I'm Sorry for Your Loss: Spectrally-Based Audio Distances Are Bad at Pitch

Anonymous Author(s)
Affiliation
Address
email

Abstract

We compare the performance commonly used auditory loss functions on a toy experiment, measuring the pitch distance between two stationary sinusoids. The results are surprising: many are bad at measuring pitch distance. The nature of their failings is qualitatively and quantitatively explored with a novel evaluation method, and the benefit of pitch-aware loss functions is motivated through a discussion of perceptual literature.

1 Introduction

Gradient-based learning using convolutions has been grown popularity since the late 90's in vision [31, 24, 4, 30], over the past decade in NLP [8, 54, 40] and more recently audio [22, 32, 27].

In audio modeling, explicitly calculated pitch is an important supplemental input feature, not just for music understanding, but also speech processing and other audio tasks [41, 26, 13, 27].

What is less appreciated is the importance of pitch at the other end of the network, in the loss function that guides the gradient through iterative learning. While there has been some attention on this problem to motivate specialized structures within speech networks [34, 56], our work demonstrates the extent to which pitch is not properly captured by common audio loss metrics based on spectral representations, exposing a weakness in their use as a sole optimization criterion.

2 Problem Setting

2.1 Pitch is a Percept

Rather than a physical quantity contained in the signal, pitch is a *percept* that is strongly correlated with the fundamental frequency of a sound. A complex harmonic tone, like a piano note, is made up of a series of sinusoidal partials whose frequencies are integer multiples of this fundamental frequency. An *A above middle C* on piano will have a fundamental frequency at 440Hz (i.e., cycles per second), with partials of various amplitudes at 880Hz, 1320Hz, 1760Hz and so on. But it is only the lowest, or fundamental frequency, to which we attribute pitch[1].

2.1.1 Pitch Perception in Speech vs. Music

While the relevance of pitch to music is obvious, pitch also carries important contextual and emotional cues in speech, and is the basis of prosody [16]. Typically, speech is more narrowly circumscribed in fundamental frequency, having a range of 70–1000Hz, compared to music which spans roughly

[1] This is slightly simplified, as partials are not always perfect integer multiples of the fundamental, nor is the fundamental always physically present in the signal–this latter phenomenon is known as "virtual pitch."

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

the range of a grand piano, 30–4000Hz [23]. Pitch in speech signals also varies more rapidly than in music [6]. In Section 5, we return to the question of why speech and other non-MIR audio researchers might nonetheless care about broad-scale robust pitch sensitivity.

2.2 Spectral Representations

Auditory neural networks commonly accept and output spectral representations of sound. One can generate the spectrum from a time-domain signal with the Fourier transform:

$$X(\omega) = \int_{-\infty}^{+\infty} x(t) \exp(-j\omega t)\, \mathrm{d}t, \tag{1}$$

where $x(t)$ is the signal, t is time, ω is frequency and $j = \sqrt{-1}$ is the imaginary number operator. In this formulation, time information is completely exchanged for frequency information. In order to keep both time and frequency information, the signal can be windowed into small segments or "frames." A Fourier transform is taken of each frame, resulting in a two dimensional function of time and frequency called the short time Fourier transform (STFT):

$$X(\tau, \omega) = \int_{-\infty}^{\infty} x(t) w(t - \tau) \exp(-j\omega t)\, \mathrm{d}t, \tag{2}$$

τ is the time index and $w(t)$ is the windowing function. As with the Fourier transform, the STFT produces a complex-valued function. Often only its magnitude is used, indicating the amplitude of frequency components over time; this magnitude-only version is commonly called the "spectrogram." The spectrogram and its variants (e.g., the Mel spectrogram, which pools spectral bins in a logarithmic fashion) are popular input and output representations for sound for neural networks.

All time-frequency representations suffer from the same fundamental trade-off, which is a direct consequence of the uncertainty principle [17]. The more that is known about a signal's location in time, the less is known about its frequency, and vice versa. If we shorten the time extent of a STFT window $w(t)$ in an attempt to gain temporal accuracy, we risk cutting off frequencies that require a longer time extent to fully oscillate (e.g., at a sampling rate of 44.1kHz, a 440Hz sinusoid requires about 100 time samples to oscillate just once).

It is worth noting that even for fairly long windows, spectral resolution can be fairly coarse in discrete time, spanning between 11–43Hz per bin in common applications (i.e., wider than a semitone at low frequencies).

