model on unseen perceptual data provides insight into model uncertainty as a function of space; by obtaining measures of confidence for the difference between priors and posteriors at a particular location in space, we draw conclusions regarding the spatial regions where perceptual data does or does not provide critical information over non-perceptual information, and where more perceptual data needs to be collected to be able to draw such a conclusion.

We feel that this is a promising attempt at using machine learning to get around the problem of HRTF dimensionality and move towards a means for determining meaningful perceptual differences between HRTFs. The details of the approach follow, along with model validation, and usage examples, some specific applications of which may be found in the supplemental materials.[10]

2. Data

Measured HRIRs: ($\mathcal{H}$) consists of a database of acoustic far-field head-related impulse response (HRIR) measurements captured from 123 subjects. The 201-point, 48 KHz-sampled HRIRs were captured along a 612 point spherical grid (denoted by $\mathcal{G}$) of directions with 36 equally spaced azimuth locations and 17 equally spaced elevation locations. For the purposes of this work, we used the corresponding magnitude spectra HRTFs, normalized by the maximum spectral energy across all of the individual's measured responses.

Localization Test: ($\mathcal{L}$) is formed by the results of a listening test conducted to evaluate localization perception, using a subset of 30 individuals from $\mathcal{H}$. The participants were presented with a series of virtual sound sources, rendered using their measured HRTF, through a pair of headphones. The test was performed in a quiet room, and the participants were seated at the center of a spherical dome mounted at the center of the room. The participants were asked to identify the sound source location by pointing with a head-mounted laser pointer to the location on the dome where they perceived the sound to be coming from. This location was registered as azimuth and elevation angles relative to the initial front direction. Further details regarding the setup, experiment protocol, and spatial processing necessary for synthesizing the virtual sound sources can be found in Ref. 11.

3. Methods

We propose a computational framework that predicts the perceived spatial location of the sound from a given pair of far-field left and right HRTFs. To do this, we first constructed a learning model that directly relates this HRTF pair to its measured source location, relying only on $\mathcal{H}$; we then modified this trained model appropriately using the perceptual data in $\mathcal{L}$. This two stage framework allowed us to validate the efficacy of the trained model with and without the perceptual data, implicitly providing insight into the biases induced by the perceptual feedback. We utilized ideas from deep learning (and more precisely, feed forward convolutional networks) to design this model.

We denote the left and right HRTFs corresponding to azimuth θ and elevation ϕ by $h^L_{\theta,\phi}$ and $h^R_{\theta,\phi}$, respectively, defined over frequency. We constructed a learning model M that maps these signals to θ and ϕ. Ideally, this is a regression prediction problem from continuous inputs to continuous outputs; however, keeping in mind the sparse structure of the spatial grid $\mathcal{G}$, and the fact that a discrete output or target space is desirable for neural network training (as highlighted in audio applications like those described in Refs. 12 and 13), we transformed the prediction problem into a classification one. The outputs are denoted by $y \in [0,1]^{612}$, where y_i represents the ith direction from the grid $\mathcal{G}$.

At prediction time, given a new pair of HRTFs, the vector entry with the greatest probability in the prediction $\hat{y}$ represents the estimated source location, denoted by $(\hat{\theta}, \hat{\phi})$. These predictions from M can be used to construct an error metric for downstream tasks. Given two HRTFs h_i and h_j from some unknown locations, M estimates $(\hat{\theta}_i, \hat{\phi}_i)$ and $(\hat{\theta}_j, \hat{\phi}_j)$, which can then be used to compute the distance metric $d(|\hat{\theta}_i - \hat{\theta}_j|, |\hat{\phi}_i - \hat{\phi}_j|)$. $d(\cdot)$ can take the form of any appropriately chosen angular distance metric in the spherical domain.

To explicitly account for the influence of perceptual feedback from $\mathcal{L}$, as mentioned earlier in this section, we first trained M with $\mathcal{H}$, followed by a fine-tuning training phase with the data from $\mathcal{L}$, resulting in a model denoted $\tilde{M}$. Details regarding the architecture, design, learning, and optimization strategies for M can be found in the supplement. Training the models followed the standard learning criterion and cross-validation principles employed in the deep learning literature.[14,15]

4. Experiments and results

4.1 Without perceptual feedback

In this first set of experiments, we examined the performance of M without the inclusion of perceptual data L. The model achieves an absolute test set classification accuracy of 65.8%, with a mean distance error (absolute difference in predicted angle) of 0.67° ($\sigma = 2.7°$) and 4.7° ($\sigma = 12.7°$) in azimuth and elevation, respectively. As a more meaningful representation of performance, however, we also report the 1-bin tolerance (1BT) measure, which corresponds to the model's accuracy in predicting the true spatial location of HRTFs within one neighboring grid location (where one grid location has a resolution of 10°). In Fig. 1, we show both the 1BT and absolute accuracy as a function of space. For simplicity, we report this measure along the elevation and azimuth axes independently, aggregating the opposite axis. We note slightly decreased performance in elevation classification at the extremes, likely due to measurement noise. We also observe lower performance in azimuth prediction overall as compared to elevation prediction likely due to smaller Cartesian spacing in azimuth at higher elevations.

