Unlocking Strong Supervision: A Data-Centric Study of General-Purpose Audio Pre-Training Methods

Supplementary Material

1. Audio Source of Training Data

We utilize 400k audio subset from *CaptionStew* [18], a composite dataset aggregating 8 open-source collections to mitigate data scarcity and enhance diversity in audi pre-training. As outlined in Table. 1, the data spans diverse acoustic domains, including environmental sounds, music, and expressive speech.

2. Label Construction

As described in main text, our label construction pipeline comprises two distinct stages: audio caption generation and LLM-aided tag parsing. In this section, we provide comprehensive implementation details and hyperparameter configurations for both steps.

2.1. Captioner Details and Examples

We employ Qwen3-Omni-30B-A3B-Captioner [20] for audio caption generation, utilizing the vLLM engine with 8-way tensor parallelism and bfloat16 precision to ensure efficiency. The data loading pipeline is managed by Lhotse [22], which applies dynamic bucketing with a batch size of 32 to optimize GPU utilization. For generation, we use nucleus sampling with a temperature of 0.6 and top-p of 0.95 to balance diversity and coherence. We provide samples of the generated captions and corresponding audio files in the supplementary material.

2.2. LLM Parser

We employ Qwen2.5-7B-Instruct [21] as a semantic parser to extract structured tags from unstructured audio captions. Utilizing vLLM with 4-way tensor parallelism, we prompt the model to generate 5–10 concise, one-word labels per sample in JSON format. The instruction ensures coverage of key acoustic dimensions—including scene, sound source, and event type—using sampling parameters of $T = 0.3$ and $p = 0.9$. We provided the full text of the instruction prompt.

3. Evaluation Details

Table. 2 provides a comprehensive breakdown of the datasets and metrics used for evaluate audio representation quality. The assessment covers three distinct protocols: linear probing, audio-language alignment and open-form question answering.

4. Complete Results

We have the complete results on three evaluation axes in Table. 3 and Table. 4. The main additions are the results for five values of K ranging from 800 to 3k, as well as the results for the multi-task setting where λ is set to 0.25 and 1, respectively.

References

[1] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023. 3

[2] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019. 3

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 3

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008. 3

[5] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014. 3

[6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 3

[7] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. In *International Conference on Machine Learning*, pages 5178–5193. PMLR, 2023. 3

[8] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 3

[9] Anuj Diwan, Zhisheng Zheng, David Harwath, and Eunsol Choi. Scaling rich style-prompted text-to-speech datasets. *arXiv preprint arXiv:2503.04713*, 2025. 3

[10] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with

Table 1. Overview of the public datasets constituting CaptionStew. The table summarizes their scale, domain coverage, audio sources, and diverse captioning pipelines (from human annotation to LLM generation).

Dataset	#audio/#cap	Domain	Audio source	Caption style	Caption generation pipeline
AudioCaps	46k/46k	general (environmental, human/animal sounds)	AudioSet	Human-annotated, short description	crowdsourced
Clotho	5k/25k	environmental sounds	FreeSound	Human-annotated, short description	crowdsourced
MusicCaps	3k/3k	music	AudioSet	Expert musician-written, multi-sentence, fine-grained description	expert curation
WavCaps	400k/400k	general (environmental, human/animal sounds)	AudioSet BBC Sound Effect FreeSound SoundBible	LLM-refined captions	three-stage pipeline: web-crawled raw descriptions → ChatGPT rewrite → filtering
AudioSetCaps	1.9M/1.9M 4.0M/4.0M 182k/182k	general (environmental, human/animal sounds)	AudioSet YouTube8M VggSound	LLM-generated, detailed, multi-sentence description	three-stage pipeline: LALM attribute extraction → LLM captioning → CLAP-based filtering
FusionAudio	1.2M/1.2M	general (environmental, human/animal sounds)	AudioSet	LLM-augmented, multi-sentence, visual-enhanced description	multimodal context fusion (audio, visual, metadata) + LLM captioning
JamendoMaxCap	360k/1.8M	music	Jamendo Platform	LLM-augmented, multi-sentence, fine-grained music description	retrieval-based metadata imputation + LLM captioning
ParaSpeechCaps	116k/116k (base) 924k/924k (scaled)	expressive speech	VoxCeleb1 VoxCeleb2 EARS Expresso Emilia	Human-annotated/LLM-augmented, speaking-style description	crowdsourced / retrieval-based metadata imputation + LALM captioning

