Next Generation Audio (NGA) API Proposal for W3C

The Case for a Standardized Interface for Personalized Audio Experiences

Executive Summary

This document outlines the need for a standardized Web API to support Next Generation Audio (NGA) codecs and their personalization capabilities. While current web platform APIs provide basic audio playback functionality, they lack the necessary interfaces to expose the rich metadata and interactive features that make NGA codecs valuable. This proposal argues that existing solutions combining WebCodecs and WebAudio are insufficient for implementing NGA features, particularly when audio components and metadata are delivered in a single stream, as is common in commercial implementations.

1. Introduction to Next Generation Audio

Next Generation Audio codecs deliver audio experiences that are more accessible, personalized, and interactive, regardless of how they are consumed. NGA frees producers from creating multiple audio mixes for different reproduction systems by allowing them to deliver a single, multi-purpose audio master instead.

The key differentiator of NGA codecs (such as AC-4, MPEG-H Audio or DTS-UHD) from traditional codecs (AAC, AC-3, MP3) is that they do not only transmit encoded audio assets for a fixed channel configuration but also include extensive metadata to enable a more suitable user experience.

1.1 Core Features of NGA

NGA codecs provide:

- Object-based audio delivery
- Metadata for personalization and adaptation
- Content creator-defined constraints on personalization
- Support for immersive audio experiences

- Accessibility enhancements

1.2 Industry Adoption

NGA codecs are being widely adopted in broadcast, streaming, and other media delivery contexts. They are supported by major industry standards like MPEG DASH and are increasingly required for next-generation content delivery.

2. Use Cases for NGA on the Web

2.1 Preselection of Audio Mixes

A fundamental NGA feature is selecting a pre-defined audio "preselection" - a set of predefined mix parameters for the included content components. Examples include:

- Different team commentary options for sports events
- Enhanced dialogue options for accessibility
- Alternative language tracks with appropriate background audio

2.2 Dialogue Enhancement

NGA allows users to adjust the prominence/gain of dialogue components within creator-defined limits. This is particularly valuable for:

- Hearing-impaired viewers
- Non-native language speakers
- Noisy viewing environments

2.3 Spatial Positioning of Audio Elements

Users can reposition audio elements within the sound stage, enabling:

- Moving audio description to a specific location for better intelligibility
- Separating commentators spatially for clearer distinction
- Creating more immersive sports experiences by repositioning crowd noise

2.4 Component Selection

NGA enables selection between multiple content components, such as:

- Language selection for dialogue

- Choice between home/away commentators
- Optional components like audio description

2.5 Narrative Importance Control

Advanced NGA features allow controlling multiple components simultaneously through a single interface, such as adjusting the balance between dialogue and background elements based on narrative importance.

3. Limitations of Current Web Platform APIs

3.1 The Integrated Stream Challenge

The central issue is that NGA has a mode where audio components and metadata are delivered in a single stream. The metadata contains crucial instructions for mixing the components, but current web APIs provide no way to surface this metadata from the decoder and use it to control audio rendering.

3.2 Why Not Use WebCodecs + WebAudio?

While WebCodecs could theoretically decode the audio and WebAudio could mix it, this approach has several critical limitations:

1. **No Metadata Bridge**: There's no standardized way to extract the embedded metadata from the decoder and use it to control WebAudio mixing parameters.
2. **Temporal Alignment**: WebAudio has no concept of time-aligned metadata that would be necessary to properly mix components according to the content creator's specifications.
3. **Industry Ecosystem Incompatibility**: The industry has standardized on MPEG DASH with MSE (Media Source Extensions) for delivery, not WebCodecs. Requiring a WebCodecs-based solution would force players to implement entirely new delivery pipelines.
4. **DRM Incompatibility**: Most commercial content requires DRM protection. WebAudio cannot process protected media streams, making it unsuitable for many real-world NGA applications.
5. **Performance Considerations**: Decoding all components in a single decoder enables crucial optimizations that may be the difference between a solution working or not working on resource-constrained devices.

3.3 Why Not Deliver Components Separately?

Delivering audio components as separate streams might seem like a solution but presents significant drawbacks:

1. **Increased Complexity and Failure Points**: Separate delivery creates new failure scenarios where some components may be received while others are dropped during transmission.
2. **Content Creator Control**: The integrated approach ensures content creators' intent is preserved through metadata-defined constraints on mixing parameters.
3. **Optimization Limitations**: Having and decoding all components in one decoder allows optimizations that are otherwise not available, which can be crucial for embedded devices.

3.4 WebAudio Limitations for Immersive Audio

WebAudio's pipeline is not designed for immersive audio experiences or object-based audio rendering, making it fundamentally misaligned with NGA's capabilities:

1. **No Object-Based Audio Support**: WebAudio is designed around channels and nodes, not audio objects with spatial properties.
2. **Limited Spatial Audio**: While WebAudio has some spatial audio capabilities, they don't match the sophistication of NGA codecs.
3. **No Standardized Personalization**: WebAudio lacks standardized interfaces for the kinds of personalization that NGA enables.

Appendix: Glossary of NGA Terms

- **Audio Object**: An individual audio component with associated metadata
- **Preselection**: A pre-defined set of audio components and mix parameters
- **Prominence**: The relative gain/volume of an audio component
- **Position Interactivity**: The ability to change the spatial position of audio components
- **Object-Based Audio**: Audio represented as individual objects with spatial properties rather than fixed channels