MULTICHANNEL AUDIO CODING USING SINUSOIDAL MODELLING AND COMPRESSED SENSING

Anthony Griffin, Toni Hirvonen, Athanasios Mouchtaris and Panagiotis Tsakalides

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

ABSTRACT

This paper explores the potential of applying compressed sensing (CS) to multichannel audio coding. In this context, we consider how sinusoidally-modelled multichannel audio signals might be encoded using compressed sensing, as opposed to directly encoding the sinusoidal parameters (amplitude, frequency, phase) as current state-of-the-art methods do. The results, obtained from listening tests using 80 sinusoids per frame with no residual noise signal, show that such a model can achieve equal or better performance to that of the 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, this can be considered as an important step towards applying the CS framework to audio coding applications.

1. INTRODUCTION

Multichannel audio allows the recreation of rich sound scenes, through the transmission of multiple audio channels. As the number of channels used can be many times that of a 2-channel stereo signal (8 channels for 7.1 multichannel audio, for example), the bitrate requirements can be considerable.

The sinusoidal model [1, 2] represents an audio signal using a small number of time-varying sinusoids. The model allows for a compact representation of the original signal and for efficient encoding and quantization. Extending the sinusoidal model to multichannel audio applications has also been proposed (*e.g.* [3]). State-of-the-art methods for encoding and compressing the parameters of the model (amplitudes, frequencies, phases) are based on directly encoding these parameters [4–7].

Compressed sensing (CS) [8, 9] 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. Thus, CS combines compression and sampling of a signal into one low-complexity step. An important restriction is that 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. This prohibits the application of CS to a large class of signals, including audio signals, which are of interest in this paper.

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 and Smart Environments".

Thus, we apply the CS framework to the sinusoidally-modelled part of an audio signal. This is a sparse signal, since by definition it contains only a small number of frequency components for each time segment. In our previous work [10, 11], we introduced a novel method of encoding the parameters of a monophonic sinusoidal model using CS. Here, we extend that work by deriving a system which applies CS to the case of sinusoidally-modelled multichannel audio. Listening tests demonstrate that the proposed system can achieve equal or better performance compared to current state-of-the-art sinusoidal coding methods. Given the advantages of the CS methodology in terms of computational complexity, applicability to sensor networks and local signal compression, as well as inherent encryption, 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 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 multichannel recording. It is noted that low-complexity 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. At the same time, the paper proposes a novel psychoacoustic modelling analysis for the selection of sinusoidal components in a multichannel audio recording.

2. SINUSOIDAL MODEL

The sinusoidal model was initially applied to the analysis/synthesis of speech [1]. A 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

$$s(t) = \sum_{k=1}^{K} \alpha_k(t) \cos[\beta_k(t)], \qquad (1)$$

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.

Each component 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 sine wave. Practically, after the sinusoidal parameters are estimated, a residual noise component is computed by subtracting the sinusoidal component from the original signal.

Current state-of-the-art methods for sinusoidal mod-

elling employ perceptual matching pursuit algorithms to determine the model parameters of each frame. To perform multichannel sinusoidal analysis, we have extended the method presented in [12] to include state-of-the-art psychoacoustic analysis [13]. At each iteration, the algorithm picks a sinusoidal component frequency that is optimal for both channels, as well as channel-specific amplitudes and phases. This choice minimizes the perceptual distortion measure

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

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.

This paper utilizes the improved masking model detailed in [13]. An important question is what masking model is suitable for multichannel 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. However, this effect in parametric coding is not so well established.

We performed preliminary experiments using: firstly, separate masking analysis, *i.e.* individual $A_{i,c}(\omega)$ based on the masker of channel c for each signal separately (see (2)), secondly, using the masker of the sum signal of all channel signals to obtain $A_i(\omega)$ for all c, and thirdly, power summation of the other signals' attenuated maskers to the masker of channel c according to

$$A_{i,c}(\omega) = 1/[M_{i,c}(\omega) + \sum_{\substack{k \\ k\neq c}} w_k M_{i,k}(\omega)], \qquad (3)$$

where $M_{i,c}(\omega)$ is the masker energy of the c-th channel after the i-th iteration, 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 multichannel sinusoidal model resulted in sparser modelled signals, increasing the effectiveness of our compressed sensing encoding.