2.3 Neural Nets that Make Sound

In this paper we focus on audio generation tasks, and their related losses. Audio generation encompasses a variety of tasks, including:

- Neural synthesis, in particular conditioned spectrogram inversion [41, 12, 38, 26]
- Audio source separation [48, 25, 21, 43]
- Denoising [19, 35]
- Upsampling [59]

2.3.1 Auditory Loss

In this work we focus on audio-to-audio losses, which are used in most audio generation and inverse synthesis tasks. Losses of this kind attempt to quantify what a human judge would consider the similarity between sounds, and are fundamentally perceptual in nature.

The notion of a distance between two sounds has a long history of investigation in perceptual research [50, 51]. While early work focused on the relative perceived magnitudes of loudness and pitch, a more recent thread in experimental psychology explores the auditory distances between instrument timbres, sounds that are more abstractly connected [39, 20, 37]. In this line of research, subjects are asked to rank the distance between multiple pairs of instrument notes. The differences are used to fit

to a "timbre space" in which all instruments are situated, using multidimensional scaling techniques. The dimensions accounting for the greatest variance in this space are then correlated to acoustic descriptors of the sound.

One particularly strong acoustical correlate is the spectral centroid, a measure that seems to show up independently in many timbre studies [36]. Its mathematical formulation is:

$$\text{Spectral Centroid} = \frac{\int \omega X(\omega)\,\mathrm{d}\omega}{\int X(\omega)\,\mathrm{d}\omega}, \tag{3}$$

expressing the weighted mean of the spectral frame. Spectral centroid is also a strong acoustical correlate of perceived "brightness" [44].

2.3.2 Differentiable Auditory Distance

While timbre spaces provide meaningful insight into the problem of auditory distance, we are interested in distances that can be computed analytically, having a differentiable calculation whose error can be back-propagated through a neural net. One prefers a loss that is also well-conditioned, perhaps a convex relation that is easy to optimize. Audio-to-audio losses can be expressed as a norm d (e.g. ℓ_1, ℓ_2, etc.) between representations of the target audio $r(x)$ and the predicted audio $r(x')$. While the raw waveform can be used to this end [41, 38, 26], a common approach is to use time-frequency representations such as the spectrogram [10, 55, 13, 14].

An augmentation of this approach uses spectrograms of multiple window-lengths simultaneously [55, 13, 14]. The Multi-scale STFT (MSSTFT) captures difference in audio phenomena at multiple scales of time and frequency, side-stepping the time-frequency trade-off discussed in Section~2.2. Interestingly, this roughly echoes a strategy used in the primary auditory cortex [7].

One can use also use features from a pre-trained neural network for auditory distances [18]. An extension of this practice, sometimes called "perceptual loss," considers the internal activations of parts or the whole of a neural net [58, 29, 14].

2.4 The Problem

Scrutinizing pitch distance in these representations is instructive. It exposes the poorly understood gap between true perceptual phenomenon and audio-to-audio distance.

From the human perspective, pitch is a natural dimension to consider when comparing sounds. Children as young as 8 months old are sensitive to melodic shape [53], which implies an early and innate sense of the relative 'up' and 'down' of pitch movement. This fundamental skill is distinct from acute pitch tuning sensitivity, which improves as a function of musical training [28].

Imagine being asked to adjust the pitch of one sine tone to match another sine tone. If the pitches start far apart, one would expect to have a fairly confident sense of the direction to tune in (i.e., up or down), and have gradually less confidence as the pitches approach in very close tuning. Figure 1a shows the ℓ_1 distance between spectrograms of sinusoids of known frequencies and a fixed amplitude. Sweeping across the spectrum, the distances from three target frequencies are shown in different colours: 130Hz, 346Hz and 921Hz. This "pitch sensitivity" is in fact the complete opposite of a natural pitch sense, locking in with extreme high confidence at short pitch distances, and being completely in the dark (i.e. having a vanishing gradient) at far distances.

This result should be obvious, given the coordinate-for-coordinate subtraction being used to measure norm distances. Comparing representations in this way allows only for very short-sighted comparison of pitch. This 'tone-deaf' quality also has significant consequences for the gradient. When the shortest path to the correct signal requires a slight spectral adjustment, the gradient often takes a circuitous path down and away from the proper solution through some other dimension.

3 Experimental Design

Our tool for this inquiry is a toy model of pure stationary sine waves. The premise is simple: in a gradient-based audio generation setting using an audio-to-audio loss, a well-behaved audio

representation should have pitch well represented, and should make gradient descent easy on a simple pitch-matching model.

