Audio Engineering Society

Convention Paper

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

A New Broadcast Quality Low Bit Rate Audio Coding Scheme Utilizing Novel Bandwidth Extension Tools

Deepen Sinha[1], Anibal J. S. Ferreira[1,2]

[1] *ATC Labs, New Jersey, USA*

[2] *University of Porto, Portugal*

Correspondence should be addressed to D. Sinha (sinha@atc-labs.com)

ABSTRACT

In this paper we describe the components of a novel audio coding algorithm capable of delivering *high-fidelity* CD-like stereo audio at the bit rates of 40-48 kbps and *natural sounding* FM grade mono at the bit rates of 18-22 kbps. Bandwidth Extension has emerged as an important tool for the satisfactory performance of low bit rate audio codecs. Recently we proposed two new bandwidth extension algorithms, Fractal Self-Similarity Model (*FSSM*) and Accurate Spectral Replacement (*ASR*), which belong to a new class of Bandwidth Extension techniques which are applied directly to the high resolution frequency representation of the signal (e.g., MDCT or ODFT). The proposed coding scheme uses *FSSM* and *ASR* in an adaptive and complementary framework. Another important component of the proposed codec is a wideband psychoacoustic model that makes an explicit use of the Comodulation Release of Masking (CMR) phenomenon. It also includes a novel parametric stereo coding technique. The proposed audio coding scheme is geared towards broadcast applications where codec latency and encoder complexity is generally not an overriding concern. In this paper we present algorithmic details of the new codec, audio demonstrations, and, comparison to other audio coding schemes. Further information and audio demonstrations are available at http://www.atc-labs.com/teslapro.

1. INTRODUCTION

In many audio broadcast applications the need for higher bandwidth efficiency is being continually felt. This is driven by many factors such as the desire to deliver more content to the listener (e.g., in Satellite and Terrestrial Digital Audio Broadcasting), or to provide content using newer media (e.g., cellular networks), or in some cases a need is being felt to improve audio quality in an existing low bit rate audio broadcast service. There appears to be a proliferation of applications demanding CD quality stereo at bit rates of

48 kbps and lower and high quality FM grade mono audio at bit rates of 20-24 kbps. These in turn continue to spur the demand for newer algorithms for audio bit rate reduction.

The field of Perceptual Audio Coding has matured over last several years and a number of audio coding technologies exist. These include proprietary schemes such as PAC (Bell Labs, Lucent) [1] and ATRAC (Sony) [2] as well as standard based codecs such as MPEG-1 Layer 3 (popularly known as MP3) [3], MPEG-2 AAC [4], Dolby AC-3 [5]. In general the established audio coding schemes fit into the framework of adaptive transform or sub-band coding whereby the output of a filter bank is quantized using quantizers driven by a perceptual model. The Modified Discrete Transformation (MDCT) [6] is a popular choice for a codec filter bank. In broadcast applications, where codec delay is not an issue the use of a high resolution MDCT (i.e., with 1024 frequency sub-bands) has become the norm and has led to higher coding efficiency in algorithms like PAC and AAC. At best these conventional audio coding techniques are capable of producing full fidelity CD quality audio in the range of 96-128kbps. Furthermore, near-CD quality audio with somewhat lower audio bandwidth (~ 15 kHz) and limited stereo is achievable in the range of 48-64 kbps.

In order to reduce the bit rate requirement further (to meet the demand for bandwidth efficiency as noted above), several parametric approaches have been proposed. These rely on a compact parametric description of all or a portion of the audio signal. One such approach that has proven to be particularly effective is the so called "Bandwidth Extension" approach. In Bandwidth Extension only a low pass filtered version of the signal is directly coded using the conventional perceptual coding paradigm. The high frequency portion of the signal spectrum is recreated at the decoder by a mapping generated from the low frequency spectrum of the signal. Typically an attempt is made to match the reconstructed high frequency spectrum to the original high frequency spectrum as closely as possible. In practice significant mismatch may remain between the two. However, the philosophy is that increased naturalness of the higher audio bandwidth signal compensates for any other perceived distortion in the (reconstructed) higher frequencies.

In [7] and [8] we introduced two novel techniques for bandwidth extension called Fractal Self-Similarity Model (*FSSM*) and *Accurate Spectral Replacement* (*ASR*) respectively. These technique offer the promise of a more accurate reconstruction of the synthesized high frequency spectrum in comparison to previously reported approaches such as the Spectral Band Replication approach [9]. We have since utilized the techniques in building two new audio coding schemes. In this paper we focus on one of the new coding schemes geared towards *broadcast applications*. The term broadcast applications signifies a class in which codec latency and *encoder* complexity is not an overriding concern. This audio coding scheme called *TeslaPro*, where Tesla stands for "Transform Domain Excitation and Self-Similarity Accumulation" incorporates several techniques for higher audio quality at lower bit rates. In particular:

- A Bandwidth Extension model that makes use of both the *FSSM* and *ASR* algorithms in an adaptive framework and also allows for the combination of aspects of *FSSM* and *ASR* synthesis.

- A new wideband psychoacoustic modeling scheme that explicitly takes into account some of the wide band phenomenon in masking and also employs a model for accurate detection and estimation of tonal components.

- A parametric stereo coding technique for higher efficiency in stereo coding.

The organization of the rest of the paper is as follows. In section 2 we take a closer look at the overall structure of the codec. Three of the key components of the coding scheme, its Psychoacoustic Model, Bandwidth Extension Model, and, Parametric stereo coding approach are described in sections 3, 4, and, 5, respectively. Audio coding results and the performance of the codec at various bit rates is discussed in section 6, followed by conclusions in section 7, and acknowledgements in section 8.

