Exploiting the Sparsity of the Sinusoidal Model Using Compressed Sensing for Audio Coding

Anthony Griffin, Christos Tzagkarakis, Toni Hirvonen, Athanasios Mouchtaris, Panagiotis Tsakalides

Exploiting the Sparsity of the Sinusoidal Model Using Compressed Sensing for Audio Coding

Anthony Griffin, Christos Tzagkarakis, 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
Email: {agriffin, tzagarak, tmhirvo2, mouchtar, tsakalid}@ics.forth.gr

***Abstract*—Audio signals are represented via the sinusoidal model as a summation of a small number of sinusoids. This approach introduces sparsity to the audio signals in the frequency domain, which is exploited in this paper by applying Compressed Sensing (CS) to this sparse representation. CS allows sampling of signals at a much lower rate than the Nyquist rate if they are sparse in some basis. In this manner, a novel sinusoidal audio coding approach is proposed, which differs in philosophy from current state-of-the-art methods which encode the sinusoidal parameters (amplitude, frequency, phase) directly. It is shown here that encouraging results can be obtained by this approach, although inferior at this point compared to state-of-the-art. Several practical implementation issues are discussed, such as quantization of the CS samples, frequency resolution vs. coding gain, error checking, etc., and directions for future research in this framework are proposed.**

I. Introduction

The growing demand for audio content far outpaces the corresponding growth in users' storage space or bandwidth. Thus there is a constant incentive to further improve the compression of audio signals. This can be accomplished either by applying compression algorithms to the actual samples of a digital audio signal, or initially using a signal model and then encoding the model parameters as a second step. In this paper, we explore a novel method for encoding the parameters of the sinusoidal model [1].

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 modelled to further improve the resulting subjective quality of the sinusoidal model [2]. The sinusoidal model allows for a compact representation of the original signal and for efficient encoding and quantization. State-of-the-art methods of encoding and compressing the parameters of the sinusoidal model (amplitudes, frequencies, phases) are based on directly encoding these parameters [3]–[6]. In this paper, we propose using the emerging compressed sensing (CS) [7], [8] methodology to encode and compress the sinusoidally-modelled 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-modelled 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.

Our method encodes the time-domain signal instead of the sinusoidal model parameters as state-of-art methods propose [3]–[6]. 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 is to our knowledge the first attempt to exploit the sparse representation of the sinusoidal model for audio signals using compressed sensing, and it is shown here that several interesting questions arise in this context.

II. Sinusoidal Model

The sinusoidal model was initially applied in the analysis/synthesis of speech [1]. 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 [2]. 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 [4]).

III. 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-modelled 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.

