Highly accurate real-time decomposition of single channel intramuscular EMG(minor revison)

Tianyi Yu, Konstantin Akhmadeev, Eric Le Carpentier, Yannick Aoustin, Dario Farina

Highly accurate real-time decomposition of single channel intramuscular EMG(minor revison)

Tianyi Yu, Konstantin Akhmadeev, Eric Le Carpentier, Yannick Aoustin, Dario Farina, *Fellow, IEEE*

***Abstract*—*Objective:* Real-time intramuscular electromyography (iEMG) decomposition, as an identification procedure of individual motor neuron (MN) discharge timings from a streaming iEMG recording, has the potential to be used in human-machine interfacing. However, for these applications, the decomposition accuracy and speed of current approaches need to be improved. *Methods:* In our previous work, a real-time decomposition algorithm based on a Hidden Markov Model of EMG, using GPU-implemented Bayesian filter to estimate the spike trains of motor units (MU) and their action potentials (MUAPs), was proposed. In this paper, a substantially extended version of this algorithm that boosts the accuracy while maintaining real-time implementation, is introduced. Specifically, multiple heuristics that aim at resolving the problems leading to performance degradation, are applied to the original model. In addition, the recursive maximum likelihood (RML) estimator previously used to estimate the statistical parameters of the spike trains, is replaced by a linear regression (LR) estimator, which is computationally more efficient, in order to ensure real-time decomposition with the new heuristics. *Results:* The algorithm was validated using twenty-one experimental iEMG signals acquired from the tibialis anterior muscle of five subjects by fine wire electrodes. All signals were decomposed in real time. The decomposition accuracy depended on the level of muscle activation and was $>$90% when less than 10 MUs were identified, substantially exceeding previous real-time results. *Conclusion:* Single channel iEMG signals can be very accurately decomposed in real time with the proposed algorithm. *Significance:* The proposed highly accurate algorithm for single-channel iEMG decomposition has the potential of providing neural information on motor tasks for human interfacing.**

***Index Terms*—Hidden Markov models, Bayes methods, Recursive estimation, Deconvolution, Electromyography decomposition, real-time decomposition, penalization.**

I. Introduction

THE electromyogram (EMG) is the electrical expression of skeletal muscle fibers during a muscle contraction. This activity originates from the neural excitation of the motor neurons (MNs) in the spinal cord. The identification of individual MN discharge timings from EMG is termed EMG decomposition [1]. The information obtained by EMG decomposition has been classically used for the diagnosis of neuromuscular disorders [2], analysis of muscle physiology [3], studies of central strategies for motor control [4], [5] as well as, more recently, for human-machine interfacing [6], [7], [8]. These applications, and especially the latter one, could greatly benefit from a highly accurate real-time decomposition method.

In the last decades, numerous approaches to EMG decomposition methods have been proposed. A significant group of strategies perform waveform detection and clustering with subsequent template matching [9], [10], [11], [12], [13], as well as validation, merging and pruning of the obtained spike trains [14], [15]. A number of studies specifically addressed resolution of action potential superpositions in such strategies [16], [17]. Another large family of EMG decomposition algorithms is based on blind source separation (BSS) of high-density multichannel data [18], [19], [20], [21]. Other types of algorithms include deep-learning [22] and Bayesian filtering [23]. Spike sorting algorithms address a similar problem of automatic extracellular recordings annotation [24], [25], [26]. Most of these methods, similar to EMG decomposition, are based on waveform clustering with subsequent template matching and, in some cases [25], resolution of superpositions.

However, most of the existing methods, both in the decomposition and spike sorting, work only in an offline manner, i.e. they require a computational time greater than the duration of the signal, or require a batch of signal that is longer than the acceptable response delay in human-machine interfaces [27]. Few exceptions are based on iterative adaptation of blind source separation for surface EMG [28], [29], and iterative clustering for intramuscular EMG [30]. However, the former only works with high-density recordings and does not adapt to ~~the drift in motor unit action potentials~~ the variation of motor unit action potentials over time , while the latter lacks R1.4

TABLE I: Main notations of Hidden Markov Model (HMM) and Bayes filter

Y	iEMG signal
H	Vector of MU action potentials shapes
Ω	Set of indexes of all MUs
A	Set of indexes of active MUs
T	Sawtooth sequences
S	Activation scenario
W	White noise
ℓ_{IR}	Maximum MUAP length
$\Theta = [t_0, \beta]$	Vector containing the discrete Weibull distribution parameters: the location parameter and the concentration parameter
t_{R}	Shifting parameter of discrete Weibull distribution, that is the refractory period
Pr	Probability
w.p.	"with probability"
$Y[n]$	iEMG signal at time index n
Y^n	Vector containing the signal from time index 1 to n
$\|^n$	"...given Y^n"
$\mathsf{Pr}(T[n] = t[n])$	Probability of the sawtooth sequences at time index n being equal to a value $t[n]$. For all elements of the state vector, the uppercase symbols (T, A, S) denote random variables, while the lowercase ones (t, a, s) stand for their realizations
n_{path}	Number of decomposition paths (scenarios) in algorithm

the resolution of temporal overlaps between action potentials. Existing real-time spike sorting algorithms, too, require multichannel input and either do not address superpositions [31] or the action potential waveform drift [32].

In our previous work, a Bayesian filtering approach for single channel iEMG decomposition that is following a time sequential approach (a sample-by-sample processing), was proposed [33], [23]. Moreover, a recruitment model for MUs was added [23] so that the algorithm could be adapted to a varying number of active MUs. Recently, we implemented this algorithm on a parallel computation framework [34] and achieved the real-time performance on a range of intramuscular signals. However, the real-time constraints led to a decrease in accuracy with respect to the original offline version. In this paper, our method is substantially extended to reach accuracy comparable to its offline implementation but maintaining the real-time performance.

In the following, our previous method, including the Hidden Markov Model (HMM) of iEMG and the Bayesian filtering procedure estimating the parameters of MUs will be briefly reviewed in Section II. Limitations of this approach that impact the accuracy when real-time constraints are posed, will also be analyzed in Section II. Then, in Section III, substantial changes in the algorithm will be proposed with the aim to boost its accuracy. Further, experimental iEMG signals and an evaluation protocol will be described in Section IV. Results of the experimental signals decomposition will be illustrated and analyzed in Section V. Finally, we will present our conclusions and future work in Section VI.

II. Review and Analysis

In this section, a short overview of the previous proposed GPU-implemented (Graphical Processing Unit) ~~on-line~~ real-time single channel iEMG decomposition algorithm, including the HMM of iEMG, the Bayes filtering, and the acceleration methods, will be provided. This description is necessary for the development of the new algorithm (Section III). Furthermore, the decomposition performance of the previous method will also be analyzed in order to point out its limitations.

A. Hidden Markov Model

1) Generation of the EMG signal: The *motor unit* (MU), as the smallest functional unit of muscle contraction, comprises a MN and the muscle fibers innervated by its axon (muscle unit). When receiving input from spinal and supraspinal circuits, the MN generates trains of *action potentials* (spikes) that propagate along the MN axon and reach the muscle fibres through the neuromuscular junction, causing the fiber depolarisation. The excited muscle fibers generate a compound potential, referred to as *motor unit action potential* (MUAP). Multiple MUs located nearby the electrodes contribute their MUAP trains to the overall interference EMG signal. Generally, the inter-spike intervals (ISI) reveal a certain regularity and have a physiological lower bound, called *refractory period*.

2) Observation model of HMM: Based on the EMG generation principles, a HMM of EMG was proposed in [33], [23]. The observation equation of HMM was derived from the linear model of EMG [35], [36]:

$$Y[n] = \sum_{i\in\Omega} \varphi_i(S[n])\, H_i[n] + W[n] \tag{1}$$

- n is the discrete time index;
- i is the index of MU;
- Ω is the set of indexes of MUs, including both active and inactive ones;
- Y denotes the observed iEMG signal;
- S represents the activation scenario, comprising two elements: $S[n] = (A[n], (T_j[n])_{j\in A[n]})$. $A[n]$ is the set of indexes of active MUs at time n; $(T_j[n])_{j\in A[n]}$ describes the time passed since its previous spike, for the active MU j;
- $\varphi_i(s)$, for each realization $s = (a, (t_i)_{i\in a})$, is a row vector of size ℓ_{IR} with all components equal to zero; except, if $i \in a$ and $t_i < \ell_{\text{IR}}$, the component in position $t_i + 1$ is equal to 1.
- H represents the MUAP waveform with finite length ℓ_{IR};
- W represents the independent identically distributed white noise sequence with unknown variance v;

3) State vectors of HMM: The statistics of the ISI for each MU are modelled by the discrete Weibull distribution [33]. This distribution is defined by the vector $\Theta_i[n]$ containing two parameters: a location parameter $t_{0_i}[n]$ and a shape parameter $\beta_i[n]$. Thus, the state vectors of HMM are as follows:

- $S[n] = (A[n], (T_i[n])_{i\in A[n]})$ the activation scenario,
- $H[n] = (H_i[n])_{i\in\Omega}$ the MUAP waveforms,
- $\Theta[n] = (\Theta_i[n])_{i\in\Omega}$ the inter-spike law parameters.

4) Transition laws of HMM: $H_i[n]$ and $\Theta_i[n]$ are firstly assumed to be constant in time. Their transition laws are as follows:

$$H_i[n+1] = H_i[n] \tag{2}$$
$$\Theta_i[n+1] = \Theta_i[n] \tag{3}$$

In practice, the steady changes of $H_i[n]$ and $\Theta_i[n]$ are considered in the Bayes filter (see II-B2).

The transition law for $S[n] = (A[n], (T_i[n])_{i\in A[n]})$ follows two transition models: the renewal model and the recruitment model, respectively for the two components $T_i[n]$ and $A[n]$.

In the renewal model, for each $i \in A[n+1] \cap A[n]$, given $\Theta_i[n]$, we have the transition distribution of $T_i[n]$:

$$T_i[n+1] = \begin{cases} 0 & \text{w.p. } r(T_i[n]+1, \Theta_i[n]) \\ T_i[n]+1 & \text{w.p. } 1 - r(T_i[n]+1, \Theta_i[n]) \end{cases} \tag{4}$$

where $r(\cdot)$ is the hazard rate function of the Discrete Weibull distribution [37]. Moreover, we should notice that $r(t, \Theta_i[n]) = 0$, if t is lower than the *refractory period* t_{R}, due to the physiological restriction.

