Re-thinking Digital Audio via Unlimited Sensing: Theory, Algorithms, and Hardware Validation

Ruiming Guo, Yuliang Zhu and Ayush Bhandari

***Abstract*—The Shannon–Nyquist principle underpins mainstream digital audio, relying on equidistant sampling and uniform quantization. This architecture fundamentally limits dynamic range and digital precision, leading to irreversible clipping when signals exceed the converter's input range and to quantization noise that caps effective resolution—constraining downstream algorithms and machine listening. We introduce Unlimited Sensing Framework–based Audio (USF-Audio), an acquisition-centric framework that jointly redesigns the analog front-end and recovery algorithms to overcome these bottlenecks at the analog–digital interface. USF-Audio employs structured non-linear acquisition via analog-domain folding (modulo wrapping) prior to digitization, replacing saturation with bounded measurements that preserve low-amplitude detail while capturing high dynamic range (HDR) signals within a fixed input range. USF-Audio reframes quantization noise as the digital representation of audio, enabling mathematical recovery to reconstruct the original waveform with high fidelity. This paradigm resolves the long-standing problem of audio declipping and yields clear improvements in effective resolution under a fixed bit budget. We establish theoretical recovery guarantees for quantized, folded measurements; develop scalable and robust algorithms suitable for audio-scale data; and implement a complete hardware acquisition pipeline. Experiments on 14 real-world datasets—including speech, music, biomedical sounds, animal vocalizations, and historical recordings, and a 52-participant subjective listening study—showcase consistent gains in dynamic range, effective number of bits, and perceptual quality over conventional ADC pipelines and state-of-the-art declipping methods. These results suggest that fundamental limitations in digital audio stem from the acquisition model itself, and that principled redesign of sensing can open new pathways for audio acquisition and processing beyond what is achievable through post-hoc processing alone.**

***Index Terms*—Audio signal processing, declipping, quantization, non-linear acquisition, sparsity and unlimited sensing.**

I. Introduction

A. Digital Audio is Mature Above the Acquisition Layer

SOUND and audio carry rich information about human communication, emotion, spatial geometry, and environmental context. Beyond speech and music, acoustic signals often encode cues that are inaccessible to other sensing modalities, making audio indispensable in applications ranging from immersive media and healthcare monitoring to bioacoustics and scientific instrumentation. The digitization of sound has therefore played a foundational role in modern signal processing, enabling scalable storage, transmission, and algorithmic analysis. From a systems perspective, contemporary audio technology can be viewed as a layered *Audio Processing Stack* (APS) in Table I, spanning physical acoustics, analog front-end conversion, digital representation, core signal processing, contextual inference, semantic understanding, and application-level integration. Over several decades, this layered architecture has supported remarkable innovation above the acquisition layer [1], including advances in time–frequency analysis, spatial audio, perceptual coding, and learning-based inference. In contrast, the principles governing how acoustic signals enter the digital domain have remained largely unchanged.

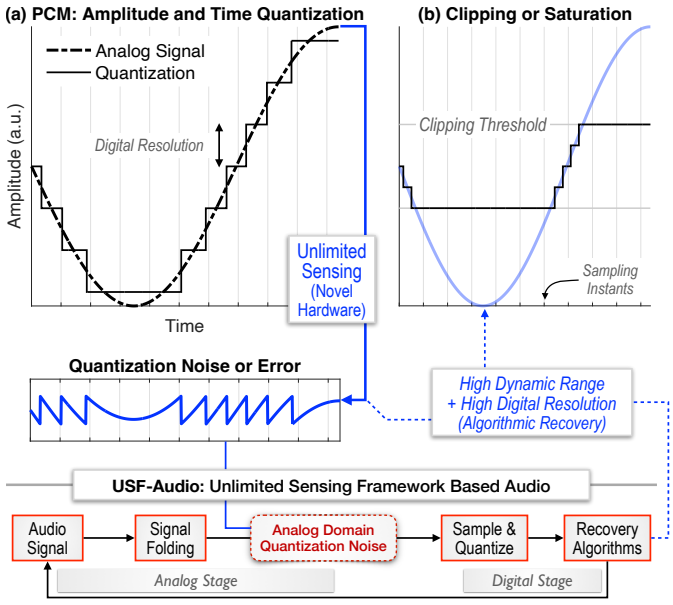


Fig. 1: Digitization via pulse–code modulation (PCM) [2] and its fundamental limitations. (a) Sampling or time quantization preserves information under the Shannon–Nyquist criterion, whereas amplitude quantization introduces irreversible loss manifested as quantization noise. (b) Signals exceeding the admissible range incur additional loss through saturation or clipping. Unlike PCM, which is prone to clipping and imposes a trade-off between dynamic range and resolution, this paper capitalizes on the Unlimited Sensing Framework (USF). The USF-Audio pipeline digitizes analog-domain quantization noise via novel hardware which eliminates clipping and enables algorithmic recovery with both high dynamic range and high digital resolution (see Fig. 6).

B. Fundamental Limits at the Analog–Digital Interface

The analog–digital interface, typically implemented via pulse–code modulation (PCM) [2]–[5] and its variants [6]–[8], forms the entry point from the physical to the digital domain. This process combines uniform sampling with amplitude

This work is supported by the European Research Council's Starting Grant for "CoSI-Fold" (101166158). Further details on Unlimited Sensing and materials on *reproducible research* are available via https://bit.ly/USF-Link.

The authors are with the Computational Sensing and Imaging Lab (CSIL) at the Department of Electrical and Electronic Engineering, Imperial College London, South Kensington, London SW7 2AZ, UK. (Emails: {ruiming.guo,yuliang.zhu19,a.bhandari}@imperial.ac.uk or ayush@alum.mit.edu).

TABLE I: Functional Audio Processing Stack (Reference Architecture).

Layer	Attribute	Primary Role	Key Functions	Standards / Frameworks
1	**Physical Acoustics** *Sound field & environment*	Governs sound generation and propagation	Sound radiation, propagation, reflection, absorption, reverberation; loudspeaker-room interaction; listening environments	ISO 3382 (Room Acoustics), ITU-R BS. 1116 (Listening Tests)
			Analog-Digital Interface	
2	**Analog Front-End & Digital Conversion** *Sensors & Circuitry*	Converts acoustic pressure to digital signals	Microphone transduction, pre-amplification, analog filtering, clocking & sync., analog-to-digital conversion / ADC (sampling & quantization)	IEC 61672 (Microphones), AES17 (Digital Audio Measurements)
3	**Digital Representation** *Signals & Transforms*	Defines digital audio representation	PCM or 1-bit samples; analysis/synthesis transforms (STFT, filter banks, wavelets, dictionaries); parametric and learned representations	ITU-T G. 711 (PCM), MPEG-1/2 Audio, MPEG-4 AAC/HE-AAC, MPEG-H 3D
4	**Core Signal Processing** *Enhancement & Separation*	Improves and manipulates audio signals	Reconstruction • Restoration • Noise suppression Dereverberation • Echo cancellation Source separation • Beamforming • Spatial filtering	ITU-T G. 168/G. 169 (Echo Control), IWAENC, REVERB Challenge, SiSEC
5	**Scene & Spatial Context** *Geometry & Environment*	Infers acoustic geometry and environment	Acoustic scene analysis, source localization & tracking, room modeling/simulation, spatial coherence analysis	ACE Challenge (Room Acoustics), DCASE (Scene/Event Analysis), Ambisonics & array reference formats
6	**Semantic & Content Understanding** *Meaning & Identity*	Extracts meaning and identity from audio	Speech recognition, speaker identification & diarization, MIR, audio event detection	MPEG-7 (ISO/IEC 15938), LibriSpeech, WSJ, CHiME, MIR benchmarks
7	**Application & Interaction** *Delivery to Users*	Delivers audio functionality to users	Streaming & VoIP, AR/VR spatial audio rendering, object-based playback, personalization, hearing devices	RTP/RTCP, WebRTC, MPEG-D SAOC, MPEG-H Immersive Audio

quantization (Fig. 1), a paradigm originating in early years of telecommunication [2]–[5] and formalized by Shannon's theory [9]. While PCM-based analog-to-digital converters (ADCs) are robust and widely adopted, they impose intrinsic limitations at digitization. Two bottlenecks are fundamental:
L1] **Limited Dynamic Range** [10]. Any ADC admits only a bounded input range. Signals exceeding this range incur saturation or clipping (Fig. 1(b)), leading to irreversible information loss; this has been a longstanding problem [11]–[13].
L2] **Range–Resolution Trade-off**. For a fixed bit budget or power envelope, dynamic range (DR) and digital resolution (DRes) cannot be simultaneously maximized. Expanding the DR coarsens quantization (Fig. 1(a)), while increasing resolution narrows the DR and exacerbates clipping (Fig. 1(b)).

These limitations originate at acquisition (Layer 2, Table I) and propagate through all subsequent processing stages or layers. No matter how sophisticated downstream algorithms may be, they operate on data whose information content has already been capped. Increasing bit-depth as a solution is impractical due to power and hardware constraints [14].

C. Algorithmic Remedies: Overview and Limits

Clipping and coarse quantization have spurred decades of algorithmic research; declipping, dequantization, dithering, noise shaping, and learning-based restoration. Moreover, these two issues—clipping and coarse quantization—are typically addressed independently, despite their coupled impact on dynamic range and resolution. Below is a brief overview.

❏ *High Dynamic Range (HDR) Audio and Declipping.* Algorithmic declipping was introduced for speech in the early 1990s by Abel & Smith [11]. Three decades later, comprehensive surveys by Záviška *et al.* [12] and Gaultier *et al.* [13] emphasize that, despite sustained progress, declipping remains a persistent and fundamentally limited problem. Beyond perceptual enhancement, clipping critically impacts domains such as healthcare monitoring [15] and audio forensics [16], [17], where high-amplitude transients carry essential information. Complementary hardware strategies include synchronized ADCs with offset gains [18], HDR-inspired fusion techniques [19], [20], and automatic gain control [21].

❏ *High Digital Resolution (HDRes).* Resolution enhancement via dithering was suggested by Jayant & Rabiner [22] and Vanderkooy & Lipshitz [23], [24] and remains widely used alongside noise shaping and oversampling [8]. Applications including healthcare [25], biodiversity monitoring [26], [27], and extra-terrestrial missions [28] routinely require 16–24 bits, highlighting the enduring need for HDRes acquisition.

Despite the advances, these methods assume a fixed acquisition model. Once amplitude information is lost during digitization, recovery is fundamentally limited. This creates a paradox: algorithmic complexity grows, yet improvements are marginal because sensing constrains the digital data. This raises a foundational question: *are current limitations algorithmic–or rooted in how we sense audio in the first place?*

D. Reframing Quantization Noise as the Signal

Shannon or PCM-style systems treat quantization noise (QN) [3] (see Fig. 1) as an undesirable artifact, discarding the fractional or modulo part during digitization[1]. Recent insights challenge this view [30]–[32]: quantization can be interpreted as a structured decomposition, where the fractional/modulo part encodes information about the signal's integer part. Under smoothness priors, *this "noise" becomes the signal*. Realizing this potential requires moving beyond linear acquisition and **introducing controlled non-linearities** at the sensing stage as well as developing algorithms that can facilitate reconstruction by stably inverting the non-linearity.

The **Unlimited Sensing Framework (USF)** [30]–[32] formalizes this idea by replacing saturation with analog-domain

[1]Note that QN, the fractional part, and the modulo operation are mathematically equivalent [29]. Conceptually, we distinguish them as follows: the fractional part or modulo representation is interpreted in the continuous domain, whereas after sampling, this interpretation manifests as QN.

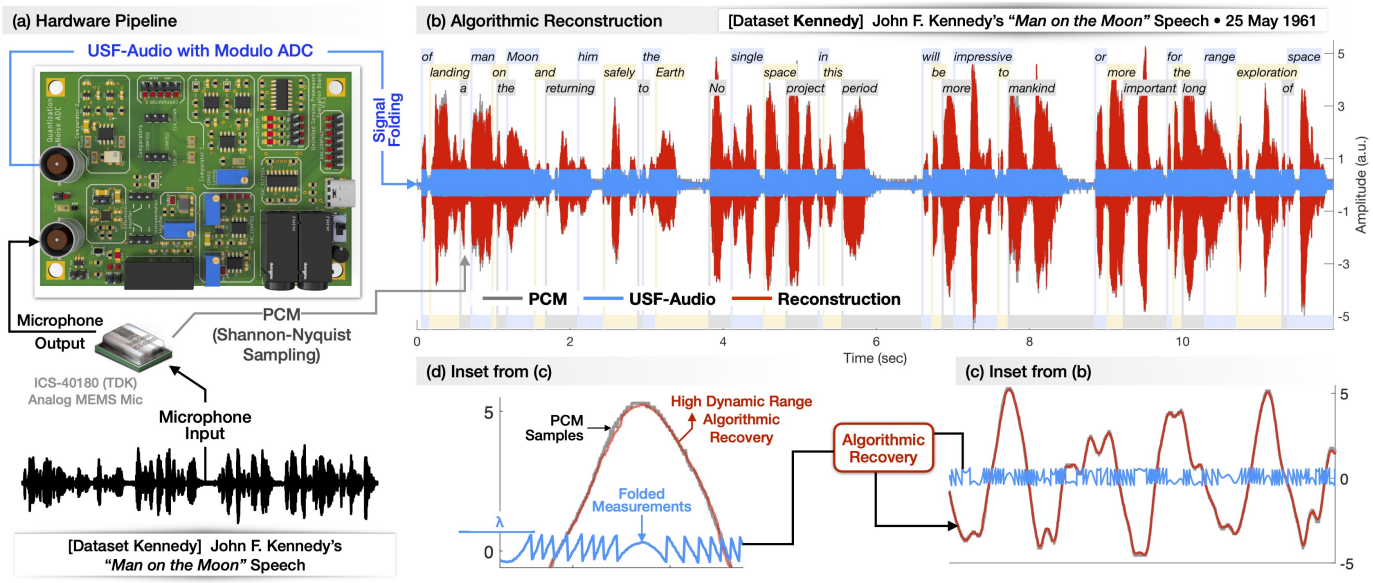


Fig. 2: USF-Audio digitization and recovery. (a) Hardware for generating folded audio. (b) Digital measurements for the Kennedy dataset with speech annotations: conventional PCM (Shannon–Nyquist), folded samples, and reconstruction via USF-ARA. Under the same 6-bit budget, USF-Audio achieves substantial gains—noise floor improvement (≈ 12 dB), dynamic range ($9.50\times$), and resolution of 0.0175 V compared to 0.1188 V for PCM. (c)–(d) Insets from (b) at different time resolutions, illustrating USF-Audio samples. The legend is shown in (d). Audio playback is available here: [◀◀ ▶ ▶▶ ■ ▮▮] https://bit.ly/USFAudio.

folding (Fig. 1). Instead of capturing the full range directly, the signal is *folded and redistributed* into a bounded interval (akin to folding a spaghetti to fit on a fork), ensuring measurements remain within the ADC's range[2]. Subsequently, algorithms recover the HDR signal from folded samples [30]–[37].

Mathematically, folded samples preserve the original signal under Shannon-Nyquist conditions, *breaking the classical range–resolution trade-off*: HDR signals can be acquired with low bit-depths while maintaining HDRes. Details on folding and distinction from related discrete methods (*e.g.* phase unwrapping and modulo PCM) can be found in [31].

For a conventional ADC (hereafter ADC) with B bits, the well-known "6-dB per bit rule" applies [3]; the signal-to-quantization-noise ratio (SQNR) is $\mathsf{SQNR}_B \propto 20\log_{10}(2^B) \approx 6.02B$ dB. In USF [38], however, the SQNR improves as

$$\mathsf{SQNR}_{\lambda,B} = \mathsf{SQNR}_B + 20\log_{10}\left(\max|g(t)|/\lambda\right) \quad (1)$$

where $\{g(t), \lambda\}$ is the input signal and folding threshold (see (2)), and the boost is attributed to the input signal's DR. This gain in SQNR leads to increased DRes and enables new capabilities in diverse application areas. The advantages of USF have been validated across multiple modalities *e.g.* tomography [39], radar [40], communications [38], imaging [41] and sub-Nyquist sampling [42], [43], demonstrating significant gains in DR and DRes. Alternative HDR acquisition techniques are still confined by the *6-dB per bit rule*; their qualitative comparison is provided in the Appendix.

