Audio Engineering Society

Convention Paper

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

A New Low-Delay Codec for Two-Way High-Quality Audio Communication

Aníbal J. S. Ferreira[1,2], and Deepen Sinha[2]

[1] *University of Porto, Portugal*

[2] *ATC Labs, USA*

Correspondence should be addressed to A. Ferreira (ajf@atc-labs.com, a.j.ferreira@ieee.org)

ABSTRACT

High-quality audio bit-rate reduction systems are widely used in many application areas involving audio broadcast, streaming and download services. With the advent of 3G mobile and wireless communication networks, there is a clear opportunity for new multimedia services, notably those relying on two-way high-quality audio communication. In this paper we describe a new source/perceptual audio coder that features low-delay, intrinsic error robustness and high subjective audio quality at competitive compression ratios. The structure of the audio coder is described and an emphasis is given on its innovative approaches to semantic signal segmentation and decomposition, independent coding of sinusoidal and noise components, and bandwidth extension using Accurate Spectral Replacement. A few test results are presented that illustrate the operation and performance of the new coder. Audio demos are available at http://www.atc-labs.com/acc/ .

1. INTRODUCTION

Several proprietary audio coding schemes such as ATRAC (Sony) [1], PAC (Bell Labs, Lucent) [2] and WMA (Microsoft), as well as standardized audio coding schemes such as Dolby AC-3 [3], MPEG-1 Layer 3 (MP3) [4], MPEG-2 AAC [5] and AMR-WB+, are successfully used in many application areas. Most of these coding schemes exhibit high compression efficiency at the cost of high system complexity, high end-to-end coding delay, and low intrinsic error robustness due to the extensive use of inter-frame coding strategies. While these last two aspects (high end-to-end coding delay, and low intrinsic error robustness) are not critical in the case of audio broadcast, streaming, messaging and download services, they represent a real barrier in the context of real-time, two-way high-quality audio communication. This particular application context is however

gaining a significant importance due to the opportunities created by 3G mobile and wireless communication networks that pave the way for new ways of multimedia human interaction and communication. In fact, while voice communication is the basic functionality that has been inherited from 1G communication networks, high-quality audio communication has not yet been deployed as a new multimedia service for example in the same way video has, in the context of 3G communication. Still, several analysts predict the next killer application in the mobile and handset industry is related to music and new audio experiences, not only because this is in line with a consumer expectation of improved communication quality and innovative functionalities (which is confirmed for example by the popularity of ring tones and phone cameras), but also because operators have strongly invested in 3G and expect the corresponding return by offering consumers new services.

We describe in this paper a new source/perceptual audio coder that operates in the frequency domain and that targets real-time or two-way audio communication. The new coder, Audio Communication Coder (ACC), has been designed to optimize the coding of high-quality monophonic audio since low bit-rate parametric methods can be used to efficiently code and recreate stereophonic and multichannel information. In particular, the new coder features:

- low-delay coding (< 50 ms) by minimizing the size of the transform, and look-ahead and bitstream buffers [6],
- intrinsic error robustness by not using interframe coding, which not only makes error concealment more effective when the communication channel is very hostile (*e.g.*, wireless) but also permits traditional functionalities such as fast-rewind and fast-forward play modes that are not easily implemented in the case for example of MP3 or AAC codecs,
- a new mechanism of bandwidth extension that looks into the fine structure of the signal so as to perform Accurate Spectral Replacement (ASR) [7] of not coded (or lost) spectral regions.

This paper is structured as follows. Section 2 describes the ACC encoder structure and approach to efficient coding. Section 3 describes the structure of the ACC decoder and approach to bandwidth extension. Section 4 describes the operation of the ACC codec with both synthetic and natural audio signals, and illustrates in detail the contribution of the ASR bandwidth extension technique. Section 5 reviews the main results presented in the paper and gives a perspective on possible applications.

2. ACC ENCODER STRUCTURE

The structure of the encoder is represented in Fig. 1. It has been designed to take advantage of different opportunities for signal compression, namely those allowed by:

- source coding tools,
- perceptual coding tools,
- bandwidth extension tools.

Source coding tools provide compression gains by reducing the redundancy of the signal. This is achieved namely by using:

- a time-frequency transformation of the audio signal and taking advantage of the associated decorrelation capability (*i.e.*, coding gain over PCM [8]),
- entropy coding techniques,
- parametric coding techniques,
- optimum quantization techniques.