3. COMPRESSED SENSING

In the compressed sensing methodology, a signal which is sparse in some basis can be represented using much fewer samples than the Nyquist rate would suggest. Given that a sinusoidally-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.

3.1 Measurements

Let $\boldsymbol{x}_l$ be the N samples of the sinusoidal component in the sinusoidal model in the l-th frame. It is clear that $\boldsymbol{x}_l$ 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 $\boldsymbol{x}_l$ 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

$$\begin{aligned} \boldsymbol{y}_l \;&=\; \boldsymbol{\Phi}_l \boldsymbol{x}_l \\ &=\; \boldsymbol{\Phi}_l \boldsymbol{\Psi} \boldsymbol{X}_l, \end{aligned} \qquad (4)$$

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 [14] for further discussion of random sampling. In this case, $\boldsymbol{\Phi}_l$ is formed by randomly selecting M rows of the $N \times N$ identity matrix.

3.2 Reconstruction

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

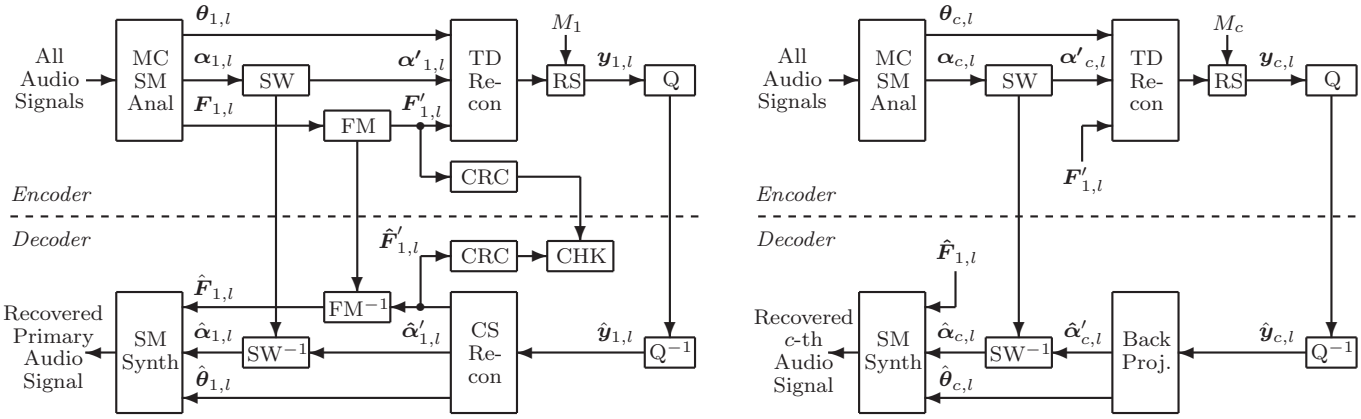
$$\hat{\boldsymbol{X}}_l \;=\; \arg\min \|\boldsymbol{X}_l\|_p \qquad \text{s.t.} \qquad \boldsymbol{y}_l = \boldsymbol{\Phi}_l \boldsymbol{\Psi} \boldsymbol{X}_l, \quad (5)$$

with $p = 1$ will recover $\boldsymbol{X}_l$ with high probability if enough measurements are taken. The ℓ_p norm is defined as $\|\boldsymbol{a}\|_p = (\sum_i |a_i|^p)^{\frac{1}{p}}$. It has recently been shown in [15] that $p < 1$ outperforms the $p = 1$ case, and it is this method that we use for reconstruction in this paper.

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

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

Let $\boldsymbol{F}_l$ be the *positive* FFT frequency indices in x_l, whose components $F_{l,k}$ are related to the frequencies



(a) Primary Audio Channel

(b) c-th Audio Channel

Figure 1: A block diagram of the proposed system. 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. In the decoder, the sinusoidal part is reconstructed from the random samples.

