ENCODING THE SINUSOIDAL MODEL OF AN AUDIO SIGNAL USING COMPRESSED SENSING

Anthony Griffin, Toni Hirvonen, Athanasios Mouchtaris and Panagiotis Tsakalides

Institute of Computer Science, Foundation for Research and Technology - Hellas (FORTH-ICS)
and Department of Computer Science, University of Crete, Heraklion, Crete, Greece
{agriffin, tmhirvo2, mouchtar, tsakalid}@ics.forth.gr

ABSTRACT

In this paper, the compressed sensing (CS) methodology is applied to the harmonic part of sinusoidally-modeled audio signals. As this part of the model is sparse by definition in the frequency domain, we investigate how CS can be used to encode this signal at low bitrates, instead of encoding the sinusoidal parameters (amplitude, frequency, phase) as current state-of-the-art methods do. We extend our previous work by considering an improved system model, by comparing our model to other schemes, and exploring the effect of incorrectly reconstructed frames. We show that encouraging results can be obtained by our approach, although inferior at this point compared to state-of-the-art. Good performance is obtained using 24 bits per sinusoid as indicated by our listening tests.

Index Terms— Audio coding, compressed sensing, sinusoidal model, signal reconstruction, signal sampling

1. INTRODUCTION

Audio compression is usually performed by either applying compression algorithms to the actual samples of a digital audio signal, or by initially using a signal model and then encoding the model parameters as a second step. In this paper, we extend our previous work [1] which introduced a novel method to encode the parameters of the sinusoidal model [2].

The sinusoidal model represents an audio signal using a small number of time-varying sinusoids. The remainder error signal—often termed the residual signal—can also be modeled to further improve the resulting subjective quality of the sinusoidal model [3]. The sinusoidal model allows for a compact representation of the original signal and for efficient encoding and quantization. State-of-the-art methods for encoding and compressing the parameters of the sinusoidal model (amplitudes, frequencies, phases) are based on directly encoding these parameters [4–7]. In this paper, we propose using the emerging compressed sensing (CS) [8, 9] methodology to encode and compress the sinusoidally-modeled audio signals.

Compressed sensing seeks to represent a signal using a number of linear, non-adaptive measurements. Usually the number of measurements is much lower than the number of samples needed if the signal is sampled at the Nyquist rate. CS requires that the signal is very *sparse* in some basis—in the sense that it is a linear combination of a small number of basis functions—in order to correctly reconstruct the original signal. Clearly, the sinusoidally-modeled part of an audio signal is a sparse signal, and it is thus natural to wonder how CS might be used to encode such a signal.

This work was funded in part by the Marie Curie TOK-DEV "ASPIRE"grant within the 6th European Community Framework Program, and in part by the FORTH-ICS internal RTD program "AmI: Ambient Intelligence Environments".

Our method encodes the time-domain signal instead of the sinusoidal model parameters as state-of-the-art methods propose [4–7]. The advantage is that the encoding operation is simplified into randomly sampling the time-domain sinusoidal signal, which is obtained after applying a psychoacoustic sinusoidal model to a monophonic audio signal. The random samples can be further encoded (here scalar quantization is suggested, but other methods could be used to improve performance). Additional advantages are that CS has inherent encryption and robustness to channel errors, and scales well to multi-channel cases. An issue that arises here is that as the encoding is performed in the time-domain—rather than the Fourier domain—the quantization error is not localized in frequency, and it is therefore more complicated to predict the audio quality of the reconstructed signal. At this point, it is noted that the paper deals only with encoding the sinusoidal part of the model.

This work extends our previous work [1] by using an improved system model with better frequency mapping, allowing us to compare our model with that of [5] through quality-rating listening tests. We also explore the effect of incorrectly reconstructed frames through preference listening tests. We show that encouraging results can be obtained by our approach, although inferior at this point compared to state-of-the-art.

