A Cross-Layer Coding for Scalable ECG streaming

Mohammed Alloulah
alloulah@outlook.com

Mark Dawkins
mtdawkins@ieee.org

Alison Burdett
alison@ieee.org

Toumaz
15 Olympic Avenue, Building 3, Milton Park
Oxford, OX14 4SA, UK

ABSTRACT

Mobile electrocardiogram (ECG) streaming in body area networks (BANs) is challenging owing to an inherently inconsistent wireless channel, which generally cannot be assumed wide-sense-stationary. Common conventional ECG compression is *entropy*-based and thus is fundamentally at odds with a BAN channel plagued with variabilities. That is, if the wireless signal experiences a deep fade regime, excessive errors, user contention, and RF interference could all combine so as to result in an interruption in ECG streaming until channel quality recovers. To mitigate against this hard limit on channel quality (i.e. the cliff effect), this paper proposes a *linear* ECG coding method whereby proneness to mis-reception due to channel errors, contention, and/or interference is traded for a soft, proportional degradation in signal definition. As such, the likelihood of ECG streaming interruption in BANs is vastly lessened while also enhancing capacity and relieving wireless medium contention. This improved robustness and scalability in the wireless network is particularly sought after in mission-critical healthcare applications with stringent QoS demands.

CCS Concepts

•**Networks** → **Network design principles; Cross-layer protocols;**

Keywords

Wireless, BAN, ECG, coding

1. INTRODUCTION

Digital healthcare technologies are at the forefront of the first wave of internet of things (IoT) devices. The healthcare demand posed by ageing populations has spurred the development of body-worn "patches" that monitor vital signs and relay them wirelessly for subsequent analysis by medical staff. The overarching aim is to supply healthcare services at a scale not currently attainable under traditional medical practice workflows. Such healthcare "automation"—both in hospitals or at home—will allow for significantly more efficient workflows, whose beneficiary is the society at large e.g. cost reduction, resource optimization, sustainable health system, etc.

A typical wireless patient monitoring system consists of ultra low-power, body-worn patches that collect vital sign measurements (e.g. heart rate, respiration rate, temperature, etc.) and a so-called off-body [15] basestation (CM4 in [27]) configured in a star topology [26]. The basestation forwards these measurements to a server unit which in turn collates medical records and generates alerts for the attention of nurses in a hospital.

An emerging body of evidence documents the clinical and economical benefits of such wireless patient monitoring [8]. However, the slow uptake of this technology in medical arenas—which are conservative and resistive to change by their very nature—has not fostered further research to revisit the system architecture and investigate avenues of potential enhancements. Specifically, medical practice is heavily regulated and standarized, which calls for reliable wireless patient monitoring wherein the scalability and robustness of diagnostic bio-waveform delivery is paramount. This stringent level of reliability is meant to reflect an overall system quality of service (QoS) which we define as: *the probability that an individual measurement data packet will be correctly transferred from patch to basestation with applicable maximum latency* [22, p. 61]. This QoS is also of utmost importance in the short-term if wireless patient monitoring were to overcome the barriers to technology adoption and penetration in medical domains.

In light of the motivation to support stringent, medical-grade QoS, a fundamental question arises: what are the issues encountered in BAN wireless patient monitoring deployments?

Robustness. BAN channels are characterized by large-scale statistics that vary not only over coarse-grained distances but also with changes in body posture, the way devices are mounted/worn, and angular antenna relative orientation [15], violating the wide-sense-stationarity assumption in on-body channel variants [20]. Together, large- and small-scale statistics in BANs give rise to more variable path loss profiles when compared to more traditional urban cellular channels for instance [21, 6]. These channel variations are abrupt and nondeterministic (e.g. a patient turning in bed) and could prove challenging (or impossible) for a system employing a low-power protocol and low-power, low-cost patches which attempt to adapt through transmitter-receiver

MobiCASE 2016, November 30-December 01, Cambridge, Great Britain
Copyright © 2016 EAI
DOI: 10.4108/eai.30-11-2016.2267099

feedback. Moreover, RF interference in industrial, scientific, and medical (ISM) bands worldwide is ever present and is continually morphing with standardization activities, which poses extra difficulties for compliance, design, and reliability.

Scalability. In a surgical ward use case, a number of patches may simultaneously stream multi-lead ECG measurements to one tethered basestation acting as an uplink. The high sampling rate of raw ECG signals as mandated by diagnostic-grade medical standards (cf. [4]) quickly saturates the link budget of co-located patches in low-power, low-throughput protocols such as IEEE 802.15.6 [1]. This is because in mission-critical medical applications, it is necessary and customary in practice to introduce a retransmission link margin over and above the nominal budget in order to guarantee a near 100% QoS (i.e. quasi error-free reception) should a worst case operating condition occur (say as a result of mobility). Such retransmission margin is typically allocated in the link budget *permanently*, which compromises the scalability of network to accommodate more participating users. This mostly idle contingency link margin can be somewhat relaxed using compression techniques, commonly entropy-based [12], which remove redundancy from the raw ECG signal. A byproduct of entropy-based redundancy removal is to fragilize the transmission stream further towards channel errors. That is, a mere unlucky flip of a channel bit at the decoder could totally destroy a portion of the transmitted ECG signal.

Recent advances in mobile video communications have shown that considerable improvements in scalability and robustness can be attained by taking a cross-layer view of the wireless design problem at hand. Specifically, it is demonstrated that a joint approach whereby source compression and channel error protection are combined results in the ability to realize graceful signal degradation that is *linearly* proportional to a given instantaneous channel quality [11, 3]. As such, the network operation is rendered self-regulating in that channel errors are tolerated at the expense of reduced video quality. This approach is in line with information-theoretic findings stating that the separation of source and channel coding is inefficient when the statistics of the channel, such as that of BAN, cannot be predicted [11, 24].

In this paper, we revisit the source-channel separation conventional wisdom in ECG streaming applications. We propose a *linear* transformation with excellent energy compaction properties at least of comparable performance to state-of-the-art *entropy*-based ECG compression techniques. This proposition is largely motivated by the desire to also enhance the reliability of ECG streaming in BANs by boosting the network's robustness and scalability. This is to be achieved capitalizing on the linearity of the proposed operator, as opposed to the nonlinearity of entropy-based compression techniques (for example) wherein there exists a hard limit on the wireless channel quality below which signal reconstruction is not possible (i.e. a cliff effect). The proposed cross-layer coding draws no distinction between the source ECG signal and the wireless channel. It is posited that such coding will pave the way for two graceful wireless ECG streaming variants, to be adapted from mobile video literature [11] & [3] (analogue and digital, respectively):

(a) Analogue-style: Transmitters may place unequal emphasis on groups of dominant, transform-domain coefficients in terms of transmission power part of a global optimization objective. Such optimization will allow the network to tolerate bit errors and packet loss at the receiver without incurring a catastrophic effect on the service; rather, the final QoS would degrade linearly commensurate with the bit errors and packet loss experienced. Clearly, this analogue, power-allocation-based scheme is linear and enhances the network's capacity (i.e. scalability) and robustness (i.e. provisions for and recovery from errors).

