Nonlinear audio processing when creating excursion material using AI content

Alexander Blinnikov[1*], *Ivan* Blinnikov[2], *Erkin* Kadirov[3], and *Nikolay* Tsybov[4]

[1]Krasnoyarsk State Agrarian University, Krasnoyarsk, Russia
[2]School No. 72 with in-depth study of subjects named after M.P. Tolstikhin, Krasnoyarsk, Russia
[3]Navoi State Mining and Technological University, Navoi, Uzbekistan
[4]Institute of Mechanical Engineering and Automation of the National Academy of Sciences of the Kyrgyz Republic, Bishkek, Kyrgyzstan

Abstract. This article examines the process of creating audio guides for the "Museum of Our Childhood" which is a unique interactive project dedicated to the engineering glory of Yenisei Siberia. Special attention is given to the stages of recording voice-overs, audio editing, and mastering of three audio tours. The paper describes approaches to non-linear processing of audio material, selection of equipment and software, as well as the specifics of integrating audio tours with museum exhibits. The importance of the audio format for preserving historical memory, engaging youth, and popularizing engineering achievements through the digitalization of cultural heritage is emphasized.

1 Introduction

Modern museums strive to keep pace with the times, implementing innovative approaches to presenting material in order to make exhibitions more engaging and accessible to a wide audience [1]. Audio guides are becoming an integral part of such solutions, combining technological capabilities with artistic expressiveness [2].

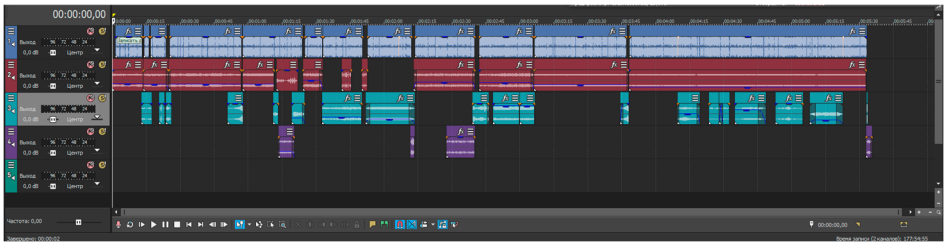

Fig. 1. Timeline on the editing table for a short museum tour (Sony Vegas PRO).

The "Museum of Our Childhood," a member of the Association of Private and Folk Museums of Russia, represents a unique cultural space that narrates the everyday life and experiences of people from the past through objects of engineering and domestic heritage of

* Corresponding author: blinshur@yandex.ru

Yenisei Siberia. To implement audio tours for the museum, a project was developed encompassing voice-over recording, audio material editing, and mastering [5]. A distinctive feature of the process was the utilization of non-linear audio processing to achieve a high level of sound quality, as well as the integration of audio tours with QR codes on exhibits (Fig. 1).

This paper is dedicated to describing all stages of creating audio guides: from equipment selection and working with voice actors to final sound processing. The applied methods and approaches are examined, emphasizing the significance of the audio format for engaging youth in the study of the region's scientific and engineering heritage.

2 Materials and methods

The first method applied in this work is the non-linear post-processing of the narrator's voice using digital transformations based on one-dimensional Fourier transforms (forward and inverse), audio filters, audio compressors, and noise reducers (Fig. 2) [5].

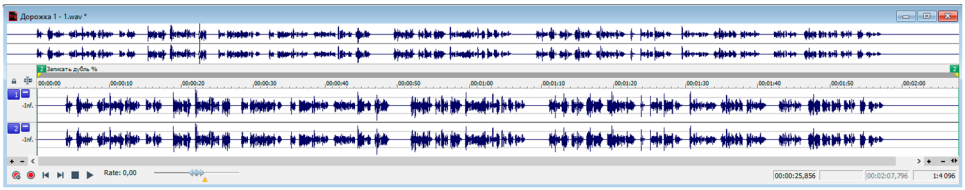

Fig. 2. Narrator's voice without processing.

The second method is software-neural network based. Using an AI generator [6], authentic audio content was created from a text prompt for background sound in the body of the finished audio tour.

The third method applied in this work was audio montage, with the addition of sound effects, layering of processed narrator's voice, and thematic musical backgrounds created using an AI generator [7].

2.1 2.1 Non-linear post-processing of the narrator's voice

The initial stage of processing the narrator's voice is its normalization in a music editor [8]. Normalization is based on equalizing the amplitude of the audio signal so that its loudness (perceptual or physical) corresponds to a given level without creating distortions.