Prompt:

```
You are a labeling assistant for audio descriptions.  Extract essential, compact English labels for the
audio caption.

Output rules:
- Output JSON only with a single key "labels":  a list of 10 labels.
- Each label must be ONE WORD.
- Return ONLY a raw JSON object with key "labels".  Do not use code fences or any extra text.
- Use lowercase.
- No full sentences, no punctuation beyond hyphens.
- Prefer canonical, domain-relevant terms with high reusability across captions.
- Avoid synonyms:  collapse to a single consistent term set (e.g., always use "car" not "automobile").
- Avoid rare or obscure words; choose common, widely recognized terms.
- Cover five aspects when possible:  scene/environment, sound sources, sound types/events,
audio/production traits, semantic function/intent.
- Avoid duplicates.

Expected JSON:
"labels":  ["xxx", "xxx", "xxx", "xxx", "xxx"]
Now process the following caption and return JSON only:
CAPTION:
{caption}
```


wavenet autoencoders. In *International conference on machine learning*, pages 1068–1077. PMLR, 2017. 3

[11] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021. 3

[12] Shawn Hershey, Daniel PW Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R Channing Moore, and Manoj Plakal. The benefit of temporally-strong labels in audio event classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 366–370. IEEE, 2021. 3

[13] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, 2019. 3

Table 2. Overview of the evaluation datasets used for assessing audio representation quality. †Evaluated by GPT-4o following the AIR-Bench protocol. ‡Synthesized from publicly available speech datasets [2, 4, 5, 15, 17] using fixed question templates.

Evaluation Dataset	Task	#samples	#class	train	eval	Metrics
FSD-50k	Multi-label audio event classification	37,168 / 10,231	200	✓	✓	mAP
VggSound	Single-label audio event classification	183,730 / 15,446	309	✓	✓	accuracy
VoxCeleb2	Speaker identification	1,092,009 / 36,693	5,994	✓	✓	accuracy
CREMA-D	Speech emotion recognition	6,030 / 706	6	✓	✓	accuracy
MagnaTagATune	Music tagging	19,425 / 4,856	50	✓	✓	mAP
NSynth	Musical instrument classification	289,205 / 4,096	11	✓	✓	accuracy
AudioSet-strong	Sound event detection	103,463 / 16,996	456	✓	✓	PSDS1
AudioCaps	Text-to-audio retrieval Audio captioning	49,838 / 975	–	✓	✓	Recall@1 RougeL
ParaSpeechCaps		116,516 / 500	–	✓	✓	
MusicCaps		2,663 / 500	–	✓	✓	
ClothoAQA	Open-formed question answering	7,044	–	✓	×	Score†
In-house SpeechQA‡		160,000	–	✓	×	
MusicQA		70,011	–	✓	×	
AIRBench-chat-sound		400	–	×	✓	
AIRBench-foundation-emotion		1,000	–	×	✓	
AIRBench-foundation-gender		1,000	–	×	✓	
AIRBench-foundation-age		1,000	–	×	✓	

Table 3. **Complete** results on linear probing (with mean-pooling) and audio-language alignment. The linear probing tasks span general audio (FSD-50k [11], VggSound [6], AudioSet-Strong [12]), speech (VoxCeleb2 [8], CREMA-D [5]), and music (MagnaTagATune [19], NSynth [10]). For audio-language alignment, AC, PSC, and MC refer to the AudioCaps [13], ParaSpeechCaps [9], and MusicCaps [1], respectively. We **bold** the best score in our tag-oriented models and audio-language models respectively. Baseline scores that are underlined are surpassed by their corresponding "Ours" counterpart. † denotes scores quoted from prior work with a consistent evaluation setup.