(b) Digital-style: Transform-domain coefficient bits may also be grouped according to their impact on the distortion of signal reconstruction, subject to a criterion e.g. mean square error (MSE). The resultant distortion groups (of bits) are then encoded using rateless codes depending on their relative importance so as to produce parity bits to be appended to the original bits, resulting in added channel protection. Such a digital scheme—similar to the above analogue-style scheme—is also linear (i.e. as opposed to entropy coding), enabling gracefully degradable medical signal delivery as a function of a given node's channel quality. Thus, this digital coding scheme also enhances the overall network scalability and robustness. When contrasted with the previous analogue scheme, such digital approach simplifies the RF resources of low-power, low-cost patches required for finer power control in favour of more tractable digital coding. The extra decode overhead, however, is shifted to the basestation side (or server side) which, by virtue of being tethered, enjoys less restrictive resource constraints within the overall network architecture design.

We evaluate the performance of the proposed linear transformation using two standard metrics commonly referred to in ECG literature: compression rate (CR) and percentage root mean square[1] difference (PRD) [14, 12]. A representative set of extensive signal morphologies including normal sinus rhythms and arrhythmias is used to characterize the proposed ECG coding performance. The following results are summarized:

- For the normal sinus rhythm dataset, the CR and PDR averaged around 11x and 3.6%, respectively. CR was well above 11x across just under half of the whole dataset. Similarly, PDR was better than its 3.6% average across half of the dataset.
- For the arrhythmia dataset, the CR and PDR averaged around 7x and 8.5%, respectively. CR fared better than its 7x average across one-third of the dataset. PDR was still better than its 8.5% average across half of the dataset.

Contribution. This paper proposes a novel cross-layer coding for ECG streaming targeted at body-worn wireless vital sign monitoring BAN systems. The cross-layer linear method is shown to perform competitively when compared to state-of-the-art ECG compression techniques, while being a key enabler towards realizing graceful signal delivery in BANs. Graceful ECG streaming is posited to help overcome barriers to uptake in medical arenas by virtue of enhanced network robustness and scalability i.e. QoS.

[1] aka RMS

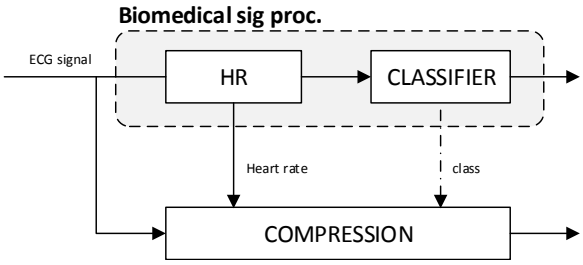


Figure 1: Overall system block diagram: ECG compression interacts with subblocks from the biomedical signal processing datapath; namely, heart rate detection and classifier subsystems.

2. ENCODER

In order to facilitate graceful wireless ECG steaming, the challenge is to cluster linear dominant coefficients that describe the ECG medical signal in a transform domain efficiently. Put differently, the key is to localize transform-domain dominant coefficients such that they can be bit-mapped in an efficient way, while also providing competitive energy-compaction performance. Mobile video systems achieve this using a 3-dimensional (3D) discrete cosine transform (DCT) [11, 3]. In 3D DCT, the idea is to exploit the inherent spatio-temporal correlations in video—in space as a result of smooth 2D pixel transitions and in time as a result of smoothly varying sequences of successive frames. Because of these correlations, 3D DCT tends to produce clustered dominant coefficients whose locations can be described concisely i.e. using little amount of information. *So how can we accomplish the same transform-domain linear dominant coefficient clustering in ECG*? We turn next to tackle this central question.

2.1 Diagonalization-based compression

We make a key observation that successive ECG cycles nominally possess a large degree of self-similarity. That is, if we take a stream of ECG data and rearrange it into a matrix whose rows are successive ECG cycles, we will readily see correlations between these rows.

In a similar vein, the proposed compression relies on casting a block of ECG data into an *approximate circulant* matrix. This casting assumes *a priori* knowledge of the ECG rate. The a priori knowledge will be supplied by a dedicated *QRS complex*[2] detection block, part of a separate biomedical signal processing pipeline. The approximate circulant matrix is subsequently diagonalized by means of the Fourier transform [7]. The resultant 2D spectrum characterizes the approximate circulant matrix's dispersion i.e. deviation from an ideal circulant matrix whose eigenvalues coincide perfectly with the main diagonal. This deviation is manifested as off-diagonals with their spread being indicative of the severity of the data block's time-frequency dispersion (i.e. relative variability).

A high-level system diagram of compression block and dependencies is shown in figure 1. The heart rate (HR) block supplies compression with the periodicity of the ECG waveform in samples. When applicable, the optional classifier block, if present part of the biomedical signal processing pipeline, may indicate to compression that the incoming signal belongs to a class of aperiodic ECG waveforms with no fixed periodicity such that compression can adapt its operation accordingly.

[2]QRS complex is a technical ECG term, see [17]

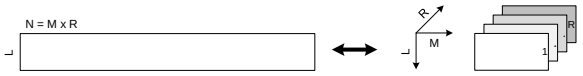


Figure 2: Interleaving original data block into subblocks to decouple circulant matrix size from ECG cycle length.

In order for this formulation to lend itself to the problem domain of compression, there are two problems that need addressing: (1) the restructuring of incoming data such that it can be cast as approximate circulant matrices, and (2) encoding the features of the 2D spectrum in a tractable manner without incurring too much overhead.

2.1.1 Matrix packing – degrees of freedom

The ECG signal has a wide-ranging rate variability i.e. typically 30 to 210 beats per minute (bpm) depending on age and physiological state. Thus depending on the sampling rate of the system, the length N of one cycle of ECG can be long or short. On the other hand, casting a block of data into an approximate circulant matrix necessitates the availability of N^2 samples. This would mean that an approximate circulant matrix will have its size dictated by the ECG rate! Such hard requirement on the approximate circulant matrix size would be problematic for two reasons. Firstly, it couples the latency of compression with ECG cycle length in samples. Secondly, it increases the resources for FFT computations as a quadratic function of N.

To mitigate against this seeming dependency on ECG rate, the data can be interleaved. This is best highlighted pictorially as shown in figure 2, where the ECG period N is rearranged by means of interleaving into R parallel shorter periods each of length M.

Specifically, the raw stream of ECG data making up a block $(L \times N)$ is downsampled by some integer factor R. The process is then repeated with successively increasing phases $\in [0, R-1]$. Hence, we have in effect restructured the original data block $(L \times N)$ into subblocks $(L \times M \times R)$ through interleaving without loss of information. There is no loss of information because we still extract correlations across the newly restructured subblocks. The nature and interpretation of this extraction may vary though depending on how we choose to pack submatrices for diagonalization. That is, after the optional reduction of the ECG rate by R, the resultant subblocks may be independently processed or packed further together.

