Praat++: Multimedia Annotation System for Speech and Vocalization

Weiran Zhang and **Kenny Q. Zhu**
University of Texas at Arlington, USA
wxz9630@mavs.uta.edu
kenny.zhu@uta.edu

Abstract

High-quality time-aligned annotation is fundamental to speech processing and animal vocalization research, yet precise boundary localization and consistent labeling remain challenging in collaborative settings. We present Praat++, a web-based multimedia annotation system designed for collaborative, video-informed, and AI-assisted timeline labeling of audio and video data. The system tightly synchronizes waveform, spectrogram, pitch, intensity, and time-aligned video playback with fine-grained region-based editing, enabling precise boundary refinement and improved label accuracy within a unified interface. Praat++ further incorporates role-aware workflow management and human-in-the-loop AI-assisted pre-annotation to enhance inter-annotator consistency and reduce labeling time. Through real-world multimodal speech and animal vocalization annotation scenarios, we demonstrate that Praat++ provides an integrated infrastructure for improving annotation quality and efficiency in dataset construction workflows. The demo video[1], website[2], and source code[3] are now publicly available.

1 Introduction

Recent advances in artificial intelligence for speech and bioacoustic analysis, including speech emotion recognition, speaker diarization, speech activity detection (SAD), and animal vocalization modeling, heavily rely on large-scale, high-quality annotated datasets. Public benchmarks such as AudioSet (Gemmeke et al., 2017), AVA (Chaudhuri et al., 2018), and RAVDESS (Livingstone and Russo, 2018), as well as dog emotion datasets such as EmotionalCanines (Dang et al., 2025), demonstrate the importance of time-aligned labeling of acoustic events and multimodal signals. Despite rapid progress in model architectures, the construction of reliable datasets remains fundamentally dependent on precise manual annotation.

Among these tasks, Sound Event Detection (SED) (Mesaros et al., 2021) represents a representative timeline-based problem that requires accurate event labeling and precise temporal boundary localization. The quality of both labels and onset–offset boundaries directly influences downstream model performance. However, timeline annotation is labor-intensive and often ambiguous, especially in acoustically complex or multimodal environments.

Accurate annotation typically requires synchronized inspection of waveform, spectrogram, pitch, intensity, and contextual visual information from video recordings. In practice, annotation quality is further influenced by inter-annotator consistency, expert supervision, and mechanisms for resolving uncertain annotated regions. Therefore, improving both boundary precision and labeling consistency while reducing annotation time is a critical challenge in dataset construction.

Existing annotation tools provide valuable functionality but reveal structural limitations for large-scale, collaborative timeline labeling with tightly synchronized audio–video and acoustic analysis. Desktop-based systems such as Praat (Boersma, 2001) offer detailed signal inspection (waveform, spectrogram, pitch, intensity) but primarily support single-user workflows and lack integrated video context and collaborative task management. Multimodal platforms such as ELAN (Wittenburg et al., 2006), Anvil (Kipp, 2001), and BORIS (Friard and Gamba, 2016) support video-aligned multi-layer annotation but provide limited fine-grained acoustic visualization for phonetic-level boundary inspection. For web-based annotation systems, Label Studio (Tkachenko et al., 2020-2025) is a general-purpose platform that emphasizes scalable task management but offers limited support for flexi-

[1] https://www.youtube.com/watch?v=YboCoBRF5lg
[2] https://redgiant.uta.edu/praat
[3] https://github.com/UTA-ACL2/PraatPlusPlus

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 812–818

ble tiered annotation and detailed acoustic feedback, and its AI assistance is typically provided via loosely coupled external components rather than tightly integrated functionality; in contrast, Whombat (Balvanera et al., 2023) provides a comprehensive collaborative workflow and robust quality control for spectrogram-centered annotation, but does not support synchronized video-informed annotation or integrated AI assistance.

As datasets continue to scale and AI assistance becomes increasingly important, the absence of unified support for synchronized audio–video inspection, detailed acoustic analysis, collaborative workflows, and integrated model assistance remains a critical gap. In practice, annotation quality is determined not only by visualization fidelity but also by structured collaboration. In group annotation settings, novice annotators may produce uncertain or inconsistent labels, particularly when acoustic cues are ambiguous. Without mechanisms for expert review, uncertainty tracking, and iterative correction, inconsistencies can accumulate and degrade dataset reliability. Furthermore, AI-assisted pre-annotation is frequently applied as a standalone preprocessing step, rather than an integrated human-in-the-loop process that supports refinement and boundary adjustment within the same interface.

