Parsing Wireless Electrocardiogram Signals with Context Free Grammar Conditional Random Fields

Thai Nguyen*, Roy J. Adams*, Annamalai Natarajan*, and Benjamin M. Marlin

Abstract— **Recent advances in wearable sensor technology have made it possible to simultaneously collect multiple streams of physiological and context data from individuals as they go about their daily activities in natural environments. However, extracting reliable higher-level inferences from these raw data streams remains a key data analysis challenge. In this paper, we focus on the specific case of the analysis of data from wireless electrocardiogram (ECG) sensors. We present a new robust probabilistic approach to ECG morphology extraction using conditional random field context free grammar models, which have traditionally been applied to parsing problems in natural language processing. We focus on ECG morphology extraction because it is a key step in higher-level detection tasks such as arrhythmia detection and the detection of drug use. We introduce a robust context free grammar for parsing noisy ECG data, and show significantly improved performance on the ECG morphological labeling task.**

I. Introduction

Wearable wireless sensors have the potential for transformative impact on the fields of health and behavioral science. Recent advances in wearable sensor technology have made it possible to simultaneously collect multiple streams of physiological and context data from individuals as they go about their daily activities in natural environments. However, extracting reliable higher-level inferences from these raw data streams remains a key data analysis challenge.

In this paper, we focus on the specific case of the analysis of data from wireless electrocardiogram sensors. An electrocardiogram (ECG) sensor produces a data stream corresponding to the electrical activity of the muscles of the heart as measured at the surface of the skin. Each normal cardiac cycle (corresponding to a single heart beat) produces a characteristic sequence of five waves (the P, Q, R, S, and T waves) as shown in Figure 1a. These waves are the result of atrial and ventricular depolarization and repolarization.

The analysis of ECG data provides fundamental information about health and activity starting with estimates of heart rate and heart rate variability [1], which are computed from the intervals between successive R wave peaks. Heart rate and heart rate variability have been used to infer stress [2], assess cognitive load [3], detect activity [4], recognize mood [5] and study mental health [6]. More detailed applications of ECG, such as arrhythmia (irregular heart beat) detection [7] and the detection of cocaine use [8], require the analysis of the morphological structure of individual cardiac cycles including the shapes of the waves and the intervals between them. For instance, myocardial infarction manifests as a series of morphological changes starting with increased T wave amplitude, followed by ST elevation, inverted T peaks and significant changes in the Q peak [9]. Morphological changes in ECG in the presence of cocaine include reduction in the RR interval (increased heart rate) modulation of the QT interval, and decrease in T wave amplitude [10].

Fig. 1: (a) Idealized ECG waveform (b) Sample data from the Zephyr BioHarness wearable chest band sensor

While low-cost off-the-shelf wearable ECG sensors can be readily deployed outside of clinical settings, the reliability of algorithms for extracting morphology from the resulting ECG data streams continues to lag behind the performance that is achievable in clinical settings. The primary reasons are that wireless ECG data are inherently noisy due to the limited number of electrodes, variable electrode placement, and issues with electrode contact with the skin. Additionally, the data are subject to artifacts like signal dropouts and baseline shifts. Refer to figure 1b for a sample ECG data gathered from the Zephyr BioHarness 3 wearable chest band sensor[1]. While some waves (*e.g., the R waves*) are more obvious, the presence of noise can make the detection of the other waves much more challenging.

In prior work, we presented an approach to the problem of ECG morphological structure extraction based on a probabilistic structured prediction model for sequence data known as a linear chain Conditional Random Field (CRF) [12]. This model generalizes the widely used logistic regression classifier by allowing for first-order Markov dependence between adjacent labels in a sequence [13]. The output of this framework is a sequence of peak locations annotated with their wave types (P, Q, R, S, T). Morphological intervals of interest (e.g, QT distance, etc.) can then be computed from the labeled peak locations. Due to the fact that noise in the data can generate spurious peaks, the model is explicitly trained on an additional label "N" that is applied to peaks

University of Massachusetts Amherst, thainguyen@umass.edu, rjadams@cs.umass.edu, anataraj@cs.umass.edu, marlin@cs.umass.edu. * These authors contributed equally to this work.

[1]**http://www.zephyr-technology.com/products/bioharness-3/**

978-1-5090-3090-3/16/$31.00 ©2016 IEEE

that do not correspond to the peaks of valid waves. While this approach significantly out-performed methods developed for clinical ECG as well as the independent logistic model when applied to wireless ECG data, it still lags behind the performance that methods developed for clinical data can achieve when applied to clinical data. Our goal in this paper is to further close this gap.

Motivated by the limitations imposed by the first order Markov assumption that underlies the linear chain CRF, in this paper we present a new approach to the ECG morphology extraction problem based on a higher-order CRF model referred to as the Context Free Grammar Conditional Random Field, or CRF-CFG model [14]. The CRF-CFG model was originally developed as a conditional text parsing model for applications in natural language processing. The model is defined through a context free grammar that consists of a set of symbols and productions that are applied recursively, producing a sequence of terminal symbols via an internal tree structure. In the CRF-CFG model, each production defines a factor that can be conditioned on data local to where the production is applied. The joint probability of a tree is then defined to be proportional to the product of the production factors used at each node in the tree. While a linear chain CRF can be thought of as inducing a joint distribution over all possible *sequences* of labels given the input features, the CRF-CFG model can be thought of as inducing a distribution over all possible *trees* allowable under the grammar that defines the model.

The advantage of using a CRF-CFG compared to a linear chain CRF is that the CRF-CFG can capture higher-order structure in the labels including long-range dependencies that span cardiac cycles. This plays an important role in ECG morphology extraction since valid ECG peaks are interspersed by noisy non-ECG peaks. In addition, it allows for models that can reason about the relative likelihood of different peak labels based on the time intervals between them.

