From HEAR to GEAR: Generative Evaluation of Audio Representations

Vincent Lostanlen, Lingyao Yan, Xianyi Yang

From HEAR to GEAR: Generative Evaluation of Audio Representations

Vincent Lostanlen VINCENT.LOSTANLEN@LS2N.FR
Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

Lingyao Yan* YANLINGYAO@BUAA.EDU.CN
Xianyi Yang* XIANYI@BUAA.EDU.CN
Sino-French Engineer School, Beihang University, Beijing 100191, China

Editors: Joseph Turian, Björn W. Schuller, Dorien Herremans, Katrin Kirchhoff, Paola Garcia Perera, and Philippe Esling

Abstract

The "Holistic Evaluation of Audio Representations" (HEAR) is an emerging research program towards statistical models that can transfer to diverse machine listening tasks. The originality of HEAR is to conduct a fair, "apples-to-apples" comparison of many deep learning models over many datasets, resulting in multitask evaluation metrics that are readily interpretable by practitioners. On the flip side, this comparison incurs a neural architecture search: as such, it is not directly interpretable in terms of audio signal processing. In this paper, we propose a complementary viewpoint on the HEAR benchmark, which we name GEAR: Generative Evaluation of Audio Representations. The key idea behind GEAR is to generate a dataset of sounds with few independent factors of variability, analyze it with HEAR embeddings, and visualize it with an unsupervised manifold learning algorithm. Visual inspection reveals stark contrasts in the global structure of the nearest-neighbor graphs associated to logmelspec, Open-L^3, BYOL, CREPE, wav2vec2, GURA, and YAMNet. Although GEAR currently lacks mathematical refinement, we intend it as a proof of concept to show the potential of parametric audio synthesis in general-purpose machine listening research.

1. Introduction

1.1. Towards machine listening

Machine listening (Rowe, 1992), also known as audio content analysis (Lerch, 2012), aims to extract the information from a digital audio signal in the same way as a human listener would. Since the 1950s and the earliest spoken digit recognizer (Davis et al., 1952), this technology has gained in sophistication, thanks to a number of factors in conjunction: the falling costs of audio acquisition hardware, the acceleration of personal computing, and the massification of user-generated content (Gemmeke et al., 2017), just to name a few. Over the past decade, the renewed interest for deep learning has spurred the development of a new generation of machine listening systems. These systems tend to have certain

* Work done during exchange with Centrale Nantes.
. Download the source code for our experiments: https://github.com/yy945635407/GEAR

traits in common: the resort to a mel-frequency or CQT representation as input, the "deep" stacking of convolutional or recurrent layers, and training from a random initialization by some variant of stochastic gradient descent (McFee, 2018). Yet, the debate surrounding the best computational architecture for machine listening remains intense: for example, recent research has shown that deep learning in the "raw waveform" domain may outperform predefined time–frequency representations (Zeghidour, 2019). The same can be said of the "Transformer," a feedforward layer with multiplicative interactions which may outperform the well-established convolutional layer (Gong et al., 2021). Lastly, many variants of pre-training, either supervised or self-supervised, have shown a strong potential (Tagliasacchi et al., 2020).

1.2. On the proliferation of tasks

Another strong tendency of the past decade in machine listening resides in the rapid diversification of its application scope. Once centered on English speech (e.g., TIMIT dataset), deep learning for machine listening has progressively taken on a broad array of other domains, such as music (Peeters and Richard, 2021), urban sounds (Bello et al., 2019), animal vocalizations (Stowell, 2022), machine sounds (Koizumi et al., 2018), and healthcare (Deshpande et al., 2022). Therefore, applied research in machine listening now benefits practitioners in many other scientific fields, from conservation ecology to digital humanities. However, at the level of fundamental research, it is difficult to assess whether the overarching quest towards human-level machine listening is making much progress, if at all (Kim et al., 2020). Indeed, at present, most of the models which are being actively maintained are trained and evaluated on a single "niche" task.

1.3. The HEAR benchmark

In this context, the "Holistic Evaluation of Audio Representations" (HEAR) initiative aims to offer a level playing field for fundamental machine listening research (Turian et al., 2022). The key idea behind HEAR is that the human auditory system is holistic (general-purpose), in the sense that it can readily learn to perform new tasks with little or no supervision. From this observation, it follows that machine listening research should be evaluated "holistically;" i.e., over as many tasks as possible. Hence, HEAR consists of a benchmark of 19 different tasks, encompassing speech, environmental sound, and music. Participants to the HEAR benchmark are not expected to solve the tasks one by one; but rather, to provide a general-purpose feature map, or *representation*. For each audio input, this representation returns a vector, known as *embedding*, which then serves to train a shallow neural network on the task of interest. We refer to the official website of HEAR for a more detailed description[1].

HEAR proposes to conduct "holistic evaluation" by juxtaposing many small niche tasks in supervised machine listening, known to have various levels of difficulty: from 10–20% mean average precision (mAP) for the recognition of vocal imitations up to 90–95% accuracy for the recognition of Beijing Opera percussion instruments. The main appeal behind such a juxtaposition is that the evaluation scores reflect a real-world purpose: for example, the two tasks above stem from content-based information retrieval and digital musicology, respectively.