To address these challenges, we present Praat++, a web-based multimedia annotation system designed to improve label accuracy, boundary precision, and collaborative consistency in timeline-based annotation tasks. The system integrates synchronized waveform, spectrogram, pitch, intensity, and time-aligned video playback to support fine-grained boundary refinement and improved label accuracy. By coupling region-based timeline editing with multimodal feedback, Praat++ facilitates precise onset–offset adjustment under complex acoustic conditions.

Beyond visualization, Praat++ introduces a structured collaborative workflow module that supports role-based annotation, expert-guided correction, and quality-controlled review. Uncertain annotated regions can be identified and refined through iterative feedback, enhancing inter-annotator consistency and improving learnability for novice annotators. This design enables expert supervision and collaborative knowledge transfer within real annotation projects.

In addition, Praat++ incorporates AI-assisted pre-annotation as an integrated enhancement module. Automatically generated silver annotations are directly editable within the timeline interface, allowing human annotators to refine both labels and boundaries while preserving full control over the final annotations. This human-in-the-loop design significantly reduces annotation time while maintaining precision and consistency.

The contributions of this demonstration are summarized as follows:

- We design a video-informed annotation interface that leverages visual cues alongside fine-grained acoustic analysis (e.g., waveform, spectrogram, pitch, and intensity) to substantially improve label accuracy and onset–offset boundary precision.
- We introduce a structured collaborative workflow with built-in quality control, supporting project management and role-aware expert review and correction, while flagging uncertain annotated regions for later review to improve efficiency and consistency in group settings.
- We provide integrated AI-assisted pre-annotation with confidence-guided correction for timeline-oriented tasks, enabling human-in-the-loop correction to significantly accelerate large-scale annotation workflows.

2 System Architecture

Praat++ is designed as a web-based, collaborative multimedia annotation system tailored for time-aligned audio–video analysis. Figure 1 summarizes the end-to-end implementation architecture. Within this architecture, Praat++ comprises three integrated modules: (i) a collaborative workflow module, (ii) a video-informed multimodal annotation module, and (iii) an AI-assisted pre-annotation module. This design separates task management, interactive annotation, and model-assisted automation, ensuring both scalability and usability in real-world annotation projects. The system is open source and supports self-hosted deployment on local or institutional servers. All data and AI models remain within the deployed environment.

2.1 Collaborative Workflow Module

The collaborative workflow module governs user roles, file-pool organization, and concurrency control, enabling large-scale team-based annotation (Fig. 2).