The primary contribution of this paper is the specification of a robust context free grammar to be used with CRF-CFG models for ECG morphology extraction. To the best of our knowledge, this is the first CFG provided for parsing ECG data. We evaluate the CRF-CFG model on data collected in both the lab and field settings, extending our prior evaluation that focused only on data collected in the lab setting. We report a mean ECG peak labeling accuracy of 94% on ECG data collected in the lab setting. The CRF-CFG achieves 22.2% reduction in error when compared to using the previous linear chain CRF model. Additionally, we evaluate the CRF-CFG model on data collected in the field setting, obtaining an accuracy of 88% under a strict lab-to-field generalization protocol, a 19.5% reduction in error relative to the previous linear chain CRF model [12].

II. Background and Related Work

In this section we briefly review the use of wireless ECG data in mobile health, related work on the problem of ECG morphological structure extraction, and the conditional random field (CRF) framework our work is based on.

A. ECG in Mobile Health

A substantial body of work has explored the use of mobile ECG sensors for applications in health and behavioral science, primarily in the context of understanding physiological stress [15], [16], assessing cognitive load [3], detection of arrhythmias caused specifically by atrial fibrillation [17], [18], [7], and detection of drug use including cocaine use [8], [19]. Nearly all of these studies have been based on features derived from R-R intervals only (heart rate, heart rate variability).

B. ECG Morphological Analysis

The problem of interest in this work is morphological labeling of the ECG trace including the identification of each P, Q, R, S and T wave, when present. The most common approach to this problem is to first identify QRS complexes, and then to apply a set of rules and a local search procedure to identify the individual waves [20], [21]. Pan and Tompkins developed a widely used QRS complex detection algorithm that achieves a QRS detection accuracy rate of 99.325% on the well-known MIT-BIH data set [22]. More recent work on QRS complex detection has focused on transform-based methods including the curve length transform [23] and the wavelet transform [21]. Both of these approaches give QRS complex identification precision and recall rates above 99.5% on standard databases. Our prior work on CRF-based methods for two-lead wireless ECG data showed that such rule-based local search methods can be very brittle in the presence of noise, leading to significantly worse performance compared to the use of probabilistic methods, specifically, the linear chain CRF model that we describe next [12].

C. Conditional Random Fields

CRFs are a sub-class of probabilistic graphical models [24] that generalize independent probabilistic classifiers like logistic regression [25] to the case of structured prediction. A CRF contains a set of label variables $\mathbf{Y} = \{Y_1, ..., Y_L\}$ and a corresponding set of feature variables $\mathbf{X} = \{X_1, ..., X_L\}$, where L is the number of variables in the model. We assume the label variables take values in the set $\mathcal{Y}$. The feature variables $\mathbf{X}_i$ are D-dimensional vectors of feature values.

A general log-linear CRF is defined through a linear energy function that takes the form of a weighted sum of K feature functions f_k involving values of $\mathbf{Y}$ and $\mathbf{X}$:

$$E_{\mathbf{w}}(\mathbf{y}, \mathbf{x}) = -\sum_{k=1}^{K} w_k f_k(\mathbf{y}, \mathbf{x})$$

These feature functions are typically sparse in the sense that they involve few label and feature variables. The set of label and feature variables referenced in function f_k is referred to as its scope S_k.

The joint probability $P_{\mathbf{w}}(\mathbf{y}|\mathbf{x})$ of a setting of the label variables $\mathbf{y} = [y_1, ..., y_L]$ conditioned on the observed feature variables $\mathbf{x} = [\mathbf{x}_1, ..., \mathbf{x}_L]$ is given below. $Z_{\mathbf{W}}(\mathbf{x})$ is referred to

as the *partition function* and is the normalization term of the probability distribution.

$$P_{\mathbf{w}}(\mathbf{y}|\mathbf{x}) = \frac{\exp(-E_{\mathbf{w}}(\mathbf{y}, \mathbf{x}))}{Z_{\mathbf{w}}(\mathbf{x})} \quad (1)$$

$$Z_{\mathbf{W}}(\mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{V}^L} \exp(-E_{\mathbf{w}}(\mathbf{y}, \mathbf{x})) \quad (2)$$

Given a data set $\mathcal{D} = \{(\mathbf{y}_n, \mathbf{x}_n)\}_{n=1:N}$ of fully labeled training sequences, the unknown parameters $\mathbf{w}$ must be learned from training data before the model can be applied for prediction. The parameters can be estimated by maximizing the ℓ_2 regularized conditional log likelihood shown below:

$$\mathcal{L}(\mathbf{w}|\mathcal{D}) = \sum_{n=1}^{N} \log P_{\mathbf{w}}(\mathbf{y}_n|\mathbf{x}_n) - \lambda \|\mathbf{w}\|_2^2 \quad (3)$$

This objective function is strongly convex, so gradient-based methods are guaranteed to find the unique optimal solution. Computing the gradients requires marginal distributions over the scope S_k of each factor f_k. If S_k contains at most two variables for all k, then S_k is referred to as a pairwise CRF, and its scopes can be represented using a graph $\mathcal{G}$ where an undirected edge connects each pair of variables that share a scope.

If the graph $\mathcal{G}$ is a tree, then the resulting CRF is referred to as a tree-structured CRF. Importantly, the marginals required to learn the parameters for a tree-structured CRF can be computed exactly in time linear in the number of variables in the model using the belief propagation algorithm [24]. Chain-structured CRFs are an important special case of tree-structured CRFs. These models are widely used for modeling sequential data, and formed the core of our previous approach to ECG morphological analysis.

Finally, we note that multinomial logistic regression (MLR) can be thought of as a special case of a CRF where each scope S_i consists of exactly one label variable Y_i and it's corresponding feature vector $\mathbf{X}_i$. In this case, the label variables are all probabilistically independent of each other.