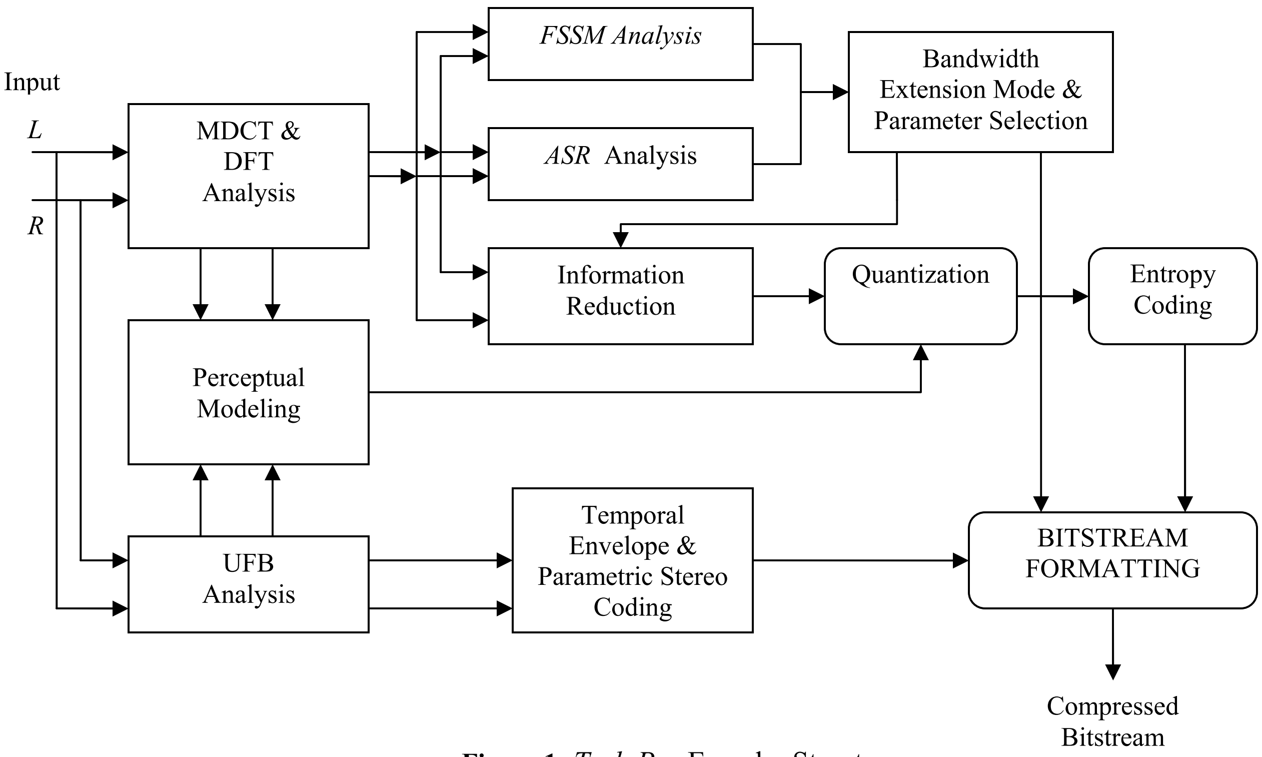

Figure 1: *TeslaPro* Encoder Structure

2. STRUCTURE OF THE CODEC

The structure of TeslaPro encoder and decoder is shown in Figures 1 and 2 respectively. At the encoder the input is analyzed in 3 domains: MDCT, DFT, and, *UFB* (*Utility Filter Bank*, see section 4 for details). For MDCT analysis one of a new class of window functions is utilized [10]. A slew of bandwidth extension tools are optionally employed depending upon the desired compression efficiency. The encoder consists of 3 broad functional areas:

- Perceptual coding including psychoacoustic modeling, quantization, and entropy coding
- Bandwidth Extension related processing
- Temporal Envelope and Parametric Stereo Coding

The decoder, as shown in Figure 2, similarly contains processing blocks corresponding to the above 3 functional areas. In the next few sections we'll take a closer look at some of the important components of *TeslaPro.* In particular its Psychoacoustic model, bandwidth extension methodology, and parametric stereo coding schemes will be elaborated upon.

3. PSYCHOACOUSTIC MODEL

The *TeslaPro* codec is a perceptual coding scheme whereby an elaborate psychoacoustic model is employed to quantize the output of an analysis filter bank. In this section we describe some of the unique components of the Psychoacoustic model used in this codec.

The field of psychoacoustic modeling for audio coding has been an active one over the past two decades. Typical configuration for the perceptual model used in audio codecs such as PAC, AAC, MPEG-LayerIII etc. may be found in [1-5]. The centerpiece of perceptual modeling is the concept of auditory masking [11 – 15, 27]. The goal is to quantize the audio signal in such a way that the quantization noise is either fully masked or rendered less annoying due to masking by the audio signal. Building of a perception model in audio codec typically involves the utilization of following four key concepts: *simultaneous masking*, *temporal masking*, *frequency spread of masking*, and, *tone vs. noise like nature* of the masker. Simultaneous masking is a phenomenon whereby a *masker* is found to mask the perception of a *maskee* occurring at the same time. Temporal masking refers to a phenomenon in which a *masker* masks a *maskee* occurring either prior to or after its occurrence. Frequency spread of masking refers to the phenomenon that a masker at a certain frequency has a masking potential not only at that frequency but also at neighboring frequencies. Finally, the masking potential of a narrow band masker is strongly dependent on the tone vs. noise like nature of the masker. These factors are utilized to estimate desired quantization accuracy, or Signal to Mask Ratio (*SMR*) for each band of frequency.

The two key aspects of *TeslaPro* Psychoacoustic model pertains respectively to the extension of a narrow band masking model to wide band audio signals and to the accurate detection of tonal components in the signal. These are described below.