1. Official website: https://hearbenchmark.com

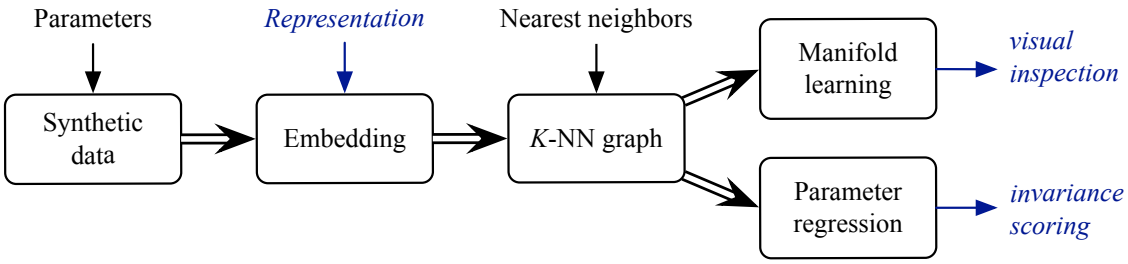

Figure 1: Flowchart of GEAR: Generative Evaluation of Audio Representations. Elements in blue depend upon the choice of audio representation, whereas elements in black are shared.

1.4. Limitations of HEAR

On the flip side, we point out that HEAR is hampered by its limited capabilities for error analysis at the task-agnostic level. The conclusion of HEAR 2021 takes the form of a matrix with 29 rows (one per submission) and 19 columns (one per task), yielding $29 \times 19 = 551$ entries in total (Turian et al., 2022, Fig. 1, p. 10). Although the benchmark organizers have derived some summary visualizations from this table (see t-SNE plots in Fig. 2, *op. cit.*), it remains difficult for the reader to figure out, at a glance, what a particular HEAR submission can and cannot do.

Of course, it would be possible to increase the level of detail and break down accuracy numbers not only per task, but also per class. However, such a procedure might not be conclusive. Specifically, if model M tends to confuse X and Y on dataset D, it is uncertain if such confusion indicates that M is invariant to an audio feature which separates X from Y; that intra-class and inter-class variabilities are strongly correlated; that M is oversensitive to small audio perturbations in D; or some combination of the above.

Another drawback of HEAR is that submissions are not compared directly; but instead, through the lens of a specific deep learning architecture, that is chosen by the organizers. Here, Occam's razor prevails: the ideal audio representation should separate classes linearly and thus obtain an excellent score, even when paired with a shallow model. We understand that such is the reason why the organizers opted for a hyperparameter selection by grid search over a fairly simple design space: one or two hidden layers, four values of initial learning rate, and two kinds of initial weight distributions (see Table 4, *op. cit*).

Yet, for all we know, expanding this grid search to deeper or wider architectures might boost classification accuracy and even rearrange the leaderboard. As a consequence, when consulting the scores of a HEAR submission, it is not clear how much of this score should be attributed to properties of the underlying representation versus to the fitness of that representation with respect to the hyperparameter grid.

1.5. Contribution

In the present paper, we propose a first attempt at remediating the two aforementioned shortcomings, which we perceive in the HEAR benchmark: lack of visual interpretation and dependency on a hyperparameter grid. To do so, we generate a synthetic dataset with known independent factors of variability. Then, we run each audio sample in the dataset

through the HEAR model under study, which we regard as a feature extractor. Thirdly, we visualize the nearest-neighbor graph between feature vectors associated to the dataset, by means of an unsupervised manifold learning algorithm. Lastly, we color the vertices of this graph according to the values of the latent variables governing the known factors of variability in the dataset. In this way, we hope to shed light on the model's ability to discover these factors of variability without any supervision.

Figure 1 illustrates our protocol. As a nod to HEAR, we call our method "Generative Evaluation of Audio Representations," or GEAR for short.

With GEAR, we do not have the ambition to compete with HEAR, nor to present an alternative leaderboard. On the contrary, we intend to provide a complementary viewpoint on the same benchmark: the strengths of GEAR are the weaknesses of HEAR and vice versa. Indeed, the main drawback of GEAR is that it is not grounded in any real-world use case or "task:" disentangling factors of variability is a necessary first step for high-dimensional representation learning (Bengio, 2013) but rarely suffices on its own to correctly assign patterns to classes. Meanwhile, GEAR operates as a visual "smoke test" which looks for some of the most basic attributes of auditory perception. In this paper, we have experimented with pitched sounds and varied three important parameters of the spectral envelope; but we stress that GEAR is a general methodology and could, in the future, apply to more sophisticated generative models than the one we present.

Our results are presented in Section 4. They reveal that the nearest-neighbor graphs which proceed from seven of the embeddings in HEAR exhibit qualitatively different topologies: some of them appear like 1-D strands; others a like 2-D sheaf; others like a 3-D dense volume Nevertheless, we acknowledge that the interpretation of point clouds in Figure 3 is hampered by the lack of shortcomings in the design of our synthetic dataset, which only became evident once the study was complete. In particular, distances in the space of synthesis parameters do not necessarily correlate with perceptual judgments of auditory similarity. Thus, although the current formulation of GEAR may serve to check local properties of continuity and independence between factors of variability, one should not make conclusions about the global geometric properties of audio representations from GEAR visualization alone. Still, we believe that finely manipulating audio data via parametric synthesizers has a strong potential towards the better interpretability of deep audio representations. Section 4.3 summarizes the known limitations of GEAR, offers some "learned lessons" after running it on HEAR challenge submissions, and offers some perspectives for future work.

1.6. Related work

The topics of "interpretability" and "explainability" are under growing attention in machine listening research, as in other subdomains of artificial intelligence. A forerunner in these topics, Sturm (2014) has proposed a witty analogy between the behavior of some high-accuracy systems for music information retrieval and that of "Clever Hans;" that is, a horse who seemingly had the ability to understand mathematics, yet was in fact responding to unintentional cues of the questioner. Indeed, these systems were only achieving a good performance by applying "tricks" such as detecting irrelevant factors of variability which happen to be confounded with the relevant ones on the dataset on which they were being evaluated. Such a concern for assessing whether machines encode audio similarity in a

human-like way is reminiscent of the philosophical and methodological framework of Wiggins (2009).

