Audio Engineering Society

Convention Express Paper 339

Presented at the 158th Convention

2025 May 22-24, Warsaw, Poland

This Express Paper was selected on the basis of a submitted synopsis that has been peer-reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This Express Paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

An audio quality metrics toolbox for media assets management, content exchange, and dataset alignment

Nikolaos Vryzas[1], Iordanis Thoidis[1], Lazaros Vrysis[1], George Papanikolaou[1], Charalampos Dimoulas[1]

[1] *Aristotle University of Thessaloniki, University Campus, 54124 Thessaloniki, Greece*

Correspondence should be addressed to Nikolaos Vryzas (nvryzas@auth.gr)

ABSTRACT

Content exchange and collaboration serve as catalysts for repository creation that supports creative industries and fuels model development in machine learning and AI. Despite numerous repositories, challenges persist in discoverability, rights preservation, and efficient reuse of audiovisual assets. To address these issues, the SCENE) project introduces an automated audio quality assessment toolkit integrated within its Media Assets Management (MAM) platform. This toolkit comprises a suite of advanced metrics, such as artifact detection, bandwidth estimation, compression history analysis, noise profiling, speech intelligibility, environmental sound recognition, and reverberation characterization. The metrics are extracted using dedicated Flask-based web services that interface with a data lake architecture. By streamlining the inspection of large-scale audio repositories, the proposed solution benefits both high-end film productions and smaller-scale collaborations. This paper presents the motivation, design, and implementation details of the toolkit, highlighting its potential to assess content quality management and contribute to more efficient content exchange in the creative industries.

1 Introduction

Content exchange and collaboration can act as a catalyst for the creation of repositories that can be used in the creative industries, as well as model creation and advances in machine learning and AI systems. While there are many available repositories, discoverability, rights preservation, evaluation, and reusability of assets are objectives that can be addressed more efficiently. The sharing of models, files, and assets can improve productivity and also enhance the democratization in content production from small-to-medium organizations and low-budget productions.

A key challenge in large-scale data exchange is the prevalence of heterogeneous schemas and ontologies, which complicate the creation of unified repositories containing multi-labeled assets from diverse origins. Although manual dataset conversion, inspection, and alignment are still common, automated approaches of metadata propagation are increasingly enabling sharing labels and consistent metadata across diverse sources. At the same time, effective content reuse depends not only on discoverability and legal compliance, but also on ensuring sufficient asset quality. While machine learning can benefit from varied-quality content to boost generalization, domains such as film production require consistently high-quality material, making manual inspection impractical at scale. These considerations underscore the importance of developing tools and strategies that enable both large-scale content sharing and effective quality control.

Besides the subjective evaluation of content quality, scalability in big repositories requires the use of automated processes, metrics and evaluation methodologies [1]. For such efforts, databases like the LIVE-SJTU Audio and Video Quality Assessment (A/V-QA) have been developed. It includes sequences generated from original content, by applying 24 different A/V distortion combinations on them [2]. ViSQOL, implemented in MATLAB and C++ is integrated for measuring audio quality in various Google projects [3].

This paper presents a web-based approach for quality assessment of audio file repositories to enable merging in a common data lake. The remainder of the paper is structured as follows: Section 2 presents the SCENE project architecture for audio quality metrics extraction, Section 3 describes the different kind of metrics that have been integrated in the current version of the platform, and in Section 4 the conclusions and future research directions are summarized.

2 A data lake architecture for repository merging

A major bottleneck in data exchange is the existence of different datasets with varying schemas and ontologies. This complicates the process of merging existing datasets into supersets. The process of adapting a dataset to merge it with other similar sets is called dataset conversion and alignment. It is a task that is often addressed manually, but automated processes have also been proposed. This allows the definition of common labels, the propagation of metadata from one set to another, for the creation of a multi-labeled set that shares some common definitions across assets from different origins.

Besides discoverability and legal issues that emerge from reusing other creators' content, the issue of the content quality is important, especially in the context of content production.