We next attempted to understand how the distance metric derived using M behaves in comparison to SDE. We did this using the following procedure: given an azimuth θ_0, we chose two possible grid locations along the elevation axis (ϕ_i and ϕ_j), and selected two HRTFs, $h^{L/R,P_1}_{\theta_0,\phi_i}$ and $h^{L/R,P_2}_{\theta_0,\phi_j}$ belonging to two random individuals P_1 and P_2 from the database $\mathcal{H}$. Using these, we computed a simple distance measure in the output space of M, namely,

$$L_M = |\hat{\phi}_i^{P_1} - \hat{\phi}_j^{P_2}|, \tag{1}$$

where $\hat{\phi}_i$ and $\hat{\phi}_j$ are predicted by M. We additionally computed an SDE measure from the magnitude frequency HRTFs. For the right HRTF, this is defined as

$$L_{SDE_R} = \frac{1}{N}\sum_{n}^{N} |20\log_{10}(h^{R,P_1}_{\theta_0,\phi_i}[n]) - 20\log_{10}(h^{R,P_2}_{\theta_0,\phi_j}[n])|, \tag{2}$$

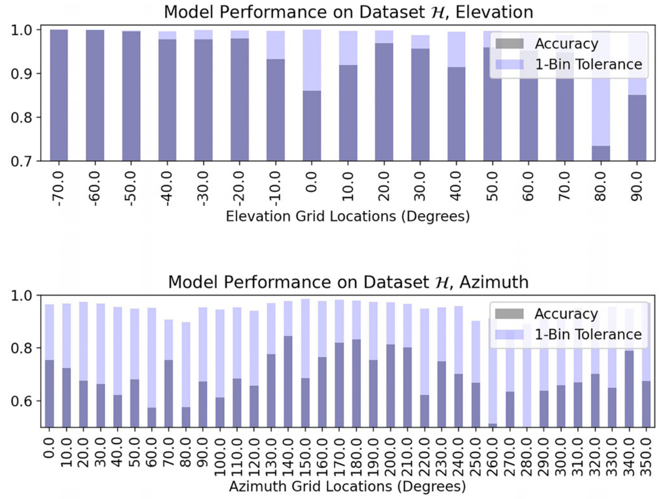


Fig. 1. Accuracy and 1BT for elevation (top) and azimuth (bottom), aggregated over azimuth and elevation, respectively.

which we repeat separately for the left HRTF; N is the number of frequency bins, equal to the tap length of the measured HRIRs (see Sec. 2). We averaged these measures (L_M, L_{SDE_R}, L_{SDE_L}) across all pairs of subjects in the test partition of $\mathcal{H}$ (approximately 50 subjects) for a given location, and repeated the procedure for every possible location along the fixed axis, choosing a few values in azimuth and elevation for the fixed axis. The results of this process are shown in Fig. 2, with trends in elevation for a fixed azimuth shown in the left image and trends in azimuth for a fixed elevation shown in the right image. We note that L_M is linear and monotonic with increasing distance in elevation and azimuth. On the other hand, while L_{SDE} for the ipsilateral ear is monotonic with increasing distance, and it lacks linearity. We also show that L_M displays significantly less inter-person variability than L_{SDE}. To ensure that these trends are robust and do not result from sampling noise, we treat each curve in Fig. 2 as a 2D distribution in angular distance and subject pair, and compute a two-sample multivariate t-test on L_M and the mean of L_{SDE_R} and L_{SDE_L}. A T^2 statistic with a p-value <0.05 suggests that the two distributions are unrelated. Taken together, these demonstrate the utility of our proposed metric—while SDE may reflect variance in distance at a course spatial resolution, our proposed metric is more robust for fine-grained angular distance comparisons, and is more robust to inter-personal spectral differences.

In Fig. 3, we provide an example to illustrate the affordances of the proposed metric. On the left, we show two magnitude HRTFs from the ipsilateral ear of two subjects which were measured 150° apart in elevation, at a fixed azimuth location of 20°; on the right, we show another pair from the same two subjects and azimuth location representing a difference of only 10° in elevation. L_M predicts a value of 150° and 10°, respectively, while L_{SDE} reports 25 dB for *both* pairs.

4.2 With Perceptual Feedback

In a second set of experiments, we explored the role of perceptual data in shaping predictions across spatial locations. To do this, we applied M and $\tilde{M}$ to the test partition of $\mathcal{L}$, and compared the performance of the two models. As an exhaustive approach, we performed an iterative hypothesis test comparing the two distributions of model predictions for each possible grid location along either the elevation or azimuth axis. This results in a measure of confidence describing whether the two distributions were drawn from the same underlying distribution. We suggest that combining this information with the models' performance as indicated by the 1BT measure provides insight into the value of perceptual data as a function of space, and we provide an illustrative example for discussion.