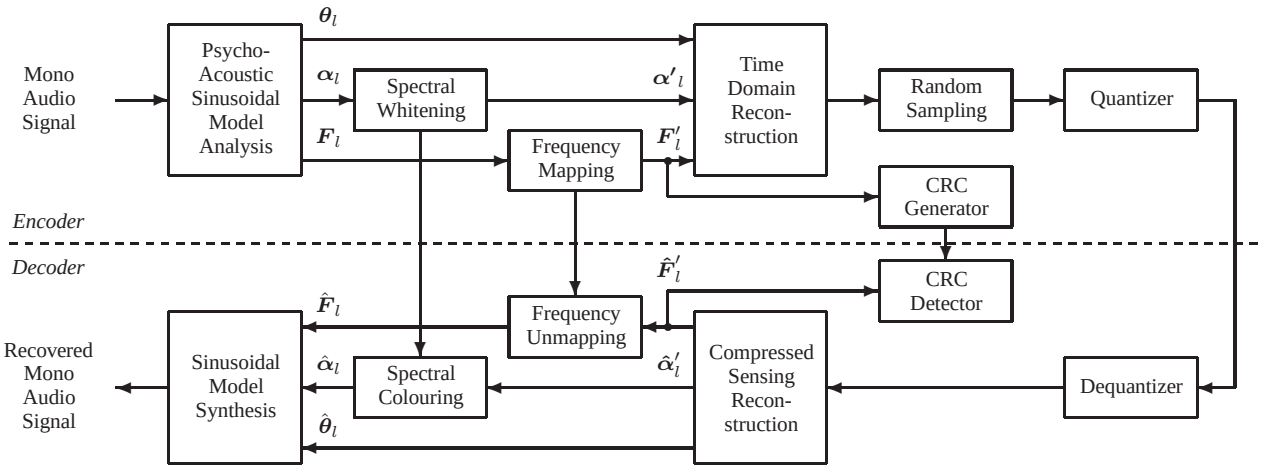

Fig. 1. A block diagram of the proposed system. In the encoder, the sinusoidal part of the monophonic audio signal is encoded by randomly sampling its time-domain representation, and then quantizing the random sample using scalar quantization.

A. 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 [9] 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.

B. 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 [7] [8] that

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

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 [10], [11] that $p < 1$ can outperform the $p = 1$ case. It is these methods that we use for reconstruction in this paper. Further discussion of the algorithms used is presented in Section IV-D

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 Q, the number of bits used for quantization.

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. For the rest of this work, we assume that any frames with error can be corrected by retransmission. Given that with a wise choice of parameters the probability of FRE (P_{FRE}) can remain quite small (*e.g.* below 10^{-2}), the additional bitrate burden due to retransmission will be negligible.

IV. System Design

A block diagram of our proposed system is depicted in Fig. 1. The audio signal is first passed through a psychoacoustic sinusoidal modelling 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—as discussed in Section IV-A—and the frequencies remapped, as discussed in Section IV-B. 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 Q 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

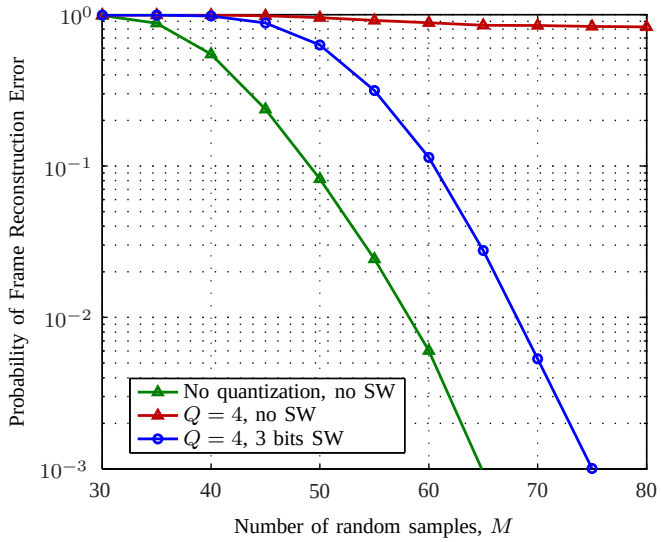

Fig. 2. Probability of frame reconstruction error *vs* the number of random samples per frame for three cases: no quantization and no spectral whitening, $Q = 4$ bits quantization and no spectral whitening, and $Q = 4$ bits quantization and 3 bits for spectral whitening.

