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

Joseph Turian
lastname@gmail.com

Max Henry
maxsolomonhenry@gmail.com

Abstract

Growing research demonstrates that synthetic failure modes imply poor generalization. We compare commonly used audio-to-audio losses on a synthetic benchmark, measuring the pitch distance between two stationary sinusoids. The results are surprising: many have poor sense of pitch direction. These shortcomings are exposed using simple rank assumptions. Our task is trivial for humans but difficult for these audio distances, suggesting significant progress can be made in self-supervised audio learning by improving current losses.

1 Introduction

Rather than a physical quantity contained in the signal, pitch is a *percept* that is strongly correlated with the fundamental frequency of a sound. While physically elusive, pitch is a natural dimension to consider when comparing sounds. Children as young as 8 months old are sensitive to melodic contour [65], implying that an early and innate sense of relative pitch is important to parsing our auditory experience. This basic ability is distinct from the acute pitch sensitivity of professional musicians, which improves as a function of musical training [36].

Pitch is important to both music and speech signals. Typically, speech pitch is more narrowly circumscribed than music, having a range of 70–1000Hz, compared to music which spans roughly the range of a grand piano, 30–4000Hz [31]. Pitch in speech signals also fluctuates more rapidly [6]. In tonal languages, pitch has a direct influence on the meaning of words [8, 73, 76]. For non-tone languages such as English, pitch is used to infer (supra-segmental) prosodic and speaker-dependent features [19, 21, 76].

In this work, we demonstrate pitch-based failure modes for audio-to-audio distances commonly used as loss functions. Negative results have potentially deleterious consequences on this benchmark: gradient-descent takes a circuitous route for some losses, implying that generalization speed and final accuracy are impaired. Worse yet, other loss functions get stuck in varying scopes of local minima. Their search spaces contain both fine-grained oscillations and broader plateaus.

1.1 Perceptual Audio-to-Audio Distances

Audio-to-audio distances attempt to quantify what a human judge would consider the similarity between sounds. Therefore, we argue that they are fundamentally perceptual in nature. The notion of a distance between two sounds has a long history of investigation in perceptual research [59, 60]. While early work focused on loudness and pitch distances in simple tones, a more recent thread in experimental psychology explores the auditory distances between instrument timbres, sounds that are more abstractly connected [26, 45, 47]. In this line of research, subjects are asked to rate the distance between multiple pairs of recorded instrument notes. The differences are used to fit to a timbre space in which all instruments are situated. The dimensions accounting for the greatest variance in this space are then correlated to acoustic descriptors of the stimuli. One particularly strong acoustical correlate is the spectral centroid, a measure that appears independently in many

1st I Can't Believe It's Not Better Workshop (ICBINB@NeurIPS 2020), Vancouver, Canada.

timbre studies [44]. Spectral centroid expresses the weighted mean of the spectral frame normalized by frame energy—in a frame containing a single sinusoid, the spectral centroid will be its frequency. It is a strong acoustical correlate of perceived "brightness" [53].

1.2 Automatic Audio-to-Audio Distances

Many important audio modeling settings rely upon a perceptually-coherent distance between two audio representations, both in generative and constrastive setttings. Generative audio losses can be viewed as divergences between audio representations. Thus, audio generation tasks use audio-to-audio distances: a) neural synthesis, in particular conditional spectrogram inversion [17, 34, 46, 66]; b) audio source separation [29, 33, 52, 58]; c) denoising [23, 43]; and d) upsampling [75], not to mention audio self-supervised learning in general.

LeCun [40] argues that most human and animal learning is self-supervised, whether generative or contrastive. Contrastive losses are based upon knowing that the representation distance between the target and a *similar* prediction is lower than the representation distance between the target and a *dissimilar* prediction. The contrastive approach involves fewer parametric assumptions than the generative one, and so we use it in our dissimilarity-ranking experiments (see Section 3.3).

1.3 Learning Representations of Pitch

Knowing that pitch is an important dimension for representation learning, there are several options for inducing pitch in learned representations: data, features, and loss functions.

Engel et al. [15] supplement their *data* with synthesized audio of a known pitch to aid in self-supervised pitch learning. One may use explicit pitch-*features* from a pitch extractor, such as CREPE [35] or SPICE [24] as input [27, 42] or conditioning information [34, 66, 73]. Nonetheless, learning new data and incorporating better features is gated by the underlying loss function.

Audio *loss* construction typically involves some prior knowledge from digital signal processing, for example in their choice of time-frequency representation. What is surprising is that these well-motivated distances perform poorly in our simple pitch experiments. This negative result explains why methods using spectral losses nonetheless end up using pitch-tracker models for conditioning [16] or reconstruction [15]. Alternately or in addition to DSP-motivated losses, knowledge from auxiliary models can be distilled using feature-matching losses (Section 2.1).