Perceptual coding tools provide compression gains by reducing the irrelevancy of the signal. This is achieved namely by using:

- psychoacoustic modeling and estimation of the maximum quantization noise (in time and frequency) at the threshold of masking (and therefore at the threshold of transparent coding),
- appropriate quantization techniques so as to shape the quantization noise at, or preferably below, the estimated threshold of masking.

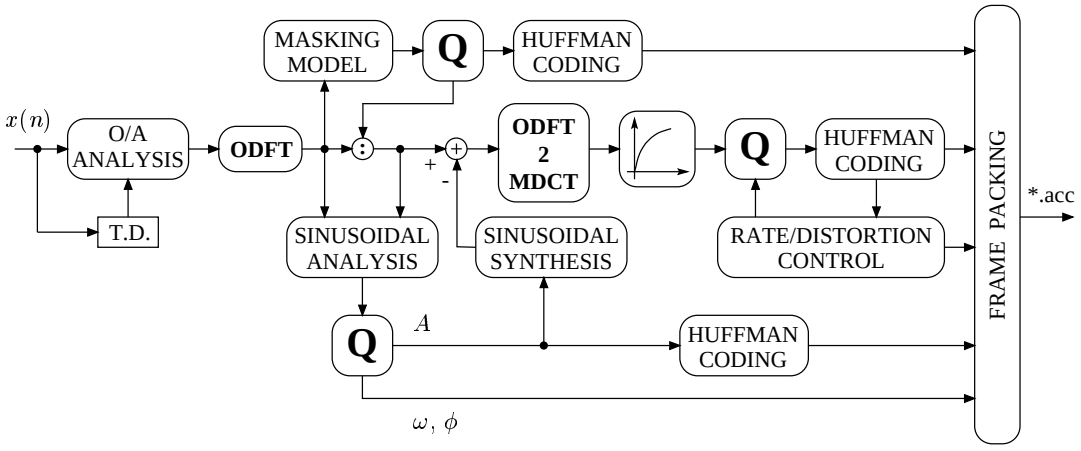

Fig. 1: Block diagram of the ACC encoder. In this diagram T.D. denotes transient detector and Q denotes quantization.

Bandwidth extension tools allow for the replacement of a spectral region of the original signal by a synthetic but perceptually similar sound signal, at the cost of a very parsimonious description (and therefore bit-rate efficient) of its perceptually important features such as spectral envelope and harmonicity.

Typically, these different aspects are addressed by different building blocks of the encoder. It also happens however that the same building block may serve different purposes. For example, in Fig. 1 the sinusoidal analysis block is used either for parametric coding or for bandwidth extension using ASR [7];

The encoder classifies first each segment of the audio signal according to its local stationary or non-stationary profile. In the latter case a window switching mechanism is activated [9, 10]. In the former case the signal is further classified as stationary noise, stationary noise with individual sinusoids, or stationary noise with sinusoids harmonically related [11]. If sinusoids are identified, they are parametrically represented and are subtracted from the audio signal, leaving a residual whose coding is more bit-rate efficient.

Irrespective of the specific classification of the signal, its threshold of masking is estimated, is coded using a short-pass liftered version of its real cepstrum [12], and is used to normalize the spectral representation of the signal, just before spectral subtraction of sinusoids. The spectral residual that is obtained after normalization and spectral subtraction, is perceptually 'white' or equally loud in frequency. In order to minimize the average power of this noise, PDF-optimized quantization is implemented through companding and uniform quantization [8, 13].

Statistical redundancies are reduced by Huffman coding of quantized coefficients, sinusoidal magnitudes and cepstral coefficients.

Distortion/rate control is implemented by means of an iterative procedure involving a single quality parameter that is transmitted to the decoder as side information [6]. The encoder operates at constant quality coding (and variable rate coding) if that parameter is constant in time, or operates at constant bit rate coding (and variable coding quality) if that parameter is made to vary in time in a suitable way.

The ACC encoder structure represented in Fig. 1 is also valid when the window switching is activated, (*i.e.*, when the encoder operates in an improved time resolution coding mode), with exception for the sinusoidal analysis that is not performed in this case due to the reduced frequency resolution.