Motivation and Vision. Eliminating clipping at acquisition (L1) and overcoming the range–resolution constraint (L2) would enable simultaneous HDR and HDRes under fixed bit budgets. This shift has far-reaching implications for digital representation and downstream processing (Layers 3 and beyond, of the APS). Yet, most literature—including recent surveys [1]—assumes a fixed acquisition layer, leaving performance capped by data quality. Revisiting *audio from an acquisition-first perspective* through hardware–algorithm co-design potentially opens fundamentally new sensing regimes.

Challenges and Prior Work in USF. While prior USF works establish theoretical, algorithmic, and hardware ingredients, they do not yet make it possible to realize a practical *end-to-end* audio pipeline. In particular, three capabilities remain unresolved. *i) Audio-specific theory.* Starting with [30], existing USF theory does not yet provide a recovery guarantee in terms of dynamic-range expansion (DRX[3]), sampling rate, and bit-depth (B). Such a result is important because, for audio acquisition, it determines when oversampling can be traded effectively against B and DRX. *ii) Practical large-scale recovery.* Existing USF recovery methods are not designed for hardware audio pipelines and voluminous audio streams. *Time-domain* methods in [30], [31], [34], [44] rely on ideal folding; the *Fourier-domain* method [32] and its variants [45] assume periodic boundary conditions. A recent approach [46] uses time-frequency domain recovery but relies on the folding-instant information for reconstruction from ideal folds. *iii) Audio-band hardware realization.* Prior single-channel MADC (modulo ADC) hardware were developed primarily as laboratory prototypes to verify the USF concept [32], [47]–[49], but they do not directly constitute an audio acquisition front-end. Audio-band sensing requires higher bandwidth, robust operation with microphone hardware, and tight integration with the recovery algorithm. Supporting signals up to 20 kHz together with substantial DRX therefore necessitates a new hardware realization rather than a direct reuse of earlier prototypes. These unresolved challenges directly motivate the contributions of this paper.

[2]A hardware demo of live folding: https://youtu.be/JuZg80gUr8M.

[3]Dynamic-range expansion (DRX) is defined as $\mathsf{DRX} = \|g\|_{\mathbf{L}_\infty}/\lambda$.

Contributions. This paper introduces USF-Audio (see Fig. 2), a principled redesign of digital audio based on the USF. We organize our contributions around a single principle: *analog-domain folding enables simultaneous gains in dynamic range (DR) and digital resolution (DRes) under a fixed bit budget.* C1 **Acquisition-Theoretic Characterization:** We develop a recovery theory for the new digital audio representation, expressed directly in acquisition parameters. In particular, we derive a quantitative bit–sampling law linking DR expansion, bit-depth, and sampling rate; establish a tighter oversampling condition removing the legacy $2e$ factor [31]; and derive an *audio-centric* refinement based on spectral moments, yielding less conservative sampling conditions for spectrally concentrated signals. C2 **New Recovery Algorithm:** We introduce a new frame-wise recovery algorithm for hardware-acquired folded audio. Our approach uses DCT-domain (discrete cosine transform) separation to suppress spectral leakage via finite-frame energy compaction, enabling scalable and robust residue estimation under quantization, noise, and non-ideal folding. C3 **End-to-End Hardware Realization:** We implement a complete USF-Audio pipeline integrating a microphone front-end with a modulo ADC. The system is designed for audio-band signals and operates under practical constraints, including finite bandwidth, noise, and non-ideal analog folding. C4 **Multi-Level Experimental Validation:** We validate USF-Audio across theory, hardware, and perception. Experiments confirm the predicted bit–sampling gains, demonstrate consistent improvements in DR, SINAD (signal-to-noise-and-distortion ratio), and ENOB (effective number of bits) under identical nominal bit-depth, and show objective and perceptual quality gains across 14 real-world recordings, including a 52-participant listening study.

Discussion and Positioning. USF-Audio is a well-positioned, holistic framework *that unifies sensing and recovery* at the acquisition level (see Layer 2, Table I). Folding before digitization ensures backward compatibility with PCM [30], [32] and its variants, [34]–[37], [50], while enabling HDR and HDRes to coexist under fixed bit budgets. Unlike declipping methods that attempt to recover lost information, USF-Audio prevents loss by design, preserving resolution through amplitude redistribution rather than stochastic techniques such as dithering or noise shaping. Hardware validation and experiments across diverse audio classes confirm that these gains stem from acquisition-level innovation, not content-specific assumptions.

E. Scope and Organization of this Paper

This paper focuses on end-to-end audio digitization under HDR conditions, emphasizing hardware–algorithm co-design. Our scope centers on acquisition-driven innovation encompassing not only the 1) sensing hardware, but also 2) the digital representations, and 3) signal processing that follow. We examine a] *how non-linear analog folding reshapes the sensing pipeline*, b] *how folded measurements can be processed directly or reconstructed* computationally, and c] h*ow these principles translate into practical audio-scale hardware and systems*. The goal is not to propose new perceptual models, codecs, or content-specific priors, but to **re-examine the digital audio front-end** and its **implications**, and demonstrate that HDR and HDRes can be achieved simultaneously in a backward-compatible fashion.

The paper is organized as follows. Section II formalizes the USF-Audio acquisition model and contrasts it with conventional PCM-based front-ends, highlighting the implications of analog-domain folding for DR and resolution. Section III presents signal recovery under quantization with theoretical guarantees. Section IV develops a recovery method tailored to folded audio measurements. Section V presents experimental validation across a wide range of audio applications, demonstrating robustness, generality, and consistency with the theoretical claims. Finally, we conclude in Section VI with broader implications for the field. The notation section is provided in the Appendix.

II. USF-Audio Acquisition Model

As shown in Fig. 2, the USF-Audio acquisition comprises: i) a microphone converting the acoustic waveform into an electrical signal g, and ii) innovative modulo ADCs (MADC) that fold the analog waveform before digitization [32], [48].

For an input $g \gg 1$, decompose $g = \lfloor g \rfloor + [\![g]\!]$, where $\lfloor \cdot \rfloor \in \mathbb{Z}$ and $0 \leqslant [\![\cdot]\!] < 1$ denote the integer and fractional parts, respectively. Conventional schemes—PCM, compressed sensing, and $\Sigma\Delta$ [51] or noise shaping—encode in terms of $\lfloor g \rfloor$, which grows with $\|g\|_{\mathbf{L}_\infty}$, discarding $[\![g]\!]$. This approximation is reasonable but introduces non-linear distortions such as clipping and coarse quantization [12]. USF exploits smoothness (*e.g.* bandlimited and spline classes) to recover the integer part from the fractional part. Thus, USF-Audio digitizes $[\![g(t)]\!] = g(t) - \lfloor g(t) \rfloor$, via modulo ADCs (MADC)—the analog counterpart of QN [52]. Note that $[\![g]\!]$ is a bounded signal, $|[\![g]\!]| \leqslant 1$. Allocating B bits to $[0, 1)$ yields higher resolution than spreading them across $[\min(g), \max(g)]$.

Problem Formulation. Let $g \in \mathcal{B}_\Omega \cap \mathbf{L}_2$ denote a continuous-time, finite-energy bandlimited signal with maximum angular frequency Ω. This signal is processed by the MADC, which applies a centered modulo non-linearity before digitization:

$$\mathscr{M}_\lambda : g \mapsto 2\lambda \left(\left[\!\!\left[\frac{g}{2\lambda} + \frac{1}{2} \right]\!\!\right] - \frac{1}{2} \right), \quad \lambda > 0, \tag{2}$$

where λ is a design parameter and $\mathscr{M}_\lambda(\cdot)$ folds $g \gg \lambda$ into the interval $[-\lambda, \lambda]$, ensuring bounded amplitude. The folded signal $y(t) = \mathscr{M}_\lambda(g(t))$ is sampled, *i.e.*, $y[n] = \mathscr{M}_\lambda(g(nT))$, with sampling step $T > 0$. Akin to Shannon-Nyquist criterion, the following sampling theorem guarantees recovery.

Theorem 1 ([30], [31]). *Let $g \in \mathcal{B}_\Omega \cap \mathbf{L}_2$ and $y[n] = \mathscr{M}_\lambda(g(nT))$. Then, $T\Omega \leqslant (2e)^{-1}$ is a sufficient condition for recovery of $g(t)$ from $y[n]$ (up to an unknown constant).*

A B-bits MADC quantizes $y(t)$ resulting in,

$$y_q[n] \stackrel{\text{def}}{=} \mathscr{Q}_{\eta_\lambda^B}\left(y[n]\right), \quad \mathscr{Q}_{\eta_\lambda^B}(\cdot) = 2\eta_\lambda^B \left\lfloor \tfrac{(\cdot)}{2\eta_\lambda^B} + \tfrac{1}{2} \right\rfloor, \tag{3}$$

where $\eta_\lambda^B = \frac{\lambda}{2^B - 1}$. As $B \to \infty$, $y_q \to y$.

Objectives. Given folded samples $\{y_q[n]\}_{n\in\mathbb{I}_N}$[4], our goals are to: 1) derive sampling conditions guaranteeing recovery, and 2) efficiently recover $g[n] \gg \lambda$ from $y_q[n]$.

III. Recovery Guarantee for USF-Audio

Theoretical Performance Bounds. Since $[\![g(t)]\!] = g(t) - \lfloor g(t) \rfloor$, in view of the modulo folding with threshold λ [30],

$$\mathscr{M}_\lambda(g) = g - \varepsilon_g, \quad \varepsilon_g(t) = \sum\nolimits_{m=0}^{M-1} c_m \mathbb{1}(t-\tau_m) \tag{4}$$

where M denotes the folding count of g, $\mathbb{1}(\cdot)$ is the indicator function, and $\{c_m \in 2\lambda\mathbb{Z}, \tau_m \in T\mathbb{Z}\}$ parameterize the residue ε_g. When working with quantized samples (3), we can write,

$$y_q[n] = y[n] + q[n], \quad \|q\|_{\ell_\infty} \leqslant \eta_\lambda^B = \tfrac{\lambda}{2^B-1} \tag{5}$$

where q denotes QN (bounded by η_λ^B). Combining (4) and (5),

$$y_q[n] = g[n] - \varepsilon_g[n] + q[n]. \tag{6}$$

In the ideal case, $B \to \infty \Rightarrow y_q \to y$, inversion [30], [31] of the modulo non-linearity is facilitated by non-linear filtering of smooth g from non-smooth ε_g. Let $\Delta^{(h)}g[n]$ denote the h-th order finite difference of g. Then, $T \leqslant 1/(2\Omega e)$ guarantees that for some $h \in \mathbb{Z}_+$, $\Delta^{(h)}g[n] < \lambda$ and $\mathscr{M}_\lambda(\Delta^{(h)}y) = \Delta^{(h)}g$. The USF algorithm in [31] then recovers $g[n]$ from $\Delta^{(h)}g[n]$ by estimating ε_g. In USF-Audio, B is finite, so the interaction among $\{B, T, \mathsf{DRX}\}$ becomes central. Unlike ideal USF [30], quantization creates a basic tension: oversampling makes $\Delta^{(h)}g$ shrink, while differencing amplifies $\Delta^{(h)}q$. Recovery is therefore possible only if, $\left\|\Delta^{(h)}g[n]\right\|_{\ell_\infty} \leqslant \frac{\lambda}{2}$, $\left\|\Delta^{(h)}q[n]\right\|_{\ell_\infty} \leqslant \frac{\lambda}{2}$. Accordingly, we seek a tighter oversampling than $T \leqslant 1/(2\Omega e)$ and, simultaneously, the minimum bit-depth B required to support a prescribed dynamic-range expansion DRX. The next theorem formalizes this result.

Theorem 2. *Let $g \in \mathcal{B}_\Omega \cap \mathbf{L}_2$ and $y_q[n] = \mathscr{Q}_{\eta_\lambda^B}(\mathscr{M}_\lambda(g(nT)))$ with $\eta_\lambda^B = \frac{\lambda}{2^B-1}$ and $B \in \mathbb{Z}_+$. Then $\widetilde{g}[n]$ can be recovered from $y_q[n]$ with error bounded by η_λ^B (up to a global offset),* viz. $\min_{\gamma\in\mathbb{Z}} \|\widetilde{g}[n] - g[n] + 2\gamma\lambda\|_{\ell_\infty} \leqslant \eta_\lambda^B$, *provided that*[5]

$$\begin{cases} T \leqslant \frac{1}{2^\alpha \Omega}, \ \alpha > 0 \\ B > \log_2\left(1 + 4\sqrt[\alpha]{2\mathsf{DRX}}\right) \end{cases}, \quad \mathsf{DRX} \stackrel{\text{def}}{=} \frac{\|g\|_{\mathbf{L}_\infty}}{\lambda}. \tag{7}$$

Proof. We have $y_q = g - \varepsilon_g + q$ (6), where $\varepsilon_g \in 2\lambda\mathbb{Z}$ and $\|q\|_{\ell_\infty} \leqslant \eta_\lambda^B$. The key step is the contraction estimates.

Signal Contraction. We begin by tightening Theorem 1, showing that $T < 1/\Omega$ suffices (ideal) recovery. By the fundamental theorem of calculus,

$$\Delta^{(1)}g[n] = g((n+1)T) - g(nT) = \int_{nT}^{(n+1)T} \partial_u g(u)du$$
$$\Longrightarrow \left|\Delta^{(1)}g[n]\right| \leqslant \int_{nT}^{(n+1)T} |\partial_t g(u)|\, du \leqslant T\|\partial_t g\|_{\mathbf{L}_\infty} \tag{8}$$

and therefore $\left\|\Delta^{(1)}g[n]\right\|_{\ell_\infty} \leqslant T\|\partial_t g\|_{\mathbf{L}_\infty}$. Since $g \in \mathcal{B}_\Omega$, Bernstein's inequality [31] gives $\|\partial_t g\|_{\mathbf{L}_\infty} \leqslant \Omega\|g\|_{\mathbf{L}_\infty}$. Combining this with the previous bound yields

$$\left\|\Delta^{(1)}g[n]\right\|_{\ell_\infty} \leqslant T\|\partial_t g\|_{\mathbf{L}_\infty} \leqslant \Omega T\|g\|_{\mathbf{L}_\infty}.$$

We use this as the induction seed to extend the results to arbitrary $h \in \mathbb{Z}_+$. To this end, define $f_0(t) \stackrel{\text{def}}{=} g(t)$ and the recurrence relation $f_h(t) \stackrel{\text{def}}{=} f_{h-1}(t+T) - f_{h-1}(t)$, so that $\Delta^{(h)}g[n] = f_h(nT)$. Applying the same to f_{h-1} yields $\left\|\Delta^{(h)}g[n]\right\|_{\ell_\infty} = \|f_h(nT)\|_{\ell_\infty} \leqslant \Omega T\left\|\Delta^{(h-1)}g[n]\right\|_{\ell_\infty}$. By induction, $\left\|\Delta^{(h)}g[n]\right\|_{\ell_\infty} \leqslant (\Omega T)^h\|g\|_{\mathbf{L}_\infty}$. By choosing

$$H \stackrel{\text{def}}{=} \left\lceil \frac{\ln(2\|g\|_{L^\infty}/\lambda)}{\ln((\Omega T)^{-1})} \right\rceil \Longrightarrow \left\|\Delta^{(H)}g[n]\right\|_{\ell_\infty} \leqslant \frac{\lambda}{2}.$$

This explicit induction is precisely what removes the legacy $2e$-factor from the oversampling bound in Theorem 1.

