AUDIOWEB-365M: A WEB-SCALE CORPUS AND SCALING STUDIES FOR CONTRASTIVE LANGUAGE-AUDIO PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in contrastive audio-language models can be attributed mainly to the scale and quality of their training data. Larger training datasets expose models to a broader range of sounds and corresponding audio captions. To exploit this, we collect 1.47 billion audio samples from academic datasets, internet datasets, web collections, Common Crawl, and YouTube, and process them with a multi-stage curation pipeline. Furthermore, we generate short and long captions for samples in our corpus that lack them. We use our dataset to study the scaling behavior of contrastive audio-language models based on the NaFlex architecture. Our experiments identify trade-offs between model size and training data for a given compute budget. Moreover, our models outperform prior audio-text models of comparable size on the MAEB and HEAR benchmarks. As a result, our models can serve as a drop-in replacement for CLAP in downstream applications. We will publicly release our dataset, captions, model checkpoints, and code.

1 INTRODUCTION

Recent advances in audio-language learning have shown that large-scale contrastive pretraining results in strong downstream performance (Radford et al., 2021; Wu et al., 2023; Elizalde et al., 2023). In contrastive language-audio pretraining (CLAP), models learn to map audio recordings and their corresponding text descriptions into a shared embedding space. This facilitates tasks such as text-audio retrieval or audio classification. In addition to that, CLAP models support text-to-audio generation (Liu et al., 2023; Hung et al., 2026) and audio source separation (Ma et al., 2024a). CLAP embeddings have also proven useful for data curation, including for detecting dataset overlap (Wijngaard et al., 2025), filtering audio-text pairs (Haji-Ali et al., 2026; Valle et al., 2025), selecting synthetic training examples (Ghosh et al., 2025), and refining audio captions (Bai et al., 2025).

CLAP models have improved as training datasets grow. LAION-CLAP used 633,526 audio-text pairs (Wu et al., 2023), M2D-CLAP used 2.6 million pairs (Niizumi et al., 2025), MS-CLAP23 increased this to 4.6 million pairs (Elizalde et al., 2024), and SLAP was trained with 109 million pairs (Mei et al., 2026). These models report increasingly strong classification and retrieval results, while also differing in model size. As training datasets grow, choosing an appropriate model size becomes increasingly important for making the best use of available data and compute.

This question has been studied through scaling laws, which show that increasing both training data and model size generally improves downstream performance (Kaplan et al., 2020; Hoffmann et al., 2022). In contrastive language-image pretraining (CLIP), scaling training data and model size has been a major source of these performance gains (Cherti et al., 2023). Recent work measured this in greater detail to compare models and datasets and estimate how many training samples to use for a given compute budget (Nezhurina et al., 2026). Moreover, data selection has been shown to be crucial (Gadre et al., 2023).

The availability of large-scale audio-text data remains a bottleneck for studying the scaling of audio models. While vision and language pretraining have scaled to open, billion-scale web-scraped datasets with transparent curation pipelines (Schuhmann et al., 2022; Gadre et al., 2023), public audio-text collections are orders of magnitude smaller. The largest open caption datasets for general audio reach about 50 million text-audio pairs (Hochlehnert et al., 2026; Haji-Ali et al., 2026).

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.

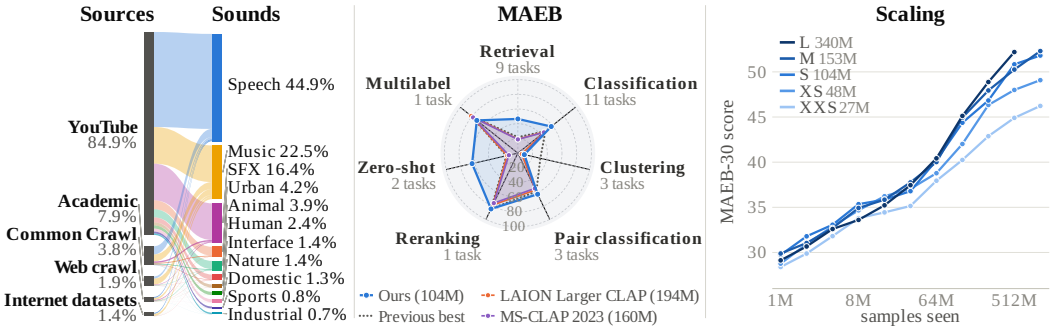


Figure 1: Our training dataset consists of many sources, spanning diverse sound categories, and our models achieve strong performance that improves with training scale. **Left**: the 365M training audio samples from source group to predicted sound category using our trained taxonomy classifier (see Section 3.3). Ribbon width is proportional to audio samples. **Middle**: MAEB (Assadi et al., 2026) of our S model, LAION-Larger-CLAP (Wu et al., 2023) and MS-CLAP-23 (Elizalde et al., 2024). Dotted: best of ten models (<500M parameters). **Right**: Scaling MAEB-30 against samples seen.

AF-Next used around 108 million samples from 1 million hours of audio (Ghosh et al., 2026), and MiDasheng trained on 1.1 million hours (Dinkel et al., 2025), both largely drawn from public datasets. We show that by combining public audio datasets with four other groups of sources, such as YouTube and audio derived from Common Crawl, this amount of data can be increased even more. In contrast, recent audio foundation models report training on private datasets of more than 100M hours, which are not publicly available (Team et al., 2025; Qwen Team, 2026).

In this work, we collected AudioWeb-365M, a training corpus of 365 million text-audio pairs from an initial collection of 1.47 billion audio samples drawn from public audio datasets, web data, Common Crawl, and YouTube (Figure 1). The dataset is processed with an extensive pipeline of multiple steps, including cleaning, deduplication, annotation, selection and captioning. From a raw pool of 1.47 billion audio samples, the final training corpus is a 365 million dataset of audio-text pairs, with audio of at most 30 seconds. We will release the indices, captions, metadata, and curation pipeline of this full training corpus to facilitate future research.

We use this corpus to train contrastive language-audio models. The audio encoder is a NaFlex audio transformer (Tschannen et al., 2025), which processes variable-length audio. The text encoder is loosely based on ModernBERT (Warner et al., 2025), but uses a SwiGLU + RMSNorm instead of GeGLU + LayerNorm. We train both encoders from scratch instead of initializing from a pretrained model. The model architecture scales more easily for testing and it allows for better display of scaling laws. Our experiments test 5 different sizes of models, from $\pm$27M to $\pm$340M parameters, and plot the results without any confounding effects from previously acquired knowledge.

For our 153M-parameter model, the average score on MAEB-30 (Assadi et al., 2026) increases by 2.2 points when the seen samples are doubled between 1M and 1024M samples. Training only on audio of at most 30 seconds raises the score by 4.1 points and requires 3.5 times less compute than training on audio of up to 120 seconds. Our best model M reaches 52.3%, compared to 32.7% for LAION-CLAP and 44.1% for GLAP, the best other contrastive audio-text model we evaluated, with 5.6 times more parameters. The model is also the best model trained from scratch on the MAEB-30 leaderboard, where other models rely on backbones whose training data is not released. On HEAR (Turian et al., 2022), our S to L models score higher than all other audio-text models.

Overall, our contribution can be summarized as follows:

- We introduce a curated audio-text dataset with 365 million audio-text pairs, selected from 1.47 billion web and dataset audio samples through a fully documented pipeline.
- We present a scaling study of contrastive audio-text training, showing a continuing gain of 2.2 MAEB-30 points per doubling of samples seen over three orders of magnitude.
- We train a 153M-parameter CLAP model that reaches 52.3% on MAEB-30, outperforming models of similar size such as M2D-CLAP (38.0%) and LAION-CLAP (32.7%).

Table 1: **AudioWeb compared to existing audio-text datasets**, including standalone dataset releases and the training data of audio encoders and audio-language models. Our AudioWeb dataset is sourced from a range of online platforms and websites, and from existing data collections, and datasets (details are provided in Appendix H). †Dasheng trains on audio only.

Dataset	Sources	Audio	Captions	Duration (h)
Clotho (Drossos et al., 2020)	1	6K	30K	44
AudioCaps (Kim et al., 2019)	1	36K	57K	101
WavCaps (Mei et al., 2024)	4	403K	403K	±7.6K
LAION-Audio-630K (Wu et al., 2023)	11	634K	634K	4.3K
Auto-ACD (Sun et al., 2024)	2	1.8M	1.8M	5.2K
SALMONN-2 (Yang et al., 2026b)	21	8.9M	8.9M	18.2K
Audio Flamingo 3 (Goel et al., 2025)	46	26.8M	26.8M	≥54.4K
AutoReCap-XL (Haji-Ali et al., 2026)	6	47M	47M	124K
LAION-BVD (Hochlehnert et al., 2026)	1	55M	55M	56K
Dasheng mix (Dinkel et al., 2024)	4	97M	–†	272K
SLAP (MovieGen Audio) (Mei et al., 2026)	1	109M	109M	±1M
Bagpiper (Tian et al., 2026)	33	157M	159M	±421K
GLAP mix (Dinkel et al., 2026)	12	445M	445M	435K
AudioWeb-500M (ours)	**553**	**500M**	**966M**	**3.34M**
AudioWeb-365M (ours)	467	365M	703M	1.38M

2 Related work

Audio-text datasets. Earlier datasets such as Clotho (Drossos et al., 2020) and AudioCaps (Kim et al., 2019) relied on a single source, often YouTube or Freesound. A recent trend shows that models are increasingly using combinations of sources (Wijngaard et al., 2025). Earlier examples of encoder models trained on a mixture of data include MS-CLAP trained on four and LAION-CLAP on 11 sources (Elizalde et al., 2023; Wu et al., 2023). Later, state-of-the-art models used 20+ sources (see Table 1 for an overview). The dataset presented in this work, AudioWeb-365M, uses a combination of 467 sources, 338 of which are either part of or derived from academic work. Other models also test varying with captioning style, similar to our work, for example AudioVerse and ACAVCaps use a combination of long and short captions, including other variants (Niu et al., 2026; Xu et al., 2026). CapTalk and LP-MusicCaps do similar splits for speech and music, respectively (Su et al., 2026; Doh et al., 2023).

Audio scaling laws. Researchers often release audio-language models in multiple sizes, where based on the scaling law properties, larger models tend to perform better than smaller models. Examples in audio encoder models include AuroLA (Xu et al., 2026) testing on 92M, 322M and 418M, and Dasheng consisting of 3 model sizes, 86M, 0.6B and 1.2B parameters (Dinkel et al., 2024). Similarly, in the audio-language learning field, models are often released in multiple sizes, for example Qwen2.5Omni with a 3B and 7B model and MOSS-Audio with a 4B and 8B model (Qwen et al., 2025; Yang et al., 2026a). While scaling model sizes, it is also important to take into account scaling the amount of data. For audio-language models, SODA (Manakul et al., 2026) shows that the compute-optimal data amount needed grows faster than the amount of training parameters. PE-AV (Vyas et al., 2026) separately compares how to scale data and model sizes in audio-visual models. Tseng et al. (2026) train both contrastive and captioning models in pairs in 0.4M, 1M, 4M, and 10M audio-language pairs, and show that on smaller scales having a pretrained model gains the most improvements, but on larger scale training from scratch helps the most. To the best of our knowledge, this work is the first to present how a fixed compute budget should be divided between model sizes and data samples seen for contrastive audio-language models.

Audio taxonomy. The Audioset ontology is a taxonomy based on 632 classes, from which the 527 labels of Audioset dataset originate. These 632 originated from a list of more than 3000 candidates (Gemmeke et al., 2017). SALT later proposed an extension to 734 sound labels (Stamatiadis et al., 2024). Universal Category System is an initiative for standardization of a set of categories of sound, with the January 2024 version spanning 753 subcategories under 82 categories (Tim Nielsen et al., 2024). This work uses a taxonomy based on 4,416 nodes consisting of 11 root nodes, and a maximum depth of 5. This taxonomy is the basis of a classifier on the set of leaf nodes of 3,718 classes to cluster a raw pool of audio data.

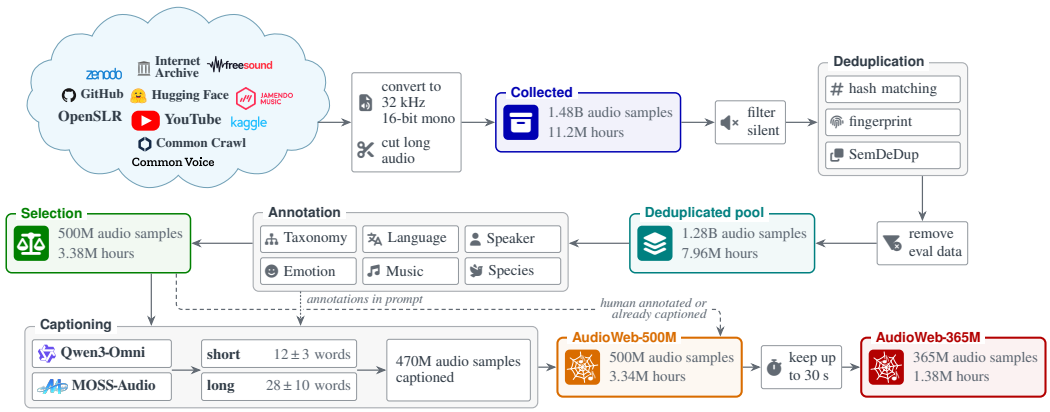


Figure 2: **Overview of the AudioWeb-365M dataset curation pipeline.** We collected 1.48B audio samples from online platforms, of which 1.47B entered deduplication. Removing silent audio, duplicates (including SemDeDup (Abbas et al., 2023)), and evaluation data leaves 1.28B audio samples. We then selected 500M audio samples based on categorical taxonomy annotations. The same annotations were used to prompt Qwen3-Omni (Xu et al., 2025b) and MOSS-Audio (Yang et al., 2026a) to generate short and long captions for 470M audio samples without textual descriptions, resulting in 500M captioned audio samples in total (AudioWeb-500M). The 365M audio samples that are up to 30 seconds long form AudioWeb-365M.

3 AudioWeb-365M Dataset

3.1 Collection

For training contrastive language-audio models, large-scale and diverse datasets are required. For the training of the model presented in this paper, a collection is aggregated over public datasets, YouTube videos, Common Crawl, among others. The final dataset contains 338 academic datasets, each tied to a published paper, preprint, challenge, or other academic release (Appendix A). Most of these datasets include valid descriptions that are also reused in the final dataset, with even a fraction of the full data being human-annotated. Table 1 compares the final dataset AudioWeb-365M with existing audio-text datasets, and Figure 2 lists the processing steps outlined in this section.

YouTube contains the majority of our dataset. All audio collected on YouTube originates from Creative Commons-licensed videos with explicit reuse permissions. This source produces 1.18B audio samples extracted from approximately 33.61M full videos. YouTube audio is extracted across the full duration of each video, with a mean source duration of 13.1 minutes. Each video can therefore contribute many audio samples. The audio is segmented at silence boundaries into audio samples between 5 and 120 seconds.

Common Crawl consists of many websites scraped from the open web (Common Crawl Foundation, 2024). A scraper extracts all mentions of audio files with valid audio extensions across all crawls through CC-MAIN-2026-08. The first extraction wave covered 35 crawl directories from 2021 to 2025 and found about 3.3 billion audio links, of which about 423M are unique. The text surrounding the audio on the website is used as a context for subsequent caption generation.

3.2 Deduplication, signal checks, and benchmark screening

The pipeline begins with removal of all empty or partial audios, or audios that cannot be read or decoded. Audios are kept if they are between 0.5 and 120 seconds, with an extra check if the audio contains at least 0.5 seconds of audible content. Exact matching groups audio samples by audio duration and the MD5 hash of uncompressed PCM bytes. This step removes 151.1M audio samples. Removing duplicate and silent recordings saves compute (Lee et al., 2022) and prevents representation collapse on repeated items (Hernandez et al., 2022).

Near-duplicate deduplication target recordings with minor acoustic variations. Two complementary strategies are used: acoustic matching through signal-processing fingerprints directly after exact matching, and semantic matching through self-supervised audio embeddings after the signal gate.

Acoustic matching uses Olaf fingerprints (Six, 2023) and removes 1.2M audio samples. Semantic near-duplicates are found using SemDeDup (Abbas et al., 2023) based on Dasheng embeddings (Dinkel et al., 2024). This stage removes 18.2M audio samples. Appendix A.4 describes both implementations in more detail.

Benchmark screening removes known evaluation material. Source identifiers, URLs, video time windows, and test-split metadata are compared with the records of MAEB (Assadi et al., 2026), HEAR (Turian et al., 2022) and MSEB (Heigold et al., 2026). Appendix A.5 reports the benchmark deduplication.

3.3 Clustering and coverage-aware selection

After filtering and deduplication, an annotation pipeline gathers information regarding the audio. These serve as filters in manual selection and as extra content for captioning models. Using the addition of this extra information during captioning helps improve the caption quality (see Appendix C.2). A prominent part of our filtering is based on a taxonomy that reduces common content while protecting rare predicted categories. A model is trained with a linear output layer of 3,718 leaf buckets to perform this classification. In addition to the taxonomy filtering, the metadata collection involves emotion classification, language, speaker age and gender, pitch, music genre and style tags, and species predictions.

The selector combines a binary validity gate, a retention fraction, a cluster-density factor, and a mixture factor. Retention floors protect rare predicted categories, languages, expressive emotions, and species, while cluster-density and mixture factors reduce common content. A deterministic hash determines which candidates are kept. Appendix B.3 gives the full calculation and thresholds.

The curated pool is 75.3% speech and 11.7% music by predicted label. After selection and exact-size trimming, 500M base rows are obtained. The final training corpus keeps the 364,955,979 selected audio samples of at most 30 seconds (Appendix F). This changes the mixture to 44.9% speech, 22.5% music, 16.4% designed sound effects, and 16.2% other classes.

3.4 Caption quality and production supervision

Caption quality is evaluated with both model judgments and human listening. Seven captioning systems produced short and long descriptions for a shared sample. The two MOSS-Audio-8B configurations Thinking and Non-Thinking ranked highest (Yang et al., 2026a), followed by Qwen3-Omni (Xu et al., 2025b). Appendix C reports the model comparison and human listening study.

The captioners Qwen3-Omni-30B-A3B-Instruct and MOSS-Audio-8B-Thinking (Xu et al., 2025b; Yang et al., 2026a) both generate a short and a long description for each audio, with 12 and 28 words on average. In total, this stage captions 470,040,216 audio samples with both caption lengths (93.99% of full AudioWeb-500M).

4 Model, training, and evaluation

Model. Our models are trained from scratch. The **audio tower** is based on a NaFlex vision transformer (Dehghani et al., 2023; Tschannen et al., 2025) over log-mel patches. The audio input (32 kHz) is converted to Mel spectrograms with 128 bins between 50 Hz and 16 kHz, and a 320-sample hop. Non-overlapping 16×16 patches give 50 tokens per second or eight tokens per 0.16 seconds. We use RoPE (Su et al., 2024) to keep the coordinates of each patch. The transformer blocks use RMSNorm (Zhang & Sennrich, 2019) and SwiGLU (Shazeer, 2020). One register token is added, and attention pooling followed by two linear layers with a GELU activation between them produces the clip embedding. The **text tower** is a bidirectional transformer with RoPE, RMSNorm, SwiGLU, and query-key normalization. It has a 50,260-entry byte-pair vocabulary and a 256-token context. The trained models span five sizes, between 27.3M (XXS) and 340.1M trainable parameters (Table 3). Smaller models enable efficient data filtering.

Training uses a standard CLIP objective on audio and text (van den Oord et al., 2019). The learnable logit scale is initialized to $1/0.07$ and has an upper bound of 300. Audio is assigned to a batch of nearest length, with buckets of 6.88, 8.96, 12.96, 22.56, and 30.08 seconds, to reduce padding and speed up training. A GPU batch holds audio samples of one bucket under a budget of 49,152 audio tokens, so a 6.88 second short-clip batch has 136 audio samples and a 30 second batch has 32 per

Table 2: MAEB benchmark results. Parameter counts (Params) are totals over both towers, including embeddings. Results for our models use the last checkpoint of each size's WSD v3 run: 1024M samples for XXS, XS, S, M and 512M samples for L, evaluated with un-normalised audio embeddings and with all-caps text inputs lowercased. We report the mean performance over 30 tasks, domain-specific (18 speech, 6 music, 6 sound tasks), and task-specific results: classification (12, including one multilabel), zero-shot classification (2), pair classification (3), clustering (3), and retrieval (10, including one reranking task). All scores are in %. Models are grouped by size, the best value in each column within a group is bolded.