2. SINUSOIDAL MODEL

The sinusoidal model was initially applied in the analysis/synthesis of speech [2]. A harmonic signal $s(t)$ is represented as the sum of a small number K of sinusoids with time-varying amplitudes and frequencies. This can be written as

$$s(t) = \sum_{k=1}^{K} \alpha_k(t)\ \cos(\beta_k(t)) \tag{1}$$

where $\alpha_k(t)$ and $\beta_k(t)$ are the instantaneous amplitude and phase, respectively. To estimate the parameters of the model, one needs to segment the signal into a number of short-time frames and compute a short-time frequency representation for each frame. Consequently, the prominent spectral peaks are identified using a peak detection algorithm (possibly enhanced by perceptual-based criteria). Interpolation methods can be used to increase the accuracy of the algorithm [3]. Each peak at the l-th frame is represented as a triad of the form $\{\alpha_{l,k}, f_{l,k}, \theta_{l,k}\}$ (amplitude, frequency, phase), corresponding to the K-th sinewave. A peak continuation algorithm is usually employed in order to assign each peak to a frequency trajectory using interpolation methods. A more accurate representation of audio signals is achieved when a model for the sinusoidal error signal is included as well. Practically, after the sinusoidal parameters are estimated, the noise component is computed by subtracting the harmonic component from the original signal. It is noted that in this paper we are only interested in encoding the sinusoidal part, and the error part is considered as available in our listening tests (as in [5]).

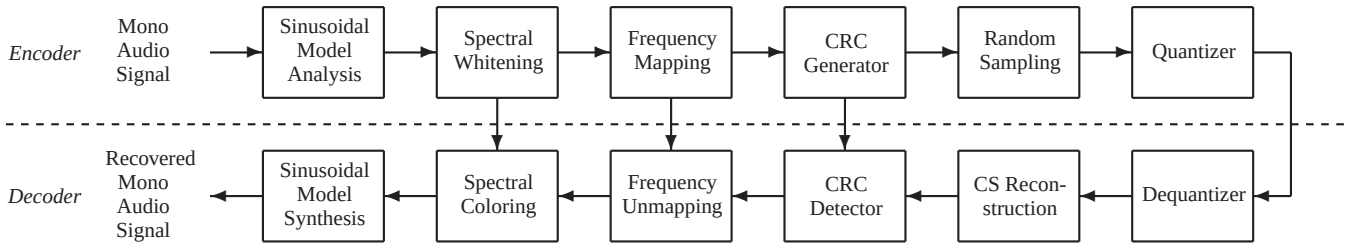

Fig. 1. A block diagram of the proposed system. In the encoder, the sinusoidal part of the monophonic audio signal is encoded by randomly sampling its time-domain representation, and then quantizing the random samples using scalar quantization.

3. COMPRESSED SENSING

In the compressed sensing methodology, a signal which is sparse in some basis can be represented using much fewer samples than the Nyquist rate would suggest. Given that a sinusoidally-modeled audio signal is clearly sparse in the frequency domain, our motivation has been to encode such signal using a small part of its actual samples, thus avoiding encoding a large degree of unnecessary information. In the following, we briefly review the CS methodology.

3.1. Measurements

Let $\boldsymbol{x}_l$ be the N samples of the harmonic component in the sinusoidal model in the l^{th} frame. It is clear that $\boldsymbol{x}_l$ is a K-sparse signal in the frequency domain. To facilitate our compressed sensing reconstruction, we require that the frequencies $f_{l,k}$ are selected from a discrete set, the most natural set being that formed by the frequencies used in the N-point fast Fourier transform (FFT). Thus $\boldsymbol{x}_l$ can be written as $\boldsymbol{x}_l = \boldsymbol{\Psi}\boldsymbol{X}_l$, where $\boldsymbol{\Psi}$ is an $N \times N$ inverse FFT matrix, and $\boldsymbol{X}_l$ is the FFT of $\boldsymbol{x}_l$. As $\boldsymbol{x}_l$ is a real signal, $\boldsymbol{X}_l$ will contain $2K$ non-zero *complex* entries representing the real and imaginary parts—or in an equivalent description, the amplitudes and phases—of the component sinusoids.