Quantization Noise Amplification. Further, $\left\|\Delta^{(h)}q[n]\right\|_{\ell_\infty} \leqslant 2^h\eta_\lambda^B$. With $T \leqslant (2^\alpha\Omega)^{-1}$, we have $2^H \leqslant 2\sqrt[\alpha]{2\mathsf{DRX}}$, and this gives the lower-bound on B (bit-depth condition),

$$B > \log_2\left(1 + 4\sqrt[\alpha]{2\mathsf{DRX}}\right) \Longrightarrow \left\|\Delta^{(H)}q[n]\right\|_{\ell_\infty} < \frac{\lambda}{2}.$$

Hence $\left\|\Delta^{(H)}g + \Delta^{(H)}q\right\|_{\ell_\infty} < \lambda$, and since $\Delta^{(H)}\varepsilon_g \in 2\lambda\mathbb{Z}$,

$$\mathscr{M}_\lambda(\Delta^{(H)}y_q) = \mathscr{M}_\lambda(\Delta^{(H)}(g - \varepsilon_g + q)) = \Delta^{(H)}g + \Delta^{(H)}q.$$

Thus $\Delta^{(H)}(g+q)$ is known exactly, and thus so is $\Delta^{(H)}\varepsilon_g$.

Recovery via Residue Estimation. Finally, given $\Delta^{(H)}\varepsilon_g$, it remains to recover ε_g. To this end, we recall the method devised in [31]. Since $\varepsilon_g[n] \in 2\lambda\mathbb{Z}$, repeated anti-differencing yields $\mathsf{S}^2(\Delta^{(h)}\varepsilon_g)[n] = \Delta^{(h-2)}\varepsilon_g[n] + 2\lambda\gamma_h n + 2\lambda\gamma_{h-1}$ with $\gamma_h, \gamma_{h-1} \in \mathbb{Z}$. Define $\kappa_{h,J} = \mathsf{S}^2(\Delta^{(h)}\varepsilon_g)[1] - \mathsf{S}^2(\Delta^{(h)}\varepsilon_g)[1+J]$ and given $\varepsilon_g = g - y$, one gets $|\kappa_{h,J} + 2\lambda\gamma_h J| \leqslant 4\|g\|_{\mathbf{L}_\infty}$. Hence, if $J \geqslant \lceil 8\|g\|_{\mathbf{L}_\infty}/\lambda \rceil$, then $\gamma_h = \mathscr{Q}_{\frac{1}{2}}\left(\frac{\kappa_{h,J}}{-2\lambda J}\right)$ is uniquely determined. Recovering $\{\gamma_h\}_{h=2}^H$ recursively yields ε_g up to $2\lambda\mathbb{Z}$, and therefore $\widetilde{g} = y_q + \varepsilon_g = g + q$ up to $2\lambda\mathbb{Z}$. Consequently,

$$\min_{\gamma\in\mathbb{Z}} \|\widetilde{g}[n] - g[n] + 2\gamma\lambda\|_{\ell_\infty} \leqslant \|q[n]\|_{\ell_\infty} \leqslant \eta_\lambda^B. \qquad \square$$

As $\alpha \to 0^+$, the sufficient condition becomes $T < 1/\Omega$ while $B = \mathcal{O}(1/\alpha)$ reflects unquantized case. Theorem 2 gives the generic acquisition law. Its sampling condition is certified using the worst-case bandwidth Ω, and is therefore *universal but conservative*. For audio, this is often pessimistic, since spectral energy is typically concentrated well below Ω. This motivates the introduction of an order-h *spectral moment*.

Definition 1. *Let $g \in \mathcal{B}_\Omega \cap \mathbf{L}_2$ with its Fourier transform $\widehat{g}$. The order-h spectral moment ($h \in \mathbb{Z}^+$) is defined as*

$$\Omega_{\mathsf{eff},h} \stackrel{\text{def}}{=} \left(\frac{\int_{-\Omega}^{\Omega} |\omega|^{2h} |\widehat{g}(\omega)|^2 d\omega}{\int_{-\Omega}^{\Omega} |\widehat{g}(\omega)|^2 d\omega}\right)^{\frac{1}{2h}} \equiv \left(\frac{\|g^{(h)}\|_{\mathbf{L}_2}}{\|g\|_{\mathbf{L}_2}}\right)^{\frac{1}{h}}. \tag{9}$$

$\Omega_{\mathsf{eff},h}$ serves as an effective bandwidth for $g^{(h)}$; for instance, let $g(t) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-t^2/\left(2\sigma^2\right)\right)$, we have $|\widehat{g}(\omega)|^2 =$

[4] N denotes the number of samples.

[5] The oversampling factor (OF) is characterized as $\mathsf{OF} = \frac{\pi}{\Omega T} \geqslant \pi 2^\alpha$.

$\exp\left(-\sigma^2\omega^2\right)$ yielding $\Omega_{\mathsf{eff},h} = \sqrt[2h]{\pi^{-\frac{1}{2}}\Gamma\left(h+\frac{1}{2}\right)}/\sigma$, where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt$ is the *Gamma function*. In particular, $\Omega_{\mathsf{eff},1} = \frac{1}{\sqrt{2}\,\sigma}$. In contrast to the worst-case support edge Ω, the descriptor $\Omega_{\mathsf{eff},h}$ is *signal-adaptive* and therefore better matched to audio structure. As a result, it yields a sharper sampling condition whenever the audio spectrum is sufficiently concentrated. For instance, speech and heart sounds yield $\Omega_{\mathrm{eff},1}/\Omega$ in the range 10^{-3} to 10^{-2}, indicating sizable savings in the sampling rates (see Table I, Supplementary Material).

Theorem 3 (Audio Centric Recovery Guarantee)**.** *Let $g \in \mathcal{B}_\Omega \cap \mathbf{L}_2$ and $y_q[n] = \mathscr{Q}_{\eta_\lambda^B}\left(\mathscr{M}_\lambda(g(nT))\right)$ with $\eta_\lambda^B = \frac{\lambda}{2^B-1}$ and $B \in \mathbb{Z}_+$. Then $\widetilde{g}[n]$ can be recovered from $y_q[n]$ with error bounded by η_λ^B (up to a global offset),* viz. $\min_{\gamma\in\mathbb{Z}} \|\widetilde{g}[n] - g[n] + 2\gamma\lambda\|_{\ell_\infty} \leqslant \eta_\lambda^B$, *provided that*

$$T \leqslant \max_{h\in\mathcal{H}_B}\left(\frac{1}{\Omega_{\mathsf{eff},h}}\right)\sqrt[h]{\frac{\lambda}{2\sqrt{\Omega/\pi}\,\|g\|_{L_2}}}, \tag{10}$$
$$\mathcal{H}_B = \left\{h\in\mathbb{Z}_+ : h < \left\lfloor \log_2\!\left(\tfrac{2^B-1}{2}\right)\right\rfloor\right\}.$$

Proof. As before, repeated use of the fundamental theorem of calculus gives $\left\|\Delta^{(h)}g[n]\right\|_{\ell_\infty} \leqslant T^h\left\|\partial_t^h g(t)\right\|_{\mathbf{L}_\infty}$. Since $\partial_t^h g \in \mathcal{B}_\Omega$, the Bernstein–Nikol'skii inequality yields

$$\left\|\partial_t^h g(t)\right\|_{\mathbf{L}_\infty} \leqslant \sqrt{\tfrac{\Omega}{\pi}}\,\left\|\partial_t^h g(t)\right\|_{\mathbf{L}_2}.$$

Further, by Parseval's theorem and Definition 1,

$$\left\|\partial_t^h g(t)\right\|_{\mathbf{L}_2}^2 = \frac{1}{2\pi}\int_{-\Omega}^{\Omega}|\omega|^{2h}|\widehat{g}(\omega)|^2\,d\omega = \Omega_{\mathsf{eff},h}^{2h}\,\|g(t)\|_{\mathbf{L}_2}^2,$$

hence $\left\|\partial_t^h g(t)\right\|_{\mathbf{L}_2} = \Omega_{\mathsf{eff},h}^h\,\|g(t)\|_{\mathbf{L}_2}$. Therefore,

$$\left\|\Delta^{(h)}g[n]\right\|_{\ell_\infty} \leqslant T^h\left(\sqrt{\tfrac{\Omega}{\pi}}\Omega_{\mathsf{eff},h}^h\right)\|g(t)\|_{\mathbf{L}_2}.$$

By the sampling condition, there exists an admissible h such that $\left\|\Delta^{(h)}g[n]\right\|_{\ell_\infty} \leqslant \lambda/2$. Further, exactly as in Theorem 2, $\left\|\Delta^{(h)}q[n]\right\|_{\ell_\infty} \leqslant 2^h\eta_\lambda^B < \lambda/2$, since $h < \left\lfloor\log_2\left(\frac{2^B-1}{2}\right)\right\rfloor$. Thus $\left\|\Delta^{(h)}g + \Delta^{(h)}q\right\|_{\ell_\infty} < \lambda$. Since $\Delta^{(h)}\varepsilon_g[n] \in 2\lambda\mathbb{Z}$, modular commutativity gives $\mathscr{M}_\lambda(\Delta^{(h)}y_q) = \Delta^{(h)}g + \Delta^{(h)}q$. The remainder of the proof, namely residue estimation, anti-differencing, and the final error bound, is identical to Theorem 2. Thus $\min_{\gamma\in\mathbb{Z}}\|\widetilde{g}[n] - g[n] + 2\gamma\lambda\|_{\ell_\infty} \leqslant \eta_\lambda^B$. □

For a fixed order h, denote by $\{T_{\max}^{(2)}(h), T_{\max}^{(3)}(h)\}$ the maximum admissible sampling steps from Theorem 2 and Theorem 3, respectively[6]. Then, the relaxation factor is

$$\frac{T_{\max}^{(3)}(h)}{T_{\max}^{(2)}(h)} = \left(\frac{\Omega}{\Omega_{\mathsf{eff},h}}\right)\sqrt[h]{\sqrt{\frac{\pi}{\Omega}}\frac{\|g\|_{\mathbf{L}_\infty}}{\|g\|_{\mathbf{L}_2}}}$$

Thus, Theorem 3 is less restrictive on sampling rate whenever $\left(\Omega_{\mathsf{eff},h}/\Omega\right)^h < \sqrt{\frac{\pi}{\Omega}}\frac{\|g\|_{\mathbf{L}_\infty}}{\|g\|_{\mathbf{L}_2}}$. The gain is strongest when the spectrum is concentrated ($\Omega_{\mathsf{eff},h} \ll \Omega$) and the peak-to-energy ratio is moderate-to-large.

Remarks. The key consequences of Theorem 2 are as follows.
R1 Oversampling lowers the bit budget. For fixed DRX, increasing OF reduces the minimum admissible bit-depth. For instance, if $\mathsf{DRX} = \|g\|_{\mathbf{L}_\infty}/\lambda = 16$, then $\alpha = \{1,2\}$ gives $B_\lambda = \{7.01, 4.56\}$ bits, so doubling the oversampling factor lowers the required resolution by 2.45 bits. By contrast, a conventional ADC would require $B \approx \log_2(\mathsf{DRX}) + B_\lambda$, thus costing an additional 4 bits for the same DR/DRes target.
R2 Few bits can support large DR. Rearranging the bit condition in Theorem 2 gives

$$\mathsf{DRX} = \frac{\|g\|_{\mathbf{L}_\infty}}{\lambda} \leqslant \frac{1}{2}\left(\frac{2^B-1}{4}\right)^\alpha. \tag{11}$$

Since $\mathsf{OF} \geqslant \pi 2^\alpha$, increasing oversampling yields an exponential gain in the achievable DR at fixed B. In particular, whenever $B \geqslant \log_2 9 \approx 3.17$, USF-Audio supports substantial HDR expansion with very few bits.
R3 Accuracy improvement by $20\log_{10}(\mathsf{DRX})$ dB. Under the standard uniform-quantization model [5],

$$\mathsf{MSE}_{\lambda,B} = \frac{\left(\eta_\lambda^B\right)^2}{3\mathsf{OF}} \Longrightarrow \boxed{\mathsf{MSE}_{\lambda,B} = (\mathsf{DRX})^{-2}\,\mathsf{MSE}_B} \tag{12}$$

with $\eta_g^B = \|g\|_{\mathbf{L}_\infty}/\left(2^B-1\right)$ and $\eta_g^B \to \eta_\lambda^B$. Thus, USF-Audio improves recovery accuracy by $20\log_{10}(\mathsf{DRX})$ dB relative to conventional quantization at the same nominal bit-depth. For example, a 15-fold DRX yields a 23 dB gain, consistent with the numerical and hardware results in Section V. Related decay laws have also been reported under additional side information, *e.g.* known folding instants [53]; such assumptions are not used in Theorem 3.
R4 Oversampling is only needed at acquisition. In Theorem 3, oversampling enables recovery at the front-end. After recovery, standard operations such as *decimation* can be applied to return to conventional audio (PCM) rates, *e.g.* 44–192 kHz.

IV. Robust, Scalable Audio Recovery Algorithm

Section III provides a reconstruction method with theoretical guarantees. At its core, the approach relies on the on-grid assumption $\varepsilon_g[n] \in 2\lambda\mathbb{Z}$, which may not hold exactly in practice due to hardware imperfections. Such model mismatch can degrade recovery accuracy and impact perceived audio quality under USF-Audio. Key practical challenges include:

- ❏ **Measurement Noise.** Noise arises in the acquisition chain, entailing (i) thermal electronic noise (approximately Gaussian) and (ii) quantization noise as in Section III.
- ❏ **Hardware Imperfections.** Analog folding can be non-ideal, yielding $\varepsilon_g[n] \notin 2\lambda\mathbb{Z}$. This introduces errors in the non-linear filtering and propagates through recovery.
- ❏ **Data Volume.** Audio sampled at 44–192 kHz yields million-scale datasets ($N \sim 10^6$), imposing stringent scalability and stability requirements on the recovery algorithm.

These challenges motivate USF-ARA (robust USF-Audio algorithm), enabling practical, high-performance audio sensing. A fully detailed derivation of the USF-ARA is provided in the Supplementary Material that accompanies this paper.

Overview of the Solution. Let $\{y_{q,w}[n]\}_{n=0}^{N-1}$ denote the noisy, folded data. To handle large-scale audio, we adopt a *frame-wise processing strategy*: partition the sequence into short segments, $\{y_{q,w}^l[n]\}_{n\in\mathbb{I}_{N_g}}^{l\in\mathbb{I}_L}$, $\{N_g, L\}$ is the frame length and count. Within each segment, we recover ε_g^l, leveraging

[6] $T_{\max}^{(2)}(h) = \frac{1}{\Omega}\left(\frac{\lambda}{2\|g\|_{\mathbf{L}_\infty}}\right)^{1/h}$ and $T_{\max}^{(3)}(h) = \frac{1}{\Omega_{\mathsf{eff},h}}\left(\sqrt{\frac{\pi}{\Omega}}\frac{\lambda}{2\|g\|_{\mathbf{L}_2}}\right)^{1/h}$.