D. CRFs for ECG Morphology Extraction

To apply the linear chain CRF model described in the previous section to ECG data, our prior work proposed a pipeline that consisted of four steps including (1) unsupervised detection of candidate locations, (2) extracting features from windows around candidate peak locations, (3) generating the structure of a linear chain CRF model given the sequence of candidate peak locations, and (4) using probabilistic inference to jointly infer labels for all candidate peak locations. Note that in Step 1, noise in the data can result in candidate peak locations that do not correspond to peaks of valid ECG waves. As a result, an additional label ("N") is included when training and predicting to denote candidate peak locations that do not correspond to valid waves.

III. CRF-CFG Models for ECG

In this section we introduce the Conditional Random Field Context Free Grammar (CRF-CFG) model and describe its application to ECG morphology extraction. We begin by describing context free grammars (CFGs), followed by the CRF-CFG model, and finally introduce the grammar we propose for ECG morphological extraction.

A. Context Free Grammars

A context free grammar (CFG) is a mathematical structure that compactly describes the set of valid strings that can occur in a language. The set of valid strings is described through a set of production rules $\mathcal{R}$ that specify transformations from a set of internal (or non-terminal) symbols $\mathcal{I}$ to sequences consisting of both non-terminal and terminal symbols. We let $\mathcal{V}$ represent the set of terminal symbols. A string in the language is simply a sequence of terminal symbols.

In this work, we will consider CFGs where each rule $r \in \mathcal{R}$ is either a triple (A, B, C) where $A \in \mathcal{I}$, $B \in \mathcal{I} \cup \mathcal{V}$, and $C \in \mathcal{I} \cup \mathcal{V}$, or a tuple (A, a) where $A \in \mathcal{I}$ and $a \in \mathcal{V}$.[2] These rules are written in the form $A \rightarrow BC$ or $A \rightarrow a$.

To generate a string, the set of rules $\mathcal{R}$ is applied recursively starting from a special "start" symbol $\alpha \in \mathcal{I}$. The generation of a terminal symbol serves as the recursion base case. The sequence of recursive production rule applications thus generates a binary tree with non-terminal symbols as the internal nodes and terminal symbols at the leaf nodes. The left-to-right sequence of terminal symbols in the tree gives the generated string.

A CFG is formally defined as a tuple $G = (\mathcal{I}, \mathcal{V}, \mathcal{R}, \alpha)$. The language defined by the grammar G consists of all strings that can be generated through the recursive application of production rules in $\mathcal{R}$ starting from α.

As an example, consider a simple CFG with $\mathcal{I} = \{\alpha, A, B\}$, $\mathcal{V} = \{a, b\}$ and the production rules $\alpha \rightarrow AB$, $A \rightarrow aA$, $A \rightarrow a$, $B \rightarrow bB$, $B \rightarrow b$.[3] The recursive application of these rules produces strings that contain any number of a's followed by any number of b's.

The problem of parsing a string is the problem of inverting the generative process defined by the grammar to infer the tree structure and production rules responsible for generating the string. In the simple example described above, every string in the language has a unique valid parse, but this is not the case in general. In such cases, weights can be attached to the productions to express their relative likelihoods, and the parsing problem can be converted into the problem of identifying the tree structure and production rules that result in the parse with the maximum total weight. The weighted CFG can equivalently be viewed as a joint probabilistic model over trees, and the maximum weighted parse can be interpreted as the maximum probability parse. The conditional random field context free grammar model presented in the next section generalizes the probabilistic CFG model to the case where the weights on the productions can depend on features of the input sequence.

[2]This is a slightly relaxed form equivalent to Chomsky normal form. It allows for more compact sets of production rules.

[3]As a notational convenience, the possible productions starting from each non-terminal are typically written together using "|" as a separator. This allows writing the last four rules as $A \rightarrow aA|A$ and $B \rightarrow bB|B$.

B. The CRF-CFG Model

The conditional random field context free grammar (CRF-CFG) model is a subclass of the general CRF model framework that specifies a probability distribution over parse trees of a string given a context free grammar $G = (\mathcal{I}, \mathcal{V}, \mathcal{R}, \alpha)$. We assume an input string of length L. We let $\mathbf{x}_i$ be a feature vector associated with position i in the string and $\mathbf{x} = [\mathbf{x}_1, ..., \mathbf{x}_L]$. We define a set of binary variables $\{y_{A,BC,i,j,l} \mid A \rightarrow BC \in \mathcal{R}, 1 \leq i \leq j < l \leq L\}$. $y_{A,BC,i,j,l}$ takes the value 1 if the rule $A \rightarrow BC \in \mathcal{R}$ is used in the parse tree with the sub-tree rooted at B covering positions i to j in the input string and the sub-tree rooted at C covering positions $j+1$ to l. The variable $y_{A,BC,i,j,l}$ takes the value 0 otherwise.

Next, we define a set of $K^{A,BC}$ scalar feature functions for every production rule $A \rightarrow BC$ in $\mathcal{R}$: $\{f_k^{A,BC}(y_{A,BC,i,j,l}, i, j, l, \mathbf{x}) \mid 1 \leq i \leq j < l \leq L\}$. Each feature function takes the form of a product of the binary indicator variable $y_{A,BC,i,j,l}$ and a function $g_k^{A,BC}(i, j, l, \mathbf{x})$ that computes a feature value from $\mathbf{x}$.

$$f_k^{A,BC}(y_{A,BC,i,j,l}, i, j, l, \mathbf{x}) = y_{A,BC,i,j,l} \cdot g_k^{A,BC}(i, j, l, \mathbf{x}) \quad (4)$$

For each production rule $A \rightarrow BC$ in $\mathcal{R}$, we have a corresponding set of weights $\{w_k^{A,BC}\}$. If there is exactly one feature function $f^{A,BC}(y_{A,BC,i,j,l}, i, j, l, \mathbf{x}) = y_{A,BC,i,j,l}$ for each production rule, then the model corresponds to the weighted CFG model described earlier. The main strength of the CRF-CFG model is that the feature extraction functions $g_k^{A,BC}$ can be any desired function of i, j, l and $\mathbf{x}$, leading to tremendous flexibility.