		MAEB								
			Domain			Task type				
Model	**Params**	**Mean**	**Speech**	**Music**	**Sound**	**Clf.**	**ZS**	**Pair**	**Clust.**	**Retr.**
Large models (≥1B parameters)										
LCO-Embedding-Omni-7B	8.9B	**53.5**	54.8	**73.7**	**29.7**	57.9	64.5	**67.3**	1.7	57.6
e5-omni-7B	8.9B	52.5	54.1	71.8	28.1	53.6	**68.8**	65.5	1.8	**59.1**
BidirLM-Omni-2.5B	2.4B	52.4	**58.6**	66.0	20.0	**59.9**	65.8	66.7	6.2	50.2
LCO-Embedding-Omni-3B	4.7B	52.3	54.0	72.0	27.3	56.1	62.2	66.7	1.3	56.6
jina-embeddings-v5-omni-small	1.6B	50.4	53.7	68.0	23.1	53.4	62.2	65.4	6.1	53.2
e5-omni-3B	4.7B	48.0	50.2	64.7	24.7	48.5	64.4	59.5	1.9	54.5
OmniEmbed-v0.1	8.9B	43.9	43.5	61.7	27.2	50.7	30.2	58.9	3.2	46.1
fusion-embedding-2-2b-preview	2.8B	40.2	38.8	60.6	23.9	48.4	55.2	61.5	3.7	31.9
Qwen2-Audio-7B	7.0B	34.5	32.6	61.9	12.9	56.2	12.4	56.9	12.7	12.8
EBind (full)	1.8B	32.8	27.6	62.4	18.4	41.4	16.3	52.3	**26.0**	21.8
Qwen2.5-Omni-3B	5.5B	23.4	23.9	41.3	3.9	31.8	14.3	51.2	0.6	13.7
Small models (<1B parameters)										
jina-embeddings-v5-omni-nano	986M	51.8	**53.5**	73.8	24.7	56.8	60.2	**65.5**	6.9	**53.4**
GLAP	855M	44.1	42.4	69.6	23.4	46.0	54.5	56.4	5.9	47.3
M2D-CLAP	199M	38.0	30.0	69.7	29.9	50.4	16.0	53.5	**31.4**	24.7
TP-CLAP	198M	37.2	28.6	67.0	**33.3**	47.8	18.2	54.4	23.0	27.4
CLAP-IPA small	174M	36.5	41.4	50.5	8.1	34.7	53.5	52.9	0.6	41.2
AudioVerse CLAP	144M	36.0	27.8	64.2	32.7	46.4	13.8	52.4	25.8	26.2
LAION Larger CLAP	194M	35.2	26.8	66.4	29.1	49.5	14.9	55.6	5.4	24.9
MS-CLAP 2023	160M	34.4	25.1	67.3	29.3	48.0	12.6	53.3	8.8	24.4
LAION-CLAP HTSAT unfused	153M	32.7	24.2	63.6	27.1	44.5	11.8	52.5	9.0	23.8
MuQ-MuLan-large	630M	29.5	24.1	65.1	10.5	42.5	12.7	52.6	4.1	18.1
wav2clip	163M	28.1	23.9	58.5	10.2	40.3	10.7	53.1	6.1	16.0
SpeechT5 multimodal	298M	26.4	25.7	46.3	8.4	37.9	16.0	57.9	1.1	12.7
Ours XXS	27M	46.2	46.0	69.2	23.9	50.9	59.9	61.4	3.1	46.2
Ours XS	48M	49.1	48.8	72.1	26.9	56.2	57.8	62.0	3.6	48.6
Ours S	104M	51.8	52.4	74.2	27.6	59.2	63.9	62.0	9.0	50.3
Ours M	153M	**52.3**	52.4	**76.1**	28.1	59.8	56.5	61.8	9.3	52.6
Ours L	340M	52.2	52.5	**76.1**	27.5	**59.9**	**64.6**	61.5	9.1	50.7

GPU (rounded down to be divisible by eight). Bucket probabilities are set so every bucket supplies about 20% of the samples, which corresponds to the training distribution. Audio samples are padded and masked inside their bucket.

All runs use AdamW with $\beta = (0.9, 0.98)$, weight decay 0.2, warmup-stable or cosine learning rate decay, gradient clipping at 1.0, and bfloat16 autocast. Peak learning rate, warmup, and batch size are set per data budget (see Appendix E for details). The caption is randomly short or long for those who have both. The training compute C is counted as three times the forward cost of both towers, adapted to the distribution mix of the dataset. Table 3 shows that around 80% of the computation is caused by the audio encoder, due to the quadratic cost of attention on longer inputs.

Evaluation: We evaluated our models on the following benchmarks. Massive Audio Embedding Benchmark (MAEB-30) (Assadi et al., 2026) is a collection of benchmarks that covers 30 individual tasks across speech, music, and environmental sounds, and spans classification, retrieval, and clustering tasks. The score is the unweighted mean over tasks. A task with several subsets, such as the languages of FLEURS, first averages within task. We group the tasks into speech (18 benchmarks), music (6 benchmarks), and sound (6 benchmarks) (Table 11). Audio samples are truncated to 30 seconds for evaluation to match the length of our training samples.

In addition to MAEB-30, we evaluate our models on the Holistic Evaluation of Audio Representations benchmark (HEAR) (Turian et al., 2022). It covers speech, music, and environmental sounds. Since this benchmark is designed for the evaluation of audio representations only, not for text-audio evaluations, we use it as a quality measure for our audio embeddings. Additionally, we evaluate on a third, more speech-oriented, Massive Sound Embedding Benchmark (MSEB) (Heigold et al., 2026).

5 Scaling Analysis

Before scaling, we fixed the training setup in ablations of our cosine decay model (Appendix F). Training only on audio of at most 30 seconds scores 4.1 MAEB-30 points higher than training on audio of up to 120 seconds while using 3.5 times less compute, and the caption length has no effect beyond run-to-run variation.

Table 3: **Model tiers for scaling study:** Hidden dimension d, parameter counts, and forward-pass compute per encoder. Audio parameters include the MAP pool and projection layers. Text parameters include the pool and output projection but *exclude* the $50{,}260 \times d_t$ token-embedding table. Total parameters refer to the full model size, including embeddings. GFLOPs are forward-pass FLOPs per clip at the training length distribution.

		Parameters (M)			GFLOPs / clip		
Tier	d	Audio	Text	Total	Audio	Text	Total
XXS	224	8.3	7.8	27.3	21.6	4.6	26.1
XS	320	16.6	15.7	48.3	38.0	8.8	46.8
S	512	41.8	36.4	103.9	83.8	19.6	103.4
M	640	64.4	56.2	152.8	123.1	29.8	152.9
L	960	166.5	125.4	340.1	292.8	65.1	357.9

Setup: Our scaling study includes five model sizes ranging from 27.3M to 340.1M parameters (Table 3). We train these models on our 365M dataset, with samples-seen scales $D = \{1M,\ 2M,\ 4M,\ 8M,\ 16M,\ 32M,\ 64M,\ 128M,\ 256M,\ 512M,\ 1024M\}$, a total of 11 data scales. We do this by sampling the training batches with replacement within each pass over the complete training pool. We train our runs with a warmup–stable–decay learning rate scheduling. To allow enough optimization steps for smaller data scales, we split the runs into one short (up to 32M samples seen) and one long (up to 1B samples seen) training run. For more details on the hyperparameters used for these runs, refer to Appendix E.

For the formulation of our scaling study, we follow Nezhurina et al. (2026), adapting $\mathcal{L} = aC^b$, with C being compute and $\mathcal{L}$ refering to the mean MAEB-30 score or subset of its 30 benchmarks, based on small compute scale saturation (B_C). This results in the following requirements for our scaling functions:

$$\begin{aligned} \mathcal{L}(C) &> 0 \quad \text{(strictly positive)}, \\ \lim_{C\to\infty} \mathcal{L}(C) &= E \quad \text{(irreducible error)}, \\ \frac{d\mathcal{L}}{dC} &< 0 \quad \text{(monotonic decrease)}. \end{aligned} \tag{1}$$

After taking these three requirements into account, we get the functional form:

$$\mathcal{L}(C) = A_c \cdot (C + B_c)^{-\alpha_c} + E_c, \quad \text{for } \alpha_c > 0, \tag{2}$$

with subscript C specifying the compute dependency.

Scaling Results: We fit our experimental data to Eq. 2 as shown in Figure 3. Our architecture shows a clear scaling trend on our 365M audio-text dataset. The fitted curve through the Pareto frontier shows that with each doubling of compute, MAEB-30 error (1 − mean score) is reduced by $\approx 4\%$ or $\approx +1.9\%$ points at the largest budget. We also tested the quality of our scaling prediction by excluding all runs with $C \geq 10^{11}$ GFLOPs. We are able to accurately predict the five runs on the compute-Pareto-frontier with $C \geq 10^{11}$ GFLOPs (see Figure 10).

Figure 3 shows that the smallest XXS model size is the computationally optimal choice up to 256M seen samples. Only at the end do we see that other model sizes can compete with it. Dataset scaling holds up to at least 256M for our smallest model size.

When splitting the MAEB-30 error score into its three domains (speech, music, and sounds) and examining their scaling behavior individually, we observe significant differences among them (Figure 4). Although sound-related benchmarks saturate after 1.5×10^{10} GFLOPs, speech and music do not reach the compute range tested with $C_{limit} = 5.5 \times 10^{11}$. We observe that there is only little

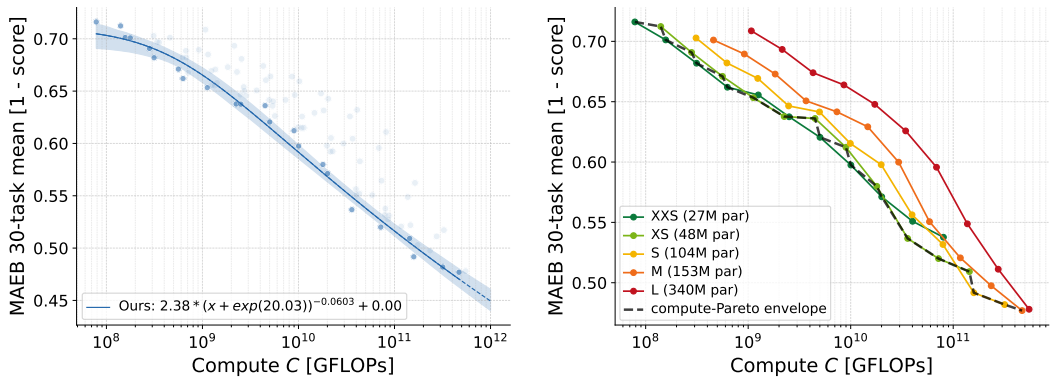


Figure 3: **Scaling behavior on our AudioWeb-365M dataset.** MAEB-30 error rate (1 − mean score) per compute C (GFLOPs). **Left**: the compute-Pareto-optimal runs (solid points) of all warmup-stable-decay and cosine learning-rate-scheduled runs, with our fitted scaling function and its 95% bootstrap CI (1000 resamples). With $\alpha = -0.0603$, the error reduces by 4.1% for each doubling of compute ($\approx +1.9$ mean MAEB-30 points at the largest budget). **Right**: each model size, from two WSD runs per size with individual cooldown phases, and the compute-Pareto envelope. Models XXS, XS, and S show slight saturation, while larger models need more compute to match the performance of smaller ones.

improvement on speech benchmarks below $\approx 10^9$ when scaling compute, while no such warmup is needed for music and sounds (Figure 4).

6 Benchmark Results

The following section discusses the benchmark results of the best and last checkpoints in our scaling study for all model sizes. Model XXS, XS, S, M were trained on 1,024M data samples while model L was only trained on 512M data samples, making it comparable to the compute used for the best model M. We compare all our models to the leading models on the MAEB-30 leaderboard while distinguishing between smaller models below 1B parameters and larger models far surpassing our size with more than 1B parameters. Table 2 shows this in terms of mean MAEB benchmark score and split into domain and task type for each benchmark.

Our models perform strongly across all MAEB domains. Looking just at the mean MAEB performance, we see that our models surpass others several times their size. Figure 5 plots the mean MAEB-30 and HEAR scores against model size. On both benchmarks, our models score higher than other audio-text models of similar size. Especially our smaller models show extraordinary performance for their parameter count. Models XXS–M are therefore only beaten by models with at least 8× more parameters. With a mean MAEB-30

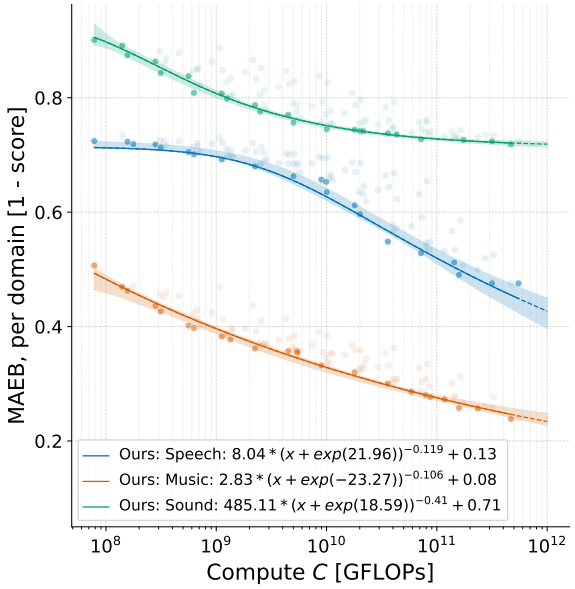


Figure 4: **Domain-specific scaling on our 365M dataset.** MAEB-30 error rate (1 − mean score) per compute C (GFLOPs) per model size and domain. Solid points are the compute-Pareto-optimal WSD and cosine LR runs used for the fit, the band the 95% bootstrap CI (1000 resamples). Speech and music scale well, sound saturates around $C = 1.5 \times 10^{10}$.

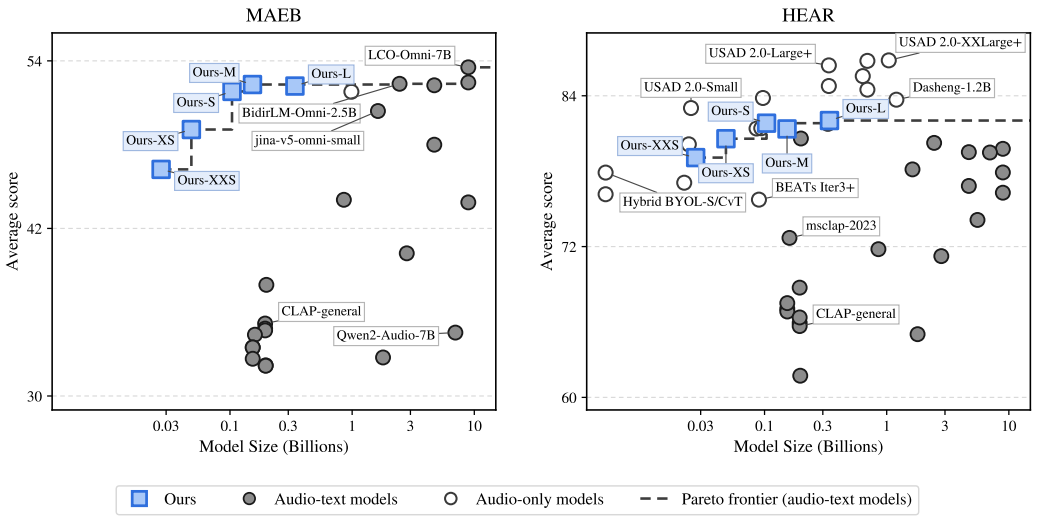


Figure 5: Embedding model performance by size. Average score across the individual tasks of MAEB (all 30 tasks) and HEAR (17 tasks). Blue squares are our own runs. Filled circles represent either audio-text multimodal models, and unfilled ones are audio-only models with no text component.

score of 52.3% model M scores fourth compared on the official leader board, rivalling multi-modal foundation models with billions of parameters.

We achieve strong performance on speech and sound tasks while beating the best model on music with three of our models (Table 2). Our training data holds much more speech than any other domain. Rebalancing it evenly between speech, music, and other classes lowers the MAEB-30 score at 256M seen samples from 43.0% to 41.7% (Appendix F). The original data distribution is therefore kept. On a per benchmark task breakdown, our models perform well on all of them except clustering which is known to misalign with other benchmarks as seen by the scores from other top-performing models. While our models outperform other CLAP models on most benchmarks, one thing to highlight is the gap in the two zero-shot tasks of MAEB-30. This gap of $\approx$45 percentage points to LAION Larger CLAP and M2D-CLAP demonstrates that our models, trained on AudioWeb-365M, are able to generalize significantly better to a wider variety of domains. For the individual results for all 30 MAEB benchmarks, refer to Table 11 in the Appendix.

The appendix reports two additional benchmarks. On HEAR (Table 13), our S, M and L models score between 81.4 to 82.0, higher than all other audio-text models (at most 80.7). Audio-only encoders such as USAD 2.0 score up to 86.8. On MSEB (Table 14), our S and L models score 29.2 and 29.1, second only to M2D-CLAP (30.3) and higher than all multimodal models with billions of parameters (at most 26.1).

7 Conclusion

Our AudioWeb-365M dataset shows that public datasets and web audio can support contrastive audio-language learning at scale with large-scale collection, filtering, deduplication, and captioning pipelines. Our focus on smaller models is backed by the observation that it is computationally more effective to train smaller models on more samples than to increase the parameter count. Training on audio of at most 30 seconds also produces a better model with less than one-third of the compute. Downstream performance follows a fitted scaling trend, with continued gains in speech and music. These results motivate further data scaling. In particular, training larger models on more than 1B samples would show whether the scaling trend continues. The planned release of dataset indices, captions, model checkpoints, and code will allow other researchers to build on both the corpus and the models. Our CLAP models can replace other CLAP models and are efficient to run, making them especially useful for data curation and text-to-audio generation.

ETHICS STATEMENT

The dataset introduced in this work includes speech, music, public sound datasets, and web recordings with varying source terms. Almost all data used in this work is licensed under Creative Commons. Our release consists of source URLs for the audio data, metadata, captions, model checkpoints, and code. It does not redistribute audio.

REPRODUCIBILITY STATEMENT

This work reports the data collection and filtering steps, captioning protocols, model configurations, and code. Appendix H summarizes the stored audio samples, and separates sources that have a citable publication from sources that do not. The selection follows from the frozen configuration in Appendix B.3 and hashed clip identifiers. Figures are built from frozen count tables and per-task result files. We will release clip metadata, with download instructions, filenames and indices, metadata, captions, model checkpoints, training configurations, and the code.

AI USE STATEMENT

In this work, we used generative AI tools for coding the analysis, for building figures, and for polishing the writing of this paper. Generative models were also part of the method. Two audio-language models were used for captioning the audio corpus, and a language model assigned captions to taxonomy leaves when the taxonomy and its classifier were built. Moreover, an audio foundation model judged caption quality. The paper describes these uses, their evaluation, and their measured errors. We reviewed all AI-assisted work. We did not use generative AI tools to formulate research questions or hypotheses. We take responsibility for the final content of this work, including any text produced with the aid of generative AI.

REFERENCES

Amro Kamal Mohamed Abbas, Kushal Tirumala, Daniel Simig, Surya Ganguli, and Ari S. Morcos. SemDeDup: Data-efficient learning at web-scale through semantic deduplication. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023.

Pablo Alonso-Jiménez, Xavier Serra, and Dmitry Bogdanov. Efficient Supervised Training of Audio Transformers for Music Representation Learning, arXiv preprint arXiv:2309.16418, September 2023.

Adnan El Assadi, Isaac Chung, Chenghao Xiao, Roman Solomatin, Animesh Jha, Rahul Chand, Silky Singh, Kaitlyn Wang, Ali Sartaz Khan, Marc Moussa Nasser, Sufen Fong, Pengfei He, Alan Xiao, Ayush Sunil Munot, Aditya Shrivastava, Artem Gazizov, Niklas Muennighoff, and Kenneth Enevoldsen. MAEB: Massive Audio Embedding Benchmark, arXiv preprint arXiv:2602.16008, February 2026.

Jisheng Bai, Haohe Liu, Mou Wang, Dongyuan Shi, Wenwu Wang, Mark D Plumbley, Woon-Seng Gan, and Jianfeng Chen. AudioSetCaps: An Enriched Audio-Caption Dataset Using Automated Generation Pipeline with Large Audio and Language Models. *IEEE Transactions on Audio, Speech and Language Processing*, 33:2817–2829, 2025. doi: 10.1109/TASLPRO.2025.3583354.

Felix Burkhardt, Johannes Wagner, Hagen Wierstorf, Florian Eyben, and Björn Schuller. Speech-Based Age and Gender Prediction with Transformers. In *Speech Communication; 15th ITG Conference*, pp. 46–50. VDE, 2023.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible Scaling Laws for Contrastive Language-Image Learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00276.