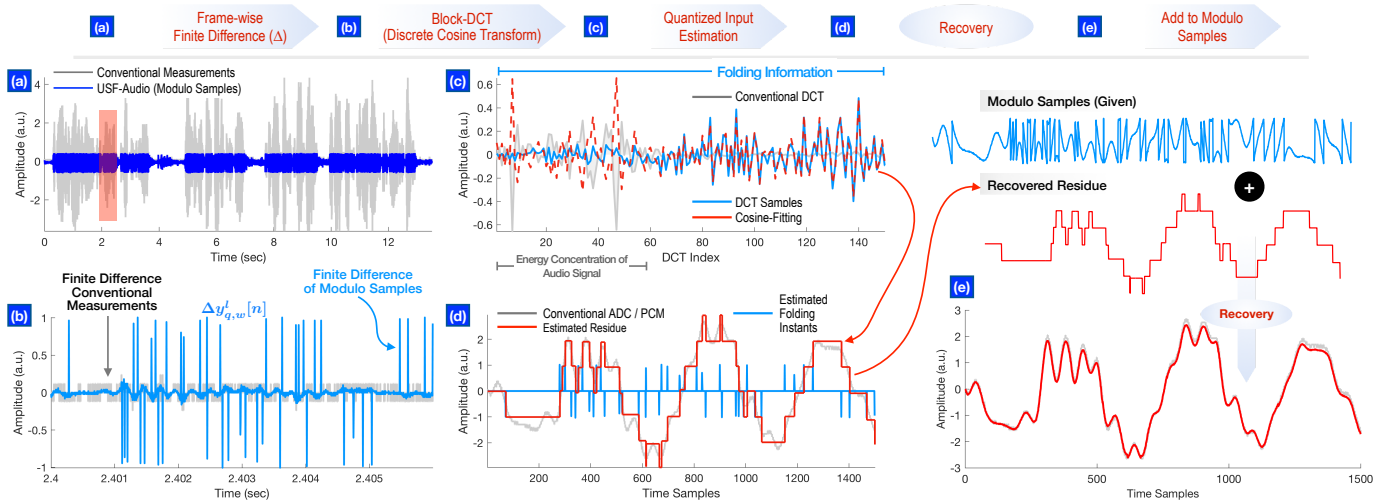


Fig. 3: Visual breakdown of USF-ARA using the dataset Kennedy for illustration. (a) The USF-Audio measurements contain $N = 3.37 \times 10^6$ samples. (b) We partition the USF-Audio measurements $y^l_{q,w}$ into frames of size $N_g = 500$ samples and apply finite differences, revealing a "smooth" + "sparse" mixture model. (c) Applying the DCT to the waveform in (b) shows that audio sequences g^l are concentrated in the lower-indexed DCT coefficients (grey ink), while the residue (sparse components) appears in higher-index coefficients (index 60 onwards). Since spikes in the time domain map to cosines in the DCT domain, spike locations can be retrieved via frequency estimation. (d) From step (c), the residue is reconstructed using spike parameters and an anti-difference operation. (e) Adding the recovered residue to the USF-Audio measurements yields the input signal.

its sparsity for robustness against measurement distortions. Concatenating the recovered residues yields ε_g, after which we reconstruct the audio signal via $\widetilde{g} \leftarrow y_{q,w} + \varepsilon_g$ (see Fig. 3).

Similar to (6), the noisy measurements can be written as $y^l_{q,w}[n] = g^l[n] - \varepsilon^l_g[n] + q^l[n] + w^l[n]$, where w^l models measurement distortions (thermal noise and folding non-idealities). Because the compound noise $q^l + w^l$ complicates recovery, we adopt a transform-domain approach that sparsifies g^l and enables direct residue extraction via spectral partitioning. Unlike the DFT (Discrete Fourier Transform) used in [32], we employ the DCT (Discrete Cosine Transform) for its superior energy compaction under finite-length truncation. Let $\underline{y}^l = \Delta^{(1)} y^l$ and denote its DCT by $\widehat{\underline{y}}^l[k]$, defined by,

$$\widehat{\underline{y}}^l[k] = \sum\nolimits_{n=0}^{N_g-2} \underline{y}^l[n] \cos\left[\tfrac{\pi}{N_g-1}\left(n+\tfrac{1}{2}\right)k\right], \quad k \in \mathbb{I}_{N_g-1}, \tag{13}$$

and we use the same convention for $\{y^l_{q,w}, g^l, \varepsilon^l_g, q^l, w^l\}$. In the DCT domain, we can derive that,

$$\widehat{\underline{y}}^l_{q,w}[k] = \widehat{\underline{g}}^l[k] - \widehat{\underline{\varepsilon}}^l_g[k] + \widehat{\underline{q}}^l[k] + \widehat{\underline{w}}^l[k], \quad k \in \mathbb{I}_{N_g-1} \tag{14}$$

where: i) $\widehat{\underline{g}}^l[k] \approx 0$ for $k \geqslant K$ (compact DCT support), ii) $\widehat{\underline{\varepsilon}}^l_g[k]$ is a sum of $2M$ sinusoids spread across the DCT band since ε_g is a sum of M spikes in time domain with $\varepsilon_g[n] = \sum_{m=0}^{M-1} c_m \delta[n - n_m]$, and iii) $\widehat{\underline{q}}^l$ and $\widehat{\underline{w}}^l$ are noise. This support separation allows direct access to $\widehat{\underline{\varepsilon}}^l_g$ from $\widehat{\underline{y}}^l_{q,w}$ in the out-of-band region $k \in [K, N_g - 2]$,

$$\widehat{\underline{y}}^l_{q,w}[k] = -\widehat{\underline{\varepsilon}}^l_g[k] + \widehat{\underline{q}}^l[k] + \widehat{\underline{w}}^l[k], \quad k \in [K, N_g - 2]. \tag{15}$$

Recovering the residue thus reduces to frequency estimation,

$$\min_{\{\mathbf{c}_l, \mathbf{n}_l\}} \sum_{k=K}^{N_g-2} \left| \widehat{\underline{y}}^l_{q,w}[k] + \sum\nolimits_{m=0}^{M^l_w - 1} c_{m,l} e^{-\jmath \frac{2k n_{m,l} \pi}{N_g - 1}} \right|^2 \tag{16}$$

where $\{\mathbf{c}_l, \mathbf{n}_l\}$ is the vector form of $c_{m,l}$ and $n_{m,l}$, respectively, parametrizing both the generalized residue and sparse outliers induced by w^l. Problem (16) has been extensively studied, yielding robust, high-precision estimators [42]. We next detail the residue estimation procedure.

Robust Residue Estimation. Accurate recovery hinges on robust per-frame residue estimation, as errors accumulate during frame concatenation. To this end, we adopt a model-fitting approach [43]: the out-of-band DCT component $\{\widehat{\underline{y}}^l[k]\}_K^{N_g-2}$,

$$f^l[i] = \sum_{k=0}^{N_1-1} \widehat{\underline{y}}^l[k+K]\, \xi_{N_1}^{k \cdot i} = \frac{\mathsf{Q}(\xi^i_{N_1})}{\mathsf{G}(\xi^i_{N_1})}, \quad \begin{matrix} \mathsf{Q} \in P_{2M-1} \\ \mathsf{G} \in P_{2M} \end{matrix} \tag{17}$$

where $\xi^n_N = e^{\jmath \frac{2\pi n}{N}}$ and the roots of $\mathsf{G}(z)$ uniquely determine the folding time stamps n_m. With this ratio-structure model, problem (16) can be equivalently reformulated as,

$$\min_{\{\mathsf{Q},\mathsf{G}\}} \sum_{i=0}^{N_1-1} \left| f^l_{q,w}[i] - \frac{\mathsf{Q}(\xi^i_{N_1})}{\mathsf{G}(\xi^i_{N_1})} \right|^2, \quad \begin{matrix} \mathsf{Q} \in P_{M^l_w - 1} \\ \mathsf{G} \in P_{M^l_w} \end{matrix}$$
$$\text{where } f^l_{q,w}[i] = \sum\nolimits_{k=0}^{N_1-1} \widehat{\underline{y}}^l_{q,w}[k+K]\, \xi^{k \cdot i}_{N_1}. \tag{18}$$

To tackle the non-convexity of (18), we use an iterative linearization and solve the weighted least-squares surrogate,

$$\left\{ \begin{matrix} \mathsf{Q}^{[j+1]} \\ \mathsf{G}^{[j+1]} \end{matrix} \right\} = \min_{\{\mathsf{Q},\mathsf{G}\}} \sum_{i=0}^{N_1-1} \left| \frac{f^l_{q,w}[i]\, \mathsf{G}(\xi^i_{N_1}) - \mathsf{Q}(\xi^i_{N_1})}{\mathsf{G}^{[j]}(\xi^i_{N_1})} \right|^2 \tag{19}$$

which provides an accurate solution provided that $\mathsf{G}^{[j]} \approx \mathsf{G}$. After $j_{\max}$ iterations, we retain the iterate that achieves the minimum MSE in (18).

Here, we provide the algorithmic implementation for solving (19): the polynomials $\{\mathsf{Q}(\xi^i_{N_1}), \mathsf{G}(\xi^i_{N_1})\}$ can be written in vector-matrix form as,

$$\left[\mathsf{Q}(\xi^i_{N_1})\right] = \mathbf{V}^{M^l_w}_{N_1} \mathbf{q}, \quad \left[\mathsf{G}(\xi^i_{N_1})\right] = \mathbf{V}^{M^l_w+1}_{N_1} \mathbf{g} \tag{20}$$

where $\mathbf{q} \in \mathbb{C}^{M^l_w}$ and $\mathbf{g} \in \mathbb{C}^{M^l_w+1}$ are the coefficients of Q and G, respectively. Having known the estimate $\mathbf{g}^{[j]}$ at $j+1$-th

Algorithm 1 USF-Audio Recovery Algorithm: USF-ARA

Input: Noisy folded measurements $\{y_{q,w}[n]\}_{n=0}^{N-1}$.

$\underline{\varepsilon}_g^0 = \emptyset$.
for $l = 1$ to L **do**
  Compute $\{\widehat{\underline{y}}_{q,w}^l[k]\}_K^{N_g-2}$ via (13).
  Estimate M_w^l via (25).
  for $i = 1$ to $i_{\max}$ **do**
    Random Initialize $\mathbf{g}^{[0]}$.
    for $j = 1$ to $j_{\max}$ **do**
      Compute $\{f_{q,w}^l[i]\}_{i=0}^{N_1-1}$ via (17).
      Construct $\{\mathbf{G}^{[j]}, \mathbf{Q}^{[j]}\}$ via (22).
      Estimate $\{\mathsf{Q}^{[j+1]}, \mathsf{G}^{[j+1]}\}$ by solving (23).
    end for
  end for
  Select $\{\mathsf{Q}, \mathsf{G}\}$ that minimizes the MSE in (18).
  Compute $\{c_{m,l}, n_{m,l}\}_{m=0}^{M_w^l-1}$ via (24).
  Recover $\{\underline{\varepsilon_g}^l[n]\}_{n=0}^{N_g-2}$ with $\{c_{m,l}, n_{m,l}\}_{m=0}^{M_w^l-1}$.
  Concatenate $\underline{\varepsilon}_g$ via $\underline{\varepsilon}_g = [\underline{\varepsilon_g}^{l-1}; \underline{\varepsilon_g}^l]$.
end for
Reconstruct ε_g from $\underline{\varepsilon}_g$ via anti-difference operation.
Recover g via $\widetilde{g} \leftarrow y_{q,w} + \varepsilon_g$.

Output: The recovered audio signal $\widetilde{g}$.

iteration, the minimization in (19) is written as,

$$\{\mathbf{q}^{[j+1]}, \mathbf{g}^{[j+1]}\} = \arg\min_{\mathbf{q},\mathbf{g}} \|\mathbf{G}^{[j]}\mathbf{g} - \mathbf{Q}^{[j]}\mathbf{q}\|_{\ell_2}^2 \quad (21)$$
$$\mathbf{G}^{[j]} \in \mathbb{C}^{N_1 \times (M_w^l+1)} \text{ and } \mathbf{Q}^{[j]} \in \mathbb{C}^{N_1 \times M_w^l}$$

where,

$$\mathbf{G}^{[j]} = \mathbf{D}\mathbf{R}^{[j]}\mathbf{V}_{N_1}^{M_w^l+1}, \quad \mathbf{Q}^{[j]} = \mathbf{R}^{[j]}\mathbf{V}_{N_1}^{M_w^l}$$
$$\mathbf{D} = \mathscr{D}_K(\mathbf{f}_{q,w}^l),\ \mathbf{R}^{[j]} = (\mathscr{D}_K(\mathbf{V}_{N_1}^{M_w^l+1}\mathbf{g}^{[j]}))^{-1}. \quad (22)$$

We use the linear constraint $\mathbf{g}^{\mathsf{H}}\mathbf{g}^{[0]} = 1$ to ensure the uniqueness of the solution to (21), where $\mathbf{g}^{[0]}$ is the initialization seed of our USF-ARA algorithm. As a result, (19) can be posed as,

$$\begin{Bmatrix}\mathbf{q}^{[j+1]}\\ \mathbf{g}^{[j+1]}\end{Bmatrix} = \arg\min_{\{\mathbf{q},\mathbf{g}\}} \left\|\mathbf{A}^{[j]}\mathbf{z}\right\|^2,\ \text{s.t. } \mathbf{z}_0^{\mathsf{H}}\mathbf{z} = 1$$
$$\mathbf{A}^{[j]} = \begin{bmatrix}\mathbf{G}^{[j]} & -\mathbf{Q}^{[j]}\end{bmatrix},\ \mathbf{z} = \begin{bmatrix}\mathbf{g}\\ \mathbf{q}\end{bmatrix},\ \mathbf{z}_0 = \begin{bmatrix}\mathbf{g}^{[0]}\\ \mathbf{0}\end{bmatrix} \quad (23)$$

which is a convex problem and can be solved efficiently. With $\{\mathsf{Q}, \mathsf{G}\}$ known, the generalized residue parameters can be retrieved by finding the roots of $\mathsf{G}(z)$,

$$\begin{matrix}\mathsf{G}(z_{m,l}) = 0\\ m \in \mathbb{I}_{M_w^l}\end{matrix} \implies \begin{matrix}c_{m,l} = \frac{-N_1 z_{m,l}^{-1}\mathsf{Q}(z_{m,l})}{\left(1-z_{m,l}^{-N_1}\right)\partial_z \mathsf{G}(z)|_{z=z_{m,l}}},\\ n_{m,l} = \mathscr{Q}_{\frac{1}{2}}\left(\frac{N_1 \mathsf{Im}(\log z_{m,l})}{2\pi}\right)\end{matrix}. \quad (24)$$

To address the potential discontinuities arising from signal segmentation, we leverage two key strategies, i) DCT and ii) sparse representation of the residue (18). This mitigates the boundary issue when recovering residue in each frame. The recovery precision and algorithm run-time can be further improved by employing windowing methods, such as overlap-add and cross-fading (see Section V-F).

Model Order Estimation. The model order M_w^l can be estimated from the out-of-band DCT sequence $\{\widehat{\underline{y}}^l[k]\}_K^{N_g-2}$ using information-theoretic criteria. A simple subspace-based alternative is to form the Toeplitz matrix from $\{\widehat{\underline{y}}^l[k]\}_K^{N_g-2}$ and estimate M_w^l as its numerical rank. However, this requires performing L singular value decompositions of matrices of size $\lfloor N_1/2 \rfloor \times \lceil N_1/2 \rceil$. In typical audio settings, that translates to roughly 2000 matrix factorizations, which burdens real-time sensing. To this end, in this paper, we employ a simple criterion by evaluating the sample variation of $\underline{y}_{q,w}^l$,

$$M_w^l = |\mathcal{M}_g|,\ \ \mathcal{M}_g = \left\{n \in \mathbb{I}_{N_g-1} \ \middle|\ \left|\underline{y}_{q,w}^l[n]\right| > \epsilon\right\} \quad (25)$$

where ϵ is the tolerance reflecting noise level[7].

Error Mitigation Strategy. With quantization, Theorem 2 guarantees stable reconstruction of g, which is corroborated numerically in Section V-A. To improve the robustness, we leverage two strategies, i) enforcing the on-grid constraint on spike locations through (24), and ii) mildly overestimating the model order M_w^l by setting $\epsilon = 0.5\lambda$. The latter expands the feasible solution space, thereby providing greater resilience to measurement noise and distortion. Experimental results suggest that the USF-Audio recovery is robust against perturbations in ϵ. This resilience is due to the fact that measurement distortion is primarily driven by non-ideal folding—manifesting as sparse outliers—while the thermal noise floor of the MADC remains negligible, as evident in Section V-B. The algorithmic procedure is summarized in Fig. 3 and Algorithm 1[8].

V. Experimental Validation and Benchmarking

Overview. The goal of this section is to validate USF-Audio as a practical acquisition pipeline and to assess the performance gains predicted by Theorem 2, namely improvements in effective resolution (Remark 1), DRX (Remark 2), and measurement precision (Remark 3). These gains are of practical interest because, in conventional ADCs, DR and DRes remain tightly coupled, so that the usable precision of the acquisition chain is often lower than the nominal quantization bit-depth. This motivates the use of the *effective number of bits* (ENOB, see [14] and Section V-B) as a practical resolution metric.

