# On-line recursive decomposition of intramuscular EMG signals using GPU-implemented Bayesian filtering

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

# 4.0 On-line recursive decomposition of iEMG using GPU-implemented Bayesian filtering

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

***Abstract*—Real-time intramuscular electromyography (iEMG) decomposition, which is largely required in the neurological studies and applications, is a complex procedure that involves identifying the motor neuron spike trains from a streaming iEMG recording. We have previously proposed a sequential decomposition algorithm based on a Hidden Markov Model of EMG, that used Bayesian filter to estimate unknown parameters of motor units (MUs) spike trains, as well as their action potentials (MUAPs). In this paper we present a parallel computation implementation of this algorithm on Graphics Processing Unit (GPU), as well as a number of modifications applied to the original model in order to achieve a real-time performance of the algorithm. Specifically, the Kalman filter, previously used to estimate the MUAPs, is replaced by a least-mean-square filter. Additionally, we introduce a number of heuristics that help to omit the most improbable decomposition scenarios while searching for the best solution. Then, a GPU-implementation of the proposed algorithm is presented. Dozens of simulated iEMG signals containing up to 10 active MUs, as well as five experimental fine-wire iEMG signals acquired from tibialis anterior, were decomposed in real time. The accuracy of decompositions depended on the level of muscle activation, but in all cases exceeded 85%.**

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

TABLE I: Main notations

| | |
|---|---|
| $Y$ | The iEMG signal |
| $\Omega$ | The set of indexes of all MUs |
| $A$ | The set of indexes of active MUs |
| $U$ | Spike trains |
| $W$ | White noise |
| $H$ | The vector of MU action potentials shapes |
| $\ell_{\text{IR}}$ | The maximum MUAPs length |
| $T$ | The sawtooth sequences |
| $S$ | The activation scenario |
| $\Theta = [t_0, \beta]$ | The vector containing discrete Weibull distribution parameters: the location parameter and the concentration parameter |
| $t_R$ | The shifting parameter of discrete Weibull distribution, that is the refractory period |
| $\mathsf{Pr}$ | Probability |
| w.p. | with probability |
| $Y[n]$ | The iEMG signal at time index $n$ |
| $Y^n$ | The vector containing the signal from time index 1 to $n$ |
| $\|n$ | Given $Y^n$ |
| $\mathsf{Pr}(T[n] = t[n])$ | The 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 denote random variables, while the lowercase ones stand for their values. |

## I. Introduction

ELECTROMYOGRAM (EMG) is a recording of an electrical activity of muscle fibers generated during their contractions, which result from the excitation originating in the motor neurons (MN) of the spinal cord in form of spike trains. The procedure of identification of these spike trains from an EMG is termed *decomposition*. Such information is crucial in scientific studies of the motor system, as well as in various neurological examinations. A real-time decomposition increases the range of its applicability, including an immediate feedback during positioning of an intramuscular electrode, fatigue assessment and control of human-machine interfaces, such as prostheses.

A majority of currently existing EMG decomposition algorithms [1], [2], [3], [4], [5], [6] are fundamentally off-line. The on-line decomposition was previously addressed in [7], where a multichannel sEMG signal was decomposed using a convolution kernel compensation approach [8]. Moreover, a real-time clustering and template matching algorithm for iEMG was presented in [9]. This algorithm was designed to estimate the cumulative discharge rate of MNs but does not provide resolution of action potentials superimposed in time. Similar challenges as in iEMG decomposition are present in spike sorting algorithms for extracellular recordings from multiple cortical neurons [10], [11], [12] and from nerves (electroneurogram) [13], [14].

Recently we proposed a Bayesian filtering approach for single-channel iEMG decomposition [15], as well as its version adapted to a case of varying number of active MUs [16]. The proposed algorithm achieves full sequential decomposition of iEMG signals. Although the proposed method requires long computation time, it can be accelerated due to its parallelizable structure. In this paper, we introduce several changes in the original algorithm, as well as its parallel implementation on GPU, which permit to achieve the real-time decomposition.

In section II, we will review the Hidden Markov Model (HMM) of iEMG. A Bayesian filtering procedure estimating the parameters of MUs will be presented in section III. Further, we will introduce methods to reduce the complexity of the original algorithm (section IV). Then we will present its parallel implementation (section V). Simulated and experimental iEMG signals used to assess the proposed approach are described in section VI. Finally, results of experimental signal decomposition will be shown and analyzed in section VII.

