AUDIO MODELING THROUGH THE LENS OF EPIPLEXITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Sound contains rich information about the surrounding environment and has become an increasingly important modality for machine learning systems that aim to understand, detect and respond to real-world events. While considerable progress has been made in audio modeling, generalization across domains remains difficult because of limited yet highly diverse training datasets, variation in recording conditions, and other factors. This motivates the search for datasets and processing techniques that support generalization across domains.

We hypothesize that the recently introduced concept of epiplexity, the amount of structural information in a dataset which can be learned by a computationally bounded observer, offers a way to assess how these datasets and processing techniques contribute to model generalization. Prior work on text, images, and video has found that models trained on datasets with higher epiplexity seem to generalize better to unseen downstream tasks (Finzi et al., 2026). This opens the door for understanding how to select datasets that support training more generalizable models. However, the behavior of epiplexity in the audio domain remains unexplored.

We aim to use epiplexity to better understand what kinds of audio datasets can improve generalization in models. We measured the epiplexity of over 40 different audio datasets and the correlations between the epiplexity of 12 training datasets and downstream model performances over three tokenizers: DAC, EnCodec, and X-Codec. We find that while tokenizers historically known to produce better performance yield lower epiplexity, for a fixed tokenizer, epiplexity is usually weakly correlated with improved model performance. We contribute a method for computing epiplexity and conditional epiplexity for audio datasets using PyTorch.

1 INTRODUCTION

Although machine learning research has recently been dominated by text and images, audio is gaining traction but remains underexplored (Liu et al., 2025). Audio is rich in spatial information not captured by visual or textual modalities (Tang et al., 2024; Zhu et al., 2025). As models begin to move increasingly into the physical world, a greater need has emerged to investigate audio data and the generalization of audio models.

Indeed, there are concerns about how well audio models generalize when compared to text models (Team et al., 2025). Different audio datasets can be highly divergent with nuances in microphone configurations, sound sources, and environmental conditions, creating challenges for training audio ML models that generalize across a wide range of domains (Kvsn et al., 2020; Xiao et al., 2025; Leem et al., 2025; Niu et al., 2025; Müller et al., 2024). In contrast, text models show far more astounding generalization capabilities as noted in early work with LLMs and their few or zero-shot capabilities (Brown et al., 2020). However, text models still struggle to perform spatial reasoning, especially when it comes to maintaining global consistency (Guo et al., 2026). Learning how to develop audio models with stronger generalization capabilities could provide a path towards improved spatial reasoning.

Although related work has primarily focused on the impact of data mixing and curation on model generalization (Chen et al., 2026; Aldeneh et al., 2026; Whetten et al., 2026), looking at the data

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

itself could also be fruitful. Epiplexity is the measure of the inherent structure in data which can be learned by a computationally bounded observer (Finzi et al., 2026). Datasets with more complex structure, i.e., greater epiplexity, are hypothesized to help models generalize better by enabling them to learn a richer understanding of the underlying data. Empirical results show that text has greater epiplexity than image or video, and data sampling and transformation techniques which increase epiplexity lead to improved model performance on unseen data (Finzi et al., 2026, Section 6.3). However, this original work did not study the epiplexity of audio data.

We seek to address the question "What audio datasets might be the best to train on from the perspective of epiplexity and generalization?" and investigate the relationship between epiplexity and audio model performance. In particular, we focus on measuring unconditional (Section 3) and conditional (Section 4) epiplexity of audio datasets tokenized with DAC, Encodec, and X-Codec, and compare epiplexity to downstream performance on next token prediction (Section 3) and classification (Section 4), respectively. The experiment methodologies are largely the same, so we offer a detailed explanation in Section 3 and then explain the adjustments needed in Section 4.

2 Related Work

2.1 Audio Complexity Research

Information-theoretic concepts such as Shannon entropy and Kolmogorov complexity formalize notions of randomness and structure, but can be difficult to estimate and fail to explain why techniques like data transformations are helpful (Finzi et al., 2026). Recent work has focused on audio-specific complexity measurements such as musical complexity, but these metrics do not provide a general account of complexity and learnable structure across speech, music, and other audio (Dubnov, 2008; Ribeiro et al., 2012; Gold et al., 2019; Guerrero et al., 2021; Ajmera et al., 2003).

The development of discrete audio representations makes more model-based measurements possible. Neural codecs including SoundStream (Zeghidour et al., 2021), EnCodec (Défossez et al., 2022), DAC (Kumar et al., 2023), and SQ-Codec (Yang et al., 2024) compress waveforms into discrete codes, while SpeechTokenizer (Zhang et al., 2024), Mimi (Défossez et al., 2024), and X-Codec (Ye et al., 2025) capture more semantic information. Recent work shows that even LLMs can be treated as compressors to make information measurement explicit to observers (Delétang et al., 2024). However, these approaches do not explicitly separate the learnable structure from the residual uncertainty under bounded computation.

2.2 Selecting Data for Model Generalization

Much prior work on generalization has focused on benchmarking (Rauch et al., 2025; Turian et al., 2022; Mesaros et al., 2024; Yang et al., 2021) or developments towards domain-invariant methods (Yu et al., 2023; Boudiaf et al., 2023; Zhang et al., 2021). However, there is limited prior work on data selection for audio. A systematic study of pre-training data curation for automatic speech recognition found that optimizing for data diversity did not improve model performance compared to random sampling (Whetten et al., 2026). When dealing with domain mismatches, careful data curation and mixing can improve generalization, suggesting that the composition of a training set matters (Chen et al., 2026; Aldeneh et al., 2026; Whetten et al., 2026). Other methods have used entropy for data selection, but remain sensitive to the structural complexity of the data (Sethy et al., 2009; Huang & Ma, 2011; Arya et al., 2026; He et al., 2014; Perry et al., 2026; Colonna et al., 2020; Finzi et al., 2026).

2.3 Epiplexity

As previously discussed, optimizing machine learning model performance through careful data selection remains a delicate problem. The authors in Finzi et al. (2026) identify three paradoxes in which theoretical predictions about how data selection affects model performance conflict with empirical observations. They introduce the notion of epiplexity to resolve these contradictions. Unlike prior theoretical concepts which lead to these paradoxes, epiplexity imposes a fixed computational constraint and therefore depends on both the data and computational resources available to the observer. The authors in Finzi et al. (2026) introduce methods for estimating epiplexity and conduct

a series of experiments, suggesting that training on datasets with higher epiplexity improves performance on downstream tasks. The resolution of these longstanding paradoxes and empirical results suggest that epiplexity could be a useful metric for data selection.

3 Experiment 1: Unconditional Epiplexity

3.1 Methodology

As defined in Finzi et al. (2026), the epiplexity of a dataset X is the length of a program P that models X, satisfies a run-time constraint, and minimizes the sum of the model's code length and the expected code length of X. The space of viable programs is too large to search in practice, so we limit ourselves to autoregressive transformers and perform a grid search over a fixed set of hyperparameter configurations with a fixed computational budget. We choose the model that yields the minimum estimated code length (up to a smoothing reduction), and use its estimated description length as our epiplexity measurement for X.

3.1.1 Computing Unconditional Epiplexity

Let X be a random variable and let $\mathcal{P}_T$ denote a set of models that can generate some random output Y given X and evaluate its conditional probability, subject to a computational budget T. In our experiments, we use autoregressive transformers, omitting the conditioning on X for unconditional epiplexity. Following Definition 8 in Finzi et al. (2026), the T-bounded epiplexity and entropy are defined as

$$S_T(X) := |P^\star| \quad \text{and} \quad H_T(X) := \mathbb{E}_X\left[\log_2 \frac{1}{P^\star(X)}\right], \tag{1}$$

where $P^\star \in \mathcal{P}_T$ is a program defined so that $S_T(X) + H_T(X)$ is minimal, with ties broken in favor of the shorter program. At a high level, $S_T(X)$ represents the size of the model bounded by some T that is small enough to be compact but large enough to capture the structure of a given dataset, while $H_T(X)$ represents the expected description length of the data under the selected program.

In practice, we train decoder-only transformers as our various programs P to perform next-token prediction, and perform a sweep over the number of layers and the total number of (non-embedding) parameters of our transformer model on each dataset using eight H100 GPUs.

To measure $|P|$, we use the technique of prequential coding described in Finzi et al. (2026, Section 4.1). This procedure gives the code length for the data and the model by encoding the procedure used to train the model. Taking inspiration from the symmetry of information, one subtracts two upper bounds (for the code length of the data and model, and the code length of the data given the model) to get a heuristic estimate for $|P|$.

We leave much of the explanation of prequential encoding to Finzi et al. (2026) and Appendix E, but in short, given batches of data Z_i at training index i, the prequential estimate of $|P|$ is the summed difference between the coding loss of the model P_i (obtained from P_{i-1} by training on Z_{i-1}) and that of the final model P_M on the same batch. The starting model P_0 is initialized randomly, and training proceeds until we reach our token limit. In other words:

$$|P_M| \approx |P_M|_{\text{preq}} := \sum_{i=0}^{M-1} \left[\log_2 \frac{1}{P_i(Z_i)} - \log_2 \frac{1}{P_M(Z_i)}\right]. \tag{2}$$

Remark 1. Note that prequential encoding only provides an estimate of $|P|$. While subtracting two upper bounds does not generally yield a bound, and the symmetry of information does not hold in the compute-limited context, prequential coding has the advantage of being computationally efficient compared to requential coding (see Finzi et al. (2026, Section 4.2)). We thus choose it to scale our experiments to a much larger number of datasets and conduct relative comparisons between datasets.

During model training, we select a number of evaluation steps to measure loss on some number of tokens from the test data. Scaling the loss to the entire test dataset, we obtain an estimate of the entropy code length (ECL) $\mathbb{E}_X\left[\log_2 \frac{1}{P_M(X)}\right]$. Following Finzi et al. (2026), rather than using the

final model to evaluate this loss, we use an exponential moving average model $\widetilde{P}_i$. Its loss on the evaluation data provides the conditional entropy code estimate:

$$\widehat{H}(P_M(X)) := \frac{\#X^{ev}}{\#E} \sum_{Z \in E} \log_2 \frac{1}{\widetilde{P}_M(Z)}, \tag{3}$$

where $\#\cdot$ denotes the number of elements in the set $\cdot$. Note that to reduce compute time, we only evaluate on a subset E of the entire evaluation split X^{ev} and then rescale to the entire split.

At each evaluation step, we also record a model description length (MDL) estimate for the current model. Adding the MDL estimate and ECL estimate gives an estimate for the two-part code length. Along the way, we track the compute budget used (in FLOPs), and thus each run produces a number of (compute, code length) pairs. After discarding points with nonpositive MDL (see Remark 2), we compute the lower convex hull $\mathcal{H}$ of the remaining points and apply the following reduction: let $\mathcal{H}_i \subseteq \mathcal{H}$ be the set of points coming from the i-th run. Assuming $\mathcal{H}_i$ is nonempty, we choose the point in $\mathcal{H}_i$ whose compute value is closest to the median compute value of points in $\mathcal{H}_i$, with ties broken chronologically. The MDL estimate of the point in the reduced hull with minimal code length is our epiplexity estimate.

Note that the MDL estimate we use is not quite the prequential estimate defined earlier, as evaluating the most recent model on prior training tokens is too computationally expensive. We follow Finzi et al. (2026) and replace $\sum_{i=0}^{M-1} \log_2 \frac{1}{P_M(Z_i)}$ in Equation (2) with the average of the $\log_2 \frac{1}{P_i(Z_i)}$ for $i = j+1, j+2, \ldots, M-1$, where j is the most recent evaluation step.

Remark 2. It is immediate from the definition that epiplexity cannot be negative. However, since our MDL estimator subtracts averaged training losses, it can be negative when the loss fails to be monotonically decreasing. We found that noisy upward spikes in loss frequently resulted in negative epiplexity estimates. Since the models we train must have positive MDLs, we discard points where the MDL estimate is nonpositive.

3.1.2 Data Preprocessing and Parameters

All of our experiments require tokenization as a preprocessing step. We first define an audio dataset X and an audio tokenizer $\mathcal{T}$. For each audio clip in X, we resample to match the target sample rate of $\mathcal{T}$ and then apply $\mathcal{T}$ to obtain its tokenized audio. In the case where $\mathcal{T}$ has multiple codebooks, we flatten the resulting array in chronological order. Each tokenized clip is standardized to a length of 5 seconds based on its clip length c (in seconds), by taking a uniformly random, contiguous 5-second segment if $c \geq 5$ and adding 0-padding otherwise. Note that before tokenizing, we split the dataset into training and evaluation splits. In the case where the dataset was shipped with official train and test splits, we respect the official splits. In the other case, we perform our own splitting (see Table 1 for dataset-specific details).

For audio tokenizers, we use DAC, EnCodec, and X-Codec (see Appendix C). For each dataset $X_{\mathcal{T}}$ tokenized by $\mathcal{T} \in \{\text{DAC}, \text{EnCodec}, \text{X-Codec}\}$, we use prequential coding to estimate the epiplexity of the tokenized dataset at a fixed compute budget. This involves a grid sweep of 48 runs, where each run fixes the number of layers $N \in \{3, 6, 12, 24\}$ and the number of parameters $P \in \{1, 2, 5, 10, 20, 40, 60, 80, 100, 120, 140, 160\}$ (in millions) of our transformer model. The sequence length is set equal to the number of tokens in 5 seconds of audio under the specific tokenizer.