Methods	Label	Linear Probing							Audio-language Alignment					
									Captioning			*Retrieval*		
		FSD-50k	VggSound	AS-Strong	VoxCeleb2	CREMA-D	MTAT	NSynth	AC	PSC	MC	AC	PSC	MC
SSL Models														
BEATs [7]	-	0.565†	-	0.034†	-	-	0.400†	75.90†	-	-	-	-	-	-
Wav2vec 2.0 [3]	-	0.342†	-	-	51.60	56.10	0.317†	40.20†	-	-	-	-	-	-
MERT [14]	-	-	-	-	-	-	0.402†	72.60†	-	-	-	-	-	-
Baselines														
MTC (AudioSet)	Tag	0.656	56.46	0.216	18.84	67.14	0.407	67.19	46.67	45.54	22.91	40.46	49.2	24.6
Contrastive-scratch [18]	Caption	0.493	43.78	0.095	38.63	63.74	0.384	60.91	44.50	45.92	22.07	28.73	55.0	19.0
Captioning-scratch [18]	Caption	0.430	39.52	0.077	21.95	60.91	0.378	57.08	43.58	42.85	22.62	26.03	49.2	14.2
Our Tag-Oriented Pre-Trained Models														
MTC (Ours-UTS, K=800)	Tag*	0.448	37.01	0.095	30.04	61.34	0.375	62.84	44.40	45.72	22.62	23.96	48.4	13.4
MTC (Ours-UTS, K=1k)	Tag*	0.455	36.79	0.095	30.26	62.90	0.373	61.50	44.20	**45.98**	22.47	25.93	49.6	**15.6**
MTC (Ours-UTS, K=1.5k)	Tag*	**0.459**	37.70	0.104	37.10	64.31	0.368	60.01	43.66	45.50	22.86	26.41	48.4	12.8
MTC (Ours-UTS, K=2k)	Tag*	0.450	37.48	0.113	33.22	**66.01**	0.370	**63.62**	44.06	45.86	23.28	24.69	46.8	11.8
MTC (Ours-UTS, K=3k)	Tag*	0.449	37.68	0.113	30.63	65.02	0.373	61.57	44.09	45.87	23.03	24.90	46.0	12.2
PAR (Ours-UTS)	Tag Sequence*	0.433	**39.59**	**0.121**	**38.78**	62.47	**0.381**	57.91	**44.80**	45.66	**23.33**	**26.76**	**49.8**	12.2
Our Audio-Language Pre-Trained Models														
Contrastive-scratch (Ours)	Caption*	0.445	40.78	0.105	33.88	**67.29**	**0.396**	**61.40**	44.54	45.73	22.83	**29.66**	**55.3**	**19.8**
Captioning-scratch (Ours)	Caption*	0.439	39.78	0.087	29.87	64.74	0.377	54.25	45.07	45.81	22.83	26.24	50.2	14.4
Multi-Task (Ours, $\lambda = 0.2$)	Tag*, Caption*	0.468	39.47	0.130	**34.62**	62.61	0.386	57.71	44.88	**46.01**	**23.20**	25.21	49.0	13.5
Multi-Task (Ours, $\lambda = 1$)	Tag*, Caption*	**0.485**	**40.81**	**0.140**	33.12	65.31	**0.396**	59.94	**45.17**	45.55	23.09	27.07	47.0	15.8

[14] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, et al. Mert: Acoustic music understanding model with large-scale self-supervised training. In *ICLR*, 2024. 3