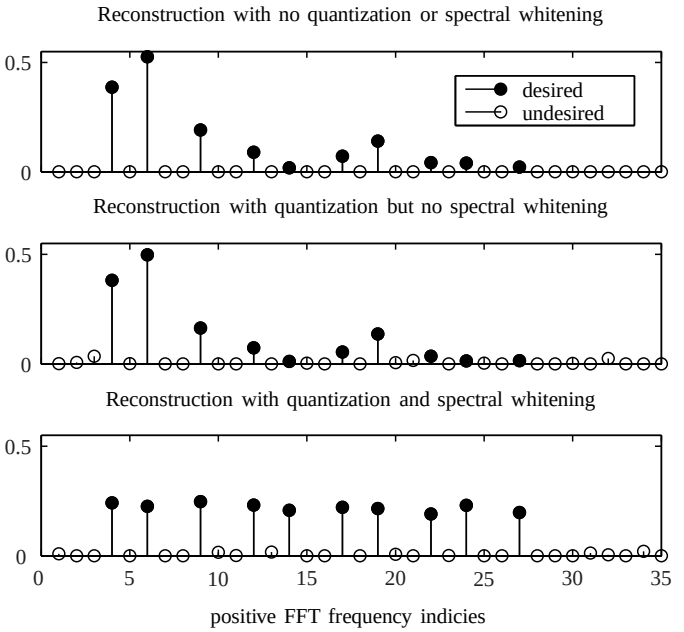

Fig. 3. Reconstructed frames showing the effects of 4-bit quantization and spectral whitening.

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 previously discussed.

In the tests employed in this paper, we investigated the performance of the proposed system using $K = 10$ sinusoid components per frame and an $N = 256$-point FFT. All the audio signals were sampled at 22 kHz with a 10 ms window and 50% overlapping between frames. The data used for the results this section are around 10,000 frames of the audio data used in the listening tests of Section V.

A. Spectral Whitening

Once we quantize the M samples that we send, we find that P_{FRE} increases significantly. Equivalently, the M required to achieve the same P_{FRE} increases. Fig. 2 illustrates this dramatically; the "$Q = 4$, no SW" curve in Fig. 2 shows that our system becomes unusable for the 4-bit quantization with no spectral whitening case.

As our quantization is performed in the time domain, it has an effect similar to adding noise to all of the frequencies in the recovered frame $\hat{\boldsymbol{x}}_l$. We must then select the K largest components of $\hat{\boldsymbol{x}}_l$ and zero the remaining components. This is illustrated in Fig. 3. The top plot shows the reconstruction without quantization, and the desired components are the K largest values in the reconstruction. The middle plot shows the effect of 4-bit quantization, where some of the undesired components are now larger than the desired ones and an FRE will occur.

To alleviate this problem we implemented spectral whitening in the encoder. We first tried to employ envelope estimation of the sinusoidal amplitudes based on [12], but we could not get acceptable performance without incurring too large an overhead. Our final choice was to simply divide each amplitude by a 3-bit quantized version of itself, and send this whitening information along with the quantized measurements. The result is seen the bottom plot in Fig. 3, where the desired components are clearly the K largest values and thus no FRE will occur. This whitening incurs an overhead of approximately $3K$ bits, but the savings in reduced M and Q allow us to achieve a lower overall bitrate for a given probability of FRE.

In the case of 4-bit quantization and 3-bit spectral whitening, our system again becomes feasible as illustrated in Fig. 2. In fact, this case only requires 10 more random samples than the case with no quantization.

B. 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 modelled 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 the resulting N_{FM} will also 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

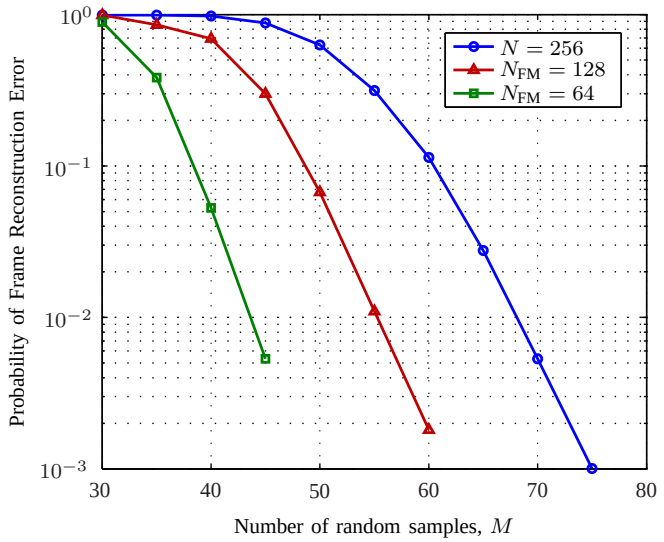

Fig. 4. Probability of frame reconstruction error *vs* the number of random samples per frame for various values of frequency mapping, with 4-bit quantization of the random samples, and 3 bits for spectral whitening.