Below is a simplified mathematical model of this process:

Let x(t) be an audio signal in the time domain, where t is time. This signal is sampled at a frequency fs, and the discrete signal is represented as x[n], where n is the time sample index.The signal level can be estimated in several ways:

A. RMS (Root Mean Square):

The root means square value of the signal over a window of length N:

$$RMS[n] = \sqrt{\frac{1}{N}\sum_{k=0}^{N-1} x^2[n-k]}, \quad (1)$$

B. Peak value:

$$\boldsymbol{Peak[n]} = \max_{\boldsymbol{k\in[n,n+N]}} |\boldsymbol{x[k]}|, \quad (2)$$

C. Perceptual loudness units (LUFS):

ITU-R BS.1770 standards are used to measure loudness taking into account human hearing perception:

$$\text{Loudness} = -0.691 \cdot \text{RMS} + \text{Offset}, \quad (3)$$

2.2 Normalization

For normalization (Fig. 3), a gain factor G is introduced that scales the signal so that its level reaches a given value L_{target} (e.g. 0 dBFS for the peak value or -14 LUFS for perceptual loudness):

$$G = \frac{L_{target}}{L_{current}}, \tag{4}$$

where $L_{current}$ is the current signal level, determined by one of the metrics above.

Application of the coefficient:

Normalized signal y[n]:

$$y[n] = G \cdot x[n], \tag{5}$$

Amplitude Limit:

After normalization, it is checked that the signal amplitude does not exceed the permissible value Amax (for example, 1 in digital representation):

$$\boldsymbol{y[n] = \min\left(\max(y[n], -A_{max}), A_{max}\right)}, \tag{6}$$

Final algorithm:

- Calculate the current signal level $L_{current}$ (RMS, Peak, or Loudness).
- Determine the gain G.
- Apply G to the signal x[n].
- Limit the amplitude of the result to A_{max}.

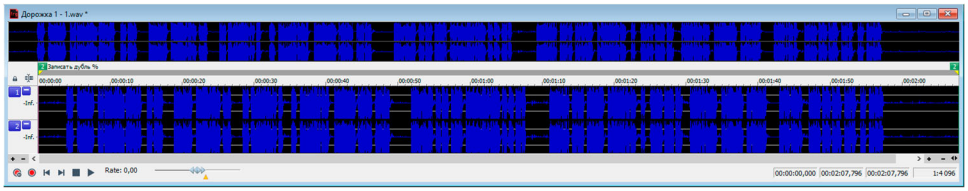

Fig. 3. Normalization of the selected voice.

2.3 Noise reduction

Noise reduction is the process of removing or reducing unwanted noise from an audio signal [9]. The mathematical model is based on the extraction of noise from the signal, its analysis and subsequent suppression:

Input signal analysis:

The audio signal x(t) is represented as the sum of the useful signal s(t) and noise n(t):

$$x(t) = s(t) - n(t), \tag{7}$$

The goal of the algorithm is to minimize n(t) while preserving s(t).

A section of the audio containing only noise n(t) is selected. This allows one to create a spectral noise profile $P_n(f)$, where f is the frequency.

The noise profile is calculated via a spectral transform, for example using the Fast Fourier Transform (FFT):

$$P_n(f) = |X_n(f)|^2, \tag{8}$$

where $X_n(f)$ is the noise spectrum calculated via the Fourier transform:

$$\boldsymbol{X_n(f) = \int_{-\infty}^{\infty} n(t) e^{-i2\pi ft} dt}, \tag{9}$$

2.3.1 Spectral subtraction

Noise is suppressed by subtracting its spectral power from the total signal spectrum Px(f):

$$\boldsymbol{P_s(f) = P_x(f) - P_n(f)}, \tag{10}$$

where: $\mathbf{P_x(f) = |X(f)|^2}$ - spectral power of the input signal x(t), $P_s(f)$ - spectral power of the useful signal s(t).

To avoid negative values of the spectral power (artifacts), a limitation is applied:

$$P_s(f) = \max(P_x(f) - P_n(f), \in), \tag{11}$$

where $\epsilon>0$ is a small threshold preventing division by zero and artifacts.

2.3.2 Construction of the suppressed signal

After noise suppression, the useful signal is reconstructed using the inverse Fourier transform:

$$s(t) = \int_{-\infty}^{\infty} \sqrt{P_s(f)} \cdot e^{i\phi_x(f)} e^{i2\pi ft} df, \tag{12}$$

where $\phi_x(f)$ is the phase of the input signal (it is preserved to minimize distortion).

2.3.3 Temporal smoothing

Temporal smoothing of the spectrum is used to prevent sharp changes in amplitude. For example, for each frequency f, exponential smoothing is calculated:

$$P_s'(f,t) = \alpha P_s(f,t) + (1-\alpha) P_s'(f,t-1), \tag{13}$$

where: α[0,1] is the smoothing coefficient, t is the current time frame.