Common Crawl Foundation. Common Crawl, 2024.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, and Others.

Patch N'pack: NaViT, a Vision Transformer for Any Aspect Ratio and Resolution. *Advances in Neural Information Processing Systems*, 36:2252–2274, 2023. doi: 10.52202/075280-0106.

Heinrich Dinkel, Zhiyong Yan, Yongqing Wang, Junbo Zhang, Yujun Wang, and Bin Wang. Scaling up Masked Audio Encoder Learning for General Audio Classification. In *Interspeech 2024*, pp. 547–551, 2024. doi: 10.21437/Interspeech.2024-246.

Heinrich Dinkel, Gang Li, Jizhong Liu, Jian Luan, Yadong Niu, Xingwei Sun, Tianzi Wang, Qiyang Xiao, Junbo Zhang, and Jiahao Zhou. MiDashengLM: Efficient Audio Understanding with General Audio Captions, arXiv preprint arXiv:2508.03983, 2025.

Heinrich Dinkel, Zhiyong Yan, Tianzi Wang, Yongqing Wang, Xingwei Sun, Yadong Niu, Jizhong Liu, Gang Li, Junbo Zhang, and Jian Luan. GLAP: General Contrastive Audio-Text Pretraining Across Domains and Languages. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 14737–14741. IEEE, 2026.

Seungheon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. LP-MusicCaps: LLM-Based Pseudo Music Captioning. In *24th International Society for Music Information Retrieval Conference*, pp. 409–416. International Society for Music Information Retrieval Conference, 2023. doi: 10.48550/arxiv.2307.16372.

Tulsee Doshi. Introducing Gemini 3.6 Flash, 3.5 Flash-Lite, and 3.5 Flash Cyber. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-6-flash-3-5-flash-lite-3-5-flash-cyber/, July 2026.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An Audio Captioning Dataset. In *ICASSP*, pp. 736–740, May 2020. doi: 10.1109/ICASSP40776.2020.9052990.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP Learning Audio Concepts from Natural Language Supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, June 2023. doi: 10.1109/icassp49357.2023.10095889.

Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. Natural Language Supervision for General-Purpose Audio Representations. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 336–340, Seoul, Korea, Republic of, April 2024. IEEE. ISBN 979-8-3503-4485-1. doi: 10.1109/ICASSP48485.2024.10448504.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, and Others. DataComp: In Search of the Next Generation of Multimodal Datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023. doi: 10.48550/arXiv.2304.14108.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. In *Proc. IEEE ICASSP 2017*, pp. 776–780, New Orleans, LA, 2017. IEEE. doi: 10.1109/ICASSP.2017.7952261.

Sreyan Ghosh, Sonal Kumar, Zhifeng Kong, Rafael Valle, Bryan Catanzaro, and Dinesh Manocha. Synthio: Augmenting Small-Scale Audio Classification Datasets with Synthetic Data. In *International Conference on Learning Representations*, volume 2025, pp. 66689–66711, 2025.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang-gil Lee, Jaehyeon Kim, Aya Aljafari, and Others. Audio Flamingo Next: Next-Generation Open Audio-Language Models for Speech, Sound, and Music, arXiv preprint arXiv:2604.10905, 2026.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio Flamingo 3: Advancing Audio Intelligence With Fully Open Large Audio Language Models. In *Neural Information Processing Systems (NeurIPS)*, July 2025. doi: 10.48550/arXiv.2507.08128.

Moayed Haji-Ali, Willi Menapace, Aliaksandr Siarohin, Guha Balakrishnan, and Vicente Ordonez. Taming Data and Transformers for Audio Generation. *International Journal of Computer Vision*, 134(3):87, 2026. doi: 10.1007/s11263-025-02632-y.

Georg Heigold, Ehsan Variani, Tom Bagby, Cyril Allauzen, Ji Ma, Shankar Kumar, and Michael D Riley. Massive Sound Embedding Benchmark (mseb). *Advances in Neural Information Processing Systems*, 38, 2026.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, Scott Johnston, Ben Mann, Chris Olah, Catherine Olsson, Dario Amodei, Nicholas Joseph, Jared Kaplan, and Sam McCandlish. Scaling Laws and Interpretability of Learning from Repeated Data, arXiv preprint arXiv:2205.10487, May 2022.

Andreas Hochlehnert, Marianna Nezhurina, Mehdi Cherti, Andrej Radonjic, Thaddäus Wiedemer, Christoph Schuhmann, Romain Beaumont, Wieland Brendel, Bernhard Schölkopf, A. Sophia Koepke, Jenia Jitsev, and Matthias Bethge. LAION-BVD: A 10-Million-Hour Open Video Dataset for Multimodal Pre-training, arXiv preprint arXiv:2608.24845, August 2026.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, et al. Training Compute-Optimal Large Language Models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Nips '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 978-1-7138-7108-8.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Zadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. TangoFlux: Super Fast and Faithful Text to Audio Generation with Flow Matching and Clap-Ranked Preference Optimization. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 150793–150816, 2026.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models, arXiv preprint arXiv:2001.08361, 2020.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating Captions for Audios in The Wild. *Proceedings of the Association for Computational Linguistics*, pp. 119–132, 2019. doi: 10.18653/v1/N19-1011.

Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. CREPE: A Convolutional Representation for Pitch Estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 161–165. IEEE, 2018. doi: 10.1109/ICASSP.2018.8461329.

Zhifeng Kong, Sang-gil Lee, Jaehyeon Kim, Boxin Wang, Zihan Liu, Sungwon Kim, Yang Chen, Arushi Goel, Rajarshi Roy, Wenliang Dai, Zhuolin Yang, Yangyi Chen, Dongfu Jiang, Sreyan Ghosh, Tuomas Rintamaki, Andrew Tao, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Unified Audio Intelligence Without Regressing on Text Intelligence, arXiv preprint arXiv:2607.05196, July 2026.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating Training Data Makes Language Models Better. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.577.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo P. Mandic, Wenwu Wang, and Mark D. Plumbley. AudioLDM: Text-to-Audio Generation With Latent Diffusion Models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023,*

Honolulu, Hawaii, USA, volume 202 of *Proceedings of Machine Learning Research*, pp. 21450–21474. PMLR, 2023.

Hao Ma, Zhiyuan Peng, Xu Li, Mingjie Shao, Xixin Wu, and Ju Liu. CLAPSep: Leveraging Contrastive Pre-Trained Model for Multi-Modal Query-Conditioned Target Sound Extraction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:4945–4960, 2024a. ISSN 2329-9290, 2329-9304. doi: 10.1109/TASLP.2024.3497586.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. Emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15747–15760, 2024b. doi: 10.18653/v1/2024.findings-acl.931.

Potsawee Manakul, Woody Haosheng Gan, Martijn Bartelds, Guangzhi Sun, William Held, and Diyi Yang. Scaling Open Discrete Audio Foundation Models with Interleaved Semantic, Acoustic, and Text Tokens, arXiv preprint arXiv:2602.16687, 2026.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354, 2024. doi: 10.1109/TASLP.2024.3419446.

Xinhao Mei, Gael Le Lan, Haohe Liu, Zhaoheng Ni, Varun Nagaraja, Yang Liu, Yangyang Shi, and Vikas Chandra. SLAP: Scalable Language-Audio Pretraining with Variable-Duration Audio and Multi-Objective Training. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 14517–14521. IEEE, 2026.

Marianna Nezhurina, Tomer Porian, Giovanni Puccetti, Tommie Kerssies, Romain Beaumont, Mehdi Cherti, and Jenia Jitsev. Scaling Laws for Robust Comparison of Open Foundation Language-Vision Models and Datasets. In *Advances in Neural Information Processing Systems*, volume 38, pp. 153979–154018, 2026.

Daisuke Niizumi, Daiki Takeuchi, Masahiro Yasuda, Binh Thien Nguyen, Yasunori Ohishi, and Noboru Harada. M2D-CLAP: Exploring General-purpose Audio-Language Representations Beyond CLAP, arXiv preprint arXiv:2503.22104, 2025.

Yadong Niu, Tianzi Wang, Heinrich Dinkel, Xingwei Sun, Jiahao Zhou, Gang Li, Jizhong Liu, Junbo Zhang, and Jian Luan. ACAVCaps: Enabling Large-Scale Training for Fine-Grained and Diverse Audio Understanding. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 15347–15351. IEEE, 2026.

Nvidia, Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, Danial Mohseni Taheri, Ilia Karmanov, Guilin Liu, Jarno Seppanen, Arushi Goel, Mike Ranzinger, Greg Heinrich, Guo Chen, Lukas Voegtle, Philipp Fischer, Timo Roman, Karan Sapra, Collin McCarthy, Shaokun Zhang, et al. Nemotron 3 Nano Omni: Efficient and Open Multimodal Intelligence, arXiv preprint arXiv:2604.24954, May 2026.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, and Others. Scaling Speech Technology to 1,000+ Languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, et al. Qwen2.5 Technical Report, arXiv preprint arXiv:2412.15115, January 2025.

Qwen Team. Qwen3.5-Omni Technical Report, arXiv preprint arXiv:2604.15804, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, pp. 8748–8763. PmLR, 2021. doi: 10.48550/arxiv.2103.00020.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and Others. Laion-5b: An Open Large-Scale Dataset for Training Next Generation Image-Text Models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. doi: 10.52202/068431-1833.

Noam Shazeer. GLU Variants Improve Transformer, arXiv preprint arXiv:2002.05202, February 2020.

Joren Six. Olaf: A Lightweight, Portable Audio Search System. *Journal of Open Source Software*, 8(87), 2023. ISSN 2475-9066. doi: 10.21105/joss.05459.

Paraskevas Stamatiadis, Michel Olvera, and Slim Essid. SALT: Standardized Audio Event Label Taxonomy. In *Detection and Classification of Acoustic Scenes and Events 2024*, Tokyo, October 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced Transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Xiaosu Su, Zihan Sun, Peilei Jia, and Jun Gao. CapTalk: Unified Voice Design for Single-Utterance and Dialogue Speech Generation, arXiv preprint arXiv:2604.08363, April 2026.

Luoyi Sun, Xuenan Xu, Mengyue Wu, and Weidi Xie. Auto-ACD: A Large-scale Dataset for Audio-Language Representation Learning. In Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (eds.), *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pp. 5025–5034. ACM, 2024. doi: 10.1145/3664647.3681472.

Xiaomi LLM-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, et al. MiMo-Audio: Audio Language Models are Few-Shot Learners, arXiv preprint arXiv:2512.23808, December 2025.

Jinchuan Tian, Haoran Wang, Bo-Hao Su, Chien-yu Huang, Qingzheng Wang, Jiatong Shi, William Chen, Xun Gong, Siddhant Arora, Chin-Jou Li, Masao Someki, Takashi Maekaku, Yusuke Shinohara, Jin Sakuma, Keita Goto, Chao-Han Huck Yang, and Shinji Watanabe. Bagpiper: Solving Open-Ended Audio Tasks Via Rich Captions. In *Third Conference on Language Modeling*, 2026.

Tim Nielsen, Justin Drury, and Kai Paquin. Universal Category System, January 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features, arXiv preprint arXiv:2502.14786, February 2025.

Wei-Cheng Tseng, Xuanru Zhou, Mingyue Huo, Yiwen Shao, Hao Zhang, and Dong Yu. Revisiting Audio-Language Pretraining for Learning General-Purpose Audio Representation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 34244–34263, 2026. doi: 10.18653/v1/2026.acl-long.1581.

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, et al. HEAR: Holistic Evaluation of Audio Representations. In *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, pp. 125–145. PMLR, July 2022.

Rafael Valle, Rohan Badlani, Zhifeng Kong, Sang-gil Lee, Arushi Goel, Sungwon Kim, Joao Santos, Shuqi Dai, Siddharth Gururani, Aya Aljafari, and Others. Fugatto 1: Foundational Generative Audio Transformer Opus 1. In *International Conference on Learning Representations*, volume 2025, pp. 4325–4353, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding, arXiv preprint arXiv:1807.03748, January 2019.

Bart van Merriënboer, Vincent Dumoulin, Jenny Hamer, Lauren Harrell, Andrea Burns, and Tom Denton. Perch 2.0: The Bittern Lesson for Bioacoustics. *arXiv Preprint arXiv:2508.04665*, 2025.

Apoorv Vyas, Heng-Jui Chang, Cheng-Fu Yang, Po-Yao Huang, Luya Gao, Julius Richter, Sanyuan Chen, Matthew Le, Piotr Dollár, Christoph Feichtenhofer, and Others. Pushing the Frontier of Audiovisual Perception with Large-Scale Multimodal Correspondence Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30172–30182, 2026.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and Others. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2526–2547, 2025. doi: 10.18653/v1/2025.acl-long.127.

Gijs Wijngaard, Elia Formisano, Michele Esposito, and Michel Dumontier. Audio-Language Datasets of Scenes and Events: A Survey. *IEEE Access: Practical Innovations, Open Solutions*, 13:20328–20360, January 2025. doi: 10.1109/ACCESS.2025.3534621.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-Scale Contrastive Language-Audio Pretraining With Feature Fusion and Keyword-to-Caption Augmentation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, Rhodes Island, Greece, June 2023. IEEE. ISBN 978-1-7281-6327-7. doi: 10.1109/ICASSP49357.2023.10095969.

Jilan Xu, Carl Thomé, Danijela Horak, Weidi Xie, and Andrew Zisserman. Scaling Audio-Text Retrieval with Multimodal Large Language Models, arXiv preprint arXiv:2602.18010, February 2026.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni Technical Report, arXiv preprint arXiv:2503.20215, March 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, and Others. Qwen3-Omni Technical Report, arXiv preprint arXiv:2509.17765, 2025b.

Chen Yang, Chufan Yu, Hanfu Chen, Jie Zhu, Jingqi Chen, Ke Chen, Wenxuan Wang, Yang Wang, Yaozhou Jiang, Yi Jiang, and Others. MOSS-Audio Technical Report, arXiv preprint arXiv:2606.01802, 2026a.

Xiaoyu Yang, Xuenan Xu, Wenyi Yu, Siyin Wang, Changli Tang, Terumi Chiba, Siyuan Hou, Ziyang Zhang, Wen Wu, Baoxiang Li, Guangzhi Sun, Chao Zhang, and Philip Woodland. SALMONN-2: Advancing General-Purpose Hearing Abilities with Self-Supervised Representations, arXiv preprint arXiv:2607.17079, July 2026b.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01100.

Biao Zhang and Rico Sennrich. Root Mean Square Layer Normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

APPENDIX

A Collection and source accounting **16**

A.1 Public datasets and web audio 16

A.2 Audio segmentation 17

A.3 Decoding and signal checks 17

A.4 Deduplication 17

A.5 Evaluation quarantine 19

A.6 Storage format 19

B Taxonomy, enrichment, and deterministic selection **19**

B.1 Taxonomy construction and classifier versions 19

B.2 Specialist enrichment 19

B.3 Selection parameters and exact-size pass 21

C Dataset Captioning **22**

C.1 Captioning Model 22

C.2 Metadata Prompt Expansion 23

D Architecture **24**

D.1 Architecture Settings 24

D.2 Audio to Text Encoder Parameter Ratio 25

E Model Scaling **26**

F Design studies **29**

G HEAR and MSEB results **31**

H Dataset catalogue **34**

A COLLECTION AND SOURCE ACCOUNTING

A.1 PUBLIC DATASETS AND WEB AUDIO

For public datasets, a search across platforms such as Hugging Face, Zenodo, and the literature focused on gathering many permissively licensed collections. Although the internet often contains noisy and unstructured datasets, reusing existing academic data brings cleaner audio, as these datasets have often been tested in practice.

Academic datasets form the largest group by number of sources, accounting for 338 of the 467 sources listed in Table 16. YouTube and Common Crawl contribute most of the data by both clip count and audio duration. Table 16 displays these datasets. A further 39 sources omitted in that table contribute between one and nine captioned training audio samples each. These audio samples remain in the training corpus but are not listed in the catalogue.

YouTube audio is extracted across the full duration of each video, with a mean source duration of 13.1 minutes. Each video can therefore contribute many audio samples.

Common Crawl releases multiple crawls per year, often containing roughly 2–3 billion web pages each. The extractor processes WARC response records because they preserve the captured HTML, including the text surrounding audio links. WAT files provide derived metadata, including links, without the full page content. The extractor streams compressed WARC response records and searches HTML for audio elements and direct audio-file links. It resolves relative URLs and records the page URL, audio URL, and up to five sentences preceding and five sentences following each audio link in the page text. When an audio element offers multiple sources, the extractor selects one using a fixed format preference that prioritizes WAV, FLAC, and AIFF. The pipeline deduplicates URLs before downloading the files, then decodes the audio and checks its duration. The first extraction wave covered 35 crawl directories from 2021 to 2025 and found about 3.3 billion audio links, of which about 423M are unique. A second wave added crawls from 2013 to 2020 and from mid-2025 to early 2026. Candidate URLs are not successful downloads: on a sample of 5.4M URLs, 54% downloaded and 44% of the downloads converted to audio, a net return of 23.7%. Files larger than 50 MB were discarded before download to prioritize shorter recordings and accelerate retrieval. Audio that took longer than 30 seconds to download were ignored. Websites that did not connect or send any packet back to the scraper within 5 seconds were skipped, after a retry 2 seconds later. The downloader rejects files longer than 120 seconds. Downloaded recordings longer than 120 seconds are subsequently segmented as described in Appendix A.2. The deduplication input includes 141,523,038 Common Crawl audio samples, of which 51,453,212 remain after exact and fingerprint deduplication.

A.2 Audio segmentation

An experiment compared silence cutting with model-based event labelling before cutting and saw silence cuts score close to Dasheng event segmentation on MAEB-30 (Figure 6). Each method used approximately 1M clips from the same YouTube source pool, captioned with Qwen3-Omni, to train our XS model with three seeds per method and approximately 8M samples seen per run. Dasheng scored 32.92 ± 0.13, silence cuts 32.57 ± 0.18, random crops 32.18 ± 0.39, and quality-filtered crops 32.33 ± 0.09 (mean $\pm$ standard deviation). Processing cost motivates silence cutting: it uses CPU frame-energy scans of 8 kHz audio and avoids applying a model to do event detection over every soundtrack.

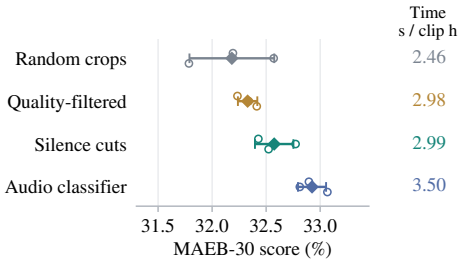


Figure 6: Segmentation comparison after 8M samples per run. Diamonds and whiskers show the MAEB-30 mean and standard deviation over three seeds. Open circles show individual runs. Time: seconds from reading the source audio to written FLAC clips, per hour of written clips, on the same 227 hours of YouTube audio using 72 CPU cores, plus one GH200 GPU for the classifier.

Silence cuts occur below -40 dB lasting at least 1 second, at the midpoints of these regions. Audio samples shorter than 5 seconds merge with adjacent segments, targeting durations near 30 seconds. Segments longer than 120 seconds split at quiet frames within ten seconds of 30-second target positions. Beyond YouTube, another pass over the other data converted 3.04M long recordings into 7.15M segments.

A.3 Decoding and signal checks

The pipeline removes audios that libraries such as `FFmpeg` and `soundfile` are unable to read or decode. It also records 40 statistics per clip: decoded duration, original sample rate and channels, RMS and peak level, dynamic range, DC offset, the proportion and position of silent frames at -60 dB, clipping proportion, bandwidth, spectral centroid and flatness, band energies, and the three 10-second windows with most energy.

A.4 Deduplication

Exact identity. All data are converted to 16 kHz mono signed 16-bit PCM signal for exact-match representation. The key combines the MD5 hash of its bytes with the audio length. This checksum

is used only for full exact matching. Audios that have the same semantic content, but are in different encodings or formats, will not match exactly. Exact matching of MD5 hash removes 151,056,336 in total. Common Crawl keeps 51.5M of its 141.5M audio samples after this stage. A small number of repeated files dominate these duplicates: ten files account for 73.6M removals, led by an audio containing a nudge to log in, downloaded 15.7M times under different URLs. The raw collection also contained 16.3M audio samples of pure silence.

Acoustic fingerprint matching. The acoustic stage links copies of the same recording after re-encoding, level changes, cropping, or padding. It uses Olaf (Six, 2023), a fingerprint technique similar to that used in Shazam. A fingerprint encodes the geometry of three spectral peaks, and two clips match when many fingerprints agree on one time offset. Fingerprints are computed on the full clip at 16 kHz mono, using a 64 ms Hamming window and an 8 ms hop. The extractor is a Rust port of Olaf's C code with bit-identical hashes. Hashes exclude absolute time but include absolute frequency, so pitch-shifted or speed-changed copies are not matched.

