CEPAT Journal of Computer Engineering: Progress, Application and Technology
Vol. 5, No. 2, November 2026, pp. 163~179
ISSN: 2963-6728, DOI: https://doi.org/10.25124/cepat.v5i02.12429

Sampling Rate and Bit Depth: Fidelity–Storage Trade-offs for Speech and Music

Salsabila HR[1], Dimas Fadhlurrahman Haryadi[1], Deny Adrian Waloya[1], Revalia Putri[1], Agustien Mulyantini[1]
[1]Department of Computer Engineering, School of Electrical Engineering, Telkom University, Indonesia

Article Info

Article history:

Received October 23, 2026
Revised November 07, 2026
Accepted November 09, 2026

Keywords:

Audio fidelity
Bit depth
Lossless compression
Pulse-code modulation
Quantization error
Resampling filter
Sampling rate

ABSTRACT

The sampling rate and bit depth of uncompressed digital audio affect its quality and also its file size. What is less clear is how much quality an extra hertz or an extra bit actually adds, and whether the answer is the same for speech and music. We tested both parameters, separately and in combination, with a Python experiment that other people can rerun. Twenty-four clips of 10 s were used (anechoic speech, orchestral or chamber music, and pop or rock from MUSDB18-HQ). Each clip was resampled to 8–48 kHz and quantized to 4–24 bit. We measured quality with the signal-to-noise ratio (SNR), spectral descriptors, and the perceptual estimator ViSQOL, and we measured storage from the size of WAV and FLAC files. A low sampling rate hurt pop the most, with 14.2 dB SNR at 8 kHz against 17.4 dB for speech and 24.9 dB for classical music. The default resampling filter turned out to be the cause of the 58 dB plateau at 44.1 kHz and of the missing gain from 24 bit; with a sharper filter the SNR at 44.1 kHz went up by 11–20 dB. If every clip had to reach 40 dB, the smallest tested setting was 44.1 kHz with 16 bit for speech and pop and 22.05 kHz with 16 bit for classical music. Using the mean as the criterion would have pointed to 32 kHz with 16 bit, even though seven of the eight pop clips stayed below 40 dB at that setting. A clip with 40 dB SNR could still score below 4.0 on ViSQOL. With FLAC, 16-bit files shrank to 0.50–0.74 of their WAV size. The material was limited to four speakers, four recordings, and four tracks, and we did not run any listening test.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Agustien Mulyantini
Department of Computer Engineering, School of Electrical Engineering, Telkom University Bandung, Indonesia
Email: agustienmulyantiniam@telkomuniversity.ac.id

1. INTRODUCTION

In digital audio, a continuous sound wave is stored as a list of numbers produced by sampling and quantization. The sampling rate (in hertz) tells how many samples are taken per second. According to the Nyquist–Shannon sampling theorem, it also limits the highest frequency that can be kept without aliasing [1], [2]. Bit depth is the number of bits for one sample, and a B-bit linear pulse-code modulation (PCM) signal has 2^B quantization levels. Consumer audio mostly follows the Compact Disc format (16 bits, 44.1 kHz), while systems that go beyond 16 bits or 44.1/48 kHz are usually called high-resolution audio [3]. Uncompressed PCM gets bigger linearly with both parameters, which means every step up in resolution costs storage and bandwidth.

Raising the sampling rate or the bit depth does not give the same improvement every time. Extra bandwidth from a higher sampling rate only helps when the source actually has energy in that band. A higher bit depth lowers the quantization error, but in practice this depends on the noise floor of the signal and of the playback chain. When the bit rate of acquisition hardware is limited, the designer has to trade sampling rate

against resolution [2]. Someone who has to choose a format for recording, archiving, or processing wants to know how much measurable quality each extra hertz or bit gives, and what it costs in storage.

Several earlier studies are related to this question. Reiss [3] did a meta-analysis of 18 listening experiments (more than 400 participants, over 12,500 trials) and found that people could tell high-resolution audio from 16-bit, 44.1 or 48 kHz audio slightly but significantly, and much better after extensive training. In a study on least-significant-bit steganography, Ramya Devi and Pugazhenthi [1] looked at how the sampling rate changes the SNR of audio with hidden data; the added noise was lower when the sampling rate followed the Nyquist–Shannon criterion. Zhao et al. [2] analyzed the recovered SNR of a compressed-sensing sub-Nyquist system for different sampling rates and quantization bits. According to their analysis, a system with high SNR benefits more from resolution, and a system with low SNR benefits more from a higher sampling rate. Liu et al. [4] built AudioSR, a diffusion model that brings back the high-frequency part of band-limited audio up to 48 kHz, which shows that losing bandwidth at a low sampling rate is a practical issue. The way audio quality is measured has also been studied. Torcoli et al. [5] found that objective perceptual measures are usually designed for one application area and may not be reliable in other areas, and Chinen et al. [6] make a distinction between physical measures (SNR, spectral distortion) and perceptual estimators such as PESQ, POLQA, and ViSQOL.

The bandwidth removed by a certain sampling rate is different for each type of content. Speech has little energy above about 5 kHz compared with the formant region, yet this part has never been shown to be unimportant. Monson et al. [7] reviewed evidence that high-frequency energy plays a role in voice and speech perception, including quality and intelligibility. Bandwidth-extension research looks at the same loss from the opposite direction. Su et al. [8] extended 8–16 kHz speech to 48 kHz using a WaveNet-based generator with a deep-feature adversarial loss, and Han and Lee [9] used one diffusion model to generate 48 kHz audio from inputs at different sampling rates. Mandel et al. [10] applied spectral-domain super-resolution to speech and music over many input rates and evaluated it with log-spectral distance, ViSQOL, and MUSHRA. Resampling can also add its own changes to the signal. Pons et al. [11] reported tonal and filtering artifacts and spectral replicas caused by upsampling, and Saito et al. [12] showed that a convolutional layer behaves like a bank of digital filters and therefore depends on the sampling frequency, so they proposed sampling-frequency-independent layers for unseen rates. On the storage side, lossless coders such as the natural-gradient predictor and NARU codec of Mineo and Shouno [13] make PCM smaller without changing any sample. Neural codecs like SoundStream [14] run at only 3–18 kbit/s, around 40 to 250 times lower than 16-bit, 48 kHz mono PCM (768 kbit/s), but the original waveform is not preserved.

To sum up, the reviewed sources deal with perceptual discrimination [3], one particular use of sampling rate [1], a theoretical acquisition model [2], whether quality measures are valid [5], [6], the perceptual role of high-frequency energy in speech [7], generative restoration of missing bandwidth [4], [8]–[10], the rate dependence of resampling operators [11], [12], and compression beyond PCM [13], [14]. In these sources we could not find a controlled comparison for ordinary uncompressed audio files where sampling rate and bit depth are varied on their own and together, for speech, orchestral music, and pop music, and where signal-level, spectral, and perceptual metrics are reported together with storage measured from real WAV and lossless FLAC files. We limit this statement to the sources we reviewed.

Part of what such an experiment shows can already be predicted from theory: the quantization SNR rises by 6.02 dB per bit [15], [16], and PCM size grows linearly with both parameters. We do not claim a new algorithm or a new theoretical result in this paper. Our aim is a reproducible empirical benchmark. All parameters, data sources, and processing steps are written down, so another researcher can run the same setup again, or apply it to different material, and compare the results. Within this benchmark we obtained four findings that do not follow from theory alone. The first is a separation between the SNR limit set by the signal bandwidth and the SNR limit set by the resampling filter; with a default filter, the plateau at high rates and the lack of benefit from 24 bit both come from the filter. The second is that the setting considered sufficient changes when a criterion is applied to the mean instead of every clip, and when a perceptual estimator is used instead of SNR. As a third result, we report lossless FLAC sizes, which make a higher sampling rate less expensive. Finally, we summarize the experiment as a five-step procedure that estimates the needed sampling rate for a quality target from the spectrum of the source and takes the bit depth from Eq. (6), so the full factorial design does not have to be repeated.

The experiment in this paper is controlled and reproducible, and it is written as a Python notebook that can be run in Google Colab or locally. We address four research questions. RQ1: How does the sampling rate affect the quality of the reconstructed audio signal, measured by SNR, RMSE, spectral characteristics, and a perceptual estimator? RQ2: How does the bit depth affect quantization error, SNR, and RMSE under controlled conditions? RQ3: How does the combination of sampling rate and bit depth affect file size, in WAV and in FLAC, and the balance between signal quality and storage? RQ4: To what extent do the resampling filter and the acceptance criterion (mean or every clip; SNR or ViSQOL) change the setting that

is judged sufficient? Our contributions are (i) a factorial design that keeps the band-limiting error and the quantization error apart, run with two resampling filters whose results are shown side by side; (ii) operating points for given SNR and ViSQOL criteria, checked on every clip, for three content categories, plus a five-step procedure to find them for other material (Section 3.8); (iii) one analysis that combines signal quality, a perceptual estimator, and measured WAV and FLAC sizes, including the gap between the actual WAV size and the theoretical PCM size; and (iv) a fully reproducible benchmark, made of an open notebook that regenerates all tables and figures (available at https://github.com/Snivoy/audio-sampling-bitdepth), a single configuration cell with every parameter of Table 1, and a manifest that records each clip together with its source, license, and excerpt offset (Section 2.4). No new algorithm is proposed and no listening test was carried out. What the study offers is a transparent quantitative baseline that can be repeated.

2. METHOD

2.1. Experimental design

Fig. 1 shows the pipeline of the experiment. From each clip we take a reference signal at its native sampling rate $f_{s,ref}$. There are three series: Series A varies only the sampling rate (no quantization), Series B varies only the bit depth at the native rate, and Series C varies both. Keeping the factors apart in this way lets us tell the band-limiting error from the quantization error, while Series C shows how they interact. Series A and C were each run twice, first with the default filter of the SciPy resampler and then with a sharper resampling filter. Both runs are reported and compared (Fig. 2(a), Table 4, Section 3.2, and Section 3.7). We kept both because they answer different questions: the default filter is what a user gets without tuning, and the sharper filter tells us how close a steep converter can come to the ideal band limit. The detailed per-clip tables (Tables 3 and 5–10) only show the sharper filter, to avoid duplication, and Section 3.2 points out where the default filter gives another result.