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 85% of the frames could be mapped by an C_{FM} equal to 4, giving an $N_{\text{FM}} = 64$.

The clear decrease in the required M for a given probability of FRE for various values of N_{FM} is illustrated in Fig. 4. The final bitrates achieved in all of the above cases are discussed in Section IV-E.

C. Quantization and entropy coding of random samples

We employed a uniform scalar quantizer to quantize the random samples. To further reduce the number of bits required for each quantization value, an entropy coding scheme [13] may be used after the quantizer. Entropy coding is a lossless data compression scheme, which maps the more probable codewords (quantization indices) into shorter bit sequences and less likely codewords into longer bit sequences. In our implementation Huffman coding is used as an entropy encoding technique. Thus it is expected that the average codeword length will be reduced after the Huffman coding. The average codeword length is defined as

$$\bar{l} = \sum_{i=1}^{2^b} p_i l_i, \tag{6}$$

where p_i is the probability of occurrence for the i-th codeword, l_i is the length of each codeword and 2^b is the total number of codewords, as b is the number of bits assigned to each codeword before the Huffman encoding.

Table I presents the percentages of compression that can be achieved through Huffman encoding for each audio signal for $Q = 3, 4$, and 5 bits of quantization. The possible compression clearly decreases as Q increases, but for our chosen case of $Q = 4$, a compression of about 8% is clearly achievable. It must be noted though that this requires significant training—something we prefer to avoid—so this is presented as an optional enhancement.

TABLE I
COMPRESSION ACHIEVED AFTER ENTROPY CODING. (Q: CODEWORD LENGTH IN BITS, $\bar{Q}$: AVERAGE CODEWORD LENGTH IN BITS AFTER ENTROPY CODING, PC: PERCENTAGE OF COMPRESSION ACHIEVED)

Signal	Q	$\bar{Q}$	PC	Q	$\bar{Q}$	PC	Q	$\bar{Q}$	PC
Violin	3	2.64	11.9%	4	3.70	7.5%	5	4.73	5.4%
Harpsichord	3	2.62	12.7%	4	3.67	8.2%	5	4.70	6.1%
Trumpet	3	2.60	13.6%	4	3.63	9.3%	5	4.66	6.8%
Soprano	3	2.59	13.7%	4	3.62	9.4%	5	4.65	7.0%
Chorus	3	2.64	12.2%	4	3.68	8.0%	5	4.71	5.9%
Female speech	3	2.60	13.2%	4	3.64	9.0%	5	4.68	6.5%
Male speech	3	2.60	13.4%	4	3.63	9.2%	5	4.66	6.8%
Overall	3	2.61	12.9%	4	3.65	8.7%	5	4.68	6.3%

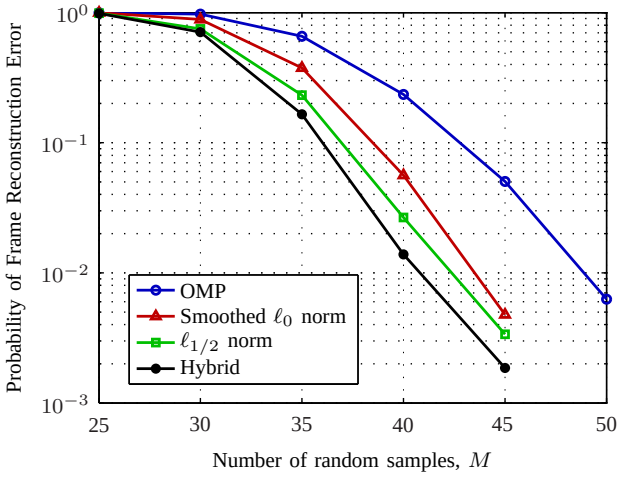

Fig. 5. Probability of frame reconstruction error *vs* the number of random samples per frame for different reconstruction algorithms, with 4-bit quantization of the random samples, 3 bits for spectral whitening, and $N_{\text{FM}} = 64$.