Each shared hash votes for the offset between its occurrences, quantized to 32 ms. The match count m counts distinct shared hashes at the offset with most votes. Chance collisions between unrelated audio spread over many offsets. For audio with n_a and n_b fingerprints, the score is

$$q(a, b) = \frac{m}{\sqrt{n_a n_b}}, \tag{3}$$

and pairs with $q \geq 0.50$ are linked. Audio with equal fingerprint counts must share at least half their fingerprints at one offset. A fully matching shorter clip must contain at least a quarter of the longer clip's fingerprints. This limits matches caused by short shared fragments.

Matching uses a distributed hash join over 100B+ fingerprint rows from 1.47 billion audio. Rows are partitioned by hash into 512 buckets and joined on shared hashes, aggregated by clip pair and offset. Counts are summed across buckets before choosing each pair's best offset. Hashes occurring more than 500 times within one source track are removed from that track before the join. This limits the quadratic join cost of frequent hashes.

Duplicate groups are connected components of edges with $q \geq 0.50$. Edges touching audio already removed by exact matching are skipped. One audio per group is kept, preferring public datasets with or without captions, Common Crawl, or video segments in that order. When audio came from the same dataset group, it prefers audio without model-generated captions, then longer text, more channels, a higher sample rate, and the lowest identifier. This stage removes 1,196,096 audio beyond exact matching.

Semantic near-duplicate screening. The semantic deduplication step is there to remove duplicates that the fingerprint stage misses. It follows SemDeDup (Abbas et al., 2023): embed every clip, cluster the embeddings with k-means, and compare pairs only inside a cluster. The encoder is Dasheng-base (Dinkel et al., 2024), a masked audio encoder trained without text, so the CLAP model under training does not shape its own data. In a smoke sample of 735K audio samples from dense archives, 30,569 pairs reached the threshold. None had been removed by the earlier stages, and only 4.6% had any fingerprint edge (14.5% among pairs of audio samples of at least 2 seconds). Most new pairs were very short audio samples such as wake words and single utterances. Other cases were one file crawled twice and a clip next to its own re-segmented copy.

Each clip is decoded to 16 kHz mono, cut to its first 10 seconds where audio after 10 seconds is not compared in this step. Audio is zero-padded if shorter than 10 seconds. Dasheng-base returns a sequence of 768-dimensional frame features, which is mean pooled over time and stored in 16-bit floating point. Embeddings exist for 1,412,277,595 stored audio samples. The screen runs on the 1,263,840,901 pool audio samples of at least 2 seconds. Shorter audio samples are not compared.

The embeddings are normalized to unit length and clustered with k-means in faiss on one GPU with $k = 65{,}536$, 20 iterations, a training sample of 20M random audio samples, seed 0, and assignment of every clip to its nearest centroid. The update is the standard Euclidean one. Centroids are not re-normalized. A target of one cluster per 2,000 audio samples would need about 632,000 clusters, and assignment cost grows with Nk, so k was capped at 65,536. The median cluster has 18,226 audio samples and the largest 95,813. The cluster sizes of this run are the density signal $n_{k(i)}$ in

Equation 4. At the selector cap of $2^{14.5}$, 17,870 clusters exceed the cap and hold 42.6% of the audio samples.

Inside each cluster, the pairwise cosine matrix of the members is computed on the GPU. This cosine similarity is set relatively high, as earlier results show false-positives. Mostly only pairs at 0.995 and above in the listening sets were repeated recordings. Clusters with more than 20,000 members are subsampled to 20,000 random members before the comparison. Within each k-means cluster, near-deduplicate samples with >0.995 cosine similarity form groups, where only 1 sample per group is kept. The keep order is by group: public datasets with captions are more important than noisy YouTube data. The run reported 113,572,831 removal decisions (9.0% of the clustered audio samples). Only 18,162,786 were applied.

A.5 Evaluation quarantine

An extensive check is applied to make sure the data in MAEB, HEAR and MSEB are not in the training data. Initially, it combines dataset and split exclusions with identifier matching to get a list of 287,076 audio samples for removal in the training data. Subsequently, another check with Olaf Fingerprints at coverage threshold of 0.8 was also performed. A third check was performed with item identifiers, MD5 hashes of decoded 16 kHz mono PCM, and waveform comparisons to remove another 15,176 samples.

A.6 Storage format

AudioWeb-500M uses WebDataset shards pairing FLAC audio with JSON metadata under a shared key, with numeric archive names and at most 4,096 samples per shard. Audio was initially normalized to 48 kHz and subsequently recoded to 32 kHz, reducing storage by approximately 16%. This choice was supported by measurements showing that 99.68% of samples contained at least 95% of their spectral energy below 13 kHz. The final dataset uses approximately 417 TB.

B Taxonomy, enrichment, and deterministic selection

B.1 Taxonomy construction and classifier versions

Figure 7 shows the top-level groups and a selection of second-level categories.

The taxonomy was expanded through caption-classification batches. Claude Sonnet 4.5 assigned each caption to a leaf and proposed additional categories. Merging removed duplicates and rejected also categories consisting of multiple events. The final hierarchy contains 4,416 nodes, including 4,085 leaves and 11 top-level groups, with depth up to five. The top-level groups are speech, music, designed sound effects, urban/transport, animal/bioacoustics, human non-speech, nature/environmental, domestic/household, interface/electronic, sports/recreation, and industrial/mechanical. For the classifier, only depths up to 3 were kept (instead of the full hierarchy of 5), resulting in a smaller prediction vocabulary of 3,718.

The original taxonomy classifier is a multilayer perceptron (MLP) trained on 1,280-dimensional mean-pooled features from a frozen Dasheng-0.6B encoder. A shared 1,024-unit hidden layer applies a linear transformation, batch normalization, GELU activation, and dropout. Separate linear output layers map the hidden representation to either leaf or top-level or second-level class logits. The labelled set has 232,728 caption-distilled audio samples, split by archive into 185,006 training, 22,404 validation, and 25,318 test audio samples. The MLP classifier is optimized for 40 epochs with AdamW, batch size 8,192, learning rate 10^{-3}, and weight decay 10^{-4}. A later improvement used a Qwen2.5-Omni model to relabel the training set, which improved top-level mapping accuracy on the held-out test set of 18,848 samples from 75.2% to 85.6%. The leaf-derived top-level mapping accuracy improved from 77.4% to 86.4%.

B.2 Specialist enrichment

Before our selection process, we gather extra data to make a better informed decision. Specialist enrichment is routed by the predicted top-level taxonomy class. Speech samples receive language, emotion, speaker age and gender, and pitch predictions. Music samples receive genre and style tags, tempo, and musical key. Animal samples receive bioacoustic class predictions. These outputs

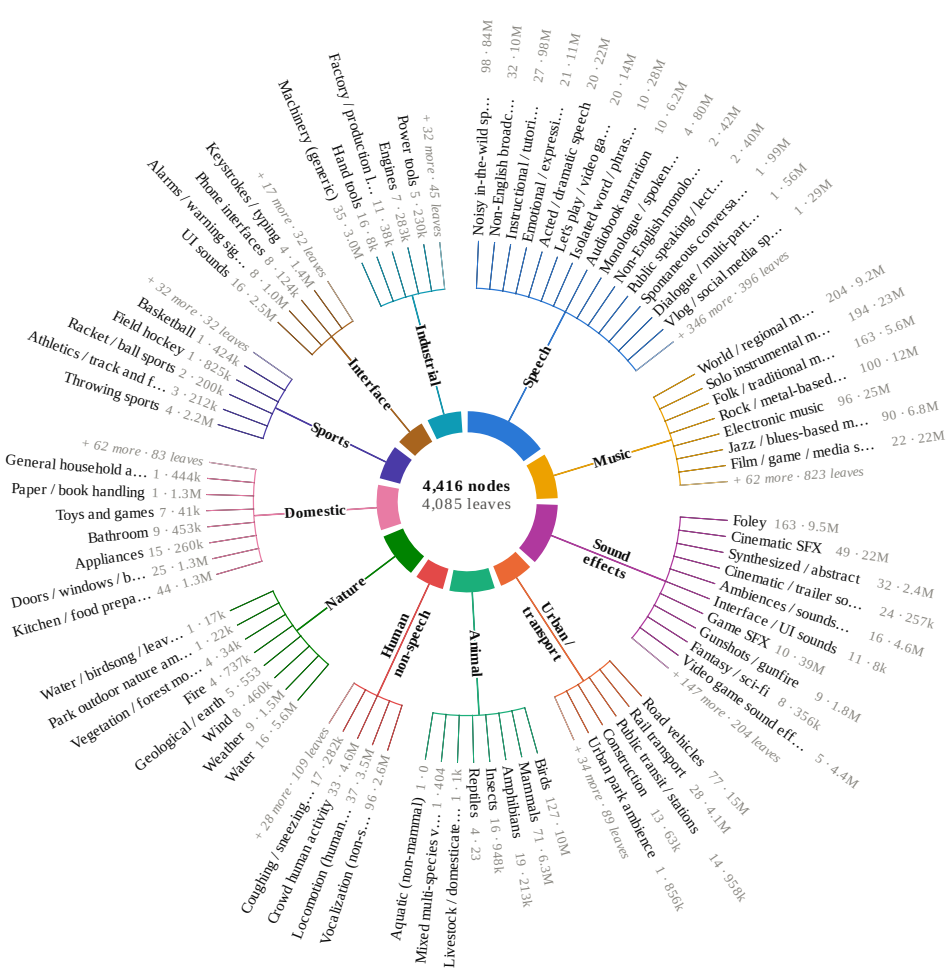


Figure 7: Visualization of the sound taxonomy. The five levels contain 11, 837, 3,150, 393, and 25 nodes, respectively, for a total of 4,416 nodes, including 4,085 leaves. Coloured inner arcs denote the 11 top-level groups, ordered clockwise from speech at the top. Branches show 77 of the 837 second-level categories, selected by descendant leaf count and audio sample count (grey numbers, respectively). Italic rows summarize omitted categories and their leaves.

accompany the taxonomy labels and signal statistics, including duration, loudness, silence fraction, clipping fraction, and bandwidth.

For speech, MMS-LID-4017 (Pratap et al., 2024) predicts one of 4,017 language labels, using the highest-scoring label and its probability. Emotion2vec+ large (Ma et al., 2024b) predicts one of nine labels: angry, disgusted, fearful, happy, neutral, other, sad, surprised, and unknown. The wav2vec2 age and gender model (Burkhardt et al., 2023) provides a continuous age estimate in years and a label of either female, male, or child. CREPE (Kim et al., 2018) estimates fundamental frequency over 50–550 Hz. The stored pitch is the median over frames with periodicity above 0.5, evaluated at 200 ms intervals.

For music, MAEST (Alonso-Jiménez et al., 2023) predicts out of 400 Discogs genre and style labels. Signal processing with librosa estimates continuous tempo in beats per minute and one of 24 musical keys, of 12 pitch classes in major or minor mode. For bioacoustics, Perch 2.0 (van Merriënboer et al., 2025) has 14,795 output classes, consisting of 14,597 species labels and 198 general sound-event classes. The five highest-scoring classes and their logits are used.

B.3 SELECTION PARAMETERS AND EXACT-SIZE PASS

For candidate clip i, the selection probability is p_i. This depends on whether the clip passes the quality checks (g_i), its minimum selection probability (r_i), the size of its acoustic cluster (c_i), and the adjustment for the target class proportions (w_i):

$$p_i = g_i \left[r_i + (1 - r_i) c_i w_i \right], \qquad c_i = \min\left(1, \frac{2^{14.5}}{n_{k(i)}}\right). \tag{4}$$

Here, $n_{k(i)}$ is the number of clips in the acoustic cluster. The factor c_i reduces the selection probability for clips in large clusters. For the 500M selection, each clip must pass all of the following checks: the audio must decode successfully, clip must contain at least 1 s of content and its original duration must be between 0.5 and 120 s. Clips with both silence fraction ≥ 0.98 and RMS ≤ -80 dB are excluded. We set $g_i = 1$ if the clip passes all checks, and $g_i = 0$ otherwise. For these correct clips, the rules below set a minimum selection probability. The value r_i is the largest probability assigned by these rules, or zero if none applies. A value of 1 means the clip is always kept.

For each category at the most specific level of the taxonomy, valid candidates are sorted from highest to lowest prediction confidence. Each receives a weight of 1 at confidence ≥ 0.5, or 0.3 otherwise. Starting at the top of the list, we add these weights until their sum reaches or exceeds 2,000. Each clip included in this sum gets its weight as a minimum selection probability. If the sum for the whole list is below 2,000, this applies to all clips in the list.

We set $r_i = 1$ for up to 50,000 valid candidates for each predicted language, chosen from highest to lowest language confidence. The same applies to up to 1M candidates per emotion label and 500 per species label, chosen from lowest to highest hash value. The emotion labels covered by this rule are angry, disgusted, fearful, happy, sad, and surprised. We also set $r_i = 1$ for valid clips whose second top-level prediction has probability ≥ 0.25 for music and designed sound effects, or ≥ 0.10 for the remaining categories other than speech. This rule also covers valid clips from CommonVoice_22_0 and EarSAVAS, to promote speaker and language diversity.

The target class proportions are applied to expected clip counts, where a candidate selected with probability 0.3 contributes 0.3 to the count. The initial budget is $T = 660$M before the final size adjustment, with shares $(0.37, 0.29, 0.34)$ for speech, music, and all remaining classes $((244.2, 191.4, 224.4)$M). Unused budget in one group is not moved to another. For second-level category h, the protected expected count P_h comes from the minimum probabilities r_i. The adjustable expected count D_h is the additional contribution after applying the cluster-size factor c_i, before adjusting the class proportions:

$$P_h = \sum_{i \in h} g_i r_i, \qquad D_h = \sum_{i \in h} g_i (1 - r_i) c_i. \tag{5}$$

For group b with share a_b, the required reduction in expected count is $E_b = \max(0, \sum_{h \in b}(P_h + D_h) - a_b T)$. Let $\mathcal{H}_b$ initially contain its categories other than language/accent and condition-specific speech. We reduce the largest adjustable category counts first, bringing them toward a common limit until the required reduction is reached or no adjustable count remains. This limit λ_b satisfies

$$\sum_{h \in \mathcal{H}_b} (D_h - \lambda_b)_+ = \min\left(E_b, \sum_{h \in \mathcal{H}_b} D_h\right), \qquad X_h = (D_h - \lambda_b)_+, \tag{6}$$

where $(x)_+ = \max(x, 0)$ and X_h is the reduction in expected count for category h. If a further reduction is needed, the same calculation is applied to the categories left out of the first step. The protected expected count P_h is never reduced. The implementation skips reductions of at most 0.5 in expected clip count.

Within each category requiring a reduction, candidates with a positive adjustable contribution $d_i = g_i(1 - r_i)c_i$ are sorted by decreasing removal score

$$s_i = 4G_i + 2q_i + \frac{8}{170} \log_2 n_{k(i)} + 2A_i - 2E_i. \tag{7}$$

Here G_i marks a YouTube-track clip. q_i is top-level prediction confidence. A_i marks an original caption, a model caption, or text of at least 25 characters. E_i marks one of the emotion labels listed

Captioner	speed ↑ samples/s per node	diversity ↑ vocab @80k	diversity ↑ vocab @400k	quality human rank ↓	quality LLM rank ↓	quality LLM BT ↑
MOSS-Audio-8B-Thinking	22.9	**5,488**	**16,613**	**2.91** ±0.39	2.76 ±0.50	1.694
MOSS-Audio-8B-Instruct	19.2	5,054	15,768	3.24 ±0.36	**2.61** ±0.40	**1.755**
Qwen2.5-Omni-7B	23.0	4,652	8,660	4.10 ±0.45	4.43 ±0.50	0.873
AudioFlamingoNext	55.5	4,743	9,858	4.26 ±0.41	4.51 ±0.53	0.714
Qwen3-Omni-30B-A3B	75.9	2,693	6,619	4.32 ±0.45	3.43 ±0.51	1.093
Nemotron-3-Nano-Omni	**96.5**	2,284	7,078	4.57 ±0.47	5.16 ±0.40	0.459
Audex-30B-A3B	36.5	5,395	9,697	4.60 ±0.45	5.10 ±0.57	0.411

Table 4: **Caption Model Comparison**. Arrows indicate whether higher (↑) or lower (↓) is better. **Speed:** Throughput is given in samples/s per node (4x GH200 GPUs). **Diversity:** *Vocab@80k* and *vocab@400k* are the number of unique words across 80k short and 400k long caption tokens, respectively. **Quality:** *Human rank* and *LLM rank (Gemini 3.6 Flash)* report the mean rank among the 7 captioners for the 50 shared audio samples (1 = best, 7 = worst) ± 95% CI. LLM BT is the Bradley–Terry strength for the Gemini 3.6 Flash judge on the full 8997 test audio samples.

above. Missing confidence and missing cluster density contribute zero to the score. Candidates with the same score are ordered by $\mathrm{SplitMix64}(\mathrm{audio_id}_i)$, from lowest to highest. Adding d_i in this order, let j be the first candidate for which the sum reaches X_h. The factor for adjusting the class proportions is

$$w_i = \begin{cases} 0, & i \text{ precedes } j, \\ 1 - \left(X_h - \sum_{\ell \prec j} d_\ell\right)/d_j, & i = j, \\ 1, & i \text{ follows } j. \end{cases} \tag{8}$$

Categories requiring no reduction have $w_i = 1$. We hash each clip's audio ID with a fixed seed and scale the result to a value between 0 and 1. The clip is selected if this value is below its selection probability p_i. Using a fixed seed makes this step reproducible.

The selected clips are matched to their current archive locations, and duplicate entries with the same archive path and key are removed. Selected clips not covered by any minimum probability rule above are labeled *fill*. The final size adjustment keeps all non-speech clips and all clips covered by these rules. Among speech clips labeled fill, it keeps those with the lowest deterministic hashes until the base index has exactly 500,000,000 rows. Hash values are grouped into ranges, and the entries in the range containing the cutoff are sorted to get the exact count.

Speech falls to 42.7%, music accounts for 27.4%, and the remaining classes account for 30.0%. The same validated additions used for all selection variants are then added. The final index labels record the selection rule for each clip: 72.307% are fill, 22.577% are assigned to the secondary-prediction or source-dataset rules, and the rest to the category, language, emotion, or species rules, or to additions. The secondary-prediction and source-dataset rules account for more selected audio samples than any one of the other minimum probability rules.

C Dataset Captioning

C.1 Captioning Model

Captioning a large training pool is a one-shot, expensive decision. When generating captions for a large dataset, we want to maximize their quality and diversity while keeping compute costs low. If the captions often contain incorrect details or are very similar to one another, we end up with a weak training signal. The biggest influence on these three factors comes from the large audio-language models used for captioning. To pick the best models for our use case, we tested seven configurations covering six model candidates: MOSS-Audio-8B (thinking on and off) (Yang et al., 2026a), Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b), Qwen2.5-Omni-7B (Xu et al., 2025a), AudioFlamingoNext (Ghosh et al., 2026), Nemotron-3-Nano-Omni-30B-A3B-Reasoning-FP8 (Nvidia et al., 2026), and Nemotron-Labs-Audex-30B-A3B (Kong et al., 2026). Model names in the following might be shortened.

Quality: In the following study (Table 4), we compare the caption quality of the candidate models. We use an automated LLM review in which all model captions are ranked against one another for each clip. Gemini 3.6 Flash (Doshi, 2026) ranked anonymized and shuffled candidates according to faithfulness, coverage, and specificity. Additional permutations, pairwise comparisons, and a claim audit measured judge consistency. We picked a subset of 8997 audio samples matching our data distribution. After generating one caption per candidate, we calculate the Bradley–Terry strength from the automated rankings. Table 4 shows that MOSS Audio 8B produces the best results by showing the highest Bradley–Terry strength in the LLM review regardless of the thinking setting. In contrast, Audex and Nemotron-3 are the two weakest models across the evaluation.

To ground these results in human perception, we rank a subset of 50 audio samples (25 short, 25 long captions) with two human annotators ("human rank") and compare the results with the LLM judge ("LLM rank"). The labeling interface anonymizes models and shuffles the captions. The annotators only agreed 12 out of 50 times on their top choice, with a mean per-clip Kendall τ_b of $+0.295$ $[+0.202, +0.387]$. Compared to the inter-annotator agreement, the LLM judge falls within the range of a human rater: $\tau_b = +0.345$ $[+0.250, +0.436]$ with one annotator and $+0.265$ $[+0.166, +0.362]$ with the other. Brackets are 95% percentile bootstrap confidence intervals from $20{,}000$ resamples over clips. Pooling the two annotators, humans and judge produce nearly the same ordering of the seven captioners (Table 4). Both put the two MOSS Audio configurations on top and Nemotron-3-Nano-Omni and Audex-30B-A3B at the bottom. The largest single disagreement is Qwen3-Omni-30B-A3B, which the judge places 3rd while our human ranking places it 5th.

