A Differentiable Perceptual Audio Metric
Learned from Just Noticeable Differences

Pranay Manocha1  Adam Finkelstein1  Richard Zhang2  Nicholas J. Bryan2  Gautham J. Mysore2  Zeyu Jin2

1Princeton University 2Adobe Research
{manocha,af}@cs.princeton.edu  {rizhang,nibryan,gmysore,zejin}@adobe.com

Abstract

Many audio processing tasks require perceptual assessment. The “gold standard” of obtaining human judgments is time-consuming, expensive, and cannot be used as an optimization criterion. On the other hand, automated metrics are efficient to compute but often correlate poorly with human judgment, particularly for audio differences at the threshold of human detection. In this work, we construct a metric by fitting a deep neural network to a new large dataset of crowdsourced human judgments. Subjects are prompted to answer a straightforward, objective question: are two recordings identical or not? These pairs are algorithmically generated under a variety of perturbations, including noise, reverb, and compression artifacts; the perturbation space is probed with the goal of efficiently identifying the just-noticeable difference (JND) level of the subject. We show that the resulting learned metric is well-calibrated with human judgments, outperforming baseline methods. Since it is a deep network, the metric is differentiable, making it suitable as a loss function for other tasks. Thus, simply replacing an existing loss (e.g., deep feature loss) with our metric yields significant improvement in a denoising network, as measured by subjective pairwise comparison.

1. Introduction

Humans have an innate ability to analyze and compare sounds. While efforts have been made to emulate human judgment via automatic methods, the gap between human and machine judgment remains large [see Figure 1]. This gap is acute in the context of synthetic audio based on deep learning, which has become so realistic that most metrics fail to reflect human perception. Many deep learning models rely on a metric for their loss functions; and misalignment between the loss and human judgment yields audible artifacts. Thus, the need for a perceptually-consistent metric hinders advancement of audio processing.

Based on human assessment studies, researchers have developed metrics that evaluate sound quality relative to a reference recording, e.g., PESQ [9], POLQA [10], and VISQOL [11]. However, these methods suffer from two general drawbacks. First, these models have acknowledged shortcomings such as sensitivity to perceptually invariant transformations [5, 6], which hinders stability in more diverse tasks such as speech enhancement. Second, these metrics are non-differentiable, and thus cannot be directly leveraged as a training objective within the context of deep learning.

Addressing the latter concern, researchers have trained differentiable neural networks that incorporate such perceptual models, for example estimating PESQ at each training iteration [7, 8]. The approach of Zhang et al. [7] encumbers training with expensive gradient estimation at each step, whereas that of Fu et al. [8] fails to model unseen perturbations.

An alternative is to learn a loss function via adversarial learning (GANs), which has shown promising results in enhancement [9], synthesis [10], and source separation [11]. Another approach adapts the deep feature loss [12] notion from the computer vision community, by using representations learned from a different task to construct similarity metrics [13]. This idea has been adopted for various audio tasks [14, 15]. However, these methods are problem-specific [16, 17] and require human assessment for accurate evaluation, particularly when small perceptual differences need to be measured.

We propose a new perceptual audio metric based on just-noticeable differences (JNDS) – the minimal change at which a difference is perceived. To do so, we first collect a large scale dataset of human judgments wherein subjects are asked an easy question: whether two audio recordings sound the same or different. Recordings are modified by injecting various perturbations characteristic of degradations commonly found in audio processing tasks, including noise, reverb, equalization distortion, and compression. The data collection process relies on active learning to efficiently sample such artifacts near the JND level. Next, we train a neural network with this data, and use the learned representation to construct a distance metric that measures the difference between two audio signals. We validate the new metric by showing that it correlates well with three diverse existing mean opinion score (MOS) datasets, as well as three two-alternative forced choice test (2AFC) datasets. Finally we show that using the new metric as a loss function improves the performance of a state of the art denoising network.

Thus, our contributions are as follows: (1) a framework for collecting crowdsourced human JND judgments for audio recordings; (2) a differentiable perceptual loss model trained on these data; (3) experiments showing that this model correlates better with MOS tests than standard metrics; (4) demonstrable improvement in a state-of-the-art speech enhancement network wherein the loss function is enhanced by our model; and (5) the dataset, code and resulting metric, as well as listening test examples – are all available from our project page:

http://pix1.cs.princeton.edu/pubs/Manocha2020ADP
2. Proposed Framework

We collect a dataset of human judgments using crowdsourcing tools, which have been shown to perform similarly to expert, in-lab tests [18,29] and then we fit a model to these data.