2 Distances based on Spectral Representations

Auditory neural networks commonly use spectral representations of sound. One can generate a (static) spectrum from a time-domain signal with the Fourier transform in Equation 1:

$$X(\omega) = \int_{-\infty}^{+\infty} x(t) e^{-j\omega t} dt \quad (1) \qquad X(\tau, \omega) = \int_{-\infty}^{\infty} x(t) w(t-\tau) e^{-j\omega t} dt \quad (2)$$

where $x(t)$ is the signal, t is time, ω is frequency and $j = \sqrt{-1}$ is the imaginary unit. The Fourier transform exchanges all time information for frequency information, which is not always desirable for analysis purposes. In order to keep both, the signal can be windowed into small segments or "frames." Taking a Fourier transform of each frame results in a two dimensional function called the short-time Fourier transform (STFT), shown in Equation 2, where τ is the time index and $w(t)$ is the windowing function. As with the Fourier transform, the STFT produces a complex-valued function. Its real-valued modulus is sufficient for many applications; this magnitude-only version is commonly called the "spectrogram," and indicates the amplitude of frequency components over time. The spectrogram and its variants (e.g., the Mel spectrogram) are among the most popular audio representations for signal analysis.

All time-frequency representations suffer from the same fundamental trade-off, which is a direct consequence of the uncertainty principle [20]: the more that is known about a signal's location in time, the less is known about its frequency, and vice versa. Shorter STFT frames gain temporal

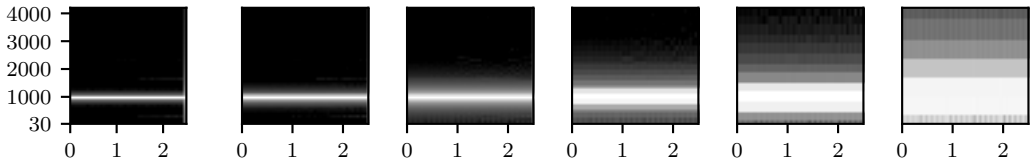


Figure 1: Spectrograms for a 1000 Hz sine tone sampled at 44.1kHz, shown for various window sizes: 2048, 1024, 512, 256, 128, and 64 (left-to-right). x-axis is time in seconds and y-axis is Hz.

accuracy, but cut off frequencies that require a longer time extent to fully oscillate (Figure 1). Window functions $w(t)$ are designed to fade frames in and out, and introduce their own spectral profile into the signal. The choice of window is a trade-off between reducing the spectral noise floor (lower "side-lobes") and gaining spectral precision (narrower "main lobes") [28]. We use a Hann window throughout.

2.1 Differentiable Audio Distances

We are interested in *differentiable* audio distances that are well-behaved and easy to optimize. Audio-to-audio distances are typically the ℓ_1 or ℓ_2 between representations of the target audio and the predicted audio. While the raw waveform can be used to this end [34, 46, 66], a more common approach is to use the ℓ_1 or ℓ_2 distance between the spectrogram or log-magnitude spectrogram [12, 55, 69, 71].

A particular loss might have certain weaknesses in certain situations, such as gradient instability. Multiple losses may be mixed to work in tandem, and hopefully add up to more than the sum of their parts. The multi-scale spectrogram (MSS) captures difference in audio phenomena at multiple scales of time and frequency in an attempt to side-step the time-frequency trade-off discussed in the previous section. Interestingly, this roughly echoes a strategy used in the primary auditory cortex [7]. As before, multi-scale losses are typically computed as ℓ_1 or ℓ_2 over linear- and log-spectrograms [5, 13, 15, 16, 68, 70, 72].

Alternately, one may use neural auditory distances. In the "knowledge-distillation" or "teacher-student" framework, one net trains the other [32]. A specific formulation of this approach, sometimes called "perceptual loss," or "feature-matching," considers the internal activations of parts or the whole of a neural net [15, 37, 55, 74]. Even specifically for pitch-learning, there are numerous examples of a feature-matching losses: a typical speech VQ-VAE is pitch-insensitive [67], but one can add a pitch reconstruction loss [21] or learn an additional pitch-only VQ-VAE [76].

3 Experimental Design

Our synthetic benchmark suggest that losses based upon common spectral representations, as well as many off-the-shelf neural representations, have difficulties tracking pitch orientation. Our tool for this inquiry is a plain model of pure stationary sine waves: a well-behaved audio-to-audio loss should have high pitch-orientation accuracy, both on a fine scale as well as a coarse one. This approach is similar in spirit to [39], which uses gradient-orientation as an evaluation criterion on a toy spheres task (albeit in the setting of adversarial learning).

3.1 Pure Sinuisoids

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

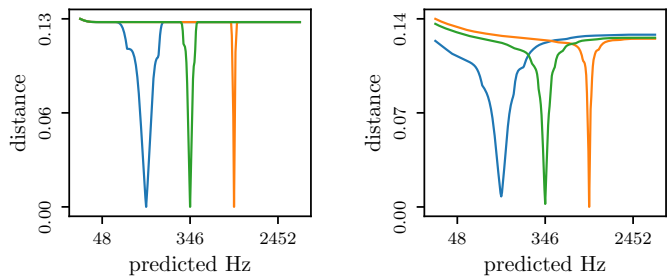
$$x_t(A, \omega, \phi) = A \cos(\omega t + \phi) \tag{3}$$

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 level as the target. Note that while linear amplitude

and frequency are used to generate the signal, these quantities are more meaningfully described by their logarithmic equivalents [60]. Thus, in our benchmark, sinusoids are generated with a uniform distribution in the *pitch space* (or logarithmic distribution in frequency) between 30–4000Hz. The *signal level* (hereafter, "level") is sampled from a normalized uniform decibel scale 0db through -25dB. We calculated normalized level of a sinusoid as

$$L_A = 25 \log_{10}(A), \tag{4}$$

such that a maximum amplitude A of 1 corresponds to a signal level L_A of 0dB. The initial phase ϕ of isolated sinusoids is not perceptible. We randomize it to remove a confounding variable.