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 detection scheme 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. Previous work has shown that a suitable target for the probability of FRE (P_{FRE}) is less than 10^{-2} [11]. Obviously, the retransmission of a frame in error requires more bandwidth compared to the interpolation option, but if the probability of FREs is kept low enough this increase should be tolerable. For instance, $P_{\text{FRE}} \leq 10^{-2}$ would incur an increase in bit-rate of approximately one percent.

In this work, the retransmission scheme is used. We note that in addition to the retransmission and the interpolation options, a third alternative is the *error-free* operation. 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 CRC and 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. Clearly, the reconstruction in the encoder dramatically increases the complexity of the encoder, and so we do not explore this mode further in this work.

4. SYSTEM DESIGN

A block diagram of our proposed system is depicted in Fig. 1. The first channel is encoded in a manner very similar to that of [10], and is shown in Fig. 1(a). The C-channel audio signal is first passed through a psychoacoustic sinusoidal modelling block to obtain the sinusoidal parameters $\{\boldsymbol{F}_{1,l}, \boldsymbol{\alpha}_{1,l}, \boldsymbol{\theta}_{1,l}\}$ for the l-th frame of the primary channel. These then go through what can be thought of as a "pre-conditioning" phase where the amplitudes are whitened (SW) and the frequencies remapped (FM). The interested reader is referred to [10] for more details. The modified sinusoidal parameters $\{\boldsymbol{F}'_{1,l}, \boldsymbol{\alpha}'_{1,l}, \boldsymbol{\theta}_{1,l}\}$ are then reconstructed into a time domain signal, from which M_1 samples are randomly selected (RS). These random samples are then quantized to Q bits by a uniform scalar quantizer (Q), 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 (Q^{-1}), and passed to the compressed sensing reconstruction algorithm, which outputs an estimate of the modified sinusoidal parameters, $\{\hat{\boldsymbol{F}}'_{1,l}, \hat{\boldsymbol{\alpha}}'_{1,l}, \hat{\boldsymbol{\theta}}_{1,l}\}$. If the CRC detector (CHK) determines that the block has been correctly reconstructed, the effects of the spectral whitening and frequency mapping are removed—(SW^{-1}) and (FM^{-1}), respectively—to obtain an estimate of the original sinusoid parameters, $\{\hat{\boldsymbol{F}}_{1,l}, \hat{\boldsymbol{\alpha}}_{1,l}, \hat{\boldsymbol{\theta}}_{1,l}\}$, 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.

Due to the fact that the sinusoidal models for all the channels share the same frequency indices,

$$\boldsymbol{F}_{c,l} = \boldsymbol{F}_{1,l} \qquad c = 2, 3, \ldots C, \tag{6}$$

$$\boldsymbol{F}'_{c,l} = \boldsymbol{F}'_{1,l} \qquad c = 2, 3, \ldots C, \tag{7}$$

$$\hat{\boldsymbol{F}}'_{c,l} = \hat{\boldsymbol{F}}'_{1,l} \qquad c = 2, 3, \ldots C, \tag{8}$$

$$\hat{\boldsymbol{F}}_{c,l} = \hat{\boldsymbol{F}}_{1,l} \qquad c = 2, 3, \ldots C, \tag{9}$$

the encoding and decoding for the other $(C-1)$ channels can be a lot simpler, as shown in Fig. 1(b). In particular, the compressed sensing reconstruction collapses to a back-projection. Let us write the measurement process of (4) as

$$\boldsymbol{y}_{c,l} = \boldsymbol{\Phi}_{c,l}\boldsymbol{\Psi}\boldsymbol{X}_{c,l} \tag{10}$$

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.

