Single-channel and Multi-channel Sinusoidal Audio Coding Using Compressed Sensing

Anthony Griffin*, Toni Hirvonen, Christos Tzagkarakis,
Athanasios Mouchtaris, *Member, IEEE,* and Panagiotis Tsakalides, *Member, IEEE*

Abstract—Compressed sensing (CS) samples signals at a much lower rate than the Nyquist rate if they are sparse in some basis. In this paper, the CS methodology is applied to sinusoidally-modeled audio signals. As this model is sparse by definition in the frequency domain (being equal to the sum of a small number of sinusoids), we investigate whether CS can be used to encode audio signals at low bitrates. In contrast to encoding the sinusoidal parameters (amplitude, frequency, phase) as current state-of-the-art methods do, we propose encoding few randomly selected samples of the time-domain description of the sinusoidal component (per signal segment). The potential of applying compressed sensing both to single-channel and multi-channel audio coding is examined. The listening test results are encouraging, indicating that the proposed approach can achieve comparable performance to that of state-of-the-art methods. Given that CS can lead to novel coding systems where the sampling and compression operations are combined into one low-complexity step, the proposed methodology can be considered as an important step towards applying the CS framework to audio coding applications.

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

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 using initially a signal model and then encoding the model parameters as a second step. In this paper, we propose a novel method for encoding the parameters of the sinusoidal model.

The sinusoidal model represents an audio signal using a small number of time-varying sinusoids [1]. The remainder error signal—often termed the residual signal—can also be modeled 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. Extending the sinusoidal model to multi-channel audio applications has also been proposed (*e.g.* [3]).

Various methods for quantization of the sinusoidal model parameters (amplitude, phase, and frequency) have been proposed in the literature. Initial methods in this area suggested quantizing the parameters independently of each other [4]–[8]. The frequency locations of the sinusoids were quantized based on research into the just noticeable differences in frequency (JNDF), while the amplitudes were quantized based either on the just noticeable differences in amplitude (JNDA) or the estimated frequency masking thresholds. In these initial quantizers, phases were uniformly quantized, or were not quantized at all for low-bitrate applications. More recent quantizers operate by jointly encoding all the sinusoidal parameters based on high-rate theory and can be expressed analytically [9]–[12]. The bitrates achieved by these methods can be further reduced using differential coding *e.g.*, [13]. It must be noted that all the aforementioned methods encode the sinusoidal parameters independently for each short-time segment of the audio signal. Extensions of these methods, where the sinusoidal parameters can be jointly quantized across neighboring segments, have recently been proposed *e.g.*, [14].

In this paper, we propose using the emerging compressed sensing (CS) [15], [16] 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 *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. We present such an investigation of how CS can be applied to encoding the time-domain signal of the model instead of the sinusoidal model parameters as state-of-the-art methods propose, extending our recent work in [17], [18]. We extend our previous work in terms of providing more results for the single-channel audio coding case, but also we propose here a system which applies CS to the case of sinusoidally-modeled multi-channel audio. At the same time, the paper proposes a psychoacoustic modeling analysis for the selection of sinusoidal components in a multi-channel audio recording, which provides a very compact description of multi-

channel audio and is very efficient for low-bitrate applications.

This is to our knowledge the first attempt to exploit the sparse representation of the sinusoidal model for audio signals using compressed sensing, and many interesting and important issues are raised in this context. The most important problems encountered in this work are summarized in this paragraph. The encoding operation is based on randomly sampling the time-domain sinusoidal signal, which is obtained after applying the sinusoidal model to a monophonic or multi-channel audio signal. The random samples can be further encoded (here scalar quantization is suggested, but other methods could be used to improve performance). An issue that arises 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; this was addressed by suggesting a spectral whitening procedure for the sinusoidal amplitudes. Another issue is that the sinusoidal model estimated frequencies should correspond to single bins of the discrete Fourier transform, or else the sparsity requirement cannot be satisfied. In practice, this translates into encoding the sinusoidal parameters selected from a peak-picking procedure (with the possible inclusion of a psychoacoustic model), without further refinement of the estimated frequencies. This important problem can be addressed (as explained in detail later) by employing zero-padding in the Fourier analysis (*i.e.,* improving the frequency resolution by shortening the bin spacing), and also by employing interpolation techniques in the decoder (since sparsity is not needed after the CS decoding). The improved frequency resolution resulted in a need to increase the number of CS measurements, and consequently the bitrate, and this problem was alleviated by employing a process termed "frequency mapping". Another important problem which was addressed in this paper is the fact that CS theory allows for signal reconstruction with high probability but not with certainty; three different ways of overcoming this problem (termed "operating modes") are suggested in this paper. In summary, several practical problems were raised during our research; by providing a complete end-to-end design of a CS-based sinusoidal coding system, this paper both clarifies several limitations of CS to audio coding, but also presents ways to overcome them, and in this sense we believe that this paper will be of interest to researchers working on applying the CS theory into signal coding.

The paper deals only with encoding the sinusoidal part of the model (*i.e.* there is no treatment for the residual signal). It is noted that other than the proposed method, the authors are only familiar with the work of [19] for applying the CS methodology to audio coding in general. While our focus in this paper is on exploiting the sinusoidal model in this context, in [19] the goal was to exploit the excitation / filter model using CS.

The importance of applying CS theory to audio coding lies mainly to the applicability of CS to sensor network applications. Sensor-based local encoding of audio signals could enable a variety of audio-related applications, such as environmental monitoring, recording audio in large outdoor venues, and so forth. This paper provides an important step towards applying CS to audio coding, at least in low-bitrate audio applications where the sinusoidal part of an audio signal provides sufficient quality. It is shown here for multi-channel audio signals that, except from one primary (reference) audio channel, a simple low-complexity system can be used to encode the sinusoidal model for all remaining channels of the multi-channel recording. This is an important result given that research in CS is still at an early stage, and its practical value in coding applications is still unclear.

The remainder of the paper is organized as follows. In Section II, background information about the sinusoidal model is given, and a novel psychoacoustic model for sinusoidal modeling for multi-channel audio signals is proposed. Background information about the CS methodology is presented in Section III. In Section IV, a detailed discussion about the practical implementation of the method is provided related to issues such as alleviating the effects of quantization (Section IV-A); bitrate improvements (Section IV-B); quantization and entropy coding (Section IV-C); CS reconstruction algorithms (Section IV-D); achieved bitrates (Section IV-E); operating modes (Section IV-F); and complexity (Section IV-G). The discussion of Section IV is then extended to the multi-channel case in Section V. In Section VI, results from listening tests demonstrate the audio quality achieved with the proposed coding scheme for the single-channel (Section VI-A) and the multi-channel case (Section VI-B), while in Section VII concluding remarks are made.

II. SINUSOIDAL MODEL

The sinusoidal model was initially used in the analysis/synthesis of speech [1]. A short-time segment of an audio signal $s(n)$ is represented as the sum of a small number of K sinusoids with time-varying amplitudes and frequencies. This can be written as

$$s(n) = \sum_{k=1}^{K} \alpha_k \, \cos(2\pi f_k n + \theta_k) \qquad (1)$$

where α_k, f_k, and θ_k are the amplitude, frequency, 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 in 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 by matching the peaks of the previous frame to the current frame, using linear amplitude interpolation and cubic phase interpolation.

A more accurate representation of audio signals is achieved when a stochastic component is included in the model. This model is usually termed as sinusoids plus noise model, or deterministic plus stochastic decomposition. In this model, the sinusoidal part corresponds to the "deterministic" part

of the signal due to the structured nature of this model. The remaining signal is the sinusoidal noise component $e(n)$, also referred to here as residual or sinusoidal error signal, which is the "stochastic" part of the audio signal, since it is very difficult to accurately model, but at the same time essential for high-quality audio synthesis. Accurately modeling the stochastic component has been examined both for the single-channel case, *e.g.* [2], [20], [21] and the multi-channel audio case [3]. Practically, after the sinusoidal parameters are estimated, the noise component is computed by subtracting the sinusoidal component from the original signal. Note that in this paper we are only interested in encoding the sinusoidal part.

A. Single-channel sinusoidal selection

To perform single-channel sinusoidal analysis, we employed state-of-the-art psychoacoustic analysis based on [22]. In the i-th iteration, the algorithm picks a perceptually optimal sinusoidal component frequency, amplitude, and phase. This choice minimizes the perceptual distortion measure

$$D_i = \int A_i(\omega)|R_i(\omega)|^2 d\omega, \qquad (2)$$

where $R_i(\omega)$ is the Fourier transform of the residual signal (original frame minus the currently selected sinusoids) after the i-th iteration, and $A_i(\omega)$ is a frequency weighting function set as the inverse of the current masking threshold energy.