In Fig. 4, we plot the 1BT measure for each possible location along the elevation axis, above the p-values (plotted as $1-$p-value) resulting from the iterative hypothesis test and the number of perceptual observations available for each location from $\mathcal{L}$. An analogous plot for this analysis in azimuth can be found in the supplement. In spatial regions where the hypothesis test shows statistical significance, and $\tilde{M}$ has outperformed M, such as where $\phi = 10°$, we draw the conclusion that perceptual observations provide critical information; in regions where the improved performance of model $\tilde{M}$ is not supported by statistical significance, such as where $\phi = 0°$, we conclude that perceptual observations do not afford additional information over that already captured by M. However, in spatial regions where the hypothesis test does not show statistical significance and very few perceptual observations have been captured relative to other locations, such as where $\phi = -40°$, there is not enough certainty to draw either conclusion; instead, we suggest that this is a useful spatial heuristic to inform the collection of perceptual observations in future iterations of participant experiments.

5. Discussion

Published datasets that include localization measurements using different HRTFs are not of sufficient size to train machine learning algorithms on their own. The proposed hybrid approach described here allows us to sidestep this problem by including perceptual and non-perceptual data in a common framework, and gives us a means by which we may begin to

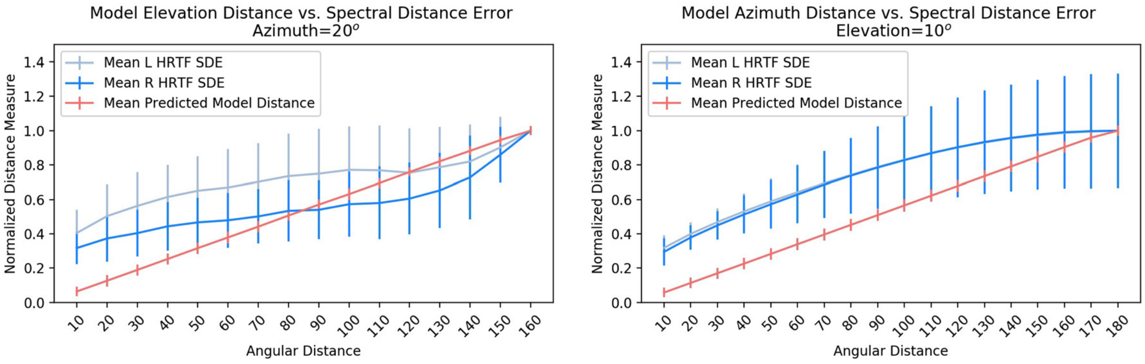


Fig. 2. We show a comparison between L_M and L_{SDE} for several pairs of HRTFs from randomly sampled subjects in $\mathcal{H}$; the vertical bars give the standard deviation across subject pairs. We show trends in elevation for a selected azimuth location (left), and trends in azimuth for a selected elevation location (right).

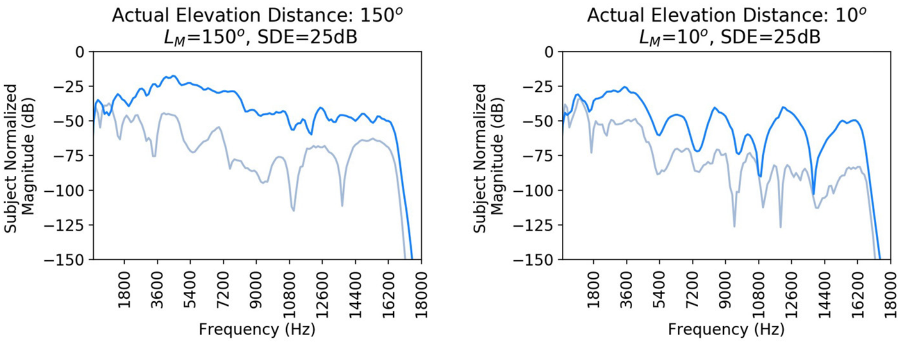


Fig. 3. A comparison of two pairs of right-HRTFs measured from the same two subjects at the same azimuth location (20°); the first pair (left) were measured 150° apart, and the second pair were 10° apart. However, both pairs result in the same L_{SDE} value.

describe the perceptual distance between two HRTFs. The proposed metric already outperforms classical measures like spectral distance. We attribute this to our choice of model—we intuit that a neural network is required to be able to robustly map HRTF signal space to spherical domain locations given the complex variance in spectral cues that determine localization across a broad population of individuals. In its current form, we posit that our proposed metric could be constructed and treated as a black-box loss function, and inserted at the tail-end of HRTF generation or selection systems to compute and propagate error back to the system in the spatial domain instead of the spectral domain.

