Audio Engineering Society

Convention Paper

Presented at the 119th Convention
2005 October 7–10 New York, New York USA

Scalability in KOZ Audio Compression Technology

Kevin M. Short[1], Ricardo A. Garcia[1], and Michelle L. Daniels[1]

[1] Chaoticom Technologies*, 3 Riverside Dr., Andover, MA, 01810, USA
{kevin, rago, michelle}@chaoticom.com

ABSTRACT

Intra-codec scalability in the KOZ audio compression technology is presented in detail. The KOZ codec uses a psychoacoustic model and high-resolution spectral analysis to create, prioritize and layer audio objects, making it inherently scalable by varying the number of layers. The layers are sufficiently fine-grained to allow both small-step and large-step bitrate variations in a light-weight, real-time process during content delivery. Decoder scalability based on availability of device resources is introduced. An overview of the architecture of the KOZ technology and some of the applications of its scalability are discussed.

1. INTRODUCTION

The KOZ scalable audio technology grew out of breakthroughs in the control of nonlinear chaotic systems. KOZ technology was originally developed to enable delivery of mobile music over cellular networks to devices such as cell phones, PDAs, and portable music players [1]. It delivers full-bandwidth, high-quality audio (music and speech) at bitrates ranging from below 12kbps all the way to lossless compression. The KOZ codec is inherently scalable throughout this broad range because of its layered design, which is based on additive re-synthesis techniques distinct from most current perceptual transform-based codecs.

While the topic of scalability has recently gained increased attention in the audio coding community, the majority of standardized scalable codecs largely depend on inter-codec scalability to cover broad ranges of bitrates, cascading different codecs to achieve a "layered" output. With these codecs, small-step and intra-codec scalability has traditionally been limited to narrow ranges of bitrates [2] [3]. Unlike these standard codecs, the KOZ codec enables both small-step (less than 1kbps increments) and large-step (greater than 32kbps increments) intra-codec scalability at a very wide range of bitrates.

The increased demand for audio and multimedia content delivered with varied mechanisms at all bitrates makes scalability an important feature for the continued value of audio codecs in the real world. Many of today's music delivery services are currently either PC-based or

* Chaoticom Technologies is a division of Groove Mobile (formerly Chaoticom).

portable device-based (but not both), and they typically have libraries consisting of thousands of tracks. With the growing popularity of mobile music players, music delivery services can benefit significantly from the ability to provide access to their entire track library at bitrates appropriate for any PC, home-based music player, or portable device. Also, streaming internet radio, online news broadcasts, and podcasting services are growing in popularity, and the use of a scalable audio codec would allow such streaming content providers to deliver the best possible quality to each customer based on network congestion and the capabilities of receiving devices.

The KOZ scalable audio codec is applicable to many of these real-world scenarios where scalability is a necessity. In addition, KOZ technology has already been implemented and deployed in numerous mobile download services worldwide with carriers in countries including the UK, Norway, Hungary, Czech Republic, Poland, and Singapore, and it provides key features such as a light-weight and flexible decoder and integrated Digital Rights Management. The nature of the KOZ technology presents unique opportunities for continued development and improvement in the future.