Comodulation Release of Masking (CMR) and Implications for Wideband Perceptual Modeling

In many audio codecs the masking model for wideband audio signals is constructed using a two step procedure. First the (short-term) signal spectrum is analyzed in multiple partitions (which are narrower than a critical band). The masking potential of each narrow-band masker is estimated by convolving it with a *spreading function* which models the frequency spread of masking. The masked threshold of the wide band audio signal is then estimated by considering it to be the superposition of multiple narrow band maskers. Recent studies suggest that this assumption of superposition may not always be a valid one. In particular a phenomenon

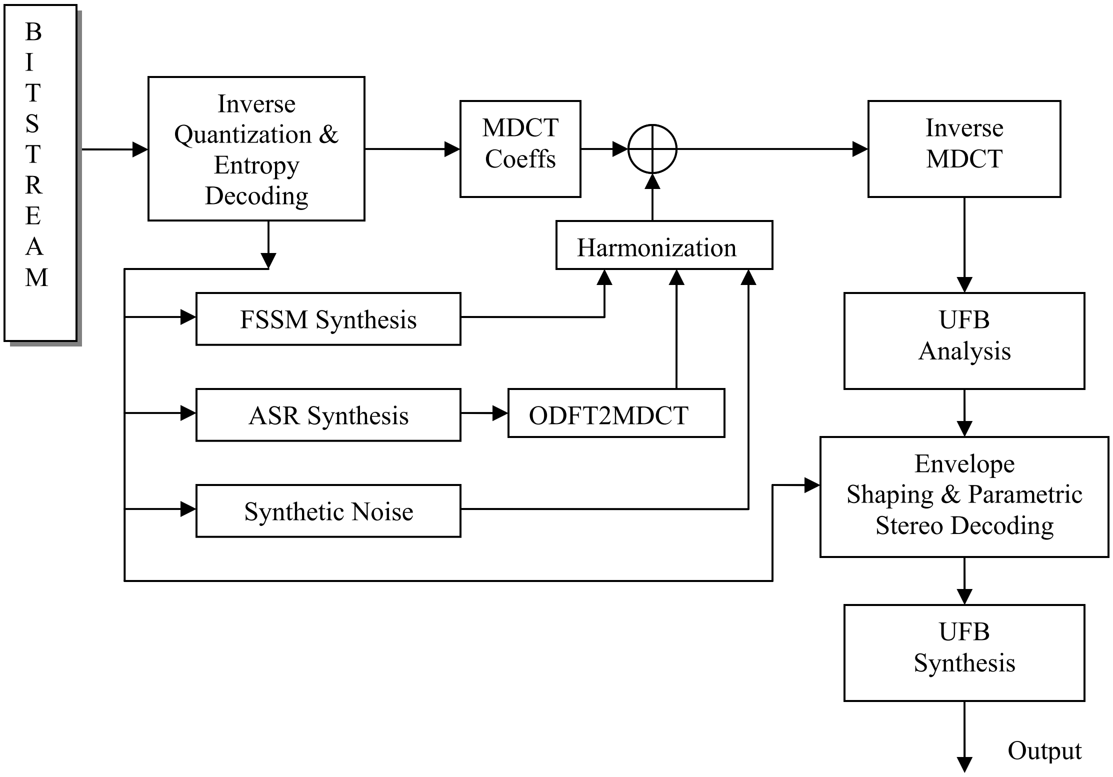

Figure 2: *TeslaPro* Decoder Structure

called Comodulation Release of Masking (*CMR*) has implication towards the extension of narrow band model to a wide band model [12, 16].

CMR is a phenomenon that describes reduced masking by a wide band (bandwidth greater than a critical band) noise like signal which is coherently amplitude modulated (*comodulated*) over the entire spectrum range. The reduction is masking has been variously reported to be between 4.0 dB to as high as 18 dB [17]. The exact physiological phenomenon responsible for CMR is still being investigated by various researchers. However, there is some evidence that *CMR* occurs due to a combination of multiple factors. It has been hypothesized that the masking release results from cues available within a critical band and from cues generated by comparisons across critical bands. In audio codecs this implies that superposition of masking does not hold in the presence of strong temporal envelope and masking of wide band signals can be significantly lower than the sum of masking due to individual narrow (sub-critical) band components depending upon the coherence of their temporal envelopes. It is tempting to think that *CMR* can be accounted for through adequate temporal shaping of the quantization noise (since the masking threshold during the dips in envelope is very likely to be lower), but experiments by Hall et al. [16] indicate that (the lack of) temporal shaping of *maskee* does not explain all (or most) of the *CMR* phenomenon. In particular masking release of about 4-8 dB should be accounted for directly in the Psychoacoustic Model.

The *TeslaPro* Psychoacoustic model incorporates a model for *CMR* which takes into account: (i) the effective bandwidth of the i^{th} critical band masker, EBM_i defined as

$$EBM_i = \frac{1}{2N} \sum_{j=i-N}^{\substack{j\neq i \\ j=i+N}} < \phi_i . \phi_j > \qquad (1)$$

where ϕ_i and ϕ_j are respectively the normalized temporal envelopes of i^{th} and j^{th} critical band maskers (a suitable value for N is about 5); and, (ii) dip in the temporal envelope of the masker, ρ (defined as the ratio between the minimum and maximum of the temporal envelope of the maker in a 20-30 *msec* window). Estimation for the reduced masking potential of the narrowband masker, *i, (CMRCOMP$_i$)* is the made as below

$$CMRCOMP_i = -10\log_{10}[\rho / N(EBM_i)] \qquad (2)$$

Where $N(\alpha)$ is a non-linearity and the *CMRCOMP$_i$* value in (2) is saturated to a minimum of 0 *dB* (a piecewise linear function with a linear rise for α below 0.7 and above 0.8 and rapid rise angle of over *80^o* for α between 0.7 and 0.8 was found suitable in our experiments). Partial support for this model is based on data in [17] and is supported by listening data based on expert listeners. The estimated *CMR* compensation is utilized when combining the masking effect of multiple bands.