The final form of the CRF-CFG energy function is shown below. We note that one additional global feature function $f^T(\mathbf{y})$ is needed to ensure that that the collection of binary variables define a valid tree structure. $f^T(\mathbf{y})$ takes the value $-\infty$ if $\mathbf{y}$ defines a valid parse tree, and the value 0 if it does not. The conditional probability of the parse given the feature vector is derived from the energy function as in the case of the general CRF shown in the previous section.

$$E_{\mathbf{w}}(\mathbf{y}, \mathbf{x}) = - f^T(\mathbf{y}) - \sum_{A \rightarrow BC \in \mathcal{R}} \sum_{1 \leq i \leq j < l \leq L} \sum_{k=1}^{K^{A,BC}} w_k^{A,BC} \cdot f_k^{A,BC}(y_{A,BC,i,j,l}, i, j, l, \mathbf{x}) \quad (5)$$

While this model is substantially richer and more complex than the linear chain CRF, it has the important property that the maximum probability parse can still be computed in polynomial time given a setting of the weights $\mathbf{w}$. Specifically, the maximum probability parse can be computed in $O(L^3)$ time using the inside-outside dynamic programming algorithm originally developed for the weighted CFG model [26]. In the next section, we describe how this model can be used to infer the morphological structure of ECG data.

C. CFGs for ECG Morphology Extraction

The ECG morphology extraction problem defined in the previous section requires the assignment of labels to peaks in the ECG waveform. We follow the same basic methodology introduced for the application of the linear chain CRF, also described in the previous section. In particular, the input to the problem is a discrete sequence of feature vectors extracted from locations that correspond to candidate peaks in the raw ECG time series. We will refer to this as the ground sequence. The label set is again $\mathcal{V} = \{p, q, r, s, t, n\}$ with n indicating a candidate peak that does not correspond to the peak of a valid wave. These labels will serve as the terminal symbols in our ECG context free grammar.

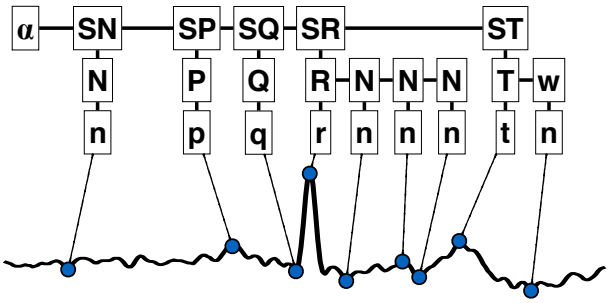

Fig. 2: A sample ECG parse. The sub-tree rooted at *SR* exhibits the grouping of true peaks with all trailing noise peaks that is enforced in the CRF-CFG model.

Unlike the case of natural language, we are free to introduce any non-terminals and production rules that we like that may help in correctly labeling the ground sequence. Since we expect that ECG waves occur in a cycle $p, q, r, s, t, p, ...$ we can use this information to define a basic grammar as seen below. The symbols P, Q, R, S, T are non-terminals that provide the structure needed to generate the desired sequence of terminals. This grammar is able to start parsing with any wave and then continues to cycle through the wave labels in sequence.

$$\begin{aligned} &\alpha \rightarrow P \mid Q \mid R \mid S \mid T \\ &P \rightarrow pQ \qquad R \rightarrow rS \qquad T \rightarrow tP \\ &Q \rightarrow qR \qquad S \rightarrow sT \end{aligned}$$

The obvious problem with the previous grammar is that it fails to account for the noise and imperfect sequences that can occur in wireless ECG data. Below, we show and alternate robust grammar that allows for arbitrary runs of noise labels that group with the previous valid wave label. It also allows for skipping between any pair of valid wave labels. At the same time, the weights associated with the production rules can be learned from data in a way that reflects the relative likelihood of their occurrence. For example, we might expect a Q to follow a P most of the time, although noise in the data may occasionally cancel out a low magnitude Q wave resulting in the need to skip from P directly to R. This kind of skipping is also needed to deal effectively with certain types of arrhythmias. In order to minimize assumptions about how long noise may last or how many peaks it may corrupt we allow for transitions between any two peak labels. A further property of this grammar is that it allows us to featurize the time gaps

between subsequent valid wave labels (p,q,r,s,t), which is not possible in a linear chain CRF. Figure 2 shows an example parse using this grammar overlayed with the input signal and detected peaks.

$$
\begin{aligned}
\alpha &\rightarrow SP \mid SQ \mid SR \mid SS \mid ST \mid SN \\
SP &\rightarrow P\ SQ \mid P\ SP \mid P\ SR \mid P\ SS \mid P\ ST \\
SQ &\rightarrow Q\ SQ \mid Q\ SP \mid Q\ SR \mid Q\ SS \mid Q\ ST \\
SR &\rightarrow R\ SQ \mid R\ SP \mid R\ SR \mid R\ SS \mid R\ ST \\
SS &\rightarrow S\ SQ \mid S\ SP \mid S\ SR \mid S\ SS \mid S\ ST \\
ST &\rightarrow T\ SQ \mid T\ SP \mid T\ SR \mid T\ SS \mid T\ ST \\
SN &\rightarrow N\ SQ \mid N\ SP \mid N\ SR \mid N\ SS \mid N\ ST \\
P &\rightarrow p \mid p\ N \qquad R \rightarrow r \mid r\ N \qquad T \rightarrow t \mid t\ N \\
Q &\rightarrow q \mid q\ N \qquad S \rightarrow s \mid s\ N \qquad N \rightarrow n \mid n\ N
\end{aligned}
$$