2.1. Data collection via active learning

We present a listener with two recordings, a reference $x_{\text{ref}}$ and perturbed signal $x_{\text{per}}$, and ask if these two recordings are exactly same or different, respectively, and a downstream task is to find if the binary response $h \in \{0, 1\}$. For the reference recording $x_{\text{ref}}$, we first sample a speech recording from a large collection and then degrade it by randomly applying a set of perturbations (e.g., noise and reverberation). To produce the perturbed recording $x_{\text{per}}$, we select a perturbation direction, or “axis” which can be one of several perturbation types or a combination applied sequentially. Figure 2 illustrates an example where the perturbation direction is an axis that a subject can just hear the difference between $x_{\text{ref}}$ and $x_{\text{per}}$, and ask if these two recordings are further described in Section 3.1. The perturbed recording $x_{\text{per}}$ is produced as a function $H$ of strength $\rho \in [0, 100]$. $x_{\text{per}} = H(x_{\text{ref}}, \rho)$. For example, if the perturbation is to add noise, then $H(x_{\text{ref}}, \rho) = x_{\text{ref}} + \rho \epsilon$ where $\epsilon$ is normalised white noise sampled from $N(0, 1e^{-4})$.

For values of $\rho$ that are too large or small, the answer is “obviously” different or the same, respectively, and a downstream metric is unlikely to gain information from such data. As such, we employ an active learning strategy to efficiently gather labelled data, in contrast to past approaches [20]. Our goal is to identify the just noticeable difference (JND) threshold, $\rho_{\text{JND}}$, such that a subject can just hear the difference between $x_{\text{ref}}$ and $x_{\text{per}}$. We attempt to sample $\rho$ to be close to the JND point, illustrated at a high-level in Figure 2.

We estimate the current subject’s most likely JND $\rho^*_{\text{JND}}$, based on all past answers, and then produce the next test case by $x_{\text{per}} = H(x_{\text{ref}}, \rho_{\text{JND}})$. We assume that human answers follow a Gaussian distribution with mean $\mu$ at the JND point and variance $\sigma^2$, representing human error. Following this, we compute the likelihood of $N$ past answers using $\mathcal{L}(\mu, \sigma^2) = \prod_{j=1}^{N} (1 - h_j)(1 - c(\rho_j | \mu, \sigma^2)) + h_j c(\rho_j | \mu, \sigma^2)$, where $\rho_1, ..., \rho_N$ are past perturbation strengths, $h_1, ..., h_N$ are the human judgments, and $c(\rho_j | \mu, \sigma^2)$ is the CDF of Gaussian $N(\mu, \sigma^2)$. After computing $\mu$ and $\sigma$ to maximize the above likelihood function, the next test case follows from $\rho^*_{\text{JND}} = \mu$. The ultimate product of our data collection is a database of triplets $\{x_{\text{ref}}, x_{\text{per}}, h\}$, which we leverage for training a perceptual metric.

2.2. Training a perceptual metric

A high quality perceptual distance metric $D$ would provide a small distance $D(x_{\text{ref}}, x_{\text{per}})$ if human judges feel they are the same recording, and a larger distance if they are judged to be different. Here, we explore four separate strategies to learn such a metric. We then investigate how well each method correlates with human judgments. All models have the same architecture for comparison, described in Section 3.3.

Using a pre-trained network. “Off-the-shelf” deep network embeddings have been used as a metric for training and have been shown to correlate well with human perceptual judgments in the vision setting [13], even without being explicitly trained on perceptual human judgments. We first investigate if similar trends hold in the audio setting. We describe the activation of layer $l$ of an L-layer deep network embedding as $F_l(x) \in \mathbb{R}^{T_l \times C_l}$, where $T_l$ and $C_l$ are the time resolution and number of channels of the layer, respectively. The distance between two audio recordings can be defined by averaging between the full feature activation stack:

$$D(x_{\text{ref}}, x_{\text{per}}) = \frac{1}{T} \sum_{l=1}^L |F_l(x_{\text{ref}}) - F_l(x_{\text{per}})|$$

We train a model (pre) on two general audio classification tasks from DCASE 2016 [21], namely acoustic scene classification (ASC) and domestic audio tagging (DAT), following the strategy in [14].