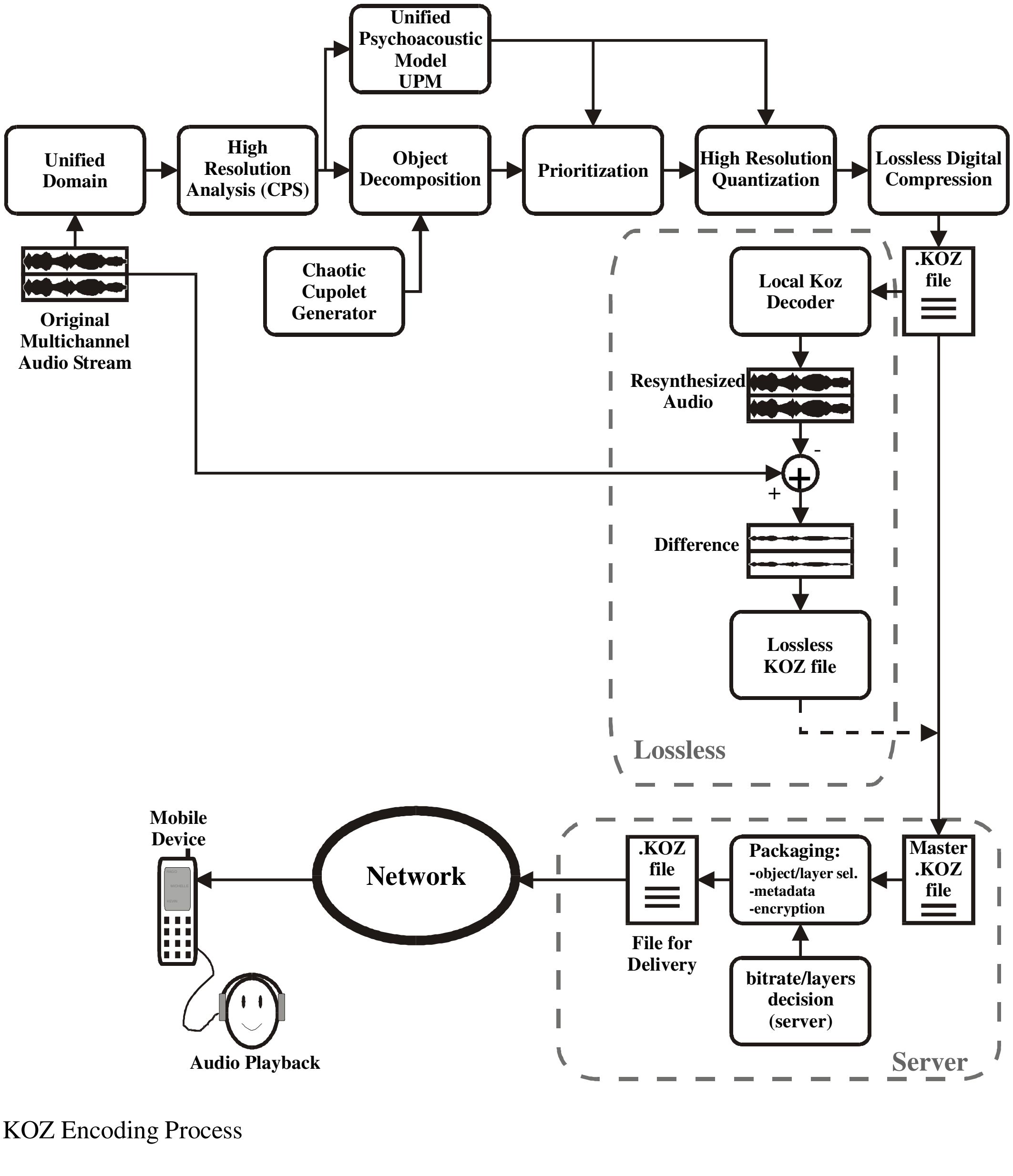

Figure 1 KOZ Encoding Process

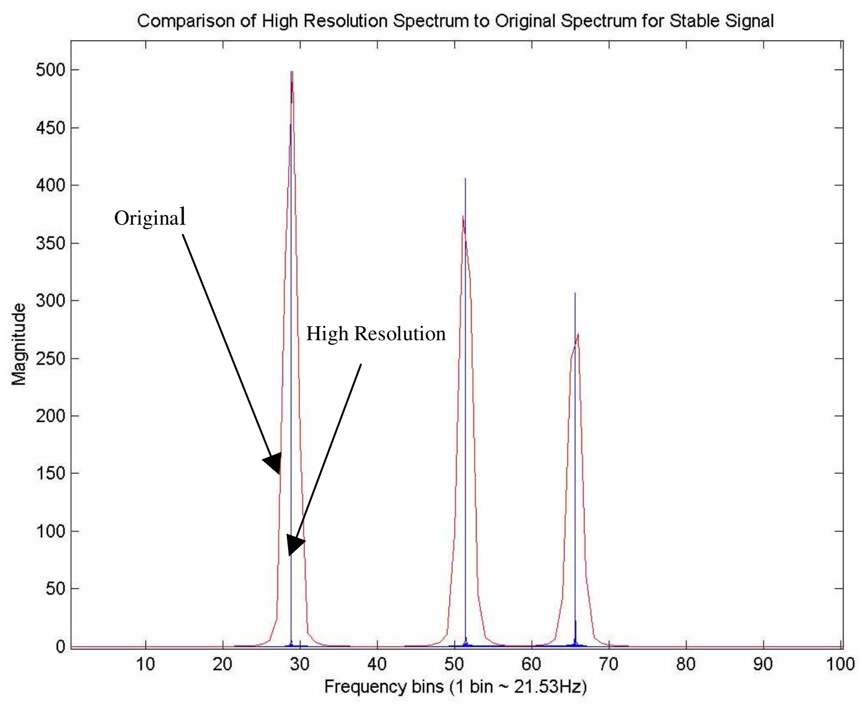

Figure 2 High-Resolution CPS Analysis

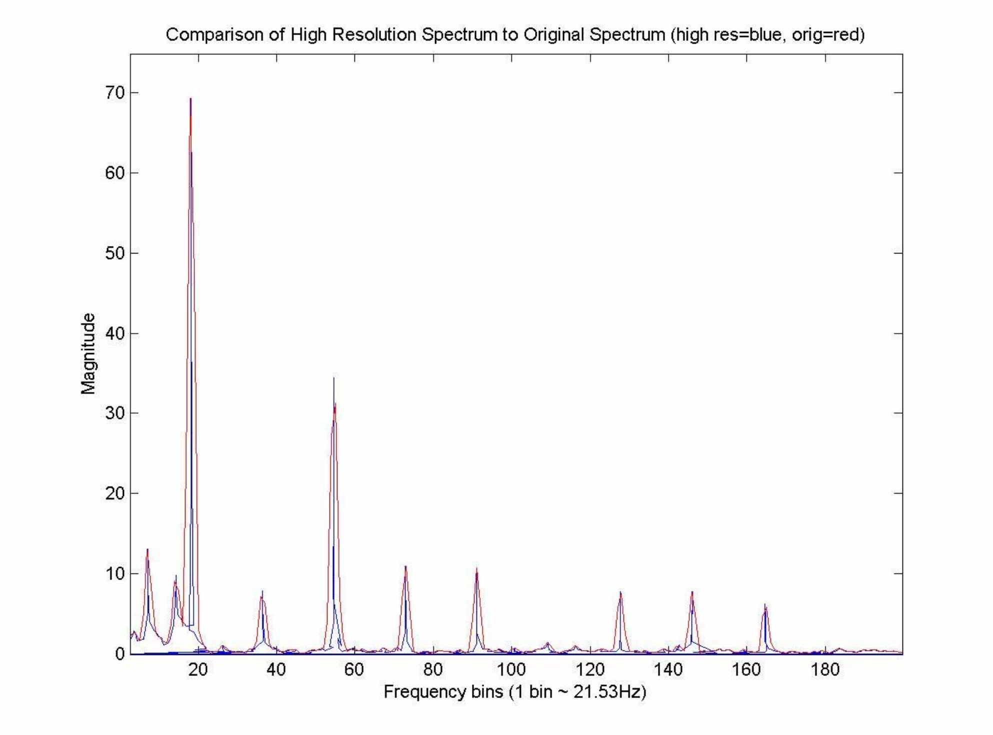

Figure 3 High-Resolution CPS Analysis of Norah Jones Track

2. OVERVIEW OF KOZ TECHNOLOGY

2.1. KOZ Encoding Process

The KOZ encoding process, outlined in Figure 1, begins with a high-resolution analysis of an audio signal in data windows of approximately 46 milliseconds (2048 samples for a typical CD sample rate of 44,100 Hz). This high-resolution stage consists of a Cross Power Spectral (CPS) analysis [4], used to differentiate tonal elements of a signal from noise-like elements, which is followed by transient analysis. Figure 2 and Figure 3 show a typical example of the outcome after the high resolution CPS analysis. Figure 2 shows the spectrum of 3 tones as estimated by an FFT, with the high resolution spectrum (now essentially a line spectrum) superimposed. In this example, the exact frequencies of the three tones are 28.7965317, 51.3764239, and 65.56498312 bins, while the estimated frequencies are 28.7960955, 51.3771794, and 65.5644420 bins.

With real-world music signals, the accuracy of the high-resolution CPS analysis of CD-sampled music is generally on the order of 0.1Hz, while the FFT resolution would be only 21.53Hz at the window size used. An example is seen in Figure 3 for a section from a Norah Jones track.

Each step of the high-resolution analysis occurs in an invertible Unified Domain representation, where independent channels of information are transformed into a multi-dimensional representation [5]. This transformation represents an arbitrary number of audio streams *n* with a single magnitude component multiplied by a complex matrix from the Special Unitary Group SU(*n*) [6], which encodes information about the spatial location of the streams and their phase relationships.

Once the transformation to the Unified Domain is performed, a Unified Psychoacoustic Model (UPM) is used to compute a Unified masking surface. This computation takes into account the spatial positions of sound sources in order to incorporate spatial masking in addition to frequency and temporal masking. While the KOZ audio codec uses the masking surface in the Unified Domain, when projected (to right and left channels, for example), the masking surface also produces masking curves consistent with standard one-dimensional masking curves used in most perceptual audio coding. An example of a UPM masking surface based on five tones appears in Figure 4.