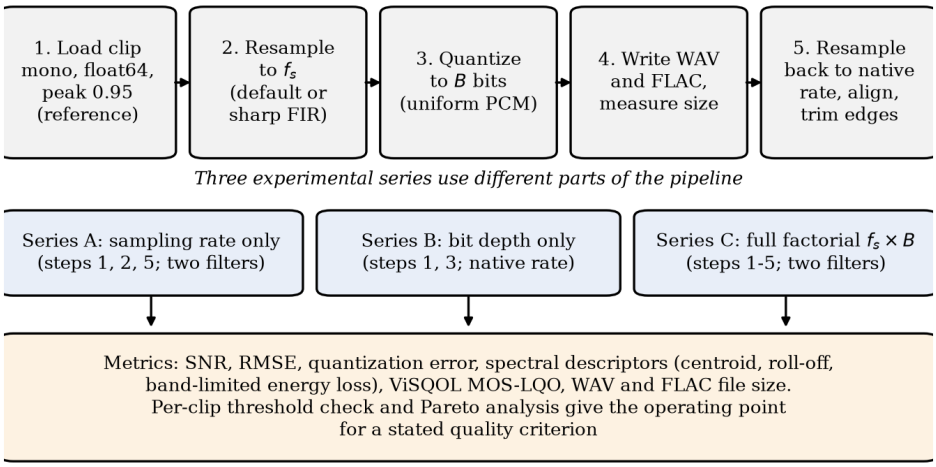


Figure 1. Experimental pipeline and the three experimental series

2.2. Audio material

In total, 24 clips of 10 s were used, eight in each of three content categories. For speech we took eight clips from the EARS dataset [17], a set of anechoic studio recordings released under the CC BY-NC 4.0 license (https://github.com/facebookresearch/ears_dataset), namely two read sentences in the regular style from each of four speakers (p001–p004; two female, two male). The classical clips come from four public-domain Musopen recordings hosted on Wikimedia Commons [18]. Three of them are symphony-orchestra works (Beethoven, Coriolan Overture, Op. 62; Borodin, In the Steppes of Central Asia; Mozart, Overture to Le nozze di Figaro, K. 492), and the fourth is a string quartet (Mendelssohn, String Quartet No. 6, Op. 80, first movement). The pop clips are taken from four tracks of the MUSDB18-HQ test set [19], the uncompressed version of the MUSDB18 corpus [20], which offers full-length pop and rock mixtures as WAV files for non-commercial research: Al James, Schoolboy Facination; AM Contra, Heart Peripheral; Georgia Wonder, Siren; and The Sunshine Garcia Band, For I Am The Moon. Two excerpts that do not overlap were cut from every music recording, one starting at 30% and one at 60% of its length.

The four pop tracks were picked with a fixed rule that we applied to all 50 test tracks. To be eligible, a track could not have a steep cliff in its long-term spectrum (see below) or any noise or tone close to the Nyquist frequency. Forty-three tracks passed, and we took the four with the largest share of power above 8 kHz (the selection table is published as pop_selection.csv). The pop category is thus rich in high-frequency

content and makes a demanding test for low sampling rates, so it should be read as a stress test and not as a random sample of pop music.

Material that was lossy coded at some point would make the damage from a low sampling rate look smaller, since encoders such as MP3, AAC, Vorbis, or Opus cut off the band above their cut-off frequency. Our first step was to check the documentation of the original repositories for every source. (i) Classical: the four files on Wikimedia Commons [18] were uploaded from the Internet Archive collection of the 2012 Musopen Kickstarter Project, which is published in FLAC [21]. According to the metadata on Commons, all four recordings are lossless FLAC streams with 48 kHz, 24 bit, and two channels. We also compared SHA-1 checksums, and the downloaded files matched the checksums listed on Commons, so they are bit-identical to the archived versions. (ii) Pop: the maintainers describe MUSDB18-HQ as the uncompressed WAV version of MUSDB18 with a bandwidth up to 22 kHz, whereas the normal MUSDB18 release contains AAC streams at 256 kbit/s band-limited to 16 kHz [19], [22]. We used the HQ release, whose files are 16-bit PCM WAV at 44.1 kHz. (iii) Speech: EARS was recorded at 48 kHz in an anechoic chamber and is distributed as WAV [17]. No lossy step is mentioned for any of the sources. Documentation alone cannot exclude processing done before the release, so a spectral check was added as well.

Lossy encoding typically leaves a steep drop in the long-term spectrum between about 11 and 20 kHz. We looked for the steepest drop of the long-term power spectrum within 1 kHz for every source (the whole recording for classical and pop, the clip itself for speech). The search ran from 11 kHz up to min(21 kHz, $f_s/2 - 1.5$ kHz), which keeps it away from the anti-aliasing roll-off of the file format. The largest drops we found were 15.7 dB/kHz (classical), 18.6 dB/kHz (pop), and 14.4 dB/kHz (speech), all under the alarm threshold of 20 dB/kHz that we set. Above 16 kHz, pop had 0.03–0.35% of its power and the other categories at most 0.02%, and the spectrograms of the pop clips go up to about 21 kHz (Fig. 2(c)). These results are in line with the documentation. Still, the check only catches the usual band-limiting signature; coding at a high bit rate without a cliff would pass, and the 20 dB/kHz limit is a heuristic. Values for every source can be found in source_integrity.csv.

Speech and classical clips have a native sampling rate of 48,000 Hz. The speech files are 32-bit floating point with one channel, and the classical files are 24-bit integer PCM with two channels. Pop clips have a native rate of 44,100 Hz, 16 bit, and two channels. We averaged every clip to a single channel so that all conditions use mono signals and the channel count plays no role in the results. After that, each signal was converted to 64-bit floating point in the range [−1, 1) and peak-normalized to 0.95 of full scale, which avoids clipping in the quantizer. Because the loss from a given sampling rate depends on how much high-frequency energy a source contains, the spectral descriptors of the references are listed here. Mean spectral centroids were 796 ± 301 Hz (speech), 705 ± 143 Hz (classical), and 839 ± 303 Hz (pop), and mean 95% roll-off frequencies were 3,293 ± 2,788 Hz, 2,045 ± 335 Hz, and 3,902 ± 1,749 Hz. Speaker p003 explains the large spread in speech, since both of their clips have a roll-off near 7.8 kHz; in pop, the spread comes from tracks whose roll-off lies between 1,561 and 6,385 Hz. Per-clip values are available in the published results file.

2.3. Parameter variation

The experimental parameters are listed in Table 1. We used six standard sampling rates: 8 kHz (narrowband telephony), 16 kHz (wideband speech), 22.05 kHz (half the CD rate, common in older multimedia), 32 kHz (broadcast and some video formats), 44.1 kHz (CD), and 48 kHz (professional audio and video). Their Nyquist frequencies run from 4 to 24 kHz, which covers the region where the three categories differ the most (Section 2.2), and the top rate is equal to the native rate of two categories. Rates above the native rate were not used, since upsampling does not add information. For pop, whose native rate is 44,100 Hz, this leaves five rates, and the pooled results therefore stop at 44.1 kHz. No resampling happens at the native rate, so this level works as the identity condition in Series A and isolates the quantizer in Series C.

Bit depths of 8, 16, and 24 were chosen because WAV and FLAC have a native container for them (8-bit PCM is unsigned in WAV and signed in FLAC, while the quantizer is identical). The quantization-only sweep from 4 to 24 bit also contains depths without a native container. We use it to test the 6 dB-per-bit law across the whole range and only look at its quantization error.

For the spectral analysis, a Welch window of 4096 samples with a Hann taper was used. It gives a resolution of 11.7 Hz at 48 kHz (10.8 Hz at 44.1 kHz), fine enough to resolve the harmonics of low voices. The 95% roll-off is the common definition in audio classification [23]. We tested whether these choices matter by recomputing the reference descriptors with 2048- and 8192-sample windows and with roll-off fractions of 0.85 and 0.99. Changing the window length moved the mean centroid by at most 0.3%, the 95% roll-off by at most 0.8%, and $\mathrm{SNR}_{\mathrm{BL}}$ by at most 0.10 dB. The roll-off fraction had a much larger effect on the absolute roll-off (for pop, 1.2 kHz at 0.85 and 10.2 kHz at 0.99), and at 0.85 the categories were ranked in a different order than at 0.95 and 0.99. For that reason, the roll-off is only used to describe the sources. The conclusions rely on E_{loss}, SNR, and ViSQOL, none of which depends on it.

Table 1. Experimental parameters

Parameter	Values or setting
Sampling rate f_s (Hz)	8,000; 16,000; 22,050; 32,000; 44,100; 48,000 (those ≤ native rate)
Bit depth B, file-based series (bit)	8, 16, 24 (WAV: unsigned 8-bit PCM, signed 16- and 24-bit PCM; FLAC: signed)
Bit depth B, quantization sweep (bit)	4, 6, 8, 10, 12, 14, 16, 20, 24
Channels	1 (mono)
Reference signal	Native rate (48,000 Hz speech and classical; 44,100 Hz pop), 64-bit float, peak 0.95
Resampling, default filter	Polyphase FIR (SciPy resample_poly), SciPy default: Kaiser window, β = 5, half-length $10 \cdot \max(p, q)$
Resampling, sharper filter	Polyphase FIR (SciPy resample_poly), Kaiser window, β = 12, half-length $50 \cdot \max(p, q)$
Quantizer	Uniform mid-tread, range [−1, 1), no dither
Spectral analysis	Welch method, Hann window, 4096 samples, 95% roll-off (sensitivity: 2048 and 8192 samples; 85% and 99%)
Edge trimming	50 ms at both ends before computing metrics
Perceptual metric	ViSQOL v3, audio mode (MOS-LQO, 1–5), at 48 kHz; Series C only
Storage formats	WAV (PCM_U8, PCM_16, PCM_24) and FLAC (libsndfile default compression level)
Acceptance criteria	SNR ≥ 30, 40, 50, 60 dB; ViSQOL ≥ 4.0, 4.5; for every clip and for the mean

2.4. Processing procedure and reproducibility

The whole procedure runs inside one Jupyter notebook, which works in Google Colab or a local Python environment and is published at https://github.com/Snivoy/audio-sampling-bitdepth. It relies on Python, NumPy, SciPy, soundfile, pandas, Matplotlib, and the ViSQOL Python package. An environment log stores the exact versions automatically; for the results in this paper they were Python 3.14.6, NumPy 2.5.4, SciPy 1.18.1, pandas 3.0.6, soundfile 0.14.0 with libsndfile 1.2.2, Matplotlib 3.11.2, and visqol-python 3.8.0. One configuration cell holds every parameter of Table 1. The first part of the notebook prepares the audio: speech and classical clips are downloaded from their public sources [17], [18], pop excerpts are read from a user-provided local copy of MUSDB18-HQ [19], and the source, license, and excerpt offset of each clip are written to data_manifest.csv. Other clips can be tested by putting them in one folder per content category. Processing then follows the steps below.

Step 1 (load and normalize). We read each clip, mix it down to mono, convert it to floating point, and peak-normalize it to obtain the reference $x_{\text{ref}}[n]$.

Step 2 (resampling). A polyphase filter converts the reference to the target rate f_s, x_s = resample_poly(x_{ref}, p, q, window = h), with $p = f_s/g$, $q = f_{s,\text{ref}}/g$, and g the greatest common divisor of the two rates. In the default case the window argument is left out, and SciPy falls back to its default (Kaiser, β = 5, half-length $10 \cdot \max(p, q)$). In the sharper case, h is a Kaiser-window low-pass FIR with β = 12 and length $2 \cdot 50 \cdot \max(p, q) + 1$, and its cut-off sits at the lower of the two Nyquist frequencies. We treat both filters as experimental conditions (Section 3.2).