Training a model on perceptual data. We add linear weights over the above model $F$ as:

$$D(x_{\text{ref}}, x_{\text{per}}) = \frac{1}{T} \sum_{l=1}^L |w_l \odot (F_l(x_{\text{ref}}) - F_l(x_{\text{per}}))|$$

where $w_l \in \mathbb{R}^{C_l}$ and $\odot$ is the Hadamard product over channels. The linear weights effectively decide which channels are more or less “perceptual”. We present three variants. First, we keep the weights of all the layers $F$ fixed and only train the linear layers. This presents a “linear calibration” of an off-the-shelf network, denoted as lin. Second, we initialize from a pre-trained classification model (pre), and allow all the weights for network $F$ (and linear layer) to be fine-tuned, denoted as fin. Third, we train both the network $F$ and the linear layer from scratch.

Training objective. Our network $F$ has a small classification network $G$ at the end, which maps this distance $D(x_{\text{ref}}, x_{\text{per}})$ to a predicted human judgment $h$. We minimize the binary cross-entropy (BCE) between this predicted value and ground truth human judgment $h$:

$$\mathcal{L}(G, D) = \text{BCE}(G(D(x_{\text{ref}}, x_{\text{per}})), h)$$

Table 1: All examined perturbations and configurations.

<table>
<thead>
<tr>
<th>Category</th>
<th>Perturbations</th>
<th>Intervals/Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pops (% audio samples)</td>
<td>0.01 % to 10 %</td>
</tr>
<tr>
<td></td>
<td>Water Drop Noise</td>
<td>2 dB to 66 dB SNR</td>
</tr>
<tr>
<td></td>
<td>Room Noise</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Direct to Reverberation Ratio (DRR)</td>
<td>-27 dB to 65 dB Reverberation Time (RT60)</td>
</tr>
<tr>
<td></td>
<td>0-60 Hz, 1-10 kHz</td>
<td>8 KHz/sec to 320 KHz/sec</td>
</tr>
<tr>
<td></td>
<td>Low Frequency Bands (cut/boost bands)</td>
<td>0 to 1</td>
</tr>
<tr>
<td></td>
<td>High Frequency Bands (cut/boost bands)</td>
<td>1 to 20 %</td>
</tr>
</tbody>
</table>

Figure 2: Depiction of active learning for data collection. From each reference, we probe along a random vector of combined perturbations (dashed line) to search for the JND. Sequential listener responses (numbered) indicate recordings sound the same as (hollow dots) – or different from (solid) – the reference.
3. Experimental Setup

3.1. Perturbation Space

We apply the proposed framework to the broad field of speech telecommunication, wherein noises like packet losses, jitter, variable delay and other channel noise artifacts like sidetones are common. Table 1 lists the perturbations and their ranges we examined in our experiment. For each listening test set, we select at most one instance from each category and sample a random order to apply these categories, e.g., white noise energy from additive, DRR from reverb, MP3 bit rate from compression, equalization (EQ) and pops as the five perturbation values ($v_1, v_2, v_3, v_4, v_5$). We permute the order to simulate different scenarios. For example, (reverb, additive, EQ, compression, pops) simulates telecommunication while (compression, EQ, reverb, pops, additive) simulates playback audio in a room environment.

3.2. Crowdsourcing for Data Collection

After determining the perturbation space, we crowdsource JND answers on Amazon Mechanical Turk (AMT). We require workers to have above 95% approval ratings. We dont require workers to have any specific device (headphones/earphones) for listening. At the beginning of the Human Intelligence Task (HIT), the subject goes through a volume level calibration test in which loud and soft sounds are played alternatively. The participants are then asked not to change the volume in the middle of the HIT. Next, an attention test is presented where the participant is asked to identify a word heard in a long sentence. This removes participants that either do not understand English or were not paying attention. Upon successfully choosing the right word, the subject goes through two teaching tests, where we train the workers on what kind of differences to look for before we move on to the actual task. Each HIT contains 30 pairwise comparisons, 10 each for one randomly chosen reference and direction. Out of these 30 comparisons, 6 (20%) tests are sentinel questions in the form of obvious audio deformations. If the participant gets any of the 6 questions wrong, we discard their data. Each audio recording is roughly 2.5 seconds long, and the subjects can replay the files if they choose to. On an average, it takes 7-8 minutes to complete a HIT. At the end, we also ask for comments/suggestions/reviews from the participants on their experience in doing this HIT. We launched 2000 HITs and retained 1812 after validation, collecting about 55k pairs of human subjective judgments.