(a) Spectrogram (b) MSS

Figure 2: In the above figures, the three curves show distances from target pitches fixed at $130, 346$, and 922 Hz. The target and prediction signal level are fixed at -12.5dB. Figure 2a shows the spectrogram-based distance between the target and prediction, as the predicted pitch ω varies. Figure 2b shows the MSS (multi-scale spectrogram) distance.

3.2 Tuning A Sine: Thought Experiment

Imagine that an experimenter asks you to adjust the pitch of one sine tone oscillator to match another. If the pitches start far apart, you would have a confident sense of the direction to tune in, up or down, and have gradually less confidence as you approach the target pitch.

Figure 2a shows the spectrogram-based distance, as predicted pitch sweeps across the spectrum. The observed pitch sensitivity is in many ways the opposite of a natural pitch sense: at great distances it has no sense of pitch orientation whatsoever (a vanishing gradient), and it locks-in with extreme confidence at fine pitch distances. This result stands to reason, given the coordinate-for-coordinate comparison implicit in norm-based distance calculations—an effect that is only exaggerated in spectrograms with large frame size, and therefore fine spectral resolution (Figure 1, left). The MSS-based distance (Figure 2b) somewhat alleviates this problem, but still suffers from a sharp gradient near the target and a vanishing gradient in high predicted pitch extrema. Finally, our distances are measured on a log-linear frequency scale. However, the spectrogram operates on a linear scale, so lower target pitches have slightly wider regions with informative gradients.

3.3 Gradient-Sign (Dissimilarity) Ranking Accuracy

In a well-behaved distance, a target sinusoid should be closer to a predicted candidate sinusoid than to a *perturbed* candidate sinuisoid that is further away by construction. This argument motivates an evaluation using a dissimilarity-ranking accuracy [1, 50], i.e. whether the gradient is pointed in the correct direction.

We randomly sample a target sinusoid with level A and pitch ω, and a prediction sinusoid with level A' and pitch ω' (where $A \neq A'$ and $\omega \neq \omega'$). We also construct a third perturbed sinuisoid. The perturbed signal level is the same as the prediction signal level. The perturbed signal pitch $\omega'' := \omega' + \varepsilon \cdot \mathrm{sgn}(\omega' - \omega)$ is, by construction, ε further away from the target ω than the prediction signal pitch ω'. From these signals we compute two distances: the distance from the target to the prediction, and the distance from the target to perturbation. In a well behaved distance, the prediction should be closer to the target than the perturbation is. This condition gives us a 0/1 error. Repeated and averaged over 1000 trials, high accuracies indicate that gradients usually point in the right

direction. We reproduce the same process using perturbations in level with a fixed pitch, using $A'' := A' + \varepsilon \cdot \mathrm{sgn}(A' - A)$. Formally:

$$\omega \text{ Acc} := \underset{\omega, A, \omega', A''}{\text{mean}} \mathbb{1}\big(d(x(A,\omega), x(A',\omega')) < d(x(A,\omega), x(A',\omega'')\big) \tag{5}$$

$$A \text{ Acc} := \underset{\omega, A, \omega', A''}{\text{mean}} \mathbb{1}\big(d(x(A,\omega), x(A',\omega')) < d(x(A,\omega), x(A'',\omega')\big) \tag{6}$$

In addition to numeric computations using explicit values of ε, a gradient may be solved and computed analytically using automatic differentiation. This is the equivalent of letting this value go to zero in Equations 5 and 6. We define $\epsilon := \varepsilon \to 0$.

We make no assumptions about pitch- and level- distance, besides our choice of range (30Hz-4KHz, -25dB to 0dB) and that sampling occurs logarithmically with respect to A and ω (Equation 3).

4 Results

We present quantitative results in Table 1. Each column indicates the average accuracy for the analytic, fine- and coarse-numeric conditions, with respect to changing pitch and changing level. The sign-accuracy of analytic gradients ($\epsilon \approx 0$) demonstrates whether the distance is capable of breaking out of very fine-grained local minima. While in practice gradient descent uses the analytic gradient, we also present numerically-computed gradients to characterize local minima at broader resolutions. The *fine* perturbation values (30 cents or 2dB) are meant to be plausibly detectable by a discerning human listener. The *coarse* perturbation values (600 cents or 10dB) are meant to be obvious to a human listener with normal hearing, and to give an indication of broad-scale local minima in the search space.

For qualitative elucidation, we provide coarse heatmaps based on a fixed target pitch and level in Figure 3. Many distances are sensitive to the choice of target signal (Figures 2a and 2b), so heatmaps based on a fixed target do not necessarily illustrate the entire search space. Supplementary Figures 5 and 6 depict coarse heatmaps for other target pitches and levels.