Compute is estimated as $6\mathcal{N}D_{tr} + 2\mathcal{N}D_{ev}$, where $\mathcal{N}$ is the number of model parameters and D_{tr} and D_{ev} are the number of training and evaluation tokens, respectively. We use a batch size of 16 and 2 accumulation steps. We warm up on 16384 tokens, linearly increasing the learning rate from 10^{-6} to $2 \cdot 10^{-6}$. We train on $3.125 \cdot 10^7$ tokens and use 20 evaluation steps, evenly spaced across all non-warmup training steps, plus an additional 50 evaluation steps in geometric progression beginning with the second training step. Note that the total number of evaluation steps may be less than 70 since these two sequences may overlap. In each evaluation step, we evaluate on 131072 tokens of the test set. We use a decay rate of 50 for the exponential moving average model.

To investigate the relationship between epiplexity and model generalization, we select two datasets from each of six domains (speech, human non-verbal sound, bioacoustics, music, urban sound, and machine noise). We fix a tokenizer $\mathcal{T}$ and tokenize each of the 12 datasets with $\mathcal{T}$. On each dataset, we train a decoder-only transformer M with 3 layers, a head dimension of 64, and an embedding

dimension of 192 for 50 epochs. We use a batch size of 64. For each of the remaining datasets, i.e. evaluation datasets, we evaluate M on the test split in a zero-shot fashion. Our evaluation consists of, for each clip, predicting the tokens corresponding to seconds 1-2 given the tokens corresponding to seconds 0-1. We measure the average cross-entropy loss achieved over the predicted tokens. We then repeat this procedure, but with the additional step of fine-tuning M for 5 epochs on the evaluation dataset before evaluating. For each fixed tokenizer, we calculate the Spearman correlation between the measured epiplexity of the pretraining dataset and the average loss on evaluation datasets, and we plot a best-fit line using Huber regression.

3.2 RESULTS

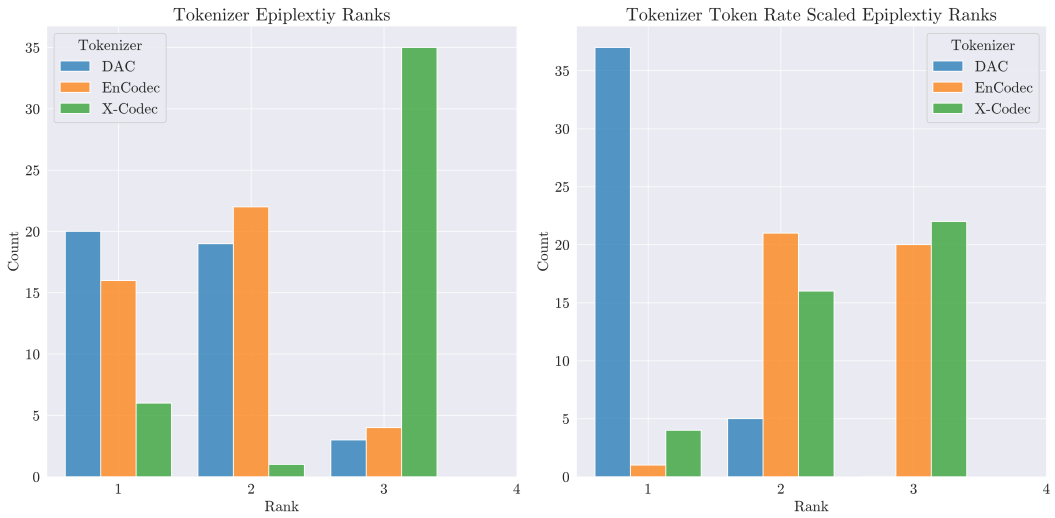


Figure 1: Left: unscaled epiplexity. Right: epiplexity divided by per-second token rate. Epiplexity measurements from Table 3.

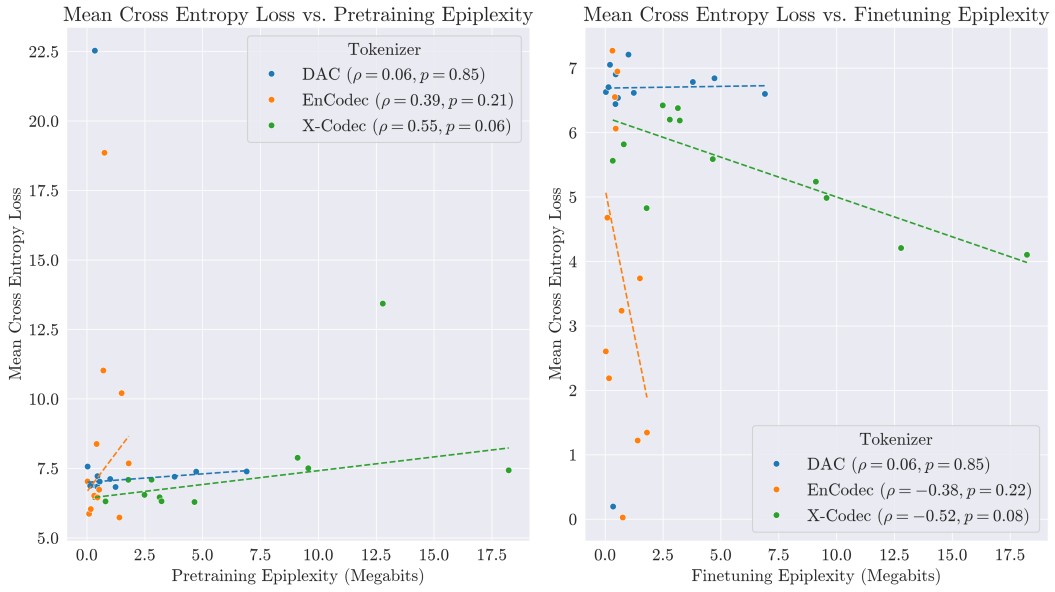


Figure 2: Left: average zero-shot loss vs. epiplexity of pretraining dataset. Right: average loss after finetuning vs. epiplexity of the evaluation dataset. Epiplexities from Table 3, zero-shot losses from Table 4, Table 5, Table 6, finetuning losses from Table 7, Table 8, Table 9.

The measured epiplexity of each dataset and tokenizer combination is reported in Table 3, and dataset metadata can be found in Table 1. For each fixed dataset, we rank tokenizers according to their corresponding epiplexity measurements, with rank 1 corresponding to the lowest epiplexity and rank 3 the highest (see Figure 1). X-Codec is commonly ranked third, while DAC and EnCodec have similarly distributed ranks, with EnCodec outranking DAC slightly more frequently. If we divide epiplexity measurements by per-second token rates and rerank the tokenizers, we find that DAC almost always produces the lowest epiplexity measurement, while EnCodec and X-Codec ranks have similarly distributed ranks, with X-Codec achieving rank 3 slightly more frequently than EnCodec, but also achieving rank 1 more frequently.

We also plot measured epiplexity densities for each tokenizer, as estimated from the data in Table 3. Interestingly, we see that the EnCodec distribution is more concentrated at lower epiplexities and has a shorter right tail than the DAC distribution, despite EnCodec ranking second and third more frequently than DAC (see Figure 5).

For our next token prediction tests on our collection of 12 datasets, the zero-shot loss for each pretraining and evaluation dataset combination is reported in Table 4, Table 5, and Table 6 for DAC, EnCodec, and X-Codec, respectively. In each table, the pretraining dataset is fixed in each row while the evaluation dataset is fixed in each column. Similarly, losses after finetuning are reported in Table 7, Table 8, and Table 9, respectively. The mean zero-shot cross entropy loss over all evaluation datasets is plotted against the measured epiplexity of the pretraining dataset (as reported in Table 3) in Figure 2 (left), along with the Huber regression line for each tokenizer. We similarly plot mean loss after finetuning vs. the measured epiplexity of the finetuning dataset in Figure 2 (right). Our plots also show the Spearman correlation coefficients between loss and epiplexity.

3.3 Discussion

A systematic study of audio tokenizers found that EnCodec outperformed DAC on a number of reconstructive benchmarks and downstream tasks (Mousavi et al., 2025). However, it is important to note that DAC produces 8 times the number of tokens per second as EnCodec does. After rescaling by token rate, EnCodec frequently produces higher epiplexity than DAC does, and DAC never achieves the highest rank. Since DAC has the highest token rate of our three tokenizers, this suggests that DAC captures structural information less efficiently than the other tokenizers. It should be noted that the performance of X-Codec, the tokenizer that produced the highest epiplexity measurements both in terms of ordinal rankings and distribution, was not measured in (Mousavi et al., 2025). However, X-Codec integrates both semantic and acoustic features, whereas EnCodec was designed with audio reconstruction in mind, suggesting that X-Codec may have benefits for downstream tasks like classification.

For EnCodec and X-Codec, we see a non-statistically significant correlation between higher epiplexity of the evaluation data and better model performance after finetuning. On the other hand, the opposite is true for zero-shot performance: for EnCodec and X-Codec, as pretraining epiplexity increases, performance worsens. As noted by Finzi et al. (2026), epiplexity is not guaranteed to correlate with model performance, as it is agnostic to the relevance of a dataset to any particular downstream task. In the case of the zero-shot attempts, the model was not trained on the same domain or data source as the test dataset used to estimate the loss. It is then entirely possible that structural information differs drastically between different audio datasets. The results here might suggest that future work studying epiplexity and covariate shifts should consider how to measure the amount of shared structural information two or more domains might contain to estimate generalization.

The relationship between performance and epiplexity was weak in both cases for DAC. Given the weak correlation and high prediction losses, it is possible that the models trained over these datasets simply failed to perform well.

4 Experiment 2: Conditional Epiplexity

Experiment 1 measures the structure within the audio representation itself. However, many practical applications of audio rely on sound event detection. Thus, in Experiment 2, we instead measure the computationally accessible structure in audio X that is useful for predicting its label Y.

4.1 METHODOLOGY

Much of Experiment 2 mirrors Experiment 1. We still use the same 48-run sweep, training configuration, and evaluation-step schedule as Experiment 1. The largest differences are how data is prepared for the sweeps and the formulation of the two-part code length from prequential coding. Given a compute budget T and a tokenizer $\mathcal{T}$, we estimate the conditional epiplexity $S_T(Y \mid X_{\mathcal{T}})$ using prequential coding and a grid search as in Experiment 1. However, our model is now $P(Y_i \mid X_i) = \mathrm{softmax}(Wh_L(X_i))$, where $h_L(X_i)$ is the hidden state at the final audio-token position and W is a zero-initialized linear readout. Since the final position has attended to the complete causal sequence, its representation is conditioned on all of X_i. While the conditioning input is still accounted for in the compute budget, only the label tokens are considered when computing the two-part code length.

4.1.1 COMPUTING CONDITIONAL EPIPLEXITY

As in Experiment 1, we begin with a randomly initialized model P_0 and train until we reach our token limit. Let $Z_0, Z_1, \ldots, Z_{M-1}$ denote ordered training minibatches, where each Z_i contains audio-label pairs (X_i, Y_i). Model P_i predicts labels in Z_i before being updated on that batch to produce P_{i+1}. Now, similar to Experiment 1, prequential coding estimates $|P_M|$ by the summed difference between the coding losses of the sequence of models P_i and the final model P_M.

$$|P_M|_{\text{preq}} = \sum_{i=0}^{M-1} \sum_{(X_i,Y_i)\in Z_i} \left[\log \frac{1}{P_i(Y_i \mid X_i)} - \log \frac{1}{P_M(Y_i \mid X_i)}\right]. \tag{4}$$

For details on this derivation, see Appendix E. As in Experiment 1, we do not actually evaluate the latest model on all prior samples at every training step. Instead, we approximate the loss of P_M on each training sample by the average loss of P_M on a held-out subset $E = \left\{(X_i^E, Y_i^E)\right\}$ of the entire evaluation split $X_{\mathcal{T}}^{ev}$.

To estimate conditional entropy, we maintain a sequence of exponential moving average models $\widetilde{P}_i$ using the same EMA convention as Experiment 1. As in Experiment 1, the EMA loss provides the conditional entropy code estimate

$$\widehat{H}(P_M(Y^{ev} \mid X_{\mathcal{T}}^{ev})) = \frac{\#X_{\mathcal{T}}^{ev}}{\#E} \sum_{(X_i^E,Y_i^E)\in E} -\log \widetilde{P}_M\left(Y_i^E \mid X_i^E\right).$$

Note that these losses (for both MDL and ECL estimation) only account for label tokens, as noted earlier. The resulting conditional two-part code length at step M is the sum of the MDL and ECL estimates as in Experiment 1. All code lengths are accumulated in nats and converted to bits by dividing by $\log 2$.

4.1.2 DATA PREPROCESSING AND PARAMETERS

Audio preprocessing and tokenization follow Experiment 1. However, we must append the multi-class label index to the flattened audio-token representation during preprocessing. When loading the data, we separate each example into an audio-token sequence X_i and a label Y_i. The class vocabulary is derived from the training split classes and remains unchanged for the evaluation split. Compute is calculated in the same way as in Experiment 1. The hyperparameter values swept over in our grid sweep, as well as batch size, sequence length, etc., remain the same as in Experiment 1. We obtain a conditional epiplexity estimate using the same Pareto front estimation procedure as in Experiment 1.

To investigate the relationship between conditional epiplexity and model performance on classification tasks, we train an encoder-only transformer M with 3 layers, a head dimension of 64, and an embedding dimension of 192 for 50 epochs. We use a batch size of 64. We pass the encoder outputs into a linear readout layer to obtain class logits. We evaluate the resulting classification model on the test split and measure the class balance-weighted average F1-score as well as accuracy, and plot

these metrics against our conditional epiplexity measurements. The datasets and tokenizers used for Experiment 2 are reported in Table 2.

4.2 RESULTS



Figure 3: Left: unscaled conditional epiplexity. Right: conditional epiplexity divided by per-second token rate. Epiplexity measurements from Table 10.