We verify that the resulting dataset has the desired properties:

1. **Balanced number of same or different answers**: Our active learning strategy predicts the JND of the listener given all their previous answers. JND is a point at which the listener is equally probable to say exactly same or different and so if our model indeed works well, we should observe that there be an almost equal number of “same” or “different” answers in our dataset. This is precisely what we observe - we observe 25782 “same” to 26130 “different” answers.

2. **Individual consistency checking**: We check consistency between all answers given by a listener and their final predicted JND level; noise levels lower than JND should say “same” and higher than JND should say “different”. Low value of user agreement would mean that the listener answered randomly and/or our method of JND prediction is not accurate, whereas a high value would mean that our model correctly predicts JND and that the users don’t answer randomly. The user agreement is around 88.3% which is high.

3.3. Training and architecture

We use a network inspired by [14] consisting of 14 convolutional layers with $3 \times 1$ kernels, batch normalisation and leaky ReLU units, and zero padding to reduce the output dimensions by half after every step. The number of channels double after every 5 layers starting with 32 channels in the first layer. We also use dropout in all convolutional layers. The receptive field of the network is $2^{14}-1$. We train the model using cross-entropy loss using a small classification model that maps distance to predicted human judgment.

We train this network for 1000 epochs, taking $\approx$ 3 days to complete using 1 GeForce RTX 2080 GPU. As part of online data augmentation to make the model invariant to small delay, we decide randomly if we want to add a 0.25s silence to the audio at the beginning or the end and then present it to the network. This helps providing shift invariance property to the model, to disambiguate that in fact the audio is similar when time shifted.

4. Results

4.1. Subjective Validation

We use previously published diverse large-scale third-party studies to verify that our trained metric correlates well on their task. We show results of our models, and compare these with embeddings obtained from self-supervised models (e.g.OpenL3 [24]) and large-scale pretrained models (e.g.VGGish [25]) trained on Audioset [26] as well as more conventional objective metrics such as MSE, PESQ [2] and VSOQL [3].

We compute the correlation between the model’s predicted distance with the publicly available MOS scores, using Spearman’s Rank order correlation (SC) and Pearson’s correlation coefficient (PC). These correlation scores are evaluated per speaker where we average scores for each speaker for each condition.

As an extension, we also check for 2-alternative forced choice test (2AFC) accuracy in which we present one reference recording and two test recordings and ask listeners which one sounds more similar to the reference. Each triplet is evaluated by roughly 10 listeners. 2AFC checks for exact ordering of similarity at per sample basis whereas MOS checks for aggregated ordering, scale and consistency. We choose four distinct classes of available datasets for our analysis:

1. **VoCo [27]**: consists of MOS tests to verify quality of 6 different word synthesis and insertion algorithms, hence not sample-aligned data.

2. **FFTNet [28]**: consists of MOS tests for synthetic audio generated by 5 different type of speech generation algorithms. It introduces artifacts specific to SE (speech enhancement), and are not sample-aligned due to phase change. The 2AFC study consists of 2050 triplets of clean reference and noisy test recordings.

3. **Bandwidth Expansion [29]**: consists of MOS tests for 3 different bandwidth expansion algorithms, aiming at increasing sample rate by filling in the missing high-frequency information. These audio samples consist of very subtle high-frequency differences. The 2AFC study consists of 1020 triplets of clean reference and noisy test recordings.

4. **Simulated**: consists of 1210 triplets of clean reference and noisy test recordings from our perturbation space described in Section 3.1.

The results are displayed in Table 2 in which our proposed method “scratch” has the best performance overall. We also summarize a few other notable observations, listed below:
We show utility of our trained metric as a loss function for Speech Models include: ours, (self)-supervised embeddings, and conventional metrics.