Our evaluation is organized in three stages. We begin with theorem-level validation in Section V-A, where *computational experiments* verify the trends predicted by Theorem 2 (at different bit-depths). We then turn to acquisition-level *hardware (HW) validation* in Section V-B, where controlled experiments quantify the *ENOB*, *SINAD*, and *noise-floor gains* of USF-Audio relative to conventional ADCs under the same nominal bit budget. Building on these results, we evaluate application-level performance on real audio in Section V-C and Section V-D, showing robustness in *severe clipping regimes* together with marked objective and perceptual quality improvements. Finally, Section V-F examines the *computational scalability* of the proposed frame-wise pipeline on large-scale datasets.

[7]We use $\epsilon = 0.5\lambda$ in this paper, which empirically works well in practice.
[8]We use $i_{\max} = 3$ and $j_{\max} = 15$ in this paper.

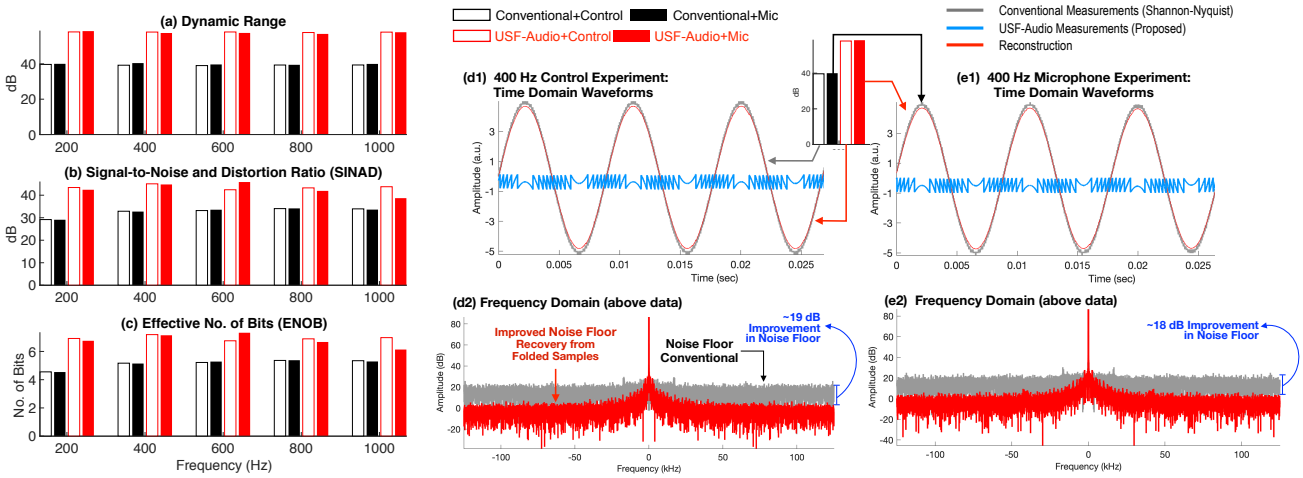


Fig. 4: Hardware evaluation on dynamic range and the effective number of bits. Controlled experiments with single-frequency audio input are performed to demonstrate the significant improvement in signal DR, SINAD, and noise floor, compared to the conventional audio sensing pipeline.

Newly Acquired Hardware Corpus. This work presents, to our knowledge, the first experimental study acquiring a diverse corpus of real-world audio end-to-end using a *HW modulo ADC* [32], [48]. All datasets are captured via the USF-Audio front-end, not by digitally folding or clipping conventional ADC outputs—ensuring reported gains reflect true acquisition-level performance under real analog non-idealities. Initial demos, audio playback and datasets are available at https://alumni.media.mit.edu/~ayush/USFAudio.

Acquisition Pipeline. The setup is shown in Fig. 2. Using off-the-shelf components, we design a novel modulo ADC, MADC (also see [32], [48]), the analog front-end of USF-Audio. A MEMS microphone SPH8878LR5H-1[9] captures Beoplay P2 speaker[10] output and feeds the analog signal to both a MADC (USF-Audio measurements) and conventional ADC (reference), under the *identical nominal quantization bit budget*. Unless stated otherwise, signals are sampled at 250 kHz with 6 bits per sample[11] on the PicoScope 3406D oscilloscope. The reference ADC input range is ± 6.5 V, while USF-Audio measurements lie in $[0, 1]$ V. For digital audio datasets (BLS and DIGI in Table III), high-precision ground truth ($\geqslant 32$ bits) are available; for the HW data, we use the reference as a proxy to the ground truth, as it records the same analog output of the microphone as USF-Audio measurements.

Evaluation Metrics. (1) ADC dynamic range (DR) is defined as $\mathsf{DR} = 20\log_{10}(g_{\max}/\text{noise floor})$ where $g_{\max} = \|g\|_{\mathbf{L}_\infty}$. (2) Reconstruction accuracy is quantified by RMSE (root mean-squared error) against the (i) high-precision ground truth for BLS and DIGI audio datasets (see Table III) and (ii) reference for the rest datasets. (3) We classify an audio frame recovery $\widetilde{g}^l$ of g^l as a failure case if $\max_n |\widetilde{g}^l[n] - g^l[n]| > 2\lambda$, indicating the issue of extra/missing folds in unfolding. (4) Accordingly, the frame-wise RMSE is calculated as $\mathsf{RMSE}_l = \sqrt{\sum_{n=0}^{N_g-1} |\widetilde{g}^l[n] - g^l[n]|^2/N_g}$. (5) The performance of MADC and conventional ADC is quantitatively assessed in terms of SINAD (signal-to-noise-and-distortion ratio), ENOB (effective number of bits), THD+N (total harmonic distortion plus noise). (6) The performance statistics of USF-Audio method is characterized by (i) ratio of failure frames, (ii) 95th percentile of RMSE_l and (iii) worst case recovery $\max_l \mathsf{RMSE}_l$. (7) The algorithm run-time is evaluated on a 2020 MacBook Pro (8-core CPU, 8GB RAM). Datasets, sources, run-time, no. of failure frames, high-percentile error metrics, and experimental configurations are summarized in Tables II—VI.

[9]Manufacturer Details: Syntiant Corp, Irvine, California.
[10]Manufacturer Details: Bang & Olufsen A/S, Struer, Midtjylland.
[11] Each sample is uniformly quantized into $\mathcal{L}$ levels, with bit-depth $\log_2 \mathcal{L}$.

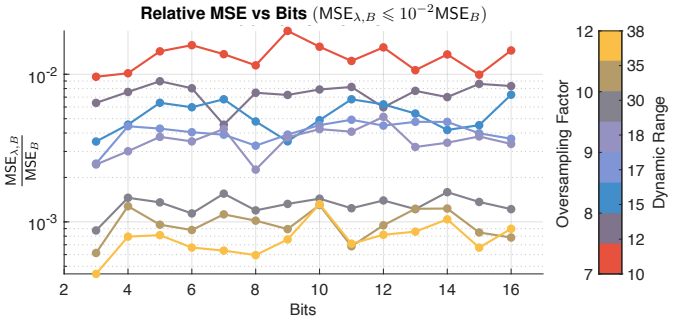


Fig. 5: Relative MSE versus quantization bit-depth. We evaluate the relative MSE normalized by the baseline MSE obtained from the ADC at varying bit-depth and *oversampling factor (OF)*. Across all assessed scenarios, USF-Audio consistently delivers a 10^2-to-10^3-fold performance gain in MSE, which effectively validates the OF-MSE law derived in Theorem 2.

A. Oversampling–MSE Scaling: Constant-Factor Gain

Theorem 2 and (12) show that USF-Audio outperforms ADCs as $\mathsf{MSE}_{\lambda,B} = (\lambda/g_{\max})^2\,\mathsf{MSE}_B$. To verify this result, we conduct the following numerical experiments: a random audio signal is generated with maximal frequency of 10 kHz, which is sampled with OF (oversampling factor) linearly spanning from 7 to 12 and DRX from 10 to 38-fold. Using the identical bit budget (3-16 bits), we evaluate relative MSE, *i.e.*, $\mathsf{MSE}_{\lambda,B}/\mathsf{MSE}_B$, after rectangular low-pass filtering (cut-off frequency of 10 kHz), as a function of bit-depth (B) and OF. This normalization underpins bit-scaling pattern of USF-Audio, exposing the excess error induced by folding and recovery.

TABLE II: Comparative Experiments: Parameters and Performance Metrics.

Frequency (Hz)	$\|g\|_\infty$	DR (dB)		SINAD (dB)		ENOB		THD+N (%)		RMSE ($\times 10^{-2}$)
		ADC	USF-Audio	ADC	USF-Audio	ADC	USF-Audio	ADC	USF-Audio	
Controlled Experiment										
200	5.23	39.67	58.06	29.18	43.42	4.55	6.92	3.48	0.67	7.04
400	5.23	39.22	57.98	32.88	45.06	5.17	7.19	2.27	0.56	7.09
600	5.23	39.01	57.86	33.18	47.16	5.22	7.54	2.19	0.44	7.25
800	5.23	39.33	57.72	34.05	43.26	5.36	6.89	1.98	0.69	7.44
1000	5.23	39.36	57.94	33.94	43.79	5.34	6.98	2.01	0.65	7.53
Speaker Experiment										
200	5.17	40.01	58.52	29.10	42.43	4.54	6.76	3.51	0.76	6.94
400	5.17	40.42	57.45	32.73	44.78	5.14	7.15	2.31	0.58	6.89
600	5.37	39.66	57.43	33.56	46.48	5.28	7.43	2.10	0.51	7.16
800	5.17	39.45	56.93	34.16	43.12	5.38	6.87	1.96	0.70	7.38
1000	5.37	39.96	57.92	33.56	45.12	5.28	7.20	2.10	0.81	7.23

As predicted by (12), USF-Audio *theoretically outperforms* ADC with MSE gain ranging from 100- to 1444-fold, consistently matching the $[10^2, 10^3]$-fold MSE reduction, as shown in Fig. 5. These numerical experiments directly validate the performance bounds derived in Theorem 2 and (12), offering a quantitative support for the resolution and noise-floor gains observed in previous subsections.

B. Effective Number of Bits (ENOB): Resolution Advantage

A defining feature of USF-Audio lies in resolution enhancement at acquisition level (Layer 2, Table I), which redistributes amplitude information in the analog domain via modulo folding and reveals fine-scale signal structure even under low quantization bit-depth. In contrast, post-hoc enhancement approaches, such as dithering [22]–[24], noise shaping and oversampling [8], operate at processing level (Layer 4, Table I) and rely on signal priors [54], [55], where *irreversible information loss* has already occurred at acquisition front-end. To quantify this resolution enhancement, we resort to ENOB[12].

Experimental Design. To isolate the source of resolution gain, we design two complementary HW experiments:

1) *Controlled Experiments*. Single-tone generated by a waveform generator, sampled directly by a MADC, bypassing the microphone. This isolates the acquisition pipeline and eliminates acoustic and transducer non-idealities.
2) *Speaker-microphone Experiments*. Single-tone signals are played through a loudspeaker and recorded using a microphone followed by the MADC, capturing realistic analog distortions and front-end non-idealities.

In each case, 5 tones ranging from 200 to 1000 Hz are tested with $N = 8918$, $N_g = 500$ and $\lambda = 0.47$. In parallel, an ADC with the same nominal bit-depth (6 bits) records reference measurements. Under this setting, the corresponding DRes of the ADC and USF-Audio are 0.20 and 0.02, respectively. Experimental parameters are summarized in Table II.

Beyond MSE, we report the signal-to-noise-and-distortion ratio (SINAD), defined as the ratio of signal power to the combined noise and distortion power. ENOB is computed as

[12]As non-idealities such as quantization noise, distortion, and limited DR reduce the number of usable bits, the nominal resolution of an ADC does not fully characterize the digital precision, motivating the use of ENOB as a practical metric of system resolution [14].

TABLE III: USF-Audio Dataset for Audio Declipping Experiments.

Dataset	Description and Source	Time (sec)
Numerical Experiments		
BLS-1	Bandlimited signal realization generated by low-pass filtering a white Gaussian noise signal.	0.03
BLS-2		0.08
BLS-3		0.16
DIGI-1	Music from the musician Yuri Korzunov (32-bit,192 kHz) • https://samplerateconverter.com/free-audio-downloads	12
DIGI-2		12
DIGI-3		12
DIGI-4		12
DIGI-5		12
Hardware Experiments		
MPA Waveform	Lab Experiment	0.05
Für Elise	**Für Elise** (or *Bagatelle No. 25* in A minor) by Ludwig van Beethoven, performed live on piano by Marnie Laird. https://youtu.be/2fcX2dWmR6g	0.05
Canon-Gromov-C	Mixed Audio: Canon + Gromov interview	9
AWM-C	Song **Always With Me** from the movie **Spirited Away** (2001) • https://youtu.be/tBSthP5LTZU	10
If-Kipling-3	Same as above but with signal clipping https://youtu.be/sqOgyNfHI1U	12
If-TWM-C	Poem: **If Tomorrow Starts without Me** read by Tom O'Bedlam • https://youtu.be/8YMJclvpUlc	11
Heart-MM2	Apex Area **S4** and **Mid-Systolic Murmur** from Heart Sound & Murmur Library • https://shorturl.at/h6kFI	9

$\text{ENOB} = (\text{SINAD} - 1.76)/6.02$ [14] and is used to quantify resolution improvement over the ADCs. Results in Table II and Fig. 4 demonstrate that USF-Audio achieves considerable improvement, at least[13]:

Dynamic Range	SINAD	ENOB
18 dB	14 dB	1.5 bits
Fig. 4 (a)	Fig. 4 (b)	Fig. 4 (c)

The ENOB gains do not arise from *dithering*, *noise shaping*, or *stochastic averaging*; instead, they reflect advantages rooted in acquisition-induced structure capture. This confirms the acquisition-level principle of USF-Audio; resolution can be reliably enhanced without increasing bit budget. This property underpins the subsequent advantages in real-audio recovery, perceptual quality, and clipping avoidance.

C. Eliminating Saturation as a Failure Mode

Declipping is a longstanding challenge that has driven extensive algorithmic research [12], [13], [56], [57], which aim to recover from clipped data using signal models, sparsity priors, or perceptual heuristics. Despite these advances, SOTA methods exhibit a well-documented performance ceiling under severe clipping [12], [13]. From the APS perspective (Table I), this limitation is structural rather than algorithmic: *clipping is inherently non-invertible*, and once saturation occurs, the information needed for recovery is lost. Conversely, USF-Audio eliminates clipping by injecting analog-domain folding prior to digitization, ensuring the recovery problem remains well-posed in HDR conditions. To highlight this distinction, we design two experiments focused on the clipping regime and

[13]These gains persist in the presence of microphone distortion and analog non-idealities, indicating that the observed resolution improvement is not an artifact of controlled laboratory conditions.

TABLE IV: Audio Declipping: Datasets, Experimental Parameters, and Performance Metrics. RMSE statistics are reported for the recovered USF-Audio signal.