In the encoder, we take M non-adaptive linear measurements of $\boldsymbol{x}_l$, where $M \ll N$, resulting in the $M \times 1$ vector $\boldsymbol{y}_l$. This measurement process can be written as $\boldsymbol{y}_l = \boldsymbol{\Phi}_l\boldsymbol{x}_l = \boldsymbol{\Phi}_l\boldsymbol{\Psi}\boldsymbol{X}_l$ where $\boldsymbol{\Phi}_l$ is an $M \times N$ matrix representing the measurement process. For the CS reconstruction to work, $\boldsymbol{\Phi}_l$ and $\boldsymbol{\Psi}$ must be *incoherent*. In order to provide incoherence that is independent of the basis used for reconstruction, a matrix with elements chosen in some random manner is generally used. As our signal of interest is sparse in the frequency domain, we can simply take random samples in the time domain to satisfy the incoherence condition, see [10] for further discussion of random sampling (RS). Note that in this case, $\boldsymbol{\Phi}_l$ is formed by randomly-selected rows of the $N \times N$ identity matrix.

3.2. Reconstruction

Once $\boldsymbol{y}_l$ has been measured, it must be quantized and sent to a decoder, where it is reconstructed. Reconstruction of a compressed sensed signal involves trying to recover the sparse vector $\boldsymbol{X}_l$. It has been shown [8] [9] that

$$\hat{\boldsymbol{X}}_l \;=\; \arg\min \|\boldsymbol{X}_l\|_p \qquad \text{s.t.} \qquad \boldsymbol{y}_l = \boldsymbol{\Phi}_l\boldsymbol{\Psi}\boldsymbol{X}_l, \qquad (2)$$

with $p = 1$ will recover $\boldsymbol{X}_l$ with high probability if enough measurements are taken. The ℓ_p norm is defined as $\|\boldsymbol{a}\|_p \;=\; \left(\sum_i |a_i|^p\right)^{\frac{1}{p}}$. It has recently been shown in [11] that $p < 1$ outperforms the $p = 1$ case, and it is this method that we use for reconstruction in this paper.

A feature of CS reconstruction is that perfect reconstruction cannot be guaranteed, and thus only a *probability* of "perfect" reconstruction can be guaranteed, where "perfect" defines some acceptability criteria, typically a signal-to-distortion ratio. This probability is dependent on M, N, K and the quantization used.

Another important feature of the reconstruction is that when it fails, it can fail catastrophically for the whole frame. Not only will the amplitudes and phases of the sinusoids in the frame be wrong, but the sinusoids selected—or equivalently, their frequencies—will also be wrong. In the audio environment, this is significant as the ear is sensitive to such discontinuities. Thus it is essential to minimize the probability of frame reconstruction errors (FREs), and if possible eliminate them.

Let $\boldsymbol{F}_l$ be the *positive* FFT frequency indices in x_l, whose components $F_{l,k}$ are related to the frequencies in the x_l by $f_{l,k} = 2\pi F_{l,k}/N$. As $\boldsymbol{F}_l$ is known in the encoder, we can use a simple forward error correction to detect whether an FRE has occurred. We found that an 8-bit cyclic redundancy check (CRC) on $\boldsymbol{F}_l$ detected all the errors that occurred in our simulations.

Once we detect an FRE, we can either re-encode and retransmit the frame in error or use some interpolation between the correct frames before and after the errored frame to estimate it. This is discussed further in Section 4.3.