<u>*Accurate Tone Detection*</u>

The *TeslaPro* codec utilizes the algorithm described in [18, 20] for the detection and accurate parameter estimation of sinusoidal components in the signal. The detected sinusoids are further analyzed for the presence of harmonic patterns using techniques similar to [19]. The output of this harmonic/tone detection and estimation algorithm is utilized in the perceptual model and also in the bandwidth extension tools (section 4). In terms of the perceptual model the position (frequency) of the detected tonal components and harmonic structure(s) plays an important role in determining the desired coding accuracy of frequency bands.

4. BANDWIDTH EXTENSION TOOLS

As noted above, the proposed coding scheme utilizes two bandwidth extension tools. Here we provide a high level description of the two tools *FSSM* and *ASR*. For a detailed description of *FSSM* the reader is referred to [7], similarly, a detailed description of *ASR* may be found in [8].

In this section we look into the two tools in the context of a bandwidth application. *ASR* and (an extension of) *FSSM* may also be used to selectively code the base band components, however, that is outside the scope of this paper. The bandwidth extension paradigm may be formalized as below.

- It is assumed that in each audio frame, the spectral representation of the signal (such as the MDCT representation) up to certain frequency f_c, denoted as $X_{LP}(f)$, is coded directly using efficient quantization and coding techniques.

- The MDCT spectrum for frequencies $f > f_c$ is to be reconstructed using a mapping BE such that

$$\overline{X}_{HP}(f) = BE(\overline{X}_{LP}(f)) \tag{3}$$

Where, $\overline{X}_{LP}$ is the quantized baseband and $\overline{X}_{HP}$ is the reconstructed higher frequencies in MDCT domain.

4.1. *FSSM* Model

In the *FSSM* technique high frequency components of the signal are reconstructed using an iterative sequence of *Expansion Operators* (EO) as below,

$$\overline{X}_{HP}(f) = \cdots EO_i \circ (\cdots (EO_1 \circ (EO_0 \circ \overline{X}_{LP}(f)) \cdots) \tag{4}$$

Where each expansion operator EO_i is assumed to have the form

$$EO_i \circ \overline{X}_{LP}(f) = H_i \bullet X_{LP}(\alpha_i f - f_i) \tag{5}$$

where α_i is a dilation parameter ($\alpha_i \leq 1$) and f_i is a frequency translational parameter. H_i is a high pass (brick-wall) filter with a cutoff frequency $f_c^{\ i} = \alpha_i * f_c^{\ (i-1)} + f_i$, with $f_c^{\ 0} = f_c$. This sequence of nested expansion operators resulting in bandwidth expansion is described further in [7]. The dilation/translation equations suggest a Fractal like Model for *FSSM* which is able to reconstruct the high frequency spectral details with a high level of accuracy across a wide range of different audio signals.

The significance of the dilation and translation terms in *FSSM* is illustrated with the help of coding examples in Figures 3 (a), (b), (c). For example, the translation term improves the accuracy of reconstruction for musical instruments with a pitch structure and also for voiced speech and vocal signals. For these classes of signals the lack of dilation terms results in a discontinuity in the pitch structure. This is illustrated in Figure 3 (a) and (b). Figure 3(a) shows the reconstructed spectrum superimposed over the original spectrum using a different bandwidth extension scheme (such as the spectrum replication approach of [9]).This is compared against the reconstruction using the *FSSM* model as shown in Figure 3(b).

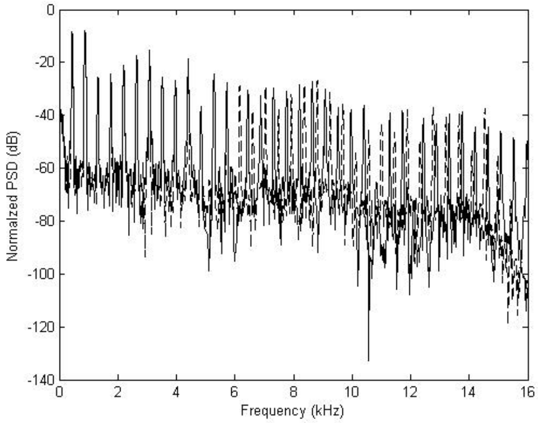

Figure 3(a): Reconstructed signal spectrum (solid line) and original spectrum (dashed line) using a spectrum replication approach.

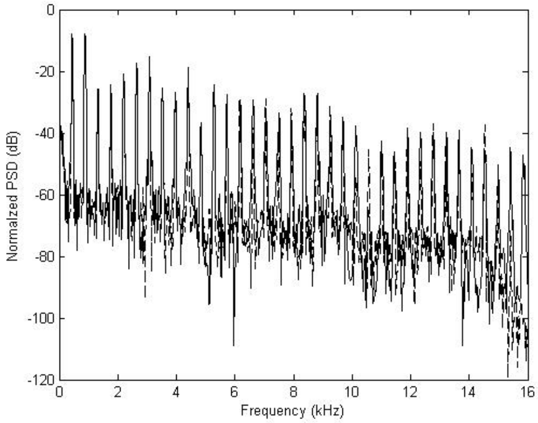

Figure 3(b): Reconstructed signal spectrum (solid line) and original spectrum (dashed line) with the *FSSM* model.

The inclusion of dilation parameter on the other hand leads to accurate signal spectrum reconstruction for a different class of audio signals, in particular for cases when the pitch structure is either not present in (part of) high frequencies or is more diffuse towards the higher frequencies. Example of a signal (“Aria”) that benefits from the inclusion of the dilation terms in *FSSM* is shown in Figure 3(c).