Conversely, a badly-behaved audio representation will add unnecessary hurdles to optimizing pitch. Figure 1 shows badly behaved losses. Gradient-based learning over these pitch-insensitive audio representation distances will be particularly onerous in high-dimensional optimization spaces [30].

Pure sine waves have only three underlying factors of variation: amplitude (A), frequency (ω), and initial phase (ϕ):

$$x_t(A, \omega, \phi) = A\cos(\omega t + \phi) \tag{4}$$

Given a target sine wave, we seek an audio distance that permits a sine-wave model to generate a predicted sine wave with the same frequency and amplitude as the target. In our experiments, sinusoids are generated with a uniform distribution in the pitch space (or logarithmic distribution in frequency Hz) between 30–4000Hz. The signal level (hereafter "level," in decibels) is sampled from a normalized uniform decibel scale 0db through -25dB. The initial phase of isolated sinusoids is not perceptible, and we choose it randomly except when stated otherwise:

$$d(x_t(A, \omega, \phi), x_t(A, \omega, \phi') = 0 \tag{5}$$

3.1 Ideal Models of Pitch Distance

There are at least two orthogonal factors when in comes to a representation of pure pitch distance:

1. Distance determined by the ratio of the fundamental frequencies of the notes in question. For example, 660/440Hz is the equivalent pitch distance to 750/500Hz. This is complicated by the fact that human hearing is not equally sensitive to pitch throughout the spectrum. We typically hear melodic information in the range from 30Hz—4KHz [42, 51].
2. Notes separated by one or more octaves (frequency ratio of powers of two, e.g. A440Hz and A880Hz) share a "pitch class" and are equivalent in this pitch dimension. This phenomenon leads to helical representations of pitch, which curls up and around once an octave [33, 47].

To the knowledge of the authors, the relative magnitude of these two dimensions in the perception of pitch distance is an open question. For simplicity, we model only the first dimension. We assume that the ideal perceived pitch distance is strictly monotonic for fixed amplitude [51]. Perceptual experiments demonstrate that loudness is similarly strictly monotonic for fixed pitch [50]:

$$d(x(A, \omega), x(A', \omega')) < d(x(A, \omega), x(A', \omega'')) \text{ if } \omega < \omega' < \omega'' \text{ or } \omega > \omega' > \omega'' \tag{6}$$

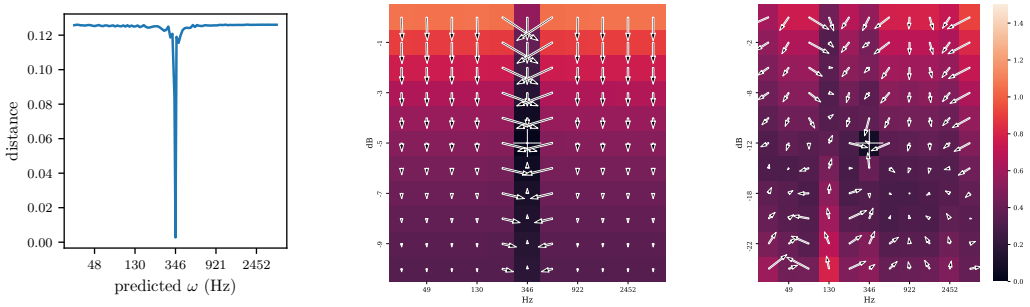
$$d(x(A, \omega), x(A', \omega')) < d(x(A, \omega), x(A'', \omega'')) \text{ if } A < A' < A'' \text{ or } A > A' > A'' \tag{7}$$

The precise parameterization of distance d is irrelevant to our evaluation measures, which is primarily concerned with the correct orientation of pitch distance. When these desiderata are violated, the gradient points in the wrong direction, which means either there is a local minimum or the loss is diverging. For example, this constraint is violated in Figure 1b in the lower left and lower right quadrants: observe that when the predicted pitch is wrong direction, the gradients push the predicted amplitude down—even if it is already too low—instead of pointing directly at the target cell in the middle. Under almost any random initialization, gradient descent will be forced to go down to a too-low amplitude, at the bottom go to the pitch, and finally crawl up the amplitude direction when the pitch is locked. Figure 1c shows the distance for the log spectrogram. Often the arrows are pointing in the wrong direction.

3.2 Evaluation Measures

In additition to properties of a distance which can be gleaned from heatmap visualization, we attempt quantify the distances in with evaluation criteria based upon the how accurately the arrows are oriented along each dimension.