3. ACC DECODER STRUCTURE

The ACC decoder structure is illustrated in Fig. 2. This structure consists of the same sequence of

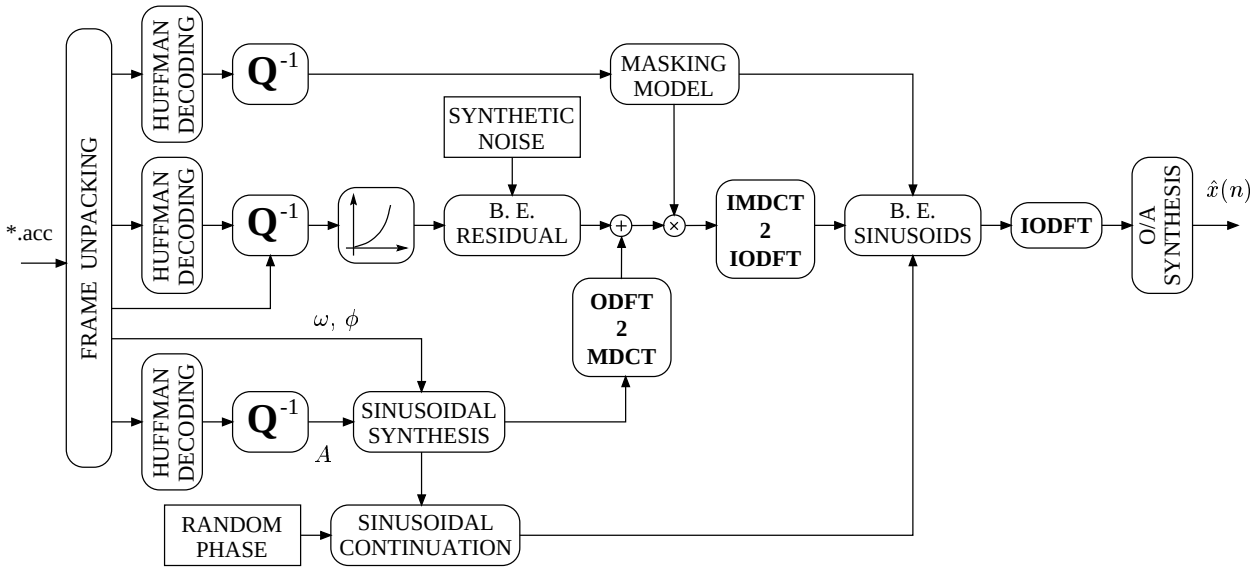

Fig. 2: Block diagram of the ACC decoder. In this diagram B.E. denotes bandwidth extension.

processing steps taken at the encoder but in a reverse order. Specifically, after Huffman decoding and inverse quantization, the spectral residual is expanded and possibly bandwidth extended. Sinusoids falling within the bandwidth of the coded residual are added back to the residual in the MDCT domain. The resulting signal is then denormalized by a model of the threshold of masking and is transformed to the ODFT domain. This real-to-complex frequency transformation is compatible with, and influenced by, the Time Domain Aliasing Mechanism (TDAC) [14] of the MDCT. Bandwidth extension of sinusoids is implemented in the ODFT domain taking advantage of the fact that it is oversampled by a factor of two relative to the MDCT and therefore the TDAC mechanism does not apply.

It should be noted that the bandwidth extension is implemented separately for noise and sinusoids according to the ASR technique [7], which means additional flexibilities exist for example in controlling the spectral tilt of the noise floor independently from the spectral tilt of sinusoidal components, whether harmonically related or not [1]. The PCM audio signal is then reconstructed after inverse ODFT transformation and overlap-add synthesis.

[1] A detailed description of the ASR concept and operation is available on-line at `http://www.atc-labs.com/asr/`

4. OPERATION AND RESULTS

In order to verify the correct operation of the ACC encoder and decoder algorithms, several tests have been performed, three of which are described in this section. Two of them involve synthetic signals and the remaining one involves a natural audio signal.

4.1. Bandwidth Extension of a Sinusoid with *Vibrato*

We have synthesized *vibrato* by modulating in frequency a single sinusoid according to the following Matlab code.

```
samples = 512*100; t=[0:samples-1]/1024;
temp=5000*sin(2*pi*25.1*t +
2.56*sin(2*pi*t/10)); fidw =
fopen('testfile.pcm','w'); fwrite(fidw, temp,
'short'); fclose(fidw);
```

Assuming the sampling frequency is 44100 Hz, and using a 1024 point ODFT/MDCT transformation, it results that the frequency resolution of the filter bank (or bin width in Hz) is about 43 Hz. An analysis of the above Matlab code reveals that the total

maximum frequency deviation is 0.512 bins or about 22 Hz, and that the central frequency of the FM signal is 25.1 bins (or about 1081 Hz).