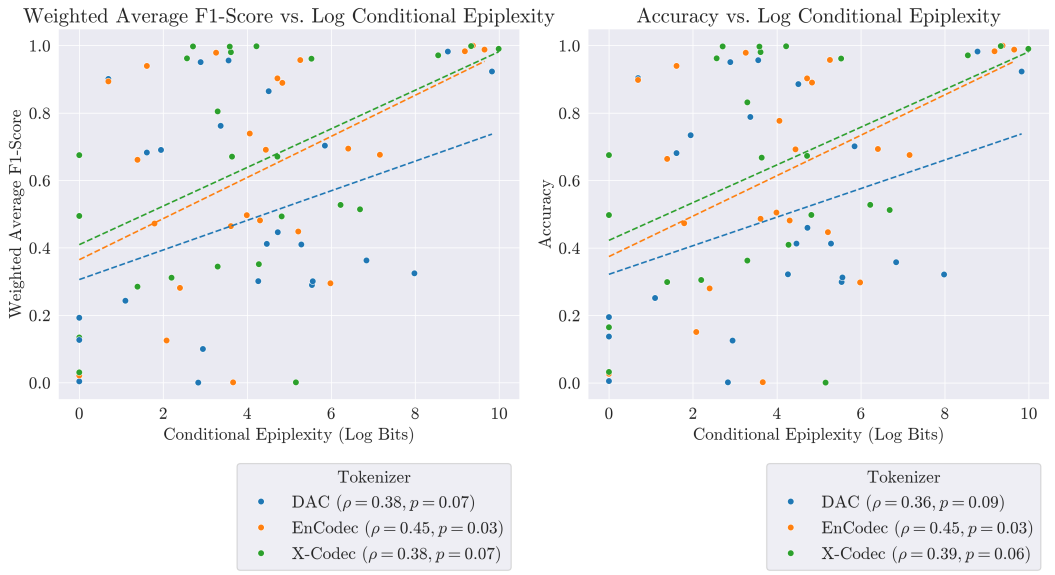


Figure 4: We use a logarithmic x-scale due to the high variance of the conditional epiplexity data. Epiplexities from Table 10, metrics from Table 11 and Table 12.

The measured conditional epiplexity of each dataset and tokenizer combination is reported in Table 10, and dataset metadata can be found in Table 2. For each fixed dataset, we rank tokenizers according to their corresponding conditional epiplexity measurements, with rank 1 corresponding to the lowest epiplexity and rank 3 to the highest (see Figure 3). All tokenizers achieve rank 1 at identical frequencies, while EnCodec is most frequently ranked third. As in the unconditional epiplexity experiment, we also rank the epiplexity measurements after rescaling by token rate. In this case, we find that EnCodec almost always achieves rank 3 and DAC commonly achieves rank 1.

We also plot conditional epiplexity densities for each tokenizer, as estimated from the data in Table 10. We note high variance in conditional epiplexity for all tokenizers.

The F1-scores and accuracies from our performance tests can be found in Table 11 and Table 12, respectively. We plot these metrics against the logarithm of the measured conditional epiplexity for the dataset. For each fixed tokenizer and each evaluation metric, we calculate the Spearman correlation between the conditional epiplexity of the pretraining dataset and the given metric, and we plot a best-fit line using Huber regression. For each metric, we also plot the density for each tokenizer (Figure 7).

4.3 Discussion

Compared to Experiment 1, the tokenizer ranks in Figure 3 are more evenly distributed, with EnCodec most frequently ranking third. However, after rescaling conditional epiplexity by token rates, EnCodec consistently produces the highest conditional epiplexity, while DAC produces the lowest.

The performance vs. conditional epiplexity graphs also tell a more consistent story than those for unconditional epiplexity do. We see from Figure 4 that the performance on classification tasks is weakly correlated with the logarithmic conditional epiplexity of the labeled data, particularly for EnCodec.

Altogether, these results suggest that at these computational budgets and dataset sizes, classification tasks might be better suited for these audio datasets than self-supervised ones. Prior work suggests that in some audio tasks, supervised learning can, at scale, match the performance of unsupervised learning tasks. For example, Perch 2.0, a SOTA bioacoustic model, is trained via supervised learning and outperforms self-supervised models across a variety of related benchmarks (van Merriënboer et al., 2026). Our results show that structure in the dataset relative to some label is much clearer. However, we observe that less structure is needed to represent that label overall. As such, it can explain why sound event detection tasks remain such a useful way to work with audio data. Future work could address the impact of larger pretraining datasets to better determine how much data might be needed for conditional epiplexity to outpace unconditional epiplexity, if at all. Answering that question could inform whether to focus on supervised or self-supervised learning in audio data.

5 Conclusion

Our work investigates unconditional and conditional epiplexity with respect to audio model performance. We find that while unconditional epiplexity has a weak relationship with performance on unseen or out-of-domain data, conditional epiplexity is more strongly correlated with performance. This suggests that audio generalizes in more task-specific directions.

To test on many datasets, we make compromises as measuring epiplexity over a hyperparameter sweep is slow even in the best case. To make things faster, we use prequential coding and reduce T to about 31 million tokens, as opposed to using more tokens and requential coding for better estimates of epiplexity. Future work therefore should consider these choices as well as means to speed up this process for more feasible experiments, including but not limited to removing the need for hyperparameter sweeps.

Open questions remain regarding cross-domain tasks. This work does not fully address the question of how audio datasets might help models generalize, if at all, to out-of-domain distributions. An interesting approach might be to measure a "joint epiplexity" between datasets to understand how much inherent structure they share and how audio models could learn from a few aggregated datasets at a time.

Answering these questions can help guide training paradigms for audio models moving towards physical AI applications. As it stands, if more audio datasets need to be labeled to achieve sufficiently good downstream performance, that will bottleneck audio models significantly, driving demand towards more human labeling. On the other hand, if it is a matter of scale, simulation may be useful. Exploring the structural information of audio will thus be a productive way to guide experiments in audio modeling.

Author Contributions

Censored for review

Acknowledgments

Censored for review

Ethics Statements

We confirm that our study complies with the ICLR Code of Ethics and does not present additional ethical issues.

AI Use

In this work, generative AI tools were not used for creation of methodology, analysis, experiment design, interpreting results or anything else in the scientific cycle. Generative AI tools were exclusively used for creating and editing software used for experiments and figure creation. All AI-assisted work was reviewed manually by humans as all code had to pass a manual review process before being included in the main branch of the git repository. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility

To run the code, please see `https://anonymous.4open.science/r/Audio_Through_Lens_Of_Epiplexity-6E59`. Note that the system will automatically download data depending on the data loader you are trying to use. We recommend using a dataset like FSD50k on smaller machines.

References

Mohammed Abdeldayem. The Fake-or-Real (FoR) Dataset (deepfake audio), 2024. URL `https://www.kaggle.com/datasets/mohammedabdeldayem/the-fake-or-real-dataset`.

Jitendra Ajmera, Iain McCowan, and Hervé Bourlard. Speech/music segmentation using entropy and dynamism features in a hmm classification framework. *Speech Communication*, 40(3):351–363, 2003. ISSN 0167-6393. doi: https://doi.org/10.1016/S0167-6393(02)00087-0. URL `https://www.sciencedirect.com/science/article/pii/S0167639302000870`.

Zakaria Aldeneh, Skyler Seto, Maureen de Seyssel, Jie Chi, Zijin Gu, Takuya Higuchi, Jee weon Jung, Shinji Watanabe, David Grangier, Barry-John Theobald, and Tatiana Likhomanenko. Which data matter? embedding-based data selection for speech recognition, 2026. URL `https://arxiv.org/abs/2603.05819`.

Anonymous. Anomalymachine-50k: Synthetic industrial machine sound anomaly dataset. `https://huggingface.co/datasets/AnomalyMachine-50K`, 2026.

Apple. Spatial librispeech, 2023. URL `https://github.com/apple-aiml-research/ml-spatial-librispeech`.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus, 2020. URL `https://arxiv.org/abs/1912.06670`.

Vipul Arya, S. H. Shabbeer Basha, Srikrishna U N, Sunainha Vijay, and Snehasis Mukherjee. Balancing uncertainty and diversity of samples: Leveraging diversity of least, high confidence samples for effective active learning, 2026. URL `https://arxiv.org/abs/2605.22169`.

Malik Boudiaf, Tom Denton, Bart van Merriënboer, Vincent Dumoulin, and Eleni Triantafillou. In search for a generalizable method for source free domain adaptation, 2023. URL https://arxiv.org/abs/2302.06658.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020. URL https://arxiv.org/abs/2005.14165.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *Oriental COCOSDA 2017*, pp. Submitted, 2017.

Mark Cartwright, Jason Cramer, Ana Elisa Mendez Mendez, Yu Wang, Ho-Hsiang Wu, Vincent Lostanlen, Magdalena Fuentes, Graham Dove, Charlie Mydlarz, Justin Salamon, Oded Nov, and Juan Pablo Bello. Sonyc urban sound tagging (sonyc-ust): a multilabel dataset from an urban acoustic sensor network, September 2020a. URL https://doi.org/10.5281/zenodo.3966543.

Mark Cartwright, Jason Cramer, Ana Elisa Mendez Mendez, Yu Wang, Ho-Hsiang Wu, Vincent Lostanlen, Magdalena Fuentes, Graham Dove, Charlie Mydlarz, Justin Salamon, Oded Nov, and Juan Pablo Bello. Sonyc-ust-v2: An urban sound tagging dataset with spatiotemporal context, 2020b. URL https://arxiv.org/abs/2009.05188.

Honglie Chen. Vggsound, 2020. URL https://github.com/hche11/VGGSound.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725, 2020. doi: 10.1109/ICASSP40776.2020.9053174.

Zhuo Chen, Yuxuan Miao, and Deyi Xiong. Data mixing for large language models pretraining: A survey and outlook. *Data Intelligence*, 8(1):15–55, March 2026. ISSN 2096-7004. doi: 10.3724/2096-7004.di.2026.0055. URL http://dx.doi.org/10.3724/2096-7004.di.2026.0055.

Lauren M. Chronister, Tessa A. Rhinehart, Aidan Place, and Justin Kitzes. An annotated set of audio recordings of eastern north american birds containing frequency, time, and species information, January 2022. URL https://doi.org/10.5061/dryad.d2547d81z.

Mary Clapp, Stefan Kahl, Erik Meyer, Megan McKenna, Holger Klinck, and Gail Patricelli. A collection of fully-annotated soundscape recordings from the southern sierra nevada mountain range, January 2023. URL https://doi.org/10.5281/zenodo.7525805.

Juan G. Colonna, José R. H. Carvalho, and Osvaldo A. Rosso. Estimating ecoacoustic activity in the Amazon rainforest through Information Theory quantifiers. *PLOS ONE*, 15(7):e0229425, July 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0229425. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0229425.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. *arXiv preprint arXiv:2205.12446*, 2022a. URL https://arxiv.org/abs/2205.12446.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech, 2022b. URL https://arxiv.org/abs/2205.12446.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis, 2017a. URL https://arxiv.org/abs/1612.01840.

Michaël Defferrard, Justin Kalavan, and FrescoFlacko. Fma: A dataset for music analysis, 2017b. URL https://github.com/mdeff/fma.

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, et al. Language modeling is compression. In *The Twelfth International Conference on Learning Representations*, volume 2024, pp. 14165–14181, 2024. URL https://openreview.net/forum?id=jznbgiynus.

Shlomo Dubnov. Unified view of prediction and repetition structure in audio signals with application to interest point detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 16 (2):327–337, 2008. doi: 10.1109/TASL.2007.912378.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression, 2022. URL https://arxiv.org/abs/2210.13438.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue, 2024. URL https://arxiv.org/abs/2410.00037.

Christine Evers, Heinrich Loellmann, Heinrich Mellmann, Alexander Schmidt, Hendrik Barfuss, Patrick A. Naylor, and Walter Kellermann. Data corpus for the ieee-aasp challenge on acoustic source localization and tracking (locata), January 2020a. URL https://doi.org/10.5281/zenodo.3630471.

Christine Evers, Heinrich W. Lollmann, Heinrich Mellmann, Alexander Schmidt, Hendrik Barfuss, Patrick A. Naylor, and Walter Kellermann. The locata challenge: Acoustic source localization and tracking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1620–1643, 2020b. ISSN 2329-9304. doi: 10.1109/taslp.2020.2990485. URL http://dx.doi.org/10.1109/TASLP.2020.2990485.

Hugging Face. Bat, 2024a. URL https://huggingface.co/datasets/zhisheng01/SpatialAudio.

Hugging Face. Commonvoice, 2024b. URL https://huggingface.co/datasets/fixie-ai/common_voice_17_0.

Facebook. Voxpopuli, 2021. URL https://huggingface.co/datasets/facebook/voxpopuli.

Marc Finzi, Shikai Qiu, Yiding Jiang, Pavel Izmailov, J. Zico Kolter, and Andrew Gordon Wilson. From entropy to epiplexity: Rethinking information for computationally bounded intelligence, 2026. URL https://arxiv.org/abs/2601.03220.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. Fsd50k, October 2020. URL https://doi.org/10.5281/zenodo.4060432.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. Fsd50k: An open dataset of human-labeled sound events, 2022. URL https://arxiv.org/abs/2010.00475.

Luca Fredianelli, Francesco Artuso, Geremia Pompei, Gaetano Licitra, Gino Iannace, and Andac Akbaba. Environmental noise dataset for sound event classification and detection. *Scientific Data*, 12(1):5991, 2025a. doi: 10.1038/s41597-025-05991-w. URL https://doi.org/10.1038/s41597-025-05991-w.

Luca Fredianelli, Francesco Artuso, Geremia Pompei, Gaetano Licitra, Gino Iannace, and Andaç Akbaba. Datased - dataset for sound event detection of environmental noise, May 2025b. URL https://doi.org/10.5281/zenodo.15346092.

Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

Benjamin P. Gold, Marcus T. Pearce, Ernest Mas-Herrero, Alain Dagher, and Robert J. Zatorre. Predictability and Uncertainty in the Pleasure of Music: A Reward for Learning? *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 39(47):9397–9409, November 2019. ISSN 1529-2401 0270-6474. doi: 10.1523/JNEUROSCI.0428-19.2019.

Yuan Gong. Vocalsound: A dataset for improving human vocal sounds recognition, 2022. URL https://github.com/YuanGongND/vocalsound#Download-VocalSound.

Yuan Gong, Jin Yu, and James Glass. Vocalsound: A dataset for improving human vocal sounds recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 151–155. IEEE, May 2022. doi: 10.1109/icassp43922.2022.9746828. URL http://dx.doi.org/10.1109/ICASSP43922.2022.9746828.

Google. Fleurs, 2022. URL https://huggingface.co/datasets/google/fleurs.

Marc Ciufo Green and Damian Murphy. Eigenscape: A database of spatial acoustic scene recordings. *Applied Sciences*, 7(11), 2017a. ISSN 2076-3417. doi: 10.3390/app7111204. URL https://www.mdpi.com/2076-3417/7/11/1204.

Marc Ciufo Green and Damian Murphy. Eigenscape, October 2017b. URL https://doi.org/10.5281/zenodo.1012809.

Daniel Guerrero, Pedro Rivera, Gerardo Febres, and Carlos Gershenson. Towards a measure for characterizing the informational content of audio signals and the relation between complexity and auditory encoding. *Entropy*, 23(12):1613, November 2021. ISSN 1099-4300. doi: 10.3390/e23121613. URL https://www.mdpi.com/1099-4300/23/12/1613.

Zhongbin Guo, Zhen Yang, Yushan Li, Xinyue Zhang, Wenyu Gao, Jiacheng Wang, Chengzhi Li, Xiangrui Liu, and Ping Jian. Can llms see without pixels? benchmarking spatial intelligence from textual descriptions, 2026. URL https://arxiv.org/abs/2601.03590.

Tianxu He, Shukui Zhang, Jie Xin, Pengpeng Zhao, Jian Wu, Xuefeng Xian, Chunhua Li, and Zhiming Cui. An active learning approach with uncertainty, representativeness, and diversity. *ScientificWorldJournal*, 2014:827586, August 2014. ISSN 1537-744X 2356-6140. doi: 10.1155/2014/827586.

Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen. Acoustic scene classification in DCASE 2020 Challenge: Generalization across devices and low complexity solutions. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, pp. 56–60, Tokyo, Japan, 2020. URL https://arxiv.org/abs/2005.14623.

Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen. Tau urban acoustic scenes 2022 mobile, development dataset, March 2022a. URL https://doi.org/10.5281/zenodo.6337421.

Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen. Tau urban acoustic scenes 2022 mobile, evaluation dataset, June 2022b. URL https://doi.org/10.5281/zenodo.6591203.

W. Alexander Hopping, Stefan Kahl, and Holger Klinck. A collection of fully-annotated soundscape recordings from the southwestern amazon basin, September 2022. URL https://doi.org/10.5281/zenodo.7079124.

Chien-Lin Huang and Bin Ma. Maximum entropy based data selection for speaker recognition. In *Interspeech 2011*, pp. 2713–2716, ISCA, August 2011. ISCA. doi: 10.21437/Interspeech.2011-145.

Il-Young Jeong and Jeongsoo Park. Cochlscene: Acquisition of acoustic scene data using crowdsourcing, November 2022a. URL https://doi.org/10.5281/zenodo.7080122.

Il-Young Jeong and Jeongsoo Park. Cochlscene: Acquisition of acoustic scene data using crowdsourcing, 2022b. URL https://arxiv.org/abs/2211.02289.

Stefan Kahl, Russell Charif, and Holger Klinck. A collection of fully-annotated soundscape recordings from the northeastern united states, August 2022a. URL https://doi.org/10.5281/zenodo.7018484.

Stefan Kahl, Connor M. Wood, Philip Chaon, M. Zachariah Peery, and Holger Klinck. A collection of fully-annotated soundscape recordings from the western united states, September 2022b. URL https://doi.org/10.5281/zenodo.7050014.

Yuma Koizumi, Shoichiro Saito, Noboru Harada, Hisashi Uematsu, and Keisuke Imoto. Toyadmos dataset, July 2019a. URL https://doi.org/10.5281/zenodo.3351307.

Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, Noboru Harada, and Keisuke Imoto. Toyadmos: A dataset of miniature-machine operating sounds for anomalous sound detection. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 313–317, 2019b. doi: 10.1109/WASPAA.2019.8937164.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, volume 36 of *NIPS '23*, pp. 27980–27993, Red Hook, NY, USA, 2023. Curran Associates Inc. URL https://dl.acm.org/doi/10.5555/3666122.3667336.

Rama Rao Kvsn, James Montgomery, Saurabh Garg, and Michael Charleston. Bioacoustics data analysis – a taxonomy, survey and open challenges. *IEEE Access*, 8:57684–57708, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.2978547.

Seong-Gyun Leem, Daniel Fulford, Jukka-Pekka Onnela, David Gard, and Carlos Busso. Describe where you are: Improving noise-robustness for speech emotion recognition with text description of the environment, 2025. URL https://arxiv.org/abs/2407.17716.

Nguyen Lim. Emotional Speech Dataset (ESD), 2024. URL https://www.kaggle.com/datasets/nguyenthanhlim/emotional-speech-dataset-esd.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. Clotho-aqa: A crowdsourced dataset for audio question answering, 2022a. URL https://arxiv.org/abs/2204.09634.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. Clotho-aqa dataset, April 2022b. URL https://doi.org/10.5281/zenodo.6473207.

Weichen Liu, Qiyao Xue, Haoming Wang, Xiangyu Yin, Boyuan Yang, and Wei Gao. Spatial reasoning in multimodal large language models: A survey of tasks, benchmarks and methods, 2025. URL https://arxiv.org/abs/2511.15722.

Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):e0196391, 2018a. doi: 10.1371/journal.pone.0196391.

Steven R. Livingstone and Frank A. Russo. The ryerson audio-visual database of emotional speech and song (ravdess), April 2018b. URL https://doi.org/10.5281/zenodo.1188976.

A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen. DCASE 2017 challenge setup: Tasks, datasets and baseline system. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, pp. 85–92, November 2017a.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Tut database for acoustic scene classification and sound event detection. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pp. 1128–1132, 2016a. doi: 10.1109/EUSIPCO.2016.7760424.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Tut acoustic scenes 2016, development dataset, February 2016b. URL https://doi.org/10.5281/zenodo.45739.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Tut acoustic scenes 2017, evaluation dataset, November 2017b. URL https://doi.org/10.5281/zenodo.1040168.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Tut acoustic scenes 2017, development dataset, March 2017c. URL https://doi.org/10.5281/zenodo.400515.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Tut sound events 2016, evaluation dataset, September 2017d. URL https://doi.org/10.5281/zenodo.996424.

Annamaria Mesaros, Romain Serizel, Toni Heittola, Tuomas Virtanen, and Mark D. Plumbley. A decade of dcase: Achievements, practices, evaluations and future challenges, 2024. URL https://arxiv.org/abs/2410.04951.

G. Meza, M. Sepúlveda, A. S. Roman, J. R. Sigal Sefchovich, and I. R. Roman. MULTIVOX: A Spatial Audio-Visual Dataset of Singing Groups. In *Proceedings of the AES International Conference on Artificial Intelligence and Machine Learning for Audio (AES AIMLA)*, London, 2025a. doi: 10.5281/zenodo.17058101.

Gerardo Meza, Mariana Sepúlveda, Adrian Roman, Jorge Rodrigo Sigal Sefchovich, and Iran Roman. [part2 of multivox: A spatial audio-visual dataset of singing groups], September 2025b. URL https://doi.org/10.5281/zenodo.17065497.

Gerardo Meza, Mariana Sepúlveda, Adrian Roman, Jorge Rodrigo Sigal Sefchovich, and Iran R. Roman. Multivox: A spatial audio-visual dataset of singing groups, September 2025c. URL https://doi.org/10.5281/zenodo.17058101.

Irene Martin Morato and Annamaria Mesaros. Macs - multi-annotator captioned soundscapes, July 2021. URL https://doi.org/10.5281/zenodo.5114771.

Veronica Morfi, Yves Bas, Hanna Pamuła, Hervé Glotin, and Dan Stowell. Nips4bplus: a richly annotated birdsong audio dataset, 2018. URL https://arxiv.org/abs/1811.02275.

Pooneh Mousavi, Gallil Maimon, Adel Moumen, Darius Petermann, Jiatong Shi, Haibin Wu, Haici Yang, Anastasia Kuznetsova, Artem Ploujnikov, Ricard Marxer, Bhuvana Ramabhadran, Benjamin Elizalde, Loren Lugosch, Jinyu Li, Cem Subakan, Phil Woodland, Minje Kim, Hung yi Lee, Shinji Watanabe, Yossi Adi, and Mirco Ravanelli. Discrete audio tokens: More than a survey!, June 2025. URL https://arxiv.org/abs/2506.10274.

Nicolas M. Müller, Nicholas Evans, Hemlata Tak, Philip Sperl, and Konstantin Böttinger. Harder or different? understanding generalization of audio deepfake detection, June 2024. URL https://arxiv.org/abs/2406.03512.

Preetum Nakkiran, Behnam Neyshabur, and Hanie Sedghi. The deep bootstrap framework: Good online learners are good offline generalizers, 2021. URL https://arxiv.org/abs/2010.08127.

Amanda Navine, Stefan Kahl, Ann Tanimoto-Johnson, Holger Klinck, and Patrick Hart. A collection of fully-annotated soundscape recordings from the island of hawai'i, September 2022. URL https://doi.org/10.5281/zenodo.7078499.

Tong Niu, Yaqi Chen, Dan Qu, and Hengbo Hu. Enhancing far-field speech recognition with mixer: A novel data augmentation approach. *Applied Sciences*, 15(4073), 2025. ISSN 2076-3417. doi: 10.3390/app15074073. URL https://www.mdpi.com/2076-3417/15/7/4073.

OpenSLR. Aishell, 2017. URL https://www.openslr.org/33/.

OpenSLR. Multilingual librispeech (mls), 2020. URL https://www.openslr.org/94/.

OpenSLR. Aishell-3, 2021. URL https://www.openslr.org/93/.

Sean Perry, Tianqi Zhang, Siya Kamboj, Anu Jajodia, Dhruv Tomar, and Ryan Kastner. Studying domain shifts in bioacoustic recording style between focal and soundscape recordings with acoustic indices. *Ecological Informatics*, 95:103739, May 2026. ISSN 1574-9541. doi: 10.1016/j.ecoinf.2026.103739. URL https://www.sciencedirect.com/science/article/pii/S1574954126001457.

Archontis Politis, Kazuki Shimada, Parthasaarathy Sudarsanam, Aapo Hakala, Shusuke Takahashi, Daniel Alexander Krause, Naoya Takahashi, Sharath Adavanne, Yuichiro Koyama, Kengo Uchida, Yuki Mitsufuji, and Tuomas Virtanen. Starss23: Sony-tau realistic spatial soundscapes 2023, March 2023. URL https://doi.org/10.5281/zenodo.7880637.

Soujanya Poria. Meld, 2018. URL https://affective-meld.github.io/.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations, 2019. URL https://arxiv.org/abs/1810.02508.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. Mls: A large-scale multilingual dataset for speech research. In *Interspeech 2020*, pp. 2757–2761. ISCA, October 2020. doi: 10.21437/interspeech.2020-2826. URL http://dx.doi.org/10.21437/Interspeech.2020-2826.

Cristian Pérez-Granados, Jon Morant, Kevin Darras, Oscar H. Marín-Gómez, Irene Mendoza, Miguel A. Muñoz-Mohedano, Eduardo Santamaría-García, Giulia Bastianelli, Alba Márquez-Rodríguez, Michał Budka, Gerard Bota, José M. De la Peña-Rubio, Eladio L. García de la Morena, Manu Santa-Cruz, Pablo de la Nava, Mario Fernández-Tizón, Hugo Sánchez-Mateos, Adrián Barrero, Juan Traba, Tomasz S. Osiejuk, Patrick J. Hart, Amanda K. Navine, Andrés F. Montoya Muñoz, Carlos B. de Araujo, Gabriel L. Medina Rosa, Ingrid M. Denóbile Torres, Ana L. Camargo Catalano, Cássio Rachid Simões, Diego Llusia, Manuel B. Morales, Pablo Acebes, Juan A. Medina, Nicholas Brown, Christos Astaras, Ilias Karmiris, Elizabeth Navarrete, Maxime Cauchoix, Luc Barbaro, David Funosas, Dominik Arend, Sandra Müeller, Fernando González-García, Alberto González-Romero, Christos Mammides, Michaelangelo Pontikis, Giordano Jacuzzi, Julian D. Olden, Sara P. Bombaci, Gabriel Marcacci, Alain Jacot, Juan P. Zurano, Elena Gangenova, Diego M. Varela, Facundo Di Sallo, Gustavo A. Zurita, Andrey Atemasov, Junior A. Tremblay, Vincent Lamarre, Anja Hutschenreiter, Alan Monroy-Ojeda, Mauricio Díaz-Vallejo, Sergio Chaparro-Herrera, Robert A. Briers, Renata Sousa-Lima, Thiago Pinheiro, Wigna C. da Silva, Alice Calvente, Anamaria Dal Molin, Alexandre Antonelli, Svetlana Gogoleva, Igo Palko, Hiêu V. Trong, Marina H. Lage Duarte, Natalia dos Santos Saturnino, Samuel R. Silva, Ana Rainho, Paula Lopes, Karl-L. Schuchmann, Marinêz I. Marques, Ana S. de Oliveira, Nick A. Littlewood, Mao-Ning Tuanmu, Yi-Ru Cheng, Hsuan Chao, Sebastian Kepfer-Rojas, Andrea L. Aguilera, Lluís Brotons, Mariano J. Feldman, Louis Imbeau, Pooja Panwar, Aaron S. Weed, Anant Deshwal, Raiane Vital da Paz, Carlos Salustio-Gomes, Dorgival D. Oliveira-Junior, Cicero S. Lima-Santos, Mauro Pichorim, Pan Wuyuan, Eben Goodale, Alfredo Attisano, Jörn Theuerkauf, and Esther Sebastián-González. Wabad: A world annotated bird acoustic dataset for passive acoustic monitoring, June 2026a. URL https://doi.org/10.5281/zenodo.20513304.