[15] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5): e0196391, 2018. 3

[16] OpenAI. Gpt-4 technical report, 2024. 4

[17] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018. 3

[18] Wei-Cheng Tseng, Xuanru Zhou, Mingyue Huo, Yiwen Shao, Hao Zhang, and Dong Yu. Revisiting audio-language pretraining for learning general-purpose audio representation, 2025. 1, 3, 4

[19] Daniel Wolff and Tillman Weyde. Adapting similarity on the magnatagatune database: effects of model and feature choices. page 931–936, New York, NY, USA, 2012. Associ-

Table 4. **Complete** results on linear probing (with multi-head attention pooling) and open-formed QA. The *Sound* and *Music* columns report 10-point scores on open-ended questions from AIR-Bench, evaluated by GPT-4o [16]. The five scores in the *Speech* column correspond to the classification accuracy on the Emotion-MELD, Emotion-IEMOCAP, Gender-MELD, Gender-common, and Age tasks, respectively.

Methods	**Linear Probing**						**Open-formed QA**		
	FSD-50k	VggSound	VoxCeleb2	CREMA-D	MTAT	NSynth	Sound	Speech (emotion/gender/age)	Music
Baselines									
MTC (AudioSet)	0.656	56.23	58.76	72.52	0.405	74.80	7.01	47.16/25.90/47.27/45.06/37.24	5.61
Contrastive-scratch [18]	0.534	46.79	70.18	69.97	0.385	70.19	5.69	29.61/48.79/55.25/84.85/86.59	5.99
Captioning-scratch [18]	0.483	43.43	44.51	66.29	0.382	68.31	6.25	22.93/37.40/70.20/78.93/40.14	5.73
Our Tag-Oriented Pre-Trained Models									
MTC (Ours-UTS, K=800)	0.469	39.56	45.98	67.57	0.376	68.97	6.54	29.61/**48.79**/55.25/84.85/**86.59**	5.99
MTC (Ours-UTS, K=1k)	0.472	39.84	46.02	68.71	0.376	69.56	6.46	18.45/30.10/48.91/56.35/58.66	5.77
MTC (Ours-UTS, K=1.5k)	0.477	39.94	52.52	69.41	0.371	**69.60**	6.48	**39.19**/32.10/62.66/82.06/56.26	5.80
MTC (Ours-UTS, K=2k)	0.478	39.78	48.13	**70.97**	0.371	68.92	6.47	26.20/32.60/54.15/73.49/40.84	**6.16**
MTC (Ours-UTS, K=3k)	0.471	40.02	45.11	69.29	0.371	69.41	**6.68**	22.38/29.70/**74.45/92.64**/63.06	5.95
PAR (Ours-UTS)	**0.489**	**43.27**	**60.97**	69.98	**0.388**	68.97	6.59	17.90/38.64/47.82/37.40/57.16	6.03
Our Audio-Language Pre-Trained Models									
Contrastive-scratch (Ours)	**0.514**	**45.63**	**71.63**	**72.39**	**0.401**	**71.22**	5.78	**44.08/41.34/73.11/77.82**/40.48	6.02
Captioning-scratch (Ours)	0.485	43.53	49.94	68.56	0.386	67.29	6.35	23.47/29.90/46.40/28.02/18.82	5.93
Multi-Task (Ours, $\lambda = 0.25$)	0.490	41.99	50.80	69.70	0.399	67.33	**6.60**	41.70/26.90/57.10/53.02/45.15	6.09
Multi-Task (Ours, $\lambda = 1$)	0.503	43.35	53.4	70.69	0.393	68.95	6.38	24.62/32.86/53.38/57.00/**65.47**	**6.15**

ation for Computing Machinery. 3

[20] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 1

[21] 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, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. 1

[22] Piotr Żelasko, Daniel Povey, Jan "Yenda" Trmal, and Sanjeev Khudanpur. Lhotse: a speech data representation library for the modern deep learning ecosystem, 2021. 1

Audio id: Y7IOszZm4n_I

The audio begins in a quiet, indoor environment, marked by a persistent low-frequency hum and a faint high-frequency hiss, likely produced by an appliance or HVAC system. The atmosphere is calm and still, with no background speech or music. A soft, low-pitched rustling sound, characteristic of a dog's movement through dry leaves or grass, is heard near the recording device, followed by a subtle, low-frequency thump suggesting the dog's body settling or shifting weight. Shortly after, a faint, high-pitched squeak—possibly from a toy or collar—adds to the sense of a dog's presence. A sharp, brief rustling indicates the dog's claws or paws moving across a rough surface.