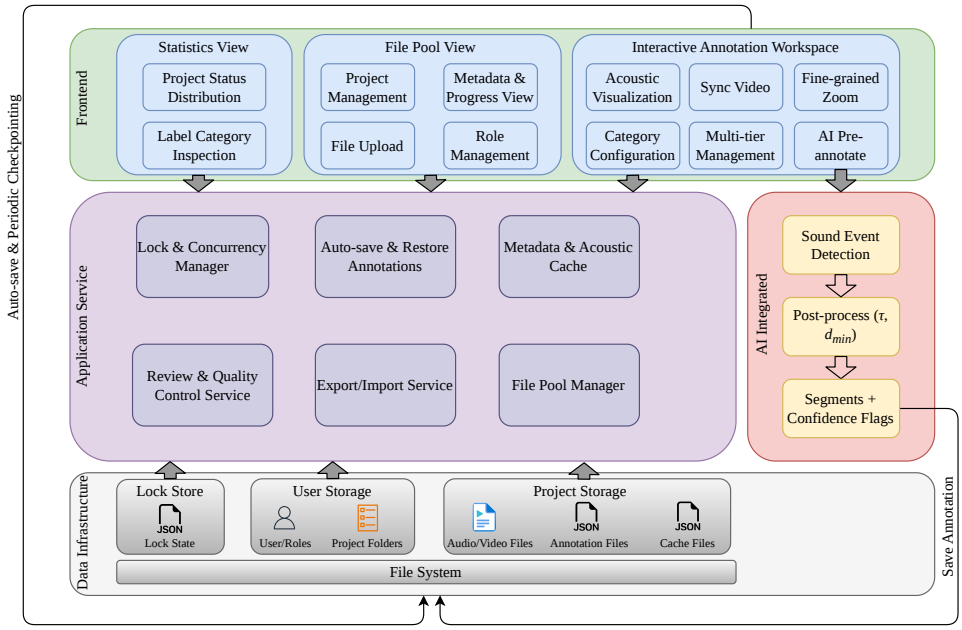

Figure 1: System implementation architecture of Praat++.

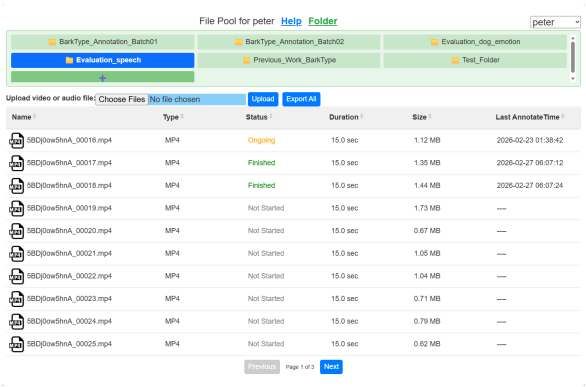

Figure 2: File pool view for Praat++.

Role-aware User Management. Praat++ supports a hierarchical user system consisting of annotators and superusers. Superusers can monitor annotation progress across users and temporarily switch into another user's workspace for inspection and quality control. This role separation enables an expert–novice workflow: experts can review and revise submitted annotations, add corrective feedback, and take over uncertain or low-confidence regions for adjudication, while novice annotators can learn from expert exemplars and previously validated labels. Such expert-in-the-loop oversight improves annotation quality and consistency, and helps increase inter-annotator agreement in collaborative projects.

Task-oriented File Pool Management. Upon login, users enter a personalized file pool interface. The system supports multi-file upload, automatic handling of duplicate filenames, and compatibility with mainstream audio and video formats. Files can be organized into user-defined folders to represent distinct annotation tasks or experimental groups. For each file, Praat++ displays structured metadata including file type, duration, size, annotation status (not started / in progress / finished), last saved timestamp, and tier-level completion status. Metadata are cached to improve responsiveness in large-scale datasets.

Concurrency Control and Data Lifecycle. To prevent conflicting edits, Praat++ enforces file-level locking. A file cannot be opened simultaneously by multiple active sessions of the same account. Annotation data are automatically versioned and can be exported in bulk. The system supports one-click export of all annotations into Praat-compatible `.TextGrid` format, ensuring interoperability with traditional acoustic analysis pipelines.

2.2 Video-informed Multimodal Annotation Module

The video-informed annotation module provides a unified workspace for fine-grained temporal labeling with synchronized acoustic visualization and time-aligned video playback (Fig. 3).

Unified Waveform-based Region Editing. Annotations are performed directly on the waveform via region-based interaction. Users can create, re-

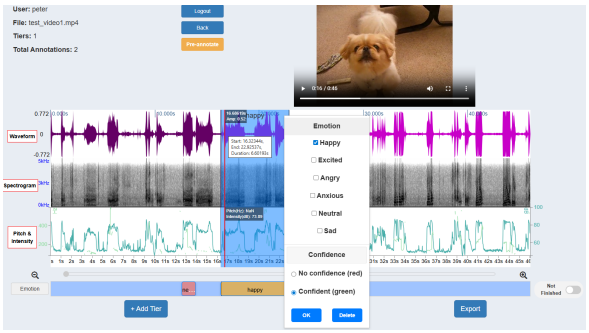

Figure 3: Annotation workspace for Praat++.

size, drag, or delete regions, and assign multi-label annotations through contextual menus. Each region supports confidence flag, allowing annotators to mark uncertain annotated regions for later refinement. Clicking a region triggers synchronized audio–video playback. Hover-based feedback displays timestamps, duration, waveform amplitude, pitch, and intensity values in real time.