4. SYSTEM DESIGN

A block diagram of our proposed system is depicted in Fig. 1. The audio signal is first passed through a psychoacoustic sinusoidal modeling block to obtain the sinusoidal parameters $\{\boldsymbol{F}_l, \boldsymbol{\alpha}_l, \boldsymbol{\theta}_l\}$ for the current frame. These then go through what can be thought of as a "pre-conditioning" phase where the amplitudes are whitened—consisting of dividing the components of $\boldsymbol{\alpha}_l$ by a 3-bit quantized version of themselves—and the frequencies remapped, as discussed in Section 4.1. The modified sinusoidal parameters $\{\boldsymbol{F}'_l, \boldsymbol{\alpha'}_l, \boldsymbol{\theta}_l\}$ are then reconstructed into a time domain signal, from which M samples are randomly selected. These random samples are then quantized to 4 bits by a uniform scalar quantizer, and sent over the transmission channel along with the side information from the spectral whitening, frequency mapping and cyclic redundancy check (CRC) blocks.

In the decoder, the bit stream representing the random samples is returned to sample values in the dequantizer block, and passed to the compressed sensing reconstruction algorithm, which outputs an estimate of the modified sinusoidal parameters, $\{\hat{\boldsymbol{F}}'_l, \hat{\boldsymbol{\alpha}}'_l, \hat{\boldsymbol{\theta}}_l\}$. If the CRC detector determines that the block has been correctly reconstructed, the effects of the spectral whitening and frequency mapping are removed to obtain an estimate of the original sinusoid parameters, $\{\hat{\boldsymbol{F}}_l, \hat{\boldsymbol{\alpha}}_l, \hat{\boldsymbol{\theta}}_l\}$, which are passed to the sinusoidal model resynthesis block. If the block has not been correctly reconstructed, then the current frame is either retransmitted or interpolated, as discussed in Section 4.3.

In the tests employed in this paper, we investigated the performance of the proposed system using $K = 25$ sinusoid components per frame and an $N = 2048$-point FFT. All the audio signals were sampled at 22 kHz with a 20 ms window and 50% overlapping between frames. The data used for the results this section are around 5,000 frames of the audio data used in the listening tests of Section 5.

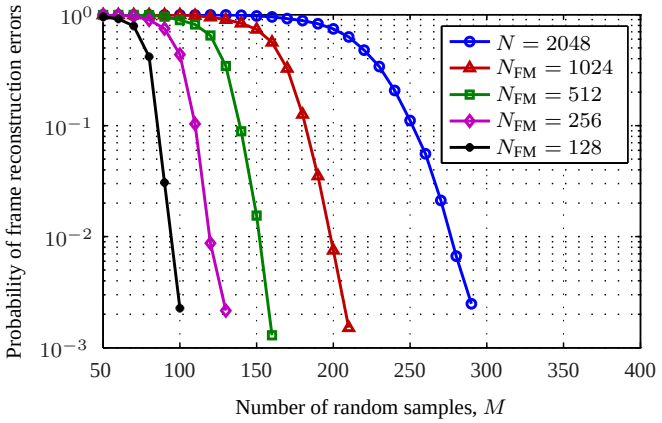

Fig. 2. Probability of frame reconstruction error *vs* the number of random samples per frame for various values of frequency mapping, with 25 sinusoids per frame, 4-bit quantization of the random samples, and 3-bit spectral whitening.

4.1. Frequency Mapping

The number of random samples, M, that must be encoded increases with N, the number of bins used in the FFT. In other words, there is a trade-off between the amount of encoded information and the frequency resolution of the sinusoidal model (which affects the resulting quality of the modeled audio signal). This effect can be partly alleviated by *frequency mapping*, which reduces the effective number of bins in the model by a factor of C_{FM}, which we term the *frequency mapping factor*. Thus the number of bins after frequency mapping is given by $N_{\text{FM}} = N/C_{\text{FM}}$.