In the recruitment model, the recruitment mechanism is interpreted as the variation of $A[n]$, the set of indexes of active MUs at time n. Given $S[n]$, we have:

$$A[n+1] = \begin{cases} A[n] \setminus i & \text{w.p. 1, if } T_i[n] = t_{\text{I}} \\ A[n] \cup i & \text{w.p. } \frac{\lambda}{\text{card}(\bar{A}[n])}, \text{ if i} \notin A[n] \\ A[n] & \text{w.p. } 1-\lambda \end{cases} \quad (5)$$

where $\text{card}(\bar{A}[n])$ is the number of inactive MUs. An i-th active MU is derecruited when $T_i[n]$ reaches a predefined limit t_{I}; a random inactive MU is recruited with predefined constant probability λ and initialized with $T[n] = 0$. Thus, $1-\lambda$ is the probability if the active MUs keep the same as before.

Given the known observed signal $Y[n]$, the rough initial waveforms $H[0]$, and the transition laws, the Bayes filter presented in the following subsection is applied for the recursive estimation of the unknown state vectors. R1.2

B. Bayes filter

1) Principles and estimations: With increasing time, the Bayes filter [33], [23], [34] recursively estimates the state vector of the HMM defined in II-A3. The posterior probability functions of the state vectors are known:

- The probability density function (PDF) of $\Theta[n]$ given S^n, H and Y^n. The regularity of ISI only depends on the state vector S^n. Furthermore, with the assumption of independence in MU activity in [33], this PDF can be written as the product of all terms $\mathsf{Pr}(\Theta_i[n] | S^n)$, for all $i \in A[n+1]$ ~~$\Theta_i[n]$ given S^n~~. The expected value of Θ_i given S^n, noted $\hat{\theta}_{i,S^n}$, is estimated by a recursive maximum likelihood (RML) estimation proposed in [23]. R1.5
- The PDF of $H[n]$ given S^n and Y^n. The expected value of this PDF, noted $\hat{H}^{|n}_{S^n}$, is estimated by a least-mean-square (LMS) filter. And the observation prediction denoted as $\hat{Y}^{|n-1}_{S^n}$ and its variance denoted as v_{S^n} are also provided by the LMS filter. Details and mathematical derivation of this procedure can be found in [34].
- The probability mass function (PMF) of S^n given Y^n. As presented in [23], the PMF of scenario is derived from an update-prediction scheme. For all possible realizations s^n of S^n, the update step is:

$$\mathsf{Pr}^{|n}(S^n = s^n) \propto \mathsf{Pr}^{|n-1}(S^n = s^n)\ g(Y[n] - \hat{Y}^{|n-1}_{s^n}, v_{s^n}) \quad (6)$$

where $g(.,v)$ is a zero-mean and variance v Gaussian PDF. And the prediction step is defined as:

$$\begin{aligned} &\mathsf{Pr}^{|n}(S^{n+1} = s^{n+1}) = \mathsf{Pr}^{|n}(S^n = s^n) \times \\ &\mathsf{Pr}(A[n+1] = a[n+1] \mid S[n] = s[n]) \times \\ &\prod_{i \in A[n+1]} \mathsf{Pr}(T_i[n+1] = t_i[n+1] \mid S^n = s^n) \end{aligned} \quad (7)$$

where $\mathsf{Pr}(A[n+1] = a[n+1] \mid S[n])$ and $\mathsf{Pr}(T_i[n+1] = t_i[n+1] \mid S^n)$ depends respectively on the recruitment model and the renewal model presented in section II-A4.

2) Adaptivity: In order to adapt to the two non-stationary state vectors: inter-spike laws parameters Θ and impulse responses H, a sliding window method [38] was used in the estimation of the two state vectors. The estimations are mainly based on the last ℓ_∞ observations ($Y[n-\ell_\infty+1:n]$) and scenarios ($S[n-\ell_\infty+1:n]$), where ℓ_∞ is the maximum window length. More details about the adaptive estimation formulas can be found in [34].

3) Initialisation: As presented in [34], no MUs are assumed to be active at the beginning of the decomposition. Thus, the set of active MUs indexes $A[1]$ and the sawtooth sequence $T[1]$ are initialized as empty vector. Initial rough MUAP waveforms $\hat{H}^{|0}_{S^1}$, that can be manually or automatically extracted based on [39], [40], [41], are supposed to be known as prior knowledge. And the two inter-spike law parameters of active MUs $\hat{\theta}_{i,S^0}$: t_0 is typically initialized as $3t_{\text{R}} \sim 4t_{\text{R}}$; β is normally set to 2 $\sim$ 4. Finally, n_{path} initial S^1 are weighted with the same initial probability $\mathsf{Pr}^{|0}(S^1 = s^1)$.

C. Acceleration

Theoretically, the state vectors of the HMM can be estimated recursively by the Bayes filter. However, this approach has a very high computational cost in practice. As described in the renewal model of section II-A4, for each $i \in A[n+1] \cap A[n]$, the sawtooth sequence $T_i[n]$ bifurcates at each time index, if $T_i[n] > t_{\text{R}}$. This implies that the number of possible scenarios for S^n grows exponentially with time. Thus, an exhaustive estimation for all scenarios is impossible. Several approaches for path pruning were proposed in [34] to reduce the number of bifurcated scenarios:

- Limiting the number of kept paths is a conventional way to deal with the large number of scenarios. The n_{path} most probable scenarios are kept at every time index, where n_{path} is chosen as a trade-off between the computational cost and the decomposition performance.
- An iEMG signal is composed of multiple prominent action potentials and many time intervals only containing background noise. A function $Z[n]$ [34] was introduced to distinguish the two cases, in order to avoid bifurcations of $T_i[n]$ in the segmentation of noise. For the segments containing action potentials, the ~~on-line~~ real-time decomposition algorithm bifurcates active scenarios into multiple new ones and keeps the n_{path} most probable scenarios, while in the segments of background noise, only the state vectors of the kept scenarios are updated at each time index. R2.1
- The simultaneous occurrence of two or more spikes at exactly the same time instant was considered as extremely low probability and thus excluded from the possible scenarios. This reduces the maximal number of possible bifurcation from $n_{\text{path}} \times 2^{\text{card}(A[n+1])}$ to $n_{\text{path}} \times (\text{card}(A[n+1]) + 1)$.

Furthermore, in order to achieve an ~~on-line~~ real-time decomposition, the algorithm was implemented into Graphics Processing Unit (GPU) parallel computing. R2.1

D. Problems with ~~online~~ real-time decomposition

The previously proposed GPU-implemented ~~on-line~~ real-time single channel iEMG decomposition algorithm [34] showed high decomposition accuracy ($> 90\%$) for experimental signals with less than seven active MUs, and large decrease in performance for a greater number of active units. The performance degradation was mainly caused by limiting the number of kept paths, which was needed for ~~online~~ real-time implementation.

Limiting the number of kept paths implies retaining the n_{path} most probable scenarios at each time index and assuming that one of them is the correct decomposition result. However, this assumption is not always valid mainly because of the high similarity in MUAP waveforms, high similarity in kept scenarios, and over explanation of strong interference, such as noise and variation of MUAP waveforms.

1) High similarity in MUAP waveforms: Due to the lack of supplementary information from other channels, single channel decomposition algorithm switches occasionally between two similar MUAP waveforms.

2) High similarity in scenarios: We define highly similar scenarios, two scenarios that have the same active part in the last ℓ_∞ time instants ($S_i[n-\ell_\infty+1:n] = S_j[n-\ell_\infty+1:n]$, where i and j are the indices of kept scenarios).

Based on the adaptivity of Bayes filter discussed in section II-B2, for any two highly similar scenarios i and j, we have:

$$\hat{H}^{|n}_{S_i^n} \approx \hat{H}^{|n}_{S_j^n} \tag{8}$$

$$\hat{\theta}_{i,S_i^n} \approx \hat{\theta}_{i,S_j^n} \tag{9}$$

which means that the two scenarios have almost the same estimated state vectors. Thus, they have great probability to provide the same results in the following decomposition, which will reduces the diversity of kept scenarios and the robustness of the decomposition algorithm.

3) Over explanation of noise and variation of MUAP waveforms: Since the ~~on-line~~ real-time decomposition algorithm in [34] follows a strict time sequential manner, when acting on sampling points with high noise and variations in MUAP waveforms, it tends to find local optimal solutions, without considering the subsequent time samples. This phenomenon is termed over explanation.

In the example shown in Figure 1, a signal was decomposed with two possible scenarios. In the first one, only one MU fires at the beginning of the interval, while the second scenario represents the superposition of two MUs firing at close instants of time. The firing of the second MU at the time index $n-2$ in the second scenario is only introduced to over explain the noise from the time sample $n-2$ to n. However, when the signal is decomposed at time index n, without knowing the following observed signal, the second scenario could have bigger posterior probability than the first one, if they have the same prior probability.

In the decomposition of experimental signals with multiple MUs, this type of over explanation in form of superposition occurs frequently, which may result in the correct scenarios not being included in the n_{path} kept ones. The worst case occurs with multi-layers over explanation, where one MU firing over explains the noise, a second MU firing is used to correct the difference between the first MUAP waveform and the observed signal, and so on. This process can increase the number of errors substantially.

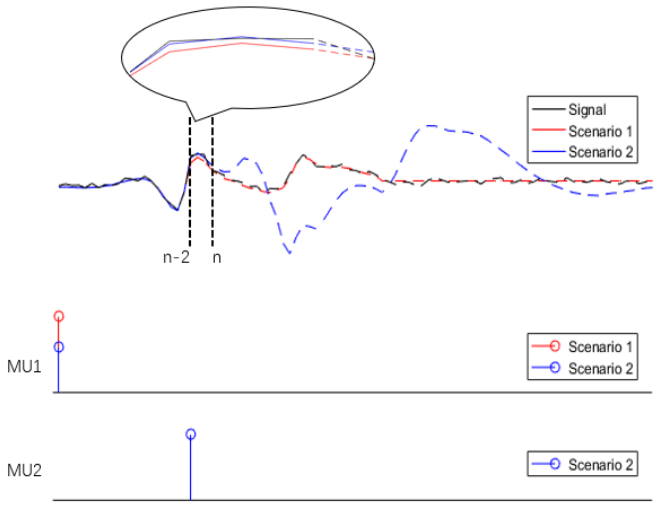

Fig. 1: Example of over explanation of noise. A signal was decomposed with two possible scenarios. Scenario 1 corresponds to the correct decomposition result, while Scenario 2 over explains the noise from time sample $n-2$ to n by introducing a firing for MU2.