Second, we believe that the proposed framework is valuable for constructing perceptually relevant models using sparse, weakly annotated, and noisy data—which are the typical attributes of audio perception datasets. Building a model first on data that is conditioned on a causal assumption—that, in this case, a statistically derived non-linear weighting function can map HRTF spectra to spatial locations—affords generalizability and robustness towards unseen subject variance. This model can then be adapted to a smaller set of data reflecting human annotations (from, for example, an individual calibration experiment), quickly and online using any number of domain adaptation strategies including the simple transfer learning approach proposed here. Last, computing measures of statistical difference between the generic and adapted models allows for an estimate of the information value of incoming perceptual observations and a heuristic to drive further querying for annotations; this is of importance given the cost of querying or human-in-the-loop approaches, especially on the scale of an individual user. As a key contribution of this work, we encourage a broader application of this approach to metric design for HRTF comparisons, with adaptations of the framework to different perceptual attributions (such as timbre/coloration perception) or different data formats (annotations captured at different spatial resolutions, with different classes of labels, etc.).

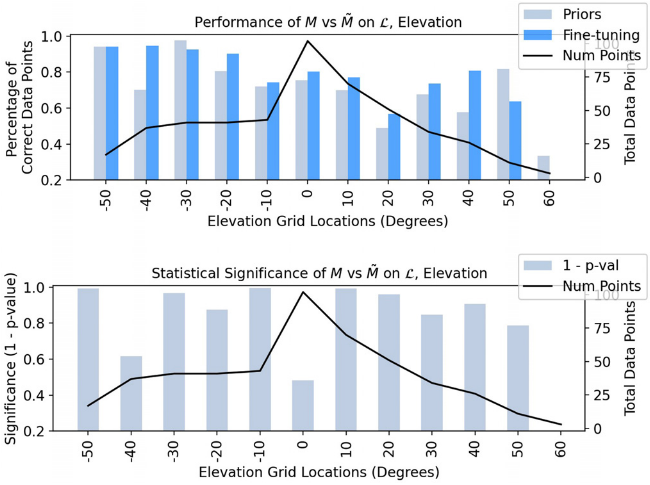


Fig. 4. We give a comparison of the performance of M and $\tilde{M}$ on the elevation axis via the 1BT measure (top); we juxtapose this with the p-values from the iterative hypothesis test comparing the two distributions (bottom).

Acknowledgments

We thank Zamir Ben-Hur, David L. Alon, Philip R. Robinson, and Ravish Mehra (from Facebook Reality Labs) for their valuable comments and suggestions regarding the dataset and for interpreting the results.

References and links

[1] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space," in *Audio Engineering Society Convention 107*, Audio Engineering Society (1999).

[2] J. C. Middlebrooks, "Individual differences in external-ear transfer functions reduced by scaling in frequency," J. Acoust. Soc. Am. **106**(3), 1480–1492 (1999).

[3] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," J. Acoust. Soc. Am. **94**(1), 111–123 (1993).

[4] C. Jenny and C. Reuter, "Usability of individualized head-related transfer functions in virtual reality: Empirical study with perceptual attributes in sagittal plane sound localization," JMIR Ser. Games **8**(3), e17576 (2020).

[5] C. Mendonça, G. Campos, P. Dias, and J. A. Santos, "Learning auditory space: Generalization and long-term effects," PloS One **8**(10), e77900 (2013).

[6] M. Geronazzo, E. Peruch, F. Prandoni, and F. Avanzini, "Applying a single-notch metric to image-guided head-related transfer function selection for improved vertical localization," J. Audio Eng. Soc. **67**(6), 414–428 (2019).

[7] F. Grijalva, L. Martini, D. Florencio, and S. Goldenstein, "A manifold learning approach for personalizing HRTFs from anthropometric features," IEEE/ACM Trans. Audio Speech Lang. Process. **24**(3), 559–570 (2016).

[8] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," IEEE Intell. Syst. **24**(2), 8–12 (2009).

[9] O. Bousquet and A. Elisseeff, "Stability and generalization," J. Mach. Learn. Res. **2**, 499–526 (2002).

[10] See supplementary material at https://doi.org/10.1121/10.0003983 for a brief overview of relevant machine learning concepts, details on constructing the presented models, extended commentary on the applications of the work, and supporting data and visualizations.

[11] Z. Ben-Hur, D. Alon, P. W. Robinson, and R. Mehra, "Localization of virtual sounds in dynamic listening using sparse HRTFs," in *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality*, Audio Engineering Society (2020).

[12] A. v d Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," arXiv:1609.03499 (2016).

[13] K. Yamamoto and T. Igarashi, "Fully perceptual-based 3D spatial sound individualization with an adaptive variational autoencoder," ACM Trans. Graph. (TOG) **36**(6), 1–13 (2017).

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (2012), pp. 1097–1105.

[15] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE (2017). pp. 131–135.