We randomly sample a target sinusoid with amplitude A and frequency ω, a prediction sinusoid with amplitude A' and frequency ω' (where $A \neq A'$ and $\omega \neq \omega'$), and a perturbed prediction with the same prediction frequency ω', but a different amplitude A''. If the sampled level A' is lower than



(a) Spectrogram (b) Spectrogram (c) Example bad distance

Figure 1: Badly-behaved distances on our task. The target pitch ω is 346 Hz. Figure 1a shows the ℓ_1 distance of spectrogram magnitude between target and predictions with the same loudness, as predicted pitch ω' varies. Figure 1b shows the distances as a heatmap as both predicted pitch and predicted loudness vary, against a pure sine wave when the target pitch is in the middle of a normalized range. Figure 1c is another badly behaved distance. The x and y axes in the latter two figures represent the predicted pitch and amplitude, respectively. For the example distance (Fig 1c) and all Figures 2, the target signal is 346Hz and -12dB. The spectrogram in Fig 1b shown here has a zoomed in level scale (y-axis) to aid in visualization, and the target amplitude is -5dB.

the target A, A'' is determined to be even lower by a fixed amount. Likewise if the sampled level is greater than the target A, we produce a signal with A'' even greater. From these three signals we generate two distances (based the representation being auditioned): $\hat{d}'$ is the distance from the target signal to the first prediction, $\hat{d}''$ is the distance from the target to the second (perturbed) prediction. $\hat{d}'$ should be lower than $\hat{d}''$ (Equations 6 and 7), which gives us a 0/1 error. We repeat this process over 300 samples, perturbing both pitch and level in the same manner. Level is perturbed on the decibal scale, while pitch is perturbed in "cents," i.e., hundredths of a semitone that scales logarithmically with frequency. Formally:

$$\omega \text{ Acc} = \underset{\omega, A, \omega', A''}{\text{mean}} \text{Err}(d(x(A,\omega), x(A',\omega')) < d(x(A,\omega), x(A', \omega' + \epsilon \cdot \text{sign}(\omega' - \omega))) \quad (8)$$

$$A \text{ Acc} := \underset{\omega, A, \omega', A''}{\text{mean}} \text{Err}(d(x(A,\omega), x(A',\omega')) < d(x(A,\omega), x(A' + \epsilon \cdot \text{sign}(A' - A), \omega')) \quad (9)$$

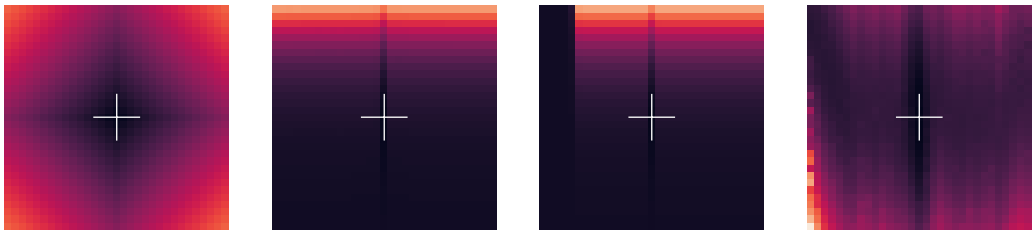
We use two magnitudes of perturbations. To measure if the gradients are correct, we measure accuracy at a fine scale (ϵ = 3 cents or 0.1dB). This is important for a gradient-learner to step in the right direction. To measure the slightly broader shape of the heatmap, we also measure accuracy at a coarse scale (ϵ = 1200 cents or 5 dB). If a gradient learner escapes local minima using stochasticity, the coarse accuracy shows if it can achieve the global minimum.

Observe in Figure 1a how there is a sharp drop at the target pitch. This behavior is present in Figure 1b through the entire y-axis. A steep slope means that a gradient learner can easily overshoot the optimum. Given that we normalize, the gradient at the optimium should be low, and definitely no more than ℓ_1's gradient of 1. As can be seen in Figure 1a and most arrows not pointing only along the y-axis in Figure 1b, there is a tendency at the extrema to have a vanishing. To quantify this behaviour we compute the slope of the distance at the high ω value of 2452 Hz.

Finally, some audio distances may be phase sensitive, which we measure by computing the median of $|d(x_t(A, \omega, \phi), x_t(A, \omega, \phi')|$, sine waves that are identical besides their phase. If this number is high, it means that the the same sine wave will appear different to the gradient-based learner, and it will never know if it has achieved a global or local optimum.