One issue with CS encoding is that no further refinement of the sinusoid frequencies can be performed in the encoder, because frequencies which do not correspond to exact frequency bins would result in loss of the sparsity in the frequency domain. This is an important problem, because it implies that we must restrict the sinusoidal frequency estimation to the selection of frequency bins (*e.g.* following a peak-picking procedure), without the possibility of further refinement of the estimated frequencies in the encoder. This can be alleviated by zero-padding the signal frame, in other words improving the frequency resolution during the parameter estimation by reducing the bin spacing. We have found, though, that for CS-based encoding this can be performed to a limited degree, as zero-padding will increase the number of measurements that must be encoded as explained in Section IV (and consequently the bitrate). Fortunately, this problem can be partly addressed by employing the "frequency mapping" procedure, described in Section IV. Furthermore, since the sparsity restriction need not hold after the signal is decoded, frequency re-estimation can be performed in the decoder, such as interpolation among frames.

B. Multi-channel sinusoidal selection

To perform multi-channel sinusoidal analysis, we have extended the sinusoidal modeling method presented in [23]—which employs a matching pursuit algorithm to determine the model parameters of each frame—to include the psychoacoustic analysis of [22]. For the multichannel case, in each iteration, the algorithm picks a sinusoidal component

frequency that is optimal for all channels, as well as channel-specific amplitudes and phases. This choice minimizes the perceptual distortion measure

$$D_i = \sum_c \int A_{i,c}(\omega)|R_{i,c}(\omega)|^2 d\omega, \qquad (3)$$

where $R_{i,c}(\omega)$ is the Fourier transform of the residual signal of the c-th channel after the i-th iteration, and $A_{i,c}(\omega)$ is a frequency weighting function set as the inverse of the current masking threshold energy. The contributions of each channel are simply summed to obtain the final measure.

An important question is what masking model is suitable for multi-channel audio where the different channels have different binaural attributes in the reproduction. In transform coding, a common problem is caused by Binaural Masking Level Difference (BMLD); sometimes quantization noise that is masked in monaural reproduction is detectable because of binaural release, and using separate masking analysis for different channels is not suitable for loudspeaker rendering. However, this effect in parametric coding is not so well established.

We performed preliminary experiments using: (a) separate masking analysis, *i.e.* individual $A_{i,c}(\omega)$ based on the masker of channel c for each signal separately (see (3)); (b) the masker of the sum signal of all channel signals to obtain $A_i(\omega)$ for all c; and (c) power summation of the other signals' attenuated maskers to the masker of channel c according to

$$A_{i,c}(\omega) = 1 / \left(M_{i,c}(\omega) + \sum_{\substack{k \\ k \neq c}} w_k M_{i,k}(\omega) \right). \qquad (4)$$

In the above equation, $M(\omega)$ indicates the masker energy, w_k the estimated attenuation (panning) factor that was varied heuristically, and k iterates through all channel signals excluding c. In this paper we chose to use the first method, *i.e.* separate masking analysis for channels ($w_k = 0$), for the reason that we did not find notable differences in BMLD noise unmasking, and that the sound quality seemed to be marginally better with headphone reproduction. For loudspeaker reproduction, the second or third method may be more suitable.

The use of this psychoacoustic multi-channel sinusoidal model resulted in sparser modeled signals, increasing the effectiveness of our compressed sensing encoding.

III. COMPRESSED SENSING

Compressed sensing [15], [16]—also known as compressive sensing or compressive sampling—is an emerging field which has grown up in response to the increasing amount of data that needs to be sensed, processed and stored. A great majority of this data is compressed as soon as it has been sensed at the Nyquist rate. The idea behind compressed sensing is to go directly from the full-rate, analog signal to the compact representation by using measurements in the sparse basis. Thus, the CS theory is based on the assumption that the signal of interest is *sparse* in some basis as it can be accurately and efficiently represented in that basis. This is not possible unless the sparse basis is known in advance, which is generally not the case. Thus compressed sensing uses *random* measurements

in a basis that is *incoherent* with the sparse basis. Incoherence means that no element of one basis has a sparse representation in terms of the other basis [15], [16]. This gives compressed sensing its *universality*, the same measurement technique can be used for signals that are sparse in different bases. This still results in the important part of signal being captured with many less measurements than the Nyquist rate.

Compressed sensing has found applications in many areas: image processing [24], spatial localization [25], [26], medical signal processing [27], to name a few. In addition, compressed sensing is particularly suited to multiple sensor scenarios, making it a good choice for wireless sensor networks [26], [28].

Although sparse representations of sound exist, for example [29]–[31], compressed sensing has not yet been particularly successfully applied to audio signals. We surmise that this is due to the fact that the sparse bases for audio do not represent audio with enough sparsity, or that they do not integrate well into the compressed sensing methodology. In this paper we take a different approach, by applying compressed sensing to a parametrically modeled audio signal that we know is sparse. This is a novel application of compressed sensing as we are using it to encode a sparse signal *that is known in advance*. We now briefly review the compressed sensing methodology and set up a more formal framework for the work in the following sections.

A. Measurements

Let x_l be the N samples of the sinusoidal component in the lth frame. It is clear that x_l is a 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 x_l can be written as

$$x_l = \Psi X_l, \tag{5}$$

where Ψ is an $N \times N$ inverse FFT matrix, and X_l is the FFT of x_l. As x_l is a real signal, 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 x_l, where $M \ll N$, which result in the $M \times 1$ vector y_l. This measurement process can be written as

$$y_l = \Phi_l x_l = \Phi_l \Psi X_l, \tag{6}$$

where Φ_l is an $M \times N$ matrix representing the measurement process. For the CS reconstruction to work, Φ_l and Ψ 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 [32] for further discussion of random sampling. Note that in this case, Φ_l is formed by randomly-selected rows of the $N \times N$ identity matrix.

B. Reconstruction

Once 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 X_l. It has been shown [15], [16] that

$$\hat{X}_l = \arg\min \|X_l\|_p \qquad \text{s.t.} \qquad y_l = \Phi_l \Psi X_l, \tag{7}$$

with $p = 1$ will recover X_l with high probability if enough measurements are taken. Note that Φ_l is considered available at the receiver as all that is required to generate it is the same seed as that used in the transmitter. It has recently been shown in [33], [34] that $p < 1$ can outperform the $p = 1$ case. It is the method of [34] that we use for reconstruction in this paper. Further discussion of the reconstruction is presented in Section IV-D .

A property 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. Aside from the effects of the reconstruction algorithm, this probability is dependent on M, N, K and Q, the number of bits of quantization used.

Another important feature of the reconstruction is that when it fails, it can fail catastrophically for the whole frame. In our case, 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 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

$$f_{l,k} = \frac{2\pi F_{l,k}}{N}. \tag{8}$$

As 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 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 interpolation between the correct frames before and after the errored frame to estimate it. These issues are discussed further in Section IV-F.

IV. SINGLE-CHANNEL SYSTEM DESIGN

A block diagram of our proposed system for single-channel sinusoidal audio coding is depicted in Fig. 1. The audio signal is first passed through a psychoacoustic sinusoidal modeling block to obtain the sinusoidal parameters $\{F_l, \alpha_l, \theta_l\}$ for the current frame. These then go through what can be thought of as a "pre-conditioning" phase where the amplitudes are whitened and the frequencies remapped. The modified sinusoidal parameters $\{F'_l, \alpha'_l, \theta_l\}$ 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
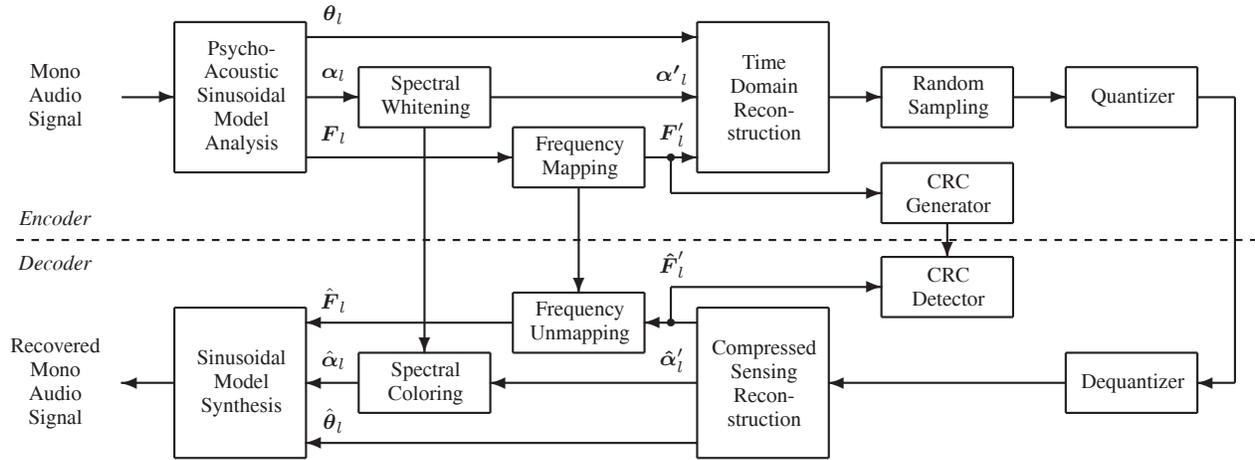
Fig. 1. Block diagram of the proposed system for the single-channel case. In the encoder, the sinusoidal part of the monophonic audio signal is encoded by randomly sampling its time-domain representation, and then quantizing the random samples using scalar quantization. The inverse procedure is then followed in the decoder.
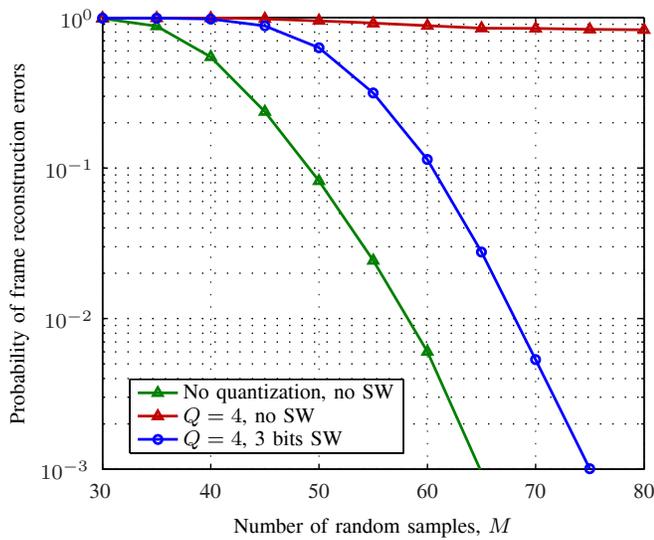