Suppression threshold adjustment

The suppression level (Fig. 4) is controlled by the suppression coefficient G(f), which depends on the signal and noise levels:

$$G(f) = \frac{P_s(f)}{P_x(f)}, \tag{14}$$

Final spectrum after suppression:

$$S(f) = G(f) \cdot X(f), \tag{15}$$

Final mathematical model:

a) Spectral transformation of the input signal:

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-i2\pi ft} dt, \tag{16}$$

b) Subtraction of the spectral power of noise:

$$P_s(f) = max(|X(f)|^2 - P_n(f), \epsilon), \tag{17}$$

c) Signal restoration:

$$s(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \sqrt{P_s(f)} \cdot e^{i\phi_x(f)} e^{i2\pi ft} df, \tag{18}$$

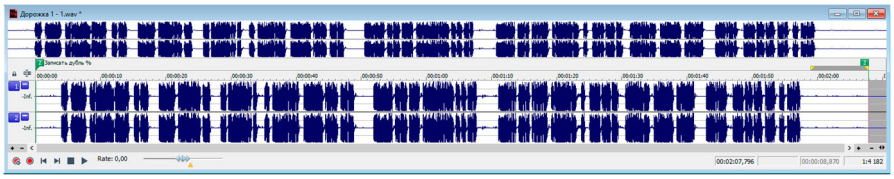

Fig. 4. Noise reduction of the announcer's voice.

This model describes a basic spectral noise reduction algorithm [10]. Real implementations include additional steps such as adaptive filtering, artifact management, and sound quality improvement.

2.3.4 Equalizer processing

The mathematical model of voice-over equalization with noise minimization and frequency response optimization includes several steps: frequency spectrum analysis, noise suppression, frequency alignment, and amplification of the required frequency ranges [11].

Mathematical representation of the original signal

Let x(t) be the original voice-over signal in the time domain, and X(f) be its transformation into the frequency domain via the discrete Fourier transform (DFT):

$$X(f) = F\{x(t)\} = \sum_{n=0}^{N-1} x[n] e^{-\frac{i2\pi fn}{N}}, \quad (19)$$

where: N is the length of the analysis window, f is the frequency.

Frequency spectrum analysis:

For the announcer's voice, the important frequency ranges are usually within 80 Hz – 12 kHz.

Let's define the energy spectrum of the signal:

$$P(f) = |X(f)|^2, \quad (20)$$

Based on the spectrum, the following are distinguished:

- Low frequencies (80−250 Hz) — the basis of the voice (fundamental frequency and lower harmonics).
- Medium frequencies (250−4000 Hz) — the area of voice intelligibility (formants).
- High frequencies (4000−12000 Hz) — add “brightness” and “clarity”.

Noise components are most often in the range of low (<80 Hz) and high (>12 kHz) frequencies.

Noise filtering:

A bandpass filter H(f) is used, which suppresses unwanted frequencies.

$$H(f) = \begin{cases} 1, & f_{low} \le f \le f_{high} \\ 0, & other\ values \end{cases}, \quad (21)$$

where: f_{low}=80 Hz, f_{high}=12 kHz - boundaries of the passable strip.

Filtered signal in frequency domain:

$$Y(f) = H(f) \cdot X(f), \quad (22)$$

Back to the time domain:

$$y(n) = F^{-1}\{Y(f)\}, \quad (23)$$

To improve the intelligibility of the announcer's voice, the equalizer parameters specified by the gain function are applied. G(f):

$$G(f) = \begin{cases} G_{low}, & f_{low1} \le f \le f_{low2} \\ G_{mid}, & f_{mid1} \le f \le f_{mid2} \\ G_{low}, & f_{high1} \le f \le f_{high2} \\ 1, & \text{иные значения} \end{cases}, \quad (24)$$

Recommended gain values:

G_{low} (80−250 Hz): − 3 dB (attenuation to reduce hum).

G_{mid} (250−4000 Hz): +3 dB (enhancement for intelligibility).

G_{high} (4000−8000 Hz): +2 dB (enhancement for «brightness»).

Processed spectrum:

$$Z(f) = G(x) \cdot Y(f), \quad (25)$$

And, accordingly, the signal in the time domain:

$$z(t) = F^{-1}\{Z(f)\}, \quad (26)$$

2.3.5 Minimization of noise parameters

Additionally, noise suppression can be applied using the spectral subtraction method (Fig. 5).

The noise model N(f) is estimated from the quiet sections of the signal and subtracted:

$$\hat{Z} = \max\left(Z(f) - N(f), 0\right), \tag{27}$$

Final model

The full processed signal of the announcer's voice after the equalizer and noise suppression:

$$\hat{z} = F^{-1}\{\hat{Z}(f)\}, \tag{28}$$

where $\hat{Z}$ is the spectrum after equalization and noise suppression.

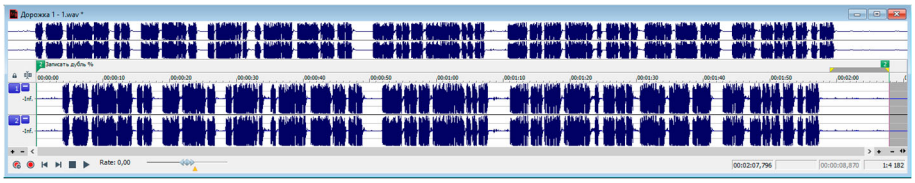

Fig. 5. Voice processing with equalizer.