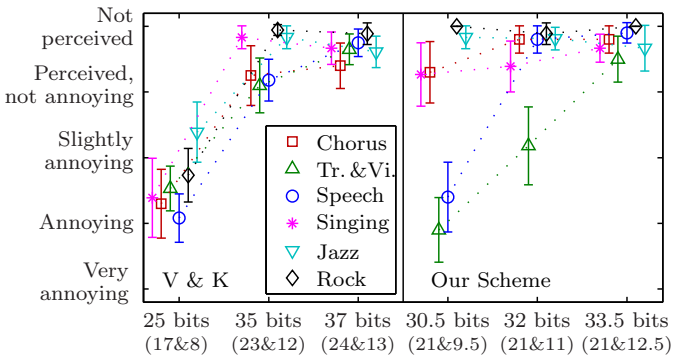


Figure 2: Results of quality rating tests for various stereo signals. "V & K" refers to the method of [6]. ("Tr.&Vi." denotes a signal containing trumpet and violin.) The bits per frame per sinusoid are given for each of the two transmitted audio channels.

Let $\mathbf{\Psi}_F$ be the columns of $\mathbf{\Psi}$ chosen using $\boldsymbol{F}_{1,l}$, and $\boldsymbol{X}_{c,l}^F$ be the rows of $\boldsymbol{X}_{c,l}$ chosen using $\boldsymbol{F}_{1,l}$. We can write (10) as

$$\boldsymbol{y}_{c,l} \;=\; \mathbf{\Phi}_{c,l}\mathbf{\Psi}_F\boldsymbol{X}_{c,l}^F. \tag{11}$$

This can then be rewritten as

$$\boldsymbol{X}_{c,l}^F \;=\; \left(\mathbf{\Phi}_{c,l}\mathbf{\Psi}_F\right)^\dagger \boldsymbol{y}_{c,l} \tag{12}$$

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 (12) gives a way of recovering $\boldsymbol{X}_{c,l}^F$ from $\mathbf{\Phi}_{c,l}$, $\boldsymbol{F}_{1,l}$ and $\boldsymbol{y}_{c,l}$. However, the decoder only has $\mathbf{\Phi}_{c,l}$, $\hat{\boldsymbol{F}}_{1,l}$ and $\hat{\boldsymbol{y}}_{c,l}$, which is $\boldsymbol{y}_{c,l}$ after it has been through quantization and de-quantization. So the decoder for the other $(C-1)$ channels can recover an estimate of $\boldsymbol{X}_{c,l}^F$ using

$$\hat{\boldsymbol{X}}_{c,l}^{\hat{F}} \;=\; \left(\mathbf{\Phi}_{c,l}\mathbf{\Psi}_{\hat{F}}\right)^\dagger \hat{\boldsymbol{y}}_{c,l}. \tag{13}$$

which is much less complex than (5).

One particular advantage of the recovery of (13) 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 allows the recovery of the original channels with a more even quality.

5. LISTENING TESTS

While the proposed multichannel 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 ITU-R BS.1116 [16]. Ten volunteers participated, and the tests took place in a quiet office room. The following six stereo signals were used: male and female speech, male and female chorus, trumpet and violin, a cappella singing, jazz and rock. For the former three stereo signals, the speech recordings were obtained from the VOICES corpus [17] of OGI's CSLU, the chorus signals were provided by Prof. Kyriakakis of the University of Southern California, and the individual instrument recordings were obtained from the EBU SQAM disc. The latter three types of recordings were obtained from popular music CDs. The test signals can be found at [1].

Table 1: Parameters used to encode the signals used in the listening tests, and their associated per-frame bitrates.

chan	M	raw bitrate	overhead CRC	overhead FM	overhead SW	final bitrate	per sine
1	240	960	8	406	320	1694	21.2
2	210	840	0	0	160	1000	12.5
2	180	720	0	0	160	880	11.0
2	150	600	0	0	160	760	9.5

The sinusoidal model analysis was performed using $K = 80$ sinusoid components per frame and an $N = 2048$-point FFT. All the audio signals were sampled at 22 kHz with a 20 ms window and 50% overlapping between frames. Using $K = 80$ provided a high-enough quality that the residual signals were not required.