6. Supplementary Text

6.1 An Overview of Machine Learning

Machine Learning is a paradigm of statistics that comprises of mechanisms for building algorithmic models from datasets. Linear models are the simplest and most widely used machine learning algorithms. For example, a given set of inputs X are transformed to Y via a simple linear transformation W, i.e., $Y = WX$. In contrast, non-linear models such as $Y = c(X)$ are defined by a function $c(\cdot)$, where c could be a logistic curve, a polynomial etc. Neural Networks are more recent, widely studied machine learning models which are a straight forward extension of such generalized linear models, where a sequential cascade of such simple functions $c(\cdot)$ are used. In this work, the proposed statistical model is a neural network.

A system level flow-chart of the neural network used in this proposal is shown in Figure 5a. The input to the model is a pair of left and right HRTFs, and the sequence of nonlinear functions applied on these inputs are denoted by $c_1(\cdot), \dots, c_6(\cdot)$, illustrated by the rectangular blocks. Each of these functions represents a set of widely-used and well-studied operations (convolutional kernels, decimation, normalization, etc) that are parametrized by unknowns. The goal of building a neural network is to estimate these parameters using a standard optimization process called backpropagation[15]. Given the inputs and learned parameters of the functions $c(\cdot)$, the network predicts as its output a category from a list of pre-specified categories each corresponding to a discrete location in azimuth and elevation space.

The training procedure for such a model involves the following steps:

(1) Splitting the dataset into training, validation and testing sets.

(2) Using the training set for estimating the unknown parameters of the functions $c_1(\cdot), \ldots, c_6(\cdot)$ using backpropagation.

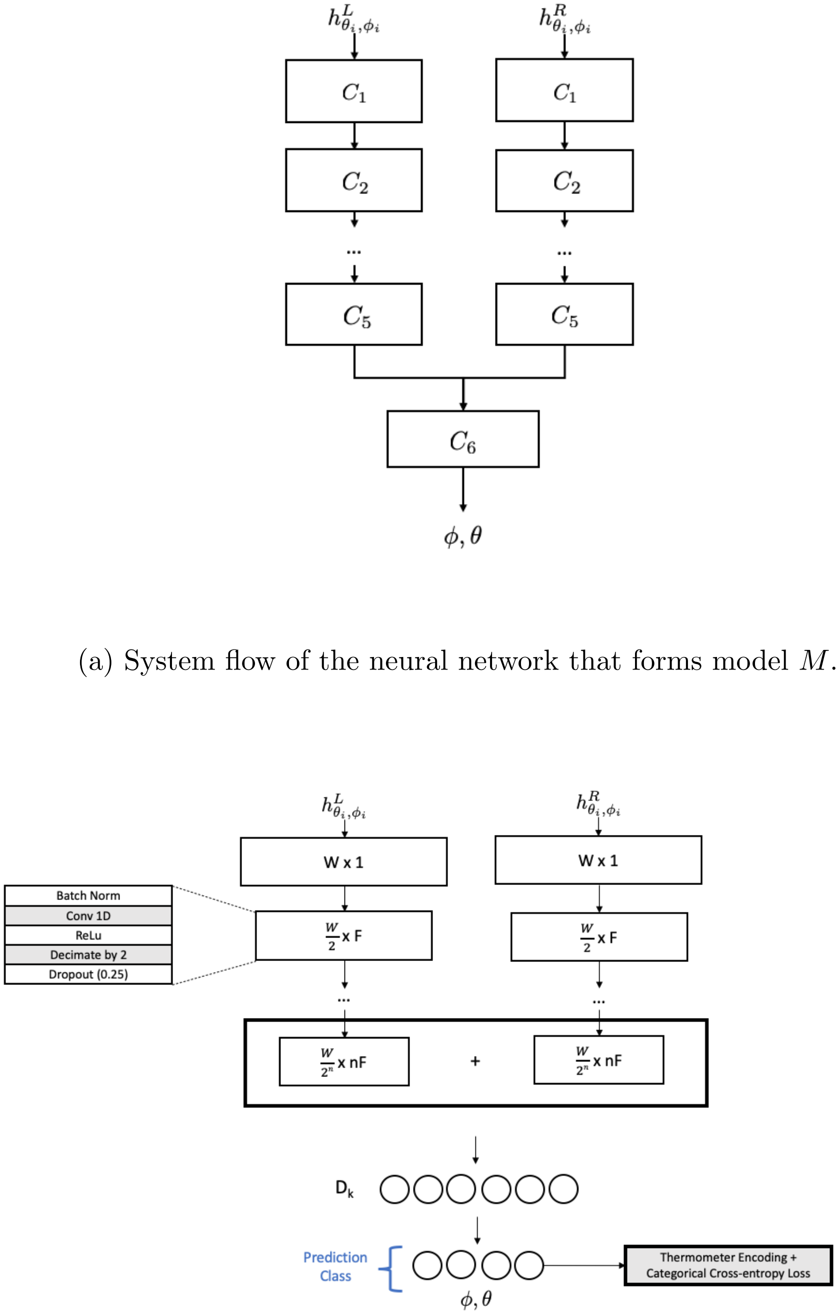
(3) Using the validation set to select the hyperparameters that control the backpropagation procedure (refer to[15] for more details).