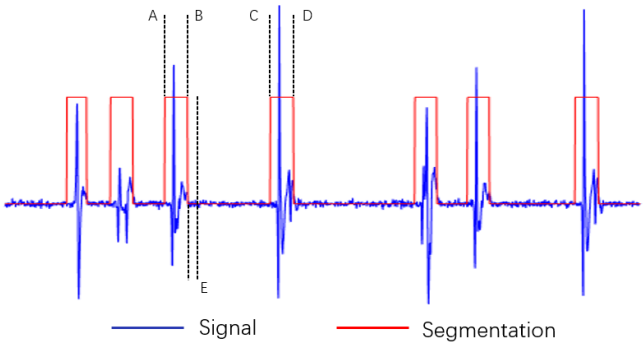

Fig. 2: Example of elimination of highly similar scenarios: in the segment AC, the high similar scenarios are eliminated at time index B, instead of checking and eliminating them at every time index. Example of the time instant for discarding scenarios with firing at the end of action potential segmentation: the scenario with a new firing at the end of segment AB will be discarded at time index E.

III. Highly Accurate ~~Online~~ Real-Time Decomposition

As shown in subsection II-D, there are substantial limitations in keeping scenarios only based on their posterior probability. In this section, we propose original approaches aimed at solving this problem and at keeping the decomposition speed within ~~online~~ real-time limits.

A. Limiting the inter-spike law parameters for highly similar MUs

Decomposition of the signal with highly similar MUAP waveforms is a common difficulty in single channel iEMG decomposition [42]. This can be partly counteracted by taking into account the regularity of firings in addition to MUAP waveform matching [34]. Here, we propose to limit one of the inter-spike law parameters β for MUs with similar MUAP shapes, in order to enhance the regularity in firing when MUAPs are very similar. β is typically limited to be more than 4.

B. Highly similar scenarios elimination

As described in subsection II-D2, highly similar kept scenarios reduce the diversity of the scenarios, leading to the decomposition being over susceptible to noise and variations in MUAP waveforms. Therefore, we propose to eliminate the highly similar kept scenarios with relative low posterior probability, and to only keep the one with highest probability.

However, an exhaustive elimination of similar scenarios at each time index is time consuming and inefficient for an R2.1 ~~on-line~~ real-time decomposition. According to the definition of high similarity of scenarios, we only need to make this elimination every ℓ_∞ time sampling. More practically, we discard the similar scenarios at the last time index of each action potentials segment, whose length is much smaller than ℓ_∞.

Based on the measurement presented in subsection II-C, the iEMG signal can be divided into sequential action potentials and noise segments with a detection function $Z[n]$. The R2.1 ~~on-line~~ real-time decomposition algorithm in [34] only bifurcates scenarios and keeps the n_{path} most probable ones in action potentials segments. Thus, the elimination at the last time index of each action potentials segment does not only ensure the diversity of kept scenarios, but also avoids the unnecessary calculation of similar scenarios in the next noise segment.

As Figure 2 shows, an iEMG signal comprises several short prominent action potentials waveforms, such as segments AB and CD, separated by segments of background noise, such as segment BC. In the segment AC, we only eliminate highly similar scenarios at time index B, instead of checking them at each time instant. In the noise segment BC, due to the elimination of similar scenarios, ~~less kept scenarios need to be updated~~ a small number of scenarios has been maintained R1.6 for the next iteration.

C. Interdiction of spike firing in the end of action potential segmentation

Sometimes, the kept scenarios may over explain the noise at the end of action potential segmentation by introducing new firings. However, a spike firing at the end of an active segment is highly improbable since the corresponding MUAP waveform cannot match the following noise segment. Thus, this type of scenarios could be discarded at the beginning of the following noise segment.

For example, in Figure 2, in the action potential segment AB, a kept scenario fires a spike near to the instant B. From time instant B to E, due to the gradually increasing difference between the estimated signal and the observed noise, the posterior probability of this scenario will show an obvious decline. Thus, the low probable scenarios are discarded at time index E. Generally, the length of BE is set to $\frac{\ell_{\text{IR}}}{5} \sim \frac{\ell_{\text{IR}}}{4}$.

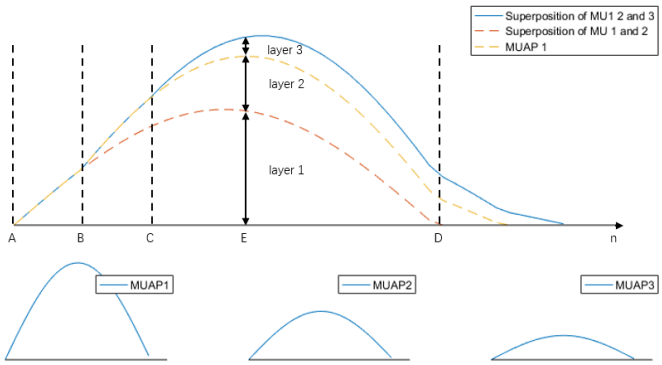

Fig. 3: Example of multiple layers superposition

D. Penalisation term of posterior probability

As presented in subsection II-D3, scenarios that over explain the observed sampling point containing strong interference by superimposing multiple MUAPs, always correspond to a locally high posterior probability, causing the possible exclusion of the correct decomposition result from the n_{path} kept scenarios. In order to resolve this issue, a penalisation term aiming at reducing local increase in posterior probability due to over explanation was added into the posterior probability formulation (6):

$$\mathsf{Pr}^{|n}(S^n = s^n) \propto \mathsf{Pr}^{|n-1}(S^n = s^n)\; g(Y[n] - \hat{Y}_{s^n}^{|n-1}, v_{s^n})\; \mathsf{Pr}_{\mathsf{pl}}{}^{|n}(S^n, \hat{H}_{S^{n-1}}^{|n-1}, \hat{\theta}_{S^n}) \tag{10}$$

where $\mathsf{Pr}_{\mathsf{pl}}{}^{|n}(S^n, \hat{H}_{S^{n-1}}^{|n-1}, \hat{\theta}_{S^n})$ is the penalisation term, calculated by the observed signal $Y[n]$, the new bifurcated scenario $s[n]$, the estimated MUAP waveforms, and estimated inter-spike law parameters.

The over explanation is normally in form of superposition composed of multiple layers of MUAP waveforms. For each layer of superposition, we penalize its gain of probability. The penalisation term can be written as the product of each layer penalisation:

$$\mathsf{Pr}_{\mathsf{pl}}{}^{|n}(S^n, \hat{H}_{S^{n-1}}^{|n-1}, \hat{\theta}_{S^n}) = \prod_{i \in \Upsilon[n]} \mathsf{Pr}_{\mathsf{pl},i}{}^{|n}(S^n, \hat{H}_{S^{n-1}}^{|n-1}, \hat{\theta}_{i,S^n}) \tag{11}$$

where $\mathsf{Pr}_{\mathsf{pl},i}{}^{|n}(S^n, \hat{H}_{S^{n-1}}^{|n-1}, \hat{\theta}_{i,S^n})$ is the penalisation term of each superposition layer; $\Upsilon[n]$ is the set of MUs superposition indexes.

An example of three layers of superposition is provided in Figure 3. Three MUs fire action potentials at the time indexes A, B, and C. Since the ~~online~~ real-time decomposition algo- R2.1 rithm respects a strict time sequential operation, the number of layers in the decomposition varies over time. If time $n \in AB$, index MU1 $\in \Upsilon[n]$; if $n \in BC$, indexes MU1 and MU2

$\in \Upsilon[n]$; if $n \in CD$, indexes MU1, MU2, and MU3 $\in \Upsilon[n]$. We define the order of MU indexes in $\Upsilon[n]$ according to its spike firing time. Thus, if $n \in CD$, $\Upsilon[n] = \{1, 2, 3\}$

Theoretically, the penalisation term should restrain MUAP waveforms from over-explaining the interference, and, at the same time, does not influence the regular decomposition of superimposed action potentials. In order to reach this objective, for all $i \in \Upsilon[n]$, the formula of $\mathsf{Pr}_{\mathsf{pl},i}{}^{|n}(S^n, \hat{H}^{|n-1}_{S^{n-1}}, \hat{\theta}_{i,S^{n-1}})$ is defined as:

$$\mathsf{Pr}_{\mathsf{pl},i}{}^{|n}(S^n, \hat{H}^{|n-1}_{S^{n-1}}, \hat{\theta}_{i,S^n}) = \frac{1}{\sqrt{2\pi v_{s^n}}} \exp(U^{|n}_{\text{pl},i}(S^n, \hat{H}^{|n-1}_{S^{n-1}}) \; N_{\text{pl},i} \; R_{\text{pl},i}(S^n, \hat{\theta}_{i,S^{n-1}}) \; \zeta_{\text{pl},i}(S^n)) \tag{12}$$

Where $U^{|n}_{\text{pl},i}(S^n, \hat{H}^{|n-1}_{S^{n-1}})$ is the function of penalisation unit; $N_{\text{pl},i}$ represents the layer coefficient; $R_{\text{pl},i}(S^n, \hat{\theta}_{i,S^{n-1}})$ denotes the regularity function of penalisation term; and $\zeta_{\text{pl},i}(S^n)$ is the attenuation function. In the following, the four terms will be described.