Since then, various methods have been proposed to assess whether a given machine listening system is a "horse"—and if so, because of what unanticipated "trick." Kereliuk et al. (2015) have proposed to approach the problem from the perspective of adversarial attacks in deep learning. More recently, Rodríguez-Algarra et al. (2019) have designed a series of intervention for a music classification pipeline, both at the level of audio content and that of class distribution, so as to characterize confounding effects. Kim et al. (2019) have proposed to evaluate the "trustworthiness" of various audio representations by verifying whether artificial perturbations in pitch or tempo yield consistent displacements in feature space across all examples of a real-world music collection. Lastly, Melchiorre et al. (2021) have paired a music recommendation system with a "listenable explanation;" i.e., a user interface revealing which parts and which instruments of a given song are predicted as most characteristic of their taste.

Another important inspiration for GEAR is the recent work of Turian and Henry (2020), who proposed to evaluate whether low-level representations of audio (some engineered, some learned) were capable of predicting which of two sine waves is higher in fundamental frequency. We concur with the idea that parametric audio synthesis may provide insight on the functioning of state-of-the-art audio representations.

Our paper extends the study of Lostanlen et al. (2020), which evaluated the ability of the scattering transform to replicate psychoacoustic masking. The main novelties of our paper reside in its systematic application to seven learned audio representations and in the quantitative measurement of invariance.

1.7. Outline

Section 2 presents our parametric generative model; our synthetic audio dataset; and our chosen graph-based algorithm for manifold learning, Isomap. Section 3 presents the application of GEAR to the HEAR baseline, as well as six open-source audio representations. Section 4 summarizes our findings in both qualitative and quantitative terms.

2. Methods

2.1. Additive Fourier synthesis

We build a dataset of complex tones according to the following additive synthesis model:

$$\boldsymbol{y}_{\boldsymbol{\theta}}(t) = \sum_{p=1}^{P} \frac{1 + (-1)^p r}{p^{\alpha}} \cos(p f_1 t) \boldsymbol{\phi}_T(t), \tag{1}$$

where $\boldsymbol{\phi}_T$ is a Hann window of duration T. This additive synthesis model depends upon three parameters: the fundamental frequency f_1, the Fourier decay α, and the relative odd-to-even amplitude difference r. We denote the triplet (f_1, α, r) by $\boldsymbol{\theta}$. The Fourier decay affects the perceived brightness of the of sound, while the relative odd-to-even amplitude difference is linked to the boundary conditions of the underlying wave equation: a value of $r = 1$ suggests a semi-open 1-D resonator, such as a clarinet, whereas a value of $r = 0$

suggests a closed resonator such as a flute. Hence, our audio signal has three degrees of freedom.

The reason why we focus on sustained pitched sounds for this study is that they are present in music, speech and some bio-acoustic sounds. Besides, note that one-dimensional audio descriptors such as spectral brightness and zero-crossing rate correlate with both f_1 and α, hence a lack of disentanglement. Likewise, a previous publication has shown that mel-frequency cepstral coefficients (MFCC), arguably the most widespread set of engineered features for speech and music processing, is incapable of disentangling f_1 from α and r (Lostanlen et al., 2020). Thus, simulating the ability of our auditory system to represent stimuli in Equation 1 is more challenging that it might seem at first glance.

2.2. Synthetic dataset

We generate $N = 2500$ signals in total, corresponding to 50 values of α between 0 and 2, 50 values of r between 0 and 1, and f_1 being an integer chosen randomly between 12 and 24. In practice, we set T to 1024 samples and P to 32 harmonics. Figure 2 shows a small subset of our synthetic dataset. In this plot, r varies horizontally, α varies vertically, and f_1 takes random values. As we can see, in time domain graph, with the increasing of r, the signal is sparser because the difference between the values in odd-even order is larger. As for α, when it increases, the signal is smoother due to the fact that α inhibits high-frequency components.

Note that human judgments of perceived dissimilarity grow monotonically with synthesis parameters (f_1, α, and r), but not linearly. In particular, it is known in psychoacoustics that perceived differences in pitch are roughly proportional to differences in $\log(1 + cf_1)$ for some constant c. Hence, there is no reason to expect that a dataset of sounds in which f_1 grow according to an arithmetic progression should map to evenly spaced points in feature space. The same can be said of parameters α and r, which are mathematically simple but are not calibrated according to a perceptually uniform psychoacoustic scale.

This lack of calibration in the design of the synthetic dataset has implications for the GEAR methodology. It implies that pairwise comparisons in feature space are only "sensible" if conducted between sounds whose variations in parameter space are small enough as to be considered infinitesimal for the listener. Since deep neural networks and audio synthesizers are both differentiable functions of their input, we may expect that these variations should map to infinitesimal variations in feature space. For this reason, we compute distances in feature space upon small Euclidean neighborhoods and rely on the Isomap data visualization algorithm to visualize the global nonlinear structure of the dataset. Section 4.3 provides additional perspectives on the problem of perceptual calibration of audio synthesizers and the geometric analysis of audio representations beyond simple nearest-neighbor graphs.