Typically, combinations are used [12]:

- Bandpass filters (IIR or FIR),
- Spectral analysis using the windowed Fourier transform (STFT),
- Dynamic processing (e.g. compression to equalize the loudness level).

2.3.6 WAV Hammer Compression

The mathematical model of its operation is based on the analysis of the amplitude of the audio signal, the use of compression and normalization.

The audio signal x(t) is analyzed in the time domain. The main parameters are calculated as follows.

- Peak value:

$$A_{peak} = max(|x(t)|), \tag{29}$$

- Root mean square (RMS):

$$A_{rms} = \sqrt{\frac{1}{T} \cdot \int_0^T x^2(t)dt}, \tag{30}$$

where T is the signal duration.

- Signal compression

Compression changes the dynamic range of the signal x(t) using the compression ratio R (for example, R=4:1) and the threshold $T_{threshold}$.

Compression algorithm: If $|x(t)|>T_{threshold}$, the amplitude is reduced by the formula:

$$y(t) = T_{threshold} + \frac{|x(t)| - T_{threshold}}{R}, \tag{31}$$

Otherwise:

$$y(t)=x(t), \tag{32}$$

- Signal limiting

After compression, the signal is normalized so that its peak amplitude reaches a given level A_{norm} (usually A_{norm}=0 dB).

- Normalization formula:

$$z(t) = \frac{y(t)}{A_{peak}} \cdot A_{norm}, \tag{34}$$

- Final mathematical formula:

$$z(t) = \begin{cases} -\frac{sign(x(t))\cdot\left(T_{threshold}+\frac{|x(t)|-T_{threshold}}{R}\right)}{A_{peak}}, & |x(t)| > T_{threshold} \\ \frac{x(t)}{A_{peak}} \cdot A_{norm}, & |x(t)| \leq T_{threshold} \end{cases}, \quad (35)$$

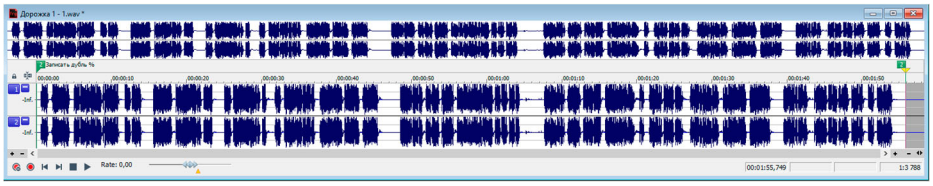

Fig. 6. Voice processing with compressor and partial trimming.

Parameters such as attack time (T_{attack}) and release time ($T_{release}$) can introduce time dependencies when changing the amplitude. This is implemented through exponential smoothing of the amplitude:

$$A' = \alpha A(t) + (1-\alpha)\cdot A'(t-1), \quad (36)$$

where $\alpha = e^{-\frac{\Delta t}{T_{attack/release}}}$.

2.4 Software-neural network generation of background content

The MUZIC AI neural network, upon receiving a text input T describing the desired audio track, including style, tempo, and instruments, generates a musical composition according to the programmed task [13]. This can be represented as a sequence of tokens **T={t₁,t₂,...,tₙ}**, where t_i is an individual word or phrase. The objective is to transform the textual description T into an audio signal S, representing a time series:

$$S = \{s_1, s_2, \ldots, s_n\}, s_i \in \mathbb{R}, \quad (37)$$

In this context, $s_i \in \mathbb{R}$ denotes the set of real numbers. This is a standard mathematical notation used to describe the data type or space in which the values reside.

- $S=\{s_1,s_2,\ldots,s_m\}$ is a time series representing the audio signal.
- Each value s_i in this series is the amplitude of the sound wave at a specific point in time, which is a real number ($s_i \in \mathbb{R}$).

An audio signal in digital format typically represents a discretized time series, where each s_i is the amplitude of the sound wave at the corresponding moment in time. For example:

- If the audio signal is recorded at a sampling rate of 44.1 kHz, there will be 44,100 s_i values per second.
- These s_i values are usually normalized to the range [−1,1] or represented in 16- or 32-bit number format, corresponding to real numbers.

Thus, $s_i \in \mathbb{R}$ emphasizes that each amplitude of the audio signal can be any real number that accurately describes the sound wave.