Step 3 (quantization). The resampled signal is quantized to B bits by the uniform quantizer of Eq. (2) and Eq. (3), with no dither. In WAV the 8-bit case is stored as unsigned integers, although the quantizer itself does not change.

Step 4 (file writing). Using the soundfile library, we write the quantized signal as a WAV file (PCM_U8, PCM_16, or PCM_24) and as a FLAC file [24] (signed 8, 16, or 24 bit, libsndfile default compression level) and read both sizes from the file system. Every file is read back and compared with the signal in memory to confirm that nothing was lost. The quality metrics themselves are computed from the in-memory signals and not from the files.

Step 5 (reconstruction). The quantized signal goes back to $f_{s,\text{ref}}$ with the same resampling method. It is trimmed to a common length, and 50 ms at each end is removed so that filter transients are excluded; the result is the reconstruction $\hat{x}[n]$. Since both signals are now at the native rate, the reference and the reconstruction share the same time grid.

Step 6 (perceptual estimate). For Series C we compare the reference and the reconstruction with ViSQOL v3 in audio mode [6]. It outputs a MOS-LQO score from 1 to 5 and expects 48 kHz input. The pop clips are at 44.1 kHz, so both of their signals are first converted to 48 kHz with the same sharper filter.

Step 7 (metrics and aggregation). Each metric is computed per clip and then summarized by its mean, standard deviation, and minimum across clips, both per content category and pooled over the configurations that all categories share. All results are exported to CSV. The operating points are chosen as described in Section 2.5.

Repeating the experiment only requires opening the notebook, running the material cell, keeping or editing the configuration cell, and running all cells. The notebook produces the raw and summary CSV files, the operating-point tables, the figures, and the environment log. No dither or other random processing is

used, so the pipeline is deterministic and a rerun with the same material and software versions reproduces every table and figure. We do not redistribute the audio. Speech and classical clips are downloaded from their public sources, but MUSDB18-HQ may only be used for non-commercial research and may not be redistributed [19], [20]. The notebook therefore reads the pop clips from the copy that every user downloads from the official Zenodo record, while the manifest fixes the exact tracks and offsets. Results are given as mean ± standard deviation across clips. We did not apply an inferential test, because each category has only eight clips taken in pairs from four sources and the clips are not independent. Differences are described together with their spread, and for the criteria we report how many clips pass.

2.5. Evaluation metrics

Consider a B-bit linear PCM signal with amplitude range [−1, 1). Its number of levels L and its quantization step Δ are

$$L = 2^B, \qquad \Delta = 2^{1-B} \tag{1}$$

and the quantizer Q_B and the quantization error e_q are

$$Q_B(x) = \Delta \,\mathrm{clip}\left(\mathrm{round}\left(\frac{x}{\Delta}\right), -2^{B-1}, 2^{B-1}-1\right) \tag{2}$$

$$e_q[n] = x_s[n] - Q_B(x_s[n]), \qquad \sigma_q^2 \approx \frac{\Delta^2}{12} \tag{3}$$

where $x_s[n]$ is the quantizer input and σ_q^2 is the quantization-noise power, under the common assumption that the error is spread uniformly over [−Δ/2, Δ/2] [15], [16].

The SNR is the ratio between the power of the reference and the power of the error:

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_n x_{\mathrm{ref}}^2[n]}{\sum_n (x_{\mathrm{ref}}[n]-\hat{x}[n])^2} \ \mathrm{dB} \tag{4}$$

where x_{ref} denotes the reference and $\hat{x}$ the reconstruction. SNR is a standard physical fidelity measure [6] and can be checked directly against quantization theory [15], [16], although it does not tell what is audible [5], [6].

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_{\mathrm{ref}}[n]-\hat{x}[n])^2} \tag{5}$$

where N is the number of samples left after trimming. We give RMSE in full-scale units and in dBFS. It describes the absolute error amplitude, which complements the relative view of SNR.

The compute the quantization SNR from the input and output of the quantizer and compare it with the uniform-noise prediction:

$$\mathrm{SNR}_q = 10\log_{10}\frac{\sum_n x_s^2[n]}{\sum_n e_q^2[n]}, \qquad \mathrm{SNR}_{q,\mathrm{pred}} = 10\log_{10}\frac{12P_x}{\Delta^2} \tag{6}$$

where P_x is the mean power of x_s. With $\Delta = 2^{1-B}$, the predicted SNR_q grows by about 6.02 dB for every added bit [15], [16] and equals 6.02 B + 10 $\log_{10}(3P_x)$. The comparison of measured and predicted values tells us where the uniform-noise assumption holds.

The estimate the power spectrum P_m at frequency f_m with the Welch method. The spectral centroid C and the roll-off frequency f_r (the frequency below which 95% of the spectral power lies) are

$$C = \frac{\sum_m f_m P_m}{\sum_m P_m} \tag{7}$$

$$f_r = \min_m f_m \quad \text{such that} \quad \sum_{k\le m} P_k \ge 0.95 \sum_k P_k \tag{8}$$

Both centroid and roll-off are standard descriptors in audio analysis [23], and a shift between reference and reconstruction indicates that high-frequency energy was lost or redistributed. The fraction of reference power above half of the target sampling rate is computed as well:

$$E_{\mathrm{loss}} = \frac{\sum_{f_m > f_s/2} P_m}{\sum_m P_m}, \qquad \mathrm{SNR}_{\mathrm{BL}} = -10\log_{10}E_{\mathrm{loss}} \tag{9}$$

$\mathrm{SNR}_{\mathrm{BL}}$ is what an ideal brick-wall band-limiter would achieve. In Series A we treat it as an approximate upper reference for the measured SNR, since orthogonal projection onto the band gives the smallest error of any band-limited signal.

For mono PCM, the theoretical data size in bytes and its relative difference from the real file size are

$$S_{\mathrm{th}} = \frac{f_s\, B\, C\, T}{8} \tag{10}$$

$$\delta = \frac{S_{\mathrm{act}} - S_{\mathrm{th}}}{S_{\mathrm{th}}} \times 100\% \tag{11}$$

where C = 1 is the channel count, T the duration in seconds, S_{act} the WAV file size reported by the file system, and δ the relative difference. We read the FLAC size S_{FLAC} the same way and express it as the

ratio $\rho = S_{FLAC}/S_{act}$. The storage ratio $R = S/S_{CD}$ relates each configuration to the CD format (44,100 Hz and 16 bit), which exists for all three categories. Table 2 gives an overview of the metrics and why they were chosen.

What a listener hears cannot be read from SNR [5], [6]. We therefore also scored Series C with ViSQOL v3 in audio mode [6]. ViSQOL compares spectrogram-based representations of the reference and the test signal and returns a MOS-LQO score between 1 and 5. Its audio mode comes from ViSQOLAudio, which was benchmarked against listening tests on music coded at low bit rates [25]. We applied the audio mode to speech as well as music. PESQ was left out because its standard versions cover only telephone-band and wideband speech up to 16 kHz [26] and cannot handle music or full-band material. Two identical signals score 4.73, the highest value of the audio mode. Since ViSQOL was built for codec artifacts [6], [25], its scores for plain requantization can only be used as an indication.

restrict the evaluation to objective measures computed with a standard model that anyone can obtain. Subjective listening tests, MUSHRA [27] for example, lie outside this scope, since the benchmark is meant to be rerun automatically by any reader. The ViSQOL values are model predictions (MOS-LQO, in the wording of ITU-T P.800.1 [28]) and not measured opinion scores, and anything we say about audibility depends on them.

We say a configuration meets a criterion when a metric reaches at least a threshold t, either for every clip (the minimum across clips is at least t) or for the mean across clips. For SNR we used 30, 40, 50, and 60 dB, corresponding to an error RMS of 3.2%, 1.0%, 0.32%, and 0.10% of the signal RMS. Forty decibels is a conventional engineering value and has no perceptual basis, which is why we vary it and also compare it with two ViSQOL thresholds. MOS-LQO scores are on the five-point absolute category rating scale of ITU-T P.800 [29], where 4 stands for “good” and 5 for “excellent”. A threshold of 4.0 is the bottom of the “good” category. The value 4.5 sits halfway between “good” and “excellent” and near the 4.73 ceiling of the audio mode, and we interpret it as near-transparent. Neither threshold was validated with listening tests in this study; both are reference points on the scale. For each criterion, the operating point is the configuration with the smallest mean file size (WAV or FLAC) that meets it, and ties go to the configuration with the higher worst-clip value. We also look at the quality–storage trade-off through Pareto dominance, where a configuration counts as efficient when no other configuration is both smaller and better in mean SNR. Pooled results only include configurations that all categories have (rates up to 44.1 kHz).

Table 2. Evaluation metrics and rationale

Metric	Rationale	Used for	Better when
SNR (dB)	Relative error power; comparable with quantization theory	RQ1–RQ4	Higher
RMSE	Absolute error amplitude in signal units	RQ1, RQ2	Lower
Quantization error	Isolates the effect of bit depth; tests the uniform-noise model	RQ2	Lower
Centroid, roll-off, E_{loss}	Reveals bandwidth loss and spectral redistribution	RQ1	Smaller shift
ViSQOL (MOS-LQO)	Perceptual estimate for speech and music; checks what SNR cannot show	RQ1, RQ3, RQ4	Higher
File size, δ, ρ	Storage cost in WAV and FLAC; validates the theoretical PCM estimate	RQ3	Smaller
Clips meeting a threshold	Per-clip check of an acceptance criterion	RQ4	More clips

3. RESULTS AND DISCUSSION

In this section we go through Series A, B, and C and then discuss them together. Results for both resampling filters are included. Fig. 2(a), Table 4, and Section 3.2 compare the two filters directly, and Section 3.7 lists the operating points for each of them; elsewhere, tables and values refer to the sharper filter unless noted. All values are means over the eight clips of a category, with ± giving the standard deviation across clips. Pooled values average all 24 clips and include configurations up to 44.1 kHz.

3.1. Effect of sampling rate (RQ1)

Series A is summarized in Table 3 and Fig. 2. The SNR rose with the sampling rate in all categories. At f_s = 8,000 Hz we measured 17.36 ± 4.31 dB for speech, 24.86 ± 4.72 dB for classical music, and 14.24 ± 3.27 dB for pop. At 32,000 Hz the values were 42.43 ± 3.70, 64.85 ± 20.30, and 30.36 ± 6.91 dB, and at 44,100 Hz speech reached 69.33 ± 2.93 dB and classical music 78.91 ± 23.72 dB (58.35 and 58.47 dB with the default filter, see Section 3.2). Without resampling, at the native rate, the error is zero. Classical music has a large spread at 32,000 Hz because the two Mendelssohn quartet excerpts reached 93.2 and 101.6 dB, while the remaining six clips were between 50.5 and 56.7 dB.