2.3. Isomap dimensionality reduction

Isomap is an unsupervised algorithm for data visualization (Tenenbaum et al., 2000). It operates in three stages: nearest-neighbor search, computation of shortest path distances, and multidimensional scaling. For the first stage, we search nearest neighbors in terms of smallest Euclidean distance in feature space. We arbitrarily set the number of neighbors

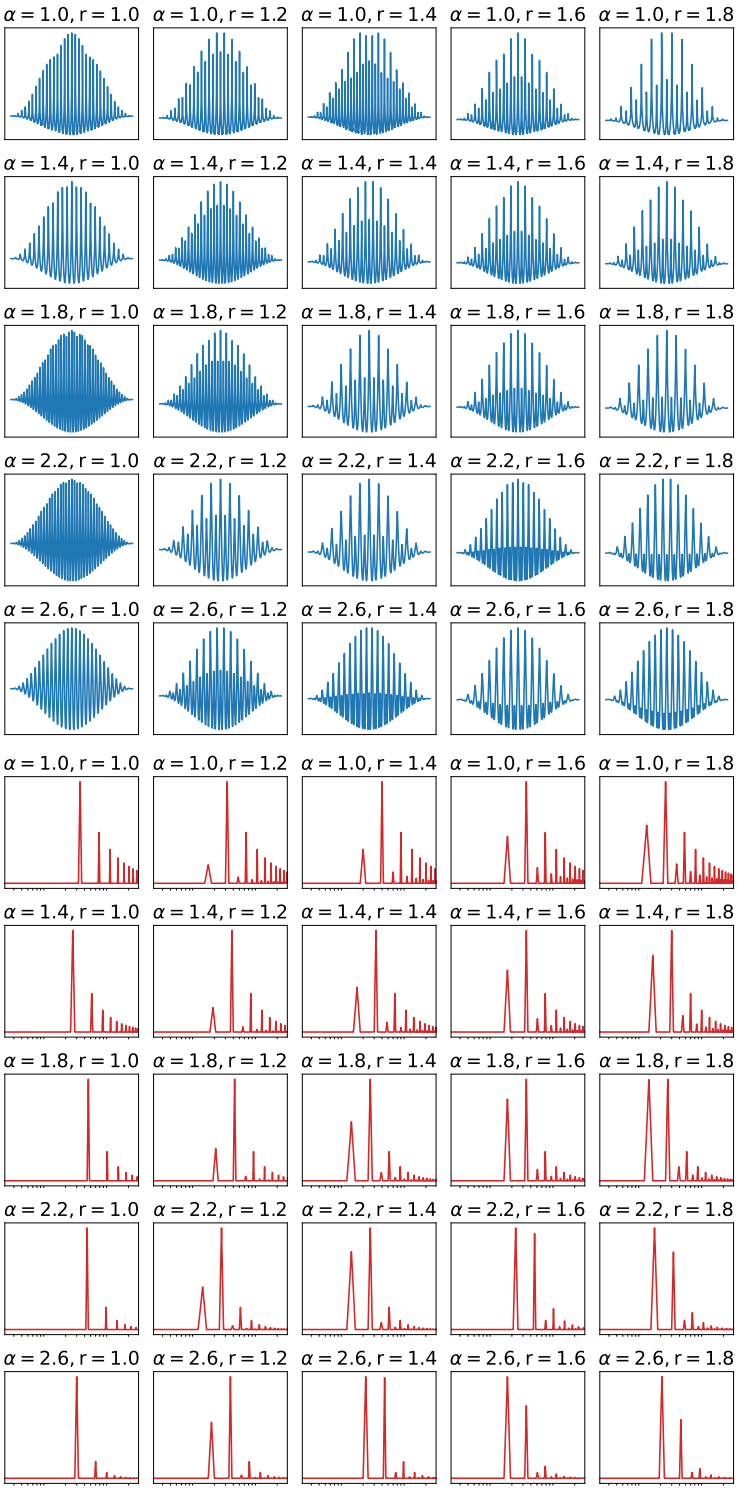

Figure 2: Visualization of 25 samples in our synthetic dataset: r varies across columns, α across rows, and f_1 varies randomly. Blue: time domain. Red: Fourier domain.

to $K = 100$. That being said, we note that the hyperparameter K has an impact on the visualization and its role in GEAR is deserving of future work.

The second stage of Isomap is classically done by a shortest path algorithm; in our case, Dijkstra's algorithm. It yields a large matrix ($N \times N$) of geodesic distances, which we square and recenter to null mean.

Lastly, multidimensional scaling (MDS) determines the top-three eigenvalues of this matrix above and its associated eigenvectors in a space of dimension N. Displaying the entries of these eigenvectors as a scatter plot produces a 3-D cloud of N points, each corresponding to a different sound, which lends itself to visual interpretation: any two points appearing nearby are similar in feature space, in the sense that there exists a short Euclidean path connecting them. Furthermore, looking up the value of α, r, or f_1 for these points assigns them a color in a continuous scale: in our case, red–white–blue.

Hence, the core hypothesis of GEAR is that a desirable audio representation should produce a dense arrangement of points in 3-D, in which all three parameters of interest appear as smooth color progressions over orthogonal coordinates. This hypothesis may be simply checked by visual inspection, or quantified automatically by defining a task of parameter regression (see Figure 1).

3. Application to HEAR embeddings

3.1. Logmelspec

The first model we choose is the "naive" baseline, provided by the organizers of HEAR. This baseline is a log-scaled mel-frequency spectrogram (logmelspec) followed by 4096 random projections, in which the random matrix weights are Gaussian and independent.

3.2. Open-L^3

Open-L^3 is a deep convolutional network (convnet) that is trained entirely by self-supervision (Cramer et al., 2019). In Open-L^3, the "open" prefix stands for open-source while the suffix L^3 is short for "Look, Listen and Learn" (Arandjelovic and Zisserman, 2017). Open-L^3 consists of two subnetworks: a video subnetwork and an audio subnetwork. The two subnetworks are trained jointly to distinguish whether a video frame and a one-second audio segment are from the same video file; a task known as audio-visual correspondence. These files are sampled from a large unlabeled dataset of 60 million videos. The audio subnetwork has reached state-of-the-art results in urban sound classification. We use this subnetwork as a feature extractor in dimension 6144. Open-L^3 is one of the three (non-naive) baselines that are provided by the organizers of the HEAR benchmark.