2.4.1 Neural Network Architecture

a) Text Encoder

The text description T is processed through an encoder that utilizes a Transformer architecture to extract semantic representations of the text. The encoder output is (38):

$$E(T) = \{h_1, h_2, \ldots, h_n\}, h_i \in \mathbb{R}^d, \quad (38)$$

where h_i is the vector representation of token t_i, and d is the dimension of the hidden layer.

b) Audio Signal Decoder

The audio signal is generated using a decoder that constructs a time series S based on the hidden representations of the text E(T). This is achieved through autoregressive generation:

$$P(S|T) = \prod_{i=1}^{m} P(s_i|s_{1:i-1}, E(T)), \quad (39)$$

where $s_{1:i-1}$ represents the previously generated audio signal values.

c) Contextual Attention Mechanism

The attention mechanism is used to map text tokens to temporal fragments of the audio signal. Attention weights αij are calculated as:

$$\alpha_{ij} = softmax\left(\frac{h_i^T \cdot q_j}{\sqrt{d}}\right), \quad (40)$$

where h_i is the text vector and q_j is the decoder query for the current time step j.

These coefficients help the decoder focus on relevant parts of the text when generating a specific section of the audio signal.

d) Audio Signal Modeling

- Mel-spectrogram Generation

Mel-spectrograms are used to represent the audio signal - two-dimensional arrays $M \in \mathbb{R}^{f \times t}$, where f is the number of frequency bands and t is the time steps. The mel-spectrogram is generated as follows:

$$M = Decoder\big(E(T)\big), \quad (41)$$

where M is interpreted as a hidden representation of the audio signal.

- Time Signal Reconstruction

The final audio signal S is reconstructed from the mel-spectrogram M using a neural vocoder, such as WaveNet or HiFi-GAN. This transformation can be described as:

$$S = Vocoder(M), \quad (42)$$

where Vocoder is trained to reproduce a high-quality signal approximating real audio data.

e) Model Training

The model is trained on data pairs (T, S), where T is the text description and S is the real audio track. The training objective is to minimize a loss function, such as the mean squared error between the real signal St_{rue} and the predicted signal S_{pred}:

$$\mathcal{L} = \frac{1}{m}\sum_{i=1}^{m}\big(S_{true,i} - S_{pred,i}\big)^2 \quad (43)$$

where $\mathcal{L}$ is the Laplace transform, m is the total number of time points in the audio signal S, i is the index of the discrete value of the time series (sound signal amplitude), S_{true} is the true amplitude value at the i-th position, and S_{pred} is the predicted amplitude value at the i-th position.

f) Sum Limits

The sum is calculated over all time points i, i.e., i takes values from 1 to m, where m is the length of the discrete time series (usually in milliseconds or seconds, depending on the sampling frequency).

For example, if an audio file has a sampling rate r (let's say 44,100 Hz), then for a 1-second signal, m = 44,100. In this case, the sum is taken over all time points of the signal.The loss function L in this form compares the true and predicted audio signals at the level of amplitudes of each time point and minimizes the mean squared difference between them.

g) Accounting for Style and Genre

To account for style or genre, the model additionally uses conditional parameters G, such as instrument type, rhythm, and tempo. These parameters are combined with the text representations E(T):

$$E'(T, G) = Concat(E(T), G), \quad (44)$$

This allows the model to adapt the generation to specified conditions.MUZIC AI 5.0 algorithms apply a combination of mathematical models and deep learning to create high-quality audio content that can be used for our instrumental tasks.

2.5 Audio Editing and Mastering

Three voice-over recordings were prepared and processed using the audio editor Sound Forge.

The editing process included trimming the initial and final sections, removing pauses, and correcting any errors or misstatements made by the narrators (Fig. 7). The recordings corresponded to three types of audio tours:

- Children's tour,
- Short tour,
- Standard tour.

Additionally, the following audio elements were prepared for each type of tour:

For the children's tour:

- 25 sound effects,
- 15 instrumental tracks generated using MUZIC AI 5.0.

For the short tour:

- 14 sound effects,
- 9 instrumental tracks generated using MUZIC AI 5.0.

For the standard tour:

- 41 sound effects,
- 29 instrumental tracks generated using MUZIC AI 5.0.

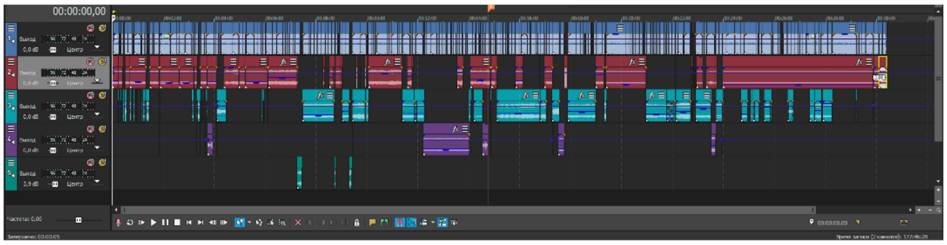

Fig. 7. Final assembly on 5 audio tracks of a standard tour (Sony Vegas PRO 16.0)

However, it is worth noting that the standard tour partially used both effects and generated instrumental compositions from the selection for the children's and short tours.