At low rates the SNR was set by the energy that the band limit removes. The fraction of reference power above $f_s/2$, E_{loss}, was 2.93%, 0.44%, and 4.49% at 8,000 Hz (speech, classical, pop), giving SNR_{BL} values of 17.44, 25.38, and 14.28 dB (Eq. (9)). With the sharper filter, every category mean stayed within

0.52 dB of SNR_{BL} for all rates up to 32,000 Hz, and within 2.11 dB (speech) and 0.48 dB (classical) at 44,100 Hz. Looking at single clips up to 32,000 Hz, the gap never exceeded 2.5 dB (Borodin, excerpt a, at 8,000 Hz). So the long-term spectrum of a source is already enough to estimate the SNR that a given rate allows (Section 3.8).

Among the three categories, pop reacted most strongly to a low sampling rate and classical music the least (Table 3). At 8,000 Hz the spectral centroid of pop dropped by −364 Hz and its roll-off by −2,147 Hz, against −215 Hz and −1,655 Hz for speech and −22 Hz and −51 Hz for classical music (Table 3; Fig. 2(c)). Cymbals, hi-hats, and synthesized sounds in the pop mixtures push their spectrograms up to about 21 kHz. Most of the power in the orchestral and chamber recordings, on the other hand, stays below about 2 kHz (mean roll-off 2,045 Hz). We think this comes from the spectra of the instruments themselves and from the loss of high frequencies due to the hall and the microphone distance.

Speech also lost more than classical music, which goes against the common idea that speech is a narrow-band signal. That idea fits telephone speech, but our recordings are full-band studio speech. Recorded anechoically and close to the microphone, they keep fricative and breath energy above 10 kHz (Fig. 2(c)); this energy is perceptually relevant [7] and disappears at low rates. Within the speech category the spread was large. Speaker p003, whose two clips have a roll-off near 7.8 kHz, lost 8.9% and 7.6% of the power and reached only 10.50 and 11.07 dB at 8,000 Hz, compared with 17.9–22.2 dB for the other speakers.

Clips that fall short are hidden by the mean. At 22,050 Hz every speech clip and every pop clip was under 40 dB. Even at 32,000 Hz, seven of the eight pop clips (22.5–45.1 dB) and three of the eight speech clips (36.9–39.7 dB) did not reach 40 dB, while no classical clip was below 40 dB from 22,050 Hz upward (Section 3.7).

Per kilohertz of added rate, speech gained 0.55, 1.31, 1.28, and 2.22 dB/kHz between 8,000 and 16,000, 16,000 and 22,050, 22,050 and 32,000, and 32,000 and 44,100 Hz. Classical music gained 1.88, 2.12, 1.21, and 1.16 dB/kHz over the same steps, and pop 0.44, 0.60, and 0.90 dB/kHz up to 32,000 Hz. Taking 1 dB/kHz as the threshold, neither speech nor classical music showed diminishing returns above 16,000 Hz. The diminishing returns for music above 22,050 Hz (0.51 and 0.37 dB/kHz) seen with the default filter are a filter effect (Section 3.2).

Table 3. Series A, effect of sampling rate with the sharper resampling filter (mean of eight clips per category; no quantization). There is no resampling at the native rate, so pop has no 44,100 Hz row and speech and classical music have no 48,000 Hz row. ΔC and Δf_r are the shifts of the spectral centroid and of the 95% roll-off

Category	f_s (Hz)	SNR (dB), mean ± SD	Min SNR (dB)	RMSE ($\times 10^{-3}$)	E_{loss} (%)	SNR_{BL} (dB)	ΔC (Hz)	Δf_r (Hz)
Speech	8,000	17.36 ± 4.31	10.50	14.56	2.932	17.44	−214.9	−1655
	16,000	21.73 ± 5.25	13.27	9.63	1.371	21.89	−128.6	−1332
	22,050	29.68 ± 4.62	22.69	3.74	0.167	29.94	−20.1	−63.0
	32,000	42.43 ± 3.70	36.89	0.80	0.007	42.69	1.5	0.0
	44,100	69.33 ± 2.93	65.84	0.03	<0.001	71.44	3.0	1.5
Classical	8,000	24.86 ± 4.72	20.22	7.99	0.443	25.38	−21.8	−51.3
	16,000	39.92 ± 6.03	32.99	1.46	0.017	40.24	−1.7	−4.4
	22,050	52.76 ± 10.65	43.09	0.45	0.002	53.03	−0.2	−4.4
	32,000	64.85 ± 20.30	50.54	0.21	<0.001	65.29	0.0	−4.4
	44,100	78.91 ± 23.72	63.79	0.05	<0.001	79.39	0.1	−4.4
Pop	8,000	14.24 ± 3.27	11.07	30.70	4.488	14.28	−363.7	−2147
	16,000	17.73 ± 4.60	14.70	21.33	2.191	17.79	−241.4	−1649
	22,050	21.36 ± 4.46	17.63	13.98	0.941	21.47	−132.6	−1116
	32,000	30.36 ± 6.91	22.52	5.59	0.174	30.58	−34.6	−300

3.2. Comparison of the default and the sharper resampling filter

This research ran Series A and C with two filters to find out how much of the outcome is decided by the filter. The default filter of resample_poly (Kaiser window, β = 5, half-length $10 \cdot \max(p, q)$) is what a user gets without tuning, and it attenuates components just under $f_s/2$ in both resampling steps. Our sharper filter is a Kaiser FIR five times as long, with β = 12. It stands for a high-quality offline converter in general and was not modeled on a specific product. The two filters are compared in Table 4 and Fig. 2(a).

The choice of filter hardly matters at low rates; at 8,000 Hz the SNR changed by only +0.10, +0.30, and +0.06 dB (speech, classical, pop). The difference grew with the rate, to 0.87 dB (speech), 10.83 dB

(classical), and 0.56 dB (pop) at 32,000 Hz, and to 10.99 dB (speech) and 20.44 dB (classical) at 44,100 Hz. With the default filter, speech and classical music both leveled off near 58 dB at 44,100 Hz (58.35 and 58.47 dB), and their gap to $\mathrm{SNR}_{\mathrm{BL}}$ widened to 13.1 and 20.9 dB even though E_{loss} was negligible. Both the plateau and the apparent equality of speech and classical music at this rate come from the filter, not from the sampling rate.

Whether a bit depth above 16 helps after resampling also depends on the filter. With the default filter, 24 bit improved on 16 bit by at most 0.03 dB at any rate, since the quantization SNR (above 80 dB) is far higher than the resampling SNR (at most 58.5 dB). With the sharper filter, 24 bit added 0.11 dB (classical, 22,050 Hz), 4.06 dB (classical, 32,000 Hz), 8.97 dB (classical, 44,100 Hz), and 0.44 dB (speech, 44,100 Hz), while pop gained nothing at rates up to 32,000 Hz, where band-limiting dominates. More than 16 bits only pays off when the resampling error itself is below about 60 dB, as happens with band-limited content and a steep filter.

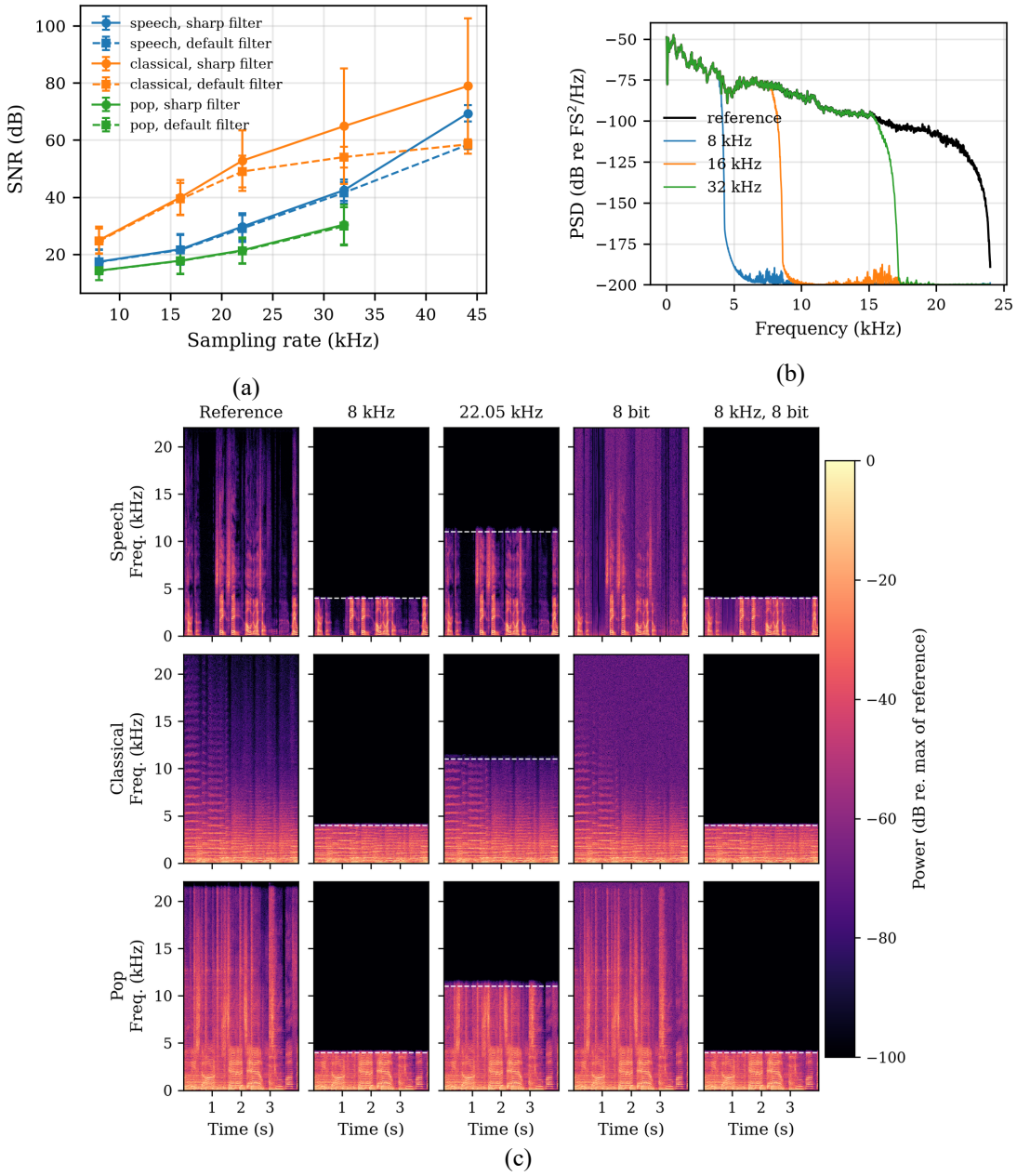


Figure 2. Effect of sampling rate. (a) SNR against sampling rate for the sharper (solid) and default (dashed) filter (mean ± SD, eight clips per category); (b) power spectral density in dB re full scale squared per hertz for the reference and for reconstructions band-limited to 8, 16, and 32 kHz, speech clip p001, sentence 1; (c) spectrograms (central 4 s) of one clip per category (speech, p001 sentence 1; classical, Beethoven excerpt a; pop, Al James excerpt a) showing the reference, band-limiting to 8 kHz and to 22.05 kHz, 8-bit quantization, and 8 kHz with 8 bit; dashed lines mark $f_s/2$