4.1 Spectral Audio Distances

The **spectrogram** (the modulus of Eq. 2) is a starting point for many kinds of spectral analysis. A spectrogram-based distance can roughly model pitch in the coarse condition (69.5%), but otherwise the gradients score near chance (i.e., they are not pointed in any meaningful direction). Level-perturbed gradients score near random in both coarse and fine conditions. In the case of both pitch and level, the analytical gradient is in agreement with the fine perturbation conditions. Interestingly, the log-magnitude spectrogram has some level sensitivity, but is worse at pitch tracking than its linear equivalent.

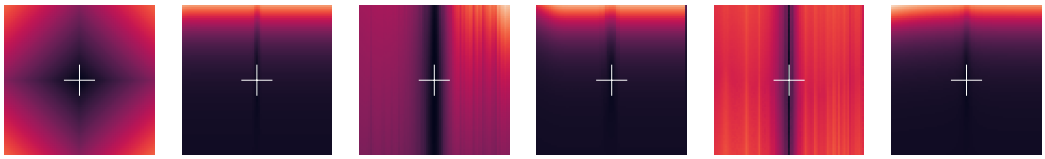
Mel spectrograms are a popular variant of the spectrogram based on the Mel scale [60]. By pooling frequency bins into progressively wider frequency spans, this representation roughly emulates the log-scaled nature of human frequency perception. As with the spectrogram, the Mel spectrogram-based distance performs near chance (if slightly above) in all conditions.

Mel-frequency cepstral coefficients (MFCCs) are a popular feature for speech and music applications based on cepstral analysis [49]. In [69], the authors argue that MFCCs are somewhat pitch invariant. Fittingly, the MFCC representation struggles with pitch, having an accuracy slightly worse than the spectrogram, though it has the best level score among the spectral representations we tested. The MFCC analytic gradient computed on level is entirely at odds with the fine perturbation results, scoring nearly at chance. This phenomenon is reflected in the extra fine-grained vertical streaks in the MFCC heatmap (Figure 3e). At a broader resolution the curves look smooth; at super-fine resolutions they are rocky.

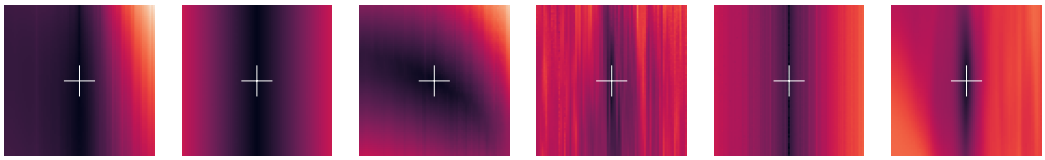
Multi-scale spectral loss (MSS) compares spectrograms at multiple scales—in our work, we use the six window sizes shown in Figure 1 for MSS. In effect, this mitigates the poor behavior of the spectrogram-based pitch distance (Figure 2). Indeed, the pitch-perturbed gradient is correctly oriented 77% of the time. Still the gradient is not entirely smooth (Figure 4a). MSS has poor analytic level gradients (55%). The heatmap in Figure 4b shows the level-gradient always drives the level down and away from the target signal, even when the prediction level is lower than than the target. Only

Candidate Distance	ω Acc $\pm\epsilon$	ω Acc $\pm$ 30 cents	ω Acc $\pm$ 600 cents	A Acc $\pm\epsilon$	A Acc $\pm$ 2 dB	A Acc $\pm$ 10 dB
Spectral representations						
Spectrogram	0.617	0.574	0.695	**0.535**	**0.518**	**0.514**
log(Spectrogram)	0.548	**0.541**	0.679	0.607	0.577	0.615
Mel	**0.511**	**0.479**	**0.580**	0.564	**0.544**	0.551
MFCC	**0.532**	0.603	0.648	0.593	*0.931*	*0.999*
MSS	0.771	*0.905*	*0.978*	0.550	**0.530**	**0.531**
log MSS	**0.532**	0.665	*0.879*	0.719	0.730	0.783
$\log_2$(Spectral Centroid)	**0.515**	*1.000*	*1.000*	**0.460**	**0.503**	**0.518**
Neural audio representations						
nsynth wavenet		0.588	*0.862*		*0.873*	*0.938*
vggish		**0.536**	0.652		0.595	0.636
openl3		0.594	*0.989*		**0.507**	**0.480**
wav2vec 2.0 Large (LV-60), no fine-tuning		0.727	*0.831*		0.682	0.738

Table 1: Accuracy of gradient orientation, at various resolutions, over 1000 trials. The worst case 95% CI is ± 0.032. For spectral representations, resolution $\epsilon \to 0$ is the analytically-computed gradient, followed by fine-grained and coarse-grained numerically-computed gradient-orientation accuracy. Bold indicates accuracy $\leq 55\%$, near random gradient orientation at this resolution. Italics indicates accuracy $\geq 80\%$, which is likely to escape local minima using stochasticity.