Dataset	$\|g\|_\infty$	Digital Resolution		Noise Floor (dB)		λ_{clip}	N	Input SDR (dB)	ΔSDR (dB)	RMSE Statistics			Run Time	No. Failure Frames / No. Frames
		ADC	USF-Audio	ADC	USF-Audio					Mean ($\times 10^{-2}$)	95th Percentile ($\times 10^{-2}$)	$\max_l$ ($\times 10^{-1}$)	(sec)	
BLS-1	10.20	6.23×10^{-5}	6.10×10^{-5}	-23.45	-52.11	0.51	20040	0.95	106.09	1.80×10^{-3}	1.78×10^{-3}	1.86×10^{-4}	1	0/40
BLS-2	10.58	6.41×10^{-5}	6.10×10^{-5}	-17.75	-47.68	0.53	50040	1.17	104.28	1.85×10^{-3}	1.85×10^{-3}	1.92×10^{-4}	1	0/100
BLS-3	10.17	6.16×10^{-5}	6.10×10^{-5}	-24.01	-44.91	0.51	100040	1.24	103.84	1.78×10^{-3}	1.76×10^{-3}	1.87×10^{-4}	1	0/200
DIGI-1	10.00	5.96×10^{-8}	5.96×10^{-8}	15.47	-6.46	0.50	2304000	1.86	159.89	1.72×10^{-6}	1.73×10^{-6}	1.86×10^{-7}	3	0/4608
DIGI-2	10.00	5.96×10^{-8}	5.96×10^{-8}	15.40	9.53	0.50	2304000	1.70	160.88	1.72×10^{-6}	1.67×10^{-6}	1.84×10^{-7}	3	0/4608
DIGI-3	10.00	5.96×10^{-8}	5.96×10^{-8}	14.72	-1.47	0.50	2304000	2.10	159.00	1.72×10^{-6}	1.75×10^{-6}	1.83×10^{-7}	3	0/4608
DIGI-4	10.00	5.96×10^{-8}	5.96×10^{-8}	13.08	-9.64	0.50	2304000	2.31	158.14	1.72×10^{-6}	1.70×10^{-6}	1.86×10^{-7}	3	0/4608
DIGI-5	10.00	5.96×10^{-8}	5.96×10^{-8}	-1.70	-7.45	0.50	2304000	2.21	158.00	1.72×10^{-6}	1.72×10^{-6}	1.96×10^{-7}	3	0/4608
Für Elise	4.13	0.01	0.01	9.09	-0.42	1.08	15615	5.17	32.17	4.30	2.17	3.75	5	0/31
MPA Waveform	3.11	0.01	0.01	15.46	0.14	0.79	31240	5.94	35.29	2.02	2.09	0.61	5	0/62
If-Kipling-3	5.32	4.80×10^{-3}	4.80×10^{-3}	41.60	16.41	0.53	2908453	4.70	27.16	3.42	3.25	4.61	193	0/5816
If-TWM-C	6.63	0.03	0.02	41.45	23.39	0.70	2833230	2.61	33.16	3.27	3.91	1.92	422	0/5666
Heart-MM2	6.10	0.01	0.01	40.20	13.14	0.73	2015435	2.56	35.68	2.99	2.91	0.79	64	0/4030
MixCannon-Gromov-C	4.36	0.03	0.02	39.16	23.47	0.66	2318684	4.36	27.28	3.59	3.72	3.09	159	0/4637
AWM-C	6.73	0.03	0.02	40.97	23.18	0.70	2506310	2.08	32.12	3.52	3.04	2.02	790	0/5012

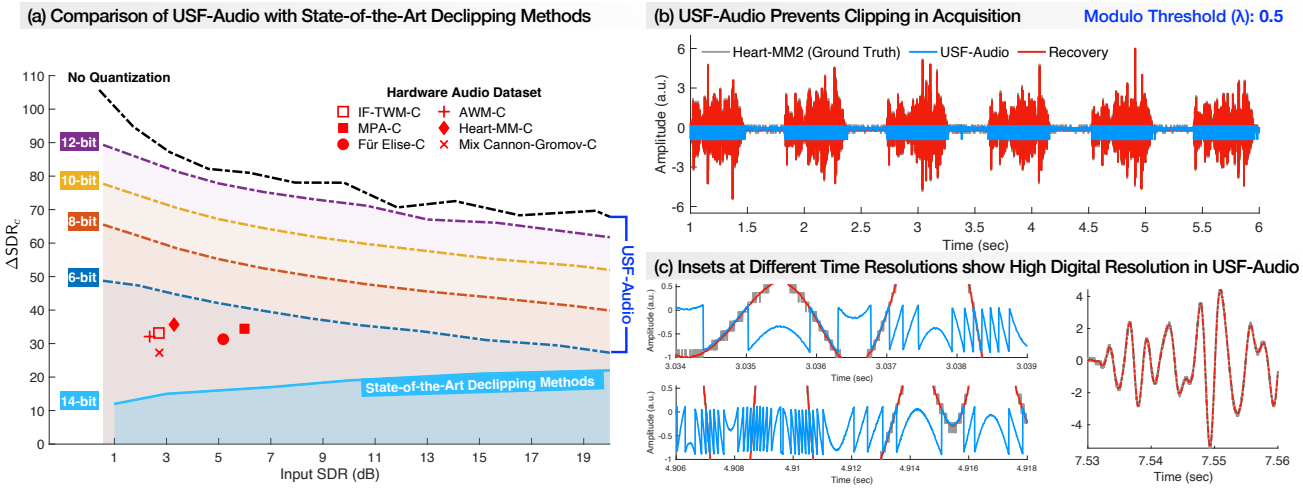


Fig. 6: USF-Audio fundamentally avoids audio clipping while simultaneously offering higher digital resolution. (a) Quantitative comparison of declipping methods quantified in terms of signal-to-distortion ratio (SDR) (see [12]). State-of-the-art declipping methods [12], [13] show their restriction to a 20 dB ceiling. For extreme clipping scenarios (input SDR: 1 to 3 dB), the improvement is at best 11 dB. By contrast, numerical experiments, based on the USF-Audio framework with different digital resolutions, consistently delivers at least a 30 dB improvement. Hardware experiments with 6 datasets with ≈ 6 bit resolution confirm our findings with numerical experiments. (b) High Dynamic Range (HDR) recovery via USF-Audio. In this experiment, we specifically use heart murmurs (mid-systolic murmur); such waveforms are naturally HDR [15]. With $\|g\|_{\mathbf{L}_\infty} = 6$ and $\lambda = 0.5$, we demonstrate HDR recovery. (c) Insets from (b) at different time resolutions to provide a visual sense improvement in digital resolution from 0.1977 (6 bit, conventional case) to 0.0051 (7.5 bit, USF-Audio).

compare against SOTA declipping algorithms, using the best reported SDR values among all methods from [12], [58], [59].

Exp-1: Numerical experiments (HDR stress test). We conduct two sets of numerical experiments: (a) Bandlimited signals (BLS in Table III with 192 kHz sampling rate) are generated by low-pass filtering white Gaussian noise and sampled using the same 14 bits for MADC and conventional ADC; (b) Digital audios (DIGI in Table III with 32 bits and 192 kHz sampling rate) are re-sampled using the identical 24 bits for both MADC and conventional ADC. In all cases, $\{N_g, \lambda\} = \{500, 0.5\}$ and reference signals are artificially clipped at threshold λ_{clip} (see Table IV) to emulate conventional saturation. In contrast to Section V-D, reference samples maintain higher digital resolution due to limited DR region $[-\lambda_{\text{clip}}, \lambda_{\text{clip}}]$. The performance metrics such as RMSE and ΔSDR are computed based on the known high-precision ground truth. This enables a direct comparison with SOTA declipping methods [12], [59] under identical input conditions.

Exp-2: Hardware experiments (real-world acquisition). Audio signals are played through a Beoplay P2 loudspeaker and acquired using the MADC with $\lambda = 0.5$. Simultaneously, two conventional ADC channels of identical nominal quantization budget (6 bits) but different input ranges are recorded: (i) a HDR ADC serving as proxy ground truth, and (ii) a low-dynamic-range ADC producing clipped data. Experiments include speech, music, instruments, bioacoustics, and mixed content data digitized via the MADC. We quantify the performance using the signal-to-distortion ratio (SDR), following established declipping benchmarks [12]. Experimental configurations and datasets are summarized in Table III. Results in Table IV and Fig. 6 show:

Numerical Experiments. Under extreme clipping conditions (input SDR $\approx$ 0.95-2.31 dB), USF-Audio achieves an average recovery performance of ΔSDR $= \{104.74, 159.18\}$ dB for $\{\text{BLS}, \text{DIGI}\}$ datasets, near machine precision (see Table IV). This corroborates exact recovery when USF feasibility conditions are satisfied and directly validates the theoretical guarantees underlying the USF-Audio (see Theorem 2). By contrast, SOTA declipping methods exhibit a performance ceiling even at higher bit-depths.

TABLE V: Hardware USF-Audio Corpus.

Dataset	Description and Source	Time (sec)
	Hardware Experiments	
AWM (Music)	Song **Always With Me** from the movie **Spirited Away** (2001) • https://youtu.be/tBSthP5LTZU	10
Canon (Music)	Pachelbel's **Canon** in D major by Marnie Laird and Patrick Laird • https://youtu.be/Ptk_1Dc2iPY	13
Jazz (Music)	Jazz piece "Take Five" by **Dave Brubeck** • https://www.youtube.com/watch?v=vmDDOFXSgAs	12
StarWars (Music)	Star Wars Theme Song (1977) • https://bit.ly/4sUZTHq	8
Kennedy (Speech)	President **Kennedy's Man on the Moon Speech** —May 1961 • https://youtu.be/GhgVZLrxiu0	14
Gromov (Speech)	Science Lives — Interview of **Mikhail Gromov** • https://youtu.be/pzW6GYrdFDg	12
Tasmania (Speech)	Tasmania by **David Attenborough** • https://bit.ly/4dwhZuw	6
Ketterle (Speech)	Wolfgang **Ketterle's Nobel Laureate Lecture** (2001) at MIT • https://youtu.be/NoO7XKVmZC8	7
If-Kipling-1 (Speech)	**If by Rudyard Kipling** — Read by Sir Michael Caine • https://youtu.be/sqOgyNfHl1U	8
If-Kipling-2 (Speech)	**If by Rudyard Kipling** — Read by Sir Michael Caine • https://youtu.be/sqOgyNfHl1U	7
If-Kipling-3 (Speech)	Same as above but with signal clipping • https://youtu.be/sqOgyNfHl1U	12
If-TWM (Speech)	**If Tomorrow Starts without Me** read by Tom O'Bedlam • https://youtu.be/8YMJclvpUlc	11
Whale (Biosounds)	Audiovisual recording of Humpback Whales Singing • https://youtu.be/UOkcvGPHsgk	11
Heart-MM (Heart sound)	Apex Area S3 Gallop from Heart Sound & Murmur Library • https://shorturl.at/h6kFI	9

Hardware Experiments. Across all real-world datasets, the end-to-end USF-Audio pipeline achieves an average performance of $\Delta\text{SDR} = 31.84$ dB using only 6 bits quantization. Under identical clipping conditions, SOTA declipping methods[14] saturate at approximately a 12 dB improvement even with 14 bits resolution. USF-Audio therefore maintains a consistent ≈ 20 dB advantage over SOTA declipping methods under severe clipping scenarios (Fig. 6, Table III). Time- and frequency-domain visualizations in Fig. 6 further reveal that USF-Audio prevents saturation entirely, while preserving fine-scale temporal and spectral structure that is irreversibly lost in conventional measurements due to clipping or low-resolution.

Instead of interpreting these results as superior algorithmic declipping, they actually demonstrate a categorical difference in acquisition philosophy. Declipping algorithms *attempt inversion after a non-invertible operation*; USF-Audio enforces invertibility by design, ensuring that amplitude information is never destroyed. Consequently, the dramatic SDR gains observed here are not attributed to improved reconstruction heuristics, but to the *removal* of clipping itself. In this sense, declipping performance serves as a secondary evidence of a deeper system-level principle: altering acquisition leads to the removal of declipping problems.

D. Objective Audio Quality Assessment

USF-Audio translates to audio quality improvements without content-specific tuning, enabling stable, large-scale recovery of real-world audio under dense folding and HW non-idealities.

Experimental Design. We evaluate USF-Audio on 14 HW datasets spanning speech, music, human, animal sounds, and

[14]When comparing with SOTA declipping methods, we pick the best reported ΔSDR for a given SDR among all methods surveyed in [12], [59].

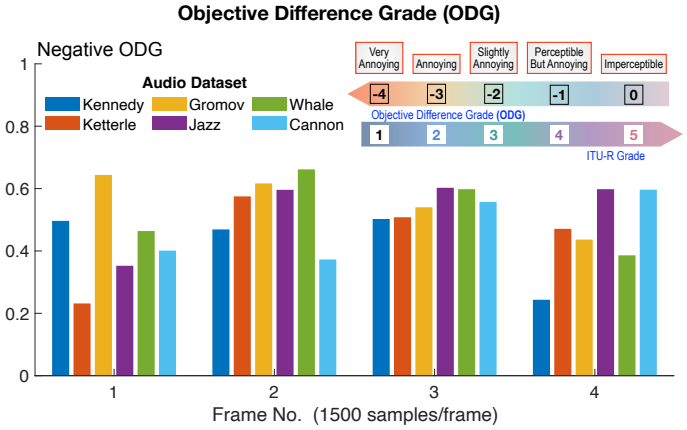


Fig. 7: We perform a frame-wise evaluation of Objective Difference Grade (ODG), a perceptual metric of audio quality, specified by the International Telecommunication Union (ITU). We consider experiments with six different USF-Audio datasets covering different audio genres: Speech (`Kennedy`, `Ketterle` and `Gromov`), Music (`Jazz`, `Canon`) and Oceanic Sounds (`Whale`), respectively. Our reconstruction achieves an average ODG value of -0.4966 which implies that the perceptual difference between ground truth and reconstruction would be "imperceptible" to the human ear.

instruments (Table V), newly acquired using the MADC with $\lambda = 1/2$. Audio durations range from 6-14 s, resulting in up to 10^6 samples, intentionally stressing long-duration recovery and dense folding. Reconstruction is performed using the *frame-wise algorithm* ($N_g = 500$) USF-ARA in Algorithm 1. An ADC pipeline of identical bit budget serves as reference.

Reconstruction accuracy is assessed via RMSE and noise-floor measurements. Objective audio quality is evaluated using the PEAQ metric (ITU-R BS.1387 1999), reported as Objective Difference Grade (ODG) [12]. ODG is evaluated frame-wise to expose potential temporal instability or local degradation. The histogram of frame-wise processing time is shown in Fig. 9.

Across all datasets, see Table VI, USF-Audio yields: 1) an average 12-fold DRX, 2) an average 18 dB improvement in noise floor, and 3) an average 11-fold boost in digital resolution. Frame-wise ODG evaluation over representative datasets yields an average score of -0.4966, lying between *imperceptible* and *perceptible but not annoying* and is closer to the former (Fig. 7). Time-frequency representations in Fig. 8 (a_5) and Fig. 8 (b_5) further underpin preservation of fine spectral structure that is suppressed in ADC measurements (Fig. 8 (a_3) and Fig. 8 (b_3)). Across all tested corpora, the average reconstruction RMSE is at the order of 10^{-2} (see Table VI), demonstrating the scalability and robustness of USF-Audio, despite the presence of dense folding, voluminous data, and HW non-idealities.

E. Perceptual Evaluation and Audio Survey

Study Objective and Scale. We conduct a controlled listening study to test whether USF-Audio improves perceived audio quality under dynamic-range-limited sensing, while remaining perceptually transparent when such limitations are absent. The study includes 52 participants, 10 audio excerpts, and 3 listening conditions: conventional audio (C-Audio), USF-Audio, and clipped audio (CLIP). Each participant evaluated every

TABLE VI: Audio Recovery Benchmarking: Datasets, Parameters, and Performance Metrics. RMSE statistics are reported for the recovered USF-Audio signal.