Cristian Pérez-Granados, Jon Morant, Kevin F. A. Darras, Oscar H. Marín-Gómez, Irene Mendoza, Miguel A. Muñoz-Mohedano, Eduardo Santamaría-García, Giulia Bastianelli, Alba Márquez-Rodríguez, Michał Budka, Gerard Bota, José M. De la Peña-Rubio, Eladio L. García de la Morena, Manu Santa-Cruz, Pablo de la Nava, Mario Fernández-Tizón, Hugo Sánchez-Mateos, Adrián Barrero, Juan Traba, Tomasz S. Osiejuk, Patrick J. Hart, Amanda K. Navine, Andrés F. Montoya Muñoz, Carlos B. de Araujo, Gabriel L. M. Rosa, Ingrid M. Denóbile Torres, Ana L. Camargo Catalano, Cássio Rachid Simões, Diego Llusia, Manuel B. Morales, Pablo Acebes, Juan A. Medina, Nicholas Brown, Christos Astaras, Ilias Karmiris, Elizabeth Navarrete, Maxime Cauchoix, Luc Barbaro, David Funosas, Dominik Arend, Sandra Müeller, Fernando González-García, Alberto González-Romero, Christos Mammides, Michaelangelo Pontikis, Giordano Jacuzzi, Julian D. Olden, Sara P. Bombaci, Gabriel Marcacci, Alain Jacot, Juan P. Zurano, Elena Gangenova, Diego Varela, Facundo Di Sallo, Gustavo A. Zurita, Andrey Atemasov, Junior A. Tremblay, Vincent Lamarre, Anja Hutschenreiter, Alan Monroy-Ojeda, Mauricio Díaz-Vallejo, Sergio Chaparro-Herrera, Robert A. Briers, Renata Sousa-Lima, Thiago Pinheiro, Wigna C. da Silva, Alice Calvente, Anamaria Dal Molin, Alexandre Antonelli, Svetlana Gogoleva, Igo Palko, Hiêu Vũ Trọng, Marina H. Lage Duarte, Natalia dos Santos Saturnino, Samuel R. Silva, Ana Rainho, Paula Lopes, Karl-L. Schuchmann, Marinêz I. Marques, Ana S. de Oliveira, Nick A. Littlewood, Mao-Ning Tuanmu, Yi-Ru Cheng, Hsuan Chao, Sebastian Kepfer-Rojas, Andrea L. Aguilera, Lluís Brotons, Mariano J. Feldman, Louis Imbeau, Pooja Panwar, Aaron S. Weed, Anant Deshwal, Raiane Vital da Paz, Carlos Salustio-Gomes, Dorgival D. Oliveira-Júnior, Cicero S. Lima-Santos,

Mauro Pichorim, Wuyuan Pan, Eben Goodale, Alfredo Attisano, Jörn Theuerkauf, and Esther Sebastián-González. Wabad: A world annotated bird acoustic dataset for passive acoustic monitoring. *Ecology*, 107(2):e70317, 2026b. doi: https://doi.org/10.1002/ecy.70317. URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecy.70317.

Muhammad Mamunur Rashid, Guiqing Li *, and Chengrui Du. Nonspeech7k dataset, June 2023a. URL https://doi.org/10.5281/zenodo.6967442.

Muhammad Mamunur Rashid, Guiqing Li, and Chengrui Du. Nonspeech7k dataset: Classification and analysis of human non-speech sound. *IET Signal Processing*, 17(6):e12233, 2023b. doi: https://doi.org/10.1049/sil2.12233. URL https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/sil2.12233.

Lukas Rauch, Raphael Schwinger, Moritz Wirth, René Heinrich, Denis Huseljic, Marek Herde, Jonas Lange, Stefan Kahl, Bernhard Sick, Sven Tomforde, and Christoph Scholz. Birdset: A large-scale dataset for audio classification in avian bioacoustics, 2025. URL https://arxiv.org/abs/2403.10380.

Ricardo Reimao and Vassilios Tzerpos. For: A dataset for synthetic speech detection. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 1–10, 2019. doi: 10.1109/SPED.2019.8906599.

Haroldo V. Ribeiro, Luciano Zunino, Renio S. Mendes, and Ervin K. Lenzi. Complexity–entropy causality plane: A useful approach for distinguishing songs. *Physica A: Statistical Mechanics and its Applications*, 391(7):2421–2428, 2012. ISSN 0378-4371. doi: 10.1016/j.physa.2011.12.009.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. Urbansound8k, November 2014. URL https://doi.org/10.5281/zenodo.1203745.

Miguel Sarabia, Elena Menyaylenko, Alessandro Toso, Skyler Seto, Zakaria Aldeneh, Shadi Pirhosseinloo, Luca Zappella, Barry-John Theobald, Nicholas Apostoloff, and Jonathan Sheaffer. Spatial librispeech: An augmented dataset for spatial audio learning. In *INTERSPEECH 2023*, pp. 3724–3728. ISCA, August 2023. doi: 10.21437/interspeech.2023-2117. URL http://dx.doi.org/10.21437/Interspeech.2023-2117.

Avi Schwarzschild, Eitan Borgnia, Arjun Gupta, Arpit Bansal, Zeyad Emam, Furong Huang, Micah Goldblum, and Tom Goldstein. Datasets for studying generalization from easy to hard examples, 2021. URL https://arxiv.org/abs/2108.06011.

Abhinav Sethy, Panayiotis G. Georgiou, Bhuvana Ramabhadran, and Shrikanth Narayanan. An iterative relative entropy minimization-based data selection approach for n-gram model adaptation. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(1):13–23, 2009. doi: 10.1109/TASL.2008.2006654.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. Aishell-3: A multi-speaker mandarin tts corpus and the baselines, 2021. URL https://arxiv.org/abs/2010.11567.

Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam, Daniel Krause, Kengo Uchida, Sharath Adavanne, Aapo Hakala, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Tuomas Virtanen, and Yuki Mitsufuji. Starss23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events, 2023. URL https://arxiv.org/abs/2306.09126.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Jun Zhang, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. Can large language models understand spatial audio?, 2024. URL https://arxiv.org/abs/2406.07914.

Core Team, Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, Xin Zhang, Xingchen Song, Yihan Yan, Yongzhe He, Cici, Bowen Shen, Chengxuan Zhu, Chong Ma, Chun Chen, Heyu Chen, Jiawei Li, Lei Li, Menghang Zhu, Peidian Li, Qiying Wang, Sirui Deng, Weimin Xiong, Wenshan Huang, Wenyu Yang, Yilin Jiang, Yixin Yang, Yuanyuan Tian, Yue Ma, Yue Yu, Zihan Zhang, Zihao Yue, Bangjun Xiao, Bingquan Xia, Bofei Gao, Bowen Ye, Can Cai, Chang Liu, Chenhong He, Chunan Li, Dawei

Zhu, Duo Zhang, Fengyuan Shi, Guoan Wang, Hailin Zhang, Hanglong Lv, Hanyu Li, Hao Tian, Heng Qu, Hongshen Xu, Houbin Zhang, Huaqiu Liu, Jiangshan Duo, Jianguang Zuo, Jianyu Wei, Jiebao Xiao, Jinhao Dong, Jun Shi, Junhao Hu, Kainan Bao, Kang Zhou, Linghao Zhang, Meng Chen, Nuo Chen, Peng Zhang, Qianli Chen, Qiantong Wang, Rang Li, Shaohui Liu, Shengfan Wang, Shicheng Li, Shihua Yu, Shijie Cao, Shimao Chen, Shuhao Gu, Weikun Wang, Wenhan Ma, Xiangwei Deng, Xing Yong, Xing Zhang, Xu Wang, Yifan Song, Yihao Zhao, Yingbo Zhao, Yizhao Gao, Yu Cheng, Yu Tu, Yudong Wang, Zhaojun Huang, Zhengju Tang, Zhenru Lin, Zhichao Song, Zhipeng Xu, Zhixian Zheng, and Zihan Jiang. Mimo-audio: Audio language models are few-shot learners, 2025. URL https://arxiv.org/abs/2512.23808.

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, 2022. URL https://arxiv.org/abs/2203.03022.

Bart van Merriënboer, Vincent Dumoulin, Jenny Hamer, Lauren Harrell, Andrea Burns, and Tom Denton. Perch 2.0: The bittern lesson for bioacoustics, 2026. URL https://arxiv.org/abs/2508.04665.

Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation, 2021. URL https://arxiv.org/abs/2101.00390.

Megan Wei. Syntheory, 2024. URL https://huggingface.co/datasets/meganwei/syntheory.

Megan Wei, Michael Freeman, Chris Donahue, and Chen Sun. Do music generation models encode music theory?, 2024. URL https://arxiv.org/abs/2410.00872.

Ryan Whetten, Titouan Parcollet, Marco Dinarelli, and Yannick Estève. A study of data selection strategies for pre-training self-supervised speech models, 2026. URL https://arxiv.org/abs/2601.20896.

Yang Xiao, Han Yin, Jisheng Bai, and Rohan Kumar Das. Dg-sed: Domain generalization for sound event detection with heterogeneous training data, 2025. URL https://arxiv.org/abs/2407.03654.

Dongchao Yang, Dingdong Wang, Haohan Guo, Xueyuan Chen, Xixin Wu, and Helen Meng. Simplespeech: Towards simple and efficient text-to-speech with scalar latent transformer diffusion models, 2024. URL https://arxiv.org/abs/2406.02328.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. SUPERB: Speech Processing Universal PERformance Benchmark. In *Interspeech 2021*, pp. 1194–1198. ISCA, 2021. doi: 10.21437/Interspeech.2021-1775. URL https://arxiv.org/abs/2105.01051.

Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, Yike Guo, and Wei Xue. Codec does matter: exploring the semantic shortcoming of codec for audio language model. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i24.34761. URL https://doi.org/10.1609/aaai.v39i24.34761.

Zhiqi Yu, Jingjing Li, Zhekai Du, Lei Zhu, and Heng Tao Shen. A comprehensive survey on source-free domain adaptation, 2023. URL https://arxiv.org/abs/2302.11803.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 30:495–507, November 2021. ISSN 2329-9290. doi: 10.1109/TASLP.2021.3129994. URL https://doi.org/10.1109/TASLP.2021.3129994.

Shiqing Zhang, Ruixin Liu, Xin Tao, and Xiaoming Zhao. Deep Cross-Corpus Speech Emotion Recognition: Recent Advances and Perspectives. *Frontiers in Neurorobotics*, 15:784514, November 2021. ISSN 1662-5218. doi: 10.3389/fnbot.2021.784514. URL https://www.frontiersin.org/journals/neurorobotics/articles/10.3389/fnbot.2021.784514/full.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech language models. In *The Twelfth International Conference on Learning Representations*, volume 2024, pp. 31798–31818, 2024. URL https://openreview.net/forum?id=AF9Q8Vip84.

Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, and David Harwath. Bat: learning to reason about spatial sounds with large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, Vienna, Austria, 2024. JMLR.org.

Hao Zhou, Xiaobao Guo, Yuzhe Zhu, and Adams Wai-Kin Kong. Macs: Multi-source audio-to-image generation with contextual significance and semantic alignment, 2025. URL https://arxiv.org/abs/2503.10287.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Emotional voice conversion: Theory, databases and esd. *Speech Communication*, 137:1–18, 2022. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2021.11.006. URL https://www.sciencedirect.com/science/article/pii/S0167639321001308.

Zhiyuan Zhu, Yu Zhang, Wenxiang Guo, Changhao Pan, and Zhou Zhao. Asaudio: A survey of advanced spatial audio research, 2025. URL https://arxiv.org/abs/2508.10924.

Álvaro Vega-Hidalgo, Stefan Kahl, Laurel B. Symes, Viviana Ruiz-Gutiérrez, Ingrid Molina-Mora, Fernando Cediel, Luis Sandoval, and Holger Klinck. A collection of fully-annotated soundscape recordings from neotropical coffee farms in colombia and costa rica, January 2023. URL https://doi.org/10.5281/zenodo.7525349.

APPENDIX A EXPERIMENT DATASETS

Table 1 and Table 2 contain the basic metadata for the datasets we experimented with.