Fig. 2. P_{FRE} *vs* M for a simple example with $N = 256$, $K = 10$ and three cases: no quantization and no spectral whitening, $Q = 4$ bits quantization and no spectral whitening, and $Q = 4$ bits quantization and 3 bits for spectral whitening.

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

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, $\{\hat{F}'_l, \hat{\alpha}'_l, \hat{\theta}_l\}$. 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, $\{\hat{F}_l, \hat{\alpha}_l, \hat{\theta}_l\}$, which are passed to the sinusoid model resynthesis block. If the block has not been correctly reconstructed, then the current frame is either retransmitted or interpolated, as discussed in Section IV-F.

In the remainder of this section, we discuss the important components of our proposed system in more detail. All the data used in the simulations discussed in this section are the audio signals that are used in the listening tests of Section VI. The audio signals were all sampled at 22kHz using a 20ms window with 50% overlapping between frames. Unless otherwise stated, the parameters used were an $N = 2048$-point FFT from which we computed a $K = 25$ sinusoid component x_l. The total number of frames of audio data in the simulations is about 5000.

As discussed in the previous section, the probability of FRE (P_{FRE}) is a key performance figure in our system. Fig. 2 presents the simulated P_{FRE} *vs* M for a simple example with $N = 256$ and $K = 10$. Let us just consider the "No quantization, no SW" curve; it is clear that P_{FRE} decreases as

M increases, due to more information being available at the decoder. Of course, a higher M requires a higher bitrate, and thus we chose to set

$$P_{\text{FRE}} \approx 10^{-2} \tag{9}$$

as a design constraint. The effects of this choice are discussed further in Sections IV-F and VI.

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 $\hat{x}_l$. We must then select the K largest components of $\hat{x}_l$ 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 [35], 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 P_{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 (and thus the bitrate) 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. In turn, lowering the frequency resolution in order to retain a low bitrate will affect the resulting quality of the modeled audio signal, since the restriction in the number of bins clearly limits the frequency estimation during the sinusoidal parameter selection. 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
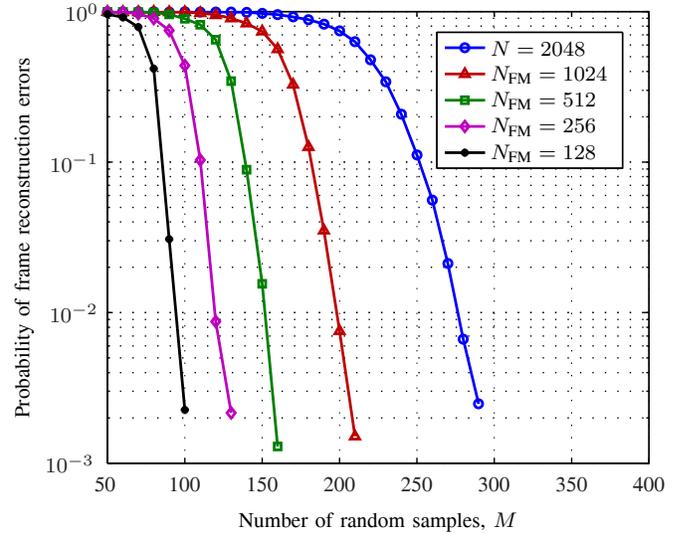

Fig. 4. P_{FRE} *vs* M for various values of frequency mapping, 4-bits of quantization of the random samples, and 3 bits for spectral whitening.

the number of bins after frequency mapping N_{FM} is given by

$$N_{\text{FM}} = \frac{N}{C_{\text{FM}}}. \tag{10}$$

We choose C_{FM} to be a power of two so that resulting N_{FM} will always be a power of two, suitable for use in an FFT.

Thus we create $\boldsymbol{F}'_l$, a mapped version of $\boldsymbol{F}_l$, whose components are calculated as

$$F'_{l,k} = \left\lfloor \frac{F_{l,k}}{C_{\text{FM}}} \right\rfloor, \tag{11}$$

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

$$\dot{F}_{l,k} = F_{l,k} \mod C_{\text{FM}}. \tag{12}$$

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

$$F_{l,k} = C_{\text{FM}} F'_{l,k} + \dot{F}_{l,k}. \tag{13}$$

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 (11) produces distinct values of $F'_{l,k}$, $k = 1, \ldots, K$.

The decrease in the required M for a given P_{FRE} for various values of C_{FM} is clearly illustrated in Fig. 4. Throughout this work, we have only presented results for which a significant number—greater than 95%—of the frames can be mapped by the given values of C_{FM}. The frames that can not be mapped to the highest value of C_{FM} are mapped to the next-highest possible value to ensure minimum impact on bitrate. The final bitrates achieved due to frequency mapping are discussed in Section IV-E.

C. Quantization and entropy coding of random samples

We employed a uniform scalar quantizer to quantize the M random samples to Q bits per sample.

The effects of quantizing the random samples cannot be analyzed in a straight-forward manner [36]–[38]. In our system, the quantization is done in the time domain, but its effects are more readily observed in the frequency domain as changes in the amplitudes and phases of the sinusoidal components. Compounding the difficulties of analysis is the fact that these changes are only visible after passing through a highly non-linear CS reconstruction algorithm. The final complication is that we are dealing with audio signals and thus psychoacoustic effects should be taken into account.

As [36]–[38] indicate, the optimal quantization of CS measurements is a very complicated problem, and one that has yet to be solved. Moreover, current work in the area suggests that quantizing the CS measurements will always have inferior performance to directly quantizing the sparse signal. We do not dispute that here, and indeed, this is not strictly what we are doing. Through the use of frequency mapping—to reduce the dimension of the sparse vector—and spectral whitening—to reduce the dynamic range of the amplitudes—we are simplifying the job that the CS reconstruction has to do. Of course, these two processes also have the side benefit of improving the quality of the reconstructed signals. All this is only possible because we know the sparse signal in advance.

For a purely objective discussion, we now consider the segmental SNR of the reconstructed audio signals. This is the mean SNR of the all the reconstructed frames, and is affected by the number of random samples M, the number of bits used for quantization Q, and the reconstruction algorithm used. The number of bits used for SW also affects the reconstructed SNR, however this dramatically affects the final bitrate, so we chose to use the minimum number of bits for SW that allows us to satisfy (9) with the lowest overall bitrate. Note that this varies with Q, and the chosen values are presented in Table I.