1) Function of penalisation unit: According to the posterior probability formula (6), the local increase in posterior probability could be originated from the prior probability, or the likelihood of observation, or both of them. However, based on the prior probability function (7), the inter-spike law distribution, and the formula (4), the prior probability of the scenario firing a new spike will be smaller than scenarios without the firing of a new spike. Thus, the likelihood of the observation $g(Y[n] - \hat{Y}^{|n-1}_{s^n}, v_{s^n})$ contributes to the overall temporary gain of the posterior probability. The likelihood function is expressed as:

$$g(Y[n] - \hat{Y}^{|n-1}_{s^n}, v_{s^n}) = \frac{1}{\sqrt{2\pi v_{s^n}}} \exp(-\frac{(Y[n] - \hat{Y}^{|n-1}_{s^n})^2}{2v_{s^n}})$$
$$\hat{Y}^{|n-1}_{S^n} = \psi(S[n]) \; \hat{H}^{|n-1}_{S^{n-1}} \tag{13}$$

where $\psi(s) = [\varphi_1(s), ..., \varphi_{\text{card}(\Omega)}(s)]$, $\text{card}(\Omega)$ denotes the number of MUs. The definition of $\varphi_1(s)$ is given in subsection II-A2. The estimated observation of the signal can be also written in form of multi-layers superposition:

$$\hat{Y}^{|n-1}_{S^n} = \sum_{i \in C[n]} \varphi_i(S_i[n]) \; \hat{H}^{|n-1}_{i,S^{n-1}} \tag{14}$$

With the formula (13) and (14), the contribution of each superposition layer for the posterior probability is clear. Thus, we have the penalisation unit of the i-th layer:

$$U^{|n}_{\text{pl},i}(S^n, \hat{H}^{|n-1}_{S^{n-1}}) = -\| \frac{(Y[n] - \sum_{k \in C[n], k=1}^{k=i} \varphi_k(S_k[n]) \; \hat{H}^{|n-1}_{k,S^{n-1}})^2}{2v_{s^n}} - \frac{(Y[n] - \sum_{k \in C[n], k=1}^{k=i-1} \varphi_k(S_k[n]) \; \hat{H}^{|n-1}_{k,S^{n-1}})^2}{2v_{s^n}} \| \tag{15}$$

where $\|.\|$ is the operator of norm.

2) Layer coefficient: Generally, multi-layers superposition scenarios better explain the signal $Y[n]$. This is because the large number of layers in the superposition easily reduces the difference between the estimated and the observed signal. However, with the growing number of layers, the generated superpositions are less and less probable. Thus, we introduce a penalization term on the number of layers. The coefficient of layer is defined as:

$$N_{\text{pl},i} = \Upsilon_{i,\text{order}} - 1 \tag{16}$$

where $\Upsilon_{i,\text{order}}$ is the order of the i-th MU in the $\Upsilon[n]$. Since the first layer of superposition is not considered as the cause of over explanation, its coefficient is supposed to be zero.

3) Regularity function: We penalize the posterior probability of superposition to prevent them from over explaining the interference in the signal, under the condition of not influencing the normal decomposition of superimposed MUAP waveforms. Therefore, the regularity parameters of spike trains are involved into regulating the penalization terms for different cases. We have the regularity function of penalization term:

$$R_{\text{pl},i}(S^n, \hat{\theta}_{i,S^{n-1}}) = 1 - \frac{\mathsf{Pr}(\Delta_i \mid \hat{\theta}_{i,S^{n-1}})}{\eta \; \mathsf{Pr}(\hat{t}_{0,i,S^{n-1}} \mid \hat{\theta}_{i,S^{n-1}})} \tag{17}$$

where $\mathsf{Pr}(\Delta|\Theta)$ is the probability mass function of the ISIs distribution, which is the discrete Weibull distribution in our algorithm; Δ_i is the length of last inter-spike of the i-th MU; $\hat{t}_{0,i,S^{n-1}}$ is the estimated location parameter of discrete Weibull distribution, representing the location of most of ISIs in the distribution; and η is the coefficient of regularity function. The coefficient η should be more than one, typically set to two in our decomposition.

The value of $R_{\text{pl},i}(S^n, \hat{\theta}_{i,S^{n-1}})$ ranges from $1 - \frac{1}{\eta}$ to 1. If Δ_i is far from $\hat{t}_{0,i,S^{n-1}}$, meaning that the superposition is probably the over explanation, $R_{\text{pl},i}(S^n, \hat{\theta}_{i,S^{n-1}})$ approaches to 1; On the contrary, if Δ_i is close to $\hat{t}_{0,i,S^{n-1}}$, $R_{\text{pl},i}(S^n, \hat{\theta}_{i,S^{n-1}})$ tends to $1 - \frac{1}{\eta}$.

4) Attenuation function: The local increase in posterior probability of scenarios that over explain the signal mainly occurs at the beginning of new layer of superposition, named penalized section. The attenuation function is to ensure the penalisation term keep working in the penalized section and reduce rapidly to one at the end of the section. We have the attenuation function:

$$\zeta_{\text{pl},i}(S^n)) = \begin{cases} 1 - (\frac{T_i[n]}{\rho_{i,\text{pl}}})^3 & \text{if } T_i[n] \leq \rho_{i,\text{pl}} \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

where $T_i[n]$ is the time passed since its previous spike of the i-th MU, defined in subsection II-A2; $\rho_{i,\text{pl}}$ is the length of penalized section, defined as the round value of the ratio between the maximum absolute value of $\hat{H}^{|n-1}_{S^{n-1}}$ and $\hat{H}^{|n-1}_{i,S^{n-1}}$, where $\hat{H}^{|n-1}_{S^{n-1}}$ is the estimated MUAP waveforms of all MUs and $\hat{H}^{|n-1}_{i,S^{n-1}}$ is the estimated MUAP of the i-th MU. Based on the decomposition results of previous algorithm, for each MU, the maximum absolute value of MUAP shapes is negatively correlated to the number of over explanation spikes. Therefore, the length of the penalized section is in an inverse ratio to the maximum absolute value of MUAP shapes.

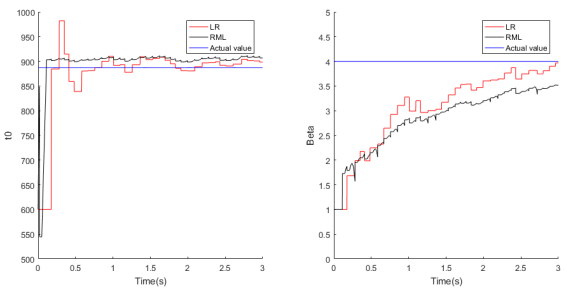

Fig. 4: Estimation of Weibull parameters with LR and RML estimator. The left figure is the estimation of location parameter t_0, while the right one is the estimation of shape parameter β [33].

E. Estimation of inter-spike law parameters

As presented in subsection II-B1, the inter-spike law parameters were estimated by the RML estimator proposed in [23]. R2.1 In the ~~on-line~~ real-time estimator RML, the inter-spike law parameters were estimated at each time index with censored data, in order to provide accurate estimation for the calculation of prior probability in formula (7). However, compared to other estimation methods based on complete data, this method is time consuming due to the recursive estimation at each ~~time~~ R1.7 ~~index~~ sample.

A number of algorithms have been proposed to estimate the discrete Weibull parameters with complete samples. Two methods proposed in [43], estimate parameters respectively based on the moments of samples and the proportion of samples. However, they present poor performance for a small number of samples. A maximum likelihood method proposed in [44] estimates recursively the parameters. When this method is implemented in the GPU parallel computation environment, it calls frequently the exponent function "double pow(double base, double exponent)" which requires high computational time. After comparing several estimation methods with inter-spikes of simulated iEMG signal (these results are not shown here), we selected a linear regression (LR) method proposed in [45]. ~~This method models the relation of the logarithm of samples proportion and a function related to cumulative density function, as well as weighing the regression residual based on Heteroscedastic Unreliability Weibull Estimation (HUWE).~~ The main idea of this LR method is to use stochastic models of the unreliability at each failure instant. The consequent heteroscedastic regression problem is solved by using Heteroscedastic Unreliability Weibull Estimation (HUWE) to weigh the regression residual. Additional mathematical R1.0 derivations can be found in [45]. Compared to the maximum likelihood estimator, this LR method is computationally more robust and efficient, because, the inversion of the approximate Hessian in the RML estimator may be close to singular. Details and mathematical derivation of this algorithm can be found in [45].

We compared the performance of RML and the selected LR estimator as follows. An active scenario S^n was generated with predefined inter-spike law parameters t_0 and β, based on the renewal model and recruitment model in subsection II-A3. Given S^n, the two estimators were used to identify the parameters. Figure 4 shows the estimation results of the two estimators, which indicate that the LR estimator with complete data provides almost the same estimates as the RML estimator. The maximum sliding window length in RML was set to 0.6 s.

In the decomposition, instead of estimating inter-spike law parameters with RML estimator, the LR estimator was implemented at the end of every action potential segment decomposition.

IV. Experimental protocols

A. Signals

~~Since the decomposition of simulated signals showed high performance in our previous work [34], we validated the new algorithm proposed in this paper only on experimental signals, which were more challenging.~~ Since the decomposition of simulated signals showed high performance in our previous work [34], the potential improvement in decomposition of simulated signals by the new proposed algorithm is limited. For these reasons, we do not report results on decomposition of simulated signals in this paper but we rather focus on the more challenging problem of decomposing experimental signals. R1.1

Two sets of experimental iEMG were acquired from the tibialis anterior (TA) muscle during ankle dorsiflexion tasks. The first set, which is the one we also used in [34], was recorded from a 26 years-old healthy man. In the recording, the subject performed five trials, each consisting of a 24-s isometric ankle dorsiflexion, following a cue on the screen. The force profile was trapezoidal with target force set to 20% or 30% of the maximal voluntary contraction (MVC) force. Additionally, in this study, we extended the validation to a greater number of experimental signals. The second set of recordings was acquired from four 20-25 years old subjects. Each subject was asked to perform four 24s trials of constant isometric contractions with target force set to, respectively, 5%, 10%, 15%, and 20% of MVC. Five minutes of rest after each trial were included in the experiment, to prevent fatigue.

All signals in both sets were recorded with fine-wire electrodes made of Teflon-coated stainless steel (50 um diameter; A-M Systems, Carlsborg, WA, USA). The EMG signals were R2.3 amplified, band-pass filtered between 100 Hz and 4 kHz, sampled at a frequency of 10kHz (OTBioelettronica MEBA amplifier), and subsequently down-sampled to 5 kHz for decomposition. The contraction force was measured by a calibrated load cell integrated in a custom-made footplate which firmly fixed the ankle. Force measurements from the load cell were pre-amplified and then recorded using the same amplifier, along with the EMG signals.

The new proposed ~~on-line~~ real-time iEMG decomposition R2.1 algorithm was used to decode the two sets of experimental signals and the results were compared with those obtained in [34]. The activation probability λ and the maximum inactive time t_I of the recruitment model were respectively set to 0.03 and $7t_{\mathrm{R}}$; The maximum window length ℓ_∞ was 1.4 s. The number of selected paths was set to 384.

TABLE II: Decomposition performance for the first set of experimental signals: 'Nb MUs' is the maximal number of MUs concurrently active in the signal; 'Nb spikes' denotes the overall number of spikes in the signal; 'Sup.' is the percentage of superposition, as defined in Section IV; 'Algorithm: P.D.' indicates the use of the parallel decomposition algorithm proposed in [34]; 'Algorithm: P.D.A.' denotes the use of the ameliorated parallel decomposition algorithm presented in this paper; 'Sens.' denotes the global sensitivity; 'Pred.' is the global predictivity.

