ADEPT

Interactive Visual Analytics for **A**udio **D**ataset **E**xploration and **P**repara**T**ion

Chen Chen, Tica Lin, Josh Kimball
Dolby Laboratories

Presented by **Liwenhan Xie (Shelly)**

Designing Interactive Systems (DIS '26) · June 13–17, 2026 · Singapore

Preparing data is where ML work actually starts

MOTIVATION

- Before training, practitioners must **understand distributions, verify labels, and select the right samples** — or risk training on poorly understood data
- Visual analytics already transformed this step for:
  - **Structured data** — charts & dashboards (Tableau, Power BI)
  - **Images** — thumbnail galleries, similarity browsing, annotation tools
- Their secret: these data types have natural visual representations

Audio (and video) remain underserved — despite their growing role in modern ML

Why is audio hard to explore?

MOTIVATION

1. **Time-based consumption** — you cannot "glance" at 10,000 clips; listening is inherently sequential and slow
2. **Multi-dimensional semantics** — spectral, temporal, and semantic content unfold together over time
3. **Hidden quality issues** — noise, clipping, wrong sample rates buried inside files
4. **Unreliable labels** — documented annotations often disagree with actual content

And: practitioner workflows for audio preparation are poorly understood — prior work focuses on domain-specific applications, not general-purpose preparation.

Formative study: how do practitioners cope today?

FORMATIVE STUDY

10 audio ML practitioners (E1–E10)
research scientists & engineers in industry; speech, music IR, multimodal

30–45 min semi-structured interviews
day-to-day tasks, tools, pain points

Current toolkit = fragmented

- librosa / torchaudio / scipy scripts
- Audacity, Adobe Audition for manual listening
- Jupyter notebooks as glue

"Switching between multiple tools and custom scripts for tasks that should be integrated."

Six design challenges

FORMATIVE STUDY

- **DC1** No visibility into **feature distributions**
- **DC2** Subsetting is **one-off scripts**, ad hoc
- **DC3** No **global overview of quality** — "listen to 10 random clips and hope"
- **DC4** **Label validity** is uncertain
- **DC5** **Fragmented data-to-code** pipelines, rewritten per project
- **DC6** **Version confusion** — "everyone has their own version of 'cleaned' "

"I've used datasets labeled as speech-only, but when I listened, I heard background noise, silence, or mouse clicks. You trust the labels, but they're not always reliable" — **E3**

Five design goals

FORMATIVE STUDY → DESIGN

DG1 **Automated feature distribution visualization** (DC1)

DG2 **Flexible filtering & subset creation** (DC2)

DG3 **Comprehensive quality assessment** at scale (DC3)

DG4 **Integrated label validity verification** (DC4)

DG5 **Unified pipelines with provenance tracking** (DC5, DC6)

ADEPT: three integrated panels

SYSTEM

Panel 1

Quality Diagnosis with Audio Features

quality metrics + 13 audio features, three coordinated views

DG1, DG2, DG3

Panel 2

Audio Semantics Verification

audio language model confidence + spectrogram for label validation

DG4

Panel 3

Tracking Data Provenance

every operation recorded; shareable specs + PyTorch loader

DG5

Web-based: React/TypeScript + D3.js frontend, Python/FastAPI backend; heavy computation precomputed for sub-second interaction

Under the hood: modeling quality & label validity

SYSTEM · MODELING

Audio quality

SDR / SNR need a clean reference — rarely available.

- **NISQA-v2**: learning-based, **no-reference**, task-agnostic
- 5 dimensions: Overall (MOS), Noisiness, Discontinuity, Coloration, Loudness

Label validity

ALM (Qwen2-Audio) — one model for any label type. Ask $N{=}10\times$ and score the agreement:

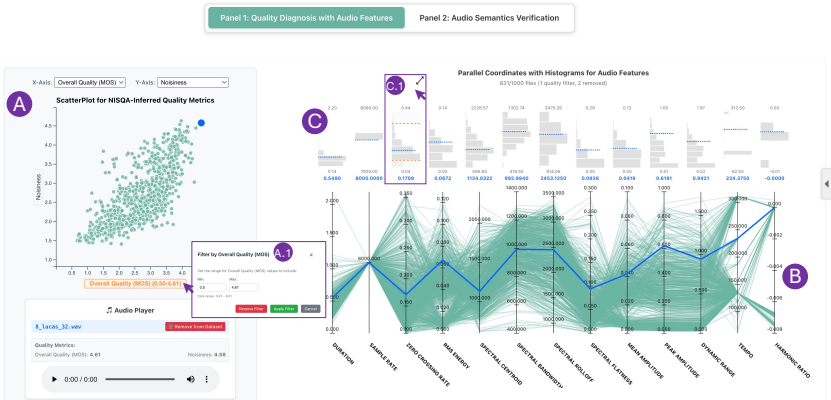
$$\frac{\gamma\, o_y}{\gamma\, o_y + (1-\gamma)(N-o_y)} \qquad (\gamma{=}0.75)$$

Traffic-light: **high** / **med** / **low**

Both are **reference-free** and **task-agnostic**, so ADEPT generalizes across audio types (DC1, DC5). The scores guide human attention — they don't replace judgment.

Panel 1: Quality Diagnosis with Audio Features

SYSTEM · INTERFACE



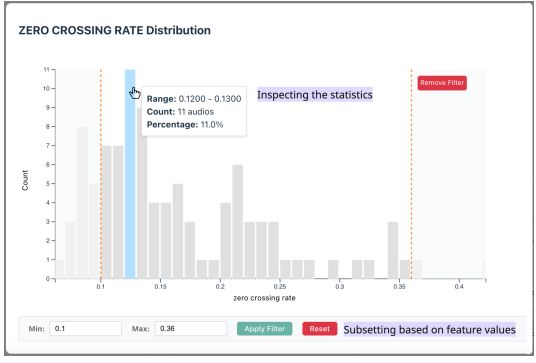
A quality scatterplot **B** 13-feature parallel coordinates **C** per-feature histograms

Three ways to subset, all in Panel 1

SYSTEM · INTERFACE

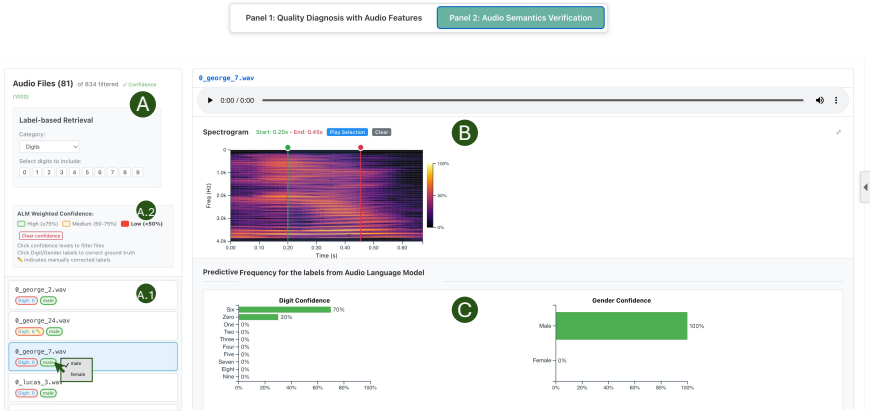
1. **Individual removal** — listen, judge, remove a single clip
2. **Quality-based range filter** — e.g. keep MOS $\in [3.5, 5]$
3. **Feature-based range filter** — e.g. keep zero crossing rate $\in [0.1, 0.36]$

Every histogram expands into a detail view with statistics and range sliders (DG1, DG2).



Panel 2: Audio Semantics Verification

SYSTEM · INTERFACE

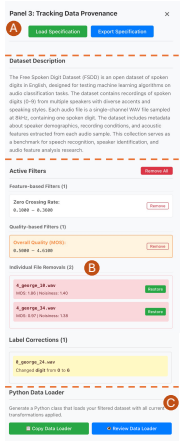


A audio list, ALM-colored **B** spectrogram + replay **C** ALM prediction frequencies

Panel 3: Tracking the Data Provenance

SYSTEM · INTERFACE

- **A Spec management** — export / import the whole session as a `.json` spec; not tied to a dataset, so teammates can replay it elsewhere
- **B Operation record** — every filter, removal, and label correction listed; remove any, or restore the initial checkpoint
- **C PyTorch data loader** — generated code that applies all recorded operations, ready for training



Reuse + sharing = DG5: reproducibility across people and projects.