Using interleaving as an extra degree of freedom, ECG data can be packed into circulant matrices in a variety of ways. For instance, figure 3 shows two colour-coded ways of packing downsampled streams: with high and low latencies. The high-latency option requires that the entire stream be downsampled first before packing can commence. While the low-latency option packs a group of downsampled phases together, allowing a submatrix to be formed much sooner. However, it is informative to emphasize the following. High-latency packing produces a time-frequency coupling representation of data that takes place within one

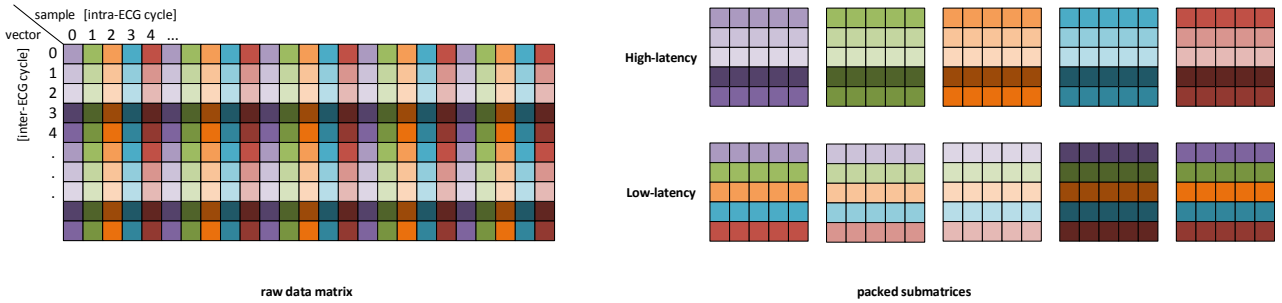


Figure 3: Illustration of high-latency and low-latency packing through interleaving. Samples from different downsampling phases within one given ECG cycle are highlighted using distinct colours (called intra-ECG cycle, horizontally in figure on the left), while samples belonging to different ECG cycles (dubbed inter-ECG cycle, vertically in figure on the left) are denoted with different shades of the same colour.

downsampling phase. Whereas low-latency packing integrates multiple downsampling phases into one time-frequency coupling analysis.

Generally, this *interleaving-based* restructuring can be used to trade off diagonalization efficacy—which necessarily incurs longer FFT sizes[3]—with compression latency. This is somewhat analogous to deep interleavers in communications, which also increase latency. Deep interleavers are at the heart of practical codes that approach the Shannon channel capacity [5].

2.1.2 Thresholding

After diagonalization, the weak coefficients of the 2D spectrum are truncated (i.e. thresholded out). That is, we discard their intangible contribution to ECG definition.

The threshold should be chosen according to a desired balance between the number of retained coefficients in the thresholded matrix (i.e. compression) and target distortion in the reconstructed signal. Further, the threshold can be derived as a function of the power of the original untransformed, time-domain ECG signal, noting the energy conservation of the linear Fourier operator. This threshold function can also be made to depend on additional parameters such as required reconstruction distortion, or extra signal features (e.g. standard deviation if the medical signal is not zero-mean). The model for the derivation of the threshold can be optionally made arbitrarily complex for dynamic reconfiguration if necessitated by the end medical application scenario. Data-driven machine learning techniques such as linear regression can be employed to best fit a multivariate model to empirical data as to produce an optimal threshold function that generalizes well.

2.1.3 Bitmap encoding – metadata

The resultant, thresholded *sparse* matrix will have large areas of zeros. It would be inefficient to tag each spectral component we are going to send over the wireless channel with its 2D coordinates. Instead, run-length encoding (RLE), is applied in the diagonal dimension to group consecutive sequences of switched-on or switched-off coefficients. For example, if in a 100×100 matrix we had the main diagonal containing 19 zeros in the middle and active coefficients either side, then its RLE description will be 1#41_0#19_1#40. Here the first binary digit 1 denotes a run of switched-on coefficients followed by how many after the hash symbol. The next sequence is inactive as indicated by 0 for 19 consecutive coefficients. The last sequence is active and comprises a stretch of 40 coefficients.

RLE is applied in the diagonal dimension for two reasons. Firstly, diagonalization by construction will tend to produce a *banded* matrix around the main diagonal for well-behaved ECG data with moderate time-frequency dispersion. Hence RLE applied to main diagonal and off-diagonals will naturally compress this representation in an optimal fashion. Secondly, since ECG data is strictly real, the 2D spectrum is symmetrical and we can discard half of the coefficients without loss of information—i.e. the symmetry around the diagonal maps nicely to encoding in the diagonal dimension.

To elaborate further, since the elements of the original square matrix are real, the DFT only needs to be performed to produce the elements of the left half of the transformed matrix. This is because the elements of the right half of the transformed matrix can be obtained from those of the left half through the conjugate symmetric property of the 2D Fourier operator. That is, if we define a centered coordinate system of the elements of the transformed matrix $H(x, y)$, then $H^*(x, y) = H(-x, -y)$ (where * denotes conjugation). This symmetry in the transformed matrix is called upon while performing RLE in the diagonal direction.

It will be demonstrated in the evaluation section that such bitmap encoding results in a very small amount of *metadata* overhead that needs to be reliably communicated to the receiver. Upon transmission and after wireless propagation, the receiver will then reconstruct the 2D spectrum from the bitmap of active (i.e. retained) coefficients and a sequence of *complex* coefficients.

In order to tie up all the concepts discussed so far, figure 4 depicts a block diagram representation of the end-to-end system which begins with ECG source coding, followed with channel transmission and reception, and onto final ECG reconstruction. Specifically, the incoming ECG stream is first packed into square matrices using either the native period of the signal or interleaved downsampled phases incorporating high- or low-latency restructuring. Secondly, the square matrices are subjected to the Fourier transform to produce 2D spectra which are in turn thresholded. The resultant sparse matrices are denoted in figure 4 by a hinton diagram with a main diagonal, a subdiagonal, and a superdiagonal. Thirdly,

[3]Without digressing from main discourse, interested reader is referred to information-theoretic concepts in [7] on the asymptotic behaviour of Toeplitz & Circulant matrices in relation to entropy.

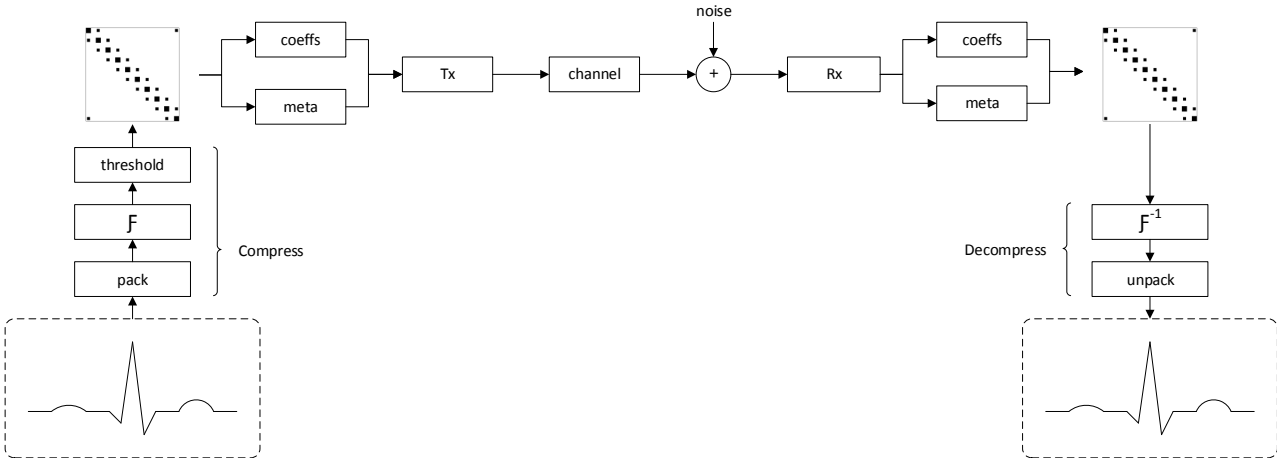


Figure 4: End-to-end system block diagram: joint source-channel coding.

the sparse matrix is next RLE-encoded to produce metadata describing the exact layout of active, retained coefficients. Fourthly, the active coefficients and metadata are transmitted over a noisy channel and subsequently received by the basestation. Upon reception, the basestation utilizes metadata to populate a matrix with the stream of transform-domain coefficients to form a reconstruction matrix. The reconstruction matrix is then subjected to the inverse Fourier operator used at the transmitter in order to obtain the ECG samples. Depending on the employed packing at the transmitter, the basestation may have to de-interleave the reconstructed ECG samples before finally arriving at the intended ECG waveform.