Dataset	DRX	Digital Resolution		Noise Floor (dB)		Bit Depth		Dynamic Range (dB)		RMSE Statistics			Run Time	No. Failure Frames / No. Frames
	η_q^B/η_λ^B	ADC	USF-Audio	ADC	USF-Audio	ADC	USF-Audio	ADC	USF-Audio	Mean ($\times 10^{-2}$)	95th Percentile ($\times 10^{-2}$)	$\max_l$ ($\times 10^{-1}$)	(sec)	
Kennedy	9.50	0.12	0.02	37.82	25.78	6.09	5.83	42.05	54.43	5.05	3.19	4.09	170	0/6758
Ketterle	6.60	0.08	0.02	32.91	21.19	6.23	5.75	40.89	52.37	2.07	1.22	1.87	31	0/3457
Gromov	14.78	0.20	0.02	41.79	23.56	6.02	6.02	41.35	59.41	3.88	3.17	3.10	402	0/5935
Whale	14.77	0.20	0.02	41.62	23.85	6.13	6.04	41.36	59.08	4.11	2.75	3.65	525	0/5705
Jazz	12.59	0.20	0.02	41.73	23.14	5.70	6.02	40.09	58.65	3.25	3.56	2.59	317	0/6013
Cannon	14.86	0.20	0.02	41.97	23.84	6.13	6.02	41.47	59.46	3.68	3.06	4.64	489	0/6313
If-Kipling-2	14.61	0.20	0.02	39.28	20.53	6.11	6.02	41.39	59.97	3.48	2.70	3.21	159	0/3436
If-Kipling-1	10.19	0.20	0.02	39.93	20.47	5.58	5.98	38.39	57.71	2.87	2.59	2.94	98	0/4116
StarWars	14.61	0.20	0.02	40.01	23.80	6.11	6.04	41.22	57.41	6.72	3.13	4.66	700	0/3918
Tasmania	13.49	0.20	0.02	38.98	20.33	5.98	6.04	40.64	59.14	3.96	3.42	3.44	187	0/3173
If-Kipling	11.35	0.20	4.80×10^{-3}	41.60	16.41	5.75	7.70	39.13	64.34	3.42	3.25	4.61	189	0/5816
Heart-MM	7.95	0.14	0.02	38.21	21.38	5.70	5.73	38.15	54.83	2.17	1.97	0.52	73	0/4298
If-TWM	13.88	0.20	0.02	41.45	23.39	5.91	5.75	40.96	58.97	3.27	3.91	1.92	422	0/5666
AWM	14.33	0.20	0.02	40.97	23.18	6.07	5.78	41.19	58.90	3.52	3.04	2.02	784	0/5012

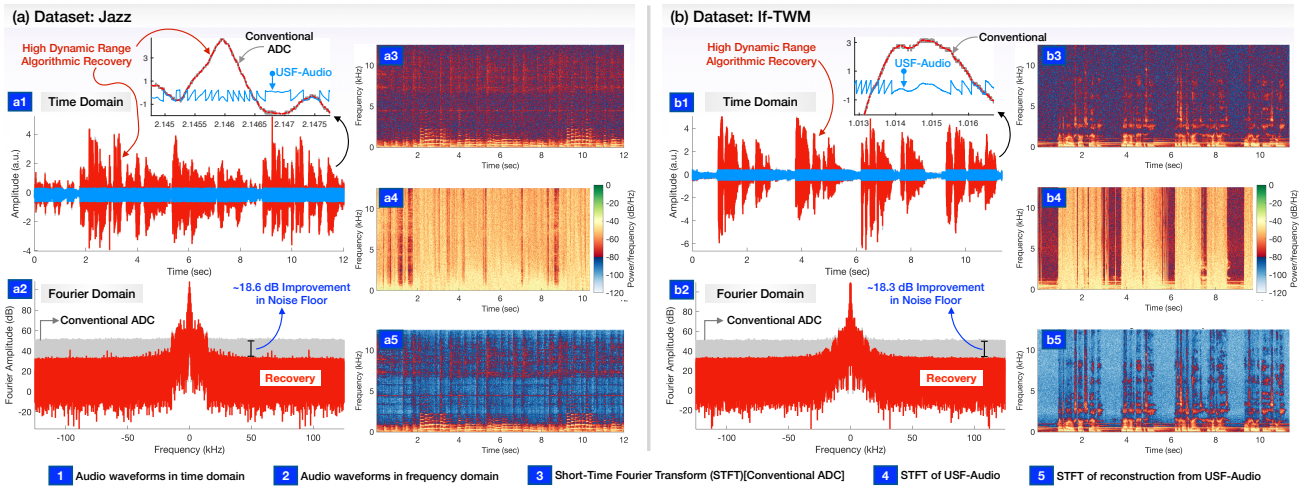


Fig. 8: Demonstration of audio recovery from $N \propto 10^6$ USF-Audio measurements on 2 of the 14 different datasets. (a) Dataset Jazz (musical piece), (b) Dataset If-TWM (combination of vocal and musical contents). Compared to the conventional Shannon-Nyquist framework, we show significant improvement in the noise floor (at least 18 dB), signal dynamic range (approximately 13-fold), and digital resolution of 0.1971 volts (best case) for conventional case while 0.0156 volts with USF-Audio, using the same 6 bit budget. Finer audio details and features are clearly visible in the Short-Time Fourier Transform (STFT) representation shown in (a5) and (b5), respectively. Initial demos and datasets are available at https://alumni.media.mit.edu/~ayush/USFAudio.

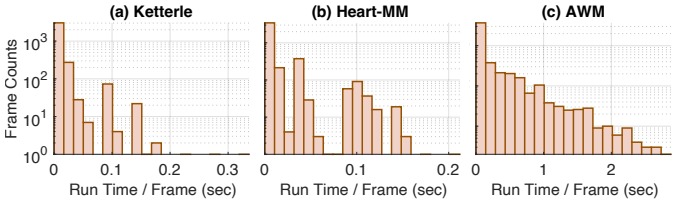


Fig. 9: Histogram of frame-wise processing time of audio corpus (a) Ketterle, (b) Heart-MM and (c) AWM reported in Table VI. The algorithm run-time primarily depends on DRX. Under moderate DRX $= 6.60$, the median run-time per frame is 0.003 sec for (a) Ketterle; while in the more challenging scenarios (DRX $= \{7.95, 14.33\}$ for (b) and (c)), more modulo-foldings occur, and finding unknown parameters $\{\mathbf{c}_l, \mathbf{n}_l\}$ in (16) requires longer processing time (median: $\{0.008, 0.02\}$ sec per frame for (b), (c)).

clip under all conditions, yielding 520 complete participant–clip observations in a balanced within-subject design.

ACR-MOS Protocol and Rationale. Perceptual quality is assessed using the ITU-T P.800 Absolute Category Rating (ACR) protocol: listeners rate each stimulus independently rather than against a reference. Raw web-interface scores $s_{i,k,c} \in [0, 100]$ are mapped to the standard five-point ACR scale via $y_{i,k,c} = 1 + 4\frac{s_{i,k,c}}{100}$, and the Mean Opinion Score (MOS) is the average of the converted scores $y_{i,k,c} \in [1, 5]$ across participants. We use ACR-MOS rather than MUSHRA because the study asks an absolute acquisition-quality question—*how good does each signal sound on its own after sensing and reconstruction*—rather than a reference-based discrimination question. This is particularly important in the low-bit HW regime, where saturation and wrap-around artifacts do not admit a single perceptually valid reference.

Statistical Analysis and Interpretation. The primary analysis is conducted at the participant level within the HW regime, where each participant contributes one mean score per condition averaged over the 5 hardware clips, avoiding pseudo-replication across clips. Pairwise comparisons use two-sided Wilcoxon signed-rank tests with Holm correction, and effect sizes are reported using paired d_z and rank-biserial correlation. As shown in Fig. 10 and summarized in Table VII, USF-Audio yields a clear perceptual improvement over conventional audio in the 6-bit HW regime, where dynamic range is the limiting factor ($\Delta\text{MOS}_{\text{HW}} = +0.592$, $d_z = 1.5275$). Equally importantly, in the high-bit DIGI control, conventional audio is given an 8-bit numerical advantage: C-Audio is represented at 32 bits, whereas USF-Audio is represented at 24 bits. Despite this advantage for the conventional baseline, the two are statistically indistinguishable ($\Delta\text{MOS}_{\text{DIGI}} = +0.011$, $p =$

TABLE VII: Descriptive statistics (MOS) for HW and DIGI regimes. Values are computed over all complete participant–clip observations.

Dataset	Condition	Mean	SD	95% CI (Lo)	95% CI (Hi)
HW	C-AUDIO	3.2155	0.8006	3.1178	3.3133
	USF-AUDIO	3.8078	0.8693	3.7017	3.9140
	CLIP	1.6166	0.5954	1.5439	1.6893
DIGI	C-AUDIO	4.1508	0.6533	4.0710	4.2306
	USF-AUDIO	4.1618	0.6490	4.0826	4.2411
	CLIP	1.4735	0.5621	1.4049	1.5422

0.7489). Thus, USF-Audio improves perceptual quality precisely when conventional acquisition is dynamic-range limited, and remains transparent even against a higher-bit conventional reference when the acquisition regime is already perceptually high fidelity. A complete description of the listening protocol, dataset, and statistical analysis (including per-clip results and robustness checks) is provided in the Supplementary Material.

F. Algorithm Run-Time Analysis

Algorithm run time is mainly governed by DRX, folding non-idealities, and data volume. When folds are ideal, as in the BLS and DIGI datasets, our method recovers million-scale audio signals within a few seconds; see Table IV.

For real-world folded audio, hardware imperfections introduce non-ideal folds, *i.e.*, $\varepsilon_g[n] \notin 2\lambda\mathbb{Z}$, increasing the number of unknowns and hence recovery time. This trend is illustrated by the frame-wise run-time histograms for Ketterle ($\{\text{DRX}, N\} = \{6.60, 1.73 \times 10^6\}$) and AWM ($\{\text{DRX}, N\} = \{14.33, 2.50 \times 10^6\}$): higher DR produces more non-ideal folds per frame and longer processing time; see Fig. 9 and Table VI.

Future work will improve audio recovery speed through both hardware and algorithmic advances: reducing non-ideal folds via improved folding-front-end design, lowering latency through embedded or chip-level implementation, and accelerating frame-wise recovery using optimized windowing and compact transforms such as Discrete Prolate Spheroidal Sequences for more efficient sparse recovery in (16).

VI. CONCLUSIONS AND LOOKING AHEAD

This work positions the Unlimited Sensing Framework (USF) as a fundamental shift in addressing limitations in Audio Signal Processing. Rather than treating clipping, coarse quantization, and restricted dynamic range as distortions to be corrected post-digitization, USF-Audio reframes them as avoidable artifacts of the acquisition model. By redesigning the sensing stage through controlled non-linear acquisition, USF-Audio demonstrates that information loss at the analog–digital interface is not inevitable.

Within USF-Audio, what is traditionally regarded as quantization noise is shown to carry recoverable structure when digitized deliberately. This perspective expands the notion of audio representation: information need not be confined to linear samples but can be encoded through controlled folding in the analog domain. Consequently, the long-standing trade-off between high dynamic range and high digital resolution—often considered fundamental—is revealed as a consequence of acquisition rather than a physical limit.

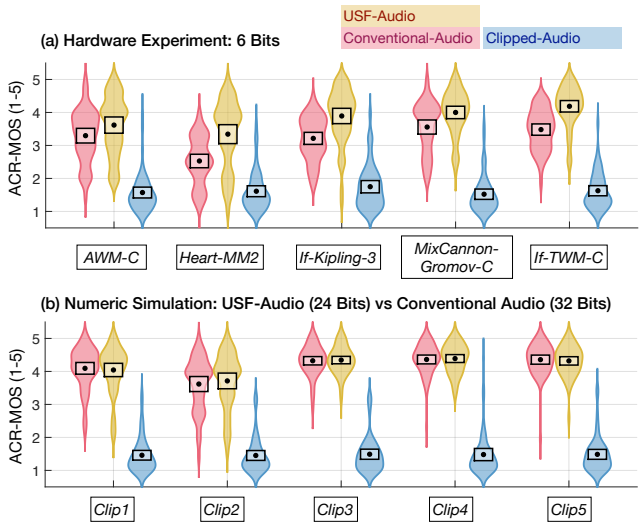


Fig. 10: ACR-MOS distributions for the listening study across (a) hardware (HW) and (b) high-bit digital (DIGI) regimes. Each violin shows the distribution of participant ratings (1–5 scale) for a given clip and condition: USF-Audio (red), conventional audio (yellow), and clipped audio (blue). Mean MOS is marked by "•" while "□" is used for 95% confidence interval. In the HW setting (6 bits), USF-Audio consistently shifts the MOS distribution upward relative to conventional audio, demonstrating improved perceptual quality under dynamic-range-limited acquisition. In the DIGI setting, conventional audio is given an explicit 8-bit advantage (32-bit vs. 24-bit for USF-Audio), yet the two distributions are nearly indistinguishable across all clips, indicating that USF-Audio remains perceptually transparent even against a higher-bit baseline. Clipped audio serves as a degraded anchor. Detailed statistical analysis, including Wilcoxon tests with Holm correction and effect sizes, is provided in the Supplementary Material.

Experimental results confirm that USF-Audio's benefits arise at the acquisition level and are largely content-agnostic. Consistent gains across speech, music, biomedical sounds, animal vocalizations, and historical recordings indicate that improvements do not rely on narrow signal priors or application-specific assumptions. Furthermore, demonstrated stability at audio scale establishes USF-Audio as not only a theoretical construct but an engineering-viable sensing approach.

Looking Ahead. More broadly, USF-Audio stresses the importance of hardware–algorithm co-design in audio signal processing. As downstream processing and learning algorithms advance, the quality and structure of data at the sensing stage increasingly determine achievable performance. By elevating acquisition from a fixed constraint to an active design dimension, USF-Audio opens new pathways for audio systems that preserve information by design rather than compensating for its loss. We anticipate that USF-Audio will open up new capabilities at Layer 3 and beyond within the APS (Table I).

Direct applications span audio and speech processing, acoustic scene analysis, bioacoustics, and medical acoustics. In this context—and in light of the rapid rise of AI and machine learning—a timely question emerges: *is the prevailing digital audio representation optimal for machine perception?* This question is meaningful because machine-perception systems do not access the acoustic scene directly; rather, they operate on the representation produced at acquisition. In USF-Audio, clipping is avoided and the classical DR–DRes trade-off is alleviated under a fixed bit budget, thereby preserving more

information for Layer 3 and beyond within the APS. This is also qualitatively reflected in the STFT visualizations in Fig. 8, where Fig. 8 (a_5) versus Fig. 8 (a_3), and Fig. 8 (b_5) versus Fig. 8 (b_3), reveal finer preserved spectral structure under USF-Audio. Beyond such qualitative evidence, early classification results [60] suggest that folded representation can improve downstream inference. Together, these observations motivate acquisition redesign as a promising direction for future machine listening, although a dedicated downstream benchmark is beyond the scope of the present work.

APPENDIX

Notation. Let $\mathbb{Z}, \mathbb{R}$ and $\mathbb{C}$ denote integers, real, and complex numbers, respectively. We use $\mathbb{I}_N = \{0, \cdots, N-1\}, N \in \mathbb{Z}^+$ to denote the set of N contiguous integers. Continuous functions and discrete sequences are written as $g(t), t \in \mathbb{R}$ and $g[n]$, respectively. The dynamic range (DR) of a signal $g(t)$ is defined as $\max_t g(t) - \min_t g(t)$. Vectors and matrices are written in bold lowercase and uppercase fonts, *e.g.* $\mathbf{g} = [g[1] \ \cdots \ g[N]]^\top \in \mathbb{R}^N$ and $\mathbf{G} = [g_{n,m}]_{n\in\mathbb{I}_N}^{m\in\mathbb{I}_M} \in \mathbb{R}^{N\times M}$. The $\mathbf{L}_p(\mathbb{R})$ space equipped with the p-norm is the standard Lebesgue space; $\mathbf{L}_\infty$ denotes the space of bounded functions. Spaces associated with sequences are written as ℓ^p. The max-norm ($\mathbf{L}_\infty$) of a function is defined as, $\|g\|_{\mathbf{L}_\infty} = \inf\{c_0 \geqslant 0 : |g(t)| \leqslant c_0\}$; for sequences, we use, $\|g\|_{\ell_\infty} = \max_n |g[n]|$. The $\mathbf{L}_2$-norm of a function is defined as, $\|g\|_{\mathbf{L}_2} = \sqrt{\int |g(t)|^2 dt}$ while for sequences, we have, $\|g\|_{\ell_2} = \sqrt{\sum_{n=0}^{N-1} |g[n]|^2}$. The vector space of polynomials with complex coefficients and degrees no greater than K is denoted by P_K, for instance, $\mathsf{Q}(z) = \sum_{k=0}^{K} h_k z^k \in P_K$. Function derivative is denoted by $\partial_t g(t)$ while for sequences, the finite-difference is denoted by $(\Delta g)[n] = g[n+1] - g[n]$. The Fourier Transform of $g \in \mathbf{L}_1$ is defined by $\widehat{g}(\omega) = \int_{\mathbb{R}} g(t) e^{-\jmath\omega t} dt$. The Discrete Fourier Transform (DFT) of a sequence $\mathbf{g} \in \ell_1$ is denoted by $\widehat{g}[m] = \sum_{n=0}^{N-1} g[n] e^{-\jmath\frac{2\pi mn}{N}}$. Let $\mathbf{V}_N^M$ denote the $N\times M$ Vandermonde matrix $\mathbf{V}_N^M = \left[\xi_N^{n\cdot m}\right]_{n\in\mathbb{I}_N}^{m\in\mathbb{I}_M}, \xi_N^n = e^{\jmath\frac{2\pi n}{N}}$. Diagonal matrices are denoted by $\mathscr{D}_K(\mathbf{h})$ with $[\mathscr{D}_K(\mathbf{h})]_{k,k} = [\mathbf{h}]_{k\in\mathbb{I}_K}$. The quantization operator is defined as $\mathscr{Q}_\lambda(g) = 2\lambda \left\lfloor \frac{g+\lambda}{2\lambda} \right\rfloor$ where $\lfloor g \rfloor = \sup\{k \in \mathbb{Z} | \, k \leqslant g\}$ is the floor function. The mean squared error (MSE) between $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ is defined as $\mathcal{E}_2(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N} |x[n] - y[n]|^2$. Given B bits quantization, the digital resolution for an input signal $g(t) \in \mathbf{L}_\infty$ is defined as, $2\eta_g^B = 2\|g\|_{\mathbf{L}_\infty} / \left(2^B - 1\right)$. With folding threshold λ, we define *Dynamic Range Expansion* of g as, $\mathsf{DRX} = \eta_g^B / \eta_\lambda^B = \|g\|_{\mathbf{L}_\infty} / \lambda$. We quantify clipping severity using the input signal-to-distortion ratio (SDR) [12], $\mathrm{SDR}(g, g_{\mathsf{clip}}) = 20\log_{10} \frac{\|g\|_2}{\|g - g_{\mathsf{clip}}\|_2}$, where g_{clip} denotes clipped version of g with threshold λ_{clip} [15]. Restoration accuracy is measured by, $\Delta\mathrm{SDR}(g, g_{\mathsf{clip}}) = \mathrm{SDR}(g, \tilde{g}) - \mathrm{SDR}(g, g_{\mathsf{clip}})$, where $\tilde{g}$ is the reconstruction of g.