D. Feature Functions for ECG Morphology Extraction

The feature functions in the CRF-CFG model framework give us great flexibility in modeling high-level structure in the ECG morphology. In this work we use the feature functions to model the characteristic time between adjacent peaks. In particular, we include feature functions of the form,

$$g_1^{SP,PSQ}(i,j,l) = 1 \tag{6}$$

$$g_2^{SP,PSQ}(i,j,l) = (t_{j+1} - t_i) \tag{7}$$

$$g_3^{SP,PSQ}(i,j,l) = (t_{j+1} - t_i)^2 \tag{8}$$

where t_i is the timestamp of peak i. The weights on these features can encode any quadratic function of the time between adjacent P and Q peaks with any number of N peaks between them. Features of this form are included for all productions starting from the symbols $\{SP, SQ, SR, ST, SS\}$.

As in the linear chain CRF, we also include features $\mathbf{x}_i$ calculated locally at each peak i in the sequence based on the input ECG signal. These features are incorporated for each rule that produces a terminal using feature functions of the following form,

$$g_k^{P,pP}(i,j,l) = \mathbf{x}_{ik} \tag{9}$$

where $i = j$ because p is a terminal. The specific features used include a sparse coding of the underlying signal as well as amplitude based features. The extraction of these features is described in more detail in Section IV-B.

E. Inference and Learning

As stated above, finding the maximum probability parse (known as maximum a posteriori (MAP) inference) in the CRF-CFG model takes $O(L^3)$ time using variations on the inside-outside algorithm [26]. Specifically, we find the maximum probability parse given a feature sequence $\mathbf{x}$ by solving the following dynamic program,

$$
\begin{aligned}
b(A,i,l) = \max_{A \rightarrow BC \in \mathcal{R}} \max_{i \le j < l} & \; b(B,i,j) + b(C,j+1,l) \\
& + \sum_{k=1}^{K^{A,BC}} w_k^{A,BC} \cdot g_k^{A,BC}(i,j,l,\mathbf{x})
\end{aligned} \tag{10}
$$

In the case where BC is only a single symbol, then the max over j can be dropped and we have the following base-case for rules producing only terminals,

$$b(a,i,l) = \begin{cases} -\infty & : i \neq l \\ \sum_{k=1}^{K^a} w_k^a g_k^{A,a}(i,i,l,\mathbf{x}) & : i = l \end{cases} \tag{11}$$

Then we have the following equivalence where c is a constant with respect to $\mathbf{y}$.

$$\max_{\mathbf{y}} \log P_{\mathbf{w}}(\mathbf{y}|\mathbf{x}) = \max_{\mathbf{y}} \log\left(-E_{\mathbf{w}}(\mathbf{y},\mathbf{x})\right) - \log Z_{\mathbf{w}}(\mathbf{x}) \tag{12}$$

$$= b(\alpha, 1, L) - c \tag{13}$$

The MAP solution, $\mathbf{y}^*$, can be found by backtracking through the dynamic programming table.

To estimate the parameters of the CRF-CFG model, we use structured maximum-margin learning [27]. In this framework, we view $-E_{\mathbf{w}}(\mathbf{y},\mathbf{x})$ as a scoring function and estimate the parameters of the scoring function by minimizing the regularized hinge loss. By observing that $E_{\mathbf{w}}(\mathbf{y},\mathbf{x})$ is a linear function of the features f_k, we can view the whole model as a linear support vector machine and use a cutting-plane method to efficiently optimize the hinge loss [27], [28]. This method requires only that we find the highest scoring $\mathbf{y}$ for a given $\mathbf{w}$, which is equivalent to finding the maximum probability parse.

Finally, we can improve performance by replacing the standard hinge loss with a loss-augmented alternative. In the loss-augmented hinge loss, we add the hamming loss between the predicted labels for the ground sequence and the true labels for the ground sequence to the loss margin [27]. This has the desirable effect of encouraging incorrect labelings that are close to the true labeling in terms of hamming loss to be scored higher than those that are not. Optimizing the loss augmented problem involves finding $\mathbf{y}$ that maximizes the score plus the hamming loss. Because hamming loss decomposes as a sum over individual positions in the base sequence, we solve this problem by augmenting the weights and feature functions appropriately and solving a modified MAP inference problem (see [27] for details). An additional desirable effect of this change is that it focuses the learning on correctly predicting the ground sequence labels, which correspond to the actual morphological structure of interest, while not penalizing for differences in the internal structure of the parse tree.

IV. Empirical Protocols

A. Datasets

The data we use to evaluate our proposed model was gathered from habituated cocaine users who participated in a NIDA-approved study that included both lab and field components. All subjects reviewed and signed a consent form approved by the local institutional review board. In the lab-based component, data were obtained from six subjects. In the field-based component, data were obtained from five subjects. In both the lab and field components, the wireless ECG sensor used was a Zephyr BioHarness chest band

Subject	# Days	Duration	# Samples	# Candidate peaks	# Labeled peaks	Labeled peaks – individual counts						# Segments
						#P	#Q	#R	#S	#T	#N	
1	–	~6h36m	5,945,750	267,233	3,266	497	499	499	494	529	748	179
2	–	~7h01m	6,314,250	268,110	4,716	608	614	620	591	510	1773	480
3	–	~7h41m	6,920,750	321,219	3,231	525	531	535	502	532	606	141
4	–	~11h2m	9,924,250	360,149	4,144	509	728	786	780	783	558	224
5	–	~11h54m	10,715,000	436,917	3,112	502	518	523	515	528	526	175
6	–	~15h45m	14,171,500	448,806	3,966	631	655	664	518	505	993	332
Lab total	–	~59h	53,991,500	2,102,434	22,435	3,272	3,545	3,627	3,400	3,387	5,204	1,531
1	3	~60h30m	54,445,000	2,154,191	543	64	77	84	75	54	189	42
2	3	~60h47m	54,706,250	2,397,117	609	68	79	80	57	75	250	29
3	3	~50h27m	45,398,250	1,793,234	411	58	61	63	61	62	106	21
4	4	~76h59m	69,292,000	2,074,527	906	89	100	103	97	97	420	31
5	5	~91h21m	82,213,000	3,904,416	949	120	131	136	136	126	300	50
Field total	18	~340h	306,054,500	12,323,485	3,418	399	448	466	426	414	1,265	173

TABLE I: Lab and field dataset statistics including number of labeled peaks and segments

sensor, which sampled data at 250Hz and transmitted it via Bluetooth to a smartphone.