We choose C_{FM} to be a power of two so that N_{FM} will always be a power of two, suitable for use in an FFT. We then create $\boldsymbol{F}'_l$, a mapped version of $\boldsymbol{F}_l$, whose components are calculated as

$$F'_{l,k} \;=\; \left\lfloor \frac{F_{l,k}}{C_{\text{FM}}} \right\rfloor, \tag{3}$$

where $\lfloor \cdot \rfloor$ denotes the floor function. We also need to calculate and send $\dot{\boldsymbol{F}}_l$ with components $\dot{F}_{l,k}$ given by

$$\dot{F}_{l,k} \;=\; F_{l,k} \mod C_{\text{FM}}. \tag{4}$$

We send $\dot{\boldsymbol{F}}_l$—which amounts to $K \log_2 C_{\text{FM}}$ bits—along with our M measurements, and once we have performed the reconstruction and obtained $\boldsymbol{F}'_l$, we can calculate the elements of $\boldsymbol{F}_l$ as

$$F_{l,k} \;=\; C_{\text{FM}} F'_{l,k} + \dot{F}_{l,k}. \tag{5}$$

It is important to note that not all frames can be mapped by the same value of C_{FM}, it is very dependent on each frame's particular distribution of $\boldsymbol{F}_l$. Essentially, each $F_{l,k}$ must map to a distinct $F'_{l,k}$. However, this can easily be checked in the encoder so that the value of C_{FM} chosen is the highest value for which (3) produces distinct values of $F'_{l,k}$, $k = 1, \ldots, K$. For the signals used in this paper, over 90% of the frames could be mapped $C_{\text{FM}} = 16$, giving $N_{\text{FM}} = 128$.

The clear decrease in the required M for a given probability of FRE for various values of N_{FM} is illustrated in Fig. 2.

4.2. Bitrates

Table 1 present the bitrates achievable for a probability of FRE of approximately 10^{-2} corresponding to the curves in Fig. 2. The overhead consists of the extra bits required for the CRC, the frequency mapping (FM) and the spectral whitening (SW). It is clear that the increasing overhead incurred from frequency mapping is more than accounted for by significant reductions in M, resulting in overall lower bitrates.

Table 1. Parameters that achieve a probability of FRE of approximately 10^{-2} with $N = 2048$ and $K = 25$

		raw	overhead			final	per
N_{FM}	M	bitrate	CRC	FM	SW	bitrate	sine
2048	275	1100	8	0	75	1183	47.3
1024	195	780	8	25	75	888	35.5
512	155	620	8	61	75	764	30.6
256	115	460	8	96	75	639	25.6
128	95	360	8	140	75	603	24.1

4.3. Operating Modes

To address the fact that we can only specify a probability of reconstruction, we propose two different operating modes to address the effect of frame reconstruction errors:

1) *Retransmission*: In the retransmission mode, any frame for which the CRC detects an FRE is re-encoded in the encoder using a different set of random samples and retransmitted. Obviously this requires more bandwidth, but if the probability of FREs is kept low enough this increase should be tolerable. For instance, we aim for $P_{\text{FRE}} \leq 10^{-2}$ in this work, which would incur an increase in bit-rate of approximately one percent.

2) *Interpolation*: For applications where retransmission is undesirable—or indeed impossible—the interpolation mode may be used. In this mode, lost frames are reconstructed using the same interpolation method as used in the regular synthesis of [2]. The assumption that a good frame is available both before and after the FRE is valid as we are considering low values of P_{FRE}. The effect of interpolation on the reconstructed signals is investigated in the listening tests of Section 5.

5. LISTENING TESTS

In this section, we examine the performance of our proposed system, with respect to the resulting audio quality. Two types of monophonic listening tests were performed, with eleven volunteers participating, using high-quality headphones in a quiet office room. The first test was based on the ITU-R BS.1116 [12] methodology, thus the coded signals were compared against the originally recorded signals using a 5-scale grading system (from 1-"very annoying" audio quality compared to the original, to 5-"not perceived" difference in quality). No anchor signals were used. The following seven signals were used (Signals 1-7): harpsichord, violin, trumpet, soprano, chorus, female speech, male speech. Signals 1-4 were obtained from the EBU SQAM disc, Signal 5 was provided by Prof. Kyriakakis of the University of Southern California (chorus recording of a classical music performance), while Signals 6-7 were obtained from the VOICES corpus of OGI's CSLU. The test signals can be found at [1].