3.3. Hybrid BYOL-S

BYOL (Bootstrap Your Own Latent) is a self-supervised learning algorithm, initially proposed for computer vision (Grill et al., 2020), and then adapted to machine listening by Niizumi et al. (2021), under the name BYOL-A (BYOL for Audio). The motivation behind BYOL is to perform contrastive learning while circumventing the problem of mining negative samples. While is trained on AudioSet (Gemmeke et al., 2017), a HEAR submission has proposed to train it on a speech-only subset of AudioSet (Elbanna et al., 2022a,b): hence, BYOL-S

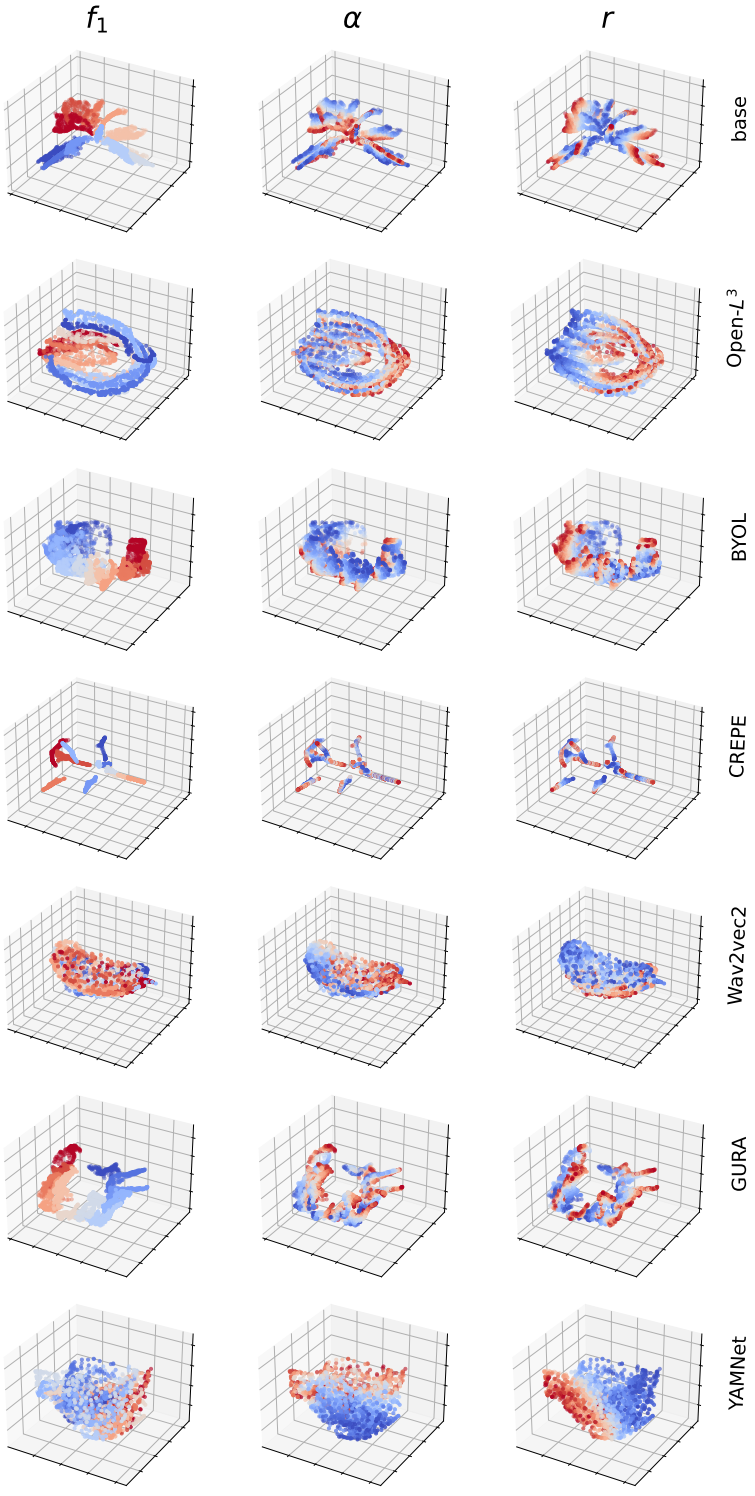

Figure 3: Isomap visualization of our synthetic dataset after feature map by seven audio representations. Top to bottom: HEAR baseline, Open-L^3, BYOL, CREPE, wav2vec2, GURA, YAMNet. Shades of red (resp. blue) denote greater (resp. lower) values of the fundamental frequency f_1 (left), the spectral decay exponent α (center), and the odd-to-even harmonic energy ratio r (right).

(BYOL for Speech). Eventually, the authors have extended BYOL-S so as to learn from handcrafted features, yielding Hybrid BYOL-S. Hybrid BYOL-S was found to outperform BYOL-S in the HEAR 2021 benchmark, which is why we prioritized it for inclusion in GEAR.

3.4. CREPE

CREPE means "Convolutional Representation for Pitch Estimation" (Kim et al., 2018): it was initially proposed to solve the problem of monophonic pitch tracking. It is a deep convolutional neural network architecture whose originality is to operate directly on the "raw waveform domain" rather than upon a time–frequency representation. After supervised learning on synthetic data, CREPE outperforms other popular models for pitch estimation on two real-world datasets. CREPE is made available as a (non-naive) baseline by the organizers of HEAR.