The sinusoidal analysis at the encoder uses a recently developed technique of instantaneous frequency estimation that is highly accurate (the estimation error is less than 0.1% of the bin width) and is robust to noise [15]. The output of this technique to the FM modulated sinusoid is depicted in Fig. 3. In this

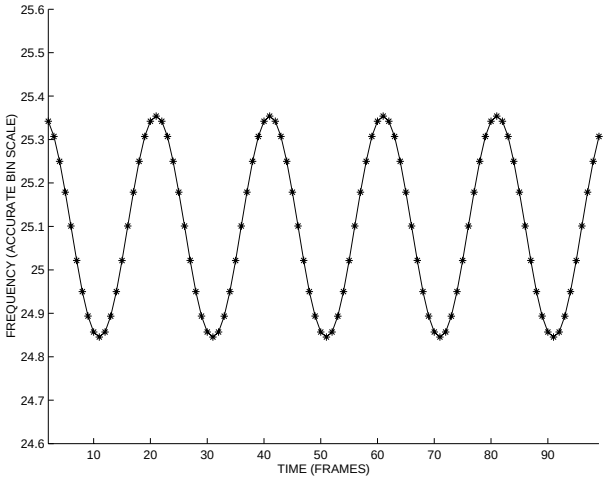

Fig. 3: Instantaneous frequency estimation of an FM modulated sinusoid.

figure stars represent instantaneous frequency estimates and solid lines connect stars for which phase coherence must be enforced. The results in this figure are consistent with the exact frequency deviation as concluded above.

Bandwidth extension has been tested by magnifying magnifying by a factor of 10 (ten) the main characteristics of the above FM signal, *i.e.*, by shifting the center frequency to 251 bins (*i.e.*, to about 10810 Hz), and by magnifying the total maximum frequency deviation to 5.12 bins (*i.e.*, to about 220 Hz). It should be noted that this differs significantly from frequency transposition[2] in which case the center frequency of the FM signal is changed but its frequency deviation remains unchanged. Fig. 4 displays the spectrogram of the bandwidth extended FM sinusoid. This representation is consistent with the expected and correct signal transformation. This test signal is available

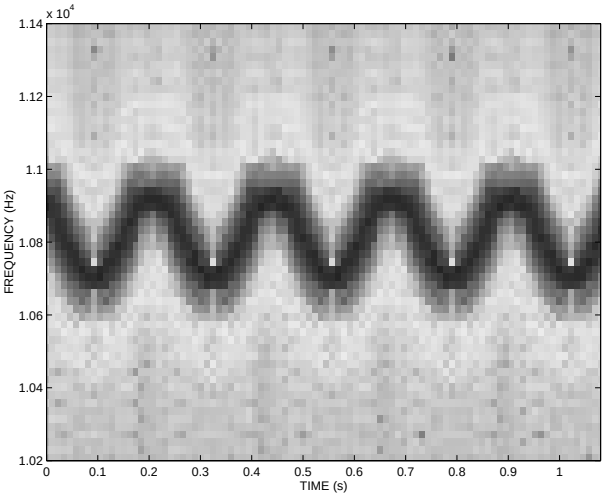

Fig. 4: Spectrogram of an FM modulated sinusoid after bandwidth extension.

at `http://www.atc-labs.com/acc/` .

4.2. Bandwidth Extension of a Harmonic Complex with *Vibrato*

Most audio signals exhibit a harmonic structure [10] and frequently this structure is modulated in frequency producing a *vibrato* effect. This is quite common in singing and musical sounds. It is therefore important to assess if the bandwidth extension of these important types of sound is correctly implemented till 20 kHz given that the frequency swing is larger for partials at higher center frequencies. We have therefore created a synthetic sound with a harmonic structure by adding three partials to the fundamental corresponding to the FM signal of the previous example. The corresponding spectrogram is represented in Fig. 5. The encoder algorithm implements accurate sinusoidal analysis whose result is reflected in Fig. 6. This figure makes clear that the frequency deviation is higher for higher order partials. The decoder algorithm has been configured so as to bandwidth extend the harmonic complex by synthesizing all partials from $11f_0$ till $18f_0$, where f_0 is the fundamental frequency. Fig. 7 represents the spectrogram of the new partials. It can

[2]Which is commonly used by other frequency extension techniques.