3. SCENARIO

To better motivate the proposed cross-layer coding, we present a concrete scenario in order to underscore the need for compression within the context of a scalable and robust medical use case.

A typical wireless vital signs monitoring system is required for patients in a hospital at risk of deterioration having undergone serious medical procedures e.g. organ transplant. A heavily instrumented intensive care unit costs an order of magnitude more per-bed and per-night compared to say a general ward. Wireless vital signs monitoring with critical ECG streaming provides an early detection mechanism in case of a relapse for subsequent rapid medical intervention. Specifically, the use of ECG streaming, rather than just a snapshot of the heart rate, permits elaborate server-based algorithms to look for additional early warning features; a recent record of ECG waveform preceding an alarm can also greatly aid initial diagnosis.

Hospitals have also a set of requirements which translate into technical specifications for a wireless vital signs monitoring system aspiring for real-world deployment. For instance, in UK national health service (NHS) hospitals, a single network access point for a general ward is needed to minimize infrastructure requirements. Fairly typical in NHS hospitals, up to 32 co-located patients can share a ward and the variable ward size implies a range of 30m can be necessary.

Having established the need for ECG streaming within a patient deterioration monitoring context, we turn next to expose numerically an example of a link budget to justify our proposal. First, we have 12-bit ECG words at 256 sample per second (sps) rate (even 512sps for diagnostic-grade monitoring, see 201.12.4.107.3 in [4]). So we have 3,072 bit per second (bps) per patch, or ∼100kbps at the access point from 32 patients—not including other vital signals or accounting for local storage when the link goes down. Including these extra items results in close to 160kbps for 32 patients. We then include margin for robustness, and assume that we only have a clear channel for 1/3 of the time, so we get close to 500kbps needed at the access point (1Mbps for remote diagnostic-grade monitoring, or 3-lead ECG requiring 2 sensor channels).

Proceeding to exact link budget numbers, in order to obtain 500kbps using the crowded 2.4G band (thus justifying at least the 3× margin above), we need a signal bandwidth in the 1MHz region. At the same time, we require an ultra low-power wireless protocol, which mandates using a relatively simple modulation & coding scheme (i.e. cannot expect to use say 16- or 64-QAM). Thus, bluetooth low energy (BLE) or IEEE 802.15.6 would suit our scenario. BLE can deliver 500kbps throughput at 30m with ∼2dB margin (under line of sight (LOS) assumption). 15.6 RATE3 (QPSK) can do the same at 30m with ∼5dB margin (under line of sight (LOS) assumption). So far, no compression has been assumed for the aforementioned ultra low-power protocols. However, the key point here is: if we want to guarantee a certain level of QoS in a medical mission-critical vital signs monitoring system, then we should aim for much more margin to allow for non-LOS channels. Hence we need more robust modulation (more coding/spreading, lower-order constellations), which means reduced throughput. Therefore, in order to enable our scenario with high reliability in the crowded 2.4G band, we must reduce our required data rate by means of ECG compression. Moreover, the linearity of our proposed cross-layer coding lends itself to realizing degradable ECG waveform delivery whereby the scalability and robustness of the vital signs monitoring system could be vastly improved—allowing for increased number of users (patients), increased sampling rate (remote diagnostic-grade), increased sensor channels (12-lead ECG), or a combination thereof.

4. EVALUATION

The proposed diagonalization-based compression is tested next on a range of ECG signals. The aim is to inform future research & development on diagonalization's ability to energy-compact ECG signals and encode their features across a representative set of morphologies. To this end, two groups of synthetic ECG signals are utilized: (1) normal sinus rhythm dataset consisting of 120 epochs[4], and (2) arrhythmia dataset consisting of 62 epochs.

4.1 Normal sinus rhythm

Method: Patient simulator datasets by Rigel Medical [18] contain 12 normal sinus rhythm conditions which are the permutations of four heart rates (HR) and three amplitudes (AMP). Specifically, $HR \in [30, 70, 120, 210]$ bps and $AMP \in [0.5, 1, 4]$ mV. Each condition has 10 epochs, making the overall epoch count equal to 120.

The characterization results are obtained per condition in groups of 10 epochs. The casting of data into circulant matrices will attempt to use as much of the available data as possible.

QRS complex detection functionality in this investigation was simulated by simply measuring the distance in samples between two consecutive ECG peaks in order to determine periodicity and match the square matrix dimensions to this periodicity. The sampling rate for all patient simulator datasets is 256 Hz.

Results: To provide a fine-grained feel for results, we will be examining the statistics of our ECG metrics of choice through the cumulative distribution function (CDF) and the complementary cumulative distribution function (CCDF). In figure 5a, the CCDF provides a quick pictorial evaluation of how often CR exceeds a certain target value across all ECG data in normal sinus rhythm. For instance, while the average CR is around 11x, it is well above 11x across just under half of the whole dataset. Similarly, the CDF of the PRD is also shown in figure 5b. The CDF elaborates on the fine-grained PRD performances across all ECG data in normal sinus rhythm, which again fared better than its 3.6% average across half of the dataset. Lastly in figure 5c, while metadata is implicitly included in CR figures, we choose to highlight the statistical trends of metadata across the dataset explicitly. Metadata amounts to under 1.6% of the total Tx data (coefficeints + bitmap) at 50 percentile and to just under 4% of total at 95 percentile confidence.

4.2 Arrhythmia

Method: SimMan® patient simulator datasets by Laerdal Medical Limited [13] contain 11 arrhythmia conditions with variable number of epochs. These are summarized in table 1 for convenience. The keen reader is referred to [17] for further medical explanation and background on these conditions.

The restructuring of signals follows the same convention described earlier. In situations where the ECG signal is aperiodic, such as some atrial or ventricular fibrillations, an arbitrary, suboptimal size is chosen for the packed square matrix prior to diagonalization.

The sampling rate for all sets is 250 Hz.

Results: In aperiodic conditions, the packed square matrix dimensions are no longer matched to a fixed period.