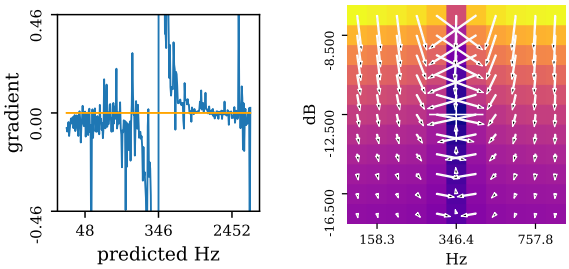


(a) Idealized ℓ_1 (b) Spectrogram (c) log Spectrogram (d) Mel (e) MFCC (f) MSS



(g) log MSS (h) Spectral Centroid (i) NSynth Wavenet (j) VGGish (k) OpenL3 (l) wav2vec2 LV-60

Figure 3: Measured distance between the target and prediction. Target pitch and level are fixed in the center of the search space (346 Hz and -12.5dB, respectively). The prediction pitch (x-axis) spans 30–4000Hz and the prediction level (y-axis) spans -25 dB–0 dB. The resolution of these heatmaps is 106 cents and 0.31 dB per cell. Figures 5 and 6 show heatmaps for different targets.



(a) Analytic MSS gradient (b) Numeric MSS gradient

Figure 4: 4a shows the analytic gradient of MSS, with levels fixed at -12.5dB and target pitch $\omega = 346$ Hz. Figure 4b shows a 2.5x zoomed in heatmap of the MSS distance. As in the heamaps of Figure 3, the target pitch and level are the center ($\omega = 346$ Hz, $A = -12.5dB$). However, the resolution of 340 cents and 1 dB per cell. Arrows illustrate the numeric gradients at this resolution.

when pitch is locked-in are the level gradients correctly oriented. Hence, MSS level accuracies are 55%. The log MSS outperforms the log-spectrogram, improving mainly on level but not pitch.

The log-scaled **spectral centroid** recovers the pitch precisely on pure sinusoids. As per its normalization factor, it is insensitive to level. However, its analytic gradients are highly chaotic for both level and pitch, scoring at chance.

4.2 Neural Audio Distances

Large-scale neural audio models are far more expressive than traditional spectral representations. Neural models have evinced uncanny ability on other zero-shot problem domains, such as uncovering analogical reasoning when trained to predict a missing word [54]. We test a variety of models under our evaluation methodology, using their final layer as an audio representation.

NSynth Wavenet is trained on musical notes from the NSynth dataset [17]. It has a strong coarse level accuracy (93%) and coarse pitch accuracy (86%). At finer pitch distances the pitch accuracy suffers (59%). This is perhaps because the dataset comprises single notes, each spaced one semitone (100 cents) apart.

VGGish [30] and **OpenL3** [10] are large-scale nets trained on a variety of audio samples taken from internet video, the latter incorporating multi-modal information from the video feed. Interestingly, both are insensitive to level, scoring just slightly above chance. OpenL3 is highly pitch accurate at a broad scale (99%), but scores near random (59%) at a fine scale.

We auditioned many pre-trained models of **wav2vec 2.0** [2], a large-scale neural net trained completely on speech; we achieved the best results using their large (LV-60) no fine-tuning model. This model has good coarse-pitch accuracy (83%), despite having never listened to music.

Overall, we were surprised to find some deficiencies of these models on our benchmark, which speaks to the difficulty of extracting a general-purpose low-dimensional representation of audio. Poor results of these representations is not meant to impugn whether pitch can be extracted from these representations using one or more layers. Fine-tuning a large-scale neural model on a simple task will likely succeed, even in a one-shot learning scenario [74].

5 Discussion

Pitch and level are both natural dimensions to consider when comparing sounds. To a listener with typical hearing, making small adjustments based on pitch or level is as simple as tuning a guitar, or turning down the radio. So why is it that common spectral-based loss functions have difficulty with our benchmark? While some distances fare relatively well on a strictly pitch-based perturbation, almost all do poorly when introducing any perturbations of level. A further mystery still, why is the search space apparently bumpy at very fine resolutions?

In any norm-based distance, one makes the implicit assumption of orthogonality; i.e., that the only important difference between two vectors lies in the difference of *corresponding* values. In the case of the spectrogram, the distance asks the question "do I need more of frequency x, or less?". But these differences contain no explicit spectral orientation—in other words, they have no sense of their spectral neighbors, and by consequence can't answer a question that is simple for humans: "is pitch x higher or lower than pitch y?".

MSS appears to mitigate this problem by taking advantage of the spectral smearing of smaller window sizes (see Figure 1). However, its analytic gradient remains unimpressive in the pitch condition, and nearly random in the level condition.

Many of the analytic gradients perform poorly. Even the 77% accurate MSS gradient is not smooth (Figure 4b). We do not completely understand the nature of this phenomenon, however in our investigations we have found a few patterns: unwindowed STFT distances result in a much bumpier gradient, and Hann windowing smooths the search space. The use of multiple analysis scales further smooths the search space, however these distances still contain many very fine-grained local minima.

Finally, pitch-recovery is an intrinsically difficult regression problem. Sinuisoidal activations can compactly construct functions with infinite VC-dimension [22], and backprop through a periodic function is difficult to train.

6 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 other machine listening communities to include broad-scale pitch perception in their training regimes?