Table 4. Series A with the default and the sharper resampling filter (mean of eight clips per category; no quantization), together with the SNR gain of 24 bit over 16 bit after resampling (Series C, last two columns)

Category	f_s (Hz)	SNR, default (dB)	SNR, sharper (dB)	SNR_{BL} − SNR, default (dB)	SNR_{BL} − SNR, sharper (dB)	24 − 16 bit, default (dB)	24 − 16 bit, sharper (dB)
Speech	8,000	17.26	17.36	0.19	0.08	0.00	0.00
	16,000	21.50	21.73	0.39	0.15	0.00	0.00
	22,050	29.01	29.68	0.93	0.26	0.00	0.00
	32,000	41.56	42.43	1.13	0.26	0.00	0.00
	44,100	58.35	69.33	13.09	2.11	0.03	0.44
Classical	8,000	24.56	24.86	0.82	0.52	0.00	0.00
	16,000	39.31	39.92	0.93	0.32	0.00	0.00
	22,050	48.97	52.76	4.06	0.27	0.01	0.11
	32,000	54.02	64.85	11.27	0.44	0.01	4.06
	44,100	58.47	78.91	20.92	0.48	0.02	8.97
Pop	8,000	14.18	14.24	0.10	0.04	0.00	0.00
	16,000	17.64	17.73	0.15	0.06	0.00	0.00
	22,050	21.14	21.36	0.34	0.12	0.00	0.00
	32,000	29.80	30.36	0.78	0.22	0.00	0.00

Looking only at the default filter would have led to three statements that the comparison now corrects. “No gain from 24 bit” holds for the default filter and for speech and pop, but classical music with the sharper filter is an exception. The 58 dB plateau turns out to be a filter artifact. As for the operating points, SNR ≥ 30 dB, SNR ≥ 40 dB, and ViSQOL ≥ 4.5 on every clip give the same result with both filters (Section 3.7). They differ for SNR ≥ 50 dB, where classical music needs 44,100 Hz with the default filter and 32,000 Hz with the sharper one, and for SNR ≥ 60 dB, where the default filter requires 48,000 Hz for classical music and speech and the sharper filter 44,100 Hz. For ViSQOL ≥ 4.0, the pop result hinges on a difference of 0.01 at the threshold, because the worst pop clip at 32,000 Hz scored 4.00 with the default filter and 3.99 with the sharper one. Our conclusions about low rates hold for both filters, whereas the absolute SNR at high rates mainly characterizes the filter. One open point is the gap of 2.11 dB between SNR and SNR_{BL} for speech at 44,100 Hz with the sharper filter, which we did not investigate further.

3.3. Effect of bit depth (RQ2)

Series B is shown in Table 5 and Fig. 3. Pooled over 24 clips, the measured quantization SNR was 34.52 dB at 8 bit, 82.68 dB at 16 bit, and 130.85 dB at 24 bit. Fitting SNR_q against B by least squares gave a slope of 6.020 dB per bit between 8 and 16 bit (6.02, 6.02, and 6.021 dB per bit for speech, classical music, and pop) and 6.006 dB per bit across the full 4–24-bit range, close to the 6.02 dB per bit predicted by the uniform-noise model [15], [16]. Measured and predicted values agreed within 0.01 dB in every category from 8 bit upward, and within 0.03 dB at 6 bit.

At 4 bit there is a deviation of +0.58 dB (pooled). The step $\Delta = 0.125$ is large relative to the signal, and the zero bin of the mid-tread quantizer ($|x| < \Delta/2$) catches 69% of the speech samples, 52% of the classical samples, and 39% of the pop samples. For those samples the error is simply the sample value, which is smaller than the uniform model assumes; we measured an error RMS of 0.0339 against $\Delta/\sqrt{12} = 0.0361$. Speech showed the biggest effect (+1.29 dB, versus +0.32 dB for classical music and +0.13 dB for pop), as it has more pauses and quiet passages. Some peaks also went above the highest 4-bit code (0.875), and on average 1.6 samples per clip were clipped. Clipping adds error instead of removing it and touches only about one sample in 300,000, so it cannot explain the deviation. From 6 bit upward nothing was clipped.

Table 5. Series B, effect of bit depth at the native sampling rate (mean of all 24 clips). The deviation column gives SNR_q measured minus predicted (dB) for speech / classical / pop

B (bit)	Δ	SNR_q (dB)	$SNR_{q,pred}$ (dB)	Measured − predicted (dB)	RMSE (dBFS)	Quantization error RMS, measured / $\Delta/\sqrt{12}$
4	2^{1-4}	11.02	10.44	+1.29 / +0.32 / +0.13	−29.43	3.389 / 3.608 ($\times10^{-2}$)
6	2^{1-6}	22.49	22.48	+0.03 / 0.00 / +0.01	−40.91	9.007 / 9.021 ($\times10^{-3}$)
8	2^{1-8}	34.52	34.52	+0.01 / 0.00 / 0.00	−52.94	2.255 / 2.255 ($\times10^{-3}$)
10	2^{1-10}	46.56	46.56	+0.01 / 0.00 / 0.00	−64.98	5.637 / 5.638 ($\times10^{-4}$)
12	2^{1-12}	58.60	58.60	0.00 / 0.00 / 0.00	−77.02	1.410 / 1.410 ($\times10^{-4}$)
14	2^{1-14}	70.64	70.64	0.00 / 0.00 / 0.00	−89.06	3.524 / 3.524 ($\times10^{-5}$)
16	2^{1-16}	82.68	82.68	0.00 / 0.00 / 0.00	−101.10	8.811 / 8.810 ($\times10^{-6}$)
20	2^{1-20}	106.77	106.77	0.00 / 0.00 / 0.00	−125.18	5.506 / 5.506 ($\times10^{-7}$)
24	2^{1-24}	130.85	130.85	0.00 / 0.00 / 0.00	−149.26	3.442 / 3.441 ($\times10^{-8}$)

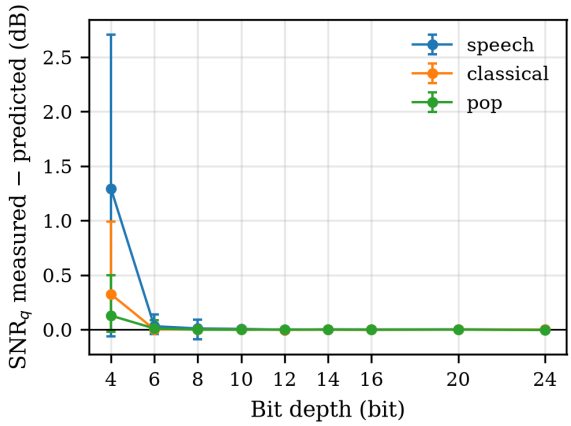


Figure 3. Quantization SNR measured minus the value predicted by the uniform-noise model, Eq. (6), against bit depth at the native rate (mean ± SD, eight clips per category)

From 6 to 24 bit, the SNR_q of speech was always about 2.5 dB lower than that of classical music, and pop was about 1.7 dB higher (80.43, 82.98, and 84.64 dB at 16 bit). Peak normalization gives each category a different mean power P_x for the same quantization noise (Eq. (6)), which explains the offset. RMSE fell from 3.39×10^{-2} at 4 bit to 3.44×10^{-8} (full scale) at 24 bit, i.e., from −29.4 to −149.3 dBFS, halving with every added bit. Fig. 2(c) also makes the quantization noise visible. At 8 bit it spreads over the whole band at an almost constant level and covers the quiet high-frequency regions of speech and classical music, while in pop the denser high-frequency content makes it harder to see.

3.4. Joint effect of sampling rate and bit depth (RQ3)

In Fig. 4, the mean SNR of every Series C configuration is plotted per category. Pooled means are left out because the categories were tested at different sets of rates. The SNR went from 14.2 dB (pop, 8,000 Hz) up to 132.8 dB (pop, 44,100 Hz, 24 bit). The top values appear only at the native rate, where no resampling filter is involved and the bit depth alone sets the SNR (83.0 and 131.2 dB for classical music at 48,000 Hz with 16 and 24 bit, 80.4 and 128.6 dB for speech, and 84.6 and 132.8 dB for pop at 44,100 Hz).

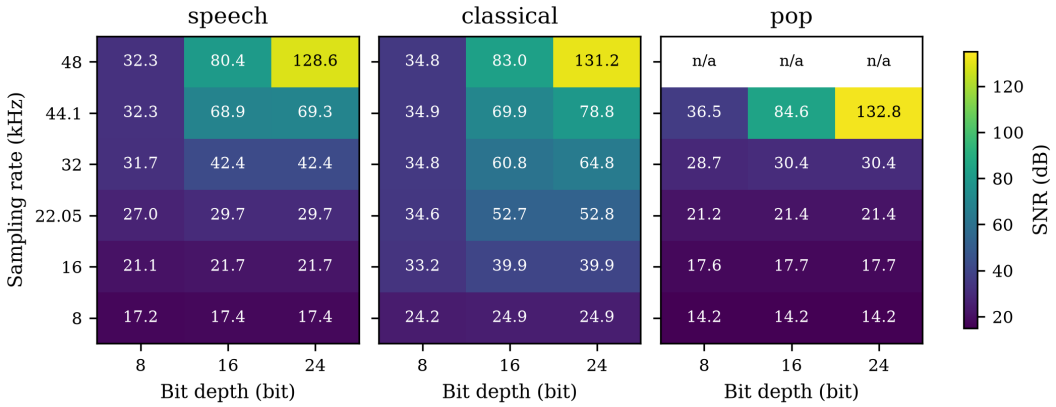


Figure 4. Mean SNR across sampling rate and bit depth (Series C, sharper filter) per content category; n/a: 48 kHz is above the native rate of the pop clips (44.1 kHz)

The results fall into two regions. At 16 and 24 bit, band-limiting was the limit for every resampled configuration up to 16,000 Hz for classical music and up to 32,000 Hz for speech and pop. There, the SNR matched the Series A value within 0.01 dB, and 24 bit brought no gain over 16 bit. With 8 bit, the quantizer took over as soon as the band-limiting SNR went above the 8-bit quantization SNR (34.8 dB for classical music, 32.3 dB for speech, and 36.5 dB for pop). From that point the SNR saturated, at 34.6–34.9 dB for classical music from 22,050 Hz and at 31.7–32.3 dB for speech from 32,000 Hz, and raising the rate gave nothing (Fig. 4). At 16 bit the quantizer only matters if the band-limiting SNR is higher than about 60 dB. This was the case for classical music at 32,000 and 44,100 Hz (60.8 and 69.9 dB instead of 64.9 and 78.9

dB), and there 24 bit does help (Section 3.2). Assuming that the resampling error (Series A) and the quantization error are uncorrelated, the total SNR can be approximated as

$$\mathrm{SNR}_{\mathrm{pred}} = -10\log_{10}\left(10^{-\mathrm{SNR_A}/10} + 10^{-\mathrm{SNR}_q/10}\right) \quad (12)$$

In this equation, $\mathrm{SNR_A}$ is the SNR measured in Series A at the same sampling rate and with the same filter (band-limiting plus resampling-filter error, before quantization), as opposed to the ideal bound $\mathrm{SNR_{BL}}$ of Eq. (9), and $\mathrm{SNR_q}$ is the quantization SNR of Eq. (6). For the sharper filter the prediction differed from the measured SNR by 0.007 dB on average (mean absolute difference 0.015 dB, maximum 0.22 dB over 408 clip–configuration pairs), and for the default filter by 0.017, 0.024, and 0.22 dB. Differences this small support the additive-error approximation. Of the resampled configurations, 8 bit showed the largest systematic deviation: the measured SNR was 0.02 dB above the prediction on average (at most 0.19 dB). A likely cause is that the reconstruction filter takes out the part of the white quantization noise that falls in its transition band below $f_s/2$. The 0.22 dB maximum was found at the native rate, most likely because edge trimming is applied to the measured SNR but not to $\mathrm{SNR_q}$.