## II. Hidden Markov Model

### A. Physiology and modeling of EMG

An elementary entity of human neuromuscular system is *motor unit* (MU). A MU comprises a MN in the spinal cord and a certain number of muscle fibers it innervates. The MNs receive the input from the upper levels of motor system and can be either in active or inactive state. While active, a MN exhibits a firing activity in form of *spike train* that propagates to the muscle fibers via MN's axon and causes their contraction. Thus, muscle fibers belonging to the same MU, are excited almost simultaneously, producing a short variation of electric potential in a nearby electrode, called motor unit action potential (MUAP). The inter-spike intervals (ISI) of the trains exhibit certain regularity and have a physiologically-inherent minimal value, called *refractory period*.

Multiple MUs located in the vicinity of the electrode simultaneously contribute to the overall signal, mixing their MUAP trains in one channel. Based on the physiological model, the EMG signal can be interpreted as a linear model [17], [18]:

$$Y[n] = \sum_{i \in A[n]} (H_i * U_i)[n] + W[n] \tag{1}$$

- $n$ is the sampling time index;
- $i$ represents the index of MU;
- $Y$ denotes the observed signal, the iEMG signal;
- $A$ is the set of indexes of active MUs;
- $H$ is the MUAP waveform with finite length $\ell_{\text{IR}}$;
- $U$ represents the spike train of MU comprising 0 and 1, where 1 denotes the discharge and 0 represents the equilibria;
- $W$ is the independent identically distributed white noise samples, with unknown variance $v$;

From (1), the decomposition problem can be interpreted as following: having the observed signal $Y^n$ and initial rough MUAP shapes $H[0]$, we estimate the unknown sequences $U[n]$ and $A[n]$, while refining the MUAP shapes $H[n]$.

### B. State vectors and transition laws of HMM

Based on the linear model presented in subsection II-A, a Hidden Markov Model (HMM) is proposed in [15], [16]. In the following part of this section, we will review the HMM.

In HMM, we introduce $\Theta_i[n]$, a vector containing two parameters of discrete Weibull distribution: a location parameter $t_{0i}[n]$ and a shape parameter $\beta_i[n]$, to describe the spike train statistics of the $i$-th MU. The inter-spike intervals (ISI) distribution respects the discrete Weibull distribution.

Then, vector $(T_i[n])_{i \in A[n]}$, related to the spike train $(U_i[n])_{i \in A[n]}$ in formula (1), is presented:

$$T_i[n] = \begin{cases} 0 & \text{if } U_i[n] = 1 \\ T_i[n-1] + 1 & \text{if } U_i[n] = 0 \end{cases} \tag{2}$$

$(T_i[n])_{i \in A[n]}$ is a discrete sequence that characterizes the time passed since the previous spike. We notice that $T_i[n]$ and $U_i[n]$ are meaningless if $i \notin A[n]$. Thus, we prefer the notation $S[n] = (A[n], (T_i[n])_{i \in A[n]})$.

Finally, the state vector in HMM is shown as following:

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

where $\Omega$ denotes the set of all MUs, including active and inactive ones.

We suppose that the $H_i[n]$ and $\Theta_i[n]$ do not change with time. Thus, we have their transition laws as following:

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

$$\Theta_i[n+1] = \Theta_i[n] \tag{4}$$

In practice, $H_i[n]$ and $\Theta_i[n]$ are not constant over time. An adaptation to their steady changes will be introduced later in subsection III-E. Transition laws for $S[n] = (A[n], (T_i[n])_{i \in A[n]})$ are presented in the following two subsections, respectively for the two components $T_i[n]$ and $A[n]$.

*1) Renewal model:* As shown in [15], the process $(T_i[n])_{n \in A[n]}$ is Markovian. For each $i \in A[n+1] \cap A[n]$, its transition distribution is:

$$T_i[n+1] = \begin{cases} 0 & \text{w.p. } r(T_i[n]+1, \Theta_i[n]) \\ T_i[n]+1 & \text{w.p. } 1 - r(T_i[n]+1, \Theta_i[n]) \end{cases} \tag{5}$$

where $r(\cdot)$ is the hazard rate function of the Discrete Weibull distribution [19].