Table 1: Experiment 1 Datasets

Name	Source	Description	Official splits
AISHELL-1	(Bu et al., 2017; OpenSLR, 2017)	Mandarin speech	Yes
AISHELL-3	(Shi et al., 2021; OpenSLR, 2021)	Mandarin speech	Yes
AnomalyMachine-50k	(Anonymous, 2026)	Machine sounds	Yes
BAT	(Zheng et al., 2024; Face, 2024a)	Spatial audio	Yes
BirdSet (HSN)	(Rauch et al., 2025; Clapp et al., 2023)[1]	Bioacoustics	Yes
Clotho-AQA	(Lipping et al., 2022a;b)	Audio Q&A	Yes
CochlScene	(Jeong & Park, 2022b;a)	Acoustic scenes	Yes
CommonVoice	(Ardila et al., 2020; Face, 2024b)	Automatic speech recognition	Yes
DataSED	(Fredianelli et al., 2025a;b)	Real-world environmental noise	No
EigenScape	(Green & Murphy, 2017a;b)	Spatial acoustic scenes	No
Emotional Speech Dataset	(Zhou et al., 2022; Lim, 2024)	English/Mandarin speech	No
Fake-or-Real (original)	(Reimao & Tzerpos, 2019; Abdeldayem, 2024)	Real/synthetic speech	Yes
Fleurs	(Conneau et al., 2022a;b; Google, 2022)	Multilingual speech	Yes
Free Music Archive (small)	(Defferrard et al., 2017a;b)	Music	Yes
FSD50K	(Fonseca et al., 2022; 2020)	Environmental sounds	Yes
LibriSpeech (Multilingual)	(Pratap et al., 2020; OpenSLR, 2020)	Multilingual read speech	Yes
LOCATA	(Evers et al., 2020b;a)	English speech, spatial audio	Yes
MELD	(Poria et al., 2019; Poria, 2018)	Friends dialogue	Yes
MultiVox	(Meza et al., 2025a;c;b)	Singing	No
Nonspeech7k	(Rashid et al., 2023b;a)	Human non-speech sounds	Yes
RAVDESS	(Livingstone & Russo, 2018a;b)	Speech, song	No
SONYC-UST	(Cartwright et al., 2020b;a)	Urban noise	Yes
Spatial LibriSpeech	(Sarabia et al., 2023; Apple, 2023)	Spatial speech	Yes
STARSS23	(Shimada et al., 2023; Politis et al., 2023)	Realistic spatial soundscapes	Yes
Syntheory (Chords)	(Wei et al., 2024; Wei, 2024)	Synthetic MIDI	No
Syntheory (Intervals)	(Wei et al., 2024; Wei, 2024)	Synthetic MIDI	No
Syntheory (Notes)	(Wei et al., 2024; Wei, 2024)	Synthetic MIDI	No
Syntheory (Scales)	(Wei et al., 2024; Wei, 2024)	Synthetic MIDI	No
Syntheory (Simple Progressions)	(Wei et al., 2024; Wei, 2024)	Synthetic MIDI	No
TAU-NIGENS 2021	(Koizumi et al., 2019b;a)	Spatial audio	Yes
TAU Urban Acoustic Scenes 2022	(Heittola et al., 2020; 2022a;b)	Urban sounds	Yes
ToyADMOS (Car)	(Koizumi et al., 2019b;a)	Machine sounds	No
ToyADMOS (Conveyor)	(Koizumi et al., 2019b;a)	Machine sounds	No
ToyADMOS (Train)	(Koizumi et al., 2019b;a)	Machine sounds	No
TUT Acoustic Scenes 2016	(Mesaros et al., 2016a;b; 2017d)	Acoustic scenes	Yes
TUT Acoustic Scenes 2017	(Mesaros et al., 2017a;c;b)	Acoustic scenes	Yes
UrbanSound8k	(Salamon et al., 2014)	Urban sounds	No
VGG-Sound	(Chen et al., 2020; Chen, 2020)	General sound	Yes
VocalSound (16 kHz)	(Gong et al., 2022; Gong, 2022)	Human sounds	Yes
VocalSound (44 kHz)	(Gong et al., 2022; Gong, 2022)	Human sounds	Yes
VoxPopuli (en)	(Wang et al., 2021; Facebook, 2021)	English speech	Yes
WABAD	(Pérez-Granados et al., 2026b;a)	Bioacoustics	No

Table 2: Experiment 2 Datasets

Name	Label	Source	Description	Official splits
AISHELL-3	speaker	(Shi et al., 2021; OpenSLR, 2021)	Mandarin speech	Yes
AnomalyMachine-50k	is anomalous	(Anonymous, 2026)	Machine sounds	Yes
CochlScene	environment	(Jeong & Park, 2022b;a)	Acoustic scenes	Yes
EigenScape	environment	(Green & Murphy, 2017a;b)	Spatial acoustic scenes	No
Emotional Speech Dataset	emotion	(Zhou et al., 2022; Lim, 2024)	English/Mandarin speech	No
Fake-or-Real (original)	is real	(Reimao & Tzerpos, 2019; Abdeldayem, 2024)	Real/synthetic speech	Yes
Fleurs	gender	(Conneau et al., 2022a; Google, 2022)	Multilingual speech	Yes
Free Music Archive (small)	top genre	(Defferrard et al., 2017a;b)	Music	Yes
LOCATA	task	(Evers et al., 2020b;a)	English speech, spatial audio	Yes
MACS	environment	(Zhou et al., 2025; Morato & Mesaros, 2021)	Acoustic scenes	No
MELD	speaker	(Poria et al., 2019; Poria, 2018)	Friends dialogue	Yes
STARSS23	origin	(Shimada et al., 2023; Politis et al., 2023)	Realistic spatial soundscapes	Yes
Syntheory (Chords)	midi program name	(Wei et al., 2024; Wei, 2024)	Synthetic MIDI, music theory	No
Syntheory (Intervals)	midi program name	(Wei et al., 2024; Wei, 2024)	Synthetic MIDI, music theory	No
Syntheory (Notes)	midi program name	(Wei et al., 2024; Wei, 2024)	Synthetic MIDI, music theory	No
Syntheory (Scales)	midi program name	(Wei et al., 2024; Wei, 2024)	Synthetic MIDI, music theory	No
TAU Urban Acoustic Scenes 2022	environment	(Heittola et al., 2020; 2022a;b)	Urban sounds	Yes
ToyADMOS (Car)	is anomalous	(Koizumi et al., 2019b;a)	Machine sounds	No
ToyADMOS (Conveyor)	is anomalous	(Koizumi et al., 2019b;a)	Machine sounds	No
ToyADMOS (Train)	is anomalous	(Koizumi et al., 2019b;a)	Machine sounds	No
TUT Acoustic Scenes 2016	environment	(Mesaros et al., 2016a;b; 2017d)	Acoustic scenes	Yes
VGG-Sound	category	(Chen et al., 2020; Chen, 2020)	General sound	Yes
VocalSound (16 kHz)	sound category	(Gong et al., 2022; Gong, 2022)	Human sounds	Yes
VocalSound (44 kHz)	sound category	(Gong et al., 2022; Gong, 2022)	Human sounds	Yes

[1]Outside of HSN, BirdSet also contains a collection of other soundscapes from Hopping et al. (2022); Álvaro Vega-Hidalgo et al. (2023); Navine et al. (2022); Morfi et al. (2018); Chronister et al. (2022); Kahl et al. (2022a;b)

APPENDIX B EXPERIMENT RESULT TABLES

In this section, we include tables showing numerical results from our experiments.

For experiment 1, Table 3 shows the raw measures of each dataset's unconditional epiplexity. Table 4, Table 5 and Table 6 all refer to the zero-shot experiments with comparing pretraining epiplexity to performance on unseen evaluation datasets. Similarly, Table 7, Table 8 and Table 9 refer to the similar experiments where the pretrained models were finetuned on the evaluation data.

APPENDIX B.1 EXPERIMENT 1

Table 3: Epiplexity Across Tokenizers

	Machine IP	DAC	EnCodec	X-Codec
AISHELL-1	79.133	0.4583	0.4208	2.7960
AISHELL-3	79.130	0.6792	1.6144	15.5588
AnomalyMachine-50k	79.133	4.7253	1.5033	9.5695
BAT	79.70	0.0071	0.3820	0.3865
BirdSet (HSN)	79.133	1.2395	1.4064	4.6513
Clotho-AQA	79.130	0.1832	0.4116	0.1752
CochlScene	79.133	1.7286	1.2985	1.6411
CommonVoice	79.130	0.8353	0.5398	9.4674
DataSED	79.130	0.0008	0.0048	0.0002
EigenScape	79.70	0.0427	0.1031	0.0405
Emotional Speech Dataset	79.133	0.1087	0.0271	3.7232
Fake-or-Real (original)	79.133	0.0151	0.5125	5.7049
Fleurs	79.130	0.5654	0.3931	0.3171
Free Music Archive (small)	79.133	0.0002	0.0930	0.3346
FSD50K	81.100	0.2725	2.2463	12.7050
LibriSpeech (Multilingual)	79.130	1.0091	0.3164	2.4905
LOCATA	81.100	0.0446	0.0805	0.2762
MELD	79.133	0.0205	0.0574	2.5864
MultiVox	81.100	0.0020	0.0135	0.0010
Nonspeech7k	81.100	0.5552	0.4588	3.227
RAVDESS	79.133	0.1765	0.1473	0.6500
SONYC-UST	81.100	0.5040	0.1560	1.2062
Spatial LibriSpeech	81.100	0.1421	1.1783	8.7652
STARSS23	79.70	0.0463	0.0267	0.0017
Syntheory (Chords)	79.133	0.8810	1.0451	4.0357
Syntheory (Intervals)	79.133	6.9062	1.8076	18.2297
Syntheory (Notes)	79.133	0.0987	0.8747	2.8560
Syntheory (Scales)	79.133	2.0172	0.5628	4.9313
Syntheory (Simple Progressions)	79.133	3.7914	0.7117	9.1079
TAU-NIGENS 2021	79.70	0.0282	0.2466	0.3719
TAU Urban Acoustic Scenes 2022	79.70	0.3457	0.7617	12.7887
ToyADMOS (Car)	79.133	0.4465	0.0303	1.7930
ToyADMOS (Conveyor)	79.133	0.5122	1.0571	3.2878
ToyADMOS (Train)	79.133	0.7586	0.4367	3.3794
TUT Acoustic Scenes 2016	79.70	0.1342	0.0335	0.4188
TUT Acoustic Scenes 2017	79.70	0.4028	0.4500	0.5699
UrbanSound8k	79.148	0.0332	0.0973	0.8063
VGG-Sound	81.100	0.0516	0.1399	0.3891
VocalSound (16 kHz)	79.133	0.2103	0.5302	3.1440
VocalSound (44 kHz)	79.133	0.9029	0.4878	3.2416
VoxPopuli (EN)	79.133	0.0913	0.3638	0.5204
WABAD	79.133	0.1453	0.1696	0.3309

Estimated epiplexity (in megabits) of each dataset + tokenizer combination, rounded to the nearest four decimal places.

Table 4: Next Token Prediction Cross Entropy Loss (Zero-shot, DAC)

	AISHELL-1	LibriSpeech (Multilingual)	VocalSound (16 kHz)	Nonspeech7k	BirdSet (HSN)	WABAD	Syntheory (Intervals)	Syntheory (Simple Progressions)	UrbanSound8k	TAU Urban Acoustic Scenes 2022	ToyAdmos (ToyCar)	AnomalyMachine-50k
AISHELL-1	7.26	7.04	7.57	6.89	7.38	7.19	7.32	7.39	6.99	7.35	7.00	7.27
LibriSpeech (Multilingual)	7.59	7.15	7.21	6.68	6.86	7.20	7.27	7.41	7.00	6.60	7.21	7.23
VocalSound (16 kHz)	7.18	6.96	7.36	6.56	6.75	6.92	6.87	6.95	6.72	6.90	6.82	7.10
Nonspeech7k	7.26	7.14	7.20	6.52	6.75	6.92	7.17	7.32	6.97	6.30	6.98	7.23
BirdSet (HSN)	6.95	6.81	7.34	6.54	6.57	6.71	6.71	6.78	6.58	7.11	6.74	6.84
WABAD	6.96	6.98	6.98	6.92	6.58	6.69	6.76	6.81	6.89	7.13	6.73	6.84
Syntheory (Intervals)	7.53	7.46	7.55	7.47	7.10	7.23	6.22	7.25	7.33	7.93	7.06	7.34
Syntheory (Simple Progressions)	7.23	7.28	7.38	7.25	7.07	7.08	6.74	6.69	7.17	7.60	7.02	7.35
UrbanSound8k	8.41	8.06	8.35	7.27	7.98	7.49	7.01	7.11	6.88	6.84	6.87	7.81
TAU Urban Acoustic Scenes 2022	23.59	20.99	22.69	18.65	23.23	23.56	23.46	23.45	21.79	0.20	22.66	23.73
ToyAdmos (ToyCar)	6.98	6.77	7.03	6.59	6.64	6.88	6.78	6.81	6.66	7.05	6.37	7.08
AnomalyMachine-50k	7.73	7.41	7.46	7.32	7.01	7.03	7.43	7.53	7.36	7.87	7.03	6.91

Table 5: Next Token Prediction Cross Entropy Loss (Zero-shot, EnCodec)

	AISHELL-1	LibriSpeech (Multilingual)	VocalSound (16 kHz)	Nonspeech7k	BirdSet (HSN)	WABAD	Syntheory (Intervals)	Syntheory (Simple Progressions)	UrbanSound8k	TAU Urban Acoustic Scenes 2022	ToyAdmos (ToyCar)	AnomalyMachine-50k
AISHELL-1	6.85	8.95	9.76	8.47	7.98	6.64	5.48	6.33	8.55	11.68	7.13	11.13
LibriSpeech (Multilingual)	7.43	7.68	7.91	7.13	4.06	4.94	6.21	6.81	7.67	6.28	5.26	7.96
VocalSound (16 kHz)	8.04	8.36	7.27	6.97	3.86	4.88	5.08	5.75	7.35	9.60	5.38	8.75
Nonspeech7k	8.10	8.60	7.55	6.34	3.19	4.66	4.95	5.22	6.36	9.74	5.64	6.94
BirdSet (HSN)	7.24	8.76	7.78	6.56	1.26	2.64	2.99	3.60	5.01	10.43	3.73	4.30
WABAD	8.09	8.67	7.80	7.15	1.02	2.16	3.09	3.81	5.99	11.51	3.75	5.47
Syntheory (Intervals)	9.97	9.64	10.09	9.13	2.52	5.25	1.00	3.98	7.96	11.05	5.39	9.46
Syntheory (Simple Progressions)	12.98	12.56	11.74	11.35	11.27	10.64	4.99	3.35	11.19	11.45	9.23	13.80
UrbanSound8k	7.93	9.39	8.27	6.69	1.91	3.09	3.87	4.54	4.81	10.64	4.03	4.13
TAU Urban Acoustic Scenes 2022	21.06	18.59	17.63	15.16	16.70	19.65	20.00	20.18	19.09	0.02	19.24	20.06
ToyAdmos (ToyCar)	7.80	9.49	9.39	7.71	2.56	5.57	3.98	4.40	6.89	10.30	2.61	9.21
AnomalyMachine-50k	12.84	12.35	11.52	10.93	2.84	6.07	12.24	12.38	10.63	13.36	7.08	3.58