4 Results

Our results are presented in Figure 2 and Table 1. We discuss them in thematic order. The constraints in Equations 6 and 7 admit a family of permissible losses, one of which we visualize in Figure 2a for didactic purposes. It computes the l_1 distance of the pitch and loudness, and makes the strong



(a) Idealized ℓ_1 (b) Spectrogram (c) Mel (d) MFCC

(e) Mag MSSTFT (f) Spec Centroid (g) NSynth Wavenet (h) VGGish

(i) OpenL3 (j) wav2vec (big960) (k) wav2vec (small) (l) JND ConvNet

Figure 2: Target is the center: pitch 346 Hz, level -12 Hz. (a) ...

Candidate Distance	ω Acc fine	A Acc fine	ω Acc coarse	A Acc coarse	target ω slope	high ω slope	phase dist
Idealized ℓ_1	1.000	1.000	1.000	1.000	1.000	1.000	0.000
Spectral							
STFT, l1	0.523	0.500	0.883	0.510	24.977	0.050	0.002
mel	0.527	0.503	0.527	0.510	32.100	-0.013	0.001
mfcc	0.490	0.600	0.690	0.987	4.137	-3.842	0.017
MSSTFT	0.850	0.513	0.963	0.517	0.147	0.001	0.002
Spec Centroid	0.717	0.530	1.000	0.573	2.018	0.873	0.002
Neural							
vggish	0.480	0.487	0.670	0.550	158.268	39.593	0.004
openl3	0.497	0.500	0.987	0.527	146.523	106.332	0.065
wav2vec (big960h)	0.567	0.447	0.817	1.234e3	133.862	-4.422	0.995
wav2vec (small)	0.673	0.467	0.923	0.453	210.123	1.196	0.167
JND ConvNet	0.473	0.503	0.460	0.463	1258.811	210.521	0.882

Table 1: Results of audio distances on the evaluation measures (Section 3.2).

assumption that both pitch and loudness differences are equal weight in the scale we have chosen. As mentioned earlier, it is an open question the relative perceptual scaling of pitch vs. loudness.

4.1 Spectral Audio Distances

The spectrogram is a starting point for many kinds of spectral analysis. As discussed earlier, there is a steep target pitch slope. The heatmap looks smooth and, accordingly, the coarse pitch accuracy is good: 0.863. Troublingly, the fine-grained pitch accuracy is almost as bad as random guessing (0.524). Zooming in on portions of different portions of the space reveals tiny sine-like oscillations, believe to be an artifact of the extremely idealized nature of pure sinusoids, and don't expect to see this pattern in more common audio settings.

Mel spectrograms are among the most popular audio features in machine learning. By pooling STFT bins into progressively wider frequency spans, this representation roughly emulates the log-scaled nature of human frequency discrimination. However, as with most time-frequency representations, the Mel spectrogram does little to explicitly quantify patterns across the frequency axis and it correlates poorly with pitch distance.

Mel-Frequency Cepstral Coeffients (MFCCs) are another popular feature for speech and music applications. MFCCs have a better course gradient with at low loudness than the spectrogram or Mel spectrogram. At higher pitch ranges, MFCC's slope is highly negative (-3.842) and thus diverges.

The multi-scale STFT (MSSTFT) [55, 13, 14] attempts to side-step the time-frequency tradeoff (see Section 2.2 by throwing computing STFTs at varying time-scales. Due to frequency information "smearing" at shorter window sizes, multi-scaling has a pleasant side-effect when averaged with longer-scale frames: The gradient is spread horizontally over the pitch axis and the course pitch accuracy improves. Interestingly, this smearing also appears to override fine-grained oscillations at the ϵ we use to compute fine gradient.

On pure sinuisoids, the spectral centroid recovers the pitch precisely, and makes for an effective pitch distance after applying a log-transform. As per its normalization factor, it is insensitive to loudness.

4.2 Neural Audio Distances

Large-scale neural audio models are far more expressive than traditional spectral representations. Moreover, large-capacity low-prior neural networks are known to excel at domain transfer. We test a variety under our evaluation methodology, using their final layer as an audio representation.

It is well understood that *fine-tuning* a large-scale neural model on a simple task will likely succeed, even in a one-shot learning scenario [58]. However, we are motivated by the zero-shot learning scenario [49] of finding a general-purpose audio similarity measure that generalizes to unknown, unspecified tasks. Neural models have evinced uncanny ability on other zero-shot problem domains, such as uncovering analogical reasoning when trained to predict a missing word [45]. We were surprised to find deficiencies of these models on our simple domain, which speaks to the difficulty of extracting a general-purpose broad-applicability low-k-dimensional representation of audio. Neural nets that are trained only on speech perform almost at random on this music task.