[4]an epoch is 30 seconds of ECG samples

Table 1: SimMan® dataset summary

Condition[a]	# of epochs	Description
AF_75	5	atrial fibrillation
AF_90	6	atrial fibrillation
AF_160	8	atrial fibrillation
AFL_150	7	atrial flutter
NSR_STdepressed_145[b]	8	normal sinus rhythm
NSR_STelevated_60[b]	5	normal sinus rhythm
NSR_STelevation_200[b]	6	normal sinus rhythm
VF_NA[c]	1	ventricular fibrillation
VF_NA[c]	2	ventricular fibrillation
VT_180	6	ventricular tachycardia
VT_240	8	ventricular tachycardia

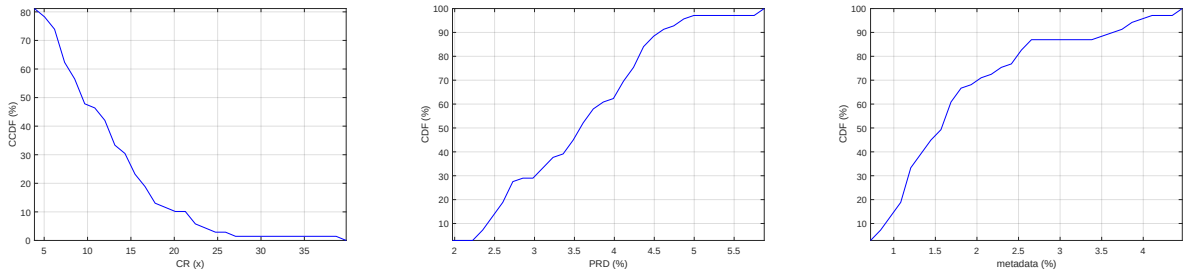
[a] Number following underscore refers to ECG rate
[b] ST denotes a segment in the ECG that can become elevated or depressed in the presence of dangerous heart conditions
[c] NA means that these conditions have no discernible fixed periodicity

This causes diagonalization to fail to localize dominant coefficients in a diagonal fashion in the transform domain. In turn, CR drops markedly compared to the earlier normal sinus rhythm results. Additionally, the cost of diagonal RLE increases too owing to the absence of diagonally extending runs of coefficients in the transform domain. However, it is interesting to observe that a useful compression performance is nonetheless obtained. This is evident from inspecting the CR CCDF of figure 6a whose average is around 7.2x, but with only one-third of the whole dataset being better than its average. The CDF of the PRD in figure 6b across all ECG data in arrhythmia averages at around 8.5% with half of the dataset faring better. The metadata statistics in figure 6c displays a strong trend of increased bitmap cost in arrhythmia as evident from the lazy rise of the CDF; metadata amounts to greater than 3% of the total Tx data at 50 percentile nearly double that of normal sinus rhythm but eventually stabilizes around the 4% mark at 95 percentile confidence.

4.3 Artefact filtering

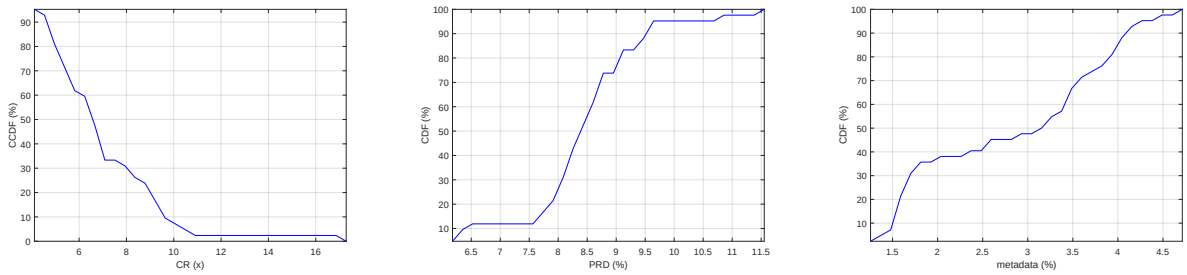
To recap, the proposed linear coding casts the ECG signal as an image to encode its features as spatial frequencies. The symmetry due to diagonalization is why metadata is generated from RLE in the diagonal direction. However, mostly due to artefacts, there are sometimes "baseline" signal features that are manifested on the first row and column (akin to DC frequencies in 1D). Strictly speaking, these are not important in the sense that they do not encode the clinically salient features in ECG. To explain visually, the following example is provided:

It can be seen from figure 7 that the matrix on the left (figure 7a) has an almost full 1st column (1st vertical black stripe) compared to the matrix on the right (figure 7b) where the 1st column and 1st row are trimmed. The conjugate symmetry around the origin described earlier applies to the matrix on the right i.e. the thresholded matrix minus "baseline features" row and column. The 1st row and 1st column are conjugate symmetric w.r.t. themselves (also the diagonal similar to a normal 1D spectrum). That is, if me-



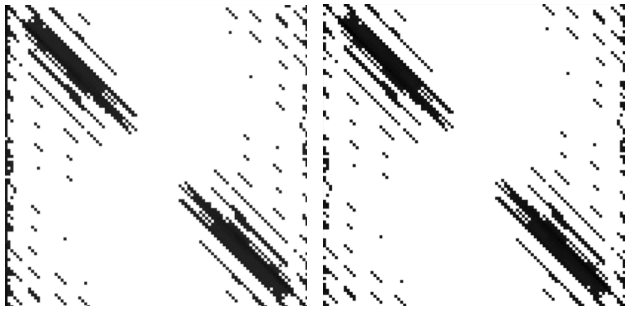
(a) CR (b) PRD (c) metadata

Figure 5: CR CCDF, PRD CDF, and metadata CDF for normal sinus rhythm conditions.



(a) CR (b) PRD (c) metadata

Figure 6: CR CCDF, PRD CDF, and metadata CDF for arrhythmia conditions.



(a) Thresholded matrix. (b) Thresholded matrix without 1st row and 1st column.

Figure 7: Removing baseline features from thresholded transform-domain matrix.

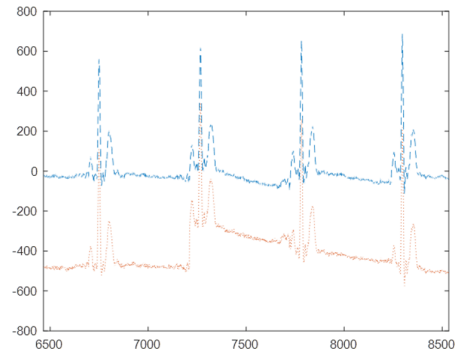


Figure 8: ECG reconstruction snippet with (red) and without (blue) baseline features.

tadata were to include these baseline features, only half of each would have to be encoded by RLE, which is a trivial addition to metadata.

The interesting issue is: do we really need these features? If anything, these baseline features are likely to be generated by ECG data artefacts (e.g. motion artefact). In this example, the artefact is a combination of: (1) a decaying electrical discharge originating from the analogue conditioning circuitry when it is first switched on per epoch and (2) DC offset level. If the ECG signal were to be reconstructed from the thresholded matrix without including the baseline features of 1st column and 1st row (which are responsible for the decaying discharge and DC, respectively), one would get the blue, zero-centered ECG signal in figure 8.

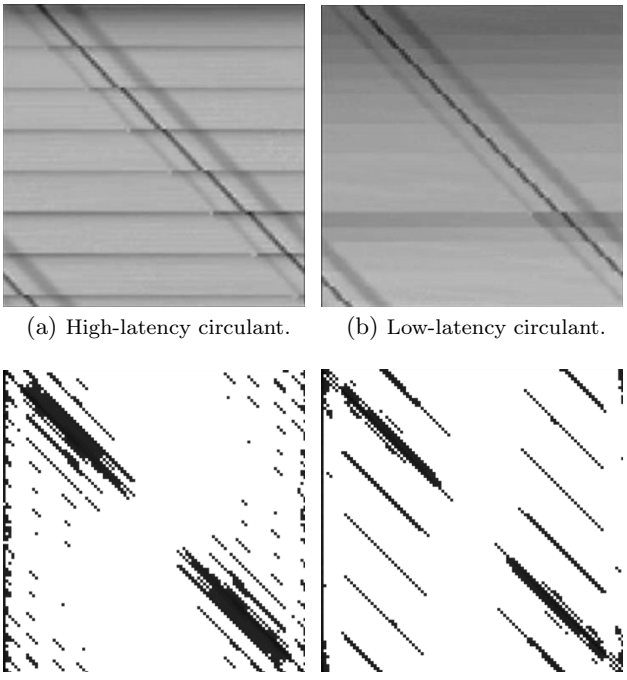
As depicted in figure 8, setting the 1st column and 1st row of the thresholded matrix to zero instead of using their actual coefficients in reconstruction has in effect "filtered" the ECG signal. The clinical features of the QRS complex of the ECG reconstructed signal are not affected; what have been filtered out are the DC level and the decaying electrical discharge.