Evaluation: 15 practitioners, one open-ended task

EVALUATION

Participants (P1–P15)

11 senior/staff researchers, 3 PhD students, 1 data operator · 9 with 5+ years of audio experience

Dataset

1,000 clips from the Free Spoken Digit Dataset — with 5 wrong labels secretly injected

Task (30 min, remote)

"Explore the dataset and curate a qualified subset characterized by high perceptual quality and accurate labeling."

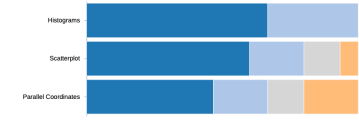
Then: questionnaire (component effectiveness, NASA-TLX-style experience, panel ranking) + semi-structured interview

How practitioners actually subset

EVALUATION · RESULTS

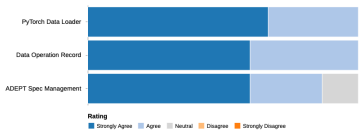
- **Feature-based filters: 73%** (11/15) — mostly used to cut outliers (e.g. the one clip longer than 2 s)
- **Individual removal: 60%** (9/15) — human ears still make the final call
- **Quality range filters: 53%** (8/15) — 7 of 8 thresholded on overall MOS
- Quality-based subsetting overall: **87%** (13/15)

Component ratings: histograms 100% positive, scatterplot 80%, parallel coordinates 67%.



(a) Effectiveness Rating of Panel 1 components.

(b) Effectiveness Rating of Panel 2 components.



(c) Effectiveness Rating of Panel 3 components.

Did they catch the wrong labels?

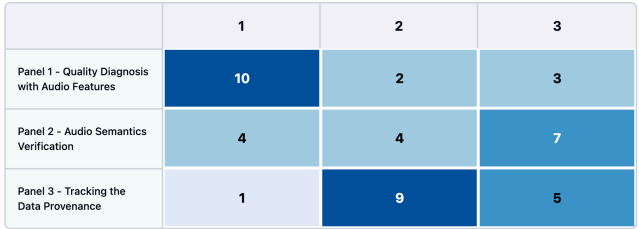
EVALUATION · RESULTS

All 15 participants found and corrected all 5 injected label errors

- ALM confidence kept the "suspicious" set **reviewable** — only ∼100 of 1,000 clips flagged **low** ($\gamma = 0.75$)
- **Label validity view**: 93% positive ratings
- **Spectrogram view**: 67% — the gap was a request for *more* types (mel, log-frequency), not a rejection

Which panel mattered most?

EVALUATION · RESULTS



Panel 1 first (10/15) · Panel 3 second (9/15) · Panel 2 last (7/15)

- Ranking tracks **frequency & breadth of use** — Panel 1 is the landing page for open-ended exploration; Panel 3 is prized for cross-session collaboration
- Panel 2's label check is a **bounded, targeted task** — done once errors are found; targeted tools may be undervalued, yet it still hit its goal (15/15)

Overall experience: low effort, high confidence

EVALUATION · RESULTS

73%

rated their **confidence** high

73%

reported **low / very low frustration**

80%

low mental & temporal demand

- NASA-TLX-style ratings across five dimensions of demand, confidence, and frustration
- Participants found ADEPT "intuitive, convenient, and easy to understand" — one likened it to a "generally intelligent 'genie' to help with looking into the data"

Four design implications beyond audio

DISCUSSION

1. **AI as attention guidance, not decision replacement** — confidence visualizations direct humans to where verification matters most
2. **Provenance is a first-class requirement** — record everything, make it shareable; not an optional add-on
3. **Progressive disclosure across granularities** — distributions → filtered subsets → individual samples, fluidly
4. **Task-agnostic foundations, domain-specific extensions** — reference-free metrics + ALMs generalize; architect for plug-in customization

Takeaways

CONCLUSION

- Audio dataset preparation is fragmented and ad hoc — we characterized **6 challenges** and **5 design goals** from 10 practitioners
- **ADEPT** integrates quality diagnosis, ALM-assisted label verification, and provenance tracking in one visual analytics workflow
- 15-practitioner evaluation: confident subsetting, **15/15 caught every injected label error**, low workload
- Transferable design implications for visual analytics on **other unstructured modalities**

Thank you!

Chen Chen · Dolby Laboratories · cchen24@terpmail.umd.edu