Table 6: Next Token Prediction Cross Entropy Loss (Zero-shot, X-Codec)

	AISHELL-1	LibriSpeech (Multilingual)	VocalSound (16 kHz)	Nonspeech7k	BirdSet (HSN)	WABAD	Syntheory (Intervals)	Syntheory (Simple Progressions)	UrbanSound8k	TAU Urban Acoustic Scenes 2022	ToyAdmos (ToyCar)	AnomalyMachine-50k
AISHELL-1	6.52	7.11	7.31	7.24	6.60	6.91	6.79	7.12	6.98	8.27	6.58	7.08
LibriSpeech (Multilingual)	6.58	6.52	6.78	6.87	6.33	6.45	6.21	6.64	6.76	6.66	6.13	6.58
VocalSound (16 kHz)	6.90	6.93	6.48	6.45	6.15	6.25	5.79	6.46	6.47	7.23	5.97	6.45
Nonspeech7k	6.80	6.73	6.62	6.28	5.82	6.06	5.75	6.43	6.02	7.35	5.83	6.04
BirdSet (HSN)	6.80	6.91	6.83	6.43	5.48	5.63	5.55	6.25	5.94	7.72	5.48	5.62
WABAD	6.73	6.99	6.76	6.76	5.39	5.42	6.25	6.75	6.13	8.29	5.46	5.56
Syntheory (Intervals)	7.65	7.85	7.70	7.88	7.12	7.17	3.59	6.25	7.29	9.50	6.29	7.01
Syntheory (Simple Progressions)	8.12	8.49	8.06	8.30	7.80	7.86	5.39	5.18	7.83	9.90	6.92	8.00
UrbanSound8k	6.68	6.82	6.96	6.40	5.86	5.90	5.31	6.05	5.71	7.97	5.66	5.79
TAU Urban Acoustic Scenes 2022	14.80	12.89	13.05	12.09	12.96	13.71	13.13	14.59	13.82	4.83	13.57	13.06
ToyAdmos (ToyCar)	7.24	7.59	7.24	7.39	6.14	6.43	6.86	7.15	6.58	8.62	4.79	6.73
AnomalyMachine-50k	8.31	7.85	7.79	7.56	6.98	6.86	7.22	7.44	7.05	9.13	6.33	5.16

Table 7: Next Token Prediction Cross Entropy Loss (Finetuning, DAC)

	AISHELL-1	LibriSpeech (Multilingual)	VocalSound (16 kHz)	Nonspeech7k	BirdSet (HSN)	WABAD	Syntheory (Intervals)	Syntheory (Simple Progressions)	UrbanSound8k	TAU Urban Acoustic Scenes 2022	ToyAdmos (ToyCar)	AnomalyMachine-50k
AISHELL-1	7.27	7.45	7.69	6.75	6.69	6.73	7.28	7.53	6.55	0.18	6.66	7.32
LibriSpeech (Multilingual)	6.94	7.10	7.16	6.56	6.59	6.66	6.48	6.71	6.66	0.19	6.41	6.91
VocalSound (16 kHz)	6.97	7.84	7.33	6.63	6.60	6.68	6.52	6.72	6.66	0.19	6.40	6.83
Nonspeech7k	6.87	7.12	6.93	6.48	6.57	6.69	6.52	6.66	6.63	0.19	6.39	6.76
BirdSet (HSN)	6.85	7.05	6.84	6.44	6.57	6.69	6.52	6.67	6.55	0.20	6.41	6.78
WABAD	6.86	6.68	6.80	6.53	6.58	6.69	6.60	6.67	6.63	0.24	6.42	6.77
Syntheory (Intervals)	6.94	7.47	7.04	6.35	6.55	6.65	6.22	6.86	6.68	0.19	6.42	6.77
Syntheory (Simple Progressions)	6.90	7.21	7.04	6.43	6.58	6.70	6.45	6.69	6.65	0.19	6.41	6.80
UrbanSound8k	6.86	6.86	6.84	6.43	6.61	6.70	6.53	6.67	6.76	0.19	6.43	6.78
TAU Urban Acoustic Scenes 2022	6.84	7.11	6.98	6.75	6.85	6.88	6.56	6.71	6.62	0.20	6.43	6.78
ToyAdmos (ToyCar)	6.89	7.31	7.13	6.65	6.57	6.69	6.48	6.64	6.66	0.19	6.36	6.77
AnomalyMachine-50k	7.02	7.21	7.12	6.37	6.59	6.68	6.63	6.78	6.61	0.18	6.46	6.93

Table 8: Next Token Prediction Cross Entropy Loss (Finetuning, EnCodec)

	AISHELL-1	LibriSpeech (Multilingual)	VocalSound (16 kHz)	Nonspeech7k	BirdSet (HSN)	WABAD	Syntheory (Intervals)	Syntheory (Simple Progressions)	UrbanSound8k	TAU Urban Acoustic Scenes 2022	ToyAdmos (ToyCar)	AnomalyMachine-50k
AISHELL-1	6.82	7.58	7.31	6.28	1.29	2.21	1.43	3.38	4.66	0.02	2.62	4.04
LibriSpeech (Multilingual)	6.67	7.68	7.31	6.21	1.11	2.17	1.33	3.24	4.70	0.02	2.61	4.35
VocalSound (16 kHz)	6.58	7.35	7.37	6.29	1.29	2.24	1.32	3.15	4.66	0.03	2.62	4.13
Nonspeech7k	6.56	7.28	6.99	6.21	1.15	2.11	1.31	3.27	4.69	0.02	2.59	3.35
BirdSet (HSN)	6.53	7.15	6.81	5.87	1.25	2.09	1.30	3.34	4.68	0.03	2.62	3.73
WABAD	6.47	7.09	6.69	6.01	1.22	2.13	1.33	3.36	4.64	0.02	2.60	3.62
Syntheory (Intervals)	6.53	7.29	6.83	5.94	1.16	2.19	1.00	2.98	4.70	0.02	2.59	3.40
Syntheory (Simple Progressions)	6.51	7.22	6.95	5.85	1.34	2.41	1.24	3.34	4.74	0.03	2.60	3.57
UrbanSound8k	6.56	7.25	6.93	6.05	1.16	2.12	1.30	3.09	4.73	0.02	2.59	3.83
TAU Urban Acoustic Scenes 2022	6.55	7.16	6.76	5.98	1.05	2.06	1.30	3.06	4.61	0.02	2.58	3.46
ToyAdmos (ToyCar)	6.50	7.26	6.75	5.86	1.10	2.19	1.40	3.31	4.56	0.03	2.61	3.61
AnomalyMachine-50k	6.59	7.34	7.10	6.31	1.53	2.25	1.48	3.39	4.82	0.02	2.59	3.63

Table 9: Next Token Prediction Cross Entropy Loss (Finetuning, X-Codec)

	AISHELL-1	LibriSpeech (Multilingual)	VocalSound (16 kHz)	Nonspeech7k	BirdSet (HSN)	WABAD	Syntheory (Intervals)	Syntheory (Simple Progressions)	UrbanSound8k	TAU Urban Acoustic Scenes 2022	ToyAdmos (ToyCar)	AnomalyMachine-50k
AISHELL-1	6.54	6.71	6.65	6.55	5.69	5.72	4.36	5.57	5.95	4.51	5.03	5.46
LibriSpeech (Multilingual)	6.29	6.52	6.41	6.32	5.55	5.57	4.20	5.35	5.84	4.42	4.87	5.09
VocalSound (16 kHz)	6.26	6.46	6.49	6.26	5.54	5.55	4.13	5.20	5.83	4.24	4.81	5.05
Nonspeech7k	6.15	6.33	6.32	6.25	5.53	5.52	4.18	5.18	5.83	4.17	4.78	4.91
BirdSet (HSN)	6.13	6.26	6.32	6.06	5.48	5.47	4.04	5.12	5.74	4.04	4.76	4.88
WABAD	6.07	6.25	6.16	5.92	5.48	5.42	4.13	5.00	5.70	4.03	4.77	4.82
Syntheory (Intervals)	6.35	6.56	6.50	6.32	5.70	5.60	3.60	5.57	5.89	4.36	4.85	5.03
Syntheory (Simple Progressions)	6.18	6.39	6.36	6.18	5.71	5.65	3.78	5.19	5.85	4.14	4.78	4.90
UrbanSound8k	6.14	6.38	6.36	6.03	5.49	5.45	4.06	5.19	5.71	4.20	4.77	4.86
TAU Urban Acoustic Scenes 2022	6.31	6.56	6.33	6.21	5.56	5.54	3.89	5.06	5.74	4.83	4.79	4.97
ToyAdmos (ToyCar)	6.09	6.30	6.33	6.10	5.61	5.53	4.14	5.07	5.78	3.98	4.79	4.86
AnomalyMachine-50k	6.23	6.43	6.40	6.10	5.59	5.56	4.19	5.31	5.83	4.18	4.88	5.16

APPENDIX B.2 EXPERIMENT 2

For data in experiment 2, conditional epiplexity measurements can be found in Table 10. The F1-scores and accuracies from the classification performance tests can be found in Table 11 and Table 12, respectively.

Table 10: Conditional Epiplexity Across Tokenizers

Dataset	Label	Machine IP	DAC	EnCodec	X-Codec
AISHELL-3	speaker	79.130	19	8	1
AnomalyMachine-50k	is anomalous	79.133	347	5	13
CochlScene	environment	79.133	2926	184	801
EigenScape	environment	79.130	71	37	72
Emotional Speech Dataset	emotion	79.133	113	85	38
Fake-or-Real (original)	is real	79.133	18	112	15
Fleurs	gender	79.130	7	2	36
Free Music Archive (small)	top genre	79.133	1	11	9
LOCATA	task	79.70	3	6	4
MACS	environment	79.133	934	54	504
MELD	speaker	79.133	17	39	174
STARSS23	origin	79.70	29	58	27
Syntheory (Chords)	midi program name	79.70	5	9727	5151
Syntheory (Intervals)	midi program name	79.70	18549	15534	21756
Syntheory (Notes)	midi program name	79.70	1	4	1
Syntheory (Scales)	midi program name	79.70	6487	11769	11291
TAU Urban Acoustic Scenes 2022	environment	79.70	255	396	27
ToyADMOS (ToyCar)	is anomalous	79.148	2	193	252
ToyADMOS (ToyConveyor)	is anomalous	79.148	35	26	68
ToyADMOS (ToyTrain)	is anomalous	79.148	91	126	37
TUT Acoustic Scenes 2016	environment	79.70	260	74	124
VGG-Sound	category	79.133	1	1	1
VocalSound (16 kHz)	sound category	79.133	198	605	112
VocalSound (44 kHz)	sound category	79.133	87	1286	1

Estimated conditional epiplexity (in bits) of each label + tokenizer combination, rounded to the nearest positive integer.

Table 11: Weighted Average F1-scores Across Tokenizers

Dataset	DAC	EnCodec	X-Codec
AISHELL-3	0.0999	0.1251	0.1340
AnomalyMachine-50k	0.7031	0.9395	0.9620
CochlScene	0.3243	0.4484	0.5142
EigenScape	0.3011	0.4643	0.3514
Emotional Speech Dataset	0.4462	0.6909	0.6705
Fake-or-Real (original)	0.9508	0.9027	0.9974
Fleurs	0.6903	0.8936	0.9969
Free Music Archive (small)	0.1925	0.2811	0.3113
LOCATA	0.2430	0.4720	0.2847
MACS	0.3626	0.4968	0.5272
MELD	0.0002	0.0009	0.0009
STARSS23	0.7618	0.7390	0.8047
Syntheory (Chords)	0.6827	0.9831	0.9710
Syntheory (Intervals)	0.9231	0.9879	0.9902
Syntheory (Notes)	0.1267	0.6611	0.4945
Syntheory (Scales)	0.9822	0.9997	0.9984
TAU Urban Acoustic Scenes 2022	0.2896	0.2947	0.3444
ToyADMOS (ToyCar)	0.9008	0.9569	0.9610
ToyADMOS (ToyConveyor)	0.9557	0.9787	0.9978
ToyADMOS (ToyTrain)	0.8645	0.8893	0.9803
TUT Acoustic Scenes 2016	0.3008	0.4813	0.4932
VGG-Sound	0.0038	0.0213	0.0305
VocalSound (16 kHz)	0.4098	0.6942	0.6705
VocalSound (44 kHz)	0.4117	0.6760	0.6749

Table 12: Classification Accuracies Across Tokenizers

Dataset	DAC	EnCodec	X-Codec
AISHELL-3	0.1253	0.1508	0.1648
AnomalyMachine-50k	0.7013	0.9396	0.9621
CochlScene	0.3213	0.4467	0.5124
EigenScape	0.3219	0.4865	0.4094
Emotional Speech Dataset	0.4597	0.6923	0.6677
Fake-or-Real (original)	0.9508	0.9027	0.9974
Fleurs	0.7342	0.8980	0.9969
Free Music Archive (small)	0.1950	0.2800	0.3050
LOCATA	0.2516	0.4732	0.2988
MACS	0.3576	0.5049	0.5278
MELD	0.0016	0.0020	0.0008
STARSS23	0.7883	0.7768	0.8316
Syntheory (Chords)	0.6811	0.9830	0.9709
Syntheory (Intervals)	0.9230	0.9879	0.9902
Syntheory (Notes)	0.1373	0.6640	0.4975
Syntheory (Scales)	0.9822	0.9997	0.9984
TAU Urban Acoustic Scenes 2022	0.2991	0.2977	0.3625
ToyADMOS (ToyCar)	0.9035	0.9573	0.9615
ToyADMOS (ToyConveyor)	0.9565	0.9786	0.9978
ToyADMOS (ToyTrain)	0.8857	0.8903	0.9805
TUT Acoustic Scenes 2016	0.3125	0.4813	0.4978
VGG-Sound	0.0054	0.0270	0.0323
VocalSound (16 kHz)	0.4129	0.6933	0.6730
VocalSound (44 kHz)	0.4129	0.6755	0.6752

APPENDIX C TOKENIZERS

We use three audio tokenizers in our experiment: DAC, EnCodec and X-Codec, all of which were designed for speech, music, and general audio. The following table details the architecture of the encoder, decoder, and quantizer of each tokenizer (Table 13).