Signal ∗	Duration (s)	Force (MVC%)	Nb MUs	Nb spikes	Sup.(%)	Algorithm: P.D.		Algorithm: P.D.A.	
						Sens. (%)	Pred.(%)	Sens. (%)	Pred.(%)
1.1	24	20	5	873	18.10	90.84	82.60	94.27	92.78
1.2	24	20	5	936	18.38	94.87	88.01	97.65	97.65
1.3	24	20	6	933	17.15	92.82	82.35	95.15	92.50
1.4	24	30	8	1282	28.39	86.04	76.70	90.48	91.41
1.5	24	30	8	1295	28.96	87.03	77.04	91.97	90.05

TABLE III: Decomposition performance for the second set of experimental signals. The meaning of indexes are the same as table II.

Signal ∗	Duration (s)	Force (MVC%)	Nb MUs	Nb spikes	Sup.(%)	Algorithm: P.D.		Algorithm: P.D.A.	
						Sens. (%)	Pred.(%)	Sens. (%)	Pred.(%)
2.1	24	20	8	1269	17.89	85.50	79.14	93.30	91.29
2.2	24	15	7	903	24.58	83.50	86.77	98.12	98.23
2.3	24	10	7	746	10.86	95.31	87.24	98.93	98.40
2.4	24	5	3	354	4.24	96.33	90.69	98.87	97.77
3.1	24	20	7	890	13.03	89.21	81.44	94.27	93.85
3.2	24	15	5	664	10.39	84.04	73.91	91.11	88.71
3.3	24	10	5	503	9.47	91.65	83.06	91.65	91.11
3.4	24	5	3	229	4.35	97.32	88.45	97.66	94.50
4.1	24	20	11	1194	27.55	53.18	43.61	83.95	84.60
4.2	24	15	11	1040	24.04	51.44	37.18	88.19	88.12
4.3	24	10	9	1010	24.95	84.75	73.67	90.10	91.83
4.4	24	5	5	481	14.35	94.39	85.34	96.88	95.69
5.1	24	20	10	1510	23.31	78.35	66.69	82.05	84.75
5.2	24	15	8	993	13.49	89.53	75.53	94.66	91.89
5.3	24	10	5	617	8.75	95.53	82.35	98.87	99.35
5.4	24	5	3	353	3.12	99.44	96.70	100	100

B. *Indexes of task and performance complexity*

1) Index of task complexity: As presented in [23], the superposition percentage was used to characterise the complexity of the decomposition task:

$$\text{Sup} = \frac{\text{Nb}_{\text{SUP}}}{\text{Nb}_{\text{SPIKES}}} \tag{19}$$

where $\text{Nb}_{\text{SPIKES}}$ is the number of spikes in the signal and Nb_{SUP} is the number of spikes involved in the superposition of MUAP waveforms. We considered a MUAP superimposed if there was at least one other MUAP within a margin of 3 ms (less than half of the average MUAP duration) around it.

2) Indexes of performance complexity: Results of automatic decomposition were evaluated with the reference spike trains identified by manual decomposition. The manual decomposition was provided off-line by an expert operator, who is not in the list of authors, using the publicly available decomposition software EMGLAB [46].

The global sensitivity and global positive predictivity were used to quantitatively evaluate the decomposition results. They were defined as follows: A MUAP was considered correctly identified (true positive) if the reference train contained a spike from the same MU within a margin of 1 ms around it. Consequently, the global sensitivity was defined as the overall number of correctly identified MUAPs from all MUs, divided by the overall number of spikes in the reference decomposition. Global positive predictivity was the number of correctly identified spikes divided by the overall number of spikes in the decomposition under evaluation.

Furthermore, an individual analysis of each MUAP train was also performed with "classification phase" indexes, including sensitivity, specificity and accuracy, as they are proposed in [47].

V. Results

All experimental signals were decomposed with the proposed algorithm programmed in C++ CUDA. All decompositions were performed on a Nvidia Tesla K80 GPU card with CUDA 9.0 and GCC 5.4.0 using single-precision floating-point format.

A. *Decomposition performance*

As shown in Table II, five experimental signals, including three recorded at 20% MVC and two recorded at 30% MVC, were decomposed respectively with the parallel decomposition algorithm (P.D.) proposed in [34] and the new algorithm (P.D.A.) proposed in this paper. For different experimental signals, the number of spikes ranged from 873 to 1295 and the percentage of superposition ranged from 18.10% to 28.96%. The decomposition results were ~~mainly~~ influenced by two factors used for the ANOVA analysis: the complexity of R1.11

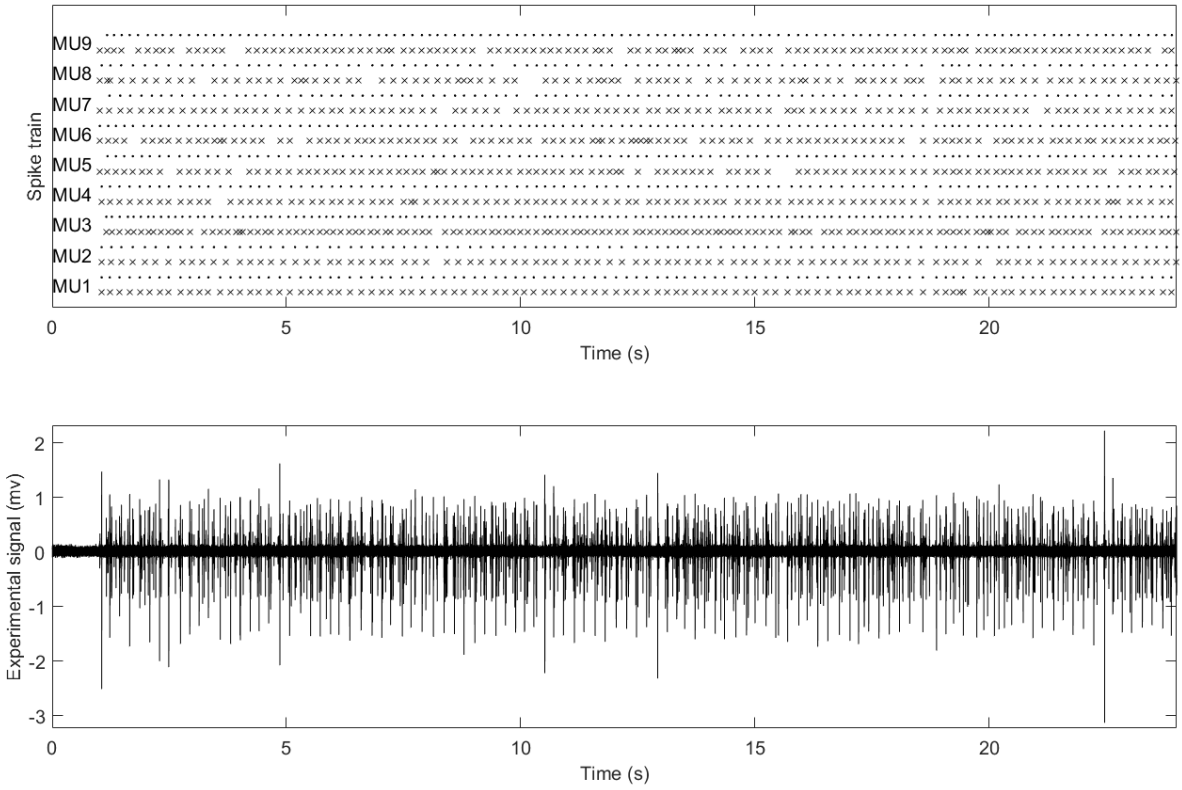

Fig. 5: Comparison of P.D.A. automatic (crosses, 'x') and reference (points, '.') decompositions (upper panel) and the experimental signal corresponding to signal 4.3 in Table II (lower panel).

the signal represented by the MVC value and the selected decomposition algorithm (P.D. or P.D.A.). In order to analyze the decomposition results, a two-way ANOVA (Analysis of variance, with P value 0.05) was applied for the first set of signals. Compared to the decomposition results of the previous P.D. algorithm, the results of the new P.D.A. algorithm showed a substantial increase in global predictivity (10% $\sim$ 15%, $F(1, 6)=5.99$, p=0.004) and a smaller but significant increase in global sensitivity (3% $\sim$ 4%, $F(1, 6)=5.99$, p=0.0131). Both the global sensitivity and predictivity of the five signals decomposed with the P.D.A. algorithm were above 90%. Furthermore, we noticed that the global sensitivity and predictivity decreased for signals with greater MVC ($F(1, 6)=5.99$, p=0.0022 for sensitivity; $F(1, 6)=5.99$, p=0.0148 for predictivity), due to the increase in decomposition task complexity, quantified by the number of MUs, spikes, and the superposition percentage. R2.2

It has to be noted that the results reported in Table II for the previous algorithm slightly differ from those reported in [34], since in this paper we did not tune any parameters, especially the lower limit for the inter-spike law parameters β (the same limits are set for the previous P.D. and the new P.D.A. algorithm).

Besides five experimental signals in Table II, 16 experimental signals were also automatically decomposed in order to validate the new P.D.A. algorithm. As shown in Table III, the 16 signals were divided in four groups. Each group contained four signals recorded from the same subject at 5%, 10%, 15%, and 20% MVC. For signals in the same group, the number of active MUs, number of spikes, and the percentage of superposition increased with increasing force level. An exception is the percentage of superposition of signal 2.2. Based on the manual decomposition results of this signal, we found that the high correlation between its 1st MU and 2nd MU led to the ~~abnormal~~ abnormally high superposition percentage. R1.13 For the 16 experimental signals, the number of MUs ranged from 3 to 11 and the percentage of superposition ranged from 3.12% to 27.55%. Since the force profile was constant, some of the signals showed a higher complexity compared to the ones in Table II.