(4) Using the testing set to evaluate the generalized performance of the learned neural network.

This learned network is the final model used to make categorical predictions.

6.2 Details on Constructing and Training M and $\tilde{M}$

In practice, model M, as shown in Figure 5b, is constructed with two stacks of convolutional blocks (shown in the inset) applied to each of the left and right magnitude HRTF inputs, and culminates in several densely connected layers. The values for the dimensional and optimization parameters in Figure 5b are given in Table 1. The categorical output of the network is first thermometer encoded following[16], and the error is computed using a categorical crossentropy loss against the ground truth. Datasets $\mathcal{H}$ and $\mathcal{L}$ are both partitioned into training, validation, and testing sets via a 70:20:10 ratio. M is first trained using only $\mathcal{H}$ for approximately 50 epochs, using early stopping criteria on the validation accuracy to end training. M is then further trained with the train and validation partitions of $\mathcal{L}$ in the same manner to obtain $\tilde{M}$.



(a) System flow of the neural network that forms model M.

(b) Detailed architecture of model M's convolutional feed-forward network.

Table 1: Structural attributes and optimization details for constructing M.

Parameter	Value
n	5
W	201
F	24
D_1	1024
D_2	512
D_3	128
D_4	64
Kernel Size	5
Kernel Stride	1
Batch Size	256
Optimizer	Adam

6.3 Extended Commentary on Motivating Applications of the Work

Here we expand on the general motivations of the work and provide guidance and procedures on how to use the proposed model for a set of 3 applications.

6.3.1 Building HRTF Personalization Systems

Over the last few years, several research groups have shown evidence that HRTF estimation entails building reasonably complex statistical models. Any such statistical model is driven by a distance function between two HRTF candidate signals. The proposal builds an opaque box that is such a metric. The model distance L_M could be used directly as a loss function to train another model predicting HRTFs from appropriate inputs, such as images of ears; it could also be used as a loss function after re-training with new localization datasets while preserving the same architecture.

6.3.2 Guiding Experimental Designs for Perceptual Tests

The proposed model provides a mechanism to design perceptual experiments that reveal valuable information about underlying perceptual attributes, helping to answer questions such as: in which locations should a localization test be performed? What should the resolution of the experiment be as a function of space? How should these questions be answered differently for different users? And given a set of experiments completed on a user, what locations should be tested in a second iteration to guarantee new information? To use the model L_M proposed here, it can be precomputed for a set of HRTFs and a set of perceptual tests. The goal of a new experiment design would be then be improve confidence in the estimation of L_M. Hence, the measured confidence intervals (such as those in Figure

2) can guide the experiments. Further, recall the analysis in Section 4.2 on using hypothesis testing to evaluate the performance of M vs. $\tilde{M}$. These statistical values are reflective of significant sample size and prediction differences. Hence, they can point to regions in space where "more information" is needed for better estimates of L_M, leading to new perception experiment designs.

6.3.3 HRTF Measurements

Measuring HRTFs is an expensive endeavor; the process is noisy and there are several constraints in terms of the equipment setup, the measurement locations, the sampling rates, the resolution of the sampling grid, etc. It would be beneficial to guide the measurement setup and process based on perception. Mathematically this guidance needs to come from the sensitivity of HRTFs given a perceptual attribute. The proposed model provides a plug-in plug-out way of constructing such sensitivity maps (and we present some such experiments via hypothesis testing in the manuscript). Firstly, the simplest such sensitivity map can be constructed from the testing set performance (1BT accuracy or a related performance measure) of $\tilde{M}$. Secondly, one can look at different trajectories in azimuth-elevation space and identify regions where L_M's performance differs significantly from SDE. This is similar to the test conducted in Section 4.2. Lastly, using multiple subsets of measured data and training strategies results in multiple approximations of L_M for the same location. The variance of these estimates can led to guidance on where more and higher-fidelity measurements of HRTFs are warranted.

6.4 Additional HRTF Comparison Examples

We show several other examples comparing L_M and L_{SDE} for pairs of HRTFs from $\mathcal{H}$ in addition to Figure 3.