For model training, having content of varying quality in the training set may enhance the aspects of generalization. However, in cases where quality is a prerequisite, like film production, the manual inspection of all available content may add prohibitive complexity to the process.

The SCENE (Searchable multi-dimensional Data Lakes supporting Cognitive Film Production & Distribution for the Promotion of the European Cultural Heritage) is a Horizon research project with the objective of supporting the European film industry through technologies that promote collaboration [4]. Among other tools, it incorporates in the platform a Media Assets Management (MAM) tool, where users can access audiovisual content and more from aligned repositories that are accessible through the platform. It covers property rights and monetization issues, providing the licence for reusing the assets.

In the current research, a toolkit for the automation of the inspection of content quality is proposed and implemented. It is designed to address the requirements that have been set in the analysis phase of the SCENE project, and to be integrated in the MAM tool. The objective is to provide through an automated process a series of audio quality indicators, using traditional signal processing and machine learning techniques, as audio assets in a searchable database that is accessible to the users and creators.

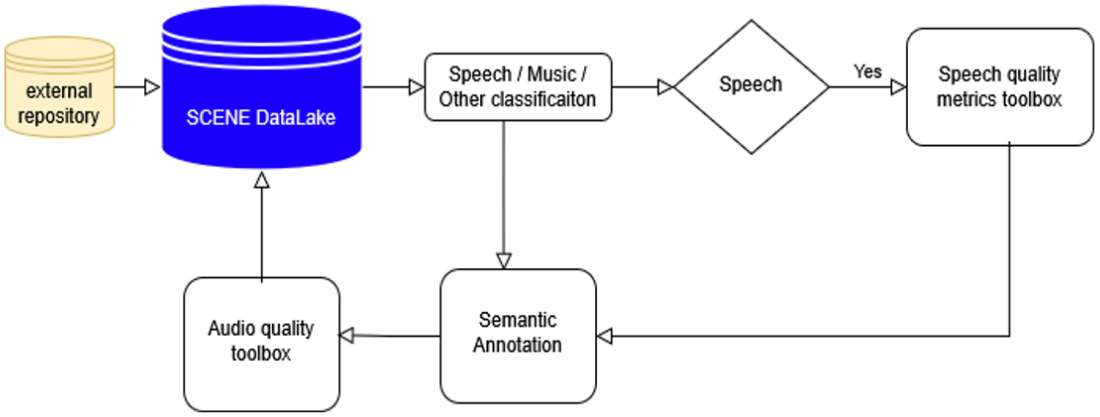


Figure 1. The DataLake architecture for semantic annotation and quality assessment of external repositories

The primary quality metrics include audio artifact detection, bandwidth estimation, and analysis of compression history regardless of current file format. Additional features incorporate noise profile estimation, speech intelligibility assessment, hierarchical classification for general audio tagging and environmental sound recognition, and reverberation parameter estimation. In terms of software architecture, every quality metric is extracted through a dedicated web service that is provided through an API. The web services are implemented using the Flask lightweight WSGI web application framework for Python. The modular architecture, along with the use of a data lake that allows storing unstructured and heterogenous data, promotes the extendability of the system in the future, as well as the update of the existing web services.

3 Audio quality metrics and semantics

The audio quality assessment toolkit integrates various quality metrics through a modular web-service architecture. These metrics are primarily non-intrusive, given that reference audio sources, such as clean speech signals or isolated sounds, are typically unavailable in practical repository contexts.

Recognizing the limitations of traditional intrusive methods, a comprehensive literature research was performed, identifying advanced neural network-based methods for reliably estimating audio quality metrics without reference signals. The toolkit is implemented using Python, leveraging PyTorch for deep learning-based metrics, and integrates existing open-source solutions such as VERSA (Versatile Evaluation of Speech and Audio) [5] and the Audio Quality Platform (AQP) framework [6]. The metrics provided are categorized hierarchically, as described below.