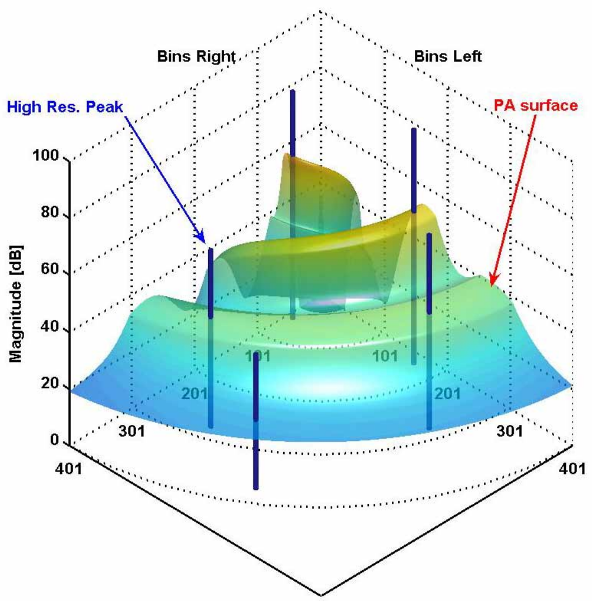

Figure 4 Unified Domain Psychoacoustic Masking Surface

Using the results of the high-resolution analysis and conversion to the Unified Domain, the audio signal is decomposed into discrete objects with mathematical representations including steady tones, noise-like elements, and transient events. These objects can be modeled with chaotic waveforms called *cupolets* in order to produce an accurate approximation of the original audio [1]. Using the UPM and principles of psychoacoustics, the cupolet objects are then ordered according to their perceptual significance and quantized. At this stage, customized variations on lossless compression techniques are applied to the quantized objects, and the result is a coded window of audio data in the fixed-point KOZ format.

Finally, a lossless KOZ window can be produced by locally decoding the lossy KOZ window and encoding the residual differences between the original window of audio data and that produced by the quantized KOZ window. These differences are embedded as an additional layer (frame) in the KOZ window using typical lossless digital compression techniques in order to improve the compression rate.

The reconstruction process from combining the signal objects produces a highly accurate representation of the audio. For example, the high resolution information in the Norah Jones track in Figure 3 can be used to reconstruct the original audio, and the frequency domain representation of the original and a 32kbps reconstruction appear in Figure 5. The differences are extremely small and are psychoacoustically insignificant.

Reconstruction is also very accurate for transient events. A time domain view of the result of the mathematical reconstruction of transient objects is shown in Figure 6, where the original signal is in blue and the reconstruction is in red.