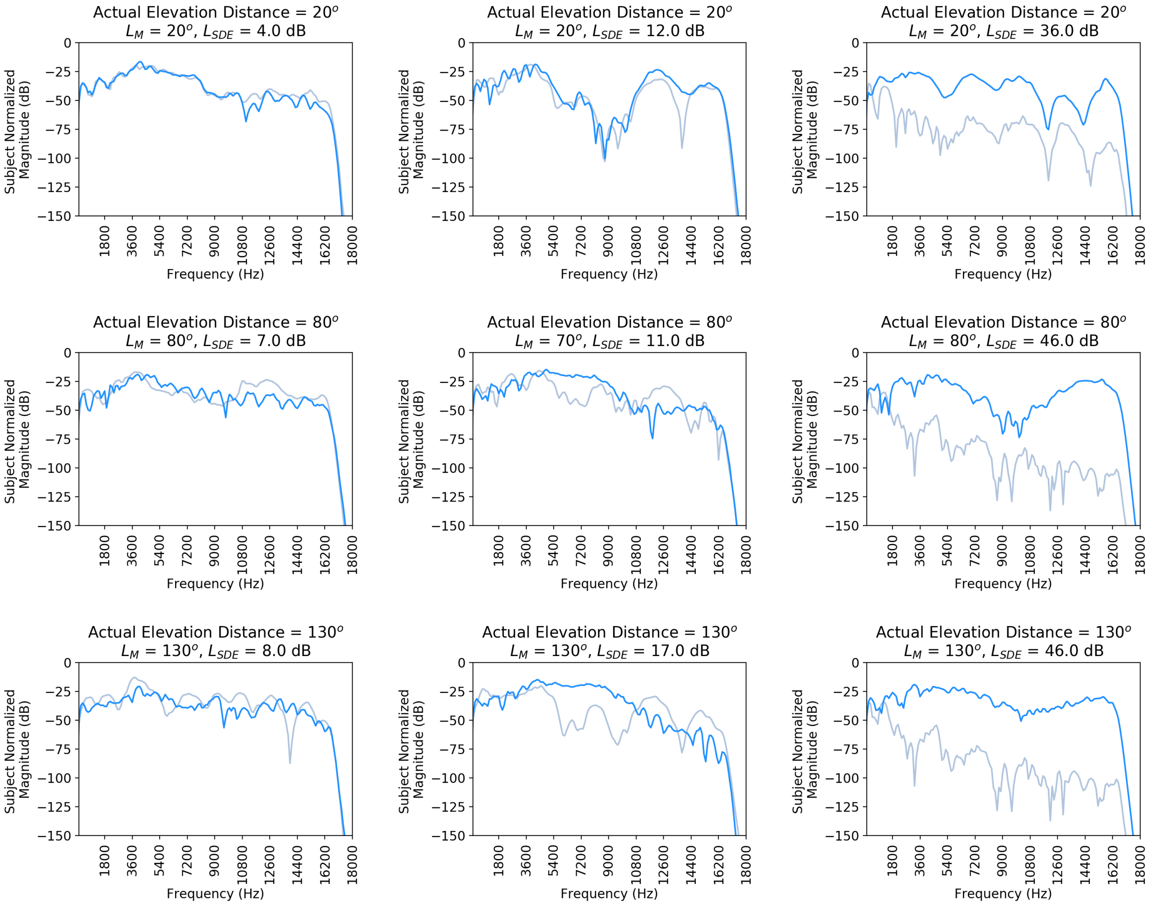


Fig. 6: **Comparing L_M and L_{SDE} for a few different ground truth angular distances in elevation; azimuth is fixed at $20°$.**