Synchronized Acoustic Visualization. Praat++ dynamically renders waveform, spectrogram, pitch, and intensity tracks aligned along a shared temporal axis. Visualization ranges are automatically adjusted according to input file characteristics (e.g., waveform amplitude normalization, spectrogram frequency bounds). Users may manually refine spectrogram display ranges for detailed inspection. Zooming and scrolling are synchronized across all acoustic layers, enabling millisecond-level boundary inspection. During playback, all synchronized visualization layers, including waveform, spectrogram, pitch, and intensity tracks, automatically scroll to maintain temporal alignment with the current playback position.

Tier and Category Control. The system supports hierarchical tier management, where each tier represents a distinct annotation dimension (e.g., emotion, behavior, event type). Users can create, rename, hide, or delete tiers, and mark individual tiers as finished. Category definitions are fully customizable, allowing annotators to define, modify, or delete label sets to match specific research needs.

Automatic Persistence. Annotation states are automatically saved when users leave the annotation page or close the browser. Upon re-entry, annotation content is restored without manual intervention. The interface displays contextual information including current user, filename, tier count, and total annotation count, ensuring transparency during long-term projects.

2.3 AI-assisted Pre-annotation Module

To reduce manual workload and accelerate dataset construction, Praat++ integrates model-assisted pre-annotation using a pretrained PANNs sound event detection (SED) model (Kong et al., 2020). The pre-annotation component is implemented as a modular backend service, so the current PANNs model can be replaced by other SED models or task-specific models in future deployments. This module does not require any additional SED training from users. Instead, annotators select a target event class from the full set of categories supported by PANNs (i.e., the AudioSet label set (Gemmeke et al., 2017)), and the system generates corresponding time-aligned segments for review and refinement.

Users select a target event class supported by PANNs, set a confidence threshold (τ), a minimum duration ($d_{\min}$), and a target tier name. The model outputs a frame-level event probability $p_t \in [0, 1]$, defined as the predicted likelihood that the selected event occurs at each time frame (e.g., speech probability for speech activity). Praat++ converts these framewise probabilities into segment-level predictions by selecting frames with $p_t \geq \tau$ and merging temporally contiguous selected frames into candidate segments; segments shorter than $d_{\min}$ are discarded. For each retained segment, the system aggregates frame probabilities (e.g., mean or max) to obtain a segment score and assigns a confidence flag: segments with an aggregated score ≥ 0.7 are marked **confident**, otherwise **not-confident**. The resulting segments (with labels, boundaries, and confidence flag) are then written to the target tier as annotated regions by clearing and overwriting its existing annotations.

Pre-annotations are inserted as editable regions, allowing annotators to refine boundaries, adjust labels, or update confidence levels. For large-scale speech annotation tasks, annotators can primarily review and listen around the pre-annotated regions instead of scanning the entire recording from scratch, which substantially reduces manual effort and improves annotation efficiency. This design maintains human oversight while leveraging automatic detection to accelerate dataset construction.

2.4 Statistical Monitoring

Praat++ includes a lightweight statistics dashboard summarizing user-level and task-level progress

Figure 4: Statistics dashboard for Praat++. **Left:** User-level annotation progress and task-level status distribution. **Right:** User-defined categories.

(Fig. 4). The system reports total file counts, folder-based task completion rates, and distribution of user-defined categories. These metrics facilitate project supervision and workload balancing in collaborative environments.

3 Evaluation

We compared Praat++ with representative audio-only and multimodal annotation tools, including Praat (Boersma, 2001), ELAN (Wittenburg et al., 2006), Anvil (Kipp, 2001), BORIS (Friard and Gamba, 2016), Raven (Charif et al., 2010), Whombat (Balvanera et al., 2023), and Label Studio (Tkachenko et al., 2020-2025), to characterize practical capabilities for time-aligned annotation.

To reflect practical annotation requirements such as efficiency, boundary precision, and labeling consistency, we organize the comparison along three dimensions. **[MM] Multimedia Annotation** includes integrated audio–video synchronization, detailed acoustic visualization for boundary inspection, and flexible tier management and label schema. **[CL] Collaborative Workflow** includes project management, role-based collaboration, confidence-aware annotations, annotation state management, and quality assurance. **[AI] AI Assistance** includes AI-assisted pre-annotation.