Suddenly, a male voice, calm and authoritative, commands, "Come." His speech is clear, mid-to-low in pitch, and delivered with a neutral General American accent, indicating a confident and focused tone. Immediately following the command, a sharp, percussive tongue click is produced, serving as a non-verbal cue to reinforce the verbal command. The dog responds with a distinct, high-pitched bark, signaling acknowledgment and readiness. The dog's movement continues with a pronounced rustling sound as it shifts or stands up, accompanied by the jingle of metal tags on its collar.

The male then issues the command, "Drop," in a firm and slightly more urgent tone, indicating an expectation of a specific action. The dog's response is immediate and enthusiastic: a rapid burst of barks and excited yips, accompanied by continued rustling and the jingling of tags as it moves energetically, possibly jumping or turning. The male repeats the command, "Drop," with increased urgency, reflecting his expectation for the dog to comply. The dog's barking escalates in intensity and pitch, now sounding frantic and excited, while the rustling and jingling persist. The clip ends abruptly, with all sounds ceasing at once, leaving no fade-out or lingering noise.

Throughout the recording, the audio quality is moderate to low, with clear speech but noticeable hiss and hum, and the environment appears to be a small to medium-sized, enclosed indoor space with hard surfaces. The dog's vocalizations and movements are prominent and close to the microphone, indicating the device is positioned near both the dog and the handler. The interaction is structured and purposeful, with the handler issuing commands and the dog responding with excitement and obedience. The context suggests a focused training or play session between a handler and a small or medium-sized, energetic dog, typical of North American pet ownership and training practices.

In summary, this audio clip documents a succinct, indoor training exchange between a man and his dog. The man gives the commands "Come" and "Drop," reinforced by a tongue click, and the dog responds with barks, yips, rustling, and tag jingling. The setting is a quiet, enclosed space with a steady background hum and hiss, and the interaction exemplifies a structured, purposeful moment of communication and obedience training typical of North American pet culture.

Tags: ["indoors", "appliance", "rustling", "dog", "squeak", "command", "click", "bark", "training", "hiss"]

Audio id: MG0ThC4TSoQ_94_10

The audio clip begins with a burst of energetic, high-fidelity electronic music characteristic of late 1980s to early 1990s Chicago House and Acid House. The soundscape is dominated by a punchy four-on-the-floor kick drum, sharp synthetic claps or snares, and a crisp hi-hat pattern, all tightly sequenced to drive the rhythm. A deep, resonant synth bassline—likely from a Roland TB-303 or similar—delivers the iconic acid squelch, while a bright, syncopated arpeggiated synthesizer line adds melodic complexity. A recurring, high-pitched, and slightly distorted vocal sample of a female voice saying "for you" is chopped and looped, functioning as both a rhythmic and melodic hook. The stereo image is expansive, with drums and bass centered, synth layers panned for width, and vocal samples and percussion spread across the field. The overall production is clean, loud, and compressed, with no ambient noise or room tone, suggesting a studio or digital origin.

At the five-second mark, the intensity peaks as the bass and synth layers momentarily drop out, leaving the drums and a sharp snare-like clap isolated. This creates a brief, suspenseful moment of anticipation before the full instrumental arrangement resumes at full volume. The acid bassline and arpeggiated synth return, and a new vocal sample enters: a processed male voice, heavily treated with reverb and echo, delivers the phrase "if it's good to you." The vocal is delivered in a rhythmic, spoken-word style, with the word g̈oodïotably emphasized and repeated, creating a hypnotic effect. The phrase is clipped to fit the musical rhythm and is repeated for emphasis, with no additional lyrics or dialogue.