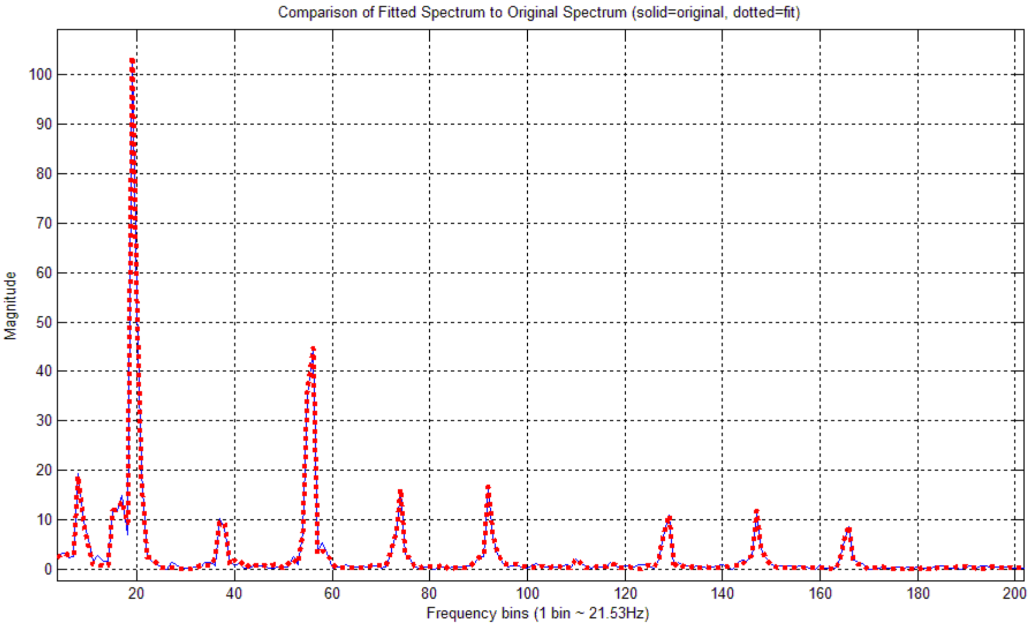

Figure 5 Encoded KOZ vs. Original Spectrum at 32kbps

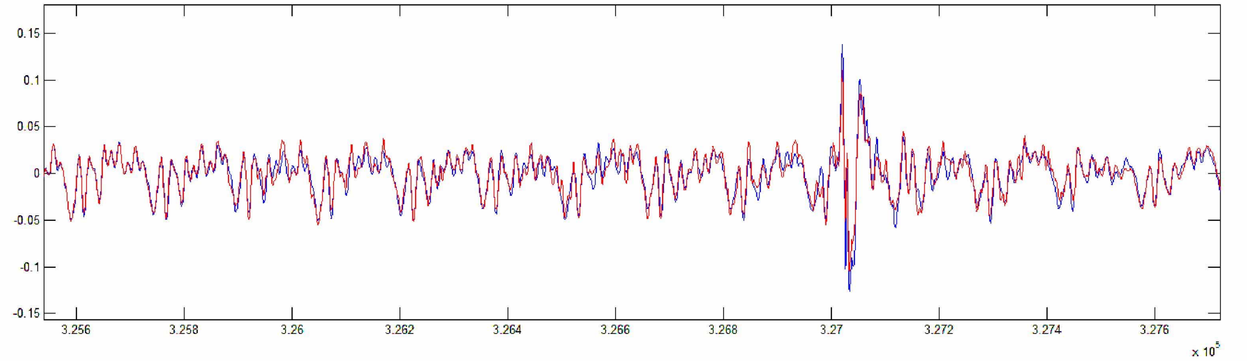

Figure 6 Transient Reconstruction at 32kbps

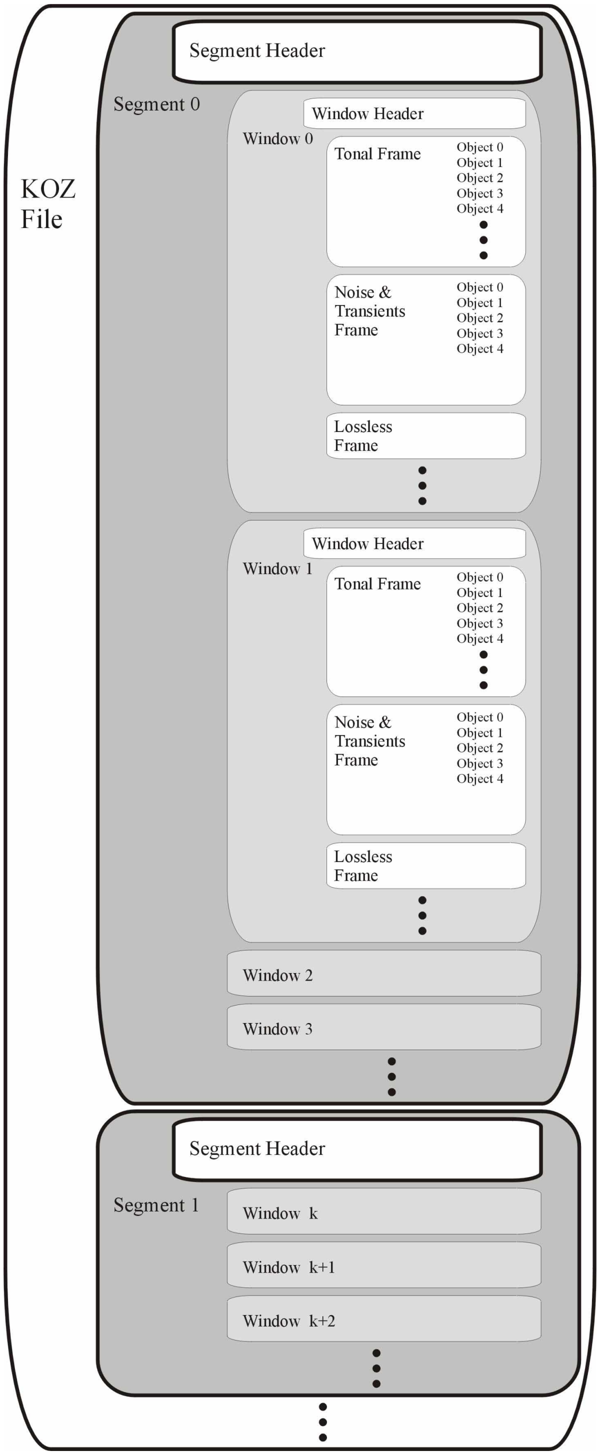

Figure 7 KOZ Bitstream Format

2.2. KOZ Bitstream Format

The structure of the KOZ bitstream, as depicted in Figure 7, is a crucial component of the KOZ scalable audio codec. The KOZ bitstream consists of nested groupings of containers, including audio objects, frames, windows, segments, segment headers, and finally a complete KOZ file.

A KOZ audio object consists of a cupolet modeling the results of the high-resolution analysis done during the KOZ encoding process. Objects in the bitstream are ordered by perceptual significance and can vary in meaning and internal structure (i.e. tonal objects, noise and transient objects).

Ordered groupings of objects of the same type are contained within KOZ frames. Types of frames include Tonal frames, Noise and Transients frames, and Lossless frames. These frames are then grouped in windows, each of which contains a header with parameters specific to that window. A window carries information in the form of frames and objects that can be decoded to produce a small segment of audio (on the order of 46 milliseconds).