Moreover, as we described previously in section II-A, ISIs have a lower bound termed *refractory period* $t_R$. We choose $t_R = 30$ ms, which is a physiologically reasonable value [20]. Thus, we have:

$$r(t, \Theta_i[n]) = 0, \text{ if } t < t_R \tag{6}$$

*2) Recruitment model:* Regulation of muscle contraction force is achieved by concurrent modulation of MN firing frequencies and recruitment of additional MUs. The recruitment mechanism is modelled as the variation of $A[n]$, which contains the indexes of all active MUs. It has the following transition law:

$$A[n+1] = \begin{cases} A[n] \setminus i & \text{w.p. 1, if } T_i[n] = t_{\text{I}} \\ A[n] \cup i & \text{w.p. } \frac{\lambda}{\text{card}(\bar{A}[n])}, \text{ if i} \notin A[n] \\ A[n] & \text{w.p. } 1 - \lambda \end{cases} \tag{7}$$

where $\text{card}(\bar{A}[n])$ denotes the number of inactive MUs. An $i$-th active MU is considered to be derecruited when $T_i[n]$ reaches a predefined limit $t_{\text{I}}$. A random inactive MU is considered recruited with predefined constant probability $\lambda$ and initialized with $T[n] = 0$. Thus, $1-\lambda$ is the probability of no MUs being activated at the instant $n$.

### C. Observation model of HMM

The observation equation can be derived from formula (1):

$$Y[n] = \sum_{i \in \Omega} \varphi_i(S[n])\, H_i[n] + W[n] \tag{8}$$

where for all $s = (a, (t_j)_{j \in a})$, $\varphi_i(s)$ is a row vector of size $\ell_{\text{IR}}$ with all components equal to zero, except, if $i \in a$ and $t_i < \ell_{\text{IR}}$, the component in position $t_i + 1$ has value 1.

## III. BAYES FILTER

### A. Principles

The state vectors of HMM $H[n]$, $\Theta[n]$ and $S[n]$ are recursively estimated by Bayes filter. In the following parts, the exponent $^{|n}$ means "given the data $Y^n$". The posterior probability functions of the state vectors are:

- The probability density function (PDF) of $\Theta[n]$ given $S^n, H$ and $Y^n$. It is obvious that $H$ and $Y^n$ are not necessary for the estimation of $\Theta[n]$. Moreover, due to the MUs independence, this PDF is the product of the PDF of $\Theta_i[n]$ given $S^n$. In section III-B, the expected value of $\Theta_i$ given $S^n$, noted $\hat{\theta}_{i,S^n}$, is approximated by a recursive maximum likelihood estimation
- The PDF of $H[n]$ given $S^n$ and $Y^n$. With the marginalization principle [21], this PDF is gaussian and is estimated by a Kalman filter as described in section III-C. The mean and the variance of this PDF will be denoted $\hat{H}^{|n}_{S^n}$ and $P_{S^n}$. Furthermore, the Kalman filter provides the observation prediction noted as $\hat{Y}^{|n-1}_{S^n}$ and its variance noted as $v_{S^n}$. To simplify the calculation complexity, a least-mean-square (LMS) filter is proposed to replace the Kalman filter in section III-C.
- The probability mass function (PMF) of $S^n$ given $Y^n$ (see part III-D).

### B. Estimation of inter-spike law parameters

As presented in our previous work on the algorithm [16], to estimate the inter-spike law parameters (discrete Weibull parameters), a recursive maximum likelihood (RML) estimator was implemented. The likelihood is optimized iteratively by the quasi-Newton method.

For all $n \geq 1$, if $i \in A[n] \cap A[n-1]$, that is, the i-th MU keeps active, we have:

$$\hat{\theta}_{i,S^n} = \hat{\theta}_{i,S^{n-1}} - \frac{1}{\tau_{i,S^n}}\ G^{-1}_{i,S^n}\ Q'_{i,S^n}(\hat{\theta}_{i,S^{n-1}}) \tag{9}$$

$$\begin{aligned} G_{i,S^n} =& \frac{1}{\tau_{i,S^n}}[Q'_{i,S^n}(\hat{\theta}_{i,S^{n-1}})][Q'_{i,S^n}(\hat{\theta}_{i,S^{n-1}})]^T + \\ & (1-\frac{1}{\tau_{i,S^n}})\ G_{i,S^{n-1}} \end{aligned} \tag{10}$$

where $\tau$ is the active time index defined with the formula:

$$\tau_{i,S^n} = \begin{cases} \tau_{i,S^{n-1}} + 1 & \text{if } i \in A[n] \\ \tau_{i,S^{n-1}} & \text{if } i \notin A[n] \end{cases} \tag{11}$$

$G_{i,S^n}$ is an approximate Hessian matrix of the maximum likelihood criterion at the current estimate, and $Q'_{i,S^n}(\theta)$ is the gradient of $Q_{i,S^n}(\theta)$ with:

$$Q_{i,S^n}(\theta) = \begin{cases} -\ln\ r(t_i[n]+1,\theta) & \text{if } t_i[n+1] = 0 \\ -\ln\ (1-r(t_i[n]+1,\theta)) & \text{if } t_i[n+1] = t_i[n]+1 \end{cases}$$