Note that the set of $n = 50$ human-annotated audio samples is meant to validate the LLM review, not to serve as a leaderboard. It can only give a rough impression of the top and bottom models while checking that the LLM judge behaves similarly to a human rater.

Speed: Another important aspect when selecting captioning models is inference speed. When captioning several hundred million samples, even slightly slower models can lead to a significant increase in compute cost. To optimize the captioning speed of each model, we considered serving layout, hyperparameter settings (e.g., maximum number of sequences), NUMA bindings, caching, output format, and more. We ended up with large differences of up to ~$5\times$ between our fastest (Nemotron, 96.5 samples/s/node) and slowest (MOSS thinking-OFF, 19.2 samples/s/node) model (Table 4). This puts the theoretical compute estimate for 470 million audio samples between 1,350 and 6800 node-hours.

Diversity: Table 4 shows the number of unique words within a limited number of text tokens. We evaluate this metric on short and long captions with 80k and 400k token limits, respectively. This way, the metric does not depend on the verbosity of each model. There is a ~$2.4\times$ difference between the most repetitive (Nemotron-3-Nano-Omni and Qwen3-Omni-30B-A3B) and the most diverse (MOSS Audio think-ON) model.

Based on the results of this study and our three main criteria (quality, diversity, and computational cost), two models stood out to us. MOSS Audio think-ON with high diversity and quality, while being slow and Qwen3-Omni for its speed, while still performing strongly on the Bradley–Terry strength metric. We decided to use a combination of Qwen3-Omni and MOSS Audio think-ON. This combination achieves a caption diversity of ~4,395 and ~11,976 unique words per 80k/400k text tokens, respectively, while balancing the quality and compute cost of MOSS Audio think-ON with the speed of Qwen3-Omni-30B-A3B. Both models then captioned disjoint, randomly selected halves of the total ~470 million audio samples, each with a long and a short caption. This results in a theoretical compute cost of about 3,710 node-hours ($4\times$ GH200 per node).

C.2 Metadata Prompt Expansion

When captioning audio samples, the prompt given to the captioning model directly affects the quality of the produced caption. In addition to the general instructions, we identified three primary approaches to include each sample's metadata generated by specialist models (Appendix B.2). They are (1) passing only general instructions with the audio, (2) attaching the whole raw JSON metadata string, or (3) adding only selected JSON fields converted to natural language to the prompt. We tested these three approaches in combination with our two captioning models (Qwen3-Omni and MOSS Audio) on 976 audio samples, selected to cover the dataset's duration and domains, resulting in six long captions per clip. Gemini 3.6 Flash then ranked all 6 captions for the same audio sample

#	Model	Caption type	Mean rank ↓	Win rate (%) ↑
1	Qwen3-Omni	Raw JSON	$\mathbf{3.20} \pm 0.10$	$\mathbf{19.9} \pm 2.5$
2	MOSS Audio	Raw JSON	3.35 ± 0.10	16.4 ± 2.4
3	Qwen3-Omni	Natural sounding	3.46 ± 0.10	14.7 ± 2.2
4	MOSS Audio	Natural sounding	3.56 ± 0.10	12.9 ± 2.1
5	MOSS Audio	Audio only	3.58 ± 0.12	19.6 ± 2.5
6	Qwen3-Omni	Audio only	3.84 ± 0.12	16.6 ± 2.4

Table 5: **Attaching raw metadata strings leads to the highest caption quality.** Overall ranking of two captioning models with three prompt types across the same 976 audio samples, evaluated with Gemini 3.6 Flash (thinking: low). Mean rank ranges from 1 (best) to 6 (worst). Win rate refers to being the best caption for a clip (higher-is-better). Brackets give 95% percentile bootstrap confidence intervals over $B = 20{,}000$ resamples of the 976 audio samples.

Tier	Audio $d_a \times L_a$, h, d_{ff}	Text $d_t \times L_t$, h, d_{ff}	d	A-Size (M)	T-Size (M)	A:T	N (M)	Total (M)	GFLOPs /clip
XXS	224×12, 4, 640	224×12, 4, 640	224	8.3	7.8	1.06	16.0	27.3	26.1
XS	320×12, 5, 896	320×12, 5, 896	320	16.6	15.7	1.06	32.2	48.3	46.8
S	512×12, 8, 1408	512×11, 8, 1408	512	41.8	36.4	1.15	78.2	103.9	103.4
M	640×12, 10, 1728	640×11, 10, 1728	640	64.4	56.2	1.15	120.6	152.8	152.9
L	1024×12, 16, 2816	960×11, 15, 2560	960	166.5	125.4	1.33	291.8	340.1	357.9

Table 6: **Model tiers for WSD scaling runs:** Audio and text columns give width × depth, number of attention heads, and SwiGLU hidden size per encoder. A-Size and T-Size are the respective encoder parameter counts in millions. Audio counts include the MAP pool and projection layers. Text counts include the pool and output projection but *exclude* the $50{,}260 \times d_t$ token-embedding table. N (non-embedding) is the sum of A-Size and T-Size. GFLOPs are total (audio + text tower) forward-pass FLOPs per clip at the training length distribution.

by accuracy, completeness, temporal details, specificity, and writing. The judge received only the audio and the six randomly ordered captions, never the metadata, so it could not simply confirm our specialists' predictions.

The results in Table 5 show that adding the raw metadata JSON string results in the best average performance across models. This observation also holds for almost all audio lengths and domains. We attribute this to the quality of the metadata and the ability of modern large audio-language models to process JSON strings well. Considering these results, we include the raw metadata JSON string in the prompts for our captioning process.

D Architecture

D.1 Architecture Settings

All models used in our experiments share a single dual-encoder layout. The layout is a contrastive audio–text model in the CLAP family: a spectrogram Vision Transformer with native variable-length input (NaFlex) (Tschannen et al., 2025) encodes audio, a bidirectional Transformer encodes text, and both are mapped into a shared embedding space trained with a standard CLIP objective on audio and text.

For the architectural design, we aimed to align with established encoder widths and depths while offering a broad range of sizes. We ended up with five model scales from XXS (27.3M parameters) up to L (340M parameters) that are aligned with known small-scale VIT dimensions and optimized for hardware alignment as seen in Table 6. For more details on the audio-to-text encoder ratio, refer to Appendix D.2. With this setup, we cover the computational range per forward pass from an average of 26.1 GFLOPs up to 357.9 GFLOPs per clip, given our training distribution.

For the architectural hyperparameters and mechanisms, we adopt proven approaches such as RoPE and SwiGLU, while motivating others through our own experiments in Section F. An overview of key design choices can be found in Table 7.

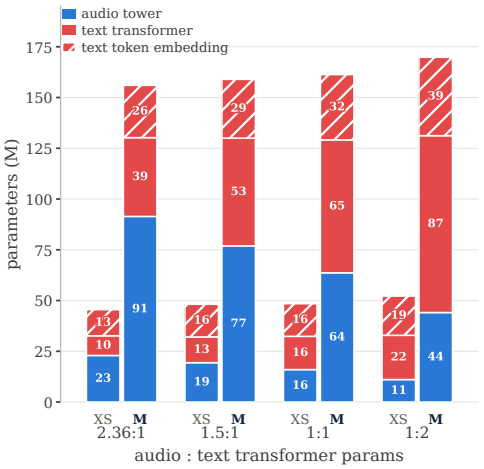
Table 7: Architecture settings shared by all tiers in the scaling study. For details on architecture shape refer to Table 6 & 8.

Audio front-end	
Sample rate / mel bins	32 kHz / 128 (50 Hz–16 kHz)
Window / hop	1024 / 320 samples (100 frames/s)
Patch (freq $\times$ time)	16×16 (160 ms), 50 tokens/s
Max. clip length	30 s (1504 tokens), train and eval
Length buckets (tokens)	344 / 448 / 648 / 1128 / 1504
Audio tokens per GPU step	49,152
Audio encoder	
Positional encoding	2-D axial RoPE (freq, time), no learned pos. emb.
Norm / MLP	pre-norm RMSNorm / SwiGLU
Extra tokens	1 register, no class token
LayerScale init	10^{-5}
Pooling / projection	MAP / Linear–GELU–Linear
Text encoder	
Tokenizer / vocab / context	GPT-2 BPE (`r50k_base`) / 50,260 / 256
Attention	bidirectional, QK-norm, no biases
Positional encoding	1-D RoPE
Norm / MLP	pre-norm RMSNorm / SwiGLU
Pooling / projection	MAP / bias-free linear
Joint space	
Temperature	learned, init $1/0.07$, clamp 100/300
Loss	symmetric InfoNCE over global batch
Precision	bf16 mixed precision

D.2 Audio to Text Encoder Parameter Ratio

A main component of the architecture design is the parameter ratio between the audio and text encoders. This ratio depends on many factors, including downstream benchmarks, model size, and the training corpus. In the following, all reported ratios and parameter shares exclude parameters related to text token embedding, which is a fixed lookup cost. When comparing other contrastive pre-trained architectures, one finds substantial discrepancies among them. While OpenCLIP (Cherti et al., 2023) and SigLIP2 (Tschannen et al., 2025) allocate up to ~78% of the model parameters to the vision encoder, SigLIP1 (Zhai et al., 2023) splits evenly, and LAION-CLAP (Wu et al., 2023) and MS-CLAP (Elizalde et al., 2023) do the opposite, by assigning between ~51% and ~73% of parameters to the text encoder. The difference is that vision-based contrastive models are trained on large vision-language datasets that provide substantial visual variety while offering rather compressed text captions. Therefore, scaling the vision encoder results in higher benchmarks. Older CLAP models, on the other hand, struggle with overfitting due to data scarcity while depending on pre-trained component sizes.

We tested four different parameter allocations on different data scales and model sizes, as shown in Figure 8. Changing the audio-to-text ratio from 2.36:1 to 1:2 does not show significant differences in the MAEB mean score with $D \leq 64M$. This observation does not hold for the two XS parameter ratios we tested on $D = 512M$, with a difference of 2.1 points. Performance differences might appear as D increases further. For computational reasons, we stopped after the runs shown in Figure 8. In general, shifting parameters to the text encoder results in faster inference speeds given that the encoder widths remain hardware-aligned. Across our model range, this results in up to ~22% faster inference speed. Considering these aspects, we chose conservative CLIP-inspired ratios by allocating ~51% to ~57% of parameters to the audio encoder. This increases inference speed compared to older LAION-CLAP and MS-CLAP ratios, while accounting for our new data scale and architectural design freedom by avoiding pre-trained components. For more details on the architecture, refer to Appendix D.1.



	audio : text			
D	2.36 : 1	1.5 : 1	1 : 1	1 : 2
XS tier				
16M	**0.346**	**0.346**	0.344	0.335
32M	0.364	0.362	**0.365**	0.359
64M	0.385	**0.388**	0.379	0.377
512M	**0.473**	–	0.451	–
M tier				
16M	0.354	0.357	**0.361**	0.346
32M	0.368	0.367	**0.373**	0.363
64M	0.395	0.387	**0.396**	0.393

Figure 8: **Audio:text capacity choice does not significantly affect performance up to 64M samples.** Left: parameter composition of the four capacity splits for model sizes XS and M. Within a tier, the non-embedding total is held constant. The 50k-vocab token embedding is adapted to the individual layer dimensions and is a fixed lookup cost, not encoder capacity. Right: Mean MAEB scores over 30 tasks per training sample seen D show that the audio-to-text tower ratio does not have a significant effect up to 64M samples.

Table 8: **Model tiers for cosine scheduled scaling runs.** Audio and text columns give width × depth, number of attention heads, and SwiGLU hidden size per encoder. A-Size and T-Size are the respective encoder parameter counts in millions. Audio counts include the MAP pool and projection layers. Text counts include the pool and output projection but *exclude* the $50{,}260 \times d_t$ token-embedding table. N (non-embedding) is the model size used in all scaling fits while total refers to the actual model size. GFLOPs are total (audio + text tower) forward-pass FLOPs per clip at the training length distribution.

Tier	Audio $d_a \times L_a$, h, d_{ff}	Text $d_t \times L_t$, h, d_{ff}	d	A-Size (M)	T-Size (M)	A:T	N (M)	Total (M)	GFLOPs /clip
XXS-C	256×12, 4, 704	192×12, 3, 512	192	10.5	5.5	1.92	15.9	25.6	29.2
XS-C	384×12, 6, 1024	256×12, 4, 682	256	22.9	9.7	2.36	32.6	45.5	58.7
S-C	576×12, 9, 1536	384×12, 6, 1024	384	51.5	21.8	2.36	73.3	92.6	117.1
M-C	768×12, 12, 2048	512×12, 8, 1365	512	91.4	38.8	2.36	130.2	156.0	194.9
L-C	896×21, 14, 2389	640×12, 10, 1706	640	211.3	60.6	3.49	271.9	304.1	431.1

E Model Scaling

Setup and hyperparameters: All our scaling plots are based on two batches of runs. One is trained with warmup-stable-decay (WSD), and the other with cosine learning-rate scheduling. Both of these batches are consistent with each other but differ in the scheduler, warmup, learning rate, batch size in addition to altering the architecture proportions. While the older cosine-scheduled runs were done a bit earlier in development, they have a more audio-encoder-heavy parameter allocation and slightly change the total number of parameters (Table 8). We mark them with a C after the model size, like XS-C. Since we compare the error rate to computation time, we can include both in the scaling plots without issue. For all other plots and tables, unless explicitly mentioned otherwise, we always refer to the warmup-stable-decay scheduled runs, as they almost always outperform the cosine runs (Appendix D.1). You can find the respective training settings for the WSD, and the cosine runs in Table 9 and 10 respectively. We did some initial learning rate and batch size experiments on 32M, 64M, and 128M data scales and extrapolated them to other scales following (Nezhurina et al., 2026). We then adapted the learning rate for WSD with a factor of $\times 0.75$ and $\times 0.80$ while choosing a data scale appropriate batch size.

Table 9: **Training configuration of the cooled-down WSD endpoints.** "XXS–L" denotes all five of our model sizes (XXS, XS, S, M, L); Samples is the number of training samples seen. Every model size shares the schedule at a given budget. Warmup and cooldown are in optimizer steps, and the cooldown decays the learning rate linearly to zero over the final 20 % of the budget. Batch size varies based on the length buckets.

				Global Batch			
Models	Samples	Warmup	Cooldown	Min	Max	Avg	LR
XXS–L	1M	350	224	512	2,176	916	4.0×10^{-4}
XXS–L	2M	350	448	512	2,176	916	4.0×10^{-4}
XXS–L	4M	350	896	512	2,176	916	4.0×10^{-4}
XXS–L	8M	350	1,792	512	2,176	916	4.0×10^{-4}
XXS–L	16M	350	3,584	512	2,176	916	4.0×10^{-4}
XXS–L	32M	350	7,168	512	2,176	916	4.0×10^{-4}
XXS–L	32M	2,800	1,744	2,048	8,704	3,683	7.5×10^{-4}
XXS–L	64M	2,800	3,488	2,048	8,704	3,683	7.5×10^{-4}
XXS–L	128M	2,800	6,976	2,048	8,704	3,683	7.5×10^{-4}
XXS–L	256M	2,800	13,952	2,048	8,704	3,683	7.5×10^{-4}
XXS–L	512M	2,800	27,904	2,048	8,704	3,683	7.5×10^{-4}
XXS–M	1,024M	2,800	55,808	2,048	8,704	3,683	7.5×10^{-4}

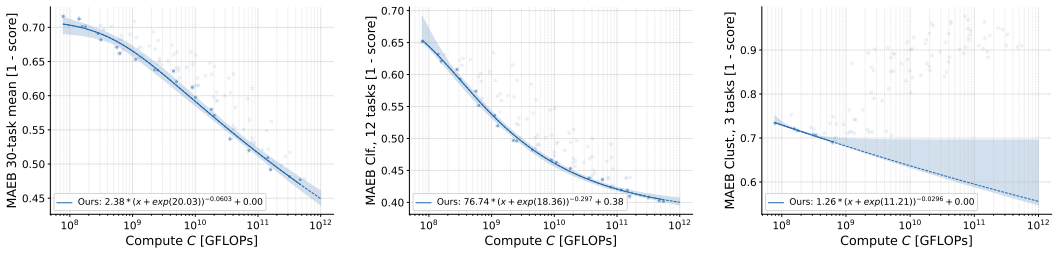
Table 10: **Training hyperparameters for the cosine learning rate scheduled scaling runs.** "XXS–L-C" denotes all five of our model sizes (XXS-C, XS-C, S-C, M-C, L-C); Samples is the number of training samples seen. Batch size varies based on the length buckets

				Global Batch			
Models	Samples	Warmup	Nodes	Min	Max	Avg	LR
XXS–L	2M	472	2	256	1,088	461	2.5×10^{-4}
XXS–L	4M	472	4	512	2,176	922	2.5×10^{-4}
XXS–L	8M	943	4	512	2,176	922	5.0×10^{-4}
XXS–L	16M	943	8	1,024	4,352	1,843	5.0×10^{-4}
XXS–L	32M	1,183	12	1,536	6,528	2,765	1.0×10^{-3}
XXS–L	64M	2,367	12	1,536	6,528	2,765	1.0×10^{-3}
XXS–L	128M	3,774	16	2,048	8,704	3,686	1.0×10^{-3}
XXS–L	256M	5,000	16	2,048	8,704	3,686	1.0×10^{-3}
XS, S	512M	5,000	16	2,048	8,704	3,686	1.0×10^{-3}

Additional scaling results: Figure 9 shows the scaling behavior of our model in contrast to different tasks of MAEB-30. It shows that, in addition to the mean MAEB-30 error rate, pair classification, retrieval, and zero-shot classification also scale well with compute. But what is not shown on the mean score scale is that classification tasks saturate at the end of our compute budget. Even more extreme are the clustering benchmarks which scale inversely to the compute. Given our equation 2 and its constraints, this leads to an almost flat line in the plot. This is a known challenge of the MAEB benchmark as noted in their conclusion (Assadi et al., 2026).

We also tested the quality of our scaling prediction by excluding all runs with $C \geq 10^{11}$ GFLOPs. As shown in Figure 10, we are able to accurately predict the five runs on the compute-Pareto-frontier with $C \geq 10^{11}$ GFLOPs. The predicted runs are marked with an $\times$ in the figure. Our predictions have a mean absolute error of 0.5 MAEB-30 points and a mean squared error of 5.22×10^{-5}.

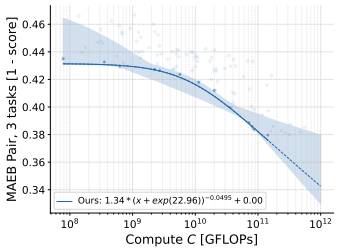
Last but not least, we plot the compute against our validation loss. Our validation set consists of ≈ 1.5k samples that were not used in training. Figure 11 shows isolines with the same data scale, while marking the lowest loss with a star. From these valleys, we can determine the most sample efficient model size that is neither over- nor under-trained. An example for this would be $D = 32M$, where M offers the best compromise between model size and data scale. To predict this for a larger D, we would need to train larger models with up to $D = 1{,}024M$ since there is no valley for them



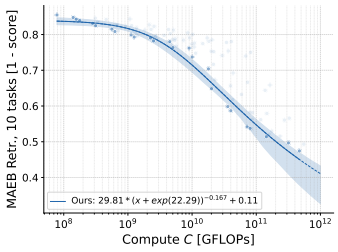
(a) MAEB (full, 30 tasks)

(b) Classification (12 tasks)

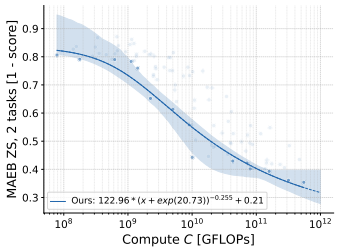
(c) Clustering (3 tasks)



(d) Pair Classification (3 tasks)



(e) Retrieval (10 tasks)



(f) ZS Classification (2 tasks)

Figure 9: **Scaling behavior on the full MAEB benchmark and per task type.** MAEB-30 error rate (1 - mean score) per compute C (GFLOPs). The band marks the 95% bootstrap CI with 1000 resamples. It shows clear trends in compute scaling for zero-shot (ZS) classification, pair classification, and retrieval. In contrast, it isolates the saturating classification task and the inversely scaled clustering benchmark subset.

yet. One important note is that validation loss does not directly transfer to downstream benchmark performance as shown in Figure 3.