- Neural-network-based metrics are more robust to non-sample-aligned data - we see that most of our models, including pre, perform better than conventional metrics on non-sample-aligned synthetic audio samples in VoCo and FFTnet.
- PESQ and VISQOL have better 2AFC accuracy on FFNet and Simulated datasets but lower MOS scores suggesting that these methods preserve order but not scale. We also observe that pre has higher 2AFC accuracy but slightly lower MOS scores suggesting that it learns perceptual ordering but not the scale. Interestingly, it performs better than scratch on the above two 2AFC datasets, suggesting that training on (related) classification may produce features that better correlates with ordering (A is closer to C than B is to C). Tuning on JND data could be considered as calibrating on the scale (how much A is close to C).
- Though the pre model is robust to non-sample-aligned data, it has problems on revealing high frequency subtle difference. It is likely because this model is trained on a task (sound classification) that does not rely on these frequency bands. On the contrary VGGish and OpenL3 perform relatively well as they are trained on much larger-scale tasks (AudioSet [25]) and thus are more reliable on high frequency perceptual features. However, they perform worse on the first two tasks, which may be because they are not trained directly on speech.
- Conventional metrics such as PESQ and VISQOL perform better on VoCo and FFTNet than BWE, indicating they are less correlated to human perception when measuring subtle differences.
- Methods relying on spectrogram differences (e.g. VGGish, OpenL3, MSE) correlate poorly with MOS. Although spectrogram is relatively robust to small shifts, the change in phase can destabilize the amplitude [30], causing random variations. When two signals are very similar, this variation becomes dominant causing spectrogram differences to fluctuate and hence, decorrelate with human perception as we see in VoCo and FFTNet cases.
- OpenL3 performs better than VGGish across tasks. It may be because OpenL3 was trained using self-supervision, mapping both visual and audio onto the same embedded space, which preserves more information than classification.

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>VoCo [25]</th>
<th>FFTNet [28]</th>
<th>BWE [29]</th>
<th>Simulated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SC</td>
<td>PC</td>
<td>2AFC</td>
<td>SC</td>
</tr>
<tr>
<td>Pre</td>
<td></td>
<td>0.60</td>
<td>0.00</td>
<td>0.07</td>
<td>0.00</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td>Lin</td>
<td>0.30</td>
<td>0.45</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fin</td>
<td>0.46</td>
<td>0.71</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Scratch</td>
<td>0.71</td>
<td>0.94</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PESQ</td>
<td>0.45</td>
<td>0.85</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>VISQOL</td>
<td>0.50</td>
<td>0.78</td>
<td>0.02</td>
</tr>
</tbody>
</table>

We propose a framework to collect human “just noticeable difference” judgments on audio signals. Directly learning a perceptual metric from our data produces a metric that correlates better with MOS tests than traditional metrics, such as PESQ [3]. We also showed that a model pre-trained on classification tasks has the correct order but incorrect perceptual scale which can be calibrated by training on our data. Furthermore, we show that the metric can be directly optimized as a loss function, in the task of speech enhancement. A similar story has emerged in the computer vision literature, where trained networks have been shown to both correlate well with human perceptual judgments [13] and serve well as an optimization objective [65], compared to traditional metrics such as SSIM [37].

We would like to extend this dataset in the future to explore a broader range of types of perturbations, and do so with a greater density of samples at a wider range of intensities. We would also like to include content beyond speech, particularly music. Such data could be leveraged to study more broadly the manifold of audio perception, and enable a broader set of applications.

5. Conclusion and Future Work

We propose a framework to collect human “just noticeable difference” judgments on audio signals. Directly learning a perceptual metric from our data produces a metric that correlates better with MOS tests than traditional metrics, such as PESQ [3]. We also showed that a model pre-trained on classification tasks has the correct order but incorrect perceptual scale which can be calibrated by training on our data. Furthermore, we show that the metric can be directly optimized as a loss function, in the task of speech enhancement. A similar story has emerged in the computer vision literature, where trained networks have been shown to both correlate well with human perceptual judgments [13] and serve well as an optimization objective [65], compared to traditional metrics such as SSIM [37].

We would like to extend this dataset in the future to explore a broader range of types of perturbations, and do so with a greater density of samples at a wider range of intensities. We would also like to include content beyond speech, particularly music. Such data could be leveraged to study more broadly the manifold of audio perception, and enable a broader set of applications.
6. References