Fig. 5. Mean segmental SNR of the reconstructed audio frames vs the number of random samples M, for varying number of bits used for quantization Q, and $N_{\mathrm{FM}} = 128$.
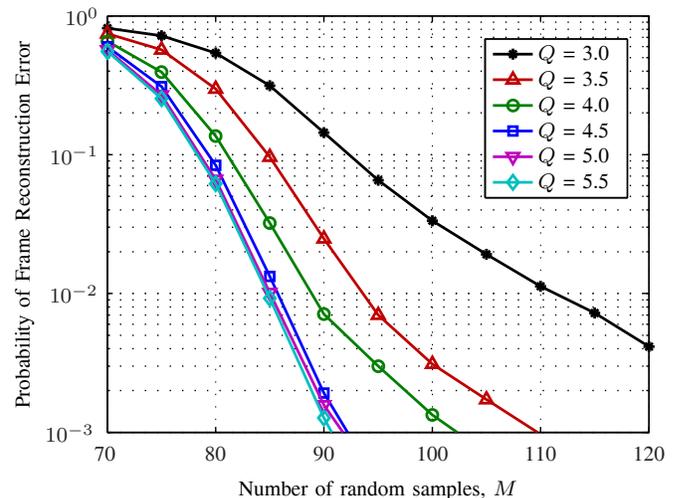
Fig. 6. P_{FRE} vs M for varying number of bits used for quantization Q, and $N_{\mathrm{FM}} = 128$.

TABLE I
NUMBER OF BITS PER SINUSOID USED FOR SPECTRAL WHITENING, FOR DIFFERENT VALUES OF Q.

Q	SW bits
3	5
3.5	4
$\geqslant 4$	3

Fig. 5 presents the mean segmental SNR of the reconstructed audio frames as M and Q are varied. The error is measured among the sinusoidal component and its quantized version in the time-domain. The SNR increases as M increases, but nowhere near as significantly as when Q is increased. We also calculated the amplitude-only SNR (ignoring the phase), which produced slightly higher, but otherwise very similar results to Fig. 5. The non-integer values of Q are achieved by a simple sharing of bits. For example, for $Q = 3.5$, 7 bits are shared over two consecutive random samples. It must also be noted that the curves in Fig. 5 were simulated using the
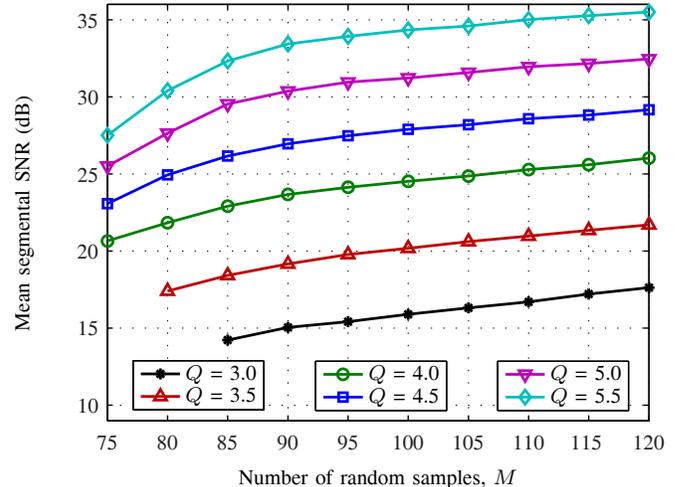
error-free mode of Section IV-F3, ensuring that there were no FREs. In fact, the choice of Q affects the P_{FRE}, and thus the choice of M that can be used, as illustrated in Fig. 6. It is for this reason that the curves for $Q = 3$ and 3.5 begin at $M = 85$ and 80 respectively in Fig. 6, as the P_{FRE} is too high at lower values of M to enable error-free reconstruction in these cases.

It is clear from Fig. 6 that increasing Q reduces the M required for a given P_{FRE}, but that there is no reduction once $Q \geqslant 4.5$. Thus one can conclude from Fig. 5 and 6 that Q is more important than M in terms of improving reconstructed SNR. However each increase in Q dramatically increases the final bitrate, so that great care must be taken in the choice of both Q and M. This is discussed further is Section IV-E and subjective results on the effects of quantization on audio quality are presented in the listening tests of Section VI.

TABLE II
COMPRESSION ACHIEVED AFTER ENTROPY CODING FOR ALL AUDIO
SIGNALS. (Q: CODEWORD LENGTH, $\bar{Q}$: AVERAGE CODEWORD LENGTH
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 sp.	3	2.60	13.2%	4	3.64	9.0%	5	4.68	6.5%
Male sp.	3	2.60	13.4%	4	3.63	9.2%	5	4.66	6.8%
Average	3	2.61	12.9%	4	3.65	8.7%	5	4.68	6.3%

To further reduce the number of bits required for each quantization value, an entropy coding scheme [39] 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

$$\bar{l} = \sum_{i=1}^{2^b} p_i l_i, \tag{14}$$

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 II 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 a training procedure— something we prefer to avoid—so this is presented as an optional enhancement. Also, the derived values correspond to the best-case scenario that the training and testing signals are of similar nature, since training was performed using the same recordings (but different segments) as the ones that were encoded.

D. "Super" Reconstruction Algorithm

In order to ensure we obtained the lowest-possible bitrate, we analyzed the performance of a variety of reconstruction algorithms. The one chose to use in our system was the smoothed ℓ_0 norm—described in [34]—as it gave the best performance and was very efficient.

The fact that our decoder can tell when an FRE has occurred, allows us to propose the use of a new reconstruction paradigm. In a sense, it can be considered as a "super" algorithm as it makes use of other reconstruction algorithms. Let us term these other reconstruction algorithms as "sub-algorithms". The super algorithm proceeds as follows: for each
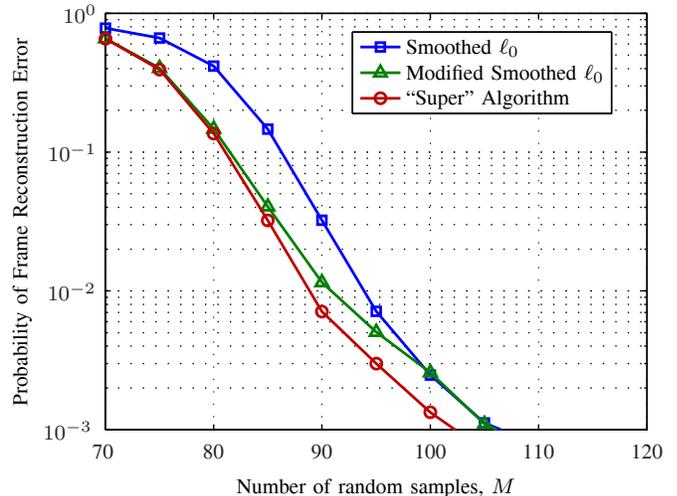

Fig. 7. P_{FRE} vs M for different reconstruction algorithms, with 4 bits for quantization of the random samples, 3 bits for spectral whitening, and $N_{\text{FM}} = 128$.

frame, we run sub-algorithm number 1 and check the CRC, if an FRE has occurred we run sub-algorithm number 2 and check the CRC, if an FRE has occurred, we run sub-algorithm number 3, and so on until the frame has been successfully reconstructed.

Thus for the super algorithm to fail *all of the sub-algorithms must fail*. At worst, the performance of the super algorithm will be that of the best sub-algorithm, but frequently it will be better, as different sub-algorithms generally fail for different frames. It must be noted that super algorithm will incur additional complexity in the decoder due to the fact that multiple sub-algorithms may need to be run, but in practice this effect could be minimised by running the best performing sub-algorithm first.

This is nicely illustrated in Fig. 7 where we consider the performance of a super algorithm based on two sub-algorithms: the smoothed ℓ_0 algorithm, and a modified smoothed ℓ_0 algorithm. The modified smoothed ℓ_0 algorithm was obtained by using a different smoothing algorithm. The super algorithm clearly provides the best possible performance, particularly when the P_{FRE} for the two sub-algorithms are less than 10^{-2}.

TABLE III
PARAMETERS THAT ACHIEVE A PROBABILITY OF FRE OF
APPROXIMATELY 10^{-2} FOR VARIOUS VALUES OF N_{FM}

N_{FM}	Q	M	raw bitrate	overhead CRC	overhead FM	overhead SW	final bitrate	per sinusoid
2048	4	275	1100	8	0	75	1183	47.3
1024	4	195	780	8	25	75	888	35.5
512	4	155	620	8	61	75	764	30.6
256	4	115	460	8	96	75	639	25.6
128	4	88	352	8	140	75	575	23.0

TABLE IV
PARAMETERS THAT ACHIEVE A PROBABILITY OF FRE OF
APPROXIMATELY 10^{-2} FOR VARIOUS VALUES OF Q

N_{FM}	Q	M	raw bitrate	overhead CRC	overhead FM	overhead SW	final bitrate	per sinusoid
128	3	109	327	8	140	125	600	24.0
128	3.5	94	329	8	140	100	577	23.1
128	4	88	352	8	140	75	575	23.0
128	4.5	84	378	8	140	75	601	24.0
128	5	83	415	8	140	75	638	25.5
128	5.5	83	456	8	140	75	680	27.2

E. Bitrates

Table III presents the bitrates achievable for a P_{FRE} of approximately 10^{-2} with $Q = 4$. 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 overhead incurred from spectral whitening and frequency mapping is more than accounted for by significant reductions in M, resulting in overall lower bitrates.