Table 13: Tokenizer architecture

Name	Encoder	Decoder	Quantization algorithm
DAC	CNN	CNN	RVQ
EnCodec	CNN + RNN	CNN	RVQ
X-Codec	CNN	CNN	RVQ

CNN and RNN refer to convolutional neural network and recurrent neural network, respectively
RVQ and FSQ refer to residual vector quantization and finite scalar quantization, respectively.

Each tokenizer name in the above table is a link to a page where the tokenizer models are published and documented.

The following table gives the tokenizers' specifications (Table 14).

Table 14: Tokenizer specifications

Name	Sample rate (kHz)	Framerate (frames/sec)	Codebooks	Vocabulary size
DAC	16	50	12	1024
EnCodec	24	75	1	1024
X-Codec	16	50	8	1024

APPENDIX D KDE PLOTS

Figure 5 and Figure 6 show the density of epiplexity and conditional epiplexity for tokenized audio data. Generally we find in both cases the distributions to be unimodal with a strong right skew, with most audio data having roughly the same epiplexity.

Figure 7 shows the densities for classification performance metrics.

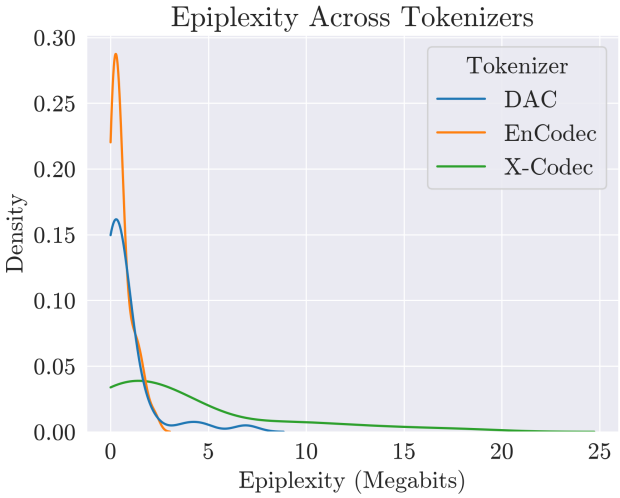


Figure 5: KDE plots of epiplexity by tokenizer, estimated from the data in Table 3. We clip the density at 0 since epiplexity cannot be negative.

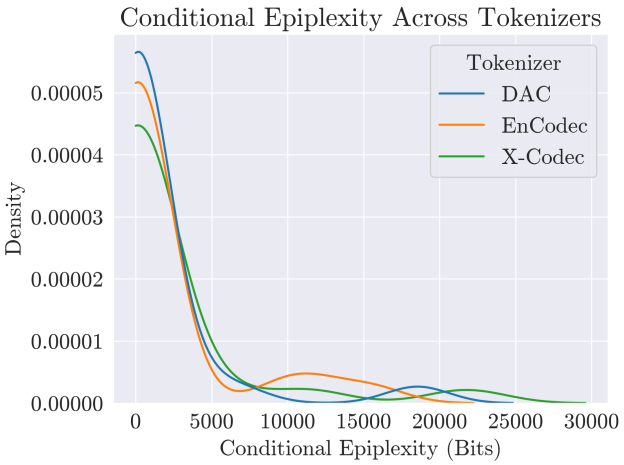


Figure 6: KDE plots of conditional epiplexity by tokenizer, estimated from the data in Table 10. We clip the sample space to $(0, \infty)$ since epiplexity cannot be negative.

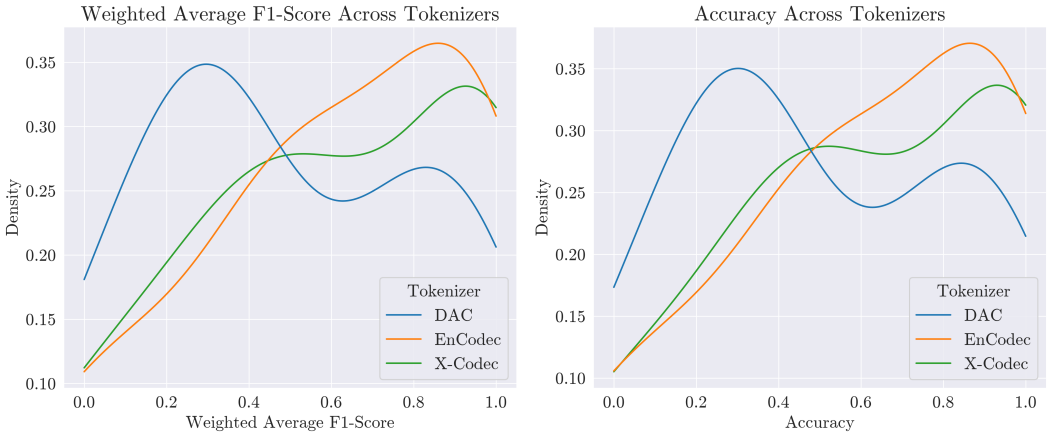


Figure 7: KDE plots of classification performance metrics, estimated from the data in Table 11 and Table 12. We clip the sample space to $(0, 1)$.

APPENDIX E PREQUENTIAL CODING

APPENDIX E.1 UNCONDITIONAL

We begin with a randomly initialized model P_0 and train until we reach our token limit. Let P_i denote the model after training on the first i portions of the data. At each step, the current model is used to encode the next portion of data, after which the model is updated. The total prequential code length can be written as

$$L_{\text{preq}}(Z_{0:M}) = \sum_{i=0}^{M-1} \log_2 \frac{1}{P_i(Z_i)}, \tag{5}$$

where Z_i denotes the data batch encoded at step i and $Z_{0:M}$ denotes the entire sequence of batches.

The final model P_M can encode the same data with a code length given by

$$L(Z_{0:M} \mid P_M) = \sum_{i=0}^{M-1} \log_2 \frac{1}{P_M(Z_i)}. \tag{6}$$

Following the approximation used in Finzi et al. (2026), the description length of the learned model can be estimated from the difference between the prequential code length and the code length obtained using the final model:

$$|P| \approx L_{\text{preq}}(Z_{0:M}) - L(Z_{0:M} \mid P_M). \tag{7}$$

Substituting Equations 5 and 6, we obtain

$$|P| \approx |P|_{\text{preq}} = \sum_{i=0}^{M-1} \left[\log_2 \frac{1}{P_i(Z_i)} - \log_2 \frac{1}{P_M(Z_i)} \right]. \tag{8}$$

Appendix E.1.1 Conditional

As in Experiment 1, we begin with a randomly initialized model P_0 and train until we reach our token limit. Let $B_0, B_1, \ldots, B_{M-1}$ denote the ordered training minibatches, where each B_j contains audio-label pairs (X_i, Y_i). Model P_j predicts the labels in B_j before being updated on that batch to produce P_{j+1}. Define

$$L(M_j, B_j) = \sum_{(X_i, Y_i) \in B_j} -\log P_j(Y_i \mid X_i).$$

The conditional label code through step M is therefore $\sum_{j=0}^{M-1} L(M_j, B_j)$. Under distributed training, the label losses and counts are aggregated across all workers before being added to this code.

At each evaluation step, we evaluate the current raw model P_M on a subset E of the tokenized evaluation split $X_{\mathcal{T}}^{ev}$:

$$\widehat{H}_{\text{raw}}(Y \mid X_{\mathcal{T}}; P_M) = \frac{1}{\#E} \sum_{(X_i^E, Y_i^E) \in E} -\log P_M(Y_i^E \mid X_i^E).$$

This gives the conditional MDL estimate

$$\sum_{j=0}^{M-1} L(P_j, B_j) - n_M \cdot \widehat{H}_{\text{raw}}(Y \mid X_{\mathcal{T}}; P_U), \tag{9}$$

where $n_M = \sum_{j=0}^{M-1} \#B_j$ is the number of labels seen up to step M.

Appendix F Potential Impact of the Token Budget, Scaling Laws and Other Modalities

We were interested in comparing the epiplexity of audio data to that of data of other modalities. As we mentioned a few times in the introduction, Finzi et al. (2026) tested epiplexity over a few modalities using data from text (OpenWebText (Gokaslan & Cohen, 2019), Lichess Chess Puzzle dataset (Schwarzschild et al., 2021)), images (CIFAR-5M (Nakkiran et al., 2021)), and videos. To see if we could replicate the results with our setup and compare audio to the other modalities, we set out to replicate the scaling law work in Finzi et al. (2026, Section 6.3).

To do this, we measured unconditional epiplexity using our standard sweep config from Section 3 over OpenWebText, Chess, CIFAR-5M, and FSD50k (tokenized with EnCodec and X-Codec). Based on our prior experiment results, we decided to remove the DAC tokenizer from the experiment. Note that because we are using the standard experiment config, we use approximately 31 million tokens instead of the 1 billion tokens used in the original work.

We also use the scaling law formulas from Finzi et al. (2026, Section 6.3) for estimating epiplexity with larger compute budgets. Much of our graphing code is also pulled from the work used to generate Finzi et al. (2026, Figure 9) as we are trying to replicate the work. As such we also try to match how OpenWebText, Chess, and CIFAR-5M were tokenized in the original study but measure epiplexity over our PyTorch pipeline instead.

What we find differs from the findings of Finzi et al. (2026, Section 6.3). In our ordering of Figure 8, OpenWebText has about as much epiplexity as compute gets asymptotically larger as CIFAR-5M,

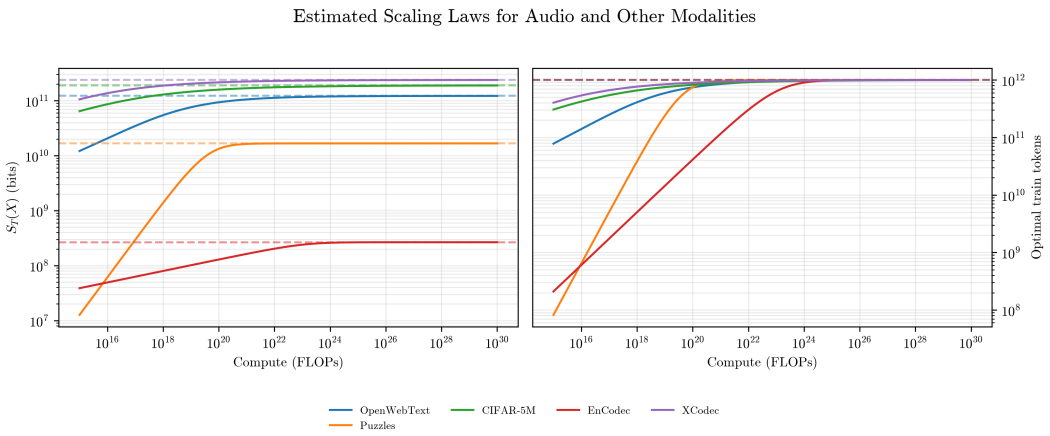


Figure 8: Scaling law estimates based on training runs of 31M tokens.

which is strikingly not what the other paper had shown. It is likely that with our restricted token budget, our fitted parameters for the scaling law model are inaccurate at higher token counts, making scaling less valuable. For one, OpenWebText should be greatly surpassing these results. Otherwise, it seems that at lower epiplexity values, X-Codec, CIFAR-5M and OpenWebText have very similar amounts of epiplexity.

Scaling law work should run larger compute budgets to estimate the scaling laws so that a better test can be done to compare the modalities to each other.

APPENDIX G NOTATION

Table 15 highlights key notation used throughout the paper. Feel free to use this in reference when reading our work.

Table 15: Notation Guide

Notation	Section	Meaning
$\mathbb{E}_X[\cdot]$	3.1.1	Expected value with respect to a random variable X
$\#\cdot$	3.1.1	Set cardinality
$P^\star$	3.1.1	Two-part code length minimizing program
$S_T(\cdot)$	3.1.1	Epiplexity with compute budget T
$H_T(\cdot)$	3.1.1	Time-bounded entropy with compute budget T
$\lvert\cdot\rvert$	3.1.1	Model description length (MDL)
$\lvert\cdot\rvert_{\text{preq}}$	3.1.1	Prequential estimate for MDL
$\widehat{H}(\cdot)$	3.1.1	Estimator for entropy code length
$\mathcal{T}$	4.1.2	Audio tokenizer
$X_{\mathcal{T}}$	3.1.2	Audio dataset X tokenized with $\mathcal{T}$
X^{ev}	3.1.1	Evaluation split of a dataset X

Section refers to the section in which the notation was first defined.