Audio editing of materials for audio excursions was performed using Sony Vegas PRO software version 16.0 [14]. Four audio tracks were used to create the shortened version of the excursion and the children's excursion, while five audio tracks were used for the standard version of the excursion. Non-linear post-processing was performed using plug-ins from the BBE Sonic Sweet series in the Sound Forge Pro 12.0 Suite (x64) environment.

Final mastering (Fig. 8) was performed using the AudioUnit automatic spectral balancer, AAX and the VST plug-in for professional audio and music applications in Sony Vegas PRO version 20.0. TEOTE, a dynamic equalizer with multi-band dynamic processing technology, was also used [15].

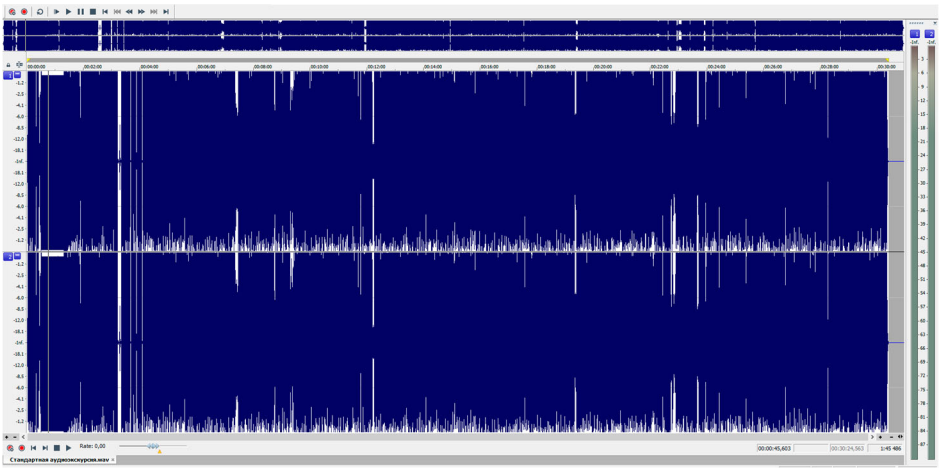

Fig. 8. Mastering the soundtrack of a standard tour (Sound Forge PRO 12.0 Suite).

Non-linear digital post-processing was performed on a workstation with a 12th Gen Intel(R) Core(TM) i7-12700K processor (3.60 GHz frequency), running Windows 10 Pro 64-bit operating system on the x64 platform.

3 Results and discussion

Considering the requirement for historically and chronologically appropriate musical accompaniment for the tour material, the authors decided to leverage generative neural network solutions.The durations of the three audio tours were as follows:

- 13 minutes and 36 seconds,
- 5 minutes and 35 seconds,
- 30 minutes and 24 seconds.

Each tour was developed with attention to the target audience and its intended context of use:

- The **children's tour** is designed to foster emotional engagement and active perception.
- The **short tour** emphasizes conciseness and focuses on key aspects.
- The **standard tour** is the most content-rich and complex to produce, incorporating elements from all versions.

Table 1. Summary table of rendering results.

Parameter	The children's tour	The short tour	The standard tour
Duration	13 min 36 sec	5 min 35 sec	30 min 24 sec
Sound effects	25	14	41
Musical compositions	15	9	29
Effect density (per 1 min)	1.84	2.5	1.35
Music density (per 1 min)	1.1	1.6	0.95
Number of audio tracks	4	4	5

Note: compiled by the authors

3.1.1 Post-processing and Sound Design for Audio Tours

This section details the post-processing techniques and sound design strategies employed in creating audio tours for the "Museum of Our Childhood" project.

3.1.2 Audio Editing and Processing

Sony Vegas Pro was utilized for multi-track editing, with the number of tracks varying from 4 to 5 depending on the specific tour. This software facilitated parallel work with multiple audio elements, allowing for precise balance between sound effects and music.

Sound Forge Pro was employed for final refinements, including:

- Volume level adjustments
- Equalization
- Application of effects such as sound brightness enhancement
- Improved perception through BBE Sonic Sweet plugins

Sound Effects and Music Utilization

The implementation of sound effects and music varied across different tour types:

Children's Tour:

- High density of sound effects (1.84 effects per minute)
- Frequent use of musical compositions (1.1 compositions per minute)
- Focus on interactivity and emotional engagement of children through frequent use of audio elements

Short Tour:

- Reduced use of sound effects (2.5 per minute) and musical compositions (1.6 per minute)
- Minimalistic audio design to align with the goal of concise information delivery

Standard Tour:

- Highest density of musical and sound content (1.35 effects per minute and 0.95 compositions per minute)
- Comprehensive use of all available resources to create a rich sound palette suitable for full listener immersion

All tours were produced in .wav format with a bitrate of 1411 kbps, ensuring high audio quality.