General media information

This includes essential information about the audio files, such as encoding format, file size, duration, sampling frequency, and the number of audio channels. This metadata facilitates basic filtering and categorization tasks within the repository.

Reverberation estimation

The toolkit employs deep learning models capable of non-intrusively estimating reverberation parameters, to provide insights into the acoustic conditions under which audio content was recorded. Specifically, a deep learning model [7] estimates reverberation characteristics, such as wet ratio, early reflection energy, and decay time. Moreover, the Speech-to-Reverberation Modulation Energy Ratio metric further quantifies speech clarity affected by reverberation.

Audio quality

Audio quality assessment metrics include deep-learning models tailored to different audio types. For music audio, the toolkit integrates the pre-trained HAAQI-Net, a neural network specialized in music audio quality evaluation. For speech quality assessment, DNSMOS is integrated, a perceptually motivated, non-intrusive objective metric specifically designed for evaluating noise suppression algorithms. Additionally, the Non-Intrusive Speech Quality Assessment (NISQA) model is incorporated for a comprehensive assessment of speech quality and naturalness [8].

Noise estimation

Noise metrics evaluate background noise conditions during speech and non-speech segments. When speech is detected, the toolkit calculates noise profiles, including Signal-to-Noise Ratio (SNR), using statistical and neural-network-based estimation techniques [8]. These metrics can be useful to users in order to identify content requiring noise suppression or audio restoration.

Audio Compression History

Transcoding to lossless formats is a very common problem in audio repositories, especially in music tracks, where the encoding information that is extracted from the file itself is not reliable. Double encoding techniques, bandwidth estimation, audio artifacts recognition [9] can act as indicators of the compression history of audio files, regardless of their current format. The current toolkit implements metrics that identify double encoding, perform bandwidth estimation, and detect audio artifacts typically associated with compression [10], [11].

Spatial Audio quality

The SCENE data lake is expected to include spatial audio files. Within the workflow of the platform, users create and submit ambisonics audio files that are used for acoustic simulation of locations related to cultural heritage. Such files may include Ambisonics Room Impulse Responses (ARIR),

sinusoidal sweeps, but also environmental audio recordings and spatial audio effects to be used in post-production. A common issue that has been discovered in the requirements analysis when working with User Generated content is the appearance of misleading audio encodings, where monophonic audio files are encoded using spatial audio formats. In the presented toolbox, a first check is performed to ensure that channels in multi-channel formats are not identical. Additionally, Spatialization Quality (SQ) and Listening Quality metrics are extracted using tools like AMBIQUAL [12] and SAQAM [13].

Environmental audio recognition

Hierarchical environmental audio classification employs a convolutional neural network to organize audio segments initially into broad categories (speech, music, other). Segments labeled "other" undergo further classification using the ESC-50 environmental sound classification scheme, enhancing discoverability and metadata accuracy. This hierarchical approach also aids in identifying mislabelled or undocumented audio assets [14].

Speech Intelligibility

For speech content, the toolkit integrates non-intrusive intelligibility metrics that estimates speech perception attributes for normal-hearing listeners. Specifically, it utilizes STOI-Net [15] a model that directly estimates the Short-Time Objective Intelligibility (STOI) scores from speech spectral features, facilitating rapid assessment of speech intelligibility in large-scale repositories.

Speaker Characteristics

Speaker-related metrics provide deeper semantic analysis whenever voice activity is detected. Leveraging open-source implementations (such as those based on Qwen2), the toolkit extracts multiple speaker-related characteristics including, speaker count, speaker gender and age, speech impairment indicators, voice pitch and pitch range, speech volume level and rate, speaking style, and detection of expressive vocalizations such as laughter or crying.

4 Conclusions and future research directions

The paper presents a toolkit for automated audio quality assessment within the context of the SCENE project, addressing the needs of filmmaking and content sharing. The toolkit is composed of various metrics investigating several quality aspects and semantics. In the following months, it will be evaluated in pilot cases by professionals of the field. The pilot phase of the toolkit will involve professional filmmakers who will provide feedback to refine post-production workflows.