Table IV shows the effect of Q on the bitrates achievable for a P_{FRE} of approximately 10^{-2}. Of interest here is that the bitrates achievable for $Q = 3$ and 4.5 are the same, similarly for $Q = 3$ and 4.5. Fig. 5 suggests that the bitrate with the higher value of Q will sound better, and this is discussed further in Section VI.

In Fig. 8 we present the P_{FRE} *vs* M for the individual signals used in our simulations and listening tests with for the case with $N_{\text{FM}} = 128$, $Q = 4$ and 3-bit spectral whitening. It is clear that for a P_{FRE} of 10^{-2} the M does not vary much, say from 87 to 96. Equivalently, with a fixed M of 88, the P_{FRE} only varies from about 0.007 to 0.04. 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 VI for more details on the signals used.

It should also be noted from Table II that the above bitrates can be reduced by about 1 bit per sinusoid if entropy coding is used, although this will require training, something we are trying to avoid.

F. Operating Modes

To address the fact that we can only specify a probability of reconstruction, we propose three 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 P_{FRE} is kept low enough this increase should be tolerable. For instance, we aim for $P_{\text{FRE}} \approx 10^{-2}$ in this work, which would incur an increase in bit-rate of approximately one percent.

2) Interpolation: In most sinusoidal coding applications, retransmission is not a viable option. For applications where retransmission is undesirable—or indeed impossible—the interpolation mode may be used. In this mode, lost frames are
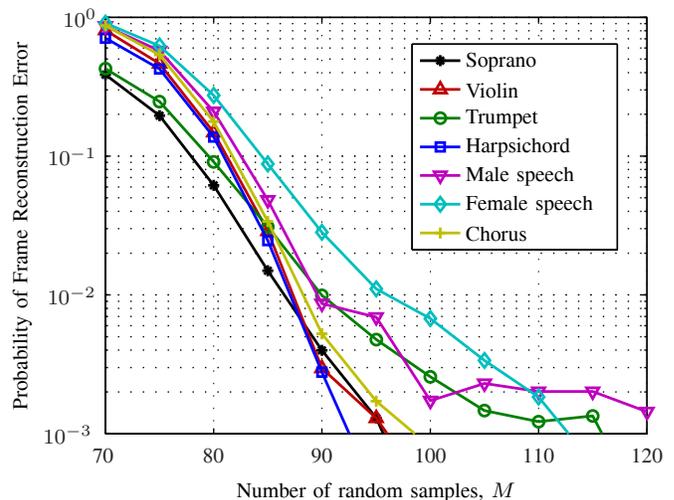

Fig. 8. P_{FRE} *vs* M for individual signals, with 4 bits for quantization of the random samples, 3 bits for spectral whitening, and $N_{\text{FM}} = 128$.

reconstructed using the same interpolation method as used in the regular synthesis of McAulay and Quatieri [1], *i.e.* using 1) linear amplitude interpolation and 2) cubic phase interpolation between matched sinusoids of different frames. Non-matched sinusoids are either "born" or "die" away (interpolated from and to zero amplitude). In case of a lost frame, a sufficient number of samples are interpolated between the previous and successive good frame. 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 VI.

3) Error-free: The final mode is one in which reconstruction is guaranteed, *i.e.* no FREs will occur. This is done by reconstructing the frame *in the encoder* using the random samples selected. If the frame is successfully reconstructed, then these random samples are transmitted. If not, then a new set of random samples are selected and reconstruction is attempted again. This process is repeated until a set of random samples that permit successful reconstruction is found.

In addition to eliminating the need for retransmission or interpolation, the error-free mode allows for a lower bit-rate, by allowing the system to operate with many less random samples than the other two modes. Of course, the reconstruction in the encoder increases the complexity of the encoder, and so we do not explore this mode further in this work.

G. Complexity

As an indication of complexity, our MATLAB CS implementation could run in real time, as the encoder and decoder take 600 μs and 4 ms per frame, respectively (only the CS encoding and decoding part, excluding the sinusoidal analysis and synthesis). With 20 ms frames and 10 ms frame advance (for 50% overlap), these equate to 6% and 40% of the available processing time. This benchmarking was performed on a Microsoft Windows XP PC with 2GB of RAM running at 2GHz.

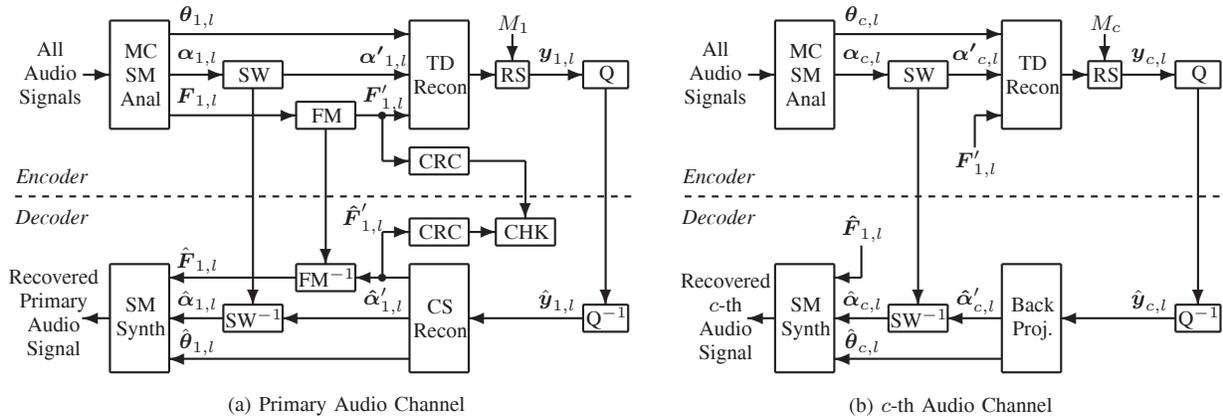(a) Primary Audio Channel   (b) c-th Audio Channel

Fig. 9. A block diagram of the proposed system for the case of multi-channel audio. In the encoder, the sinusoidal part of each audio channel is encoded by randomly sampling its time-domain representation, and then quantizing the random samples using scalar quantization. The single-channel system is fully applied to one of the audio channels (primary channel) in (a), while for the remaining channels (b) only a subset of the quantization process is needed. In the decoder, the sinusoidal part is reconstructed from the random samples of the multiple channels.

V. MULTI-CHANNEL SYSTEM DESIGN

A block diagram of our proposed system for the case of multi-channel audio is depicted in Fig. 9. The primary channel is encoded in a manner very similiar to that described in the previous section, and is shown in Fig. 9(a), which corresponds to the block diagram of Fig. 1. The only differences are that the psychoacoustic sinusoidal modeling block now takes all C audio channels as an input, as discussed in Section II-B, and that many quantities now have an extra subscript specifying which of the C channels they belong to.

For the encoding and decoding of the remaining channels (excluding the primary channel) we propose performing the following procedure. Due to the fact that the sinusoidal models for all the channels share the same frequency indicies,

$$\boldsymbol{F}_{c,l} = \boldsymbol{F}_{1,l} \qquad c = 2, 3, \ldots C, \tag{15}$$

$$\boldsymbol{F}'_{c,l} = \boldsymbol{F}'_{1,l} \qquad c = 2, 3, \ldots C, \tag{16}$$

$$\hat{\boldsymbol{F}}'_{c,l} = \hat{\boldsymbol{F}}'_{1,l} \qquad c = 2, 3, \ldots C, \tag{17}$$

$$\hat{\boldsymbol{F}}_{c,l} = \hat{\boldsymbol{F}}_{1,l} \qquad c = 2, 3, \ldots C, \tag{18}$$

the encoding and decoding for the other $(C - 1)$ channels can be a lot simpler, as shown in Fig. 9(b). In particular, the compressed sensing reconstruction collapses to a back-projection. Let us write the measurement process of (6) as

$$\boldsymbol{y}_{c,l} = \boldsymbol{\Phi}_{c,l} \boldsymbol{\Psi} \boldsymbol{X}_{c,l} \tag{19}$$

where $\boldsymbol{y}_{c,l}$, $\boldsymbol{\Phi}_{c,l}$ and $\boldsymbol{X}_{c,l}$ denote the c-th channel versions of $\boldsymbol{y}_l$, $\boldsymbol{\Phi}_l$ and $\boldsymbol{X}_l$, respectively.