The comparison results are summarized in Table 1. Overall, Praat++ provides comprehensive support across multimedia annotation, collaborative workflow, and AI-assisted pre-annotation. Compared with desktop tools such as Praat and ELAN, Praat++ extends fine-grained acoustic inspection with synchronized audio–video interaction and tier-oriented timeline editing, while additionally supporting project-centric workflows for collaborative dataset construction. While tools such as Anvil, BORIS, and Raven support multimodal or timeline-based annotation, they generally lack integrated web-based workflow management and quality assurance mechanisms designed for large-scale collaboration. Label Studio offers flexible multimodal annotation with extensible external AI integrations; however, it does not natively provide Praat-style acoustic visualization or multi-tier timeline management required for precise boundary refinement in speech and animal vocalization annotation. In contrast, Praat++ integrates fine-grained acoustic visualization with structured collaborative workflow (including annotation states, quality assurance, and confidence-aware flags) and tightly coupled AI-assisted pre-annotation to improve annotation accuracy, consistency, and efficiency.

4 Related Work

Time-aligned annotation tools are fundamental to building reliable datasets for speech processing and bioacoustic research. Existing systems typically emphasize either fine-grained acoustic inspection, video-aligned multi-tier coding, or general-purpose web labeling, leaving a practical gap for collaborative, video-informed, and AI-assisted timeline labeling workflows with scalable quality control.

Acoustic analysis and browser-based Praat variants Praat (Boersma, 2001) remains a standard tool for phonetic analysis and TextGrid-based tier annotation with detailed acoustic views. Praat on the Web (Domínguez et al., 2016) brings Praat-style visualization and script-based semi-automatic processing to the browser and extends feature-oriented views, but provides only limited in-browser annotation editing and does not target

Table 1: Feature comparison between Praat++ and representative annotation tools.

Grp.	Features	Praat	ELAN	Anvil	BORIS	Raven	Whombat	Label Studio	Praat++
[MM]	Integrated sync video	–	✓	✓	✓	–	–	✓	✓
[MM]	Detailed acoustic visualization	✓	–	–	–	✓	–	–	✓
[MM]	Tier management & label category	✓	✓	✓	–	–	–	–	✓
[CL]	Project management	–	–	–	–	–	✓	✓	✓
[CL]	Role-based collaboration	–	–	–	–	–	–	✓	✓
[CL]	Confidence-aware annotation	–	–	–	–	–	–	–	✓
[CL]	Annotation state management	–	–	–	–	–	✓	✓	✓
[CL]	Quality assurance	–	–	–	–	–	✓	–	✓
[AI]	AI-assisted pre-annotation	–	–	–	–	–	–	✓	✓

project-level collaboration such as task management, concurrency control, or role-aware review.

Video-aligned multimodal timeline annotation Video-aligned temporal annotation frameworks (e.g., ELAN (Wittenburg et al., 2006), Anvil (Kipp, 2001), BORIS (Friard and Gamba, 2016)) support multi-layer coding over synchronized media, but typically offer limited Praat-style acoustic inspection for SED-style boundary correction and are not designed around integrated, role-aware quality-control workflows within a unified web environment.

Collaborative web-based annotation platforms Collaborative web platforms have emerged to support scalable annotation workflows for bioacoustics and ecological monitoring. Whombat (Balvanera et al., 2023) provides a web-based interface for managing audio annotation at scale, including project-level organization and multi-user workflows. However, it does not target Praat-style acoustic inspection for fine-grained boundary refinement, nor does it tightly couple synchronized audio–video context, tier-oriented timeline schemas, and confidence-/state-aware quality control within a unified interface.

Datasets motivating speech and canine vocalization annotation Large labeled datasets have driven progress in audio and speech understanding, including AudioSet (Gemmeke et al., 2017) and densely labeled speech activity datasets such as AVA-Speech (Chaudhuri et al., 2018). For affective analysis, RAVDESS (Livingstone and Russo, 2018) provides audio–visual expressions. In the animal domain, EmotionalCanines (Dang et al., 2025) introduces arousal–valence annotations for dog vocalizations, while DogSpeak (Lekhak et al., 2025) scales to in-the-wild bark sequences with rich dog identity metadata, highlighting the need for robust annotation interfaces and consistent quality control under real-world variability.