3.2 Comparative analysis of results

The main parameters used in the analysis are as follows:

Integrated Loudness (LUFS) represents the average loudness of audio throughout the entire playback time, measured in Loudness Units Full Scale. Loudness Range (LU) indicates the difference between the maximum and minimum loudness in an audio track, measured in Loudness Units. RMS Level (dB) is the root mean square level of the audio signal, measured in decibels, reflecting the average loudness of the audio. Average Value (dB) represents the mean sound level over a specific period, measured in decibels.

Zero Crossings (Hz) count the number of times the sound wave crosses the zero line per second, potentially indicating the complexity of the audio signal. Maximum True Peak Sample Position (Time) denotes the time at which the maximum true peak of the audio signal is reached. Maximum True Peak Sample Value (dB) measures the level of the maximum true peak of the audio signal in decibels. Maximum Short-Term Loudness (LUFS) represents the maximum loudness measured over a short period. Maximum Momentary Loudness (LUFS) indicates the peak loudness measured over a very short time interval.

All loudness level values are typically expressed in decibels (dB), where 0 dB represents the maximum signal level without clipping. LUFS and LU are units of measurement that take into account the perception of loudness by the human ear.

After adapting the audio files for the audio guide within the excursion activities, the .wav versions were converted to .mp3 format at 320 kbps, 44100 Hz, 16-bit, stereo. Spectral analysis was performed using SPAN Plus, a real-time Fast Fourier Transform (FFT) audio frequency spectrum analyzer (Fig. 10).

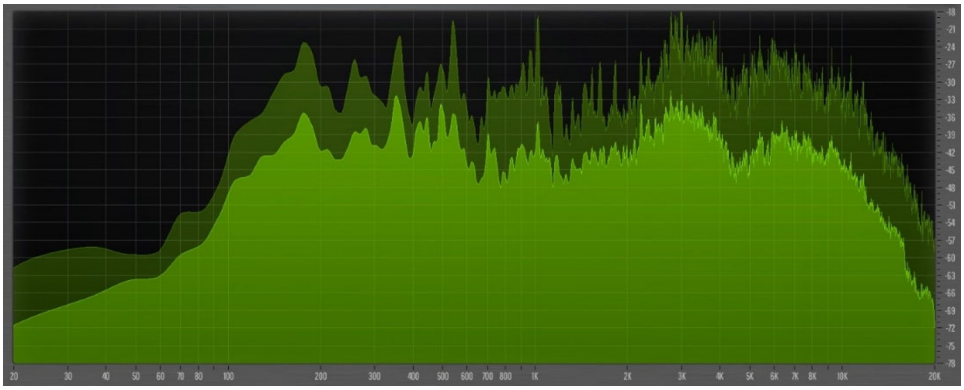

Fig. 10. Spectral analysis of RT Avg and Max (Short tour).

3.2.1 General Overview

The audio characteristics of three distinct tour types were analyzed: the Children's Tour, the Short Tour, and the Standard Tour. For the Children's Tour, the Integrated Loudness was measured at -8.85 LUFS, with a Loudness Range of 2.60 LU. The RMS Level was recorded at -12.001 dB for the left channel and -11.966 dB for the right channel, while the Average Value was -78.268 dB and -70.309 dB for the left and right channels, respectively.

The Short Tour exhibited an Integrated Loudness of -8.49 LUFS and a Loudness Range of 1.90 LU. Its RMS Level was measured at -11.665 dB for the left channel and -11.664 dB for the right channel. The Average Value for this tour was -80.767 dB for the left channel and -73.407 dB for the right channel.

Lastly, the Standard Tour demonstrated an Integrated Loudness of -8.82 LUFS and a Loudness Range of 2.40 LU. The RMS Level for this tour was recorded at -11.418 dB for the left channel and -11.401 dB for the right channel. The Average Value was measured at -58.713 dB and -58.053 dB for the left and right channels, respectively.

These measurements provide a comprehensive overview of the audio characteristics for each tour type, allowing for a detailed comparison of their loudness, dynamic range, and overall audio levels.

3.2.2 Comparative Analysis

The comparative analysis of the three audio tours reveals several key aspects of their audio characteristics. In terms of volume level, all three tours exhibit similar Integrated Loudness values ranging from -8.49 to -8.85 LUFS, indicating that all audio tours maintain a relatively high-volume level. This consistency in loudness across the tours ensures a uniform listening experience for visitors.

The Loudness Range varies among the tours, with the children's tour displaying the highest value at 2.60. This suggests that the children's tour incorporates more diverse dynamic changes in its audio material, potentially to maintain younger listeners' engagement. In contrast, the short tour has the lowest Loudness Range at 1.90, which may indicate a more consistent dynamic throughout its duration.

Regarding RMS levels and mean values, the analysis shows that the RMS level for all tours is approximately -12 dB. However, a notable observation is the significant difference in mean loudness values between the left and right channels across all tours. For instance, the children's tour exhibits a difference of about 8 dB between its left and right channels. This disparity in channel loudness could be intentional, possibly to create a more immersive audio experience or to emphasize certain elements of the tour.