3.5. Storage requirements: WAV and FLAC

Table 6 puts the theoretical PCM size from Eq. (10) next to the actual WAV size. All clips are equally long, so clips with the same sampling rate and bit depth gave identical file sizes. The relative difference δ, Eq. (11), went from 0.0031% (48,000 Hz, 24 bit) to 0.055% (8,000 Hz, 8 bit). In absolute terms the difference was always exactly 44 bytes, the size of the canonical WAV header (a 12-byte RIFF descriptor, a 24-byte format chunk, and an 8-byte data-chunk header), and its relative weight drops for longer clips or higher data rates. We read every WAV and FLAC file back and found no difference from the signal in memory, confirming lossless storage. Moving from 8 to 16 bit made the actual file 99.95–99.99% larger, and moving from f_s = 8,000 to 48,000 Hz made it 499.7% larger at 8 bit, in line with the linear scaling of Eq. (10). Against the CD configuration (44,100 Hz and 16 bit; 882,044 bytes or 0.841 MiB), the storage ratio R varied from 0.091 (8,000 Hz, 8 bit) to 1.633 (48,000 Hz, 24 bit).

Every configuration became smaller with FLAC (Table 7). Averaged over the sampling rates, a FLAC file was 0.38, 0.61, and 0.74 of the WAV size at 8, 16, and 24 bit. Speech compressed best (0.50 at 16 bit, range 0.36–0.64 over clips) and pop worst (0.74, range 0.60–0.85), with classical music in between (0.60,

Table 6. Theoretical (th) and actual (act) WAV file size of a 10-s mono clip (bytes; the same for all clips)

f_s (Hz)	8 bit th	8 bit act	16 bit th	16 bit act	24 bit th	24 bit act
8,000	80,000	80,044	160,000	160,044	240,000	240,044
16,000	160,000	160,044	320,000	320,044	480,000	480,044
22,050	220,500	220,544	441,000	441,044	661,500	661,544
32,000	320,000	320,044	640,000	640,044	960,000	960,044
44,100	441,000	441,044	882,000	882,044	1,323,000	1,323,044
48,000	480,000	480,044	960,000	960,044	1,440,000	1,440,044

Table 7. FLAC size relative to WAV size (mean ratio of eight clips) at 8, 16, and 24 bit, with the mean FLAC size at 16 bit

Category	f_s (Hz)	ρ, 8 bit	ρ, 16 bit	ρ, 24 bit	FLAC, 16 bit (MiB)
Speech	8,000	0.358	0.593	0.729	0.091
	16,000	0.308	0.540	0.693	0.165
	22,050	0.282	0.515	0.677	0.217
	32,000	0.249	0.477	0.651	0.291
	44,100	0.222	0.434	0.622	0.365
	48,000	0.216	0.418	0.611	0.383
Classical	8,000	0.579	0.782	0.855	0.119
	16,000	0.453	0.697	0.798	0.213
	22,050	0.385	0.636	0.757	0.267
	32,000	0.316	0.549	0.699	0.335
	44,100	0.273	0.477	0.644	0.402
	48,000	0.263	0.459	0.629	0.421
Pop	8,000	0.603	0.798	0.866	0.122
	16,000	0.542	0.765	0.843	0.233
	22,050	0.517	0.749	0.833	0.315
	32,000	0.478	0.722	0.815	0.441
	44,100	0.426	0.681	0.788	0.573

range 0.32–0.87); our guess is that the dense, loud pop mixtures leave the predictor less redundancy to exploit. Higher sampling rates also gave lower ratios, for example 0.78 at 8,000 Hz and 0.46 at 48,000 Hz for classical music at 16 bit, probably because the additional samples of a band-limited signal are easy to predict. Going from 32,000 to 44,100 Hz at 16 bit thus costs 20% (classical), 25% (speech), and 30% (pop) more storage in FLAC, compared with 38% in WAV.

3.6. Perceptual check with ViSQOL

The ViSQOL scores of Series C are shown in Fig. 5. At 16 bit, the mean score climbed from 2.28, 2.07, and 1.51 at 8,000 Hz (speech, classical, pop) to 4.34, 4.40, and 4.28 at 32,000 Hz, and reached 4.73 at 44,100 Hz, which is the score of an identical pair and the ceiling of the audio mode. The worst clip at 32,000 Hz had 4.25 (speech), 4.09 (classical), and 3.99 (pop). Scores at 16 and 24 bit were nearly identical, matching what we saw for SNR.

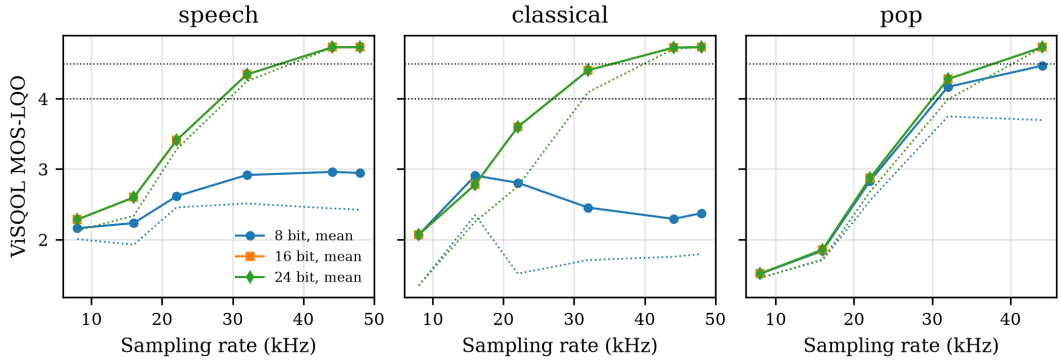


Figure 5. ViSQOL score (MOS-LQO) against sampling rate (Series C, sharper filter), with the mean of eight clips (markers) and the worst clip (dotted lines) at 8, 16, and 24 bit; dotted horizontal lines are drawn at 4.0 and 4.5

SNR and ViSQOL were closely but not perfectly related, with Spearman correlations over all Series C results of 0.93 for speech, 0.86 for classical music, 0.91 for pop, and 0.82 for the pooled data. A good perceptual score was not guaranteed at 40 dB SNR. Among the 130 clip–configuration pairs with SNR ≥ 40 dB, 18 scored below 4.0 on ViSQOL (lowest 2.37) and 42 below 4.5, and among the 106 pairs with SNR ≥ 50 dB, 6 were below 4.0 and 18 below 4.5. Borodin excerpt a at 16,000 Hz is the clearest example, with 51.3 dB SNR but a ViSQOL score of only 2.37. Less than 0.005% of its power lies above 8 kHz, so the missing band barely affects the SNR, while ViSQOL compares the spectrogram band by band and penalizes it. Every pair with SNR ≥ 60 dB (88 in total) scored at least 4.63. The round value of 40 dB is not a perceptual limit, and for this reason Section 3.7 reports the criteria for both metrics.

Care is needed with the 8-bit results. For classical music at 8 bit the mean score went down from 2.91 at 16,000 Hz to 2.29 at 44,100 Hz. Our explanation is that white quantization noise, which a lower rate cuts off above $f_s/2$, fills the quiet high-frequency bands that ViSQOL compares. Speech never scored above

Table 8. Per-clip check at 16 bit (sharper filter), giving the number of clips (out of eight) that meet each criterion and the worst clip. * native rate (no resampling)

Category	f_s (Hz)	Clips with SNR ≥ 40 dB	Worst SNR (dB)	Clips with ViSQOL ≥ 4.0	Clips with ViSQOL ≥ 4.5	Worst ViSQOL
Speech	8,000	0	10.5	0	0	2.13
	16,000	0	13.3	0	0	2.34
	22,050	0	22.7	0	0	3.28
	32,000	5	36.9	8	0	4.25
	44,100	8	65.7	8	8	4.72
	48,000*	8	76.5	8	8	4.73
Classical	8,000	0	20.2	0	0	1.35
	16,000	3	33.0	0	0	2.25
	22,050	8	43.1	2	2	2.76
	32,000	8	50.5	8	2	4.09
	44,100	8	63.7	8	8	4.71
	48,000*	8	79.9	8	8	4.73
Pop	8,000	0	11.1	0	0	1.45
	16,000	0	14.7	0	0	1.72
	22,050	0	17.6	0	0	2.68
	32,000	1	22.5	7	0	3.99
	44,100*	8	80.8	8	8	4.73

2.96 at 8 bit, at any rate. Pop at its native rate reached 4.47 (worst clip 3.70), presumably because the dense high-frequency content of the mixtures masks the noise. Note that ViSQOL was designed for codec artifacts [6] and that we did not validate it for requantization without dither.

3.7. Quality–storage trade-off and operating points (RQ3, RQ4)

Fig. 6 plots mean SNR against file size for Series C and marks the Pareto-efficient configurations of each category for WAV and FLAC sizes. In WAV, speech had efficient configurations at 8,000 to 44,100 Hz with 8 bit, at 32,000, 44,100, and 48,000 Hz with 16 bit, and at 48,000 Hz with 24 bit. Classical music had them at 8,000 to 22,050 Hz with 8 bit, 16,000 to 48,000 Hz with 16 bit, and 48,000 Hz with 24 bit, and pop at 8,000 to 44,100 Hz with 8 bit and at 44,100 Hz with 16 and 24 bit. The only efficient 24-bit configurations were at the native rate, where no resampling is done. Every resampled 24-bit configuration was dominated: it needs 50% more storage than the 16-bit file at the same rate, and for classical music the 16-bit file at the next higher rate was better anyway (32,000 Hz with 24 bit gave 64.9 dB at 0.92 MiB, whereas 44,100 Hz with 16 bit gave 69.9 dB at 0.84 MiB). The FLAC fronts contain mostly the same configurations, about 30–60% smaller, along with a few extra 8-bit configurations at high rates.