Wavenet trained on musical notes from the NSynth dataset [12] have some pitch perception, but break down when there are large pitch or loudness discepancies.

VGGish [22] and OpenL3 [9] are large-scale nets trained modal-modally on audio-visual classification across all kinds of web video. On coarse pitch distance accuracy, VGGish's distance curve is a bit bumpy (0.670), more sophisticated OpenL3 does exceedingly well (0.987). They are both insensitive to absolute stationary loudness.

We auditioned two models of wav2vec [2], an even larger net trained completely on speech. As suggested by the authors, the big fine-tuned model did not perform outside the domain of interest of the fine-tuning, and was incognizant to pitch and loudness. Their pretrained small model could recognize loudness but no pitch information.

By training a convolutional neural network directly on perceptual results of just noticeable difference (JND) experiments, [35] have closed the loop on delimiting the boundaries of what phenomena are perceptual. The authors suggest using their pretrained network as a loss function. However, their

experiments are based on speech stimuli using no explicit perturbations of pitch or absolute stationary loudness, and thus this network is also pitch blind in our domain.

Neural networks that have only heard speech react in a completely chaotic way to this task, suggesting that music encodes completely orthogonal signals. The simple act of listening to music is appreciated by Wavenet, VGGish, and OpenL3, who develop at least some sensitivity to pitch.

5 Does my non-music model really need robust pitch sensitivity?

A pessimistic perspective might be that if a particular task is almost solved, it is not useful to pursue broader machine understanding. If so, what could motivate, the speech, NLP, and vision communities to include broad-scale pitch perception and musical audio in their training regimes?

Task-specific approaches can produce one-trick ponies with unexpected failure modes. [52] describes Clever Hans, a horse that ostensibly could do basic arithmetic, fooling even a zoologist. In fact, Clever Hans had no mathematical understanding and instead relied upon subconscious visual cues from his interrogator. [3] describes an analogous failure-mode: a model that predicted the tango genre exceptionally well, but only because all the tango recordings in its corpus were old and characteristically band-limited. This is similar to certain patch-based face recognition approaches where a face is recognized when presented with an adversial example where the eyes and mouth are switched. The study of adversarial examples is a growing field [1, 57].

[52] argues that evaluation should not only quantify performance on a specific task, but more aid the improvement of systems in general. Training on broad open-domain data, like the recently released FSD50K corpus [15] exposes the model to the full gamut of natural phenomena. Multi-task learning demonstrates that sample complexity reduces, training accelerates, and generalization improves by including data that surfaces more information than is present in the target task's data [8, 46].

Training on more data and more tasks is better, but there might be a ceiling [5]. In the limit, there are bounds on understanding when training even on all audio from the entire internet. In NLP, the nuances of a question like "Is an orange more like a baseball or more like a banana?" cannot be fully comprehended by a model trained purely on text, even the entire web. Learning the size of objects using pure text analysis requires significant gynmastics [11]; vision demonstrates physical size more easily.

This school of thought suggests that listening to all the audio on the web can only learn so much; true general purpose learning can only happen through grounding in more modalities than just hearing. For example, NLP language models are useful in speech recognition. Multi-*modal* learning, e.g. learning audio and visual representations simultaneously, has recently accelerated [9].

More controversial [5] is the argument that passively consuming the internet's entirety of media data in all modalities (text, vision, and audio) provides insufficient context and the broader contextual scopes of embodiment and social interaction are needed for deeper understanding.

6 Conclusion

The spectral audio distance measures are bad at measuring pitch distance, particularly when loudness is a dimension of variability, in stark contrast to the typical human experience of sound. Neural models exposed only to speech are completely pitch-deaf; those that have been exposed to music develop some pitch sensitivity.

Gradient-based methods have transformed vision, NLP and audio learning. However, strong optimization is only as good as the underlying optimization criterion. Deep learning on loss functions that behave in ill-conditioned or counter-intuitive ways just means we learn the wrong thing more effectively. Our negative results demonstrate that even in low dimensions (i.e., pitch and level), pitch distance optimizations are at best ill-conditioned. The problem is only compounded in higher dimensions, and raises the spectre that even more sophisticated audio loss functions are deficient in unknown ways and don't capture other as-yet-unknown important underlying factors of variation. The question remains what is necessary to coax auditory perception into a general-purpose differentiable loss.

Broader Impact

We present this work as an invitation to researchers to scrutinize auditory representations at large, and to hold loss functions to a higher standard, one closer to what is perceptually salient. We do not expect any negative outcomes from this work.