The results of this test are given in Fig. 3, where the vertical lines indicate the 95% confidence limits. The *retransmission* mode was employed for this quality test. The proposed method was implemented operating with $M = 95$ samples, corresponding to the last row of Table 1, which translates into using 24 bits per sinusoid. This was compared to a popular sinusoidal coding method, namely that of [5], operating at the rates of 24 and 21 bits per sinusoid, denoted as "V&K 24 bits" and "V&K 21 bits" in the figure, respectively. From the figure it can be clearly seen that the proposed method achieves similar quality to state-of-the-art sinusoidal coding methods for the same bitrate. The results for the lower bitrate of 21 bits per sinusoid indicate that listeners can distinguish the reduction in bitrate and thus in quality for some of the signals. Consequently, it was sensible to compare our method operating at 24 bits per sinusoid with the method of [5] at that same rate and not lower. It is noted that more recent methods such as [6, 7] perform better than the method

[1] http://www.ics.forth.gr/~mouchtar/cs4sm/

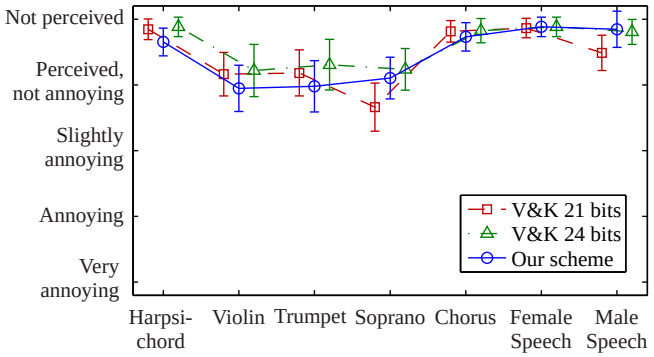

Fig. 3. Results of quality rating tests for $K = 25$ sinusoids per frame. V&K refers to the method of [5] with the given target entropy.

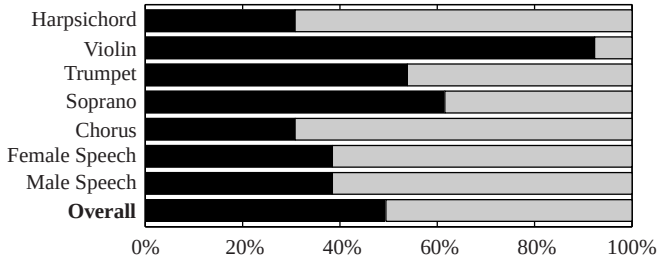

Fig. 4. Results of the preference listening tests for retransmitted signals (black) over 1% FRE interpolated signals (grey).

of [5] and thus could achieve similar performance to our method using less bits per sinusoid, depending though on the particular signal used. It is not claimed here that the proposed approach can result in lower bitrates than current state-of-the-art methods. Rather it shown that is possible to achieve similar performance, with a system which is based on a novel approach and can possibly be improved in terms of bitrate, while introducing the advantages due to the CS methodology, as stated in Section 1. It is noted that for all listening tests the sinusoidal error signal was obtained and added to the sinusoidal part, so that audio quality is judged without placing emphasis on the stochastic component, similarly to other tests in this area [5, 7].