D. Reconstruction Algorithms

In order to ensure we obtained the lowest-possible bitrate, we analyzed the performance of a variety of reconstruction algorithms. The ones we found to perform best for our system were the ℓ_p norm with $p = 1/2$ and the smoothed ℓ_0 norm, described in [10] and [11], respectively. Fig. 5 presents the results of simulations with our finally-chosen parameters. We have included the results obtained using orthogonal matching pursuit (OMP) [14] for reference. The smoothed ℓ_0 norm is the best choice of algorithm as it is least complex—being the same order of complexity as OMP—and performs almost as well as the $\ell_{1/2}$ norm. The $\ell_{1/2}$ norm is about 1000 times as complex as the other two algorithms, although the authors do state that [10] is a relatively naïve implementation.

The final curve in Fig. 5—labelled "Hybrid"—is new reconstruction algorithm that we are proposing. In a sense, it can be considered as a "super" algorithm as it makes use of all the other algorithms. As we can tell whether or not a particular algorithm has successfully reconstructed a frame—by checking the CRC to see if an FRE has occurred—we can then try a different algorithm and check whether that succeeds. This is only possible as different algorithms fail for different frames. Thus for the hybrid algorithm to fail, *all three* of the other algorithms must fail. This clearly provides the best possible performance, but incurs additional complexity due to the fact that

TABLE II
PARAMETERS TO ACHIEVE A PROBABILITY OF FRE OF APPROXIMATELY 10^{-3}, FOR $N = 256$, $N_{\text{FM}} = 128$, $K = 10$

N_{FM}	Q	M	raw bitrate	overhead CRC	overhead FM	overhead SW	final bitrate	per sinusoid
128	5	60	300	8	11	50	369	36.9
128	4	60	240	8	11	50	319	31.9
128	3	70	210	8	11	50	279	27.9

TABLE III
PARAMETERS THAT ACHIEVE A PROBABILITY OF FRE OF APPROXIMATELY 10^{-2} WITH $N = 256$ AND $K = 10$

N_{FM}	Q	M	raw bitrate	overhead CRC	overhead FM	overhead SW	final bitrate	per sinusoid
256	4	68	272	8	0	30	310	31.0
128	4	55	220	8	11	30	269	26.9
64	4	43	172	8	23	30	233	23.3

multiple algorithms may need to be run.

In practice, this effect could be minimised by running the smoothed ℓ_0 norm—the least-complex algorithm—and only running the others if this fails. It is clear from Fig. 5 that using the hybrid algorithm would save about 2 random samples, and with $Q = 4$, and $K = 10$, this equates to almost 1 bit per sinusoid. Nevertheless, we chose not to use this algorithm in the majority of our simulations due to the increased complexity.

E. Bitrates

In Table II, three sets of M and Q are given (per audio frame) that achieve a probability of FRE of approximately 10^{-3}, for the $N = 256$, $N_{\text{FM}} = 128$, and $K = 10$ case with differing values of Q. The overhead consists of the extra bits required for the CRC, the frequency mapping and the spectral whitening. These are the parameters that were used for the listening tests of Section V. Note that at this point in the research we were aiming for a probability of FRE of approximately 10^{-3} rather than 10^{-2} and were using 5 bits for spectral whitening instead of 3 bits.

After the results of the first set of listening tests, we moved to focus on $Q = 4$, and Table III presents the bitrates achievable for a probability of FRE of approximately 10^{-2} corresponding to the curves in Fig. 4. It is clear that the overhead incurred from spectral whitening and frequency mapping is more than accounted for by significant reductions in M, resulting in overall lower bitrates.

In Fig. 6 we present the P_{FRE} *vs* M for the individual signals used in our simulations and listening tests with for the case with $N_{\text{FM}} = 64$, $Q = 4$, 3-bit spectral whitening and the smoothed ℓ_0 norm reconstruction algorithm. It is clear that for a P_{FRE} of 10^{-2} the M does not vary much, say from 43 to 44. Equivalently, with a fixed M of 43, the P_{FRE} only varies from about 0.008 to 0.018. This supports our claim that our system does not require any training, as this is a wide variety of signals that perform similarly. See Section V for more details on the signals used.