3.5. Wav2vec2

Wav2vec2 (Baevski et al., 2020) is a self-supervised learning model built based on a contrastive learning task. It is also one of the baselines in HEAR competition. Its feature encoder consists of a convnet in the raw waveform domain, followed by a Transformer and a vector quantization module for sequence modeling. Wav2vec2 has proven to extract cross-linguistic speech units. On a downstream task of automatic speech recognition, wav2vec2 matches the previous state of the art model with 100 times fewer labeled samples.

3.6. GURA Fusion

GURA [2] is a set of ensemble methods applied on three models: HuBERT (Hsu et al., 2021), wav2vec2, and CREPE. The authors have considered several aggregation strategies: feature concatenation, averaging, and fusion. For GEAR, we choose to visualize the feature concatenation variant, referred to as "GURA Cat H+w+C" on the leaderboard and "fusion_cat_xwc" in their repository. It concatenates three 1024-dimensional embeddings and produces a 3072-dimensional embedding.

3.7. YAMNet

Developed by Google, YAMNet (Yet another Audio MobileNet) [3] is an instance of MobileNet (Howard et al., 2017): it composes a depthwise convolution and a pointwise convolution, hence a "depthwise separable" variant of the 2-D convnet which significantly reduces the number of parameters. YAMNet is pre-trained on log-mel-spectrograms from AudioSet in a supervised way. It produces a 1024-dimensional embedding.

2. GURA GitHub repository: https://github.com/tony10101105/HEAR-2021-NeurIPS-Challenge---NTU-GURA
3. YAMNet website: https://www.tensorflow.org/hub/tutorials/yamnet

4. Results

4.1. Visual inspection

Figure 3 shows the result of Isomap for all seven embeddings listed in the previous section. Each row corresponds to a different embedding while each columns corresponds to a different synthesis parameter: i.e., f_1, α, r.

We make the following observations:

base The logmelspec baseline is strongly sensitive to pitch (f_1), especially for bright spectra (low α) comprising a sparse harmonic series (high r).

Open-L^3 Discontinuities in the color scale associated to f_1 indicate that the topology of the pitch axis is not preserved. This is consistent with the previous findings of Lostanlen et al. (2020).

BYOL A 2-D manifold, roughly indexed by f_1 and α coordinates. Clarinet-like sounds ($r = 1$) appear as outliers.

CREPE Strongly sensitive to pitch (f_1) and quasi-invariant to spectral envelope (α and r). This is to be expected since CREPE was trained for fundamental frequency estimation.

wav2vec2 A 3-D manifold but whose coordinates do not align with the original degrees of variability of the data.

GURA Similarly to BYOL, a 2-D manifold in which f_1 is the dominant factor of variability.

YAMNet Arguably the best representation in the GEAR benchmark: YAMNet produces a dense 3-D manifold in which the underlying parameters of audio synthesis (f_1, α, and r) appear as smoothly changing over perpendicular directions.

4.2. Nearest-neighbor regression

We complement the qualitative method above by a quantitative method: namely, nearest-neighbor regression. Given a tuple of parameters $\boldsymbol{\theta}_i$, let us denote by $\mathcal{N}_K(\boldsymbol{\theta}_i)$ the set of its K nearest neighbors. We look up the parameters values $\boldsymbol{\theta}_j$ in this set and compute an unweighted mean over each of them, yielding the 3-D vector:

$$\tilde{\boldsymbol{\theta}}_i = \frac{1}{K} \sum_{\boldsymbol{\theta_j} \in \mathcal{N}_K(\boldsymbol{\theta_i})} \boldsymbol{\theta_j}. \tag{2}$$

Our postulate is that the K nearest neighbors $\boldsymbol{\theta}_j \in \mathcal{N}_K(\boldsymbol{\theta}_i)$ of a desirable audio representation should be evenly distributed around the value $\boldsymbol{\theta}_i$, and thus yield a regression estimate $\tilde{\boldsymbol{\theta}}_i$ which is close to $\boldsymbol{\theta}_i$ itself. We measure the logarithm of the element-wise ratio between estimated parameter and true parameter:

$$\log\left(\frac{\tilde{\boldsymbol{\theta}}_i}{\boldsymbol{\theta}_i}\right) = \log \tilde{\boldsymbol{\theta}}_i - \log \boldsymbol{\theta}_i \tag{3}$$

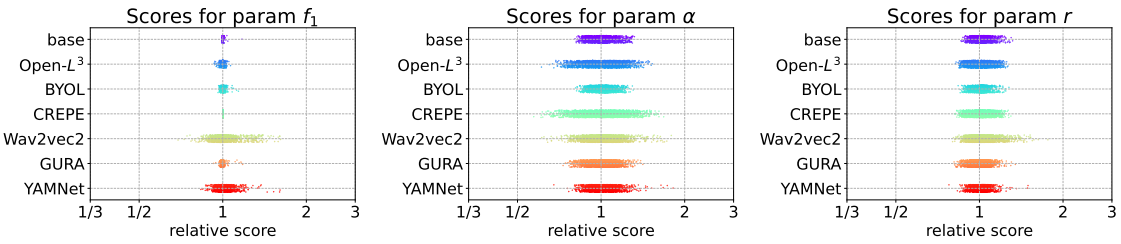

Figure 4: Unsupervised benchmark results of our synthetic dataset after feature map by different embedding models. subplot a corresponds to the unsupervised results of parameter f_1, subplot b corresponds to α, and subplot c corresponds to r. Different colors represent the results of different models. The x-axis is the score on a logarithmic scale while the y-axis lists model names. Scores closer to 1 (vertical center of the plot) are judged to be better.

and repeat the same operation for every synthetic sample i. Intuitively, an ideal audio representation should yield small absolute values for the log-ratio above; i.e., ratios $\tilde{\boldsymbol{\theta}}_i / \boldsymbol{\theta}_i$ that are all close to one.