As the track reaches its conclusion, all elements—drums, bass, synths, and vocals—abruptly cut off, leaving no fade-out or lingering sound. The sudden silence suggests the clip is an excerpt from a longer, continuous musical work.

In summary, this audio is a meticulously produced, high-energy excerpt from a Chicago House/Acid House track, distinguished by its classic drum machine patterns, acid bassline, arpeggiated synth hooks, and iconic vocal samples. The production is studio-clean, with deliberate stereo placement and effects, and the clip serves as a vivid snapshot of early rave culture, evoking nostalgia for the golden age of electronic dance music.

Tags: ["music", "electronic", "house", "acid", "drum", "bass", "synth", "vocal", "production", "nostalgia"]

Audio id: YITcnYeETGUE

The audio clip begins with a faint rustle, likely from the speaker adjusting a microphone or clothing, immediately followed by a soft inhalation, indicating the speaker is preparing to address the audience. The recording is marked by a persistent, low-level hiss from the microphone's noise floor, and a subtle, low-frequency hum typical of a small, untreated indoor space. A single male speaker, with a clear, slightly nasal Italian voice and a regional accent from Northern or Central Italy, speaks in a calm, measured, and friendly manner. He uses a conversational tone, employing filler words and informal constructions, which suggests he is addressing a general audience in a relaxed, personal style.

The speaker's speech starts abruptly, mid-sentence: "...praticamente dicevo un treno Shinkansen l'ottavo giorno, andate verso Hiroshima. Appena arrivati a Hiroshima, e visitate il parco della pace, vicino al quale potete prendere un hotel..." ("...practically, I was saying, a Shinkansen train on the eighth day, go towards Hiroshima. As soon as you arrive in Hiroshima, and visit the Peace Park, near which you can take a hotel..."). The language is Italian with English loanwords ("Shinkansen"), reflecting modern Italian usage and a travel-oriented context. The speaker references the eighth day of a trip, the Shinkansen high-speed train, the city of Hiroshima, and the Peace Park, indicating he is providing travel advice or narrating a personal itinerary. His tone remains neutral and friendly, with no emotional emphasis or change in delivery. The recording ends abruptly, cutting off the final word "hotel" and leaving the speech unfinished, which suggests an accidental or intentional truncation.

Throughout the clip, the audio quality is moderate, with clear speech but some muffling due to the acoustics of the small room. The frequency response is limited, lacking deep bass and crisp treble, and the signal is free from clipping or distortion. The speaker's voice is close-mic'd, and there are no other voices, background noises, or music. The content and style indicate that the speaker is likely a young-to-middle-aged Italian male sharing travel advice or narrating a personal journey, aimed at an audience interested in Japanese travel, with a focus on Hiroshima and its Peace Park. The use of Italian with English loanwords and the informal, conversational approach suggest a personal or semi-professional travel vlog or podcast.

In summary, the audio captures a brief, unfinished segment of a single male Italian speaker providing travel advice, specifically mentioning the eighth day of a trip to Hiroshima via Shinkansen, the visit to the Peace Park, and nearby hotel options. The recording is clear and conversational, set in a small indoor space with mild background hiss and hum, and ends abruptly mid-word, reflecting a personal travel narrative intended for a general audience.