Now let $\boldsymbol{\Psi}_F$ be the columns of $\boldsymbol{\Psi}$ chosen using $\boldsymbol{F}_{1,l}$, and $\boldsymbol{X}^F_{c,l}$ be the rows of $\boldsymbol{X}_{c,l}$ chosen using $\boldsymbol{F}_{1,l}$. We can then write (19) as

$$\boldsymbol{y}_{c,l} = \boldsymbol{\Phi}_{c,l} \boldsymbol{\Psi}_F \boldsymbol{X}^F_{c,l}. \tag{20}$$

Which can then be rewritten as

$$\boldsymbol{X}^F_{c,l} = \left(\boldsymbol{\Phi}_{c,l} \boldsymbol{\Psi}_F\right)^{\dagger} \boldsymbol{y}_{c,l} \tag{21}$$

where $(\mathbf{B})^{\dagger}$ denotes the Moore-Penrose pseudo-inverse of a matrix $\mathbf{B}$, defined as $(\mathbf{B})^{\dagger} = \left(\mathbf{B}^H \mathbf{B}\right)^{-1} \mathbf{B}^H$ with $\mathbf{B}^H$ denoting the conjugate transpose of $\mathbf{B}$.

Thus (21) gives a way of recovering $\boldsymbol{X}^F_{c,l}$ from $\boldsymbol{\Phi}_{c,l}$, $\boldsymbol{F}_{1,l}$ and $\boldsymbol{y}_{c,l}$. However, the decoder only has $\hat{\boldsymbol{\Phi}}_{c,l}$, $\hat{\boldsymbol{F}}_{1,l}$ and $\hat{\boldsymbol{y}}_{c,l}$, which is $\boldsymbol{y}_{c,l}$ after it has been through quantization and dequantization.

So the decoder for the other $(C - 1)$ channels can recover an estimate of $\boldsymbol{X}^F_{c,l}$ using

$$\hat{\boldsymbol{X}}^{\hat{F}}_{c,l} = \left(\boldsymbol{\Phi}_{c,l} \boldsymbol{\Psi}_{\hat{F}}\right)^{\dagger} \hat{\boldsymbol{y}}_{c,l}. \tag{22}$$

One particular advantage of the recovery of (22) is that it is only the primary ($c = 1$) audio channel that determines whether or not an FRE occurs. The number of random samples required for the other $(C - 1)$ channels can be significantly less than that for the primary channel, and thus $M_c < M_1$, $c = 2, 3, \ldots C$. Decreasing M_c only decreases the signal-to-distortion ratio, which the ear is much less sensitive to than the effect of FREs. This of course means that the primary channel will be the best quality channel, with the other $(C - 1)$ being of lower quality. This may or may not be desired, and if not, sum and differences of the channels may be sent instead of the actual channels. This still allows the recovery of the original channels, but with a more even quality.

VI. LISTENING TESTS

In this section, we examine the performance of our proposed system, with respect to the resulting audio quality. Listening tests were performed in a quiet office space using high-quality headphones (Sennheiser HD650), with the participation of ten volunteers (authors not included). Monophonic audio files were used for the single-channel algorithm, and stereophonic files were used for the multi-channel algorithm. Two types of tests were performed. The first test was based on the ITU-R BS.1116 [40] 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). Low-pass filtered (with 3.5 kHz cutoff) versions of the original audio recordings were used as anchor signals. This test is referred to as the quality rating test in the following paragraphs. 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.

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 [10], [12]. The signals were downsampled to 22 kHz, so that the stochastic component affects the resulting quality to a lesser degree compared to the 44.1 kHz case. This is because the stochastic component is particularly dominant in higher frequencies—thus its effect would be more evident at the 44.1 kHz than the 22 kHz sampling rate—and the focus of the paper is on the sinusoidal rather than the stochastic component.

The sinusoidal analysis/synthesis window was 20 ms, using $K = 25$ sinusoids per frame, with 50% overlapping. Our proposed algorithm used an $N = 2048$-point FFT and $N_{FM} = 128$. The values of Q and M are different for the single-channel and multi-channel cases. Due to the use of the psychoacoustic model, for some frames less than 25 sinusoids were selected. Thus, the bitrate results in the following paragraphs are given in terms of bits per sinusoid. The parameters were chosen so as to always obtain $P_{FRE} \approx 10^{-2}$.

A. Single-channel Case

For the single-channel case, 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 [41] of OGI's CSLU. The audio signals used in the tests can all be found at our website[1].

The results of the single-channel listening test are given in Fig. 10. The *retransmission* mode was employed for this quality test. The proposed method was implemented operating at the following rates

- $Q = 3.0$ bits per sample, $M = 109$ samples, resulting in 24 bits per sinusoid,
- $Q = 4.0$ bits per sample, $M = 88$ samples, resulting in 23 bits per sinusoid,
- $Q = 4.5$ bits per sample, $M = 84$ samples, resulting in 24 bits per sinusoid.

These values correspond to the first, third, and fourth rows of Table IV, where the additional details of the implementation parameters (overhead) can be found. We note that the resulting bitrates (23 or 24 bits per sinusoid) could be further reduced by employing the entropy coding of Section IV-C. The proposed method was compared to a state-of-the-art method, namely that of [10], operating at the same rate of 24 bits per sinusoid (denoted as V&K 24 in the figure), and also at a lower bitrate of 21 bits per sinusoid (denoted as V&K 21 in the figure).

From the results of Fig. 10 it can be clearly seen that the proposed method achieves similar quality to state-of-the-art sinusoidal coding methods for the same bitrate. More
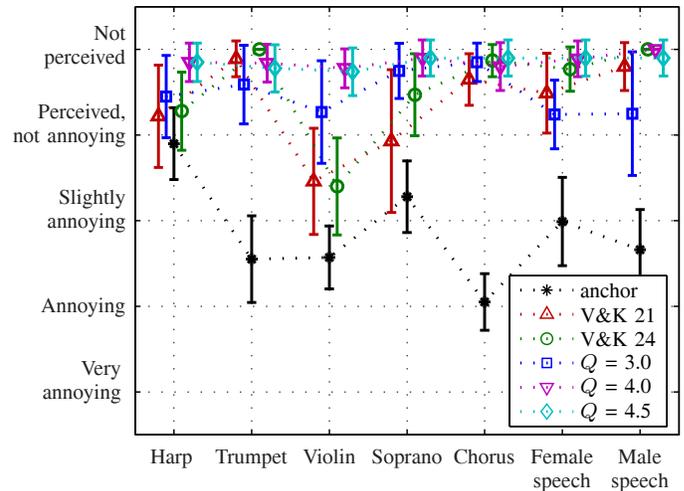
[1]http://www.ics.forth.gr/~mouchtar/cs4sm/

Fig. 10. Results of quality rating listening tests. V&K refers to the method of [10] with the given target entropy (21 or 24 bits per sinusoid). $Q = 3.0$ and $Q = 4.5$ correspond to the proposed method with 24 bits per sinusoid (different choices of implementation parameters), while $Q = 4.0$ corresponds to 23 bits per sinusoid.

specifically, comparing the $Q = 4.0$ and $Q = 4.5$ cases with the similar bitrate of V&K 24, we can see that the proposed method results in comparable audio quality, and in fact outperforms V&K 24 for some signals. 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 [10] at that same rate and not lower. On the other hand, the $Q = 3.0$ case, is clearly seen to result in lower audio quality for most audio signals, although it also operates at the same bitrate of 24 bits per sinusoid as $Q = 4.5$.

The $Q = 3.0$ case was included in the test so as to verify our expectation from Fig. 5, that using more bits per 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. This fact, which was indicated by the SNR results in that figure, was verified in this listening test. Given that the results for the proposed method are similar for the $Q = 4.0$ and $Q = 4.5$ choices, we use $Q = 4.0$ for the remaining listening test results in the single- and multi-channel cases since it provides a slightly lower bitrate than $Q = 4.5$.

At this point, it must be noted that more recent methods may perform better than the method of [10] (such as [12]) and thus could achieve similar performance to our method using less bits per sinusoid, depending though on the particular signal used. As explained, our general objective is to examine the applicability of the CS framework to sinusoidal audio coding—and thus we do not claim here superiority of the proposed approach compared to current state-of-the-art methods.