These findings provide valuable insights into the audio characteristics of the tours, highlighting both consistencies and variations in their sound design. Such analysis can inform future refinements and ensure that the audio quality aligns with the intended visitor experience for each tour type (Fig. 11).

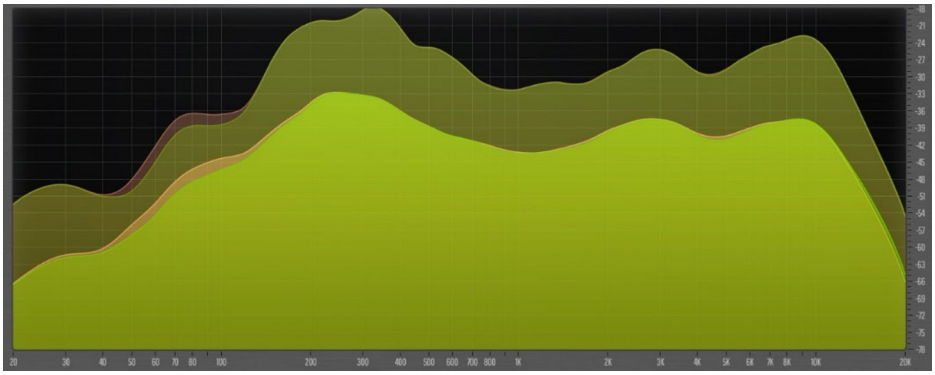

Fig. 11. Spectral analysis of RTAvgMax Stereo Slow (Children's excursion).

Peak Values:

- All tours have maximum true peaks at 0 dB TP. This indicates that none of the audio exceeds the permissible clipping level.

- The audio tours exhibit similar loudness and dynamic characteristics. However, the children's tour stands out with a wider loudness range and significant differences between channels. The short tour demonstrates more consistent dynamics with a narrower loudness range. All three tours are within safe limits in terms of peak signal level.

4 Conclusion

The created audio tours represent an important step in the development of museum experiences, ensuring accessibility and engagement for diverse target audiences. The use of modern technologies, such as neural network generators and non-linear sound processing, allows for the creation of high-quality audio content that not only informs but also entertains listeners.

The children's tour, with its high level of interactivity and emotional richness, aims to attract the younger generation to cultural heritage. In contrast, the short tour offers a concise presentation of key facts, making it ideal for a quick overview of the exhibition. The standard tour, possessing the richest sound palette, provides an in-depth immersion into the subject matter and enables listeners to better understand the context of the presented objects.

Thus, each of the tours fulfills its unique role in the educational process and contributes to a deeper perception of the museum content at the "Museum of Our Childhood" in the House of Science and Technology. The implementation of such innovative solutions makes museums more modern and appealing to a broader audience, which is a crucial aspect in the context of growing competition for visitor attention.

The project was developed with the support of the Presidential Grants Foundation No. 24-2-005304 «Interactive Museum of the Engineering Glory of Yenisei Siberia».

References

1. A.A. Voroshilova, I.V. Kovalev, A.V. Bagachuk, Y.Y. Bocharova, Informatics. Economics. Management **2,** 0311–0320 (2023)
2. E.V. Yudina, Acoustical Physics **66,** 213-230 (2020)
3. E.A. Kozlova, Issues of Cognitive Linguistics **2,** 103-111 (2018)
4. G.A. Volkov, K.R. Nazarova, Student Science and XXI Century **13,** 14-16 (2016)
5. T.M. Tatarnikova, E.D. Poimanova, P.Yu. Bogdanov et al., Software Products and Systems **1,** 145-150 (2021)
6. V.A. Zverev, A.I. Malekhanov, Acoustical Physics **70,** 283-288 (2024)
7. I.G. Ilyukhina, S.G. Richter, DSPA: Issues of Digital Signal Processing Application **12,** 31-38 (2022)
8. V.N. Sorokin, Acoustical Physics **70,** 778-794 (2024)
9. A.A. Andreev, M.S. Shapovalova, RSUH/RGGU Bulletin. Series: Informatics. Information Security. Mathematics **1,** 35-45 (2022)
10. V.G. Shamaev, A.B. Gorshkov, Acoustical Physics **65,** 122-144 (2019)
11. V.I. Dzhigan, Digital Signal Processing **3,** 14-23 (2022)
12. V.A. Chastikova, Z.Ya. Tugusheva, F.R. Gunai, Electronic Network Polythematic Journal "Scientific Works of KubSTU" **2,** 326-332 (2016)
13. E.N. Zeynaliev, Actual Research **19-1,** 37-44 (2024)
14. A.V. Gevorsky, M.S. Kostin, K.A. Boykov, Russian Technological Journal **12,** 30-58 (2024)
15. P.S. Ladygin, A.A. Lependin, A.V. Mansurov, High-Performance Computing Systems and Technologies **7,** 46-52 (2023)