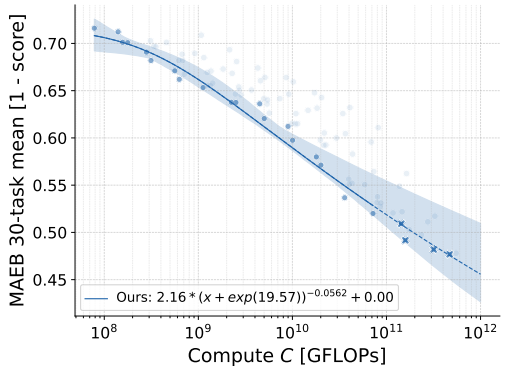


Figure 10: **Scaling predictions based on less than 10^{11} GFLOPs.** We are able to accurately predict our five best performing frontier models with more than 10^{11} GFLOPs of compute. We can do this with a mean squared error of 5.22×10^{-5} or a mean absolute error of 5.1×10^{-3}.

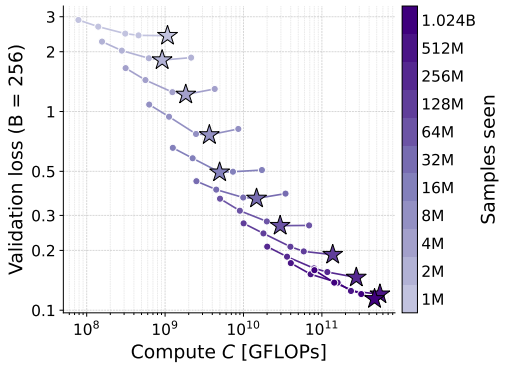


Figure 11: **Validation loss per compute C.** We plot our models based on compute and data scale against their validation loss with batch size $B = 256$. We can see clear valleys in the loss at smaller compute scales, indicating the optimal compute for a given data scale. This plot only contains runs trained on a warmup-stable-decay learning rate schedule.

Table 11: MAEB-30 task scores (%) of the final scaling-grid models, each the largest cooled-down checkpoint that has finished training for its tier. Parameter counts are totals over both towers, including embeddings. All models are evaluated with native MTEB 2.21.0 on the same embedding wrapper, with audio capped at 30 s (1504 tokens), un-normalised audio embeddings, and all-caps text inputs lowercased. Multi-subset tasks average all subsets. Bold marks the best value per row.

Task	Type	XXS (27M)	XS (48M)	S (104M)	M (153M)	L (340M)
Samples seen		1024M	1024M	1024M	1024M	512M
Speech (18 tasks)						
SpokenSQuAD T2A	Retr.	34.3	39.0	38.0	**41.7**	36.3
CommonVoiceMini21 T2A	Retr.	60.8	65.6	68.0	**71.9**	71.3
FLEURS T2A	Retr.	47.3	50.7	55.2	**61.8**	55.8
GigaSpeech T2A	Retr.	59.1	64.3	**71.6**	70.4	65.6
MInDS-14	Clf.	47.1	61.9	70.6	**71.7**	70.2
VoxCeleb SA	Clf.	29.4	30.7	**33.4**	33.3	32.7
IEMOCAP gender	Clf.	80.1	85.9	90.4	90.6	**92.6**
VoxPopuli language ID	Clf.	92.5	96.4	98.0	98.1	**98.6**
CREMA-D	Clf.	30.8	34.2	37.4	38.4	**40.8**
CommonLanguage age	Clf.	15.4	15.6	**16.0**	15.8	**16.0**
VoxPopuli gender	Clust.	5.6	8.1	23.8	23.6	**23.9**
CREMA-D (clustering)	Clust.	**0.7**	0.6	0.6	0.6	**0.7**
SIB-FLEURS	Clf.	21.3	23.6	26.8	**27.2**	26.3
NMSQA	Pair	75.6	76.8	**76.9**	76.4	74.8
VoxPopuli accent	Pair	53.6	**53.9**	53.5	53.1	53.4
CREMA-D (pair)	Pair	55.0	55.4	55.6	55.9	**56.2**
RAVDESS	Zero-shot	28.4	21.4	**32.2**	21.0	32.1
SpeechCommands v0.02	Zero-shot	91.3	94.1	95.6	92.0	**97.1**
Mean		46.0	48.8	52.4	52.4	**52.5**
Music (6 tasks)						
JamAlt artist A2A	Retr.	88.3	88.7	88.9	89.5	**89.6**
JamAlt lyrics A2T	Retr.	41.3	44.2	47.4	**54.8**	52.9
Beijing Opera	Clf.	85.2	91.3	93.3	94.5	**95.6**
GTZAN genre	Clf.	83.6	**87.2**	87.1	86.9	86.1
Mridangam tonic	Clf.	32.8	37.7	44.0	46.6	**48.3**
GTZAN reranking	Rerank	84.0	83.4	84.2	**84.5**	84.0
Mean		69.2	72.1	74.2	**76.1**	**76.1**
Sound (6 tasks)						
Clotho T2A	Retr.	28.3	31.1	31.2	**31.6**	30.4
MACS T2A	Retr.	18.3	18.1	17.8	18.6	**20.1**
UrbanSound8K T2A	Retr.	0.8	0.8	**0.9**	**0.9**	0.7
BirdCLEF	Clf.	31.4	39.4	41.8	**42.9**	41.7
VehicleSound	Clust.	3.1	2.1	2.6	**3.8**	2.6
FSDKaggle2019	Multi.	61.5	70.0	**71.4**	71.1	69.5
Mean		23.9	26.9	27.6	**28.1**	27.5
MAEB-30 mean		46.23	49.08	51.81	**52.31**	52.20

F DESIGN STUDIES

Audio duration: Training on audio of at most 30 seconds produces a better model with less than one-third of the compute. At 256 million training samples seen, dropping longer audio samples scores 47.1%, cropping them to 30 seconds scores 46.5%, and training on audio up to 120 seconds scores only 43.0% on the mean MAEB benchmark (Table 12). Allowing batches to hold up to 120-second audio samples leads to far more audio tokens per clip. At 50 audio tokens per second, the maximum number of audio tokens increases from 1504 (30 seconds) to 6008 (120 seconds), leading to smaller batches per GPU and slower training. The 120-second bucket holds only 8 audio samples per GPU batch but accounts for about three-quarters of the training FLOPs because attention cost grows with the square of the token count. Removing it reduces the cost per sample from 1.94×10^{12} to 0.56×10^{12} FLOPs, a factor of 3.5, and increases throughput on 128 GPUs from 6,600 to 12,300 samples per second. A compromise for this slowdown could be to oversample shorter buckets while

Table 12: Impact of training on different audio lengths, and caption types on mean MAEB-30 performance for our model of size M-C. Domain columns average the 18 speech, 6 music, and 6 sound tasks. All models were trained on AudioWeb-500M with a 120 second audio length limit unless stated otherwise.

	Samples	Variant	FLOPs (10^{20})	MAEB-30	Speech	Music	Sound
Baseline	256M	up to 120 s	4.97	43.0	40.2	68.4	26.2
Audio length	256M	drop audio above 30 s	**1.44**	**47.1**	**45.8**	**72.3**	26.1
	256M	crop to 30 s	1.76	46.5	45.1	70.6	**26.4**
	256M	short-clip oversampling	3.18	41.9	39.0	67.3	24.8
Captions	128M	50% short, 50% long captions	2.48	**39.1**	**34.9**	66.3	**24.6**
	128M	100% short captions	2.48	38.6	34.1	**66.7**	23.9
	128M	100% long captions	2.48	38.5	34.5	65.7	23.4

keeping the 120-second bucket. As shown in Table 12, this oversampling approach does not produce better results with 41.9% mean MAEB score. We fix the policy that drops longer audio samples for the model-size grid. It leaves 364,955,979 of the 500,104,811 audio samples.

Caption ratio: Training on shorter or longer captions typically leads to performance trade-offs in downstream benchmarks. We tested to train only on the short or long caption provided by our dataset. We observed that no caption policy is better than another within run-to-run variation. At 128M samples the fixed per-clip mix of short and long captions scores 39.1%, short captions only 38.6%, and long captions only 38.5% on the MAEB benchmark (Table 12). Given these results and the fact that an even split offers the most caption diversity, we fix the default to an even split for further experiments.

Data domain: Our training dataset has significantly more speech data than any other domain. To test whether this is positive or negative, we tried balancing the dataset evenly between speech, music and others. We do this by dropping samples during loading, with probabilities of 0% for music, 40% for speech, and 4% for other classes. This keeps the data pool the same and oversamples to achieve a more even training data distribution. Training a model on 256 million rebalanced samples, scores 41.7% while reaching 43.0% on MAEB without it. We therefore preserve the dataset's original data distribution for other experiments.

G HEAR AND MSEB RESULTS

Table 13 reports the HEAR scores of Figure 5 by domain and task type. Table 14 reports the English tasks of the Massive Sound Embedding Benchmark (MSEB) (Heigold et al., 2026) for most models of Table 2.

Table 13: HEAR benchmark results for the models of Table 2 and the audio-only encoders of Figure 5, with the same embeddings as there. We report the mean over 17 tasks (all HEAR tasks except Beehive and Maestro 5h), domain-specific means and the three HEAR task types: clip classification, multi-label tagging (FSD50K) and timestamp event detection (DCASE 2016). Rows marked * are scores reported by the model's authors or the official HEAR 2021 leaderboard. Within each group models are sorted by the mean, and only models with every value available are listed (GURA: H = HuBERT, w = wav2vec2, C = CREPE). All scores are in %. The best value in each column is bolded, separately for the audio-only models and for the audio-text models together with ours.

			Domain			Task type		
Model	**Params**	**Mean**	**Sound**	**Speech**	**Music**	**Clf.**	**Tagging**	**Events**
Audio-only models								
USAD 2.0-XXLarge+*	1.0B	**86.8**	**74.7**	**90.8**	92.6	**87.9**	62.5	95.0
USAD 2.0-XLarge+*	695M	**86.8**	**74.7**	90.6	92.7	87.8	62.8	95.3
USAD 2.0-Large+*	336M	86.4	**74.7**	90.3	92.1	87.4	62.2	95.3
AF3 Whisper encoder*	637M	85.6	74.6	89.2	90.9	86.6	61.4	93.8
USAD 2.0-Large*	336M	84.8	72.1	85.4	**93.4**	86.0	56.0	95.0
USAD 2.0-XLarge*	695M	84.5	72.3	84.7	93.0	85.5	57.1	**96.2**
USAD 2.0-Base*	97M	83.8	70.4	84.8	92.8	84.9	55.7	95.1
Dasheng-1.2B*	1.2B	83.7	70.4	87.0	90.8	85.0	54.2	94.2
USAD 2.0-Small*	25M	83.0	69.5	84.3	91.7	84.2	53.7	94.3
USAD-Large*	330M	81.8	69.3	85.7	87.9	82.9	53.0	93.9
USAD-Base*	94M	81.4	68.3	85.8	87.7	82.6	52.2	93.6
Dasheng-Base*	86M	81.4	67.5	83.6	89.7	82.6	51.0	93.9
USAD-Small*	24M	80.1	67.1	84.4	86.4	81.4	51.1	89.5
GURA Fuse Cat H+w+C*	1.3B	79.6	62.2	81.9	90.4	81.9	42.0	82.6
RedRice ced_base*	86M	79.0	73.3	70.4	89.2	79.0	**65.5**	92.2
Hybrid BYOL-S/CvT*	5M	77.9	67.5	76.3	86.5	78.9	52.0	88.9
GURA Fuse Cat H+w+C (time)*	1.3B	77.8	59.5	79.0	90.1	80.2	37.4	82.6
RedRice ced_small*	22M	77.1	73.5	67.0	86.9	77.0	64.3	91.6
BYOL-S/CvT*	5M	76.2	64.6	74.7	85.5	77.2	48.3	89.1
BEATs Iter3+*	90M	75.7	62.3	71.2	88.6	79.0	63.6	39.6
RedRice ced_mini*	9.6M	75.4	71.3	64.7	85.9	75.1	63.9	90.7
GURA Fuse Hubert*	963M	75.2	61.9	81.1	80.5	77.0	41.3	82.6
CP-JKU mn40_as (e1+l+all_se)*	68.4M	73.4	70.4	66.1	80.9	73.6	63.1	81.3
Logitech AI SERAB BYOL-S*	5.3M	73.4	59.5	75.2	82.0	75.5	50.9	64.2
GURA Fuse wav2vec2*	315M	73.2	60.7	79.5	77.5	74.9	40.3	79.8
GURA Cat H+w+C*	1.3B	72.3	50.0	72.2	88.3	75.3	31.4	68.1
OpenL3*	4.7M	70.8	61.2	59.3	86.0	71.7	44.7	83.3
CP-JKU PaSST 2lvl+mel*	86M	70.0	72.7	57.0	77.2	68.8	64.1	92.5
CP-JKU PaSST 2lvl*	86M	69.9	72.5	57.0	77.2	68.8	64.1	91.3
CP-JKU PaSST base*	86M	69.1	70.0	57.0	77.2	68.8	64.1	78.8
GURA Avg H+w+C*	1.3B	68.1	45.6	64.1	87.1	71.3	26.4	62.4
wav2vec2*	315M	67.7	49.9	71.2	78.0	70.0	34.2	66.3
GURA Hubert*	963M	67.5	51.7	77.6	71.5	70.5	31.4	58.4
GURA Avg Hubert+Crepe*	985M	66.7	44.5	60.3	87.1	69.8	25.3	61.0
GURA Cat Hubert+wav2vec2*	631M	66.3	47.9	77.9	71.2	70.0	32.3	45.2
GURA Cat wav2vec2+crepe*	338M	66.1	41.4	62.9	86.1	69.5	23.4	58.5
RedRice EfficientNet-B2*	7.7M	66.0	67.0	54.6	73.6	65.5	60.7	79.0
GURA Avg Hubert+wav2vec2*	631M	65.5	45.9	77.4	71.0	69.4	31.8	41.7
IUT-CSE MLP (keyword)*	0.42M	61.1	41.2	60.2	75.8	64.5	18.7	51.8
CP-JKU mn04_as (e1+l)*	0.98M	60.6	63.1	44.4	70.4	60.6	56.7	65.0
Stellenbosch LSL DBERT*	317M	58.3	40.2	51.0	76.3	62.6	26.3	24.6
CREPE*	22.2M	54.6	37.6	28.3	85.6	57.5	15.9	50.4
AMAAI wav2vec2 music+speech*	95.0M	48.6	42.8	45.2	55.3	50.4	19.8	51.0
IUT-CSE MLP (audio)*	0.21M	44.3	22.4	36.8	65.4	49.3	8.6	5.2

Table 13, continued

Model	Params	Mean	Domain			Task type		
			Sound	Speech	Music	Clf.	Tagging	Events
Audio-text models								
jina-embeddings-v5-omni-nano	986M	80.7	71.7	81.8	86.4	81.5	59.6	89.8
M2D-CLAP	199M	80.6	**75.2**	72.2	**90.4**	80.7	68.1	92.3
BidirLM-Omni-2.5B	2.4B	80.3	69.0	82.3	86.8	81.5	52.0	90.5
LCO-Embedding-Omni-7B	8.9B	79.8	70.3	80.1	86.3	80.1	65.9	88.4
LCO-Embedding-Omni-3B	4.7B	79.5	72.4	77.5	86.1	79.9	64.3	89.2
Qwen2-Audio-7B	7.0B	79.5	68.5	81.8	85.7	80.0	62.2	88.8
jina-embeddings-v5-omni-small	1.6B	78.1	69.9	78.9	83.5	78.6	58.4	90.3
e5-omni-7B	8.9B	77.9	72.3	81.2	79.5	78.1	59.0	**93.1**
e5-omni-3B	4.7B	76.8	70.2	78.4	80.4	77.2	58.4	90.1
OmniEmbed-v0.1	8.9B	76.3	70.3	80.2	77.8	76.4	58.4	91.8
Qwen2.5-Omni-3B	5.5B	74.1	66.2	72.8	80.7	74.8	52.7	84.8
MS-CLAP 2023	160M	72.7	70.7	62.3	81.5	72.1	69.2	85.6
GLAP	855M	71.8	67.2	70.7	75.8	71.4	61.7	87.8
fusion-embedding-2-2b-preview	2.8B	71.2	69.0	73.0	71.6	71.0	63.1	82.3
MuQ-MuLan-large	630M	69.3	57.7	58.6	85.1	70.1	41.7	84.4
SpeechT5 multimodal	298M	67.3	57.3	74.5	69.3	68.5	31.8	84.9
LAION-CLAP HTSAT unfused	153M	67.1	70.2	56.0	72.7	65.3	70.5	89.3
LAION-CLAP HTSAT fused	154M	66.8	70.9	56.6	71.2	65.2	68.6	88.8
LAION Larger CLAP (music+speech)	194M	66.4	72.6	49.3	74.1	64.5	**70.7**	89.5
LAION Larger CLAP	194M	65.7	70.5	48.2	74.7	63.9	70.6	87.6
EBind (full)	1.8B	65.0	61.4	48.3	79.6	64.7	54.2	81.2
MS-CLAP 2022	196M	61.7	64.9	49.8	68.0	60.6	57.6	82.8
wav2clip	163M	56.5	54.5	37.4	71.4	57.0	35.1	70.1
Descript/MARL Wav2CLIP*	163M	53.5	42.7	37.9	72.5	58.3	36.2	0.0
LAION Larger CLAP (music)	194M	49.3	33.0	36.6	70.1	51.8	9.8	52.3
Ours XXS	27M	79.1	71.2	81.6	82.9	79.8	59.3	87.5
Ours XS	48M	80.6	73.5	83.5	83.6	81.4	60.1	88.8
Ours S	104M	81.8	73.4	84.5	85.9	82.7	60.9	89.9
Ours M	153M	81.4	73.5	84.6	84.7	82.1	61.5	89.5
Ours L	340M	**82.0**	73.3	**84.8**	86.3	**82.8**	61.2	91.3

Table 14: MSEB benchmark results on the English tasks, for most models of Table 2 (same parameter counts and checkpoints). All rows are our runs of the MSEB evaluation code with L2-normalised audio embeddings. We report the mean over 62 higher-is-better scores from 52 tasks (speaker clustering scores age, gender and speaker identity separately), domain-specific means (40 speech, 2 environmental sound, 20 bird sound scores) and task-specific means: classification (16), clustering (26), reranking (5), reasoning (5), and retrieval (10). Stability (Stab., 5 tasks) is the continuous edit distance between the embeddings of clean and noise-augmented speech; it is lower-is-better and not part of the mean. Speech transcription and salient-term segmentation are not reported, since they score text output that an embedding model does not produce. Multilingual tasks, the image–speech retrieval tasks, and the full-index document retrieval (we use the small index) are not run. All scores are in %. The best value in each column is bolded.

			Domain			Task type					
Model	**Params**	**Mean**	**Speech**	**Sound**	**Bird**	**Clf.**	**Clust.**	**Rerank.**	**Reason.**	**Retr.**	**Stab.↓**
M2D-CLAP	199M	**30.3**	**35.7**	52.5	**17.1**	**34.6**	**50.6**	1.4	0.0	0.2	9.4
LAION-CLAP HTSAT unfused	153M	28.0	32.7	56.5	15.8	33.8	45.7	1.7	0.0	0.0	13.0
LAION Larger CLAP	194M	27.8	32.0	**57.1**	16.4	32.0	46.2	1.6	0.0	0.0	23.2
MS-CLAP 2023	160M	26.1	29.8	54.4	15.8	27.1	45.2	1.6	0.0	0.1	14.3
LCO-Embedding-Omni-7B	8.9B	26.1	30.5	55.7	14.2	23.4	12.7	41.7	0.8	**69.8**	13.5
e5-omni-7B	8.9B	25.7	31.2	27.3	14.5	21.0	16.1	**62.8**	14.2	45.4	10.7
LCO-Embedding-Omni-3B	4.7B	25.6	29.8	53.7	14.4	20.2	13.0	55.5	0.5	64.7	13.8
e5-omni-3B	4.7B	25.4	30.8	26.7	14.4	18.6	15.7	53.7	**16.5**	51.5	9.3
BidirLM-Omni-2.5B	2.4B	24.5	28.7	34.1	15.0	33.4	15.0	32.5	0.0	43.1	13.7
jina-embeddings-v5-omni-nano	986M	20.9	22.8	43.5	14.9	25.0	14.8	41.7	0.0	30.4	15.3
jina-embeddings-v5-omni-small	1.6B	20.6	22.9	41.9	13.8	25.9	12.4	42.5	0.2	32.4	16.5
SpeechT5 multimodal	298M	15.9	18.0	16.6	11.8	18.7	26.1	1.7	0.0	0.2	9.5
GLAP	855M	14.9	15.0	35.5	12.8	20.0	16.5	34.6	0.0	0.5	12.6
Ours XXS	27M	27.4	32.4	36.2	16.4	32.9	39.0	30.2	0.0	0.8	**8.1**
Ours XS	48M	27.0	31.6	36.6	16.8	34.3	37.6	28.6	0.0	0.6	8.4
Ours S	104M	29.2	35.0	37.5	16.7	33.6	39.7	39.9	0.0	4.1	8.7
Ours M	153M	28.7	34.4	37.1	16.6	33.5	37.8	44.3	0.0	4.2	8.8
Ours L	340M	29.1	35.1	38.8	16.2	33.6	40.8	34.2	0.0	3.5	9.9

H DATASET CATALOGUE

The catalogue lists each of the 467 sources with at least 10 AudioWeb-365M audio samples at three stages: as stored audio in the archive, as selected audio in AudioWeb-500M, and as audio samples of at most 30 seconds in AudioWeb-365M. The archive counts come from the collection of stored audio samples with their measured durations. The archive lists 1.46B of the 1.48B stored samples. Rows that link to the same source page are combined into one row.