The operating points are given in Table 9. Requiring SNR ≥ 40 dB for every clip, the smallest configuration was 44,100 Hz with 16 bit for speech (0.84 MiB in WAV, 0.36 MiB in FLAC) and pop (0.84 and 0.57 MiB), and 22,050 Hz with 16 bit for classical music (0.42 and 0.27 MiB). Had the same threshold been applied to the mean, the choice would have been 32,000 Hz with 16 bit for speech, and also 32,000 Hz with 16 bit (0.61 MiB in WAV) for the pooled categories instead of 44,100 Hz with 16 bit (0.84 MiB). This setting saves 27%, but its worst clip, a pop excerpt, only reached 22.5 dB, and just 14 of the 24 clips passed 40 dB. For mixed material, a criterion on the mean is too optimistic.

Under ViSQOL ≥ 4.0 for every clip, the operating point was 32,000 Hz with 16 bit for speech and classical music and 44,100 Hz with 16 bit for pop (the worst pop clip had 3.99 at 32,000 Hz). Under ViSQOL ≥ 4.5 it was 44,100 Hz with 16 bit for all categories. Of the tested configurations, the CD format was therefore the smallest one that met SNR ≥ 60 dB and ViSQOL ≥ 4.5 for all 24 clips. No criterion in Table 9 called for 48,000 Hz or 24 bit.

The default filter gave the same operating points as Table 9 for SNR ≥ 30 dB, SNR ≥ 40 dB, and ViSQOL ≥ 4.5 on every clip, including 44,100 Hz with 16 bit for speech and pop and 22,050 Hz with 16 bit

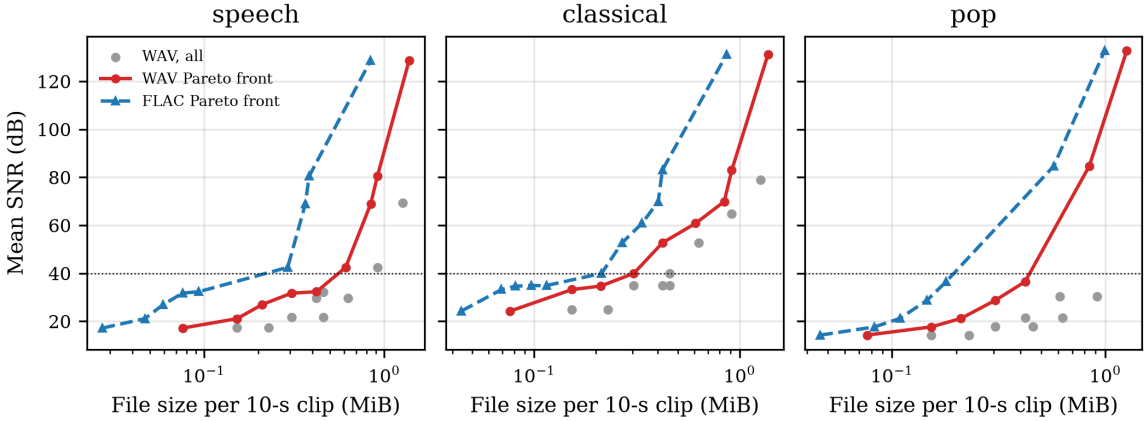


Figure 6. Quality–storage trade-off of Series C (sharper filter, mean SNR of the eight clips in each category) with the Pareto-efficient configurations for WAV and FLAC sizes; the dotted line is at 40 dB, and grey points show all WAV configurations

for classical music at 40 dB. For SNR ≥ 50 dB, classical music needed 44,100 Hz with 16 bit (32,000 Hz with the sharper filter). For SNR ≥ 60 dB, speech and classical music needed 48,000 Hz with 16 bit, and since the pop clips stop at 44,100 Hz, the pooled data had no common configuration. Under ViSQOL ≥ 4.0 the default filter chose 32,000 Hz with 16 bit for all three categories, with the worst pop clip at 4.00 instead of 3.99 (Section 3.2). Across all clips, the CD configuration had a worst-clip SNR of 56.5 dB and a ViSQOL of 4.58 with the default filter, against 63.7 dB and 4.71 with the sharper one.

3.8. Guidance for specific applications

In Table 10 the operating points are turned into recommendations for common uses; the number in brackets is the WAV size relative to the CD format. Many recorders save 48,000 Hz, 24-bit WAV by default, which takes 1.37 MiB for a 10-s mono clip. Compared with that, 32,000 Hz with 16 bit for speech is 56% smaller as WAV and 79% smaller as FLAC and still meets ViSQOL ≥ 4.0 on every clip. The CD format, as a

Table 9. Smallest tested configuration meeting each criterion (sharper filter; configuration and, in brackets, WAV / FLAC size in MiB per 10-s mono clip). "All" pools the 24 clips at rates up to 44,100 Hz

Criterion	Speech	Classical	Pop	All
SNR ≥ 30 dB, every clip	32 kHz / 16 bit (0.61 / 0.29)	16 kHz / 8 bit (0.15 / 0.07)	44.1 kHz / 8 bit (0.42 / 0.18)	44.1 kHz / 16 bit (0.84 / 0.45)
SNR ≥ 40 dB, every clip	44.1 kHz / 16 bit (0.84 / 0.36)	22.05 kHz / 16 bit (0.42 / 0.27)	44.1 kHz / 16 bit (0.84 / 0.57)	44.1 kHz / 16 bit (0.84 / 0.45)
SNR ≥ 50 dB, every clip	44.1 kHz / 16 bit (0.84 / 0.36)	32 kHz / 16 bit (0.61 / 0.33)	44.1 kHz / 16 bit (0.84 / 0.57)	44.1 kHz / 16 bit (0.84 / 0.45)
SNR ≥ 60 dB, every clip	44.1 kHz / 16 bit (0.84 / 0.36)	44.1 kHz / 16 bit (0.84 / 0.40)	44.1 kHz / 16 bit (0.84 / 0.57)	44.1 kHz / 16 bit (0.84 / 0.45)
ViSQOL ≥ 4.0, every clip	32 kHz / 16 bit (0.61 / 0.29)	32 kHz / 16 bit (0.61 / 0.33)	44.1 kHz / 16 bit (0.84 / 0.57)	44.1 kHz / 16 bit (0.84 / 0.45)
ViSQOL ≥ 4.5, every clip	44.1 kHz / 16 bit (0.84 / 0.36)	44.1 kHz / 16 bit (0.84 / 0.40)	44.1 kHz / 16 bit (0.84 / 0.57)	44.1 kHz / 16 bit (0.84 / 0.45)
SNR ≥ 40 dB, mean only	32 kHz / 16 bit (0.61 / 0.29)	22.05 kHz / 16 bit (0.42 / 0.27)	44.1 kHz / 16 bit (0.84 / 0.57)	32 kHz / 16 bit (0.61 / 0.36)

general-purpose choice, is 39% smaller as WAV and 67% smaller as FLAC (mean over all categories) and meets SNR ≥ 60 dB and ViSQOL ≥ 4.5 on every clip.

For other material, the operating point can be found with the five-step procedure below, which follows the logic of this benchmark without the full factorial design. Every step uses quantities defined in Section 2.5 and calculated by the published notebook, so other researchers can apply it to their own data. (1) Fix the criterion: the metric, the threshold, and whether it must hold for every item or for the mean. In our results this choice alone moved the answer by one standard sampling rate. (2) Pick the sampling rate from the spectrum. Compute the fraction of power above $f_s/2$ for representative items and the bound $\mathrm{SNR}_{\mathrm{BL}} = -10 \log_{10} E_{\mathrm{loss}}$ (Eq. (9)), then take the lowest rate at which the worst item still reaches the target. With the sharper filter, this rule matched the full experiment in all 12 cases (3 categories × 4 SNR thresholds).

Table 10. Recommended configuration per application (sharper filter; WAV size in MiB per 10-s mono clip with the ratio to the CD format in brackets; mean FLAC size in MiB)

Use	Content	Criterion	Configuration	WAV, MiB (× CD)	FLAC, MiB
Voice over a narrow link, intelligibility only (not measured here)	Speech	none met: SNR 17.4 dB, ViSQOL 2.28	8 kHz / 16 bit	0.15 (0.18)	0.09
Speech archive, podcast, lecture	Speech	ViSQOL ≥ 4.0, every clip	32 kHz / 16 bit	0.61 (0.73)	0.29
Speech master, every speaker	Speech	SNR ≥ 40 dB and ViSQOL ≥ 4.5, every clip	44.1 kHz / 16 bit	0.84 (1.00)	0.36
Orchestral or chamber music, streaming or archive	Classical	ViSQOL ≥ 4.0, every clip	32 kHz / 16 bit	0.61 (0.73)	0.33
Pop or rock music, streaming or archive	Pop	ViSQOL ≥ 4.5, every clip	44.1 kHz / 16 bit	0.84 (1.00)	0.57
Mixed library, one setting	All	SNR ≥ 60 dB and ViSQOL ≥ 4.5, every clip	44.1 kHz / 16 bit	0.84 (1.00)	0.45
Classical music judged by SNR ≥ 40 dB only (not recommended)	Classical	SNR ≥ 40 dB met, but worst ViSQOL 2.76	22.05 kHz / 16 bit	0.42 (0.50)	0.27

This check was in-sample, though, and a 1 dB margin would have picked a higher rate in one case. The rule also assumes a steep converter, since with the default filter the measured SNR can be up to 21 dB below the bound. (3) Pick the bit depth from Eq. (6), $\mathrm{SNR}_q = 6.02\,B + 10 \log_{10}(3P_x)$. In our data 16 bit gives 80–85 dB, which is enough unless the band-limited SNR goes above about 60 dB (Eq. (12) then shows how much is lost), and 8 bit is only good for targets of about 30 dB or lower because it caps the SNR at 32–36 dB. (4) Check the result on a few representative items with a perceptual estimator, since 40 dB SNR did not always mean a ViSQOL score of 4.0 (Section 3.6). (5) Choose WAV or FLAC, depending on which size matters for the application.