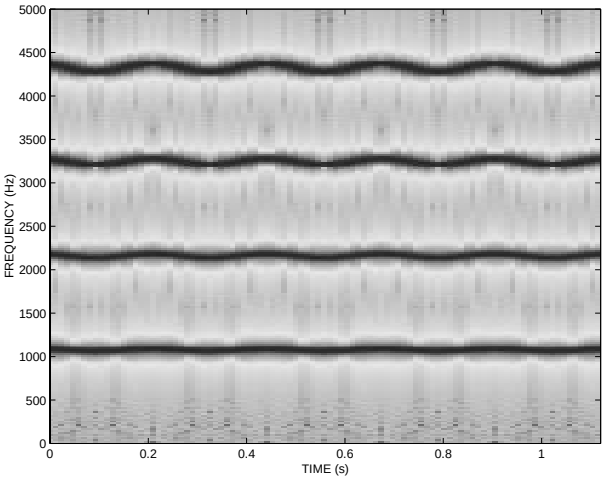

Fig. 5: Spectrogram of an FM modulated harmonic complex with four partials.

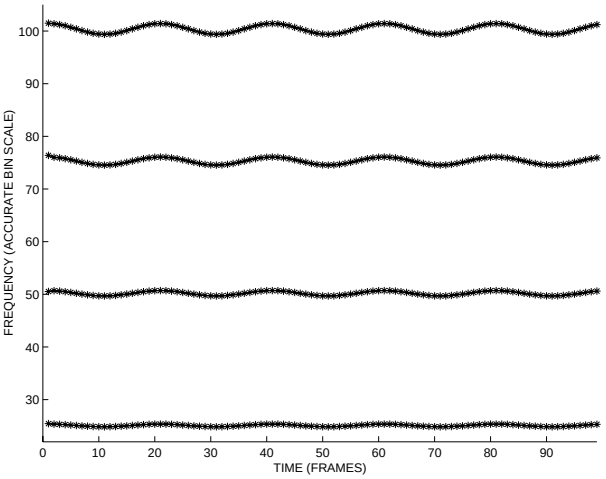

Fig. 6: Accurate estimation of the instantaneous frequencies of four partials pertaining to a harmonic complex.

be seen that even the highest order partial whose center frequency is near 20 kHz, is correctly synthesized despite the fact that its total maximum frequency deviation (or swing) is the largest: 9.2 bins or about 397 Hz. Both audio files are available at `http://www.atc-labs.com/acc/` .

4.3. Coding and Bandwidth Extension of a Natural Music Signal: *trompet*

A short excerpt of a single pitched sound (*i.e.*, a single musical note) produced by a trompet, has been used for coding and bandwidth extension. The encoder (a non-optimized version) has been configured for constant bit-rate coding at 24 kbit/s. This bit rate allows the explicit coding of the lowest 5 kHz frequency range of the audio signal. This is clear in Fig. 8 where the short-time power spectral density (PSD) of the original signal is represented by a dotted line, and the PSD of the ACC decoded signal (without bandwidth extension) is represented by a solid line. This figure also makes clear that in the coding only the first five (5) partials or the original signal are preserved. Fig. 9 shows the PSD of the original signal (dotted line) and the PSD of the ACC decoded signal with bandwidth extension of sinusoids only (solid line). This figure illustrates the realization of the conceptual approach of the ASR technique described in [7], and highlights the fact that frequency precise bandwidth extension of sinusoids is possible given the high sub-bin accuracy of the ASR analysis/synthesis technique. As expected, this figure also shows that the magnitudes of the bandwidth extended sinusoids are not exact to the original since only a merely indicative and scarce information is used to drive their reconstruction. Fig. 10 depicts the decoded and bandwidth extended signal including the sinusoidal and noise components. These two components are synthesized independently according to the ASR technique, which means their spectral tilt can be independently controlled to better replicate the spectral profile of the original signal. For example, in Fig. 10, the noise component that extends from about 5 kHz till about 20 kHz is completely synthetic[3] and for that reason it has been deliberately placed a few dBs below the noise floor of the original signal.

For comparison, the exact same original musical excerpt has been coded at 64 kbit/s (stereo, by duplicating the monophonic signal) with the MP3Pro encoder/player available at `http://www.mp3prozone.com` . The short-time PSD of the resulting decoded signal is depicted in Fig. 11 together with the short-time PSD of the

[3] This has been a particular coding option; in fact, other options for the bandwidth extension of the residual are available.