Windows occur sequentially in segments, each of which begins with a segment header. Segment headers can vary in size depending on their contents, and they define parameters required for the decoding of subsequent segments until the next segment header is reached. These parameters can include file meta-data, encryption information for Digital Rights Management, global parameters related to the lossless encoding of KOZ objects, and even Huffman tables. Segments can be anywhere from one window to an entire track in length, and the break-in delay of an already-streaming KOZ file is directly correlated to the length of the file's segments. Finally, one or more segments representing sequential windows of audio data make up a complete KOZ file.

2.3. Encoder Scalability in the KOZ Codec

The object decomposition enabled by high-resolution signal analysis and the structure of a KOZ bitstream with its objects ordered by perceptual relevance make the KOZ codec inherently scalable. A "master" KOZ file can be created with a high number of object layers and the best possible quantization for each layer. The master file format allows the music to be accessed in layers, and a given number of layers can be selected to achieve a target bitrate. This selection of layers can be done in real-time and can be varied on a window-by-window basis even while streaming a track for real-time playback.

For low bitrate applications, such as delivery of music over wireless networks to cell phones, the least significant layers (and even entire frames, such as lossless frames) are omitted before transmission of the file, with the selection of transmitted layers based on the capacity of the transmission channel or the storage constraints imposed by the wireless device. It is also possible to vary the quantization depth of the underlying parameters based on psychoacoustic information such that at high bitrates a highly accurate waveform reconstruction is achieved, while at lower bitrates slightly less accuracy is required. Varying the quantization in this manner can improve the performance of the lossless coding employed for KOZ objects.

This scalability can produce a variable bitrate output on a window-by-window basis, with no knowledge of previous or future windows required. Because the Unified Domain already provides the KOZ codec with an extremely compact representation of spatial aspects of an audio signal, it is not necessary to decrease the stereo image of a KOZ file in order to decrease the bitrate. Similarly, the chaotic elements used in the KOZ codec to represent objects can cover broad spectral regions with very few bits, and therefore it is also not necessary to reduce the bandwidth of a KOZ file in order to decrease the bitrate. In this way, a full-bandwidth KOZ file with an accurate spatial image can be delivered even at the lowest bitrates.

Scalability at delivery time eliminates the need for multiple copies of each track to exist on the server with different bitrates or formats. This kind of scalability is a very lightweight process – the server is not required to decode the master KOZ file, only parse it and re-package it for delivery. Truncation and removal of layers requires only computations comparable to the frame-truncation methods described in other approaches, and minimal calculations are involved with changing quantization levels. Individual layers can be removed to allow small-step scalability in increments of less than 1kbps, or many layers can be removed at once in order to produce large steps in a wide range of sizes (up to 32kbps or more).

2.4. Decoder Scalability in the KOZ Codec

The KOZ decoder itself also contributes to the scalability of the KOZ audio codec. For client devices with either limited processing power or specific requirements for number of channels, bit depth, or sampling rate of the output audio, it may be necessary to make tradeoffs between processing power and perceived audio quality at decoding time. The KOZ decoder can maximize quality while reducing the number of CPU operations by limiting the number of objects decoded in each frame (only decoding the most perceptually relevant objects) or by changing the accuracy of the reconstruction of individual objects. Because of the underlying mathematical representation, discrete KOZ objects can be reconstructed at any sampling rate or bit depth, and they can be projected to any number of output channels.

For client devices with sufficient processing power, KOZ objects can also be decoded to support more output channels than the original source file contained (up-mixing). For example, because the Unified Domain objects carry specific information about their spatial position, a stereo file can be up-mixed to surround sound.

3. POSSIBLE APPLICATIONS OF KOZ SCALABILITY

There are many possible applications of the KOZ scalable audio codec. For anyone seeking to provide optimal media delivery over a broad range of networks and/or to a diverse set of devices, scalability is essential.

For example, one potential application of the scalable KOZ technology is a “networked home jukebox” scenario, where a central content server on a home network stores scalable “master” KOZ files which can be simultaneously streamed or downloaded to a variety of wired and wireless devices such as a home stereo system, a desktop or laptop computer, or a portable music player. Because each of these devices has different processing capabilities and a different amount of available storage, in order to deliver the best possible quality to each device, the central server must be able to scale its master files to any given target bitrate. The server must also consider the overall traffic on the home network in order to determine the constraints on delivery bandwidth for real-time streaming audio playback. A priority could even be assigned to each client device connected to the home server to influence bandwidth distribution.