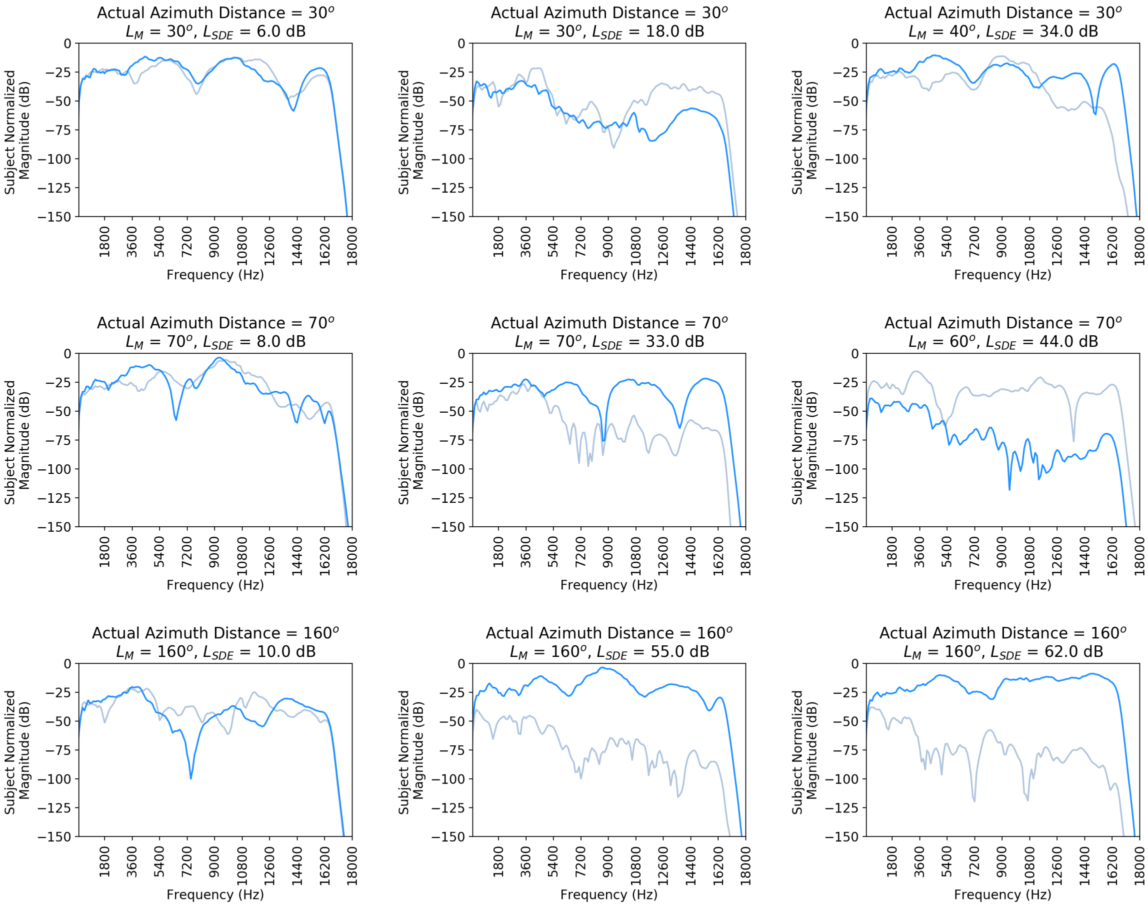


Fig. 7: **Comparing L_M and L_{SDE} for a few different ground truth angular distances in azimuth; elevation is fixed at $-30°$.**

6.5 Comparing M and $\tilde{M}$ in Azimuth

We provide the same analysis as in Figure 4, but instead plotted along the azimuth axis and aggregated over the elevation axis.

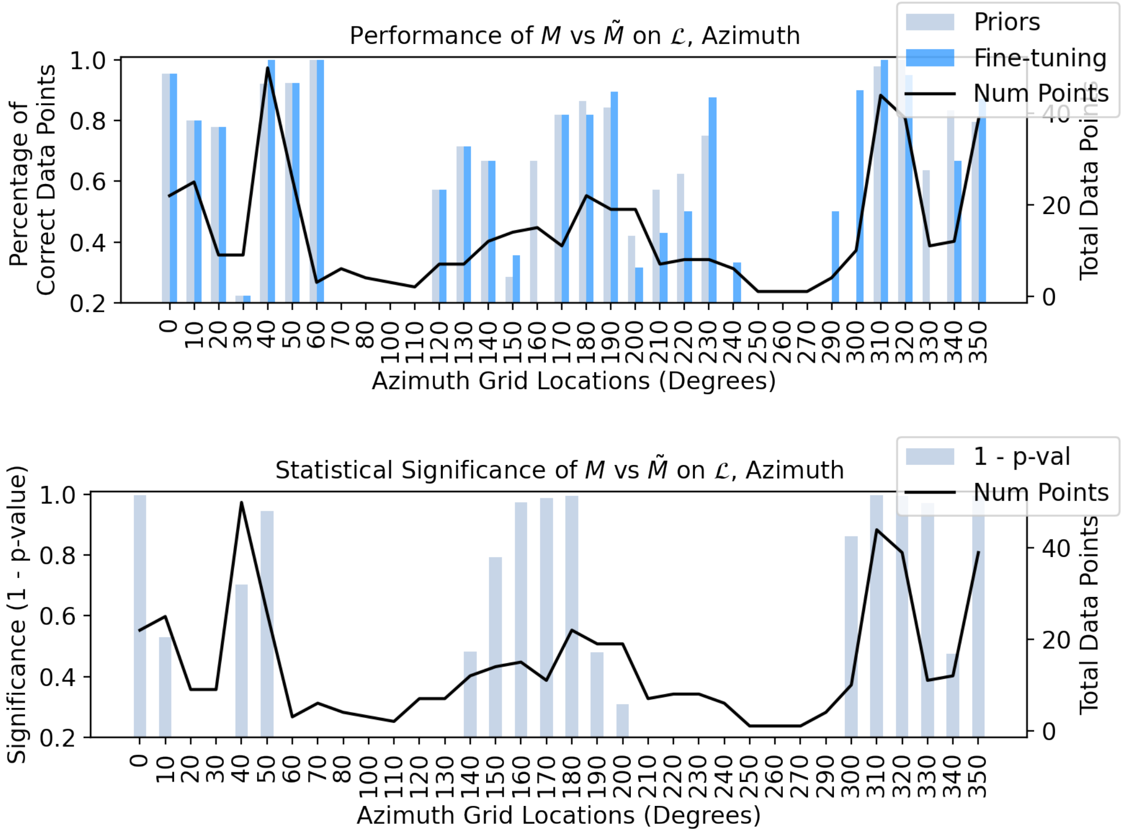


Fig. 8: **A comparison of the performance of M and $\tilde{M}$ on the azimuth axis via the 1BT measure (top); we juxtapose this with the p-values from the iterative hypothesis test comparing the two distributions (bottom).**