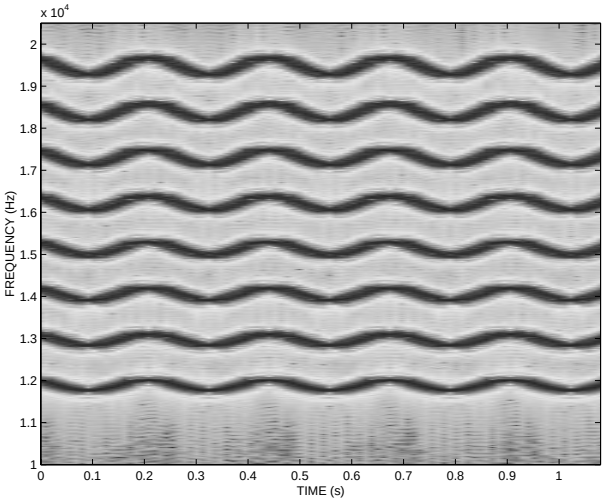

Fig. 7: Spectrogram of the bandwidth extended region of a harmonic complex with *vibrato*.

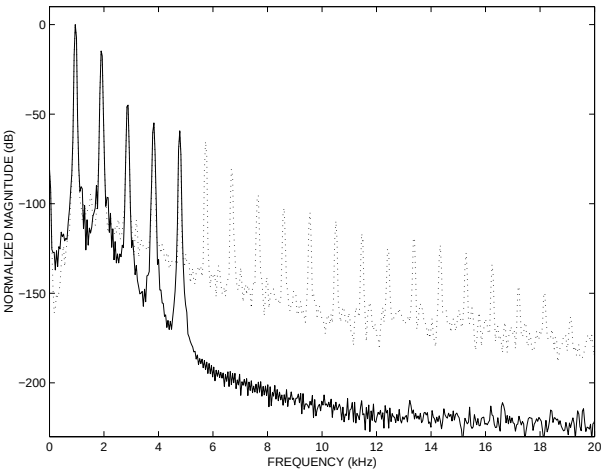

Fig. 8: Short-time PSD of the original *trompet* audio excerpt (dotted line) and of the resulting ACC decoded audio at 24 kbit/s (solid line) without bandwidth extension.

original signal. This figure reveals that:

- the audio signal is coded with good quality between DC and about 4.5 kHz (*i.e.*, probably this is where most of the available bits are spent)
- the spectral region of the audio signal between about 4.5 kHz and about 9 kHz is coded with less quality than the previous spectral region given that the noise floor of the input signal is not reproduced at the output,
- the spectral region of the audio signal between about 9 kHz and 16 kHz has been bandwidth extended with evident mismatches in terms of frequency location and magnitude of the higher order partials (of the harmonic complex), relative to the original signal.

All audio files addressed in this subsection are available at `http://www.atc-labs.com/acc/` .

5. CONCLUSION

We have described in this paper the structure of the encoder and decoder of a new low-delay audio codec, Audio Communication Coder or ACC, that has been designed to target real-time two-way high-quality audio communication. The operation of the new codec has been illustrated with both synthetic and natural audio signals. The distinctive features of this codec include low end-to-end communication delay (< 50 ms), moderate and approximate symmetric encoder-decoder complexity, intrinsic error robustness since no inter-frame coding is used, accurate bandwidth extension using the Accurate Spectral Replacement technique, and detailed signal segmentation and classification allowing semantic inference, which paves the way for future applications related to semantic access, filtering and retrieval of audio material on the coded (*i.e.*, bit stream) domain. Immediate application scenarios include real-time 3G mobile and wireless communication of high-quality audio.

6. REFERENCES

[1] Kyoya Tsutsui, Hiroshi Suzuki, Mito Sonohara Osamu Shimoyoshi, 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.

[2] 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.