Another potential application of the scalable KOZ technology is a “wide area network jukebox” scenario, where one or more servers delivers content over many sub-networks to a wide range of devices with many different communication mechanisms (e.g. DSL, GPRS, wireless LAN). In this case, the content server monitors its network status and determines the amount of bandwidth to allot to each client device depending on fluctuations in network traffic and statistics measured by the server. For the most part, this scenario requires the same scalability features as the “networked home jukebox” scenario, but because of the device and network diversity, it demands a broader range of possible bitrates. In both cases, the servers act as central content storage devices with the capability of establishing, monitoring, and maintaining content delivery connections to multiple clients simultaneously. A content server storing scalable “master” KOZ files for each track in its library may monitor network conditions, and can scale the bitrate of transmitted files “on the fly” on a window-by-window basis, or at the beginning of a connection, in order to meet various Quality of Service targets.

Within these and other similar content distribution scenarios, there are multiple types of interactions between client devices and the content server which can be greatly facilitated by the KOZ technology. These include the following:

3.1. Full-Track Download

If a client requests a file which is to be downloaded in its entirety before playback, then the server will determine the ideal bitrate for the client device based on the device’s storage capacity and the network bandwidth available for the connection. The master KOZ file is then parsed and downloaded to the client using the available connection, while the client stores the entire file for later listening. Because real-time streaming playback during the download is not required in this case, it is not necessary for the server to change the bitrate of the downloading track on a window-by-window basis. Therefore any bitrate can be set at the beginning of the download, and minimal processing is required on the part of the server. In this case, the ability to scale one KOZ master file to any bitrate eliminates the need for multiple copies of each available track living on the server at a wide variety of bitrates in order to accommodate different devices.

3.2. Full-Track Streaming Playback

If the client requests a file for streaming playback in real time, then the server must not only consider the client device’s capabilities and the type of connection from the server to the device. It must also monitor the available network bandwidth during the streaming process and vary the bitrate of content delivered to the client in order to insure maximum possible quality without breaks in the audio. With the KOZ scalable audio codec, a master KOZ file can be processed window-by-window in real time while it is streamed to the client device in order to provide optimal audio quality and distribution of network resources.

When new network Quality of Service requirements are presented (e.g. a connection from the server to the client becomes unstable due to interference or lack of signal strength, or an additional demand is placed on the network because of increased bandwidth usage from other devices), it may be necessary to reduce the bitrate of the streamed KOZ file on the fly by truncating layers from the master KOZ file. Similarly, if the available bandwidth increases, the server can send more layers from the KOZ file to the client. With the KOZ codec, these decisions can be made on a window-by-window basis with no dependence on past or future network conditions. For example, the perceptually least significant layers of objects in each KOZ frame can be truncated and removed from the bitstream as it is parsed and passed on to the client device. In this way, the bitrate of the resulting KOZ stream can be varied by small or large steps (from less than 1kbps to greater than 32kbps steps). This is a very light-weight process and the server can perform this function for many devices, with each device playing different tracks simultaneously. Also, with minimal calculations, the quantization levels of the remaining KOZ layers can be decreased compared to the master file during the same parsing and delivery process.

If the client also wants to store the file for later playback while it is being streamed, the server must also consider the storage limitations of the device when determining bitrate. In this case, however, a significant advantage of the layered nature of the KOZ codec is that, if desired, it allows additional supplemental layers of information to be sent after the streaming is complete. In this way, if the client device has enough available storage space, a track can be streamed and played at a low bitrate and later enhanced with additional layers to create a higher-quality file for later playback.

3.3. Streaming Playback from Existing Stream

The client is not limited to downloading full tracks. It can also "break-in" to an existing multicast stream (e.g. the content server is already streaming a radio broadcast or live performance). In this scenario, a client must wait for the next segment header of the KOZ stream, containing crucial information for proper decoding, to be sent before decoding can begin (see Section 2.2). When the client requests a complete track, then the server can send a KOZ file with only one segment header at the beginning of the file, and playback can begin immediately. However, in the break-in scenario, the server must send segment headers at any desired regular or irregular interval. In that case, the break-in delay before a client can begin decoding the KOZ bitstream is related to the number of windows sent per segment. Each time a segment header is sent, the bitrate of the KOZ stream increases slightly. Because of this, it is necessary to achieve a balance between the size of each segment and the desired break-in delay of a stream.

4. ADVANCED FEATURES OF THE KOZ TECHNOLOGY

4.1. Audio Processing Techniques

Each high-resolution KOZ object has a well-defined mathematical representation, allowing many audio processing techniques to be easily applied to a file in the compressed KOZ format without decoding to PCM data beforehand. These methods can be applied during the decoding process to produce output PCM data, or they can be applied to create a modified KOZ file without re-encoding from the original source file. These processing techniques include:

- Up-mixing to multiple channels from the Unified Domain
- Down-mixing from multiple channels to stereo or mono
- Up-sampling and Down-sampling to any desired sample rate
- Reconstruction at different bit depths, e.g. 16-bit or 24-bit audio
- Pitch-Shifting
- Time-Stretching