It should also be noted that the lowest bitrate for the $N_{\text{FM}} = 64$ case can be reduced to under 21 bits per sinusoid if entropy coding and the hybrid reconstruction algorithm are used, although this will require training and an increase in complexity in the decoder.

V. LISTENING TESTS

In this section, we examine the performance of our proposed system, with respect to the resulting audio quality. Two types of

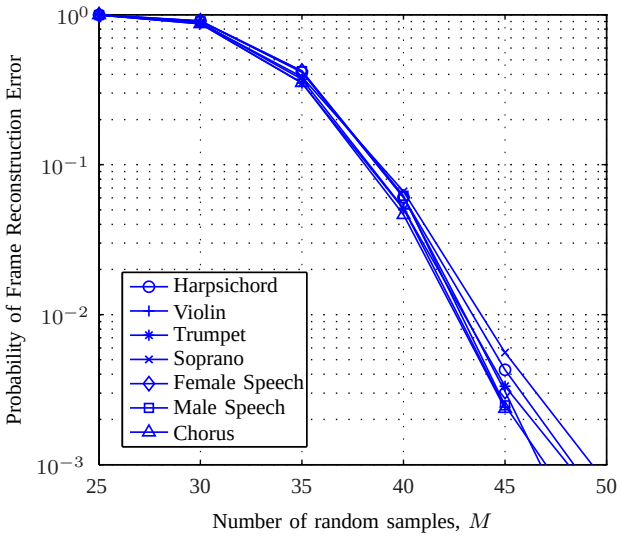

Fig. 6. Probability of frame reconstruction error *vs* the number of random samples for individual signals, with 4-bit quantization of the random samples, 3 bits for spectral whitening, and $N_{\text{FM}} = 64$.

monophonic listening tests were performed, where volunteers were presented with audio files using high-quality headphones in a quiet office room. The first test was based on the ITU-R BS.1116 [15] 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 (a recording of the chorus of a classical music performance), while Signals 6-7 were obtained from the VOICES corpus [16] of OGI's CSLU. The audio signals used in the tests can all be found at our website[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, and this is similar to other tests in this area [4], [6]. The signals were downsampled to 22 kHz, so that the stochastic component does not affect the resulting quality to a large degree. This is because the stochastic component is particularly dominant in higher frequencies, thus its effect would be more evident in the 44.1 kHz than the 22 kHz sampling rate, while the focus of the paper is on the sinusoidal rather than the stochastic component. The second type of test employed was a preference test (forced choice), where listeners indicated their preference among a pair of audio signals at each time, in terms of quality. The sinusoidal analysis/synthesis window was 10 ms long, with 50% overlapping.

One quality and one preference test were conducted to evaluate the quality of the audio signals when modelled by $N = 256$-point FFT and $K = 10$ sinusoids per frame (no psychoacoustic model employed). The goal was to evaluate the resulting quality in this case, regarding the effect of the number of bits of quantization and number of random samples in the resulting audio quality. Eleven volunteers participated in this pair of listening tests. The results of the *quality* test are shown in Fig. 7, where the vertical lines indicate

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

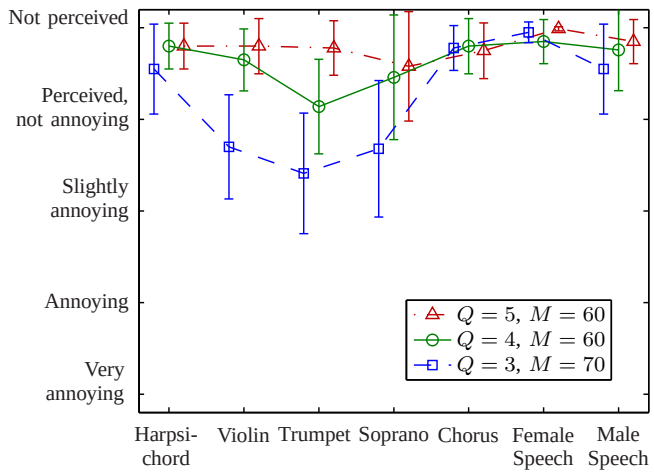

Fig. 7. Results of quality rating listening test for 10 sinusoids per frame, for various choices of bits per sample (Q) and number of random samples (M).