Communicating these baseline features part of the transform-domain coefficients with the requisite addition to metadata is optional. The preconditioning of ECG signals prior to diagonalization possibly with enhanced sensor interface analogue front-end circuitry may alleviate to a large extent this issue altogether. Ultimately, this is likely to be a biomedical signal processing decision since medical practitioners tend to at times insist on replicating faithfully the raw ECG signal along with its nuances and artefacts.

4.4 Latency reduction

As discussed in section 2.1.1, we next turn to demonstrate the flexibility afforded by the interleaving-based packing approach. Specifically, we give a flavour for the high-latency and low-latency modes of interleaved packing depicted in figure 3 by presenting a concrete numerical example.

In the first three sets of normal sinus rhythm conditions ({1-10}, {11-20}, and {21-30}), the ECG data has a cycle periodicity of 515 which can be downsampled by 5 in order to transform the original stream into 5 substreams with 103 periodicity corresponding to phases 0, 1, 2, 3, and 4. These substreams are then packed individually into 5 approximate circulant matrices. With a sampling rate of 256 Hz, the block latency of such packing is therefore $5 \times 103^2/256 \approx 207$ seconds. Such high latency may or may not be appropriate for a given use case.



(a) High-latency circulant. (b) Low-latency circulant.

(c) High-latency thresholded. (d) Low-latency thresholded.

Figure 9: Matrices: high-latency vs. low-latency.

Instead of downsampling the entire ECG stream and packing the resulting phases per matrix, we can further interleave the downsampled phases into one matrix. Such packing has the advantage of filling a circulant matrix sooner so that diagonalization may commence, reducing compression latency. Specifically, the latency now becomes $103^2/256 \approx 41$ seconds—a factor of 5 reduction compared to earlier. Conceptually, this interleaved packing is also justified on the grounds of effectively extracting "intra-cycle" correlations and encoding these correlations as spatial frequencies. This is a valid assumption in biomedical oversampled ECG waveforms as mandated by medical standards.

Figure 9a illustrates one circulant submatrix packing in high-latency mode. The horizontal stripes here corresponds to the decaying electrical discharge we discussed earlier, occurring regularly every epoch. In contrast, the low-latency packing mode of figure 9b depicts a circulant submatrix with only one decaying electrical discharge visible. This is because five downsampling phases now participate in low-latency packing, allowing for only one decaying electrical discharge to take place before the submatrix is filled. Inspecting the corresponding thresholded matrices of the high-latency and low-latency packing of figures 9c & 9d (respectively), we notice a "continually banded" energy in diagonalization in low-latency mode as opposed to a dominant cluster around the main diagonal in high-latency mode. This is due to the intra-ECG cycle repetitive pattern being decomposed into significant 2D spatial frequencies.

To investigate any potential performance implications when performing lower-latency, "intra-cycle" diagonalization, table 2 presents the compression metrics, again performed across the first three sets of normal sinus rhythm conditions.

Table 2: Performance comparison: high-latency vs. low-latency

Set	CR (x)		PRD (%)	
	Latency			
	high	low	high	low
{1-10}	11.8	14.4	3.55	3.23
	12.0	14.6	2.68	3.21
	12.5	11.8	3.16	3.32
	12.4	17.2	2.48	3.01
	12.2	11.9	4.09	3.36
{11-20}	7.3	9.8	4.32	4.41
	7.4	9.1	3.24	3.77
	7.1	7.0	3.58	3.57
	7.2	7.5	3.58	3.69
	7.2	7.8	4.21	3.65
{21-30}	3.9	4.5	5.94	5.25
	4.0	4.3	4.37	5.26
	4.0	3.5	4.63	5.08
	3.9	4.2	3.95	4.87
	3.9	3.3	5.88	4.73

Comparing the compression and distortion figures of high-latency and low-latency packing, no immediate performance penalty seems to have been incurred. In fact, fluctuations aside, low-latency has at times—e.g. 4th subblock in 1st set and 1st subblock in 2nd set—improved CR markedly. This is attributed to reduced fast transitions (i.e. decaying electrical discharge) within a subblock as a result of denser interleaved packing as visible from figures 9a & 9b.

4.5 Summary

The average performance in terms of CR and PRD for the two datasets studied earlier is summarized in table 3.

Table 3: Overall performance summary

Dataset	CR (x)		PRD (%)	
	μ	σ	μ	σ
Normal sinus rhythm	11.64	7.08	3.60	0.89
Arrhythmia	7.22	2.47	8.46	1.11

The combined average CR of around 9x outperforms the much lighter-weight CORTES algorithm [2] which has been

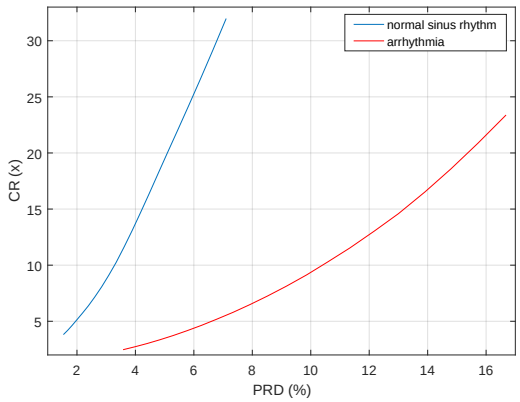


Figure 10: CR and PRD trade-off curves for normal sinus rhythm (blue) and arrhythmia (red).

Table 4: Overall latencies

	Set	Latency (sec)		Set	Latency (sec)
Normal sinus rhythm	Epochs{1-10}	207.2	Arrhythmia	Epochs{1-5}	80.0
	Epochs{11-20}	207.2		Epochs{6-11}	129.6
	Epochs{21-30}	207.2		Epochs{12-19}	168.1
	Epochs{1-10}[a]	41.4		Epochs{20-26}	40.0
	Epochs{11-20}[a]	41.4		Epochs{27-34}	41.6
	Epochs{21-30}[a]	41.4		Epochs{35-39}	125.0
	Epochs{31-40}	189.0		Epochs{40-45}	22.5
	Epochs{41-50}	189.0		Epochs{46}	25.6
	Epochs{51-60}	189.0		Epochs{47-48}	40.0
	Epochs{61-70}	64.0		Epochs{49-54}	28.9
	Epochs{71-80}	64.0		Epochs{55-62}	16.9
	Epochs{81-90}	64.0			
	Epochs{91-100}	20.8			
	Epochs{101-110}	20.8			
	Epochs{111-120}	20.8			

[a] attempt to lower latency

evaluated internally by the biomedical team on the same datasets with an average CR of around 5x. Interestingly, the real-time computational requirements of the proposed ECG compression were found to be of the same order as that of CORTES which performs minimal digital signal processing (DSP). The evaluation was conducted on a leading-edge embedded low-power platform with dedicated built-in DSP acceleration capabilities. However, the early stage evaluation did not account for matrix packing overhead required when FFT sizes (i.e. ECG periodicities) were not in exact power of two increments. Ultimately such issues can be largely mitigated by purpose-built hardware acceleration. CORTES, however, has the advantage of being amenable to compressing shorter segments of ECG data. Conversely, CORTES falls short at faithfully reproducing certain arrhythmia waveforms, impeding many diagnostic functionalities. A comparative study of the proposed cross-layer ECG coding is beyond the scope of this paper; we simply touch on certain relevant aspects in the course of the discussion to give the reader a feel for issues often encountered in this large body of literature.