Task-specific approaches can produce one-trick ponies with unexpected failure modes. Sturm [61] describes Clever Hans, a horse that could ostensibly do basic arithmetic, fooling even a zoologist. In fact, Clever Hans had no mathematical understanding at all, and instead relied upon subconscious visual cues from his interrogator. Bello [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 phenomenon is similar to certain patch-based face recognition approaches, where a face is recognized when presented with an adversarial example where the eyes and mouth are switched. Even a few pixel change can trick a neural network [62, 63].

There is a growing understanding that uncovering unusual failure modes is not butterfly hunting, but indeed is deeply connected with generalization. Adversarial examples occur in nature [38]. Gilmer et al. [25] demonstrate a tight correlation between vision performance on a toy adversarial sphere dataset with generalization on natural tasks. Spurious correlation [48] and under-specification [11] also lead to poor generalization.

Evaluation should not only quantify performance on a specific task, but aid the improvement of systems in general [61]. Training on broad open-domain data, like the recently released FSD50K corpus [18], exposes the model to the full gamut of natural phenomena. Sufficiently diverse synthesized data might encompass real-world data as a special case [64]. 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 [9, 56].

Incorporating more training data may lead to a better result, but there might be a ceiling to what data alone can bring to the table [4]. 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 gymnastics [14]; 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 an experiential notion of sound, i.e., one that is perceptually, and perhaps even cognitively, based. By extension, this implies incorporating more modalities than just hearing [10].

More controversial still 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 [4].

7 Conclusion

Learning is only as good as the underlying optimization criterion. Our synthetic benchmark exposes a surprising gap between automatic audio-to-audio distances and human perception. In contrast to the human experience of sound, spectrally-based audio representations—as well as several off-the-shelf neural representations—have difficulties with basic pitch orientation. When signal level is included as a dimension of variability, gradient orientation accuracy decreases. The confounding effects of level are perplexing, and further work is necessary to identify the cause.

Our negative results demonstrate that even for a low-dimensional problem (i.e., pitch and level), distances based upon these audio representations are perceptually unintuitive and bad for learning. Their search spaces have local minima at fine and coarse resolutions. These issues are surely compounded in higher-dimensional problems, and raise the spectre that they cannot model as-yet-unknown underlying factors of variation. The question remains: what is necessary to induce general-purpose audio representations similar to human auditory perception?

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.

Acknowledgments and Disclosure of Funding

The authors would like to thank the following readers for their feedback: Alex Nisnevich, Alexandre Passos, Edouard Oyallon, Felipe Tobar, Harri Taylor, Jesse Engel, Lamtharn (Hanoi) Hantrakul, Nicolas Pinto, Oleg Polosin, Pranay Manocha, Richard Socher, Richard Zhang, as well as the workshop reviewers and chairs.

References

[1] Sameer Agarwal, Josh Wills, Lawrence Cayton, Gert R. G. Lanckriet, David J. Kriegman, and Serge J. Belongie. Generalized non-metric multidimensional scaling. In Marina Meila and Xiaotong Shen, editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, AISTATS 2007, San Juan, Puerto Rico, March 21-24, 2007*, volume 2 of *JMLR Proceedings*, pages 11–18. JMLR.org, 2007.

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

[5] Adrien Bitton, Philippe Esling, and Tatsuya Harada. Neural granular sound synthesis. *CoRR*, abs/2008.01393, 2020.

[6] Michael J. Carey, Eluned S. Parris, and Harvey Lloyd-Thomas. A comparison of features for speech, music discrimination. In *Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '99, Phoenix, Arizona, USA, March 15-19, 1999*, pages 149–152. IEEE Computer Society, 1999.

[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] Jan Chorowski, Ron Weiss, Samy Bengio, and Aäron van den Oord. Unsupervised speech representation learning using wavenet autoencoders. *IEEE Transactions on Audio, Speech, and Language Processing*, 2019.

[9] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, 2011.

[10] 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.

[11] Alexander D'Amour, Katherine A. Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yi-An Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *CoRR*, abs/2011.03395, 2020.

[12] 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.

[13] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *CoRR*, abs/2005.00341, 2020.

[14] Yanai Elazar, Abhijit Mahabal, Deepak Ramachandran, Tania Bedrax-Weiss, and Dan Roth. How large are lions? inducing distributions over quantitative attributes. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3973–3983. Association for Computational Linguistics, 2019.

[15] Jesse Engel, Rigel Swavely, Lamtharn Hanoi Hantrakul, Adam Roberts, and Curtis Hawthorne. Self-supervised pitch detection by inverse audio synthesis. In *ICML Self-supervised learning in Audio and Speech Workshop*, 2020.

[16] Jesse H. Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. DDSP: differentiable digital signal processing. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[17] Jesse H. Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1068–1077. PMLR, 2017.

[18] 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.

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

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

[21] Cristina Gârbacea, Aäron van den Oord, Yazhe Li, Felicia S. C. Lim, Alejandro Luebs, Oriol Vinyals, and Thomas C. Walters. Low bit-rate speech coding with VQ-VAE and a wavenet decoder. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 735–739. IEEE, 2019.

[22] R. J. Gaynier and Tom Downs. Sinusoidal and monotonic transfer functions: Implications for VC dimension. *Neural Networks*, 8(6):901–904, 1995.

[23] François G. Germain, Qifeng Chen, and Vladlen Koltun. Speech denoising with deep feature losses. In Gernot Kubin and Zdravko Kacic, editors, *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 2723–2727. ISCA, 2019.