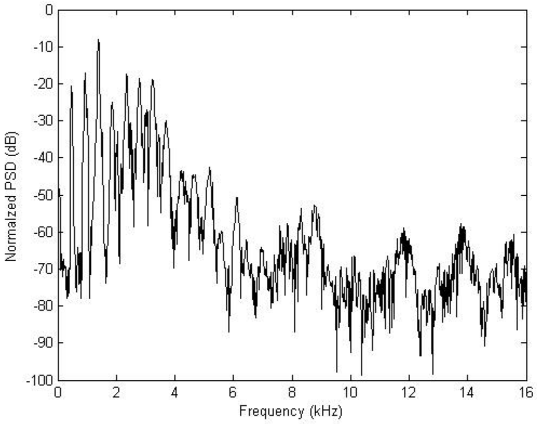

Figure 3(c): Example of a signal (short-term PSD) that benefits from the inclusion of the dilation term in the *FSSM* model.

The FSSM model parameters are estimated using a combination of 3 criteria: (1) Maximization of a Self-similarity coherence (*SSC*) function as defined below:

$$\Phi(\alpha_i, f_i) = \left\langle X\ (f) \cdot X(\alpha_i f - f_i) \right\rangle \qquad (6)$$

(2) A harmonic continuity criterion to ensure the accuracy of the dominant harmonic structure in the signal, (3) Consistency criterion over time (multiple audio frame) to ensure steady alias-free reconstruction of steady harmonics. Furthermore, the quality of the estimates improves significantly if the MDCT spectrum is normalized by the coarse envelope prior to the estimation of these parameters.

The *FSSM* model in general is a *FSSM+Isolated Tones+Noise* model. Part or all of short term signal spectrum which does not fit the *FSSM* model above is modeled either as isolated (non-harmonic) tones or synthetic noise which is added to the reconstructed high frequency signal.

An interesting observation related to the *FSSM* model is that the temporal envelope of the reconstructed high frequency components using the *FSSM* model shows a high level of coherence with the temporal envelope of the base band components. This observation is illustrated with the help of a synthetic narrowband noise signal in Figure 4. The figure shows the base band signal (Figure 4a), the *FSSM* constructed high frequency signal (Figure 4b) and the Hilbert envelopes of the two signals superimposed on each other (Figure 4c).

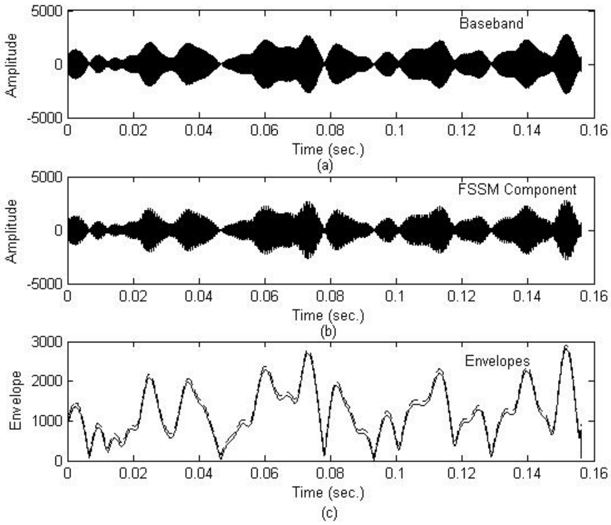

Figure 4: (a)Base band noise signal, (b)*FSSM* constructed high frequencies, (c) Envelopes of (a) & (b)

4.2. The *ASR* Bandwidth Extension Model

The *ASR* Model for bandwidth extension is described in detail in [8]. It takes into account the specificity of the coherent (i.e., sinusoidal) components of an audio signal, as well as the specificity of the incoherent (i.e., noise) components of an audio signal, namely with respect to their different perceptual impact and their different spectral nature and fine spectral structure. At the heart of *ASR* is a sinusoidal analysis and synthesis algorithm *with sub-bin accuracy*. The *ASR* model is particularly effective when the audio signal exhibits a well defined harmonic structure of sinusoids. In this case a bandwidth extension technique based on the replication of base band components may not provide satisfactory reconstruction of higher order partials. A replication model in this case, as noted above, has a significant deficiency in the sense that it may either break the organization of the harmonics in frequency which is likely to be noticeable to the human auditory system in the form of a pitch shift or the appearance of several pitches instead of a single one. *ASR* also allows sufficient and flexible control over the phase of the synthesized higher order partials which may not be possible in techniques utilizing mapping based on the lower frequencies (base band). The most general form of *ASR* processing consists of the following steps.

1. Normalization of the audio spectrum by a model of the smooth spectral envelope, the noise part of the resulting flattened spectrum is very approximately white.

2. Segmentation of the flattened spectrum into sinusoids and a residual (or noise), this residual results by removing (i.e., by subtracting) sinusoids directly from a complex discrete frequency representation of the audio signal, presuming that this representation is able to resolve all existing sinusoidal components.

3. Synthesis and bandwidth extension of sinusoids with sub-bin accuracy and using a reduced set of parameters (frequencies, magnitude, or phases) describing the original audio sinusoidal components.

4. Synthesis and bandwidth extension of noise with bin accuracy and using a reduced (low-band) spectral portion of the original audio residual.

5. Sum of both bandwidth extended components and inverse normalization in order to recover the spectral envelope model of the original spectrum.

The *ASR* model is highly flexible in terms controlling the spectral balance of the reconstructed high frequency components. For example, the spectral tilt affecting the incoherent components, and the spectral tilt controlling the sinusoidal components can be shaped and controlled in an independent way.

Further details on ASR may be found in [8] and at http://www.atc-labs.com/asr. Here we focus on aspects of *ASR* which are related to the coding of sinusoidal components. In the *ASR* model the parameters necessary for the synthesis of harmonic partials are suitably reduced. For example in many cases the phase information may be completely discarded, or in other cases it is transmitted only at the time of harmonic birth and used in conjunction with a synthesis technique that insures frame continuity from frame to frame.