If $i \notin A[n] \cap A[n-1]$, we have:

$$\begin{cases} \hat{\theta}_{i,S^n} = \hat{\theta}_{i,S^{n-1}} \\ G_{i,S^n} = G_{i,S^{n-1}} \end{cases} \tag{12}$$

### C. Estimation of impulse responses

*1) Kalman filter:* Given $S^n$, the Markov model for impulse responses reduces, for all $n \geq 1$:

$$\begin{cases} H[n+1] = H[n] \\ Y[n] = \sum_{i\in\Omega} \varphi_i(S[n])\ H_i[n] + W[n] \end{cases} \tag{13}$$

If $H[1]$ is Gaussian, formula (13) is a standard linear Gaussian model. $H[n] \mid S^n, Y^n$ is Gaussian with mean $\hat{H}^{|n}_{S^n}$ and covariance matrix $P_{S^n}$, $Y[n] \mid S^n, Y^{n-1}$ is Gaussian with mean $\hat{Y}^{|n-1}_{S^n}$ and variance $v_{S^n}$. These means and variances are estimated recursively by the Kalman filter. With the initial prior $\hat{H}^{|0}_{S^0}$ and $P_{S^0}$, we have, for all $n \geq 1$:

- Prediction of observation:

$$\begin{aligned} \hat{Y}^{|n-1}_{S^n} &= \psi(S[n])\ \hat{H}^{|n-1}_{S^{n-1}} \\ v_{S^n} &= \psi(S[n])\ P_{S^{n-1}}\ \psi(S[n])^\top + v \end{aligned} \tag{14}$$

- Estimation of state:

$$\begin{aligned} K_{S^n} &= P_{S^{n-1}}\ \psi(S[n])^\top\ v^{-1}_{S^n} \\ \hat{H}^{|n}_{S^n} &= \hat{H}^{|n-1}_{S^{n-1}} + K_{S^n}\ (Y[n] - \hat{Y}^{|n-1}_{S^n}) \\ P_{S^n} &= P_{S^{n-1}} - K_{S^n}\ v_{S^n}\ K^\top_{S^n} \end{aligned} \tag{15}$$

where $\psi(s) = [\varphi_1(s), ..., \varphi_{\text{card}(\Omega)}(s)]$, $\text{card}(\Omega)$ denotes the number of MUs.

The variance $v$ of the noise is unknown. A heuristic approach is proposed to estimate it with the square of the estimation error $Y[n] - \psi(S[n])\ \hat{H}^{|n}_{S^n}$.

$$\hat{V}^{|n}_{S^n} = (1-\frac{1}{n})\hat{V}^{|n-1}_{S^{n-1}} + \frac{1}{n}(Y[n] - \psi(S[n])\ \hat{H}^{|n}_{S^n})^2 \tag{16}$$

And its global estimation is:

$$\hat{V}^{|n} = \sum_{S^n}\ \hat{V}^{|n}_{S^n}\ \mathsf{Pr}^{|n}(S^n = s^n) \tag{17}$$

where $\hat{V}^{|n}$ replaces $v$ in the formula (14).

*2) Least mean square filter:* Due to the size of matrix $P_{S^n}$, which is $(\text{card}(\Omega) \times \ell_{\text{IR}}) \times (\text{card}(\Omega) \times \ell_{\text{IR}})$, the Kalman filter requires a large computational power. The least-mean-square filter (LMS) is proposed to replace the Kalman filter to accelerate the estimation.

The derivation procedure from Kalman filter to the LMS filter is justified in appendix A. With the rough initial prior $\hat{H}^{|0}_{S^0}$, for all $n \geq 1$, we have the formula of the LMS filter:

$$\begin{aligned} \epsilon[n] &= Y[n] - \psi(S[n])\ \hat{H}^{|n-1}_{S^{n-1}} \\ m_{\Delta,i}[n] &= \frac{\sum_j \Delta_i[j]}{\text{card}(\Delta_i)} \\ \tilde{v}[n] &= 1 + \sum_i m_{\Delta,i}[n]\ \varphi_i(S[n])\varphi_i(S[n])^\top \\ \hat{H}^{|n}_{i,S^n} &= \hat{H}^{|n-1}_{i,S^{n-1}} + \frac{m_{\Delta,i}[n]\varphi_i(S[n])\epsilon[n]}{n\ \tilde{v}[n]} \end{aligned} \tag{18}$$

where $\Delta_i[j]$ denotes the $j$-th inter-spike interval of the $i$-th MU; $\text{card}(\Delta_i)$ is the number of inter-spike intervals of the $i$-th MU; $m_{\Delta,i}[n]$ is the expectation value of the inter-spike intervals of the $i$-th MU at the time index $n$; and $\tilde{v}[n]$