[24] Beat Gfeller, Christian Havnø Frank, Dominik Roblek, Matthew Sharifi, Marco Tagliasacchi, and Mihajlo Velimirovic. SPICE: self-supervised pitch estimation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:1118–1128, 2020.

[25] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S. Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian J. Goodfellow. Adversarial spheres. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net, 2018.

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

[27] Lamtharn Hantrakul, Jesse Engel, Adam Roberts, and Chenjie Gu. Fast and flexible neural audio synthesis. *Ismir 2019*, pages 524–530, 2019.

[28] F.J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978.

[29] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5:2154, 06 2020.

[30] 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 W. Wilson. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 131–135. IEEE, 2017.

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

[32] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

[33] 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.

[34] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aäron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2415–2424. PMLR, 2018.

[35] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. CREPE: A convolutional representation for pitch estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 161–165. IEEE, 2018.

[36] 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.

[37] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C. Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 14881–14892, 2019.

[38] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *ICLR Workshop*, 2017.

[39] Ricardo Bigolin Lanfredi, Joyce D. Schroeder, and Tolga Tasdizen. Quantifying the preferential direction of the model gradient in adversarial training with projected gradient descent. *CoRR*, abs/2009.04709, 2020.

[40] Yann LeCun. The future is semi-supervised. `https://iclr.cc/virtual_2020/speaker_7.html`. ICLR Turing Award talk.

[41] Vincent Lostanlen, Sripathi Sridhar, Brian McFee, Andrew Farnsworth, and Juan Pablo Bello. Learning the helix topology of musical pitch. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 11–15. IEEE, 2020.

[42] 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. In B. Yegnanarayana, editor, *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 37–41. ISCA, 2018.

[43] 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. In Helen Meng, Bo Xu, and Thomas Fang Zheng, editors, *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 2852–2856. ISCA, 2020.

[44] 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.

[45] 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.

[46] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron C. Courville, and Yoshua Bengio. SampleRNN: An unconditional end-to-end neural audio generation model. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

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

[48] Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *CoRR*, abs/2010.15775, 2020.

[49] A. Michael Noll. Cepstrum pitch determination. *The Journal of the Acoustical Society of America*, 41(2):293–309, 1967.

[50] Hua Ouyang and Alexander G. Gray. Learning dissimilarities by ranking: from SDP to QP. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 728–735. ACM, 2008.

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

[52] 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. In Helen Meng, Bo Xu, and Thomas Fang Zheng, editors, *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 2637–2641. ISCA, 2020.

[53] 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.

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

[55] Wei Ping, Kainan Peng, and Jitong Chen. Clarinet: Parallel wave generation in end-to-end text-to-speech. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

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

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

[58] 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.

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

[60] 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.

[61] 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.

[62] J Su, D V Vargas, and K Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, October 2019.

[63] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[64] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2017, Vancouver, BC, Canada, September 24-28, 2017*, pages 23–30. IEEE, 2017.

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

[66] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. In *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016*, page 125. ISCA, 2016.

[67] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 6306–6315. Curran Associates, Inc., 2017.

[68] Xin Wang, Shinji Takaki, and Junichi Yamagishi. Neural source-filter waveform models for statistical parametric speech synthesis. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:402–415, 2020.

[69] Yuxuan Wang, R J Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, and Quoc Le. Tacotron: Towards end-to-end speech synthesis. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017-Augus:4006–4010, 2017.

[70] R Yamamoto, E Song, and J Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with Multi-Resolution spectrogram. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6199–6203, May 2020.

[71] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Probability density distillation with generative adversarial networks for high-quality parallel waveform generation. In Gernot Kubin and Zdravko Kacic, editors, *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 699–703. ISCA, 2019.

[72] Geng Yang, Shan Yang, Kai Liu, Peng Fang, Wei Chen, and Lei Xie. Multi-band MelGAN: Faster waveform generation for high-quality text-to-speech. *CoRR*, abs/2005.05106, 2020.

[73] 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.

[74] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595. IEEE Computer Society, 2018.

[75] 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. In *International Conference on Learning Representations (ICLR)*, 2020.

[76] Yi Zhao, Haoyu Li, Cheng-I Lai, Jennifer Williams, Erica Cooper, and Junichi Yamagishi. Improved prosody from learned F0 codebook representations for VQ-VAE speech waveform reconstruction. In Helen Meng, Bo Xu, and Thomas Fang Zheng, editors, *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 4417–4421. ISCA, 2020.

A Supplementary Material

A.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, the ratio of 660:440Hz is equivalent to the ratio 750:500Hz, and therefore their pitch distances are the same. However, this rule 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 [51, 60].
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 [41, 57].