Models of auditory perception are implicitly based on prototypical notion of "healthy" hearing. We acknowledge that hearing ability varies widely from listener to listener, and that this line of research may be considered exclusionary from the perspective of the hearing impaired. To this extent, research in this direction may leverage data that are not wholly representative of the population. We should be cognizant of this bias and—indeed—it should be possible for machines to model different people's pitch perception.

References

[1] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018.

[2] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*, 2020.

[3] Juan Pablo Bello. Some thoughts on the how, what and why of music informatics research. `https://www.youtube.com/watch?v=mx7P3ELHsC8#t=31m47s`.

[4] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In *Advances in neural information processing systems*, pages 153–160, 2007.

[5] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience Grounds Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[6] M.J. Carey, E.S. Parris, and H. Lloyd-Thomas. A comparison of features for speech, music discrimination. pages 149–152, Phoenix, AZ, USA, 1999. IEEE.

[7] Taishih Chi, Powen Ru, and Shihab A. Shamma. Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2):887–906, August 2005.

[8] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537, 2011.

[9] Jason Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. Look, listen, and learn more: Design choices for deep audio embeddings. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3852–3856. IEEE, Brighton, United Kingdom, May 2019.

[10] Alexandre Defossez, Neil Zeghidour, Nicolas Usunier, Leon Bottou, and Francis Bach. Sing: Symbol-to-instrument neural generator. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9041–9051. Curran Associates, Inc., 2018.

[11] Yanai Elazar, Abhijit Mahabal, Deepak Ramachandran, Tania Bedrax-Weiss, and Dan Roth. How large are lions? inducing distributions over quantitative attributes. *arXiv preprint arXiv:1906.01327*, 2019.

[12] J Engel, C Resnick, A Roberts, and others. Neural audio synthesis of musical notes with wavenet autoencoders. *International*, 2017.

[13] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. DDSP: Differentiable digital signal processing. *arXiv:2001.04643 [cs, eess, stat]*, January 2020.

[14] Jesse Engel, Rigel Swavely, Lamtharn Hanoi Hantrakul, Adam Roberts, and Curtis Hawthorne. Self-supervised pitch detection by inverse audio synthesis. June 2020.

[15] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: an open dataset of human-labeled sound events. *CoRR*, abs/2010.00475, 2020.

[16] Robert W Frick. Communicating emotion: The role of prosodic features. *Psychological Bulletin*, 97(3):412–429, 1985.

[17] D. Gabor. Theory of communication. *Journal of the Institution of Electrical Engineers - Part I: General*, 94(73):58, 1947.

[18] Liang Gao, Haibo Mi, Boqing Zhu, Dawei Feng, Yicong Li, and Yuxing Peng. An adversarial feature distillation method for audio classification. *IEEE Access*, 7:105319–105330, 2019.

[19] Francois G Germain, Qifeng Chen, and Vladlen Koltun. Speech denoising with deep feature losses. June 2018.

[20] John M. Grey. Multidimensional perceptual scaling of musical timbres. *The Journal of the Acoustical Society of America*, 61(5):1270–1277, May 1977.

[21] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. Spleeter: A fast and state-of-the art music source separation tool with pre-trained models. 2019.

[22] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. CNN architectures for large-scale audio classification. *arXiv:1609.09430 [cs, stat]*, 2017.

[23] Wolfgang Hess. *Pitch Determination of Speech Signals: Algorithms and Devices*, volume 3. Springer Science & Business Media, 2012.

[24] G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

[25] Andreas Jansson, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde. Singing voice separation with deep U-Net convolutional networks. *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017*, pages 745–751, 2017.

[26] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. volume 80 of *Proceedings of Machine Learning Research*, pages 2410–2419, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[27] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. CREPE: A Convolutional Representation for Pitch Estimation. *arXiv:1802.06182 [cs, eess, stat]*, February 2018.

[28] L. Kishon-Rabin, O. Amir, Y. Vexler, and Y. Zaltz. Pitch discrimination: Are professional musicians better than non-musicians? *Journal of Basic and Clinical Physiology and Pharmacology*, 12(2), 2001.

[29] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron Courville. MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. *arXiv:1910.06711 [cs, eess]*, December 2019.

[30] Yann LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. Efficient backprop. In Grégoire Montavon, Genevieve B. Orr, and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade - Second Edition*, volume 7700 of *Lecture Notes in Computer Science*, pages 9–48. Springer, 2012.

[31] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324, 1998.