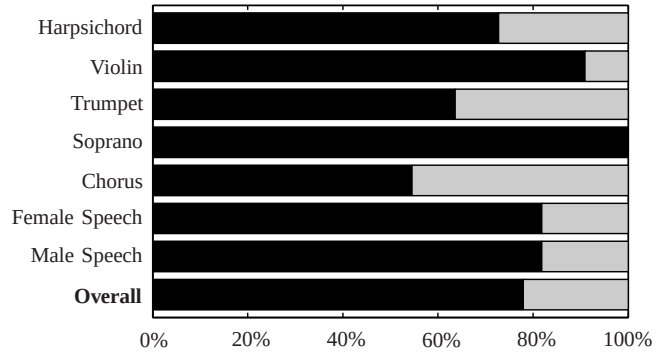

Fig. 8. Results of the preference listening tests for 10 sinusoids with $Q = 4$, $M = 60$ signals (black) over $Q = 3$, $M = 80$ signals (grey).

the 95% confidence limits. Three different cases of encoding were used. The resulting bitrates per audio frame for these three cases are given in Table II. It is clear from Fig. 7 that the quality for the $Q = 5$, $M = 60$ and $Q = 4$, $M = 60$ cases remains well above 4.0 grade (perceived, not annoying), even for the more complex chorus signals, while for the $Q = 3$, $M = 70$ case—which represents the lowest bitrate of the three cases—the quality deteriorates. Thus we can conclude that with a bitrate of 300 bits per audio frame we can achieve very good quality (above 4.0). 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 I.

It is also interesting to investigate whether for a fixed bitrate, more bits should be put into the number of bits/sample Q or the number of (random) samples M. A *preference* listening test was conducted for this purpose, with audio signals encoded with $Q = 4$, $M = 60$ and $Q = 3$, $M = 80$. It is clear from Fig. 8 that $Q = 4$, $M = 60$ was a preferred distribution of available bits, although this was more significant for some signals over others. We can conclude from this test that using more bits/sample is more important than increasing the number of samples (for a constant bitrate), especially at low bitrates where the effect of quantization is more evident.

VI. Conclusions

In this paper, an initial investigation was performed into whether the Compressed Sensing framework can be employed to encode the harmonic part of audio signals which are modelled by the sinusoidal model. This was proposed based on the fact that CS results in fewer measurements than the Nyquist rate for sparse signals, and the harmonic part of audio signals is sparse by definition in the Fourier domain. The results obtained are encouraging, and at the same time raise many issues for further investigation such as quantization of the samples, addressing incorrectly reconstructed audio frames, the tradeoff between frequency resolution and number of samples needed, improving the spectral whitening, and reducing the decoder complexity.

Acknowledgment

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.

The authors would like to thank all the volunteers who partcipated in the listening tests.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] R. Chartrand, "Exact reconstructions of sparse signals via nonconvex minimization," *IEEE Signal Proc. Lett.*, vol. 14, no. 10, 2007.

[11] G. 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] O. Cappe, J. Laroche, and E. Moulines, "Regularized estimation of cpestrum envelope from discrete frequency points," in *IEEE ASSP Workshop on App. of Sig. Proc. to Audio and Acoust.*, October 1995.

[13] K. Sayood, *Introduction to data compression*. Morgan Kauffman, 2000.

[14] J. Tropp and A. Gilbert, "Signal recovery from partial information via orthogonal matching pursuit," 2005, preprint.

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

[16] A. Kain, "High resolution voice transformation," Ph.D. dissertation, OGI School of Science and Engineering at Oregon Health and Science University, October 2001.