To create a labeled dataset, a large number of segments consisting of two to four cardiac cycles were selected at random from the complete timeseries. The candidate peak generation process was then run over these segments to over-generate possible ECG peak locations using an implementation of Billauer's PeakDet method with the threshold set at $t = 0.02$ [29]. These candidate peaks were then labeled by the research staff.[4] Approximately $20,000$ peaks were labeled in the lab dataset and 3000 in field dataset. These labeled peaks are roughly organized into ~1500 and ~170 segments respectively. The details of the data from the lab and field components of this study are summarized in Table I.

We selected these data sets because the problem domain is highly challenging. The lab data exhibits significant variability due to cocaine use. The field data set is even more variable as it was recorded in the natural environment from subjects engaged in a wide range of activities including significant drug use.

B. Feature Extraction

We begin by defining a window of length 51 centered on each candidate peak in the raw ECG signal (25 data samples on either side of candidate peak location; 204 milliseconds). To account for baseline drift, we normalized the signal within each candidate peak window by subtracting off the mean within the segment.

As in [12], we apply sparse coding to learn a shape-based representation of the signal within each window. We learn a sparse coding dictionary with $K = 100$ basis elements using the data extracted from all of the six lab subjects' candidate peaks. We use a value of $\alpha = 1$ for the sparse coding sparsity regularization term. The same dictionary (learned only on lab subjects) is used for both the lab and field datasets.

We use the learned dictionary to perform sparse coding on candidate peak windows for both the training and test subjects. We supplement the sparse coding features with the height and height squared of the candidate peak within each window since this information is factored out in the normalization process before sparse coding is applied. We also add a bias feature (a constant feature equal to 1). The resulting feature representation for each candidate peak is thus a length 103 vector.

For the CRF-CFG model, we also include features between all pairs of candidate peaks within a segment. In particular, we use the time between the peaks, the time squared, and a constant bias feature.

C. Morphology Extraction Methods and Evaluation

In both lab and field datasets, we consider three different methods for extracting morphology in ECG data. We compare the performance of the CFG based CRF model (CRF-CFG) to the linear chain CRF model (CRF-LC) from [12], and the standard multinomial logistic regression model (MLR). The MLR model uses only candidate peak features to predict and predicts labels independently. The linear chain CRF model uses both candidate peak features and transition information between peaks to predict peak labels. The CRF-CFG model extends the CRF-LC model as described above. The MLR and CRF-LC models are trained using maximum likelihood estimation while the CRF-CFG model is trained to minimize hinge loss. The linear chain model and the CFG model are both trained on subsequences of the data.

D. Train-Test Protocols

We performed two, slightly different, prediction experiments on the lab and field datasets. On the lab data, we used a *leave-one-subject-out* protocol. Specifically, for each lab subject, we tuned and trained a model using all data from the other five subjects and tested base sequence prediction accuracy on the held out subject. To tune hyper-parameters for each model, we performed a further cross-validation on the training set by randomly generating three leave-one-subject-out train/validation splits from the remaining five train subjects.

To test the models on field data, we made the strict assumption that no field data is available at train time. In this protocol, all field subjects are put into a single test set and a model is tuned and trained using all lab subjects. The

[4]The research staff had basic knowledge of the ECG structure, but were not experts. However, the subjects were pre-screened for arrhythmias by clinical staff, and any label errors affect all algorithms equally.