The primary filterbank domain for ASR processing is the Odd-DFT (ODFT). At the encoder sinusoidal components are estimated from the ODFT spectrum and removed by direct synthesis of ODFT spectral bins using a model of the frequency response of sine window [18, 20] and the estimated frequency, magnitude, and phase parameters. It has been concluded that only a small number of frequency bins per sinusoidal component are needed to generate a good quality sinusoid and to effectively remove it from the ODFT spectrum. The sinusoidal components are further analyzed to detect the presence of one or more harmonic patterns (including harmonics with missing fundamentals) as well as isolated (non-harmonic) sinusoidal components. Parameters necessary for the synthesis of high frequency sinusoidal components are then analyzed and suitably reduced (e.g., by discarding the phase components). The reduced parameters are forwarded to the decoder.

In the decoder the high frequency sinusoidal components can be synthesized directly in the ODFT domain, avoiding the TDAC mechanism associated with MDCT. A sinusoidal continuation algorithm is used to generate sinusoidal trajectory using only the transmitted frequency and magnitude parameters. In most cases phase information is only needed at the time of harmonic birth. Furthermore, in most cases a reduced level of magnitude information in the form of a smooth spectral envelope is needed for the sinusoidal continuation algorithm.

The accuracy of sinusoidal synthesis using the ASR model is depicted in [8] using a synthetic FM modulated sinusoid and natural audio signals.

4.3. Adaptive Combination of *FSSM* and *ASR* Models

In the TeslaPro codec the *FSSM* and *ASR* bandwidth extension tools may be deployed independently or in combination with each other. The actual adaptation mechanism as well as relative strengths of the two models for different types of audio signals is currently under investigation. This adaptation, however, is performed at the encoder. The decoder is capable of flexibly combining the two models. In particular for each frame of audio the following possibilities for high frequency reconstruction are permitted:

- The high frequencies are synthesized using a *FSSM* + noise + isolated (non-harmonic) tone model. The isolated tones in this case are synthesized using the sinusoidal synthesis techniques used in the *ASR* decoder. Furthermore, at the encoder a sinusoidal and harmonic detection analysis similar to *ASR* encoder is performed and the resulting information is utilized in estimating the *FSSM* model parameter.

- As a second alternative higher frequencies are reconstructed using the full *ASR* model which consists of bandwidth extension of coherent components and bandwidth extension of incoherent components using transposition and synthetic noise addition

- As a third alternative, the *ASR* model is used for the synthesis of harmonics and tonal components, while, the *FSSM* model is utilized to generate the *residual* which may still include significant non noise-like components such as pitch structures found in vocals, etc. Remaining components which are neither tonal nor fit the *FSSM* model are then replaced by synthetic noise.

In audio frames where both *FSSM* model and *ASR* sinusoidal synthesis model is active, certain harmonization of the composite synthesized spectrum is necessary to ensure that components added by the two models do not adversely effect the perceived quality of the component added by the other model.

4.4. Time-Frequency Envelope Shaping Considerations for Synthesized High Frequency Components

The *TeslaPro* codec utilizes the following techniques for suitable shaping of the time-frequency envelope of the synthesized high frequency components.

- The spectral smooth envelope in frequency (MDCT) domain can be coded and transmitted directly using a differential spectral envelope coding techniques.

- The relative magnitude of *FSSM* and sinusoids can be controlled in one or more frequency bands.

- Since the default frequency resolution of the primary coding and bandwidth extension filter bank is quite high additional temporal shaping is performed using a secondary filter bank. This aspect is discussed in more detail below.

TeslaPro employs a higher time resolution “Utility Filter Bank” (*UFB*) for the desired temporal shaping of the reconstructed higher frequencies. The *UFB* is a complex, over-sampled modulated filter bank [7]. An over sampling ratio between 2 and 8 can be employed and temporal shaping with a 4-5 *msec* resolution is possible. Depending upon the complexity profile of the decoder and bit rate of operation the *UFB* may take one of the following 3 forms.

- Discrete Fourier Transform (DFT) with a higher time resolution (compared to MDCT): A DFT with 128-256 size *power complementary* window may be used in a sequence of overlapping blocks (with a 50% overlap between 2 consecutive windows). This represents an over-sampling ratio of 2.

- A complex modulated filter bank with an over-sampling ratio between 4 and 16 and sub-band filters of the form

$$h_i = h_0 \cdot e^{j\frac{2\pi}{N}\cdot i\cdot n} \qquad (7)$$

  where h_0 is a suitably optimized prototype filter.

- A complex non-uniform filter bank; e.g., one with two uniform sections and transition filters to link the 2 adjacent uniform sections as described in [7]. This filter bank is designed using the technique described in [22]. The sub-bands in the lower sections have ½ the bandwidth of the sub-bands at higher frequencies. The higher frequency resolution at lower frequencies is useful, for example, in parametric stereo coding (see next section).

Multi-band temporal envelope information to perform the temporal shaping is computed by analyzing the output of the *UFB* and transmitting a suitable representation as side information. The overhead for this information can be reduced by utilizing the temporal shape that may already exist and by grouping the information in adjacent time and frequency bands

5. PARAMETRIC CODING OF STEREO INFORMATION

TeslaPro incorporates techniques for efficient coding of stereo signals using parametric representation. Parametric stereo coding can be used for all or part of the spectrum. In general parametric coding techniques attempt to code inter-aural (inter-channel) localization cues using as few bits as possible. Several different techniques have been proposed in recent years [23-26]. In general an attempt is made to encode and transmit the following three sets of localization cues [23].

- Inter-aural intensity difference (*IID*): the intensity differences are the primary localization mechanism at higher frequencies (dominant above about 5 kHz) and continue to play an important role at lower frequencies, particularly in terms of consistency with other cues.

- Inter-aural time differences (*ITD*): The inter-channel time differences in the form of inter-channel phase difference cues play an important role in the localization of sound below 1500 Hz. The *ITD* in the form of inter-channel "*envelope delay*" continues to be significant till about 5 kHz (or higher in the form of *precedence effect*).