Sources are grouped as academic datasets, internet datasets, web crawl and web collections, or YouTube-derived data. A source is academic when a citable publication was found for it: an arXiv preprint, a DOI that resolves to a paper, or a citation found on the source or derivative page. A source is an internet dataset when the same search did not find such publication, which is the usual case for a Hugging Face upload released without a paper. Websites such as Kaggle uploads count as web-crawled data unless the Kaggle page cites a paper that introduces the data, and so do uploads whose page says nothing about the data. The distinction of academic dataset, internet dataset, etc. describes the evidence we could find, and a later publication can move a source between the two groups. A paper that only uses the data does not count. Only the YouTube entry forms the YouTube-derived group, as other datasets that built from YouTube but are their own publication, such as AudioSet and VGGSound, are academic.

In total, about 88.3% (323.3 million) of the AudioWeb-365M contains a Creative Commons license. When counting relatively per data source, about 59% have a Creative Commons declaration: 23% CC BY, 12% CC BY-SA, 15% variants with non-commercial or no-derivatives restrictions, 5% CC0, and 4% mixed or unspecified Creative Commons terms. About 8% list software licenses such as Apache, MIT or GPL, 1.5% public-domain declarations, and 19% other or source-specific terms. The remaining 12% do not have a verified license declaration. These percentages count each source equally, without weighting by audio samples or hours. They summarize recorded declarations, which can include audio, annotations, metadata, or code. About 12% of all sources rely on inventory declarations that have not been verified on a source page.

Table 15: Sources, stored audio samples and indexed audio hours by group, at three stages: the archive as collected, AudioWeb-500M (the selection index), and AudioWeb-365M (audio samples at most 30 seconds long). A clip re-cut from a recording above 120 seconds is charged once to its source in the last two stages. The archive counts the recording as collected. Rows count the sources listed in Table 16, those with at least 10 AudioWeb-365M audio samples. The deduplicated pool between the first two stages is given in Figure 2.

		Archive		AudioWeb-500M		AudioWeb-365M	
Source group	Sources	Audio	Hours	Audio	Hours	Audio	Hours
Academic datasets	338	84,008,598	230,340	32,903,945	84,218	31,329,824	58,894
Internet datasets	53	12,063,788	31,509	4,442,125	13,813	4,064,528	10,515
Web crawl and web collections	75	185,052,887	1,386,211	38,695,929	427,660	20,774,516	101,168
YouTube-derived data	1	1,179,581,779	9,420,695	423,263,373	2,814,586	309,935,327	1,204,563
Total	467	1,460,707,052	11,068,755	499,305,372	3,340,276	366,104,195	1,375,140

Table 16: Complete dataset catalogue, alphabetized inside each group. *Archive* counts every clip stored in the collection before deduplication. *AudioWeb-500M* counts the selected audio samples. *AudioWeb-365M* counts the subset of audio samples of at most 30 seconds. This catalogue includes 2,723,019 recovered audio samples. A source counts as academic if it is linked to a publication, in the broadest sense. Otherwise it counts as an internet source. Only sources with at least 10 AudioWeb-365M audio samples are listed. Another 39 sources contribute between one and nine audio samples each. Names link to source pages where the dataset is found, or in some cases the paper itself.

	Archive		AudioWeb-500M		AudioWeb-365M	
Dataset	Audio	Hours	Audio	Hours	Audio	Hours
Academic datasets — 338 sources						
4Q Audio Emotion	2,878	23.8	140	1.2	58	0.5
AAM	14,741	464.6	8,576	253.7	340	1.9
ACalt4	120,464	330.0	1,434	3.9	1,434	3.9

Table 16 continued.

	Archive		AudioWeb-500M		AudioWeb-365M	
Dataset	Audio	Hours	Audio	Hours	Audio	Hours
ACM MIRUM	1,312	13.5	1,199	12.3	691	5.8
AdoVoc Pro	2,111	1.4	20	< 0.1	20	< 0.1
AED	10,083	26.3	4,561	12.1	4,561	12.1
AESDD Greek	604	0.7	222	0.3	222	0.3
AfriSpeech 200	57,762	170.2	9,658	34.1	9,415	31.0
AIR Aachen	214	0.2	51	0.1	51	0.1
AISHELL 3	88,035	85.6	53,072	54.6	53,072	54.6
AISHELL-1	141,525	178.9	72,343	97.9	72,343	97.9
AISHELL-4	79,402	84.2	35,998	45.9	35,995	45.9
ALFFA	41,403	53.9	24,890	34.1	24,890	34.1
AliMeeting	195,672	157.1	71,713	77.6	71,713	77.6
Althingi Icelandic	199,609	542.4	34,196	85.9	34,196	85.9
AMI Corpus	134,237	95.5	4,682	3.2	4,682	3.2
AnimalSpeak	853,724	5,733.6	149,656	1,289.8	100,195	348.1
AnuraSet	90,919	75.8	73,543	61.3	73,543	61.3
Arabic Speech Corpus	3,000	5.2	398	0.5	398	0.5
ARCA23K	17,919	39.4	5,153	11.8	5,150	11.8
ArcticBirdSounds	796	3.6	416	3.3	386	3.0
Argentine Spanish	8,048	11.3	34	< 0.1	34	< 0.1
Armenian Speech Crowdsourcing	28,548	67.3	12,430	29.6	12,430	29.6
ASVP ESD	12,619	15.1	5,910	9.3	5,910	9.3
Att-HACK	35,694	26.5	1,700	1.6	1,700	1.6
AudioAlpaca	15,025	42.7	7,974	22.7	7,974	22.7
AudioCaps	40,934	112.1	1,792	4.9	1,792	4.9
AudioCaption	23,049	67.9	1,879	5.5	1,879	5.5
AudioEgoVLP	8,938	6.2	3,259	4.4	3,225	3.9
AudioGrounding	4,750	13.2	264	0.7	264	0.7
AudioSet	1,953,487	5,379.0	1,027,258	2,828.6	1,027,258	2,828.6
AudSem	2,036,420	1,704.0	904,625	853.4	904,625	853.4
Auto-ACD	1,920,699	5,293.4	302,128	833.5	302,128	833.5
AVCaps	6,099	85.8	1,621	22.4	374	2.6
AVE Dataset	4,142	9.5	2,089	4.4	2,089	4.4
BD-4SK-ASR	1,090	1.1	394	0.4	394	0.4
Bengalese Finch and Munia Song	3,110	15.8	1,182	6.6	1,064	4.1
Bengalese Finch song repository	3,546	14.7	3,484	14.2	3,358	12.8
Bengalese finches 2016	2,965	9.8	2,939	9.8	2,939	9.8
Bengali ASR	218,703	215.5	95,792	103.3	95,792	103.3
Bengali TTS	6,514	9.9	411	0.6	411	0.6
BibleTTS	128,840	362.8	91,988	254.8	91,988	254.8
BirdSet	597,703	3,516.7	447,242	2,647.8	445,425	2,632.6
BirdVox-ANAFCC	67,167	5.2	33	< 0.1	33	< 0.1
BirdVox-full-night	11,079	18.9	8,668	15.3	8,586	13.9
BRSpeech-DF	479,058	1,032.9	132,728	323.8	132,697	323.4
BSD10k	10,201	32.4	7,207	27.4	7,207	27.4
Burmese Speech	2,530	4.1	1,355	2.3	1,355	2.3
BVD	51,488,014	156,045.0	18,827,986	51,645.6	17,541,678	30,968.6
CadenzaCLIP	19,954	24.8	5,994	7.9	5,994	7.9
CaFE French	936	1.2	314	0.4	314	0.4
CAPTDURE	8,034	20.3	2,004	4.7	2,004	4.7
Catalan Speech	4,240	9.4	1,330	2.8	1,330	2.8
CatMeows	483	0.3	183	0.2	183	0.2
ccMixter	439	10.9	361	8.1	30	0.1
CCMusic	225	1.2	61	0.5	49	0.2
CCOM HuQin	64,230	21.9	161	1.3	108	0.1
CETUC	100,000	143.3	25,584	40.2	25,584	40.2
Chilean Spanish	7,136	11.8	90	0.2	90	0.2
CHiME-Home	12,241	13.6	7,366	8.2	7,366	8.2
Choral Singing Dataset	222	5.8	179	3.6	20	0.1
ChoraleBricks	392	4.4	337	3.8	60	0.4
CitySpeechMix	1,484	4.0	657	1.8	657	1.8
CloserMusicDB	375	8.2	342	6.9	14	0.1
CMU AN4	1,078	0.8	182	0.2	182	0.2
CMU-MOSEI	22,838	48.4	528	1.4	521	1.3
CMU-MOSI	6,597	7.8	37	0.1	37	0.1
CN-Celeb	668,901	1,381.4	285,833	594.3	281,207	533.1
CochlScene	76,089	211.4	67,526	187.6	67,526	187.6
Colombian Spanish	7,847	12.1	140	0.2	140	0.2
Common Voice 22.0	4,063,218	5,526.7	3,889,615	5,319.9	3,889,520	5,318.1
CommonVoice FR	651,691	897.2	284,723	445.1	284,721	445.0
CompMusic	6,519	7.2	654	5.5	481	1.3
Coswara	23,508	65.8	15,942	47.1	15,914	46.8
COUGHVID	46,955	106.4	28,320	66.9	28,304	66.6
covers80	537	12.2	521	11.1	19	0.1
Crowd Noise Density	99	0.4	97	0.4	97	0.4
CSS10	128,586	281.3	22,611	51.0	22,610	51.0
Dagstuhl ChoirSet	2,389	19.9	1,406	14.6	883	2.6
DataSEC	3,226	14.6	2,124	12.8	1,586	5.0
DataSED	717	14.9	594	12.8	113	0.6
DAVE	2,426	16.4	1,771	11.9	1,506	9.2
DB3V	11,469	25.5	10,795	24.0	10,795	24.0
DCASE 2013 Event Detection	11,946	12.6	1,972	3.2	1,972	3.2
DCASE 2013 Scenes	198	1.6	126	1.1	126	1.1
DCASE 2024 Task 7	310	0.3	299	0.3	299	0.3

Table 16 continued.

	Archive		AudioWeb-500M		AudioWeb-365M	
Dataset	Audio	Hours	Audio	Hours	Audio	Hours
DCASE2018 Task5	250,125	694.7	107,264	297.9	107,264	297.9
DCASE2025 ASD	17,306	44.8	13,135	33.6	13,135	33.6
DCASE23 Task2	7,425	20.6	155	0.4	155	0.4
DEMAND	288	2.4	249	2.1	249	2.1
DESED Synthetic	2,045	5.7	1,950	5.4	1,950	5.4
DISCO	8,095	2.2	1,683	0.5	1,683	0.5
Domestic Canary	456	3.7	456	3.7	237	1.3
DSD100	405	12.8	403	11.3	38	0.2
E-GMD	45,193	319.0	41,779	303.2	30,203	41.3
EARS	17,227	87.8	6,434	26.1	6,313	23.9
EarSAVAS	121,737	127.0	56,181	112.6	53,319	63.9
EasyCall Corpus	21,385	12.3	3,043	2.3	3,041	2.3
EDANSA-2019	10,771	29.9	10,154	28.2	10,154	28.2
El Bongosero	8,629	13.1	139	0.2	139	0.2
Emo-Soundscapes	6,463	11.0	129	0.2	129	0.2
EMODB	535	0.4	79	0.1	79	0.1
EmoGator	32,054	16.7	6,015	3.2	6,015	3.2
EmoMusic	408	1.4	36	0.1	36	0.1
EmoSynth	933	1.9	37	0.1	37	0.1
EmoV-DB	6,893	9.5	1,703	2.9	1,703	2.9
Emozionalmente	34,987	32.0	69	0.1	69	0.1
English Dialects	17,877	31.3	1,547	3.7	1,547	3.7
EPIC-Sounds	12,880	52.3	11,985	48.8	11,985	48.8
ErhuPT	2,523	0.8	650	0.3	650	0.3
European starling	1,255	15.3	1,199	14.5	276	1.8
EuroSpeech	1,288,190	5,359.5	139,236	580.8	139,236	580.8
Expresso	39,426	68.4	6,247	20.9	5,904	14.3
FakeAVCeleb	21,560	31.9	97	0.2	97	0.2
FAVDBench	9,643	41.6	7,951	16.7	7,951	16.7
FLEURS	37,327	118.1	12,818	40.7	12,810	40.6
FMA	1,152	9.5	1,023	8.5	580	4.8
Freefield1010	13,842	38.5	2,624	7.3	2,624	7.3
Freiburg106	1,516	0.9	771	0.5	767	0.4
FSD50K	51,090	108.2	38,296	95.4	38,289	95.2
FSDKaggle2018	11,073	20.2	4,384	10.1	4,384	10.1
FSDKaggle2019	29,233	103.6	16,435	66.4	16,435	66.4
Galician Speech	5,587	10.3	373	0.8	373	0.8
Giant Otter	453	0.1	104	< 0.1	104	< 0.1
Golos	400,000	431.8	42,163	56.0	42,153	55.9
Gram Vaani	199,999	1,966.0	62,136	489.9	34,197	128.9
Greatest Hits	10,642	4.6	5,463	3.6	5,463	3.6
Grid AV Speech	34,000	17.4	448	0.2	448	0.2
Groove MIDI Dataset	1,083	7.8	1,068	7.7	781	1.1
GuitarSet	3,163	26.9	244	2.1	112	0.8
Gujarati Speech	7,409	13.7	1,155	2.0	1,155	2.0
HD-EPIC	50,968	50.9	19,153	30.0	18,499	19.8
Hi-Fi TTS	323,978	291.7	19,341	24.0	19,341	24.0
HI-MIA	1,322,843	533.6	429,880	180.1	429,880	180.1
HI-MIA-CW	16,290	6.6	3,700	1.5	3,700	1.5
HLS-CMDS	538	2.2	290	1.2	290	1.2
Homburg	5,869	16.4	304	0.8	304	0.8
Hospital Ambient Noise	1,124	1.6	964	1.3	964	1.3
HumBugDB	8,830	31.7	6,648	26.4	5,564	12.4
Iban ASR	5,444	14.9	2,673	8.3	2,668	8.3
ICBHI Respiratory	2,760	16.5	899	5.5	847	4.7
IDMT-SMT-Audio-Effects	54,014	30.0	38,644	21.5	38,644	21.5
IDMT-SMT-Bass	18,174	12.6	364	0.2	364	0.2
IDMT-SMT-Drums	2,795	10.2	108	0.4	108	0.4
IDMT-SMT-Guitar	5,898	28.3	159	1.6	55	0.3
IISc MILE Kannada	122,052	402.4	61,293	200.2	61,205	199.3
IISc-MILE Tamil	54,000	92.9	8,370	14.1	8,366	14.0
Improvising Duos	31	0.3	28	0.3	12	0.1
iNatSounds	231,088	1,223.2	214,186	1,135.3	177,562	640.3
Individual Bird Identification	28,001	39.5	13,125	20.7	13,084	20.1
InsectSet32	480	1.4	292	0.9	286	0.8
InsectSet47 & InsectSet66	7,956	54.4	309	8.9	18	0.1
IRMAS	7,201	33.7	834	3.9	834	3.9
ISMIR 2004 Genre	1,375	40.5	1,392	37.0	84	0.5
ISMIR2004Tempo	3,221	23.1	129	0.9	86	0.5
Isolated Urban Sound	393	4.7	212	3.0	88	0.4
IsoVAT	353	1.1	24	0.1	24	0.1
Jam-ALT	152	4.8	149	4.1	15	0.1
jazznet	18,464	45.3	31	0.1	31	0.1
JGDB	3,124	8.7	47	0.1	47	0.1
JL Corpus	12,160	7.1	115	0.1	115	0.1
JSUT Japanese	7,696	10.3	6,200	8.2	6,200	8.2
Kannada Speech	8,800	17.0	363	0.4	363	0.4
Kazakh Speech	153,853	332.6	84,831	197.3	84,830	197.3
Khmer TTS	2,906	4.0	1,216	1.7	1,216	1.7
kinetics	550,525	1,372.5	438,720	1,115.8	438,720	1,115.8
Kitchen20	1,070	1.5	944	1.3	944	1.3
Knocking Emotional	500	0.2	375	0.2	375	0.2
laions got talent	1,685,807	5,318.4	782,227	2,714.5	719,555	2,103.4

Table 16 continued.

	Archive		AudioWeb-500M		AudioWeb-365M	
Dataset	Audio	Hours	Audio	Hours	Audio	Hours
Libriheavy HQ	200,000	825.2	9,916	39.6	9,524	36.3
LibriSpeech	39,665	121.8	982	2.9	982	2.9
LibriTTS	247,318	397.1	3,258	5.7	3,255	5.7
LinTO Arabic Tunisian	6,144	2.8	203	0.1	203	0.1
LM-SSD	346	11.0	338	8.9	54	0.2
LP-MusicCaps-MTT	21,637	175.0	16,086	130.1	16,086	130.1
Macaques	5,018	0.6	21	< 0.1	21	< 0.1
MACS	20,825	57.8	864	2.4	864	2.4
MAD Military	7,466	11.0	6,074	9.2	6,074	9.2
MagicData-RAMC	71,572	52.4	32,408	31.8	32,408	31.8
MagnaTagATune	25,858	209.2	7,529	60.9	7,529	60.9
Malayalam Speech	4,126	5.5	1,225	1.9	1,225	1.9
Malromur Icelandic	108,566	136.1	38,054	49.0	38,054	49.0
Manatee Vocalizations	7,392	8.3	6,374	8.2	6,365	8.1
ManaTTS	19,303	25.7	4,481	7.4	4,481	7.4
Marathi Speech	1,569	3.0	489	1.1	489	1.1
Marmosets2016	321	0.3	262	0.3	262	0.3
MAST Rhythm	17,262	22.0	307	0.4	307	0.4
MAVD Traffic	165	1.9	73	0.8	65	0.5
MC Speech	24,018	22.1	987	1.2	987	1.2
MDBDrums	517	6.7	230	3.7	28	0.2
MediaSpeech	10,023	40.0	4,320	17.2	4,320	17.2
Medley-solos-DB	21,566	17.8	12,293	10.1	12,293	10.1
MELD	12,453	11.1	1,752	2.0	1,752	2.0
MIMII	54,056	150.2	51,792	143.9	51,792	143.9
MIMII DUE	36,433	101.2	35,442	98.5	35,442	98.5
MiniMax TTS Multilingual	144	0.4	20	0.1	20	0.1
MIR-1K	6,240	27.1	396	4.4	97	0.2
MMAR	1,000	5.5	607	3.4	592	3.3
MobvoiHotwords	286,580	262.1	93,384	108.1	93,384	108.1
MS-SNSD	24,337	24.8	24	< 0.1	24	< 0.1
MSWC	200,000	62.3	9,425	2.8	9,425	2.8
MUCS Multilingual ASR	268,658	324.4	124,374	157.1	124,372	157.1
Multilingual LibriSpeech	40,000	166.5	13,356	56.1	13,356	56.1
Multilingual TEDx	41,791	67.8	7,958	17.5	7,847	16.1
MUSAN	1,516	23.6	1,325	19.2	640	2.0
MUSDB18-HQ	750	23.5	518	16.7	10	0.1
Music Arena	5,703	114.7	5,033	102.6	2,224	15.7
MusicBench	53,168	144.9	34,543	94.1	34,543	94.1
MusicCaps	5,020	26.2	3,678	19.6	3,678	19.6
MUSIFEAST-17	650	5.4	181	1.5	181	1.5
NABat	20,652	35.2	20,310	34.7	20,310	34.7
NAR	852	0.1	46	< 0.1	46	< 0.1
NCHLT Afrikaans	3,953	3.2	1,057	0.9	1,057	0.9
NCHLT isiXhosa	3,119	2.9	2,583	2.5	2,583	2.5
NCHLT isiZulu	3,066	3.6	2,868	3.4	2,868	3.4
NCHLT Sesotho	4,160	3.7	3,167	3.0	3,167	3.0
NCHLT Setswana	3,866	3.5	3,246	3.0	3,246	3.0
NCHLT Tshivenda	3,476	4.2	3,035	3.7	3,035	3.7
nEMO	4,481	3.1	274	0.2	274	0.2
Nepali ASR	157,905	154.6	86,447	89.0	86,447	89.0
Nepali TTS	2,064	2.8	246	0.3	246	0.3
Nepali TTS (Male and Female)	675	1.2	210	0.5	210	0.5
Neural Audio Fingerprint	25,602	193.1	19,556	147.8	17,003	125.8
NIGENS	893	4.5	532	3.5	372	1.0
Nigerian Pidgin ASR	3,633	7.3	1,156	2.5	1,153	2.4
NonSpeech7k	7,014	6.8	3,501	3.6	3,501	3.6
NonverbalTTS	9,894	26.9	323	1.1	319	1.1
North American Bird Species	1,064	0.1	43	< 0.1	43	< 0.1
NSynth	432,593	480.7	234,252	260.3	234,252	260.3
ODAQ	377	1.3	204	0.7	204	0.7
Onset Leveau	40	0.2	10	< 0.1	10	< 0.1
Open-AoE	8,901	69.0	8,884	68.9	8,366	64.3
OpenAIR IRs	26	< 0.1	22	< 0.1	22	< 0.1
OpenMIC-2018	20,008	55.6	17,853	49.6	17,853	49.6
OpenSet FewShot Audio	2,720	3.0	522	0.6	522	0.6
Orchset	744	4.6	32	0.2	29	0.2
Panama Katydid	3,029	5.3	1,930	3.9	1,851	2.4
Pansori-TEDxKR	5,414	4.9	1,672	1.8	1,672	1.8
People's Speech	200,000	757.6	3,264	10.4	3,264	10.4
Percussion Articulation	5,233	17.5	185	0.6	185	0.6
Peruvian Spanish	22,936	38.8	91	0.2	91	0.2
Phonation Modes	22,371	35.8	242	5.5	58	0.3
PhysioNet Heart2016	3,240	20.2	3,196	19.9	2,354	11.0
PicidaeDataset	1,654	1.4	1,404	1.2	1,394	1.1
Porjai Thai voice	4,941	10.5	1,587	4.1	1,585	4.1
Pragmatic Similarity	1,542	2.2	202	0.3	202	0.3
Quran-MD	268,969	999.8	135,281	738.1	111,139	424.9
Quranic Audio	6,828	8.7	2,563	5.0	2,556	4.9
ReaLISED	2,479	1.0	1,258	0.6	1,258	0.6
ReefSet	3,902	2.1	3,224	1.7	3,224	1.7
RescueSpeech	2,412	1.6	232	0.2	232	0.2
RIRS_NOISES	120,488	48.5	537	2.2	485	0.9