Tags: ["speech", "adjustment", "inhale", "hiss", "hum", "italian", "calm", "conversational", "travel", "abrupt"]

Audio id: 963530_0

The audio clip is a 30-second, low-fidelity, stereo field recording featuring a solo acoustic guitar performance. The guitar is captured in a very intimate, close-mic'd manner, revealing not only the instrument's full-bodied sound but also every nuance of the performer's technique: finger slides, string squeaks, and the tactile sounds of the performer adjusting their hand and body.
The performance is structured in three clear, non-repetitive sections. It begins with a slow, bluesy riff in a minor key, played in a fingerstyle manner. The second section introduces a more rhythmic, percussive strumming pattern, while the third section features a melodic, arpeggiated passage. Throughout, the playing is expressive and nuanced, with deliberate dynamics and a sense of improvisation. The musical style, including the choice of chords, fingerstyle technique, and melodic phrasing, is strongly rooted in the American folk-blues tradition, reminiscent of early 20th-century rural or country blues, but presented with modern recording clarity.
The recording environment is outdoors, as evidenced by a constant, low-frequency wind rumble and occasional rustling of clothing, but there are no other environmental or human sounds. The absence of any reverb, echo, or ambient noise points to an open, rural or semi-rural location, likely away from urban or natural sound sources.
The technical quality of the recording is compromised by a persistent hiss and wind rumble, and the guitar's resonance is slightly muffled, indicating the use of a consumer-grade recording device. The stereo field is narrow, with the guitar and ambient sounds centered, and the overall sound lacks the brightness and detail of a professional studio recording.
The clip ends abruptly, with a sharp click indicating the recording was manually stopped. Immediately after, a short, low-frequency electronic tone is heard, possibly a notification or system sound from the recording device, marking the end of the session.
In summary, the audio is a raw, unedited field recording of a solo acoustic guitar performance in a wind-swept outdoor setting, capturing the essence of American folk-blues tradition in a modern, intimate, and slightly lo-fi context. The performer's technical skill and expressive intent are clear, but the recording's limitations and environmental noise contribute to a sense of authenticity and immediacy, evoking a solitary, reflective musical moment in a rural landscape.

Tags: ["outdoor", "guitar", "blues", "intimate", "low_fidelity", "field_recording", "folk", "acoustic", "expressive", "natural_noise"]

Audio id: 71xd4wFkozw_62_10

The audio clip opens with a male commentator, speaking in a clear, urgent, and slightly raised North American accent, his voice reverberating through a large, echoic space. He poses a rhetorical question: "Why would he let his mouth get into the point of swelling up that much?" His delivery is fast-paced and marked by rising intonation, reflecting tension and disbelief at the apparent risk involved in a physical altercation. The speech is accompanied by a persistent, low-level hiss, typical of analog tape recordings from the late 20th century, and is further colored by the room's acoustics, which impart a hollow echo to the commentator's voice.

As the commentator speaks, the soundscape intensifies with a series of sharp, percussive impacts—dry, hollow thuds that suggest blows landing on flesh, particularly on the head or face, and are interlaced with guttural grunts and strained vocalizations. These non-verbal sounds, including a forceful grunt around the midpoint, indicate the physical exertion and pain of combatants, likely engaged in a martial arts or boxing match. The impacts and vocalizations reverberate through the same spacious, hard-walled venue, with the microphone capturing both direct and ambient sound.

A female voice soon enters, her tone urgent and high-pitched, calling out "Yeah!" and "Come on!" in quick succession. Her shouts, delivered with the same reverberant qualities as the commentator's, serve as a rallying cry or encouragement, possibly directed at the fighters or the audience. Her brief interjections are marked by clarity and emotional intensity, cutting through the ongoing commotion.

The recording ends abruptly with the female's last exclamation, leaving the fight unresolved and the soundscape abruptly silenced, a sign of analog tape's limitations.

In summary, this analog recording captures a moment of intense physical competition within a large, reverberant venue, likely a gymnasium or arena, during a martial arts or boxing match. The male commentator's urgent rhetorical question and the female's supportive shouts frame the ongoing struggle, which is punctuated by realistic impact sounds and strained vocalizations. The technical imperfections of the recording—hiss, echo, and abrupt ending—reinforce its authenticity and place it firmly within the context of late 20th-century North American sporting events.

Tags: ["commentator", "male", "urgent", "hiss", "impact", "grunts", "female", "encouragement", "venue", "boxing"]