- Inter-aural coherence cues (*ICC*): It has been observed that even if there is no intensity or time difference between channels, stereo image may still sound diffuse depending upon the coherence between the two channels. Coherence between the channels is known to have an impact on stereo image width and stability.

In *TeslaPro* the parametric stereo information is coded in the domain of the *UFB* output. Depending upon the frequency range over which parametric coding is being used either a uniform or non-uniform configuration of *UFB* is employed (the latter non-uniform case is used if parametric coding to be used at frequencies below 1500 Hz). The parameters used in stereo coding consist of following two set of information extracted from the *UFB* output for the two channels

- *Stereo (L/R) multi-band time-frequency envelopes*. These envelopes are computed over an adaptive time-frequency grid with a sub-critical band partition. Grouping over uniform regions in the time frequency plane and joint encoding is used to improve coding efficiency of the two channel envelope. The stereo envelope captures the *IID* cues as well as the some of the most significant inter-envelope delay cues.

- *Inter-channel time differences*. The time differences for lower frequencies are coded as phase differences between the two channel values. In the mid and high frequency range a determination is made if additional "group delay" differences are needed (keeping in view the time-frequency envelope information that has already been coded).

If it is necessary to reduce the coherence between the two channels in one or more frequency bands a randomized incoherent phase component in the corresponding *UFB* bands is introduced.

6. CODING RESULTS

The *TeslaPro* codec is capable of operating across a wide range of bit rates. Several audio samples at various bit rates are available at http://www.atc-labs.com/teslapro. In particular, samples at the following bit rates are available

- 48 kbps. At this bit rate full stereo dynamic range of the original music CD and a 16 kHz audio bandwidth is maintained. Bandwidth extension tools are used to reconstruct frequencies above 8 kHz. The audio quality is CD-like

- 40 kbps. At this bit rate 16 kHz audio bandwidth is maintained and a stereo dynamic range that is somewhat lower than the original CD is maintained. Bandwidth extension is use for frequencies starting at 6 kHz. Audio quality continues to be CD-like or near-CD.

- 32 kHz. At this bit rate bandwidth extension is used between 6 and 16 kHz and full parametric stereo coding (till the lower end of the spectrum) is employed. The audio quality may be characterized as "FM quality stereo".

- 24 kbps. At this bit rate 16 kHz audio bandwidth with stereo reconstruction (albeit limited stereo) is achieved.

- 22 kbps. Audio quality at this bit rate may be characterized as high quality full bandwidth (16 kHz) mono.

The accuracy of bandwidth extension using the TeslaPro codec at lower bit rates is illustrated in Figures 5 and 6 using a sample coded at 40 kbps. In Figure 5, the frequency spectrum of original and coded audio is shown. It is evident that all the spectral features (e.g., harmonics and tones etc.) are reconstructed in the decoded signal. In Figure 6, the temporal envelope of original and decoded signal in a randomly chosen critical band is shown. Once again there is a close match between the original and the decoded audio.

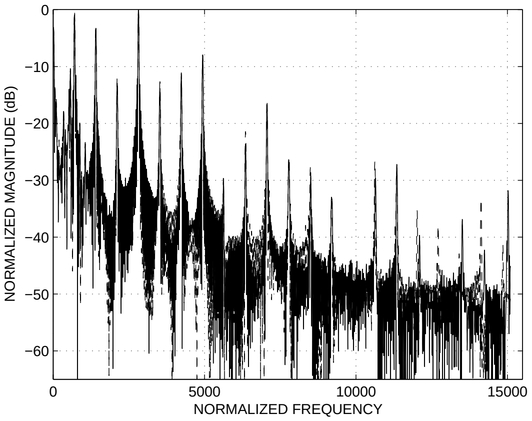

Figure 5: Decoded signal spectrum (solid line) and original spectrum (dashed line) for a frame of Harpsichord signal coded at 40 kbps.

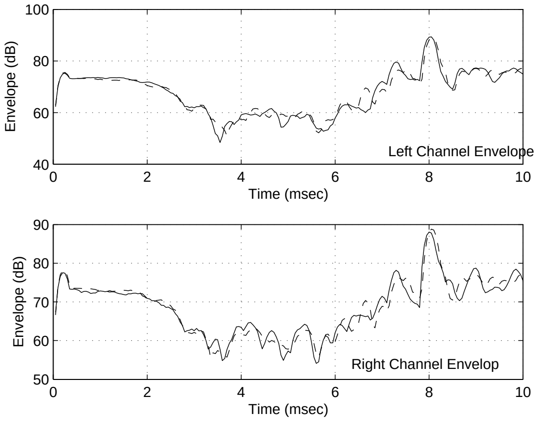

Figure 6: Decoded signal (40 kbps) Temporal Envelope (solid line) and Original Temporal Envelope (dashed line) for a critical band around 10 kHz. (Left channel envelope is shown on top and the Right channel envelope is at the bottom of the figure).

For comparison audio files encoded with *TeslaPro* at 40 and 48 kbps are also encode with MP3Pro at 64 kbps using the encoder/player available at http://www.mp3prozone.com.

From these samples preliminary conclusions regarding *TeslaPro* audio quality can be made. The audio at 48 kbps sounds less “flat” with more of the features of the original audio compared to other compression technologies utilizing bandwidth extension. Also at lower bit rates (22, 24 etc.), the voices sound natural and do not exhibit ringing type artifacts found in some of the other bandwidth extension schemes. Formal comparison of the TeslaPro codec with other coding schemes is currently in progress.

7. CONCLUSIONS

Some of the key components of the TeslaPro codec are described. This new coding schemes utilizes a combination of powerful bandwidth extension techniques. It employs a truly wide band Psychoacoustic Model and an efficient parametric stereo coding technique. The resulting coding scheme provides excellent audio quality across a wide range of bit rates.