For completeness, the various latencies incurred as a result of the diagonalization-based compression for all data is tabulated in table 4. The trade-off curves between CR and PRD are depicted in figure 10 for the two studied datasets. A lazier CR rise when relaxing PRD can be noted for arrhythmia. This is attributed to the non-ideal diagonalization and its reduced compression performance in aperiodic cases.

4.6 Discussion

Some commentary on the proposed cross-layer coding is supplied next.

Employing interleaved matrix packing has been shown to afford significant latency reduction compared to populating period squared number of samples. However, the method still incurs of the order of seconds worth of latency in order to extract intra- and inter-ECG period correlations. Therefore, in applications where true real-time ECG waveform is required—e.g. a medical telemetry system aimed to replace an expensive bedside monitor in an intensive care unit—other low- or zero-latency methods will be more appropriate. See for instance [23] for a cautionary note targeted at medical staff to disseminate awareness regarding latency of ECG hospital telemetry systems. Thus the proposed cross-layer coding would be more approperiate for a class of ECG streaming applications where higher latencies can be tolerated such as Holter systems [12] or the use case scenario elaborated on in section 3.

Metadata is key for the successful reconstruction of the diagonalized matrices. Two approaches can be utilized to guarantee the integrity of metadata: (1) forward error correction (FEC) and (2) added link budget. The latter is perhaps easiest to realize in a prototype system since metadata constitute a tiny fraction of the transmission stream. As such metadata can be allocated redundant budget in the most simplistic of implementations. In our system, we have found empirically that 3x link budget allocation probabilistically guarantees a near perfect information delivery.

In principle, the proposed diagonalization-based compression framework can be further leveraged for other periodic medical bio-waveforms such as a PPG (photoplethysmogram) signal—i.e. values which encode estimates of the amount of oxygen in the patient's blood—especially within the context of BANs. As such, BANs proneness to large pathloss variabilities can be combated by means of degradable signal delivery in order to enhance the overall network robustness and scalability.

5. RELATED WORK

A variety of compression methods for ECG signals are treated in prior art. Broadly, these methods fall under two categories: lossless and lossy. Examples of lossless algorithms include Huffman [10]. Lossy algorithms are subdivided further into waveform-based and transform-based methods. In methods based on the ECG waveform, clinically salient features are identified in order to retain only a subset of samples sufficient to approximate the original ECG waveform. Examples of a waveform-based methods include CORTES [2]. Transform-based methods take the ECG signal to another domain for systematic, feature-agnostic processing such as wavelet [9, 25] and Fourier [14]. A combination thereof is also possible in a hybrid fashion [12]. To the best of our knowledge, *degradable* joint source-channel coding has hitherto not been treated in the context of wireless ECG streaming.

We draw inspiration from two seminal works on graceful wireless mobile video, both utilizing 3-dimensional (3D) discrete cosine transform (DCT) in their respective systems [11, 3]. We also draw inspiration from the domain of wireless OFDM equalization, and specifically the step of building interference matrices, time-selective in [19] and frequency-selective in [16]. In a nutshell, the work presented herein is a synthesis of all-wireless concepts from [19, 16, 11, 3].

6. CONCLUSIONS

This paper presents a linear coding method that lends itself to cross-layer design for mobile wireless ECG streaming in BANs. The method is a key enabler for realizing a graceful ECG wireless monitoring wherein BAN QoS is enhanced beyond current state-of-the-art systems. This enhancement will come about through the ability of BAN to self-adjust to infrequent erroneous transmissions by introducing proportional degradation to ECG, thereby maximizing scalability and robustness. This is to be achieved capitalizing on our proposed linear joint source-channel ECG coding, which we show to perform competitively in terms of ECG compression metrics.

7. REFERENCES

[1] IEEE Standard for Local and metropolitan area networks – Part 15.6: Wireless Body Area Networks. *IEEE Std. 802.15.6-2012*, 2012.

[2] J. P. Abenstein and W. J. Tompkins. A New Data-Reduction Algorithm for Real-Time ECG Analysis. *IEEE Trans. on Biomedical Engineering*, BME-29(1):43–48, Jan 1982.

[3] S. Aditya and S. Katti. FlexCast: Graceful Wireless Video Streaming. In *Proc. of the 17th Annual Int'l. Conf. on Mobile Computing and Networking*, MobiCom '11, pages 277–288, New York, NY, USA, 2011. ACM.

[4] EESTI EVS-EN 60601-2-25. Medical electrical equipment – Part 2-25: Particular requirements for the basic safety and essential performance of electrocardiographs. Standard, Oct. 2015.

[5] ETSI EN 302 755. Digital Video Broadcasting (DVB); Frame structure channel coding and modulation for a second generation digital terrestrial television broadcasting system (DVB-T2). Standard, July 2015.

[6] A. Fort, C. Desset, P. Wambacq, and L. V. Biesen. Indoor body-area channel model for narrowband communications. *IET Microwaves, Antennas Propagation*, 1(6):1197–1203, Dec 2007.

[7] R. Gray. *Toeplitz and Circulant Matrices: A Review.* Foundations and Trends in Technology. Now Publishers, 2006.

[8] M. Hernandez-Silveira, K. Ahmed, S.-S. Ang, F. Zandari, T. Mehta, R. Weir, A. Burdett, C. Toumazou, and S. J. Brett. Assessment of the feasibility of an ultra-low power, wireless digital patch for the continuous ambulatory monitoring of vital signs. *BMJ Open*, 5(5), 2015.

[9] M. L. Hilton. Wavelet and wavelet packet compression of electrocardiograms. *IEEE Trans. on Biomedical Engineering*, 44(5):394–402, May 1997.

[10] ISO 11073-91064. Health informatics – Standard communication protocol – Part 91064: Computer-assisted electrocardiography. Standard, May 2009.

[11] S. Jakubczak and D. Katabi. A Cross-layer Design for Scalable Mobile Video. In *Proc. of the 17th Annual Int'l. Conf. on Mobile Computing and Networking*, MobiCom '11, pages 289–300, New York, NY, USA, 2011. ACM.

[12] H. Kim, R. F. Yazicioglu, P. Merken, C. V. Hoof, and H. J. Yoo. ECG Signal Compression and Classification Algorithm With Quad Level Vector for ECG Holter System. *IEEE Trans. on Information Technology in Biomedicine*, 14(1):93–100, Jan 2010.

[13] Laerdal Medical Limited. SimMan® Patient Simulator. http://www.laerdal.com/gb/doc/86/SimMan.

[14] M. S. Manikandan and S. Dandapat. ECG Signal Compression using Discrete Sinc Interpolation. In *2005 3rd Int'l. Conf. on Intelligent Sensing and Information Processing*, pages 14–19, Dec 2005.

[15] D. Miniutti, L. Hanlen, D. Smith, A. Zhang, D. Lewis, D. Rodda, and B. Gilbert. Narrowband on-body to off-body channel characterization for BAN. Technical report, IEEE 802.15-08-0559-00-0006, August, 2008.