For the decomposition performance of the 16 experimental signals in Table III, both the global sensitivity and predictivity of 13 of them with number of active MUs less than 10 were above 90%, while the global sensitivity and predictivity of the other three with 10 or 11 active MUs were above 80%. As for the first set of signals, a two-way ANOVA was applied to the results for statistical analysis. Consistent with the decomposition performance shown in Table II, the performances of the new P.D.A. algorithm (Table III) was much superior than for the previous P.D. algorithm ($F(1,$

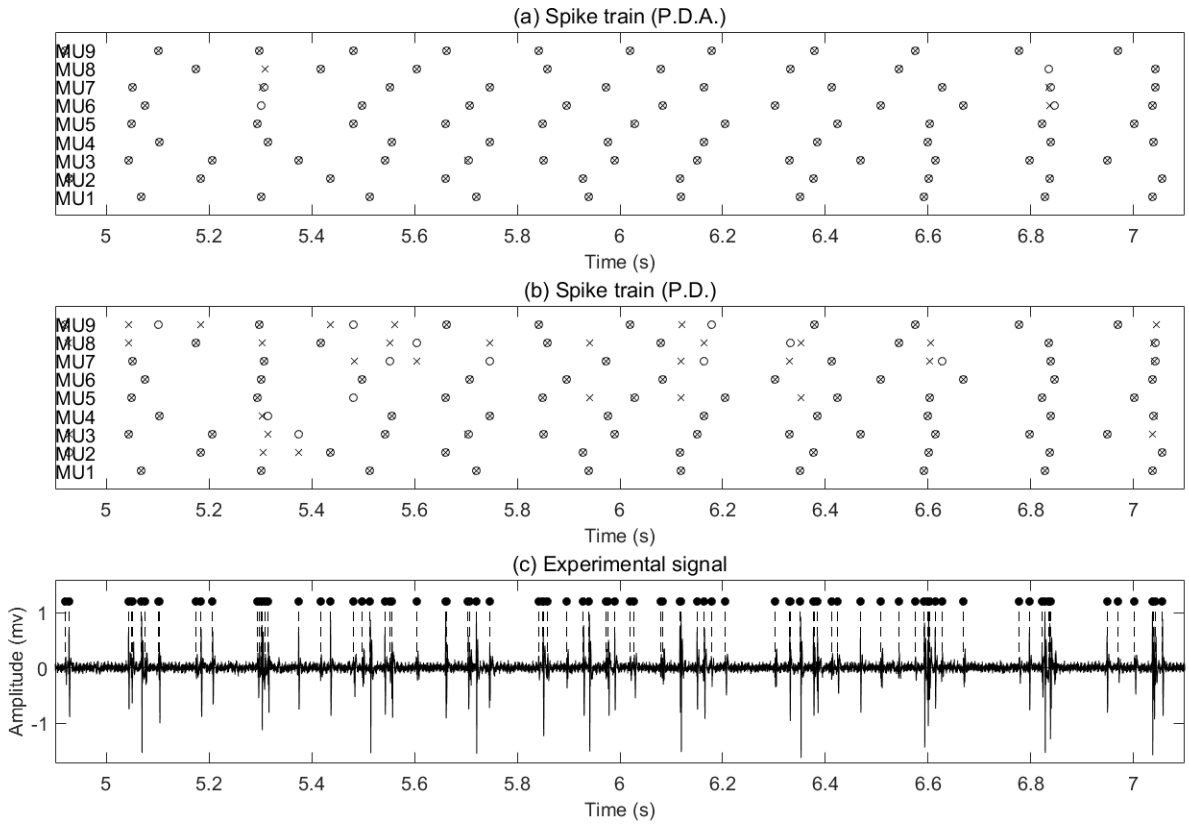

Fig. 6: An extract of the experimental signal decomposition shown in figure 5: (a) Comparison between the identified spike timings (crosses, 'x') by P.D.A. and the reference results (circles, '○'); (b) Comparison between the identified spike timings (crosses, 'x') by P.D. algorithm and the reference results (circles, '○'); (c) Experimental signal and the location of spikes in the reference result represented by black dot

R2.2 24)=4.26, p=0.0277 for sensitivity; $F(1, 24)=4.26$, p=0.0004 for predictivity). Furthermore, the decomposition performance significantly depended on the signal complexity as quantified by the MVC value ($F(3, 24)=3.009$, p=0.0159 for sensitivity; $F(3, 24)=3.009$, p=0.0343 for predictivity). A post-hoc LSD test (Fisher's least significant difference) identified significant differences for the sensitivity between signals at 5% MVC and 15% MVC, 5% MVC and 20% MVC, and 10% MVC and 20% MVC; and a significant difference for the predictivity between the signals at 5% MVC and 20% MVC. Therefore, the decomposition performance decreased with the increase in complexity.

In the following, we analyse in detail the results of the decomposition for the signal 4.3, as a representative case, which is one of the most complex signals we decomposed, having two very similar MUAP waveform.

A global view of the decomposition results is given in Figure 5. In the upper panel, nine decomposed spike trains (crosses) of each MU of P.D.A. are correlated with the reference results (points); in the lower panel, the decomposed iEMG signal is shown.

Figure 6 provides a detailed view of the decomposition results containing approximately two seconds of recorded signal. The first panel shows the comparison between the identified spike timings (crosses) by P.D.A. and the reference results (circles); in the second panel, the identified spike timings by P.D. algorithm correlate to the reference; the last panel shows the corresponding two seconds interval of recorded signal and the location of spikes in the reference result. Although two mistakes occured at 5.3 s and 6.83 s, P.D.A.

TABLE IV: Decomposition performance for the experimental signal decomposition shown in figure 5: for each MU, 'Sens.' denotes the sensitivity; 'Pred.' is the predictivity; 'Acc.' represents the accuracy.

MU	Sens.	Spec.	Acc.
MU1	99.07	99.75	99.67
MU2	92.93	99.39	98.70
MU3	92.96	98.98	98.06
MU4	94.34	99.14	98.59
MU5	91.38	99.38	98.38
MU6	89.83	99.14	97.95
MU7	82.35	97.98	96.30
MU8	80.21	97.54	95.79
MU9	86.29	98.89	97.22

Fig. 7: Nine MUAP shapes (manually-extracted dictionary) for the signal presented in Figure 5, and a comparison between the 7th one and the 8th one. The length of each MUAP waveform is 150 ms.

algorithm performed generally well, successfully processing several complex superpositions (see first panel of Figure 6). However, P.D. algorithm failed to process several superpositions containing small amplitude MUAP waveforms, where MUAP waveforms are illustrated in Figure 7. In addition, we note multiple problems in the decomposition with the P.D. algorithm, such as: over explanations at 5.05 s, 5.55 s, 5.95 s , 6.13 s, 6.36 s, and 6.60 s, and switches between the 7th MU and the 8th MU at 5.57 s, 5.60 s, and 5.74 s, were all corrected in the decomposition by the P.D.A. algorithm.

For the classification phase of P.D.A., Table IV shows the individual (per MU) performance indexes. Figure 7 shows the MUAP waveforms of nine MUs with a detailed comparison between MU 7 and 8. Based on Table IV and the Figure 7, we notice that the first four MUs were well classified, due to their larger MUAP waveforms, while the relative lower sensitivity for the 7^{th} and 8^{th} MU was caused by their smaller and similar MUAP waveforms, which may lead to the over explanation problem and switches with each other. Compared to the P.D. algorithm, the P.D.A. algorithm substantially improves the decomposition performance, but still cannot eliminate all classification mistakes.

The algorithm estimates the parameters of the inter-spike intervals distribution, used to calculate the firing rates. The corresponding firing rates are illustrated in Figure 8. Empirical ones were estimated as the inverse of the moving average of subsequent inter-spike intervals in the reference decomposition. The estimated ones of P.D.A. were calculated with the parameters t_0 and β estimated by the LR estimator. Based on the decomposed spike trains of P.D.A., we also estimated the parameters t_0 and β with RML estimator used in P.D. and calculated its firing rates. Considering the empirical firing rates as the correct ones, we calculated the NRMSE (normalized root mean square error) of the estimated firing rates for the RML estimator and the LR estimator using the formula:

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{i=n}(\hat{f}[i]-f[i])^2}}{\frac{1}{n}\sum_{i=1}^{i=n} f[i]} \qquad (20)$$

where $\hat{f}$ is an estimate of the firing rate made by RML or LR estimator; f denotes the actual firing rate calculated from the ~~ground truth~~ reference decomposition. Here, the RMSE (root mean square error) of the estimates were normalised by the average of the actual firing rates. Since the data were collected from a brief constant isometric contraction, the actual firing rates can be assumed virtually constant and the RMSE can be normalized by their average values. Table V reports the average of correct firing rates as well as the NRMSE by LR and RML estimators. The three highly coincident R1.17

TABLE V: Evaluation for Inter-spike laws parameters estimator through NRMSE: 'Aver.' represents the average of actual firing rates; 'LR.' denotes the NRMSE of LR estimator; 'RML.' is the NRMSE of RML estimator.

MU	Aver.	LR.	RML.
MU1	4.65	0.023	0.040
MU2	4.26	0.023	0.057
MU3	6.21	0.032	0.047
MU4	4.62	0.033	0.049
MU5	5.06	0.056	0.057
MU6	5.10	0.067	0.056
MU7	4.50	0.038	0.041
MU8	4.21	0.042	0.062
MU9	5.42	0.075	0.042

firing rates lines in Figure 8 and the NRMSE value shown in Table V demonstrate that the replacement of inter-spike law parameters estimator has little influence on the decomposition performance.