AI-assisted annotation and human–AI collaboration Human-in-the-loop sound event labeling has been explored to reduce the cost of annotating recordings by using machine-driven region recommendations to guide users to promising candidate segments and supporting iterative boundary refinement (Kim and Pardo, 2018). Meanwhile, pretrained audio models such as PANNs (Kong et al., 2020) can provide frame-level event scores as pre-annotation cues; however, existing workflows are often fragmented across separate inference and annotation stages. Peanut (Zhang et al., 2023) demonstrates a human–AI collaborative annotation paradigm for audio–visual data by combining mixed-initiative interaction with active learning to reduce user effort, although it targets frame/object-level audio–visual grounding rather than timeline segmentation and boundary refinement.

5 Conclusion

We presented Praat++, a web-based system for time-aligned audio–video annotation in speech and animal vocalization research. This work highlights that high-quality boundary labeling at scale typically relies on: (i) tightly coupled acoustic–visual inspection to resolve ambiguous annotated regions, (ii) role-aware collaboration with built-in quality control to maintain consistency, and (iii) in-editor AI pre-annotation to focus human effort on the most critical regions.

Acknowledgment

This work was partially supported by NSF Award No. 2349713.

References

Santiago Martinez Balvanera, Oisin Mac Aodha, Matthew J Weldy, Holly Pringle, Ella Browning, and Kate E Jones. 2023. Whombat: An open-source annotation tool for machine learning development in bioacoustics. arXiv preprint arXiv:2308.12688.

Paul Boersma. 2001. Praat, a system for doing phonetics by computer. Glot. Int., 5(9):341–345.

RA Charif, Amanda M Waack, and Laura M Strickman. 2010. Raven pro 1.4 user's manual. Cornell lab of ornithology, Ithaca, NY.

Sourish Chaudhuri, Joseph Roth, Daniel PW Ellis, Andrew Gallagher, Liat Kaver, Radhika Marvin, Caroline Pantofaru, Nathan Reale, Loretta Guarino Reid, Kevin Wilson, et al. 2018. Ava-speech: A densely labeled dataset of speech activity in movies. arXiv preprint arXiv:1808.00606.

Tuan M Dang, Theron S Wang, Hridayesh Lekhak, and Kenny Q Zhu. 2025. Emotionalcanines: A dataset for analysis of arousal and valence in dog vocalization. In Proceedings of the 33rd ACM International Conference on Multimedia, pages 13281–13288.

Monica Domínguez, Iván Latorre, Mireia Farrús, Joan Codina-Filba, and Leo Wanner. 2016. Praat on the web: An upgrade of praat for semi-automatic speech annotation. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, pages 218–222.

Olivier Friard and Marco Gamba. 2016. Boris: a free, versatile open-source event-logging software for video/audio coding and live observations. Methods in ecology and evolution, 7(11):1325–1330.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 776–780. IEEE.

Bongjun Kim and Bryan Pardo. 2018. A human-in-the-loop system for sound event detection and annotation. ACM Transactions on Interactive Intelligent Systems (TiiS), 8(2):1–23.

Michael Kipp. 2001. Anvil-a generic annotation tool for multimodal dialogue. In Proc. Eurospeech 2001, pages 1367–1370.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28:2880–2894.

Hridayesh Lekhak, Theron S Wang, Tuan M Dang, and Kenny Q Zhu. 2025. Dogspeak: A canine vocalization classification dataset. In Proceedings of the 33rd ACM International Conference on Multimedia, pages 13369–13375.

Steven R Livingstone and Frank A Russo. 2018. 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.

Annamaria Mesaros, Toni Heittola, Tuomas Virtanen, and Mark D Plumbley. 2021. Sound event detection: A tutorial. IEEE Signal Processing Magazine, 38(5):67–83.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. Elan: A professional framework for multimodality research. In 5th international conference on language resources and evaluation (LREC 2006), pages 1556–1559.

Zheng Zhang, Zheng Ning, Chenliang Xu, Yapeng Tian, and Toby Jia-Jun Li. 2023. Peanut: A human-ai collaborative tool for annotating audio-visual data. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, pages 1–18.