[32] Jongpil Lee, Jiyoung Park, Keunhyoung Kim, and Juhan Nam. SampleCNN: End-to-End Deep Convolutional Neural Networks Using Very Small Filters for Music Classification. *Applied Sciences*, 8(1):150, January 2018.

[33] Vincent Lostanlen, Sripathi Sridhar, Brian McFee, Andrew Farnsworth, and Juan Pablo Bello. Learning the helix topology of musical pitch. *arXiv:1910.10246 [cs, eess]*, February 2020.

[34] Hieu-Thi Luong, Xin Wang, Junichi Yamagishi, and Nobuyuki Nishizawa. Investigating accuracy of pitch-accent annotations in neural network-based speech synthesis and denoising effects. *arXiv:1808.00665 [cs, eess, stat]*, August 2018.

[35] Pranay Manocha, Adam Finkelstein, Richard Zhang, Nicholas J Bryan, Gautham J Mysore, and Zeyu Jin. A differentiable perceptual audio metric learned from just noticeable differences. January 2020.

[36] Stephen McAdams, Anne Caclin, and Bennett K. Smith. A confirmatory analysis of four acoustic correlates of timbre space. *The Journal of the Acoustical Society of America*, 112(5):2239, 2002. 2239.

[37] Stephen McAdams, Suzanne Winsberg, Sophie Donnadieu, Geert De Soete, and Jochen Krimphoff. Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, 58(3):177–192, December 1995.

[38] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. Samplernn: An unconditional end-to-end neural audio generation model, 2016. cite arxiv:1612.07837.

[39] Paul Mermelstein. Distance measures for speech recognition, psychological and instrumental. *Pattern Recognition and Artificial Intelligence*, 116:374–388, 1976.

[40] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[41] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. *arXiv:1609.03499 [cs]*, September 2016.

[42] Andrew J Oxenham. Pitch perception. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 32(39):13335–13338, September 2012.

[43] Manuel Pariente, Samuele Cornell, Joris Cosentino, Sunit Sivasankaran, Efthymios Tzinis, Jens Heitkaemper, Michel Olvera, Fabian-Robert Stöter, Mathieu Hu, Juan M Martín-Doñas, David Ditter, Ariel Frank, Antoine Deleforge, and Emmanuel Vincent. Asteroid: the PyTorch-based audio source separation toolkit for researchers. May 2020.

[44] Geoffroy Peeters, Bruno L. Giordano, Patrick Susini, Nicolas Misdariis, and Stephen McAdams. The Timbre Toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, 130(5):2902–2916, November 2011.

[45] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[46] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv:1706.05098 [cs, stat]*, June 2017.

[47] R N Shepard. Geometrical approximations to the structure of musical pitch. *Psychological review*, 89(4):305–333, July 1982.

[48] Andrew J. R. Simpson, Gerard Roma, and Mark D. Plumbley. Deep Karaoke: Extracting Vocals from Musical Mixtures Using a Convolutional Deep Neural Network. In Emmanuel Vincent, Arie Yeredor, Zbyněk Koldovský, and Petr Tichavský, editors, *Latent Variable Analysis and Signal Separation*, volume 9237, pages 429–436. Springer International Publishing, Cham, 2015.

[49] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943, 2013.

[50] S. S. Stevens. A scale for the measurement of a psychological magnitude: Loudness. *Psychological Review*, 43(5):405–416, 1936.

[51] Stanley Smith Stevens, John Volkmann, and Edwin B Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

[52] Bob L. Sturm. A simple method to determine if a music information retrieval system is a "horse". *IEEE Transactions on Multimedia*, 16(6):1636–1644, 2014.

[53] Sandra Trehub, Dale Bull, and Leigh Thorpe. Infants' perception of melodies: The role of melodic contour. *Child Development*, 55(3):821–830, 1984.

[54] Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394, 2010.

[55] Xin Wang, Shinji Takaki, and Junichi Yamagishi. Neural source-filter waveform models for statistical parametric speech synthesis. *arXiv:1904.12088 [cs, eess, stat]*, 2019.

[56] Yusuke Yasuda, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6905–6909, Brighton, United Kingdom, May 2019. IEEE.

[57] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9):2805–2824, 2019.

[58] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *arXiv:1801.03924 [cs]*, April 2018.

[59] Zhoutong Zhang, Yunyun Wang, Chuang Gan, Jiajun Wu, Joshua B Tenenbaum, Antonio Torralba, and William T Freeman. Deep audio priors emerge from harmonic convolutional networks. page 12, 2020.