3.9. Relation to previous studies and limitations

Our measured slope of SNR_q versus bit depth can be compared with the quantization theory used in [2], [15], [16]. It was 6.02 dB per bit between 8 and 16 bit and 6.01 dB per bit across the full range, consistent with the theoretical 6.02 dB per bit. For a compressed-sensing sub-Nyquist system, Zhao et al. [2] derived that high SNR favors resolution and low SNR favors sampling rate. Ordinary audio files do not allow

a direct test of this, but Fig. 4 shows a similar pattern in which the limiting parameter changes across the grid: at low rates only the sampling rate matters, and once the band-limiting error is small, the bit depth sets the ceiling. Ramya Devi and Pugazhenthi [1] worked on a different task (the SNR of steganographic embedding), and their numbers cannot be compared directly with ours. The direction of the change was the same, though, since in our data the SNR increased monotonically with the rate in all three categories.

All metrics in this study are signal-level or objective estimates and cannot say with certainty whether a difference is audible. We can neither confirm nor reject the listening results summarized by Reiss [3]; what we provide is the size of the physical differences. After resampling to 44,100 Hz, 16 bit reached 68.9 dB (speech) and 69.9 dB (classical), and 24 bit added 0.44 and 8.97 dB, while at the native rate 16 and 24 bit gave quantization errors of −101.1 and −149.3 dBFS RMS. The energy lost at low rates (2.93% for speech, 0.44% for classical music, and 4.49% for pop at 8,000 Hz) gives an idea of how much content a restoration model such as AudioSR [4] or the bandwidth-extension methods in [8]–[10] would have to regenerate, and that amount is clearly larger for pop and full-band speech than for orchestral music.

Our results should be read with several limitations in mind. The resampling filter influences every result at high rates (Section 3.2), and the two filters we tested do not represent all converters. The number of sources is small (four speakers, four classical recordings, and four pop tracks, with clips taken in pairs). Because the pop tracks were selected for their high-frequency content, none of the categories is a representative sample of its genre, and other material such as electronic music with energy up to 20 kHz, solo instruments, or noisy field recordings was not tested. The lossless status of the sources is based on repository documentation and a spectral heuristic; neither can completely exclude lossy processing before release, although no such step is documented. The reference is itself a digital recording with its own bit depth and noise floor, so when B exceeds the native depth (16 bit for pop), a higher SNR only reflects a closer copy of the reference, noise included, and no extra information about the original sound. We did not use dither or noise shaping, which can change the character of the quantization error at low bit depth. SNR and RMSE are waveform-matching metrics, and ViSQOL, designed for codec artifacts, was not validated here for requantization or resampling. Since the study is limited to objective evaluation by design (Section 2.5), its thresholds are reference values on the MOS scale, and any claim about audibility needs listening tests [3], [27]. FLAC sizes depend on the encoder and its compression level (we used the libsndfile default), and headers and container rules make the actual file size differ from the theoretical PCM size; bit depths without a native container (for example 12 bit) were only analyzed for quantization error. Lastly, all clips were mono and 10 s long, library defaults were used apart from the sharper filter, and other software versions may give different results.

4. CONCLUSION

We presented a reproducible empirical benchmark of how sampling rate and bit depth affect audio quality and storage for speech, classical music, and pop music. Regarding RQ1, the SNR increased with the sampling rate, from 17.4, 24.9, and 14.2 dB (speech, classical, pop) at 8 kHz to 69.3 dB (speech) and 78.9 dB (classical) at 44.1 kHz with the sharper resampling filter, and to about 58.4 dB with the default filter. With the sharper filter it stayed within 0.52 dB of the band-limited bound (category means up to 32 kHz). Pop was the most sensitive to a low rate and classical music the least. Regarding RQ2, the quantization SNR increased by 6.02 dB per bit and the RMSE halved with every bit, agreeing with the uniform-noise model within 0.01 dB from 8 bit upward. Regarding RQ3, file size grew linearly with both parameters and was larger than the PCM size only by a 44-byte header. FLAC brought 16-bit files down to 0.50–0.74 of the WAV size, and 8 bit capped the SNR at 32–36 dB. Regarding RQ4, the resampling filter changed the SNR at high rates by up to 20 dB as well as the benefit of 24 bit (0.03 dB with the default filter; 0.44 dB for speech and 9.0 dB for classical music with the sharper one), and the acceptance criterion shifted the operating point by one standard rate. Requiring SNR ≥ 40 dB for every clip gave 44.1 kHz with 16 bit for speech and pop and 22.05 kHz with 16 bit for classical music. A criterion on the mean would have selected 32 kHz with 16 bit, where seven of eight pop clips and three of eight speech clips stayed under 40 dB. An SNR of 40 dB did not guarantee a ViSQOL score of 4.0, and 18 of 130 clip–configuration pairs fell below it.

In practical terms, the CD format (44.1 kHz, 16 bit) was the smallest tested configuration that met SNR ≥ 60 dB and ViSQOL ≥ 4.5 for all 24 clips, and no criterion on resampled material required 48 kHz or 24 bit. When only speech is stored and ViSQOL ≥ 4.0 is acceptable, 32 kHz with 16 bit is 27% smaller than the CD format. Relative to a 48 kHz, 24-bit WAV file, these settings save 39–56% of storage as WAV and 67–79% as FLAC. The notebook, its configuration, and the data manifest are public, so the benchmark can be rerun as it is or applied to new material. Using the five-step procedure in Section 3.8, other users can estimate the operating point of their own material from its spectrum and Eq. (6) and confirm it with a few measurements, keeping in mind that we only checked the procedure on the material of this study. Our conclusions are valid for the tested content, the two resampling filters, the signal-level and ViSQOL metrics,

and sources whose documentation and spectra showed no lossy history. Future work could add listening tests, dithering and noise shaping, more resampling filters and genres, multi-channel material, and compressed formats.

REFERENCES

[1] R. Ramya Devi and D. Pugazhenthi, "Ideal sampling rate to reduce distortion in audio steganography," *Procedia Comput. Sci.*, vol. 85, pp. 418–424, 2016, doi: 10.1016/j.procs.2016.05.185.

[2] Y. Zhao, H. Wang, Y. Zheng, Y. Zhuang, and N. Zhou, "High sampling rate or high resolution in a sub-Nyquist sampling system," *Measurement*, vol. 166, Art. no. 108175, 2020, doi: 10.1016/j.measurement.2020.108175.

[3] J. D. Reiss, "A meta-analysis of high resolution audio perceptual evaluation," *J. Audio Eng. Soc.*, vol. 64, no. 6, pp. 364–379, Jun. 2016, doi: 10.17743/jaes.2016.0015.

[4] H. Liu, K. Chen, Q. Tian, W. Wang, and M. D. Plumbley, "AudioSR: Versatile audio super-resolution at scale," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2024, pp. 1076–1080, doi: 10.1109/ICASSP48485.2024.10447246.

[5] M. Torcoli, T. Kastner, and J. Herre, "Objective measures of perceptual audio quality reviewed: An evaluation of their application domain dependence," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1530–1541, 2021, doi: 10.1109/TASLP.2021.3069302.

[6] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "ViSQOL v3: An open source production ready objective speech and audio metric," in *Proc. 12th Int. Conf. Quality Multimedia Experience (QoMEX)*, Athlone, Ireland, 2020, pp. 1–6, doi: 10.1109/QoMEX48832.2020.9123150.

[7] B. B. Monson, E. J. Hunter, A. J. Lotto, and B. H. Story, "The perceptual significance of high-frequency energy in the human voice," *Front. Psychol.*, vol. 5, Art. no. 587, 2014, doi: 10.3389/fpsyg.2014.00587.

[8] J. Su, Y. Wang, A. Finkelstein, and Z. Jin, "Bandwidth extension is all you need," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2021, pp. 696–700, doi: 10.1109/ICASSP39728.2021.9413575.

[9] S. Han and J. Lee, "NU-Wave 2: A general neural audio upsampling model for various sampling rates," in *Proc. Interspeech*, 2022, pp. 4401–4405, doi: 10.21437/Interspeech.2022-45.

[10] M. Mandel, O. Tal, and Y. Adi, "AERO: Audio super resolution in the spectral domain," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10095382.

[11] J. Pons, S. Pascual, G. Cengarle, and J. Serrà, "Upsampling artifacts in neural audio synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2021, pp. 3005–3009, doi: 10.1109/ICASSP39728.2021.9414913.

[12] K. Saito, T. Nakamura, K. Yatabe, and H. Saruwatari, "Sampling-frequency-independent convolutional layer and its application to audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2928–2943, 2022, doi: 10.1109/TASLP.2022.3203907.

[13] T. Mineo and H. Shouno, "Improving sign-algorithm convergence rate using natural gradient for lossless audio compression," *EURASIP J. Audio, Speech, Music Process.*, vol. 2022, Art. no. 12, 2022, doi: 10.1186/s13636-022-00243-w.

[14] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "SoundStream: An end-to-end neural audio codec," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 495–507, 2022, doi: 10.1109/TASLP.2021.3129994.

[15] W. R. Bennett, "Spectra of quantized signals," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 446–472, Jul. 1948, doi: 10.1002/j.1538-7305.1948.tb01340.x.

[16] B. Widrow and I. Kollár, *Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[17] J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, "EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation," in *Proc. Interspeech*, 2024, pp. 4873–4877, doi: 10.21437/Interspeech.2024-153.

[18] Musopen, "Musopen Symphony and Musopen String Quartet recordings," Wikimedia Commons. [Online]. Available: https://commons.wikimedia.org/wiki/Category:Musopen (accessed Nov. 6, 2026).

[19] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "MUSDB18-HQ – an uncompressed version of MUSDB18," Zenodo, 2019, doi: 10.5281/zenodo.3338373.

[20] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," Zenodo, Dec. 2017, doi: 10.5281/zenodo.1117372.

[21] Musopen, "Musopen Collection as FLAC (Musopen Kickstarter Project, 2012)," Internet Archive. [Online]. Available: https://archive.org/details/MusopenCollectionAsFlac (accessed Nov. 6, 2026).

[22] SigSep, "MUSDB18 dataset documentation: compressed STEMS and uncompressed WAV (MUSDB18-HQ)," 2022. [Online]. Available: https://sigsep.github.io/datasets/musdb.html (accessed Nov. 6, 2026).

[23] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Munich, Germany, 1997, vol. 2, pp. 1331–1334, doi: 10.1109/ICASSP.1997.596192.

[24] M. van Beurden and A. Weaver, "Free Lossless Audio Codec (FLAC)," IETF, RFC 9639, Dec. 2024, doi: 10.17487/RFC9639.

[25] A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram, and N. Harte, "ViSQOLAudio: An objective audio quality metric for low bitrate codecs," *J. Acoust. Soc. Am.*, vol. 137, no. 6, pp. EL449–EL455, Jun. 2015, doi: 10.1121/1.4921674.

[26] *Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*, ITU-T Recommendation P.862.2, Nov. 2007.

[27] *Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems*, ITU-R Recommendation BS.1534-3, Oct. 2015.

[28] *Mean Opinion Score (MOS) Terminology*, ITU-T Recommendation P.800.1, Jul. 2016.

[29] *Methods for Subjective Determination of Transmission Quality*, ITU-T Recommendation P.800, Aug. 1996.