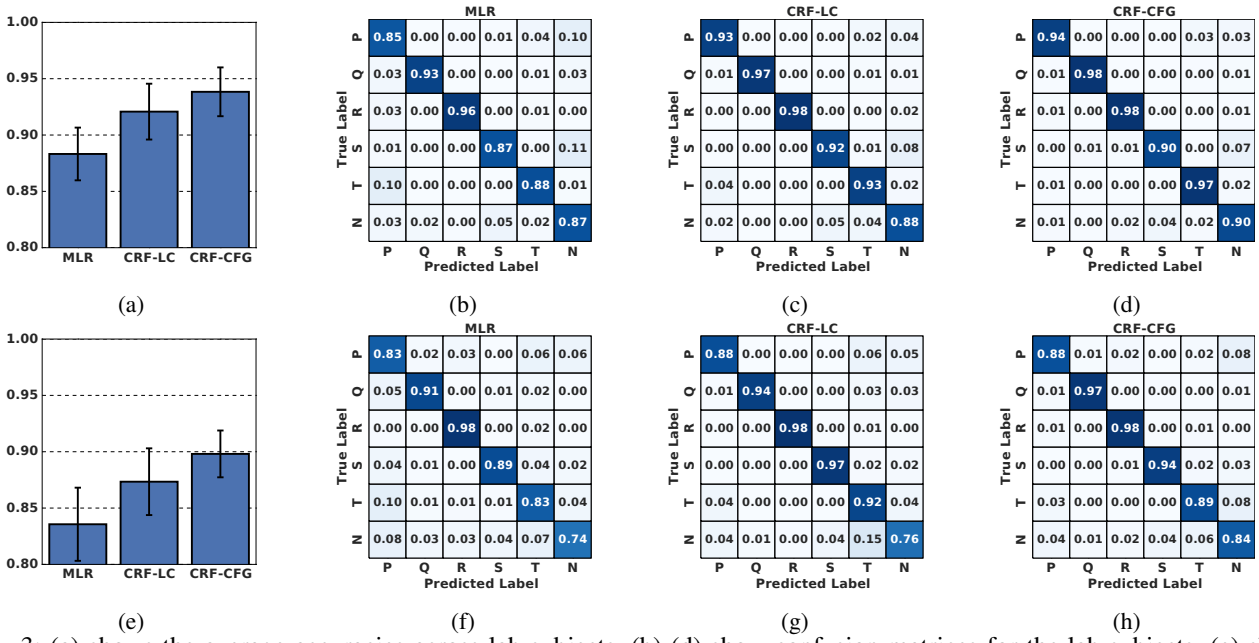

Fig. 3: (a) shows the average accuracies across lab subjects, (b)-(d) show confusion matrices for the lab subjects, (e) shows the average accuracies across field subjects, and (f)-(h) show confusion matrices for the field subjects.

same validation protocol was used to tune hyper-parameters as was used in the lab protocol.

V. Results

The leave-one-subject-out prediction accuracy for each model averaged across subjects is shown in Figure 3(a) for the lab data. Error bars show one standard error calculated with respect to the subjects. These results show that the CRF-CFG model outperforms the CRF-LC model by a margin of 1.8% for a relative error reduction of 22.2%. These results are significant at the $p = 0.05$ using a paired t-test with Bonferroni correction. Additionally, the confusion matrices for each method are shown in Figures 3(b)-(d). The results show that the CRF-CFG method is at least as good as the linear chain CRF model for most peak types, and provides significantly improved performance for T-waves.

The average classification accuracies on the field subjects are shown in Figure 3(e). The confusion matrices for the field setting are shown in Figures 3(f)-(h). We can see that while the CRF-CFG model does not have uniformly better performance for all peak types compared to the linear chain CRF in the field setting, it does have superior overall performance obtaining a 2.2% improvement in accuracy for a relative error reduction of 19.5%. These results are significant at the $p = 0.1$ level using a paired t-test with Bonferroni correction. We attribute this drop is significance, at least in part, to the much smaller amount of labeled field data. We also note that while there is a drop in performance between the lab and field data sets due to the strict across-subjects protocol used, the overall performance gap between the CRF-CFG and CRF-LC models is consistent in both the lab and field settings.

VI. Conclusions, Discussion, and Future Work

We have presented a new approach to ECG morphological structure extraction based on the use of conditional random field context free grammar models originally developed for probabilistic parsing within the natural language processing community. We have developed a robust grammar for parsing noisy wireless ECG data that allows us to represent some of the higher-order temporal structure contained in the data when compared to the previous linear chain CRF approach, which was based on a first-order Markov assumption. The CRF-CFG model achieves relative error reductions of 22.2% (lab) and 19.5% (field) when compared to the linear chain CRF model. These results further close the gap between the performance of methods for extracting morphological structure from low-cost wireless ECG sensor data and the performance of methods for extracting morphological structure from clinical ECG data.

One trend of note is that the CRF-CFG outperforms the linear chain CRF at classifying T waves on the lab data, but not on the field data. This appears to be explained by a number of extremely noisy samples in the field data. In these cases, the linear chain CRF tends to label many of the peaks as T waves resulting in high recall, but low precision. The CRF-CFG, on the other hand, chooses a single peak to call the T wave. While the choice is sometimes incorrect, the predicted structure makes much more sense than having a number of adjacent T waves. The result is that the CRF-CFG model has lower T wave performance, but higher overall accuracy.

It is important to note that the improved performance obtained by the CRF-CFG model comes at an increased computational cost. The linear chain CRF requires only linear time during inference while, in general, the CRF-

CFG model requires cubic time. In the experiments we present, inference is run on smaller sub-sequences, and not on full data sequences. The average number of peaks per sub-sequence is approximately 20, so the additional compute time required by the CRF-CFG model is not prohibitively large. In practice, cubic time inference would be prohibitive if applied to long input sequences. This issue could be overcome by using the linear chain model to provide a fast initial parse while identifying candidate peaks with highly confident predictions. The candidate peak locations with highly confident predictions could then be used to partition the data into smaller segments for parsing with the CRF-CFG model.

There are many possibilities to expand on the approach presented here. First, the CRF-CFG model can use much more complex features for higher-level productions than the simple temporal features we have used. For example, shape information for intervals between pairs of peaks (as opposed to centered around single peaks) could be extracted using sparse coding and used as features. Alternatively, if we believe that there are recurring, recognizable types of noise, we can easily instantiate multiple types of noise labels. There is also tremendous freedom to define improved grammars. Of particular interest are beat-based grammars that add production rules to explicitly group labeled ECG peaks together into complete cardiac cycles.

Acknowledgments: This work was partially supported by the National Institutes of Health under awards 1U54EB020404-01, and the National Science Foundation under Grant No. 1350522. The authors would like to thank Robert Malison, Gustavo Angarita, Edward Gaiser and the staff at the Connecticut Department of Mental Health and Addiction Services for data collection.

REFERENCES

[1] M. Malik, "Heart rate variability," *Annals of Noninvasive Electrocardiology*, vol. 1, no. 2, pp. 151–181, 1996.

[2] E. Ertin, N. Stohs, S. Kumar, A. Raij, M. al'Absi, and S. Shah, "Autosense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field," in *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*, 2011, pp. 274–287.