B. Decomposition time

Table VI provides the execution time for all experimental signals for P.D.A. and P.D.. The execution time was measured with the CUDA event API (Application Programming Interface). We notice that all signals, except the signal 5.1 with 10 MUs, could be decomposed in real time using 384 paths. The long decomposition time of signal 5.1 was due to the large maximum length of MUAP waveforms (around 10 ms), compared to others (6 ms $\sim$ 8 ms). However, this signal could still be decomposed in real time with the P.D.A. algorithm using 256 paths. The execution time was 23.82 s. Moreover, we notice that the execution time of P.D.A. is almost the same as the one of P.D. for all signals. This is because in the P.D.A., time saved from the inter-spike law parameters estimation was used to regulate the diversity of scenarios and penalize the posterior probabilities. R1.20

VI. Conclusion and Perspectives

In our previous works [33], [23], [34], a real-time decomposition algorithm based on a HMM of the EMG, using Bayesian filtering to estimate the unknown parameters of discharge series of MUs, and accelerated in the parallel computation environment, was proposed. Several iEMG signals with number of MUs up to ten were successfully decomposed. However, ~~that~~ the previous version of the algorithm required delicate manual parameter tuning before decomposition and showed relatively low performance for experimental signals with number of active MUs greater than six. R1.19

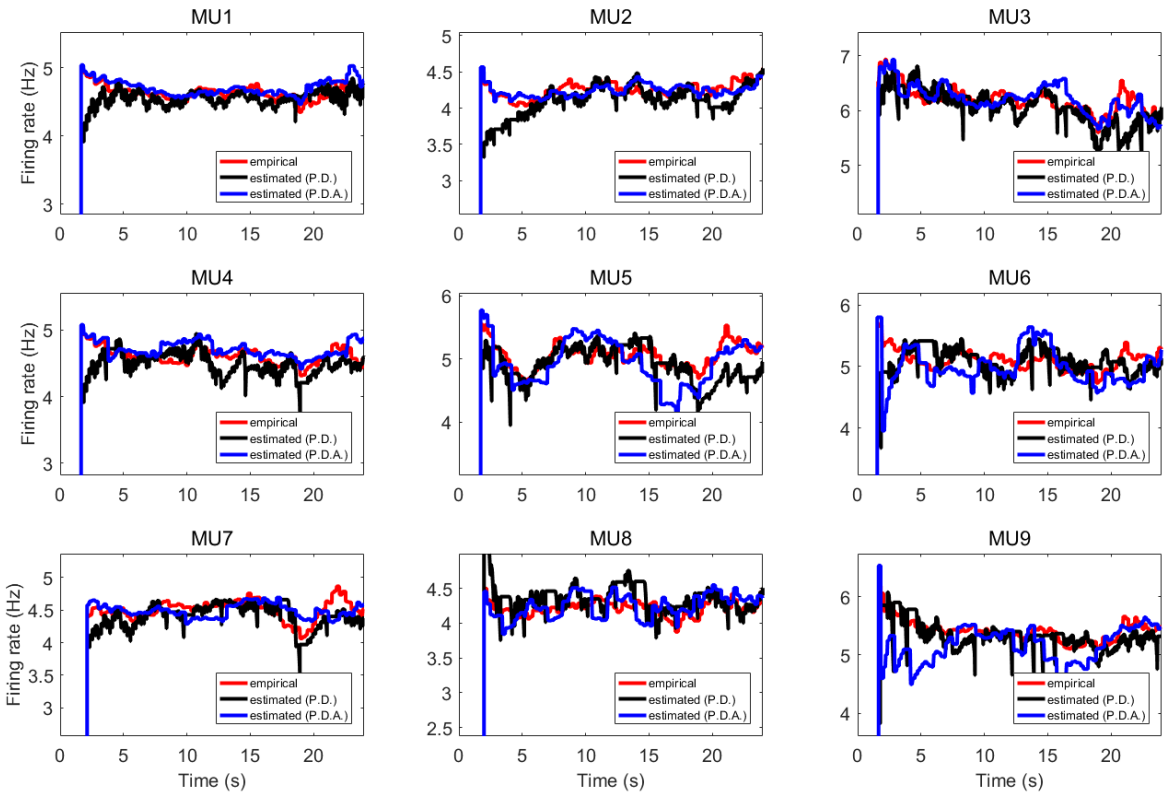

Fig. 8: Firing rates for the iEMG presented in figure 5: the red line (empirical) represents the firing rates estimated using reference decomposition; black line (estimated P.D.) represents the firing rates calculated via parameters of discrete Weibull distribution estimated with RML estimator [33], [23]; blue line (estimated P.D.A.) denotes the firing rates calculated via parameters of discrete Weibull distribution estimated with LR estimator (see section III-E). All the three firing rates were zeros in the beginning of decomposition.

TABLE VI: Execution time for decomposition of experimental signals. The meaning of indexes are the same as table II

Signal ∗	Duration (s)	Force (MVC%)	Nb MUs	Nb spikes	Sup.(%)	Time(s)	
						Algorithm: P.D.	Algorithm: P.D.A.
1.1	24	20	5	873	18.10	18.04	17.80
1.2	24	20	5	936	18.38	18.31	17.85
1.3	24	20	6	933	17.15	17.36	16.16
1.4	24	30	8	1282	28.39	18.46	18.16
1.5	24	30	8	1295	28.96	19.53	18.81
2.1	24	20	8	1269	17.89	18.44	18.52
2.2	24	15	7	903	24.58	16.84	16.60
2.3	24	10	7	746	10.86	18.40	16.98
2.4	24	5	3	354	4.24	14.32	15.43
3.1	24	20	7	890	13.03	20.33	18.50
3.2	24	15	5	664	10.39	18.47	18.38
3.3	24	10	5	503	9.47	17.18	17.24
3.4	24	5	3	229	4.35	13.49	14.86
4.1	24	20	11	1194	27.55	22.50	22.84
4.2	24	15	11	1040	24.04	20.58	20.71
4.3	24	10	9	1010	24.95	19.27	17.97
4.4	24	5	5	481	14.35	18.26	17.33
5.1	24	20	10	1510	23.31	26.50	26.96
5.2	24	15	8	993	13.49	20.82	19.93
5.3	24	10	5	617	~~97.08~~ 8.75	18.16	19.32
5.4	24	5	3	353	~~99.72~~ 3.12	13.46	16.13

In this paper, we presented several methods to solve the main issues of the previous algorithm, including the replacement of time-consuming RML estimator by a more computationally efficient LR estimator for the inter-spike law parameters, as well as four new heuristics. The proposed approach was validated with 21 experimental signals acquired from five subjects, thus substantially enlarging the validation dataset with respect to our previous work. The results showed that the new algorithm can provide much higher accuracy in decomposition in real time, compared to our previous work [34]. Moreover, the proposed algorithm could decompose experimental signals with a larger numbers of MUs than the real-time decomposition method proposed in [28]. For comparison, the real-time decomposition method proposed in [30] does not process ~~superposed~~ superimposed waveforms and thus provides much lower performance than the proposed one.

As indicated in the analysis of decomposition results, although this approach is able to improve greatly the decomposition performance, some mistakes still remain in the classification. The main reason is due to the limited information of single channel signals. Thus, a multi-channel version of the presented algorithm, which may further boost the decomposition accuracy, is our planned future objective. Another limitation is the relatively low decomposition speed for signals with a large number of active MUs, which restricts the approach to relatively low contraction forces. Tracking large numbers of MUs requires a greater number of scenarios to ensure high accuracy, leading to a large computational complexity, which currently cannot be achieved in real time. This difficulty may be overcome by further simplification of our model or implementation in multi-GPUs.

Development of real-time decomposition algorithms will contribute to the emergence of new types of human-machine interfaces. The presented algorithm processes intramuscular EMG signals to provide access to neural cells in the spinal cord. Intramuscular EMG may become a reliable and safe technology [48] due to the advances in the development of chronically implantable electrodes [49], [50], epimysial electrodes (i.e., placed directly on the muscle surface) [51], as well as of thin-film electronics [52].

A control signal for human-machine interfaces can be derived from either the firing rates or the spike trains estimated by the presented algorithm. Despite the variability of a single MU spike train (see Figure 8), we believe that by pooling information from several MUs and using it to produce a single estimate of contraction force or user intent would produce a reliable communication channel.

In this study, we used signals recorded during constant force contractions. However, it has been previously shown [23] that the algorithm adapts to the dynamic contractions, due to its ability to track the number of active MUs and to adjust to the variations in MUAP templates. However, the possible extent of this adaptability, with respect to the velocity of contraction and the rate at which the geometry of the muscle may change, is a topic for future investigation.

ACKNOWLEDGMENT

This work was supported by the European Research Council synergy project "Natural BionicS" (#810346).

REFERENCES

[1] B. Mambrito and C. D. Luca, "A Technique for the Detection, Decomposition and Analysis of the EMG Signal," *Electroencephalography and Clinical Neurophysiology*, vol. 58, pp. 175–188, 1984.
[2] T. Kamali, R. Boostani, and H. Parsaei, "A multi–classifier approach to MUAP classification for diagnosis of neuromuscular disorders," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 22, pp. 191–200, 2014.
[3] Z. C. Lateva, K. C. McGill, and M. E. Johanson, "The innervation and organization of motor units in a series-fibered human muscle: the brachioradialis," *European J. of Applied Physiology*, vol. 108, pp. 1530–1541, 2010.
[4] C. J. D. Luca and Z. Erim, "Common drive of motor units in regulation of muscle force," *Trends in Neurosciences*, vol. 17, pp. 299–305, 1994.
[5] A. Adam and C. J. D. Luca, "Recruitment order of motor units in human vastus lateralis muscle is maintained during fatiguing contractions," *Trends in Neurosciences*, vol. 90, pp. 2919–2927, 2003.
[6] D. Farina and A. Holobar, "Human machine interfacing by decoding the surface electromyogram," *IEEE Signal Processing Magazine*, vol. 32, pp. 115–120, 2015.
[7] F. Negro, S. Muceli, M. Castronovo, and D. Farina, "Multi-channel intramuscular and surface EMG decomposition by convolutive blind source separation," *J. of Neural Engineering*, vol. 13, no. 2, 2016.
[8] D. Farina, I. Vujaklija, and M. Sartori, "Man/machine interface based on the discharge timings of spinal motor neurons after targeted muscle reinnervation," *Nature biomedical engineering*, vol. 1, 2017.
[9] R. S. LeFever and C. J. De Luca, "A procedure for decomposing the myoelectric signal into its constituent action potentials-part i: technique, theory, and implementation," no. 3, pp. 149–157. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/4121379/
[10] K. C. McGill, K. L. Cummins, and L. J. Dorfman, "Automatic decomposition of the clinical electromyogram," *IEEE Transactions on Biomedical Engineering*, no. 7, pp. 470–477, 1985. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/4122096/
[11] J. Florestal, P. Mathieu, and K. McGill, "Automatic decomposition of multichannel intramuscular EMG signals," *J. of Electromyography and Kinesiology*, vol. 19, pp. 1–9, 2009.
[12] S. H. Nawab, S.-S. Chang, and C. J. De Luca, "High-yield decomposition of surface EMG signals," *Clinical Neurophysiology*, vol. 121, no. 10, pp. 1602–1615, Oct. 2010.
[13] H. R. Marateb, S. Muceli, K. C. McGill, R. Merletti, and D. Farina, "Robust decomposition of single-channel intramuscular emg signals at low force levels," *J. of Neural Engineering*, vol. 8, no. 6, p. 066015, 2011.
[14] C. J. De Luca, "Decomposition of surface EMG signals," vol. 96, no. 3, pp. 1646–1657. [Online]. Available: http://jn.physiology.org/cgi/doi/10.1152/jn.00009.2006
[15] H. Parsaei and D. W. Stashuk, "EMG signal decomposition using motor unit potential train validity," vol. 21, no. 2, pp. 265–274. [Online]. Available: http://ieeexplore.ieee.org/document/6313920/
[16] K. C. McGill, "Optimal resolution of superimposed action potentials," vol. 49, no. 7, pp. 640–650. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/1010847/
[17] H. Marateb and K. McGill, "Resolving superimposed MUAPs using particle swarm optimization," vol. 56, no. 3, pp. 916–919. [Online]. Available: http://ieeexplore.ieee.org/document/4636707/
[18] A. Holobar and D. Zazula, "Multichannel blind source separation using convolution kernel compensation," *IEEE Trans. Signal Process*, pp. 55:4487–96, 2007.
[19] M. Chen and P. Zhou, "A novel framework based on fastica for high density surface emg decomposition," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 1, pp. 117–127, 2016.
[20] F. Negro, S. Muceli, A. M. Castronovo, A. Holobar, and D. Farina, "Multi-channel intramuscular and surface emg decomposition by convolutive blind source separation," *J. of Neural Engineering*, vol. 13, no. 2, p. 026027, 2016.
[21] J. Roussel, P. Ravier, and M. Haritopoulos, "Decomposition of Multi-Channel Intramuscular EMG Signals by Cyclostationary-Based Blind Source Separation," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 2035–2045, 2017.