The results of this test are given in Fig. 2, where the vertical lines indicate the 95% confidence limits. Our proposed method was implemented using 4-bit quantization of the random samples and the parameters given in Table 1. The primary channel was the sum of the left and the right channels, and the secondary channel their difference. The primary channel had 4 bits per sinusoid of spectral whitening (SW) and approximately 5 bits per sinusoid for frequency mapping (FM), and required 240 random samples to achieve a P_{FRE} of less than 10^{-2}, giving a required bit rate of 21.2 bits per sinusoid. The secondary channel had 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 $\{150, 180, 210\}$, giving $\{9.5, 11.0, 12.5\}$ bits per sinusoid respectively.

In Fig. 2, the notation *e.g.* 21 & 9.5 bits in the x-axis, corresponds to using 21 bits for the primary channel and 9.5 bits for the secondary channel *per sinusoid*, while 30.5 is the total number of bits per sinusoid used (the summation of all channels). Note that for each additional audio channel in this example, 9.5 bits per sinusoid would be required. We used the *retransmission* mode to ensure no FREs occurred.

The signals generated by our method were compared to a popular sinusoidal coding method, namely that of [6], denoted as "V&K", operating at the rates of 17 & 8, 23 & 12, and 24 & 13 bits per sinusoid, for the left and right channels respectively. Both channels are coded separately, and no frequency information is sent for the right channel as it is the same as that used in the left channel. Thus, the fact that our multichannel

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

sinusoidal model uses the same frequency indices for all channels, which was exploited in our multichannel CS coding method as explained, is also exploited for the method of [6], so that the comparison provided is fair. In this case though, the left and right channels (and not their sum and difference) are encoded. As previously stated, the notation *e.g.* 17 & 8 in Fig. 2 corresponds to using 17 bits per sinusoid for the left channel (used as primary), and 8 bits for the right channel (used as secondary).

The signal with 17 & 8 bits per sinusoid was used as an anchor signal, and it is clear that the listeners could distinguish the reduction in bitrate and thus quality. It can be clearly seen in Fig. 2 that our proposed method achieves a similar quality to that of [6] for the slightly lower bitrate, 33.5 vs 37 bits per sinusoid for the stereo signal. These rates were chosen for comparison as they achieve a consistent quality for all signals. Since at this rate the proposed method performs slightly better than [6] for some signals and slightly worse for others, we claim that their performance is comparable. Of interest is also the lower rate of 32 bits for the proposed method compared to the 35 bits of [6], where it can be seen that, with the exception of the trumpet/violin signal, the proposed method performs very well, and more consistently compared to [6]. We are confident that these gains extend to the multichannel case (more than 2 channels). More generally, our interest in this paper is to provide a study as to whether CS can be applied to audio coding, and in this sense the results in this paper are quite encouraging.

6. CONCLUSIONS

We have presented a new method for encoding a multichannel signal that has been modelled using the sinusoidal model, making use of compressed sensing (CS). The complexity of the secondary channels is significantly lower than that of the primary channel. Through listening tests, we have shown that our method can achieve equal or better performance to that of other state-of-the-art sinusoidal coding methods. This can be considered an important result towards the final objective of being able to apply the CS framework to audio signals, where the challenge is that of addressing the sparsity requirement. Application of CS to audio coding, even for low-bitrate applications such as those examined here (*i.e.* parametric modelling) can lead to novel systems for analog-to-digital conversion of audio signals and allow for local coding of audio signals (*e.g.* using a sensor network for recording and compressing the audio information in large outdoor spaces).

7. ACKNOWLEDGMENTS

The authors would like to thank the listening test volunteers.

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] 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] 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.

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

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

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

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

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

[10] 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.

[11] 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 Engineering (ICME'09), New York, NY, USA*, June 2009.

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

[13] 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.

[14] 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.

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

[16] ITU-R Recommendation BS.1116, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," 1997.

[17] A. Kain, *High Resolution Voice Transformation*, Ph.D. thesis, OGI School of Science and Engineering at Oregon Health and Science University, October 2001.