[3] E. Haapalainen, S. Kim, J. F. Forlizzi, and A. K. Dey, "Psychophysiological measures for assessing cognitive load," in *Ubiquitous computing, Proceedings of the 12th ACM international conference on*, 2010, pp. 301–310.

[4] J. Pärkkä, M. Ermes, P. Korpipää, J. Mäntyjärvi, J. Peltola, and I. Korhonen, "Activity classification using realistic data from wearable sensors," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 10, no. 1, pp. 119–128, 2006.

[5] G. Valenza, M. Nardelli, A. Lanata, C. Gentili, G. Bertschy, R. Paradiso, and E. P. Scilingo, "Wearable monitoring for mood recognition in bipolar disorder based on history-dependent long-term heart rate variability analysis," *Biomedical and Health Informatics, IEEE Journal of*, vol. 18, no. 5, pp. 1625–1635, 2014.

[6] A. Gaggioli, G. Pioggia, G. Tartarisco, G. Baldus, D. Corda, P. Cipresso, and G. Riva, "A mobile data collection platform for mental health research," *Personal and Ubiquitous Computing*, vol. 17, no. 2, pp. 241–251, 2013.

[7] S. Hu, Z. Shao, and J. Tan, "A real-time cardiac arrhythmia classification system with wearable electrocardiogram," in *Body Sensor Networks, Proceedings of the 2011 International Conference on*, 2011, pp. 119–124.

[8] A. Natarajan, A. Parate, E. Gaiser, G. Angarita, R. Malison, B. Marlin, and D. Ganesan, "Detecting cocaine use with wearable electrocardiogram sensors," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, 2013, pp. 123–132.

[9] ECGPedia, "Myocardial infarction," http://en.ecgpedia.org/index.php?title=Myocardial_Infarction.

[10] A. Magnano, N. Talathoti, R. Hallur, D. Jurus, J. Dizon, S. Holleran, B. D. M., E. Collins, and H. Garan, "Effect of acute cocaine administration on the QTc interval of habitual users," *The American journal of cardiology*, vol. 97, no. 8, pp. 1244–1246, 2006.

[11] K. Levin, M. Copersino, D. Epstein, S. Boyd, and D. Gorelick, "Longitudinal ECG changes in cocaine users during extended abstinence," *Drug Alcohol Depend*, vol. 95, no. 1-2, pp. 160–163, 2008.

[12] A. Natarajan, E. Gaiser, G. Angarita, R. Malison, D. Ganesan, and B. Marlin, "Conditional random fields for morphological analysis of wireless ecg signals," in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2014, pp. 370–379.

[13] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[14] J. R. Finkel, A. Kleeman, and C. D. Manning, "Efficient, feature-based, conditional random field parsing." in *ACL*, vol. 46, 2008, pp. 959–967.

[15] F. Alamudun, J. Choi, R. Gutierrez-Osuna, H. Khan, and B. Ahmed, "Removal of subject-dependent and activity-dependent variation in physiological measures of stress," in *Pervasive Computing Technologies for Healthcare, Proceedings of the 6th International Conference on*, 2012, pp. 115–122.

[16] J.-H. Hong, J. Ramos, and A. K. Dey, "Understanding physiological responses to stressors during physical activity," in *Ubiquitous Computing, Proceedings of the 2012 ACM Conference on*, 2012, pp. 270–279.

[17] R. Bouhenguel and I. Mahgoub, "A risk and incidence based atrial fibrillation detection scheme for wearable healthcare computing devices," in *Pervasive Computing Technologies for Healthcare, Proceedings of the 6th International Conference on*, 2012, pp. 97–104.

[18] R. Bouhenguel, I. Mahgoub, and M. Ilyas, "An energy efficient model for monitoring and detecting atrial fibrillation in wearable computing," in *Body Area Networks, Proceedings of the 7th International Conference on*, 2012, pp. 59–65.

[19] S. M. Hossain, A. A. Ali, M. M. Rahman, E. Ertin, D. Epstein, A. Kennedy, K. Preston, A. Umbricht, Y. Chen, and S. Kumar, "Identifying drug (cocaine) intake events from acute physiological response in the presence of free-living physical activity," in *Proceedings of the 13th international symposium on Information processing in sensor networks*, 2014, pp. 71–82.

[20] R. Jané, A. Blasi, J. García, and P. Laguna, "Evaluation of an automatic threshold based detector of waveform limits in holter ecg with the qt database," in *Computers in Cardiology 1997*, 1997, pp. 295–298.

[21] J. P. Martínez, R. Almeida, S. Olmos, A. P. Rocha, and P. Laguna, "A wavelet-based ecg delineator: evaluation on standard databases," *Biomedical Engineering, IEEE Transactions on*, vol. 51, no. 4, pp. 570–581, 2004.

[22] J. Pan and W. J. Tompkins, "A real-time qrs detection algorithm," *Biomedical Engineering, IEEE Transactions on*, vol. 32, no. 3, pp. 230–236, 1985.

[23] W. Zong, G. Moody, and D. Jiang, "A robust open-source algorithm to detect onset and duration of qrs complexes," in *Computers in Cardiology, 2003*, 2003, pp. 737–740.

[24] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[25] D. W. Hosmer Jr and S. Lemeshow, *Applied logistic regression*. John Wiley & Sons, 2004.

[26] K. Lari and S. J. Young, "The estimation of stochastic context-free grammars using the inside-outside algorithm," *Computer speech & language*, vol. 4, no. 1, pp. 35–56, 1990.

[27] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," in *Journal of Machine Learning Research*, 2005, pp. 1453–1484.

[28] A. C. Müller and S. Behnke, "pystruct - learning structured prediction in python," *Journal of Machine Learning Research*, vol. 15, pp. 2055–2060, 2014. [Online]. Available: http://jmlr.org/papers/v15/mueller14a.html

[29] E. Billauer, "Peak detection," http://billauer.co.il/peakdet.html.