[16] A. F. Molisch, M. Toeltsch, and S. Vermani. Iterative Methods for Cancellation of Intercarrier Interference in OFDM Systems. *IEEE Trans. on Vehicular Technology*, 56(4):2158–2167, July 2007.

[17] A. T. Reisner, G. D. Clifford, and R. G. Mark. *The physiological basis of the electrocardiogram.* Artech House Publishers, 2006.

[18] Rigel Medical. Rigel 333 ECG Patient Simulator. http://www.rigelmedical.com/products/simulators/ecg-patient-simulator.

[19] P. Schniter. Low-complexity equalization of OFDM in doubly selective channels. *IEEE Trans. on Signal Processing*, 52(4):1002–1011, April 2004.

[20] D. B. Smith and L. W. Hanlen. Channel modeling for wireless body area networks. In *Ultra-Low-Power Short-Range Radios*, pages 25–55. Springer, 2015.

[21] D. B. Smith, D. Miniutti, T. A. Lamahewa, and L. W. Hanlen. Propagation models for body-area networks: A survey and new outlook. *IEEE Antennas and Propagation Magazine*, 55(5):97–117, 2013.

[22] Toumaz Healthcare. 2nd generation SoC Functional Specifications. Technical report, July 2016.

[23] M. P. Turakhia, N. M. Estes, B. J. Drew, C. B. Granger, P. J. Wang, B. P. Knight, and R. L. Page. Latency of ECG Displays of Hospital Telemetry Systems. *Circulation*, 126(13):1665–1669, 2012.

[24] S. Vembu, S. Verdu, and Y. Steinberg. The source-channel separation theorem revisited. *IEEE Trans. on Information Theory*, 41(1):44–54, Jan 1995.

[25] X. Wang and J. Meng. A 2-D ECG Compression Algorithm Based on Wavelet Transform and Vector Quantization. *Digit. Signal Process.*, 18(2):179–188, Mar. 2008.

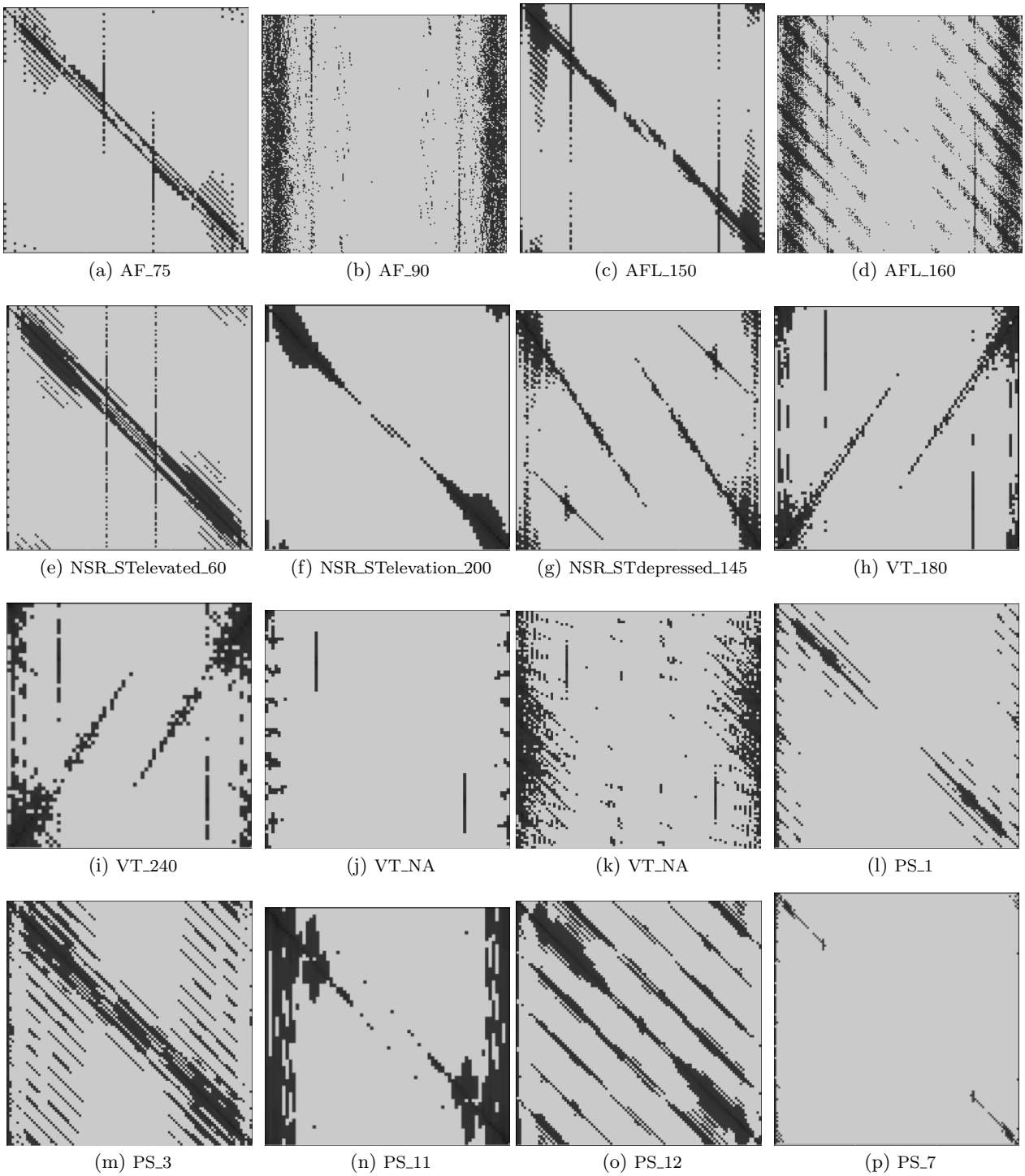
[26] A. C. W. Wong, D. McDonagh, O. Omeni, C. Nunn, M. Hernandez-Silveira, and A. J. Burdett. Sensium: an ultra-low-power wireless body sensor network platform: Design & application challenges. In *2009 Annual Int'l. Conf. of the IEEE Engineering in Medicine and Biology Society*, pages 6576–6579, Sept 2009.

[27] K. Y. Yazdandoost and K. Sayrafian-Pour. Channel model for body area network (BAN). Technical report, IEEE P802.15-08-0780-09-0006, April, 2009.

APPENDIX

A. ASSORTED SELECTION OF DIAGONALIZED MATRICES

For the benefit of interested parties, we provide an assorted selection of examples of diagonalized transform-domain matrices in figure 11 as encountered during the characterization conducted on normal sinus rhythm and arrhythmia datasets. Emphasis is placed on the more exotic-looking examples to cement the reader's intuition on inner-workings of the proposed coding method. For instance, it can be readily seen in figures 11b, 11j, and 11k that diagonalization has failed to produce diagonally-extending runs of transform-domain coefficients in one atrial fibrillation and two ventricular fibrillation cases, respectively. Nonetheless, large areas in these matrices are thresholded out and as such useful compression is obtained.



(a) AF_75 (b) AF_90 (c) AFL_150 (d) AFL_160

(e) NSR_STelevated_60 (f) NSR_STelevation_200 (g) NSR_STdepressed_145 (h) VT_180

(i) VT_240 (j) VT_NA (k) VT_NA (l) PS_1

(m) PS_3 (n) PS_11 (o) PS_12 (p) PS_7

Figure 11: Examples of diagonalized transform-domain matrices (not to scale).