As described in Section 2.1, the representation of an audio signal as a collection of high resolution objects in the Unified Domain contains clear information describing the spatial position of each object. For example, if a stereo track is encoded, the audio objects are represented with a single source magnitude at a spatial position, with corresponding phase information representing the phase state of the source signal as it hits the transducers. This Unified Domain data therefore contains all of the information necessary to project an audio object onto any number of output channels for any speaker configuration (or for headphones). Because mono, stereo, or multi-channel source material can all be represented in the Unified Domain in the same fashion, accurate up-mixing and down-mixing becomes trivial during the decoding process. For example, a stereo source signal encoded with the KOZ format can be decoded in stereo by projecting the magnitude of each KOZ object onto left and right channels. The same stereo stream could be decoded in mono with a projection to the center channel, or it could be up-mixed to 5.1 channel surround sound by projecting each object onto multiple output channels. A special down-mixing technique from stereo to mono has been used with high success. Using knowledge of the spatial position of the sources, the output stream is focused on the dominant event position for every window. This technique has the benefit of emphasizing the most significant audio objects regardless of where they were located in the original audio stream.

KOZ files encoded from source material at any sampling rate can also be easily decoded at any other sampling rate. Up-sampling and down-sampling at decoding time is simplified by the mathematical representation of KOZ audio objects, which allows the functional objects to be sampled in any desired manner during the reconstruction process.

Re-synthesis of KOZ audio objects at different bit depths is trivial because the mathematical representations of the objects can be performed with any output precision.

Pitch shifting and time stretching algorithms require objects to be re-synthesized with different parameters for frequency and duration in time. The decomposition of audio into tonal, noise-like and transient objects allows KOZ files to be easily pitch shifted and/or time stretched at reconstruction time. A pitch-shifted audio stream has the same duration as the original audio stream, but its components are shifted in frequency in an ordered manner (for example, by a whole number of semi-tones). The mathematical representation of each type of KOZ object allows efficient pitch-shifting by simple transformations altering the parameters describing each object. Time-stretched audio streams preserve the pitch of the original stream, but their duration is either longer or shorter than the original. Both processes require the same type of signal decomposition in tone-like, noise-like and transient elements. KOZ files and the efficient mathematical reconstruction algorithms for KOZ audio objects make these kinds of processes very straightforward.

4.2. Decoding in Either Frequency or Time Domain

The signal objects used in the KOZ codec have equivalent representations in both the frequency and time domains, and therefore the decoding process can operate in either domain. In particular, for extremely limited devices with slow processors, decoding in the time domain (or partial decoding of a reduced-resolution signal) has proven to be very effective.

4.3. KOZ Video and Image Compression

Just as audio signals can be decomposed into discrete objects that can easily be represented by chaotic cupolets, both still image and video signals can also be encoded in this manner. Streaming, variable bitrate, high-quality image and video compression with the KOZ technology retains the scalable properties of the KOZ audio codec.

5. CONCLUSIONS

The KOZ audio compression technology provides efficient intra-codec small-step and large-step scalability at a broad range of bitrates. This scalability is enabled by the high-resolution analysis performed during the KOZ encoding process, which produces layered audio objects ordered by perceptual significance within a KOZ file. The ability to scale master KOZ files from any given bitrate (including lossless) to any lower bitrate makes the KOZ technology a powerful tool for a wide variety of media content storage and distribution models. Added functionality allowing the KOZ decoder to apply various audio processing techniques or to scale its functionality based on the capabilities of each client device makes this scalability practical for a large number of wired or wireless, portable or fixed platforms including home stereo systems, portable music players, and personal computers. KOZ technology has already been implemented and deployed as a key enabler in many mobile music download services worldwide. Because of the KOZ technology's unique analysis and decomposition processes, it has the potential to continue to expand and improve in many ways, making exciting new applications both practical and possible.

6. REFERENCES

[1] K. M. Short, R. A. Garcia, M. Daniels, J. Curley and M. Glover, "An Introduction to the KOZ Scalable Audio Compression Technology", AES 118th Convention Paper, Barcelona, May 2005, Preprint 6446.

[2] ISO/IEC JTC1/SC29/WG 11 N7040, "Call for Information on Scalable Speech and Audio Coding", January 2005, Hong Kong, CN.

[3] M. Bosi and R. E. Goldberg, *Introduction to Digital Audio Coding and Standards,* Kluwer Academic Publishers. (2003)

[4] D. J. Nelson and K. M. Short. "A channelized cross spectral method for improved frequency resolution. Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis," IEEE Press, October 1998.

[5] K. M. Short, R. A. Garcia and M. L. Daniels, "Multi-channel Audio Processing Using a Unified Domain Representation", AES 119th Convention Paper, New York City, October 2005.

[6] Bhandari, "Polarization of Light and Topological Phases," Physics Reports 281 (1997) 1-64.