5 Acknowledgement

This research is part of the SCENE project and has received funding from the European Union's Horizon research and innovation programme under grant agreement No 101095303.

References

[1] Torcoli, M., Kastner, T., & Herre, J. (2021). Objective measures of perceptual audio quality reviewed: An evaluation of their application domain dependence. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 1530-1541.

[2] Min, X., Zhai, G., Zhou, J., Farias, M. C., & Bovik, A. C. (2020). Study of subjective and objective quality assessment of audio-visual signals. IEEE Transactions on Image Processing, 29, 6054-6068.

[3] Chinen, M., Lim, F. S., Skoglund, J., Gureev, N., O'Gorman, F., & Hines, A. (2020, May). ViSQOL v3: An open source production ready objective speech and audio metric. In 2020 twelfth international conference on quality of multimedia experience (QoMEX) (pp. 1-6). IEEE.

[4] Nikolaos Vryzas, Lazaros Vrysis, Marina Eirini Stamatiadou, Charalampos Dimoulas and George Kalliris An acoustic simulation interface for cultural heritage locations scouting.

[5] Shi, J., Shim, H. J., Tian, J., Arora, S., Wu, H., Petermann, D., ... & Watanabe, S. (2024). VERSA: A Versatile Evaluation Toolkit for Speech, Audio, and Music. *arXiv preprint arXiv:2412.17667*.

[6] Geraghty, J., Li, J., Ragano, A., & Hines, A. (2022, June). AQP: an open modular Python platform for objective speech and audio quality metrics. In Proceedings of the 13th ACM Multimedia Systems Conference (pp. 191-196).

[7] Thoidis, I., Vryzas, N., Vrysis, L., Kotsakis, R., Kalliris, G., & Dimoulas, C. (2022, May). Disentangled estimation of reverberation parameters using temporal convolutional networks. In Audio Engineering Society Convention 152. Audio Engineering Society.

[8] Mittag, G., Naderi, B., Chehadi, A., & Möller, S. (2021). NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. *arXiv preprint arXiv:2104.09494*.

[9] Vryzas, N., Katsaounidou, A., Vrysis, L., Kotsakis, R., & Dimoulas, C. (2022). A prototype web application to support human-centered audiovisual content authentication and crowdsourcing. Future Internet, 14(3), 75.

[10] Vryzas, N., Katsaounidou, A., Kotsakis, R., Dimoulas, C. A., & Kalliris, G. (2018, May). Investigation of audio tampering in broadcast content. In *Audio Engineering Society Convention 144*. Audio Engineering Society.

[11] Li, X., Leglaive, S., Girin, L., & Horaud, R. (2019). Audio-noise power spectral density estimation using long short-term memory. IEEE Signal Processing Letters, 26(6), 918-922.

[12] Narbutt, M., Allen, A., Skoglund, J., Chinen, M., & Hines, A. (2018, May). AMBIQUAL-a full reference objective quality metric for ambisonic spatial audio. In 2018 tenth international conference on quality of multimedia experience (QoMEX) (pp. 1-6). IEEE.

[13] Manocha, P., Kumar, A., Xu, B., Menon, A., Gebru, I. D., Ithapu, V. K., & Calamia, P. (2022). SAQAM: Spatial audio quality assessment metric. arXiv preprint arXiv:2206.12297.

[14] Sarris, A. L., Vryzas, N., Vrysis, L., & Dimoulas, C. (2024). Investigation of Data Augmentation Techniques in Environmental Sound Recognition. Electronics, 13(23), 4719.

[15] R. E. Zezario, S.-W. Fu, C.-S. Fuh, Y. Tsao, and H.-M. Wang, “STOINet: A deep learning based non-intrusive speech intelligibility assessment model,” in Proc. APSIPA ASC, 2020