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. Under this assumption, the ideal perceived pitch distance is strictly monotonic for fixed level [60]:

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

$$d(x(A,\omega), x(A',\omega')) < d(x(A,\omega), x(A'',\omega')) \text{ if } A < A' < A'' \text{ or } A > A' > A'' \tag{8}$$

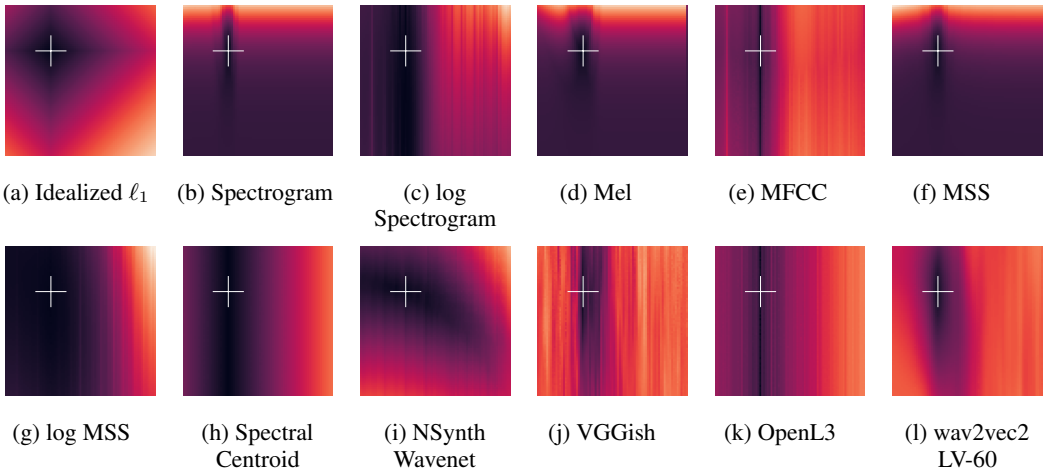
which lead to Equations 5 and 6.

A.2 Hyperparameters

All spectral distances are implemented in PyTorch using a sampling rate of 44.1KHz and Hann windowing. Other windowing methods available in `scipy` performed equivalently or worse. Although hyperparameter selection on evaluation data is against the spirit of zero-shot learning, we tuned hyperparameters to concisely demonstrate upper-bounds on performance. These are the hyperparameters used in our results section:

Spectrogram	ℓ_1 distance, $\lvert$Spectrogram$(x)\rvert$, nfft=2048, frame_overlap=0.75
log(Spectrogram)	ℓ_2 distance, $\log(\lvert$Spectrogram$(x)\rvert+$1e-4$)$, nfft=2048, frame_overlap=0.75
Mel	ℓ_1 distance, nfft=1024, frame_overlap=0.5, nmels=1024, fmin=30, fmax=4000
MFCC	ℓ_1 distance, nfft=1024, frame_overlap=0.5, nmels=128, nmfcc=128, norm=None, fmin=30, fmax=4000
MSSTFT and log MSSTFT	Computed as Spectrogram and log(Spectrogram) respectively, with nffts=[2048, 1024, 512, 256, 128, 64]
Spectral Centroid	$\ell 1$ distance, $\log_2$ spectral centroid, with nfft=2048, frame_overlap=0.75, power=1.0

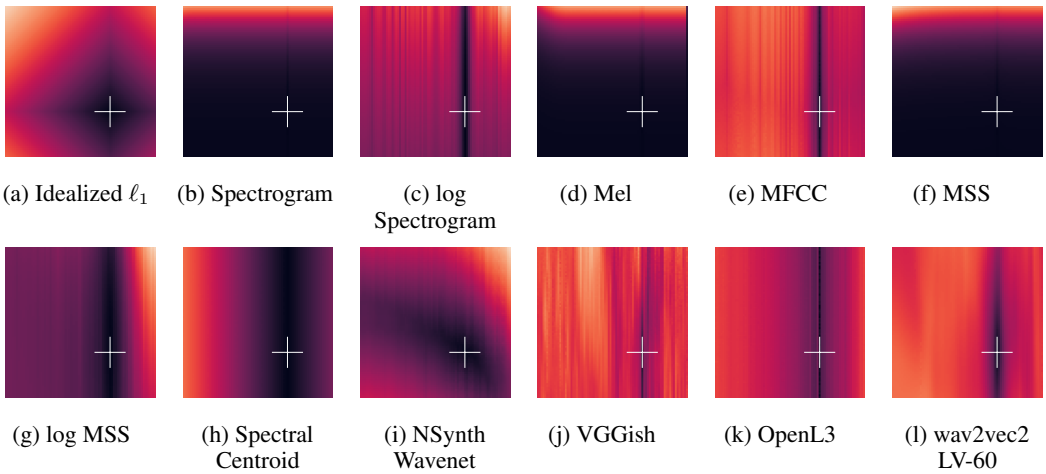
Table 2: Hyperparameters used in spectral distances.



(a) Idealized ℓ_1 (b) Spectrogram (c) log Spectrogram (d) Mel (e) MFCC (f) MSS

(g) log MSS (h) Spectral Centroid (i) NSynth Wavenet (j) VGGish (k) OpenL3 (l) wav2vec2 LV-60

Figure 5: As Figure 3, but with target pitch and level fixed at 130 Hz and -7.5dB, respectively.



(a) Idealized ℓ_1 (b) Spectrogram (c) log Spectrogram (d) Mel (e) MFCC (f) MSS

(g) log MSS (h) Spectral Centroid (i) NSynth Wavenet (j) VGGish (k) OpenL3 (l) wav2vec2 LV-60

Figure 6: As Figure 3, but with target pitch and level fixed at 922 Hz and -17.5dB, respectively.