Table 16 continued.

	Archive		AudioWeb-500M		AudioWeb-365M	
Dataset	Audio	Hours	Audio	Hours	Audio	Hours
Rockefeller Zebra Finch Songs	30,941	109.2	30,071	107.8	30,071	107.8
Russian LibriSpeech	92,000	159.6	17,175	32.2	17,175	32.2
Sagalee Oromo ASR	53,572	100.2	41,521	78.2	41,516	78.2
Samromur Icelandic	100,000	148.7	18,957	31.8	18,957	31.8
Samromur L2	143,030	151.8	25,744	31.5	25,744	31.5
Samromur Unverified	200,000	201.3	38,313	45.2	38,310	45.1
SASPEECH	16,512	25.0	8,007	14.9	8,007	14.9
SeniorTalk	60,029	37.8	19,805	17.0	19,805	17.0
Seoul Corpus	57,658	24.2	21,780	13.5	21,780	13.5
SHALCAS22A	14,580	20.0	6,184	8.8	6,184	8.8
Silbo Gomero	3,427	2.4	1,718	1.4	1,718	1.4
SINGA:PURA	10,320	28.7	1,000	2.8	1,000	2.8
Sinhala ASR	185,293	224.5	101,048	131.5	101,046	131.5
Sinhala TTS	2,496	4.1	933	1.6	933	1.6
Slakh2100	22,360	705.2	17,344	507.4	883	4.2
SMD	209	3.6	149	2.6	12	0.1
SMIIP-TV	17,476	18.5	7,380	9.6	7,380	9.6
SOMOS	20,100	26.2	88	0.1	88	0.1
SONYC-UST	18,510	51.4	13,498	37.5	13,498	37.5
SoulX-Duplug-Eval	5,063	13.9	1,022	3.1	1,022	3.1
Sound Effect Variations	3,790	4.3	2,389	2.7	2,389	2.7
SoundingEarth	14,167	166.4	11,759	135.0	1,870	15.4
Speech Commands	235,277	64.3	4,073	1.2	4,067	1.1
speechocean762	8,200	9.2	1,710	2.0	1,710	2.0
STARSS22	652	12.1	106	2.7	20	0.2
SUBESCO	7,000	7.8	3,744	4.2	3,744	4.2
Sundanese ASR	219,156	332.6	152,226	229.5	152,226	229.5
Sundanese TTS	4,213	5.4	1,886	2.6	1,886	2.6
SynBAD	11,813	7.0	3,926	2.4	3,926	2.4
TACOS	8,676	44.1	6,933	35.8	6,933	35.8
Tamil Speech	4,291	7.1	1,398	2.3	1,398	2.3
TAU SEBin 2021	98	0.4	93	0.4	93	0.4
TAU Urban 2025	178,913	49.7	96,531	26.8	96,531	26.8
TAU Urban Acoustic Scenes 2019	29,987	83.3	420	1.2	420	1.2
TAU Urban Acoustic Scenes 2019 Mobile	55,920	155.3	9,577	26.6	9,577	26.6
TAU Urban Acoustic Scenes 2019 OpenSet	15,824	44.0	2,833	7.9	2,833	7.9
TAU Urban Acoustic Scenes 2020 3Class	14,400	40.0	3,727	10.4	3,727	10.4
TAU Urban Acoustic Scenes 2020 Mobile	46,435	129.0	28,902	80.3	28,902	80.3
TAU Urban Acoustic Scenes 2022 Mobile	230,350	64.0	127,093	35.3	127,093	35.3
TEDLIUM3	269,925	458.0	6,933	13.4	6,933	13.4
Telugu Speech	4,448	5.7	2,142	3.1	2,142	3.1
THCHS30	13,388	34.2	6,586	16.9	6,586	16.9
THUYG-20	16,753	44.9	10,338	27.4	10,288	26.8
TIBMD@MUC	68,465	84.2	33,485	42.2	33,456	41.9
Timers and Such	254,149	174.1	4,294	4.2	4,294	4.2
TinySOL	14,321	26.3	128	0.2	128	0.2
TORGO	16,954	14.2	1,316	1.9	1,316	1.9
ToyADMOS	76,708	227.2	69,428	205.7	69,428	205.7
TUT Acoustic Scenes 2017	13,796	38.3	3,249	9.0	3,249	9.0
TUT Sound Events 2017	122	1.0	99	0.8	64	0.2
TUT Urban Acoustic Scenes 2018	24,880	69.1	940	2.6	940	2.6
TUT Urban Acoustic Scenes 2018 Mobile	52,790	146.6	9,387	26.1	9,387	26.1
Tuvan Throat Singing	28	0.1	25	0.1	21	0.1
Ugandan Birds	549	3.9	511	3.6	398	2.0
URBAN SED	10,000	27.8	9,547	26.5	9,547	26.5
Urbansas	1,298	1.9	1,077	1.6	1,077	1.6
URDU Emotional	400	0.3	85	0.1	85	0.1
USM-SED	10,010	13.9	7,670	10.7	7,670	10.7
VAD Multi-Species	3,782	8.1	2,633	6.3	2,559	5.4
VCC2016 Edinburgh	2,160	2.1	10	< 0.1	10	< 0.1
Venezuelan Spanish	6,188	8.9	55	0.1	55	0.1
VGGSound	196,972	545.9	71,846	199.1	71,846	199.1
VIVAE	1,565	0.4	200	0.1	200	0.1
VIVOS	24,839	31.3	8,564	11.2	8,564	11.2
VOC-ALS	1,224	5.0	1,049	4.4	999	3.9
VocalSet	11,700	28.4	727	1.7	721	1.6
VocalSketch	16,619	18.6	10,575	14.4	10,575	14.4
VocalSound	20,776	24.1	13,360	17.4	13,360	17.4
VoiceAssistant	8,192	9.2	62	0.1	62	0.1
VoxCeleb1	153,516	351.6	2,309	4.0	2,302	4.0
VoxCeleb2	1,092,000	2,360.0	78,060	127.6	77,907	125.8
Vystadial	58,918	46.3	5,153	4.8	5,152	4.8
WABAD	39,809	42.5	25,002	38.8	24,200	27.6
Watkins Marine Mammals	14,932	19.4	10,881	17.8	10,474	9.1
WavCaps	138,186	838.9	9,505	202.3	3,287	8.7
WaveFake	134,266	216.2	7,502	10.4	7,502	10.4
WavText5K	4,476	20.4	764	4.0	650	0.9
West African Radio Corpus	17,390	144.9	14,124	117.7	14,124	117.7
WLASL	7,020	5.0	549	0.3	549	0.3
WMW Bird	870	9.2	791	8.3	446	1.7
XBMU-AMDO31	22,630	31.1	19,183	26.6	19,183	26.6
YBSS-200	1,999	2.2	592	0.7	592	0.7
Yellow-cheeked crested gibbon	7,752	25.8	5,048	16.8	5,048	16.8

Table 16 continued.

	Archive		AudioWeb-500M		AudioWeb-365M	
Dataset	Audio	Hours	Audio	Hours	Audio	Hours
Yoruba Speech	3,795	4.3	1,365	1.7	1,365	1.7
YouTube-8M	1,079,666	2,999.1	662,465	1,840.2	662,465	1,840.2
Zebra Finch Vocal Repertoire	1,309	0.3	318	0.2	318	0.2
Internet datasets						
African Accented French	16,637	19.9	7,496	9.7	7,470	9.4
Animal Sound Archive	140,958	693.0	7,165	35.3	7,164	35.3
Arabic Audio Deepfake	19,559	14.3	5,391	4.4	5,391	4.4
ARU Speech Corpus	8,640	7.1	48	< 0.1	48	< 0.1
Audio Commons	52,048	83.7	38,000	70.3	37,440	52.6
Barkopedia Dog	2,235	1.1	782	0.8	779	0.7
Berklee Samples	2,378	4.8	2,011	4.5	1,936	3.6
captioned-ai-music-snippets	2,461,724	9,237.5	2,311,170	8,609.1	1,952,373	5,619.0
cc0-music-captioned	7,682	173.7	3,492	90.7	439	1.5
Conet Project	150	3.5	112	2.7	10	0.1
CWRU BearingFault	95	0.3	94	0.3	92	0.3
dawn chorus en	899	3.0	29	0.1	29	0.1
DCASE 5-Class 3-Source Separation	12,760	37.8	1,426	4.3	1,415	3.9
Deeply Korean	1,697	2.9	980	1.8	980	1.8
Deeply Nonverbal Vocal	726	0.6	506	0.4	506	0.4
Deeply Parent Child Vocal Interaction	3,634	2.8	1,408	1.8	1,387	1.5
Drone Audio Detection	180,309	59.9	13,676	28.5	13,600	26.3
EchoThief IR	115	< 0.1	55	< 0.1	55	< 0.1
emolia-balanced-5M	5,256,681	14,833.3	918,310	2,598.4	915,546	2,575.4
Emotional Speech (EQ4You)	1,796,333	2,298.7	89,605	114.5	89,605	114.5
EmotionTTS Korean Speech	9,002	13.3	115	0.2	115	0.2
emoUERJ	377	0.3	48	0.1	48	0.1
ESCorpus-PE	3,764	5.0	838	1.4	838	1.4
FolkMusic	11,966	9.9	9,544	7.9	9,544	7.9
Free ST American English	13,867	16.7	276	0.5	276	0.5
Free ST Chinese	102,600	109.7	33,481	38.0	33,481	38.0
Galgame Speech	343,189	490.8	222,929	350.0	222,929	350.0
Gowajee Thai	20,307	17.2	5,393	4.9	5,392	4.9
Heroico Spanish	37,942	33.2	4,849	4.2	4,849	4.2
Human Vocality Primitives	20	< 0.1	15	< 0.1	15	< 0.1
Khmer ASR Cultural	13,767	31.0	7,880	17.9	7,875	17.8
Kikuyu Audio	128	0.3	127	0.3	125	0.3
LJ Speech	13,100	23.9	21	< 0.1	21	< 0.1
MAGICDATA Read Speech	200,000	249.2	108,385	140.0	108,380	139.9
Malayalam SMC	1,541	1.6	331	0.4	331	0.4
Nigerian English	6,718	11.5	495	1.0	495	1.0
ParlamentParla	200,000	572.8	65,431	188.7	65,431	188.7
Primewords Chinese	49,881	98.0	36,081	70.3	36,081	70.3
ProgrammaticAudio	103,038	304.4	78,204	229.7	78,204	229.7
Quran Multi Reader	123,821	648.1	57,468	325.5	46,394	175.9
Rasheed Mudasiru	189	1.0	21	0.1	21	0.1
Samromur Mimic	124,300	81.2	10,743	9.7	10,743	9.7
synthetic vocal bursts	325,548	729.9	306,265	686.7	306,265	686.7
TEDx Spanish	11,243	24.5	3,196	7.1	3,196	7.1
Tepetzintla Nahuatl	427	13.0	346	10.7	12	< 0.1
Thorsten Emotional	11,780	14.3	152	0.3	152	0.3
Thorsten German	41,598	42.1	3,495	4.8	3,495	4.8
Tibetan Greetings	3,022	2.8	1,802	1.7	1,802	1.7
Tunisian MSA	12,288	12.6	3,287	4.1	3,285	4.1
University of Iowa MIS	2,339	23.7	970	11.6	377	2.2
VOA Ukrainian	107,912	128.5	37,555	43.7	37,513	43.2
VoxForge	202,794	320.0	40,578	73.6	40,530	72.5
Yesno	60	0.1	48	0.1	48	0.1
Web crawl and web collections						
78rpm Collection (1920s--1930s)	844	27.7	606	19.8	12	< 0.1
Adobe Audition SFX	9,294	22.1	6,037	17.5	5,483	7.1
AI Mechanic	39	0.3	23	0.2	19	0.1
AIRS Respiratory	1,757	5.4	1,677	5.2	1,630	4.7
Animal Sounds	746	0.6	89	0.1	89	0.1
Animal Voice Sound	1,300	1.0	222	0.2	221	0.2
Animals Sound	874	1.0	297	0.4	297	0.4
archive.org	166,168	2,467.7	113,613	1,937.0	44,134	111.3
Assamese TTS	3,746	2.2	370	0.2	370	0.2
Audio Classifier	7,755	10.8	485	0.7	485	0.7
Audiostock	251,634	9,919.3	94,963	1,565.6	27,182	132.8
Baby Crying	1,313	2.3	662	1.1	662	1.1
BAVED Arabic Emotions	3,870	1.8	491	0.2	491	0.2
BBC Sound Effects	7,994	55.1	6,892	47.1	4,665	20.4
Bird Species Audio	9,107	7.6	7,519	6.3	7,519	6.3
BirdCLEF 2021	18,020	236.9	16,411	213.2	7,132	34.3
BirdCLEF 2022	14,853	148.3	12,828	128.7	7,629	29.1
BirdCLEF 2023	16,942	163.8	15,498	150.9	9,015	38.0
Common Crawl	153,310,921	1,250,898.9	29,895,480	366,294.2	13,966,931	79,341.5
CommonVoice EN	195,179	225.6	2,599	3.9	2,592	3.7
Cornell Birdcall Identification	21,042	257.3	18,879	224.1	9,092	37.4
Emergency Vehicle Siren	600	0.5	367	0.3	367	0.3

Table 16 continued.

	Archive		AudioWeb-500M		AudioWeb-365M	
Dataset	Audio	Hours	Audio	Hours	Audio	Hours
Epidemic Sound	208,295	314.1	36,381	79.0	33,903	45.3
Fake-or-Real (FoR)	169,736	152.5	1,713	1.4	1,712	1.4
Fan Anomaly Sound	9,738	27.1	9,305	25.8	9,305	25.8
FindSounds	5,715	5.0	838	0.8	838	0.8
Freesound	979,059	5,606.3	85,782	92.9	85,782	92.9
FreeToUseSounds	2,844	15.3	2,326	13.7	1,634	5.2
Generic Audio Samples	17,368	24.1	11,502	16.0	11,502	16.0
Goebl Music Examples	270	1.3	235	0.6	212	0.1
Heartbeat Sounds	832	1.6	573	1.2	573	1.2
Hebrew Speech Kan	10,000	8.7	2,434	2.5	2,434	2.5
Hindi-1482Hrs	1,339	1.9	767	1.1	767	1.1
Indian Languages Audio	10,000	13.8	4,410	6.1	4,410	6.1
Jamendo	1,809,152	14,654.9	1,561,839	12,643.0	1,561,839	12,643.0
Kaggle Animal Sounds	24	< 0.1	12	< 0.1	12	< 0.1
Kaggle Music Recommendation	1,500	12.5	1,385	11.5	669	5.6
Khwarizmi Wake Word	2,586	0.7	496	0.1	496	0.1
Kids Speech	4,979	10.1	2,750	5.5	2,719	5.2
LiveATC Air Traffic Control	65	0.6	27	0.3	12	0.1
Lung Sound	336	1.6	301	1.5	301	1.5
majestrino	972,658	1,284.4	152,356	234.9	152,300	233.8
MER500	3,027	10.4	44	0.1	44	0.1
MLSP 2013 Bird	645	1.8	621	1.7	621	1.7
multilingual-in-the-wild	719,307	1,293.6	245,001	475.8	245,001	475.8
Music Berkeley Emotions	1,638	2.3	1,541	2.2	1,541	2.2
Musical Instruments	2,708	15.7	2,059	12.0	1,506	4.5
NOAA Whale	22	0.2	22	0.2	14	0.1
Nokia Ringtones	1,343	8.5	1,274	8.0	911	4.4
ODEON Sound Effects	5,089	18.3	3,098	12.8	2,681	7.2
Public Domain Sounds	600	6.1	553	5.7	323	1.2
Radio Commercials	242	4.1	103	1.6	25	0.1
SDH	18,709,371	12,234.2	2,915,977	2,150.3	2,915,973	2,150.2
SFX	521,558	2,440.6	258,196	1,463.9	196,188	427.4
Sonniss GDC Game Audio	10,733	81.8	4,161	35.7	2,914	6.0
SoundBible	1,230	3.7	252	0.3	248	0.2
sounddomains	5,650,325	82,483.9	2,939,784	39,397.8	1,191,390	4,981.3
SoundJay	1,093	3.6	109	0.3	100	0.1
Soundly	2,685	49.3	1,282	33.4	91	0.3
Speech Accent Archive	2,138	16.5	291	2.4	183	1.2
Speech Emotion Kaggle	68	0.1	21	< 0.1	21	< 0.1
Spoken Language ID	73,620	204.5	10,272	28.5	10,272	28.5
Starrail Voice	4,942	8.5	2,665	5.1	2,664	5.1
Submarine Sonar	11	0.1	10	0.1	10	0.1
TESS	5,600	3.2	79	< 0.1	79	< 0.1
TF Speech Recognition Challenge	223,265	61.9	17,645	4.9	17,645	4.9
Toads and Frogs (Anuran)	3,069	2.3	1,392	1.0	1,392	1.0
Turkish Podcasts	18,432	51.2	8,826	24.3	8,826	24.3
Ukrainian STT	690,353	450.6	108,415	111.2	108,038	106.3
USDA Insect Sounds	23	0.1	23	0.1	22	0.1
Vehicle Type Sound	1,705	1.9	1,668	1.9	1,668	1.9
We Sound Effects	457	6.5	388	5.4	192	0.5
Whale Detection Challenge	84,513	47.0	64,671	35.9	64,671	35.9
Whale Sound	49	< 0.1	10	< 0.1	10	< 0.1
Zapsplat	62,532	106.0	34,036	82.0	31,815	37.4
YouTube-derived data						
YouTube	1,179,581,779	9,420,695.0	423,263,373	2,814,585.9	309,935,327	1,204,563.5