8. ACKNOWLEDGEMENTS

. The authors would like to gratefully acknowledge the support of Sirius Satellite Radio in making this work a possibility. Parts of the new coding scheme were developed in collaboration with Sirius Satellite Radio. The support and feedback from Mr. Mark Kalman, and, Mr. Ray Lowe is much appreciated. The contributions of Ms. Bhadresha Dedhia are also thankfully acknowledged.

9. REFERENCES

[1] J. D. Johnston, D. Sinha, S. Dorward, and S. R Quackenbush, “AT&T Perceptual Audio Coding (PAC),” *in AES Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. 1996, pp. 73-82.

[2] Kyoya Tsutui, Hiroshi Suzuki, Mito Sonohara Osamu Shimyoshi, Kenzo Akagiri, and Robert M.Heddle, “ATRAC: Adaptive Transform Acoustic Coding for MiniDisc,” *93rd Convention of the Audio Engineering Society*, October 1992, Preprint n. 3456.

[3] K. Bradenburg, G. Stoll, et al. “The ISO- MPEG-Audio Codec: A Generic-Standard for Coding of High Quality Digital Audio,” in *92nd AES Convention*, 1992, Preprint no. 3336.

[4] Marina Bosi et al., "ISO/IEC MPEG-2 Advanced Audio Coding," *101st Convention of the Audio Engineering Society*, November 1996, Preprint n. 4382.

[5] Mark Davis, "The AC-3 Multichannel Coder," *95th Convention of the Audio Engineering Society*, October 1993, Preprint n. 3774.

[6] J. P. Princen, A. W. Johnson, and A. B. Bradley, "Subband/Transform Coding Using Filter Bank Designs Based on Time Domain Alias Cancellation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1987, pp. 2161-2164.

[7] Deepen Sinha, Anibal Ferreira, and, Deep Sen "A Fractal Self-Similarity Model for the Spectral Representation of Audio Signals," *118th Convention of the Audio Engineering Society*, May 2005, Paper 6467.

[8] Anibal J. S. Ferreira and Deepen Sinha, "Accurate Spectral Replacement," *118th Convention of the Audio Engineering Society*, May 2003, Paper 6383.

[9] M Dietz, L. Liljeryd, K. Kjorling, and O. Kunz, "Spectral Band Replication, a novel approach in audio coding," *112th Convention of the Audio Engineering Society*, May 2002, Paper 5553.

[10] Deepen Sinha and Anibal Ferreira "A New Class of Smooth Power Complementary Windows and their Application to Audio Signal Processing," *to be presented at the 119th Convention of the Audio Engineering Society*, October 2005.

[11] Joseph L. Hall, *"Auditory Psychophysics for Coding Applications,"* Section IX, Chapter 39, The Digital Signal Processing Handbook, CRC Press, Editors: Vijay K. Madisetti and Douglas B. Williams, 1998.

[12] B.C.J. Moore, *An Introduction to the Psychology of Hearing, 5th Ed.*, Academic Press, San Diego (2003).

[13] Eberhard Zwicker, and Hugo Fastl, *Psychoacoustics: Facts and Models*, Springer Series in Information Sciences (Paperback), Second updated edition.

[14] Anibal J. S. Ferreira, *Spectral Coding and Post-Processing of High Quality Audio*, Ph.D. thesis, Faculdade de Engenharia da Universidade do Porto-Portugal, 1998, http://telecom.inescn.pt/doc/phd_en.html.

[15] D. Sinha, *Low bit rate transparent audio compression using adapted wavelets.* Ph.D. thesis, University of Minnesota, 1993.

[16] Hall JW, Grose JH, Mendoza L (1995) Across-channel processes in masking. In: Hearing (Moore BCJ, ed), pp 243–266. San Diego:Academic.

[17] Jesko L. Verhey, Torsten Dau, and Birger Kollmeier "Within-channel cues in comodulation masking release (CMR): Experiments and model predictions using a modulation filter bank model" Journal of the Acoustical Society of America, 106(5), p. 2733-2745.

[18] Anibal J. S. Ferreira and Deepen Sinha, "Accurate and Robust Frequency Estimation in ODFT Domain," *in 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 16-19, 2005, submission accepted.

[19] Anibal J. S. Ferreira, "Combined Spectral Envelope Normalization and Subtraction of Sinusoidal Components in the ODFT and MDCT Frequency Domains," in 2001 *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 21-24 2001, pp. 51-54.

[20] Anibal J. S. Ferreira, "Accurate Estimation in the ODFT Domain of the Frequency, Phase and Magnitude of Stationary Sinusoids," in 2001 *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 21-24 2001, pp. 47-50.

[21] Anibal J. S. Ferreira, "Perceptual Coding Using Sinusoidal Modeling in the MDCT Domain," *112th Convention of the Audio Engineering Society*, May 2002, Paper 5569.

[22] Z. Cvetkovic and J. D. Johnston, "Nonuniform Oversampled Filter Banks for Audio Signal Processing," *IEEE Transactions on Speech and Audio Processing*, Vol. 11, No. 5, September 2003.

[23] Jens Blauert, *Spatial Hearing, Revised Ed.* MIT Press (1996). ISBN 0-262-02413-6.

[24] D. Sinha, "Technique for parametric coding of a signal containing information", *U.S. Patent No.* US6539357, Filed 1999 (Published March, 2003).

[25] F. Baumgarte and C. Faller, ``Binaural Cue Coding Part I: Psychoacoustic fundamentals and design principles," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, no. 6, Nov. 2003.

[26] C. Faller and F. Baumgarte, ``Binaural Cue Coding - Part II: Schemes and applications, "*IEEE Trans. on Speech and Audio Proc.*, vol. 11, no. 6, Nov. 2003.

[27] Nikil Jayant, James Johnston, and Robert Safranek, "Signal Compression Based on Models of Human Perception," *Proceedings of the IEEE*, vol. 81, no. 10, pp. 1385-1422, October 1993.

[28] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Prentice-Hall, 1975.

[29] ITU-R Recommendation BS.1534, "Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems," June 2001.