Comparison with Alternative Hardware Solutions. Alternative HDR front-end strategies, such as dual-gain or multi-range ADCs [19], [21], limiter/headroom approaches [61], and AGC (automatic gain control)-based systems [21], improve robustness to overload by either adding hardware paths or modifying the signal prior to quantization. Dual-gain architectures require multiple gain paths, calibration, and fusion; limiter/headroom methods suppress peaks at the cost of altered dynamics; and AGC avoids saturation through time-varying gain control, but modifies the signal itself, and in doing so may distort or obscure semantically important cues [62], [63]. Fundamentally, in all such cases, the "6-dB per bit rule" [3] remains a digital bottleneck. By contrast, USF-Audio changes the acquisition itself via a single-channel analog-domain folding, yielding a new digital representation. Its advantage is thus structural: rather than managing the signal so that a linear ADC does not clip, USF-Audio relieves the quantizer from representing the full input DR, enabling simultaneous gains in DRX and effective resolution under the same nominal bit budget.

[15] Smaller λ_{clip} leads to more severely clipped measurements; empirically speaking, $\lambda_{\mathsf{clip}} \leqslant 0.1\|g\|_{\mathbf{L}_\infty}$ results in an approximately 1-3 dB SDR, suggesting over 50% samples are clipped [12].

ACKNOWLEDGEMENT

The authors thank the Prof. Bernardini (AE), for the careful handling of the manuscript, and the reviewers for their constructive and encouraging comments. We are grateful to Alfred M. Bruckstein, Aurélien Bourquard, Mike Brookes, Frédéric Bimbot, Thierry Blu, Mario A. T. Figueiredo, and K. V. S. Hari for their encouraging remarks on this work. We thank Aurélien Bourquard, Mayank Manas and Václav Pavlíček for valuable feedback on earlier versions of the manuscript. We also thank Simon Arberet, Matthias Beckmann, Miguel C. Heredia, Aidan Hogg, Matthieu Kowalski, Lorenzo Picinali, Chandra S. Seelamantula, Salman Asif, José Luis Romero, Irina Shafkulovska, and Siddharth Somasundaram for participating in the initial audio survey prior to its public release.

REFERENCES

[1] G. Richard, P. Smaragdis, S. Gannot, P. A. Naylor, S. Makino, W. Kellermann, and A. Sugiyama, "Audio signal processing in the 21st century: The important outcomes of the past 25 years," *IEEE Signal Processing Magazine*, vol. 40, no. 5, pp. 12–26, Jul. 2023.

[2] H. S. Black and J. O. Edson, "Pulse code modulation," *Transactions of the American Institute of Electrical Engineers*, vol. 66, no. 1, pp. 895–899, Jan. 1947.

[3] F. F. E. Owen, *PCM and Digital Transmission Systems*, ser. Texas Instruments Electronics Series. McGraw-Hill, 1982.

[4] W. Bennett, "Secret telephony as a historical example of spread-spectrum communication," *IEEE Transactions on Communications*, vol. 31, no. 1, pp. 98–104, Jan. 1983.

[5] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

[6] N. Jayant, "Digital coding of speech waveforms: PCM, DPCM, and DM quantizers," *Proc. IEEE*, vol. 62, no. 5, pp. 611–632, May 1974.

[7] J. Vanderkooy and S. Lipshitz, "Why 1-bit sigma-delta conversion is unsuitable for high-quality applications," in *Audio Engineering Society Convention*, May 2001.

[8] M. J. Hawksford, "SDM versus LPCM: The debate continues," in *Audio Engineering Society Convention*, 2001.

[9] C. Shannon, "Communication in the presence of noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, Jan. 1949.

[10] B. Blesser, "Audio dynamic range compression for minimum perceived distortion," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 1, pp. 22–32, Mar. 1969.

[11] J. Abel and J. Smith, "Restoring a clipped signal," in *IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*. IEEE, Apr. 1991.

[12] P. Zaviska, P. Rajmic, A. Ozerov, and L. Rencker, "A survey and an extensive evaluation of popular audio declipping methods," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 1, pp. 5–24, Jan. 2021.

[13] C. Gaultier, S. Kitić, R. Gribonval, and N. Bertin, "Sparsity-based audio declipping methods: Selected overview, new algorithms, and large-scale evaluation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1174–1187, Feb. 2021.

[14] R. Walden, "Analog-to-digital converter survey and analysis," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 4, pp. 539–550, Apr. 1999.

[15] P. Gupta, M. J. Moghimi, Y. Jeong, D. Gupta, O. T. Inan, and F. Ayazi, "Precision wearable accelerometer contact microphones for longitudinal monitoring of mechano-acoustic cardiopulmonary signals," *npj Digital Medicine*, vol. 3, no. 1, Feb. 2020.

[16] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *IEEE Conf. on Advanced Video and Signal Based Surveillance*. IEEE, Sep. 2007.

[17] R. C. Maher, "Audio forensic examination," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 84–94, Mar. 2009.

[18] P. Smaragdis, "Dynamic range extension using interleaved gains," *IEEE Audio, Speech, Language Process.*, vol. 17, no. 5, pp. 966–973, Jul. 2009.

[19] R. Janzen and S. Mann, "Extreme-dynamic-range sensing: Real-Time adaptation to extreme signals," *IEEE Multimedia*, vol. 24, no. 2, pp. 30–42, Apr. 2017.

[20] ——, "High dynamic range simultaneous signal compositing, applied to audio," in *IEEE Canadian Conf. on Electrical and Computer Engg. (CCECE)*. IEEE, Apr. 2012.

[21] J. Hernandez, D. E. Garcia, and S. Mann, "HDR AGC signal optimization applied to audio," in *2020 IEEE SENSORS*, Oct. 2020, pp. 1–4.

[22] N. S. Jayant and L. R. Rabiner, "The application of dither to the quantization of speech signals," *Bell Syst. Tech. J.*, vol. 51, no. 6, pp. 1293–1304, Jul. 1972.

[23] J. Vanderkooy and S. P. Lipshitz, "Resolution below the least significant bit in digital systems with dither," *Journal of the Audio Engineering Society*, vol. 32, no. 3, pp. 106–113, Mar. 1984.

[24] S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy, "Quantization and dither: A theoretical survey," *Journal of the Audio Engineering Society*, vol. 40, no. 5, pp. 355–375, May 1992.

[25] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1264–1271, May 2012.

[26] A. P. Hill, P. Prince, E. P. Covarrubias, C. P. Doncaster, J. L. Snaddon, and A. Rogers, "AudioMoth: Evaluation of a smart open acoustic device for monitoring biodiversity and the environment," *Methods in Ecology and Evolution*, vol. 9, no. 5, pp. 1199–1211, Jan. 2018.

[27] S. S. Sethi *et al.*, "Characterizing soundscapes across diverse ecosystems using a universal acoustic feature set," *Proceedings of the National Academy of Sciences*, vol. 117, no. 29, pp. 17 049–17 055, Jul. 2020.

[28] S. Maurice *et al.*, "In situ recording of Mars soundscape," *Nature*, vol. 605, no. 7911, pp. 653–658, Apr. 2022.

[29] V. Pavlíček and A. Bhandari, "1-bit unlimited sampling beyond Fourier domain: Low-resolution sampling of quantization noise," *IEEE J. Sel. Topics Signal Process.*, pp. 1–12, 2025.

[30] A. Bhandari, F. Krahmer, and R. Raskar, "On unlimited sampling," in *Intl. Conf. on Sampling Theory and Applications (SampTA)*, Jul. 2017.

[31] ——, "On unlimited sampling and reconstruction," *IEEE Trans. Signal Process.*, vol. 69, pp. 3827–3839, Dec. 2020.

[32] A. Bhandari, F. Krahmer, and T. Poskitt, "Unlimited sampling from theory to practice: Fourier-Prony recovery and prototype ADC," *IEEE Trans. Signal Process.*, pp. 1131–1141, Sep. 2021.

[33] S. Rudresh, A. Adiga, B. A. Shenoy, and C. S. Seelamantula, "Wavelet-based reconstruction for unlimited sampling," in *IEEE Intl. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, Apr. 2018, pp. 4584–4588.

[34] O. Ordentlich, G. Tabak, P. K. Hanumolu, A. C. Singer, and G. W. Wornell, "A modulo-based architecture for analog-to-digital conversion," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 5, pp. 825–840, Oct. 2018.

[35] Y. Gong, L. Gan, and H. Liu, "Multi-channel modulo samplers constructed from Gaussian integers," *IEEE Signal Processing Letters*, vol. 28, pp. 1828–1832, 2021.

[36] A. Eamaz, K. V. Mishra, F. Yeganegi, and M. Soltanalian, "UNO: Unlimited sampling meets one-bit quantization," *IEEE Trans. Signal Process.*, vol. 72, pp. 997–1014, 2024.

[37] E. Azar, S. Mulleti, and Y. C. Eldar, "Unlimited sampling beyond modulo," *Appl. Comput. Harmon. Anal.*, vol. 74, p. 101715, Jan. 2025.

[38] Z. Liu, A. Bhandari, and B. Clerckx, "λ-MIMO: Massive MIMO via modulo sampling," *IEEE Trans. Commun.*, pp. 6301 – 6315, Nov. 2023.

[39] M. Beckmann, A. Bhandari, and M. Iske, "Fourier-domain inversion for the modulo Radon transform," *IEEE Trans. Comput. Imaging*, vol. 10, pp. 653–665, Apr. 2024.

[40] T. Feuillen, B. S. M. R. Rao, and A. Bhandari, "Unlimited sampling radar: Life below the quantization noise," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023.

[41] B. Monroy and J. Bacca, "Deep lightweight unrolled network for high dynamic range modulo imaging," *IEEE Trans Comput Imaging*, vol. 12, pp. 406–415, 2026.

[42] R. Guo, Y. Zhu, and A. Bhandari, "Sub-Nyquist USF spectral estimation: K frequencies with $6K + 4$ modulo samples," *IEEE Trans. Signal Process.*, vol. 72, pp. 5065–5076, 2024.

[43] R. Guo and A. Bhandari, "Unlocking off-the-grid sparse recovery with unlimited sensing: Simultaneous super-resolution in time and amplitude," *IEEE J. Sel. Topics Signal Process.*, pp. 1–13, 2025.

[44] E. Romanov and O. Ordentlich, "Above the Nyquist rate, modulo folding does not hurt," *IEEE Signal Processing Letters*, vol. 26, no. 8, pp. 1167–1171, Aug. 2019.

[45] S. B. Shah, S. Mulleti, and Y. C. Eldar, "Lasso-based fast residual recovery for modulo sampling," in *IEEE Intl. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, Jun. 2023.

[46] N. I. Bernardo, "Short-time Fourier transform-based signal recovery for modulo analog-to-digital converters," *IEEE Access*, vol. 14, pp. 7308–7324, 2026.

[47] S. Mulleti, E. Reznitskiy, S. Savariego, M. Namer, N. Glazer, and Y. C. Eldar, "A hardware prototype of wideband high-dynamic range analog-to-digital converter," *IET Circuits, Devices & amp; Systems*, vol. 17, no. 4, pp. 181–192, Jun. 2023.

[48] Y. Zhu and A. Bhandari, "Unleashing dynamic range and resolution in unlimited sensing framework via novel hardware," in *2024 IEEE SENSORS*. IEEE, Oct. 2024, pp. 1–4.

[49] J. Zhu, J. Ma, Z. Liu, F. Qu, Z. Zhu, and Q. Zhang, "A modulo sampling hardware prototype and reconstruction algorithms evaluation," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–11, 2025.

[50] D. Prasanna, C. Sriram, and C. R. Murthy, "On the identifiability of sparse vectors from modulo compressed sensing measurements," *IEEE Signal Process. Lett.*, vol. 28, pp. 131–134, 2021.

[51] R. Schreier, S. Pavan, and G. C. Temes, *Understanding Delta-Sigma Data Converters*. Wiley, Mar. 2017.

[52] A. Bhandari and F. Krahmer, "HDR imaging from quantization noise," in *IEEE Intl. Conf. on Image Processing (ICIP)*, Oct. 2020, pp. 101–105.

[53] N. I. Bernardo, S. B. Shah, and Y. C. Eldar, "Modulo sampling: Performance guarantees in the presence of quantization," *IEEE Trans. Signal Process.*, vol. 73, pp. 3829–3842, 2025.

[54] S. J. Godsill, P. J. Wolfe, and W. N. Fong, "Statistical model-based approaches to audio restoration and analysis," *J New Music Res*, vol. 30, no. 4, pp. 323–338, Dec. 2001.

[55] C. Brauer, T. Gerkmann, and D. Lorenz, "Sparse reconstruction of quantized speech signals," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016.

[56] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, "Audio inpainting," *IEEE Audio, Speech, Language Process.*, vol. 20, no. 3, pp. 922–932, Mar. 2012.

[57] H. Liu, X. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang, "VoiceFixer: A unified framework for high-fidelity speech restoration," in *Interspeech 2022*, ser. interspeech 2022. ISCA, Sep. 2022, pp. 4232–4236.

[58] V. Sechaud, L. Jacques, P. Abry, and J. Tachella, "Equivariance-based self-supervised learning for audio signal recovery from clipped measurements," in *2024 European Signal Processing Conference (EUSIPCO)*. IEEE, Aug. 2024, pp. 852–856.

[59] ——, "Learning to reconstruct from saturated data: audio declipping and high-dynamic range imaging," *arXiv*, 2026.

[60] V. Pavlíček and A. Bhandari, "Gesture recognition with USF-radars: Modulo pre-processing enhances classification," in *2025 IEEE Radar Conference (RadarConf25)*. IEEE, Oct. 2025, pp. 1–6.

[61] B. A. Blesser, "Digitization of Audio: A Comprehensive Examination of Theory, Implementation, and Current Practice," *Journal of the Audio Engineering Society*, vol. 26, no. 10, pp. 739–771, oct 1978.

[62] M. A. Stone and B. C. J. Moore, "Effect of the speed of a single-channel dynamic range compressor on intelligibility in a competing speech task," *The Journal of the Acoustical Society of America*, vol. 114, no. 2, pp. 1023–1034, july 2003.

[63] R. C. Maher, *Principles of Forensic Audio Analysis*. Springer International Publishing, 2018.