For the regression task, we lower the value of the number of neighbors: $K = 40$, compared to $K = 100$ in Isomap. Figure 4 displays the resulting distribution of log-ratios for each parameter and each representation in GEAR. Unfortunately, the regression benchmark remains inconclusive: all embeddings fare similarly in terms of estimation error for α and r. The only assured finding is that the naive logmelspec baseline and CREPE are highly sensitive to f_1, which was to be expected. Hence, future work should not evaluate regression error at the local scale of nearest neighbors; but rather, at the global scale of invariance and disentanglement.

4.3. Perspectives

GEAR is an attempt at evaluating audio representations via generative models. In this way, it takes a different approach than HEAR, which is based on real-world classification tasks. Our paper has shown that GEAR is feasible in practice while incurring a moderate workload. Specifically, this paper was achieved by two MSc students, working part time under the supervision of one faculty member. In comparison, data collection and preparation for HEAR required the work of 23 challenge organizers. Hence, we believe that the conceptual simplicity of the GEAR methodology has the potential to expand the accessibility and attractiveness of HEAR to newcomers, particularly at the undergrad and grad student levels.

However, we acknowledge that GEAR currently suffers from a lack of direct applicability to research questions in machine listening. This leads to two avenues of methodological clarification. First, to be meaningful for hearing scientists, parameters in the GEAR synthesizer ought to be sampled according to a perceptually uniform progression. These progressions are often nonlinear in terms of physical units: for example, the human perception of pitch is nonlinear with respect to fundamental frequency in Hertz. In order to account for these distinctions, it would be necessary to include some prior knowledge about auditory perception into the design of the synthesizer. Such prior knowledge could take the form of just-noticeable differences (JND) and could apply to low-level attributes such as pitch, loudness, or roughness. It could also involve relative dissimilarity judgments such as those

involved in the study of qualitative timbre. This former avenue of research connects with ongoing work about the perceptual control of differentiable synthesizers.

A second avenue of research arises from the fact that GEAR is currently tied to Euclidean nearest-neighbor searc in feature space. Meanwhile, the HEAR benchmark is not based on nearest-neighbor classification but on some form of representation learning: namely, a shallow neural network. Even though this shallow neural network is continuous with respect to its input, it stretches distances nonuniformly and non-linearly, thus yielding a higher-level metric space in which comparisons are no longer Euclidean in terms of embeddings. In order to align better with the formulation of HEAR, GEAR should not be restricted to the raw feature space but should also be performed at deeper levels of representation. The shallow neural network could be trained in a supervised way, by performing parameter regression; or in an unsupervised way, e.g., via self-supervised contrastive learning. Following this protocol would incur a stage of neural architecture search, It would certainly be heavier in terms of workload and computation than nearest-neighbor search in feature space; but also more informative and more consistent with real-world audio classification in HEAR.

5. Conclusion

The HEAR benchmark provides a common API for sharing and improving general-purpose machine listening models. In this paper, we have taken this opportunity to download HEAR submissions *en masse* and run them as feature extractors to a synthetic dataset of pitched sounds. In doing so, we have visualized deep audio embeddings as points in a 3-D space, with colors denoting the parameters underlying synthesis. Our contribution, named GEAR, serves as a qualitative counterpoint to HEAR: although it does not fulfil any real-world "task," it sheds light on the respective abilities of audio representations to disentangle auditory attributes, without depending on a choice of supervised learning architecture downstream. The companion website of our paper (see footnote of first page) contains all the necessary source code to reproduce our findings, as well as to replicate them on future editions of HEAR.

One limitation of GEAR in its current formulation is that it largely relies on visual inspection. Our parameter regression benchmarks from Section 4 provide some quantitative evidence for local neighborhoods, but not for global disentanglement. As such, GEAR would not easily scale to hundreds of audio representations, nor to dozens of degrees of freedom in the synthetic data. Furthermore, we note that Isomap may occasionally produce spurious graphical patterns which were not structural properties of the underlying manifold (Donoho and Grimes, 2003), hence deceiving the human eye. In this context, it would be interesting to perform topological data analysis (TDA) on the nearest-neighbor graphs so as to automate the characterization of the feature space (Hensel et al., 2021).

Although we only have experimented with fundamental frequency and two low-level attributes of spectral envelope (α and r, see Section 2), we stress that the GEAR methodology is very generic and could easily be transferred to different synthetic datasets in the future, insofar that the underlying synthesizer is parametric with continuous independent parameters. For example, beyond the case of sustained harmonic tones, one might run GEAR on a physical synthesis model for virtual drum shapes (Han and Lostanlen, 2020); on a neural audio synthesizer with perceptually relevant control (Roche et al., 2021); or on a text-to-

speech rendering engine with global style tokens for expressive conditioning of prosody (Wang et al., 2018).

6. Acknowledgment

We thank Gaëtan Garcia, Mathieu Lagrange, Mira Rizkallah, Joseph Turian, and Cyrus Vahidi for helpful discussions.

References

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

Juan P Bello, Claudio Silva, Oded Nov, R Luke Dubois, Anish Arora, Justin Salamon, Charles Mydlarz, and Harish Doraiswamy. SONYC: A system for monitoring, analyzing, and mitigating urban noise pollution. *Communications of the ACM*, 62(2):68–77, 2019.

Yoshua Bengio. Deep Learning of Representations: Looking Forward. In *Proceedings of the International Conference on Statistical Language and Speech Processing*, pages 1–37. Springer, 2013.

Jason Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3852–3856. IEEE, 2019.

Ken H Davis, R Biddulph, and Stephen Balashek. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 24(6):637–642, 1952.

Gauri Deshpande, Anton Batliner, and Björn W Schuller. Ai-based human audio processing for covid-19: A comprehensive overview. *Pattern recognition*, 122:108289, 2022.