The preference tests in this section were performed in order to examine the quality that can be achieved in the *interpolation* mode of operation for our system, compared to the *retransmission* mode. In the interpolation case, delays for
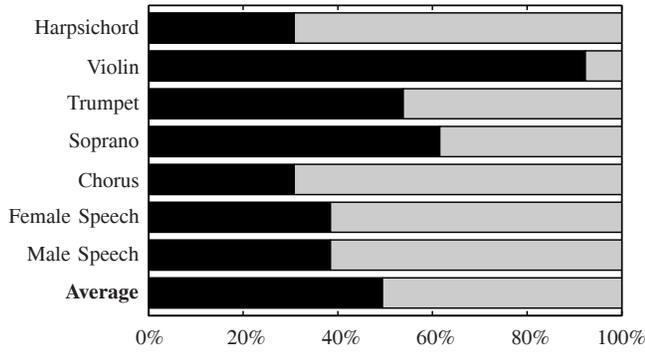
Fig. 11. Results of the preference listening tests for retransmitted signals (black) over 1% FRE interpolated signals (grey).
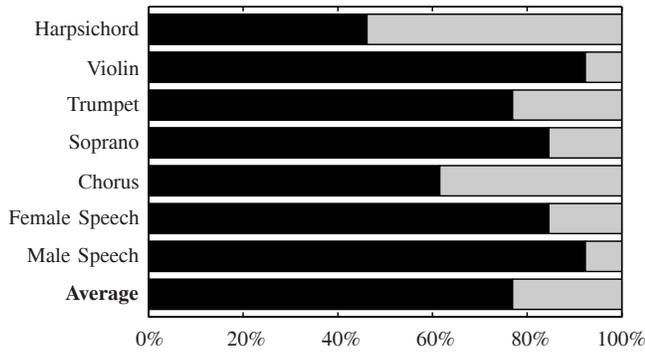

Fig. 12. Results of the preference listening tests for retransmitted signals (black) over 10% FRE interpolated signals (grey).

retransmission of the frames in error are avoided, and thus it is of practical value to investigate its performance. Thus, we evaluated the interpolation case for a 1% probability of frame error, and for a 10% probability of error. The audio signals were encoded using $Q = 4.0$ resulting in 23 bits per sinusoid. The results of the comparison of the 1% case to the retransmission case (*i.e.*, no FREs) are given in Fig. 11. It can be seen from this test 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 in general 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. 12, the preference is clear towards the retransmission operation mode, which indicates that the interpolation method can no longer conceal the frame errors to an acceptable degree.

B. Multi-channel Case

While the proposed multi-channel coding scheme operates in principle regardless of the number of channels, and in fact becomes more beneficial in terms of total bitrate when the number of channels is high, it was convenient for us to perform listening tests using headphones and stereo signals, following the ITU-R BS.1116 methodology as previously. The following six stereo signals were used: male and female speech, male and female chorus, trumpet and violin, a cappella singing, jazz
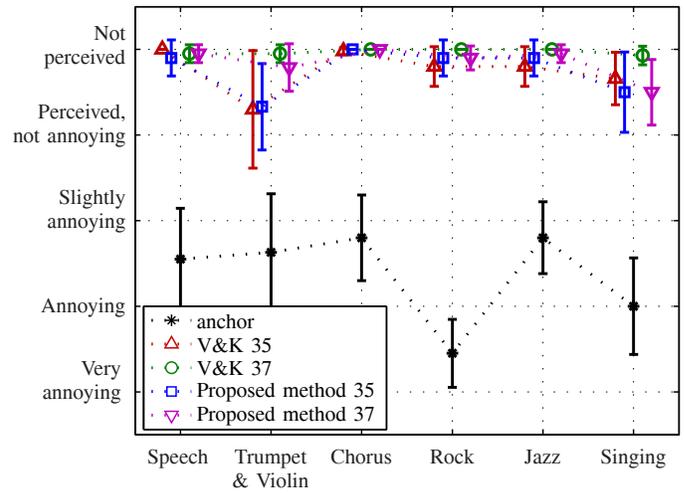

Fig. 13. Results of quality rating tests for various stereo signals. "V&K" refers to the method of [10]. The bits per sinusoid values (35 or 37) are given as the sum for the two encoded audio channels.

and rock. The latter three types of recordings were obtained from popular music CDs, while the remaining audio signals are the same used in the previous section. The test signals used in this section can be found at our aforementioned website (multi-channel case).

The results of this test are given in Fig. 13, where the vertical lines indicate the 95% confidence limits. Again, V&K refers to the method of [10]. Our proposed method was implemented using 4-bit quantization of the random samples, and the parameters given in Table V. In order to balance the overall quality of the reconstructed stereo signals, the primary and secondary channels were alternated every other frame, so that the left and right channels were each encoded as the primary channel 50% of the time. This was found to produce signals with a slightly higher quality than using a sum and difference method.

The primary channel had 3 bits per sinusoid of spectral whitening (SW), 5.6 bits per sinusoid for frequency mapping (FM), and required 88 random samples to achieve a P_{FRE} of approximately 10^{-2}, giving a required bit rate of 23.0 bits per sinusoid. The secondary channel had 3 or 2 bits per sinusoid of spectral whitening and no bits were required for frequency mapping. The number of random samples for the secondary channel were 69 or 62, resulting in 14.04 and 11.92 bits per sinusoid respectively.

In Fig. 13, the value of 35 bits per sinusoid for the proposed method corresponds to the parameters leading to 23 for the primary and 11.92 for the secondary channel in Table V, while the 37 total bits per sinusoid correspond to the 14.04 bits per sinusoid value in the table for the secondary channel. Note that if additional channels were to be encoded, they would be considered as secondary channels. We used the *retransmission* mode to ensure no FREs occurred.

The signals generated by our method were compared again to the method of [10], denoted as "V&K", operating at the rates of 23 & 12 (total 35), and 24 & 13 (total 37) bits per sinusoid, for the left and right channels respectively. These

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING

13

TABLE V
PARAMETERS USED TO ENCODE THE SIGNALS USED IN THE
MULTI-CHANNEL LISTENING TESTS, AND THEIR ASSOCIATED PER-FRAME
BITRATES.

chan	Q	M	raw bitrate	overhead			final bitrate	per sine
				CRC	FM	SW		
1	4.0	88	352	8	140	75	575	23.00
2	4.0	69	276	0	0	75	351	14.04
2	4.0	62	248	0	0	50	298	11.92

values were selected so as to achieve the best possible sound quality for this method, considering that in the single-channel case 21 bits per sinusoid were found to provide inadequate quality. For this method, both channels were coded separately, and no frequency information was sent for the right channel as it were the same as that used in the left channel. Thus, the fact that our multi-channel sinusoidal model uses the same frequency indices for all channels, which was exploited in our multi-channel CS coding method as explained, is also exploited for the method of [10], so that the comparison provided is fair.

It can be seen in Fig. 13 that our proposed method achieves a similar quality to that of [10] for the total rate of 35 bits per sinusoid. This is also true for the 37 bits per sinusoid case, although in this latter case one can observe a very slight but consistent preference towards the method of [10]. Overall, the results obtained for the multi-channel case can be considered as quite encouraging, given that our interest in this paper is to provide a study as to whether CS can be applied to audio coding, which is in principle verified by our results. Further work certainly remains given that the decoding complexity for the proposed method remains significantly higher than state-of-art sinusoidal coding methods.

VII. CONCLUSIONS

The methodology of compressed sensing was introduced to the long examined problem of encoding the sinusoidal parameters of an audio signal. Current state-of-the-art methods directly encode these parameters based on the high-rate theory, minimizing the distortion using a jointly optimal scalar quantizer for these parameters. In contrast, based on CS theory we propose using a small subset of the samples of the sinusoidal part, given that this part is sparse in the frequency domain. These samples are randomly chosen and subsequently quantized using a scalar quantizer. The complexity in the encoder is similar to state-of-the-art methods, while it is higher in the decoder. The methodology was examined both for monophonic as well as multi-channel audio signals, and it was found that comparable performance in terms of audio quality of current sinusoidal coding methods can be achieved with the proposed methodology. It is noted that our interest is not to claim at this point superiority of the proposed method in terms of bitrate or complexity compared to state-of-the-art methods for sinusoidal audio coding. Our interest is to examine the applicability of the CS theory to this area, as a first step towards examining the challenging problem of applying CS to audio coding in a more general context. Given the important property of CS to combine sampling and compression in a single step and move the complexity from

the encoder to the decoder, our final long-term objective is to design an audio coding methodology based on CS which is applicable for the case that the encoder operates on a resource-constrained platform, such as a wireless sensor network. In that case, audio-related applications such as environmental monitoring, remote presence, and obtaining field recordings would be greatly facilitated. At the same time, more near-term but certainly non-trivial issues such as combining the sinusoidal modeling with the random measurement procedure, and including psychoacoustic analysis in the quantization of the random measurements will be examined in our future work.

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 synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14(4), pp. 12–24, Winter 1990.

[3] C. Tzagkarakis, A. Mouchtaris, and P. Tsakalides, "A multichannel sinusoidal model applied to spot microphone signals for immersive audio," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 8, pp. 1483–1497, Nov. 2009.

[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] S. N. Levine and J. O. Smith III, "A switched parametric and transform audio coder," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, pp. 985–988, 1999.

[6] T. S. Verma and T. H. Y. Meng, "A 6 kbps to 85 kbps scalable audio coder," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, pp. 877–880, 2000.

[7] H. Purnhagen and N. Meine, "HILN the MPEG-4 parametric audio coding tools," *Proc. IEEE Int. Symp. Circuits, and Systems (ISCAS)*, pp. 201–204, 2000.