[22] A. K. Clarke, S. F. Atashzar, A. D. Vecchio, D. Barsakcioglu, S. Muceli, P. Bentley, F. Urh, A. Holobar, and D. Farina, "Deep learning for robust decomposition of high-density surface emg signals," *IEEE Trans. on Biomedical Engineering*, vol. 68, no. 2, pp. 526–534, 2021.

[23] T. Yu, K. Akhmadeev, E. Le Carpentier, Y. Aoustin, R. Gross, Y. Péréon, and D. Farina, "Recursive decomposition of electromyographic signals with a varying number of active sources: Bayesian modeling and filtering," *IEEE Trans. on Biomedical Engineering*, vol. 67, no. 2, pp. 428–440, 2019.

[24] J. E. Chung, J. F. Magland, A. H. Barnett, V. M. Tolosa, A. C. Tooker, K. Y. Lee, K. G. Shah, S. H. Felix, L. M. Frank, and L. F. Greengard, "A fully automated approach to spike sorting," *Neuron*, vol. 95, no. 6, pp. 1381 – 1394.e6, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0896627317307456

[25] P. Yger, G. L. Spampinato, E. Esposito, B. Lefebvre, S. Deny, C. Gardella, M. Stimberg, F. Jetter, G. Zeck, S. Picaud, J. Duebel, and O. Marre, "A spike sorting toolbox for up to thousands of electrodes validated with ground truth recordings in vitro and in vivo," *eLife*, vol. 7, p. e34518, mar 2018. [Online]. Available: https://doi.org/10.7554/eLife.34518

[26] F. J. Chaure, H. G. Rey, and R. Quian Quiroga, "A novel and fully automatic spike-sorting implementation with variable number of features," *Journal of Neurophysiology*, vol. 120, no. 4, pp. 1859–1871, 2018, pMID: 29995603.

[27] T. R. Farrell and R. F. Weir, "The optimal controller delay for myoelectric prostheses," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 15, no. 1, pp. 111–118, 2007.

[28] V. Glaser, A. Holobar, and D. Zazula, "Real-Time Motor Unit Identification From High-Density Surface EMG," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 21, no. 6, pp. 949–958, 2013.

[29] D. Y. Barsakcioglu and D. Farina, "A real-time surface EMG decomposition system for non-invasive human-machine interfaces," in *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2018, pp. 1–4.

[30] S. Karimimehr, H. R. Marateb, S. Muceli, M. Mansourian, M. A. Mananas, and D. Farina, "A Real-Time Method for Decoding the Neural Drive to Muscles Using Single-Channel Intra-Muscular EMG Recordings," *Int. J. of Neural Systems*, vol. 27, no. 6, p. 1750025, 2017.

[31] J. J. Jun, C. Mitelut, C. Lai, S. L. Gratiy, C. A. Anastassiou, and T. D. Harris, "Real-time spike sorting platform for high-density extracellular probes with ground-truth validation and drift correction," *bioRxiv, [Preprint]*, 2017.

[32] M. Pachitariu, N. Steinmetz, S. Kadir, M. Carandini, and H. Kenneth D., "Kilosort: realtime spike-sorting for extracellular electrophysiology with hundreds of channels," *bioRxiv, [Preprint]*, 2016. [Online]. Available: https://www.biorxiv.org/content/early/2016/06/30/061481

[33] J. Monsifrot, E. Le Carpentier, Y. Aoustin, and D. Farina, "Sequential Decoding of Intramuscular EMG Signals via Estimation of a Markov Model," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 5, pp. 1030–40, 2014.

[34] T. Yu, K. Akhmadeev, É. Le Carpentier, Y. Aoustin, and D. Farina, "On-line recursive decomposition of intramuscular emg signals using gpu-implemented bayesian filtering," *IEEE Trans. on Biomedical Engineering*, vol. 67, no. 6, pp. 1806–1818, 2019.

[35] D. Farina, A. Crosetti, and R. Merletti, "A model for the generation of synthetic intramuscular EMG signals to test decomposition algorithms," *IEEE Trans. on Biomedical Engineering*, vol. 48, no. 1, pp. 66–77, Jan. 2001.

[36] D. Stashuk, "EMG signal decomposition: how can it be accomplished and used?" *J. of Electromyography and Kinesiology*, vol. 11, no. 3, pp. 151–173, 2001.

[37] V. Barbu and N. Limnios, "Reliability theory for discrete-time semi-Markov systems," in *Semi-Markov Chains and Hidden Semi-Markov Models toward Applications*, ser. Lecture Notes in Statistics. Springer New York, 2008, vol. 191, pp. 1–30.

[38] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification*. Massachusetts and London: The MIT Press, 1983.

[39] J. Florestal, P. A. Mathieu, and A. Malanda, "Automated Decomposition of Intramuscular Electromyographic Signals," *IEEE Trans. on Biomedical Engineering*, vol. 53, no. 5, pp. 832–839, 2006.

[40] K. C. Mcgill, K. L. Cummins, and L. J. Dorfman, "Automatic decomposition of the clinical electromyogram," *IEEE Trans. on biomedical engineering*, vol. 32, no. 7, 1985.

[41] C. Katsis, Y. Goletsis, A. Likas, D. Fotiadis, and I. Sarmas, "A novel method for automated EMG decomposition and MUAP classification," *Artificial Intelligence in Medicine*, vol. 37, pp. 55–64, 2006.

[42] S. H. Nawab, R. P. Wotiz, and C. J. De Luca, "Decomposition of indwelling emg signals," *J. of Applied Physiology*, vol. 105, no. 2, pp. 700–710, 2008.

[43] M. Khan, A. Khalique, and A. Abouammoh, "On estimating parameters in a discrete weibull distribution," *IEEE Trans. on Reliability*, vol. 38, no. 3, pp. 348–350, Aug. 1989.

[44] K. Kulasekera, "Approximate MLE's of the parameters of a discrete Weibull distribution with Type-I censored data," *Microelectronics Reliability*, vol. 34, no. 7, pp. 1185–1188, 1994.

[45] A. Fernandez and M. Vazquez, "Improved estimation of weibull parameters considering unreliability uncertainties," *IEEE Trans. on Reliability*, vol. 61, no. 1, pp. 32–40, 2011.

[46] K. McGill, Z. Lateva, and H. Marateb, "EMGLAB: An interactive EMG decomposition program," *J. of Neuroscience Methods*, vol. 149, no. 2, pp. 121–133, 2005.

[47] D. Farina, R. Colombo, R. Merletti, and H. B. Olsen, "Evaluation of intra-muscular EMG signal decomposition algorithms," *J. of Electromyography and Kinesiology*, vol. 11, pp. 175–187, 2001.

[48] K. D. Bergmeister, I. Vujaklija, S. Muceli, A. Sturma, L. A. Hruby, C. Prahm, O. Riedl, S. Salminger, K. Manzano-Szalai, M. Aman, M.-F. Russold, C. Hofer, J. Principe, D. Farina, and O. C. Aszmann, "Broadband prosthetic interfaces: Combining nerve transfers and implantable multichannel EMG technology to decode spinal motor neuron activity," *Frontiers in Neuroscience*, vol. 11, p. 421, 2017. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnins.2017.00421/full

[49] P. F. Pasquina, M. Evangelista, A. Carvalho, J. Lockhart, S. Griffin, G. Nanos, P. McKay, M. Hansen, D. Ipsen, J. Vandersea, J. Butkus, M. Miller, I. Murphy, and D. Hankin, "First-in-man demonstration of a fully implanted myoelectric sensors system to control an advanced electromechanical prosthetic hand," *J. of Neuroscience Methods*, vol. 244, pp. 85–93, 2015. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0165027014002672

[50] S. Salminger, A. Sturma, C. Hofer, M. Evangelista, M. Perrin, K. D. Bergmeister, A. D. Roche, T. Hasenoehrl, H. Dietl, D. Farina, and O. C. Aszmann, "Long-term implant of intramuscular sensors and nerve transfers for wireless control of robotic arms in above-elbow amputees," *Science Robotics*, vol. 4, no. 32, p. eaaw6306, 2019. [Online]. Available: http://robotics.sciencemag.org/lookup/doi/10.1126/scirobotics.aaw6306

[51] S. Lewis, M. Russold, H. Dietl, R. Ruff, J. M. C. Audi, K.-P. Hoffmann, L. Abu-Saleh, D. Schroeder, W. H. Krautschneider, S. Westendorff, A. Gail, T. Meiners, and E. Kaniusas, "Fully implantable multi-channel measurement system for acquisition of muscle activity," *IEEE Trans. on Instrumentation and Measurement*, vol. 62, no. 7, pp. 1972–1981, 2013. [Online]. Available: http://ieeexplore.ieee.org/document/6514117/

[52] J. Kim, M.-S. Lee, S. Jeon, M. Kim, S. Kim, K. Kim, F. Bien, S. Y. Hong, and J.-U. Park, "Highly transparent and stretchable field-effect transistor sensors using graphene-nanowire hybrid nanostructures," *Advanced Materials*, vol. 27, no. 21, pp. 3292–3297, 2015. [Online]. Available: http://doi.wiley.com/10.1002/adma.201500710