David L Donoho and Carrie Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10): 5591–5596, 2003.

Gasser Elbanna, Alice Biryukov, Neil Scheidwasser-Clow, Lara Orlandic, Pablo Mainar, Mikolaj Kegler, Pierre Beckmann, and Milos Cernak. Hybrid handcrafted and learnable audio representation for analysis of speech under cognitive and physical load. *arXiv preprint arXiv:2203.16637*, 2022a.

Gasser Elbanna, Neil Scheidwasser-Clow, Mikolaj Kegler, Pierre Beckmann, Karl El Hajal, and Milos Cernak. BYOL-S: Learning self-supervised speech representations by bootstrapping. In Joseph Turian, Björn W. Schuller, Dorien Herremans, Katrin Kirchoff, Paola Garcia Perera, and Philippe Esling, editors, *HEAR: Holistic Evaluation of Audio*

Representations (NeurIPS 2021 Competition), volume 166 of *Proceedings of Machine Learning Research*, pages 25–47. PMLR, Dec 2022b.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An Ontology and Human-labeled Dataset for Audio Events. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

Yuan Gong, Yu-An Chung, and James Glass. PSLA: Improving audio tagging with pretraining, sampling, labeling, and aggregation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3292–3306, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Han Han and Vincent Lostanlen. wav2shape: Hearing the shape of a drum machine. In *Proceedings of Forum Acusticum*, pages 647–654, 2020.

Felix Hensel, Michael Moor, and Bastian Rieck. A survey of topological machine learning methods. *Frontiers in Artificial Intelligence*, 4:681108, 2021.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Corey Kereliuk, Bob L Sturm, and Jan Larsen. Deep learning, audio adversaries, and music content analysis. In *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5. IEEE, 2015.

Jaehun Kim, Julián Urbano, Cynthia CS Liem, and Alan Hanjalic. Are nearby neighbors relatives? Testing deep music embeddings. *Frontiers in Applied Mathematics and Statistics*, 5:53, 2019.

Jaehun Kim, Julián Urbano, Cynthia Liem, and Alan Hanjalic. One deep music representation to rule them all? A comparative analysis of different representation learning strategies. *Neural Computing and Applications*, 32(4):1067–1093, 2020.

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)*, pages 161–165. IEEE, 2018.

Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, Yuta Kawachi, and Noboru Harada. Unsupervised detection of anomalous sound based on deep learning and the neyman–pearson lemma. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27 (1):212–224, 2018.

Alexander Lerch. *An introduction to audio content analysis: Applications in signal processing and music informatics*. Wiley-IEEE Press, 2012.

Vincent Lostanlen, Alice Cohen-Hadria, and Juan Pablo Bello. One or Two Frequencies? The Scattering Transform Answers. In *Proceedings of the European Signal Processing Conference*, 2020.

Brian McFee. Statistical methods for scene and event classification. In *Computational Analysis of Sound Scenes and Events*, pages 103–146. Springer, 2018.

Alessandro B Melchiorre, Verena Haunschmid, Markus Schedl, and Gerhard Widmer. Lemons: Listenable explanations for music recommender systems. In *European Conference on Information Retrieval*, pages 531–536. Springer, 2021.

Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. Byol for audio: Self-supervised learning for general-purpose audio representation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

Geoffroy Peeters and Gaël Richard. Deep learning for audio and music. In *Multi-Faceted Deep Learning*, pages 231–266. Springer, 2021.

Fanny Roche, Thomas Hueber, Maëva Garnier, Samuel Limier, and Laurent Girin. Make that sound more metallic: Towards a perceptually relevant control of the timbre of synthesizer sounds using a variational autoencoder. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, 4:52–66, 2021.

Francisco Rodríguez-Algarra, Bob Sturm, and Simon Dixon. Characterising confounding effects in music classification experiments through interventions. *Transactions of the International Society for Music Information Retrieval*, pages 52–66, 2019.

Robert Rowe. *Interactive music systems: machine listening and composing*. MIT press, 1992.

Dan Stowell. Computational bioacoustics with deep learning: A review and roadmap. *PeerJ*, 10:e13152, 2022.

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.

Marco Tagliasacchi, Beat Gfeller, Félix de Chaumont Quitry, and Dominik Roblek. Pre-training audio representations with self-supervision. *IEEE Signal Processing Letters*, 27: 600–604, 2020.

Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Joseph Turian and Max Henry. I'm Sorry For Your Loss: Spectrally-Based Audio Distances Are Bad at Pitch. In *Proceedings of the NeurIPS "I Can't Believe It's Not Better" Workshop*, 2020.

Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W. Schuller, Christian J. Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, Max Henry, Nicolas Pinto, Camille Noufi, Christian Clough, Dorien Herremans, Eduardo Fonseca, Jesse Engel, Justin Salamon, Philippe Esling, Pranay Manocha, Shinji Watanabe, Zeyu Jin, and Yonatan Bisk. HEAR: Holistic Evaluation of Audio Representations. In Douwe Kiela, Marco Ciccone, and Barbara Caputo, editors, *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, volume 176 of *Proceedings of Machine Learning Research*, pages 125–145. PMLR, 06–14 Dec 2022. URL https://proceedings.mlr.press/v176/turian22a.html.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-end Speech Synthesis. In *Proceedings of the International Conference on Machine Learning*, pages 5180–5189. PMLR, 2018.

Geraint A Wiggins. Semantic gap?? Schemantic schmap!! Methodological Considerations in the Scientific Study of Music. In *Proceedings of the IEEE International Symposium on Multimedia*, pages 477–482. IEEE, 2009.

Neil Zeghidour. *Learning representations of speech from the raw waveform*. PhD thesis, PSL Research University, 2019.