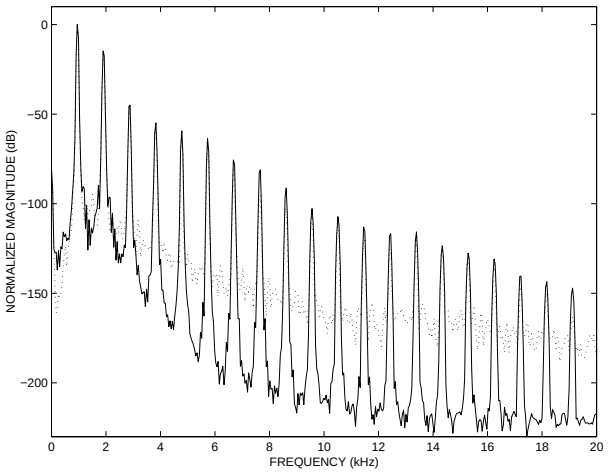

Fig. 9: Short-time PSD of the original audio excerpt (dotted line) and of the resulting ACC decoded audio at 24 kbit/s (solid line) with bandwidth extension of the sinusoidal part only.

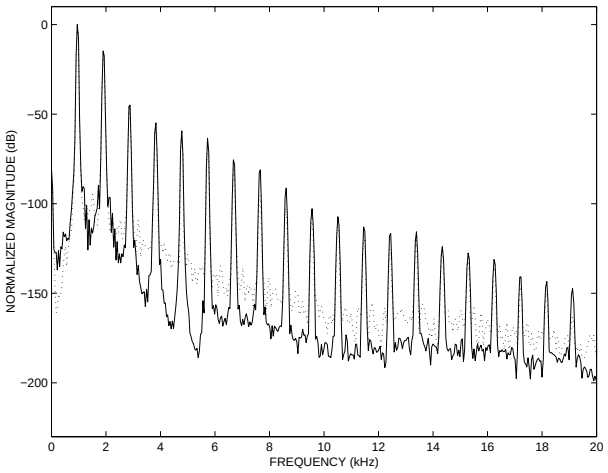

Fig. 10: Short-time PSD of the original audio excerpt (dotted line) and of the resulting ACC decoded audio at 24 kbit/s (solid line) with bandwidth extension of the sinusoidal and noise components.

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

[4] 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.

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

[6] Aníbal J. S. Ferreira, "Efficient Intraframe Coding of Monophonic Audio," *116th Convention of the Audio Engineering Society*, May 2004, Paper 6166.

[7] Aníbal J. S. Ferreira and Deepen Sinha, "Accurate Spectral Replacement," *118th Convention of the Audio Engineering Society*, May 2005, Paper 6383.

[8] N. S. Jayant and Peter Noll, *Digital Coding of Waveforms*, Prentice-Hall, 1984.

[9] B. Edler, "Coding of Audio Signals with Overlapping Block Transform and Adaptive Window Function. (in German)," *Frequenz*, vol. 43, pp. 252–256, September 1990.

[10] Aníbal 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.

[11] Aníbal J. S. Ferreira and André C. Rocha, "Combined Source and Perceptual Audio Coding," *115th Convention of the Audio Engineering Society*, October 2003, Paper 5982.

[12] Aníbal 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.

[13] Aníbal J. S. Ferreira, "Optimum Quantization of Flattned MDCT Coefficients," *115th Convention of the Audio Engineering Society*, October 2003, Paper 5988.

[14] J. P. Princen, A. W. Johnson, and A. B. Bradley, "Subband/Transform Coding Using Filter Bank Designs Based on Time Domain

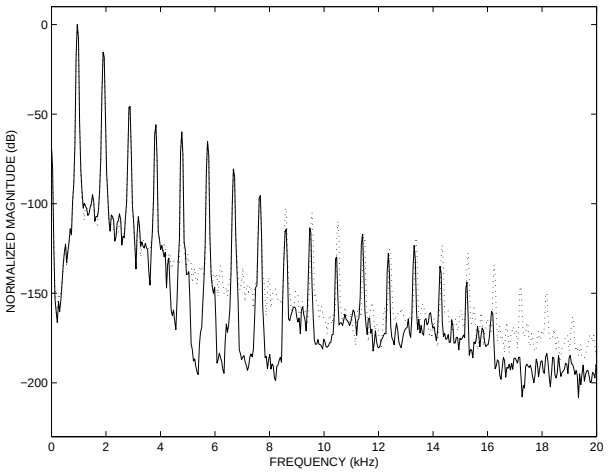

Fig. 11: Short-time PSD of the original audio excerpt (dotted line) and of the resulting MP3+SBR (stereo) encoded audio at 64 kbit/s (solid line).

Aliasing Cancellation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, April 1987, pp. 2161–2164.

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