[8] A. C. den Brinker, E. G. P. Schuijers, and A. W. Oomen, "Parametric coding for high-quality audio," 112th *Audio Engineering Society (AES) Convention*, p. 5554, 2002.

[9] R. Vafin and W. B. Kleijn, "Entropy-constrained polar quantization and its application to audio coding," *IEEE Trans. Speech and Audio Process.*, vol. 13(2), pp. 220–232, 2005.

[10] 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.

[11] R. Vafin and W. B. Kleijn, "Jointly optimal quantization of parameters in sinusoidal audio coding," *Proc. IEEE Workshop on Applications of Signal Process. to Audio and Acoust. (WASPAA)*, pp. 247–250, 2005.

[12] P. Korten, J. Jensen, and R. Heusdens, "High resolution spherical quantization of sinusoidal parameters," *IEEE Trans. Speech and Audio Process.*, vol. 13(3), pp. 966–981, 2007.

[13] R. Heusdens, J. Jensen, P. Korten, and R. Vafin, "Rate-distortion optimal high-resolution differential quantisation for sinusoidal coding of audio and speech," *Proc. IEEE Workshop on Applications of Signal Process. to Audio and Acoust. (WASPAA)*, pp. 243–246, 2005.

[14] M. H. Larsen, M. G. Christensen, and S. H. Jensen, "Variable dimension trellis-coded quantization of sinusoidal parameters," *IEEE Signal Proc. Lett.*, vol. 15, pp. 17–20, 2008.

[15] 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.

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

14

IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING

[17] 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.

[18] A. Griffin, T. Hirvonen, A. Mouchtaris, and P. Tsakalides, "Encoding the sinusoidal model of an audio signal using compressed sensing," in *Proc. IEEE Int. Conf. on Multimedia and Expo (ICME'09), New York, NY, USA*, June 2009.

[19] T. Sreenivas and W. B. Kleijn, "Compressive sensing for sparsely excited speech signals," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, pp. 4125–4128, 2009.

[20] M. Goodwin, "Residual modeling in music analysis-synthesis," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, vol. 2, pp. 1005–1008, May 1996.

[21] R. C. Hendriks, R. Heusdens, and J. Jensen, "Perceptual linear predictive noise modelling for sinusoid-plus-noise audio coding," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, vol. 4, pp. 189–192, May 2004.

[22] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP J. Appl. Signal Process.*, vol. 2005, no. 1, pp. 1292–1304, January 2005.

[23] M. Goodwin, "Multichannel matching pursuit and applications to spatial audio coding," in *Asilomar Conf. on Signals, Systems, and Computers*, October 2006.

[24] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, and R. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Processing Magazine*, vol. 25(2), pp. 83–91, March 2008.

[25] V. Cevher, M. Duarte, and R. Baraniuk, "Distributed target localization via spatial sparsity," in *Proc. European Signal Processing Conf. (EU-SIPCO), Lausanne, Switzerland*, August 2008.

[26] A. Griffin and P. Tsakalides, "Compressed sensing of audio signals using multiple sensors," in *Proc. European Signal Processing Conf. (EUSIPCO), Lausanne, Switzerland*, August 2008.

[27] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magnetic Resonance in Medicine*, vol. 58(6), pp. 1182–1195, December 2007.

[28] D. Baron, M. B. Wakin, M. F. Duarte, S. Sarvotham, and R. G. Baraniuk, "Distributed compressed sensing," 2005, preprint.

[29] M. Kowalski and B. Torrésani, "Random models for sparse signals expansion on unions of bases with application to audio signals," *IEEE Trans. on Signal Process.*, vol. 56(8), pp. 3468–3481, August 2008.

[30] R. Grosse, R. Raina, H. Kwong, and A. Y. Ng, "Shift-invariant sparse coding for audio classification," in *Proc. 23rd Conference on Uncertainty in Artificial Intelligence*, 2007.

[31] M. D. Plumbley, S. A. Abdallah, T. Blumensath, M. G. Jafari, A. Nesbit, E. Vincent, and B. Wang, "Musical audio analysis using sparse representations," in *Proc. 17th COMPSTAT 2006 (17th Symposium of IASC-ERS), Rome, Italy*, August 2006.

[32] 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.

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

[34] 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.

[35] O. Cappe, J. Laroche, and E. Moulines, "Regularized estimation of cepstrum envelope from discrete frequency points," in *IEEE Workshop on Applications of Signal Process. to Audio and Acoust. (WASPAA)*, October 1995.

[36] V. Goyal and A. F. S. Rangan, "Compressive sampling and lossy compression," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 48–56, March 2008.

[37] J. Sun and V. Goyal, "Optimal quantization of random measurements in compressed sensing," in *IEEE Int. Symp. on Inf. Theory (ISIT)*, March 2009.

[38] P. Boufounos and R. Baraniuk, "Quantization of sparse representations," Rice ECE Department, Tech. Rep. TREE 0701, 2007.

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

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

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

PLACE PHOTO HERE

Anthony Griffin received his PhD in Electrical & Electronic Engineering from the University of Canterbury in Christchurch, New Zealand in 2000. He then spent three years programming DSPs for 4RF, a Wellington-based company selling digital microwave radios. He subsequently moved to Industrial Research Limited—also based in Wellington—focusing on signal processing for audio signals and wireless communications. In 2007, he joined the Institute of Computer Science, Foundation for Research and Technology-Hellas (FORTH-ICS), Heraklion, Greece as a Marie Curie Fellow, where he is working on compressed sensing for audio signals and wireless sensor networks. He also occasionally teaches a postgraduate course in Applied DSP at the University of Crete.

PLACE PHOTO HERE

Toni Hirvonen received M.Sc and Ph.D degrees from the Helsinki University of Technology (TKK, currenty Aalto University), Finland in 2002 and 2007, respectively. He has worked as a postdoctoral researcher at TKK and as a Marie Curie Fellow at FORTH-ICS, Greece. Since 2010, he is with Dolby Laboratories, Sweden. His main research areas are signal processing and psychophysics.

PLACE PHOTO HERE

Christos Tzagkarakis received the B.Sc. and M.Sc. degrees in computer science from the University of Crete, Heraklion, Greece, in 2005, and 2007, respectively. He is currently pursuing the Ph.D. degree at the Computer Science Department of the University of Crete, in the area of audio signal processing. His research interests include signal processing for immersive audio environments, audio modeling and coding, music information retrieval, and speaker recognition.

PLACE PHOTO HERE

Athanasios Mouchtaris (S'02-M'04) received the Diploma degree in electrical engineering from Aristotle University of Thessaloniki, Greece, in 1997 and the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, CA, USA in 1999 and 2003 respectively.

From 2003 to 2004 he was a Postdoctoral Researcher in the Electrical and Systems Engineering Department of the University of Pennsylvania, Philadelphia. From 2004 to 2007 he was a Postdoctoral Researcher in the Institute of Computer Science of the Foundation for Research and Technology Hellas (FORTH-ICS), Heraklion, Crete, and a Visiting Professor in the Computer Science Department of the University of Crete. Since 2007 he has been an Assistant Professor in the Computer Science Department of the University of Crete, and an Associated Researcher in FORTH-ICS. His research interests include signal processing for immersive audio environments, spatial audio rendering, multichannel audio modeling, speech synthesis with emphasis on voice conversion, and speech enhancement.

Dr. Mouchtaris is a member of the IEEE and Eta Kappa Nu.

PLACE
PHOTO
HERE

Panagiotis Tsakalides (M95) received the Diploma in electrical engineering from the Aristotle University of Thessaloniki, Greece, in 1990, and the Ph.D. degree in electrical engineering from the University of Southern California (USC), Los Angeles, in 1995.

He is a Professor of Computer Science at the University of Crete, and a Researcher with the Institute of Computer Science, Foundation for Research and Technology-Hellas (FORTH-ICS), Greece. From 2004 to 2006, he served as the Department Chairman. From 1999 to 2002, he was with the Department of Electrical Engineering, University of Patras, Patras, Greece. From 1996 to 1998, he was a Research Assistant Professor with the Signal and Image Processing Institute, USC, and he consulted for the U.S. Navy and Air Force. His research interests lie in the field of statistical signal processing with emphasis in non-Gaussian estimation and detection theory, and applications in sensor networks, audio, imaging, and multimedia systems. He has coauthored over 100 technical publications in these areas, including 25 journal papers. He is the PI of the 1.3 M euros FP7 MC-IAPP "CS-ORION" project (2010-2014) conducting research on compressed sensing for remote imaging in aerial and terrestrial surveillance.

Dr. Tsakalides is a member of the ERCIM Network of Innovation/Technology and Knowledge Transfer Experts (I-Board), and of IEEE.