The second part of the listening tests was a *preference* test, which indicates the quality that can be achieved in the *interpolation* mode of operation for our system, compared to the *retransmission* mode. The results of a comparison of the interpolation mode with a 1% probability of FRE to the retransmission case (*i.e.*, no FREs) are given in Fig. 4. It can be seen that in this case, there is a preference towards the retransmission mode signals but not in all seven signals. For this purpose, the overall preference is also given which indicates only a small preference to the retransmission mode signals, and indicates that in most cases the 1% frame errors can be acceptably corrected with the interpolation method. In contrast, for the case of 10% frame errors shown in Fig. 5, the preference is clear towards the retransmission operation mode, which indicates that interpolation can no longer conceal the frame errors to an acceptable degree.

6. CONCLUSIONS

We have shown that our method of using compressed sensing to encode a sinusoidally-modeled audio signal can achieve comparable performance in terms of audio quality of current sinusoidal coding methods, although superiority of the proposed method in terms of bitrate is not claimed. The advantages of the proposed approach are due to the application of CS to the sinusoidal audio coding, namely inherent encryption, robustness to network errors, and lack of need

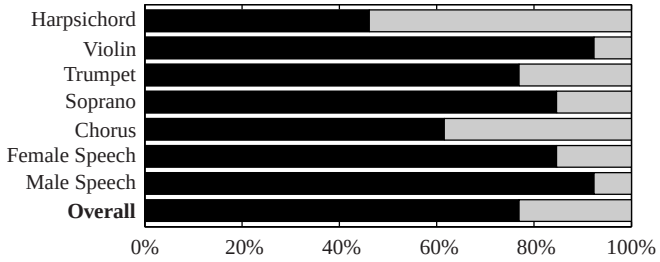

Fig. 5. Results of the preference listening tests for retransmitted signals (black) over 10% FRE interpolated signals (grey).

for training of the system. Furthermore, since this an early attempt to apply the CS methodology to the problem of sinusoidal audio coding, it is anticipated that further performance improvement can be achieved in subsequent work on this subject.

7. ACKNOWLEDGMENTS

The authors would like to thank Christos Tzagarakis for his help organizing the listening tests and all the volunteers who participated.

8. REFERENCES

[1] A. Griffin, C. Tzagarakis, T. Hirvonen, A. Mouchtaris, and P. Tsakalides, "Exploiting the sparsity of the sinusoidal model using compressed sensing for audio coding," in *Proc. Workshop on Signal Processing with Adaptative Sparse Structured Representations (SPARS'09), St. Malo, France*, April 2009.

[2] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. ASSP-34, no. 4, pp. 744–754, August 1986.

[3] X. Serra and J. O. Smith, "Spectral modeling sythesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14(4), pp. 12–24, Winter 1990.

[4] K. N. Hamdy, M. Ali, and A. H. Tewfik, "Low bit rate high quality audio coding with combined harmonic and wavelet representation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Atlanta, Georgia, USA*, May 1996.

[5] R. Vafin, D. Prakash, and W. B. Kleijn, "On frequency quantization in sinusoidal audio coding," *IEEE Signal Proc. Lett.*, vol. 12, no. 3, pp. 210–213, March 2005.

[6] R. Vafin and W. B. Kleijn, "Jointly optimal quantization of parameters in sinusoidal audio coding," in *Proc. IEEE Workshop on Applications of Signal Process. to Audio and Acoust. (WASPAA)*, October 2005.

[7] P. Korten, J. Jensen, and R. Heusdens, "High resolution spherical quantization of sinusoidal parameters," *IEEE Trans. Speech and Audio Process.*, vol. 13, no. 3, pp. 966–981, 2007.

[8] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, February 2006.

[9] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[10] J. Laska, S. Kirolos, Y. Massoud, R. Baraniuk, A. Gilbert, M. Iwen, and M. Strauss, "Random sampling for analog-to-information conversion of wideband signals," in *Proc. IEEE Dallas Circuits and Systems Workshop (DCAS), Dallas, TX, USA*, 2006.

[11] G.H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "Complex-valued sparse representation based on smoothed ℓ_0 norm," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Las Vegas, Nevada, USA*, April 2008.

[12] ITU-R, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," 1997.