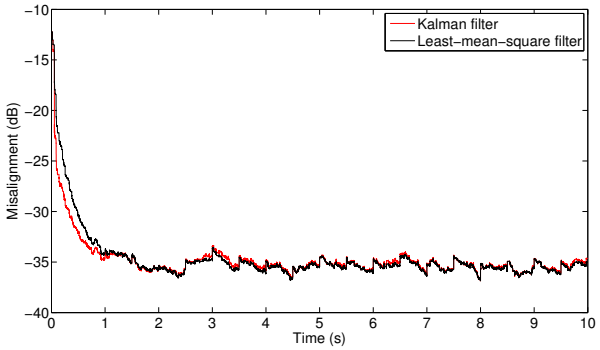



Fig. 1: Misalignment of the Kalman filter algorithm and least-mean-square filter algorithm

represents the ratio of the variance of innovation $v_{S^n}$ to the variance of noise $\hat{V}^{|n}$.

The prediction of observation $\hat{Y}_{S^n}^{|n-1}$ is the same as the formula (14) and the prediction of the variance of innovation $v_{S^n}$ is:

$$v_{S^n} = \tilde{v}[n]\ \hat{V}^{|n}. \tag{19}$$

The performance of the Kalman filter and the LMS filter were evaluated as follows. A simulated signal of five MUs was generated by the HMM model with the time varying impulse responses $H[n]$. Given the scenario $S^n$ and rough initial impulse responses $\hat{H}_{S^0}^{|0}$, the two filters were used to identify $H[n]$. The measure of performance was the normalized misalignment (in dB), defined as $20\log_{10}[\|H[n] - \hat{H}_{S^n}^{|n}\|_2/\|H[n]\|_2]$. Figure 1 shows the misalignment of the two filter algorithms. They have almost the same performance. Thus, the LMS filter is preferred because of the computation time gain.

### D. Posterior probability of scenario

As proposed in our previous work [16], the posterior probability recursion was derived by means of an update-prediction scheme. As follows from the Bayes' theorem, for all possible realizations $s^n$ of $S^n$, the update step is:

$$\mathsf{Pr}^{|n}(S^n = s^n) \propto \mathsf{Pr}^{|n-1}(S^n = s^n)\ g(Y[n] - \hat{Y}_{s^n}^{|n-1}, v_{s^n}) \tag{20}$$

where $g(.,v)$ is a zero-mean and variance $v$ Gaussian PDF.

The prediction step is:

$$\begin{aligned}&\mathsf{Pr}^{|n}(S^{n+1} = s^{n+1}) = \mathsf{Pr}^{|n}(S^n = s^n)\times\\ &\mathsf{Pr}(A[n+1] = a[n+1] \mid S[n] = s[n])\times\\ &\prod_{i\in A[n+1]} \mathsf{Pr}(T_i[n+1] = t_i[n+1] \mid S^n = s^n)\end{aligned} \tag{21}$$

where $\mathsf{Pr}(A[n+1] = a[n+1] \mid S[n])$ is the transition probability of the recruitment model. The elements $\mathsf{Pr}(T_i[n+1] = t_i[n+1] \mid S^n)$, for all $i \in A[n+1]$, are calculated in a different ways for the two following cases:

If $i \in A[n+1] \cap A[n]$, meaning that the MU keeps active, we have:

$$\mathsf{Pr}(T_i[n+1] = t_i[n+1] \mid S^n) \approx \begin{cases} r(T_i[n]+1, \hat{\theta}_{i,S^n}) & \text{if } t_i[n+1] = 0\\ 1 - r(T_i[n]+1, \hat{\theta}_{i,S^n}) & \text{if } t_i[n+1] = T_i[n]+1\\ 0 & \text{otherwise}\end{cases} \tag{22}$$

where $\hat{\theta}_{i,S^n}$ is the inter-spike law parameters of the RML estimation provided in part III-B.

If $i \in A[n+1]\backslash A[n]$, meaning that the MU is activated at the time index $n+1$, according to the model, the inter-spike law parameters are not necessary and we have:

$$\mathsf{Pr}(T_i[n+1] = t_i[n+1] \mid S^n) = 1 \tag{23}$$

The above calculation imply that because of the lack of information about the inter-spike law, the contribution of the MU activation in the posterior probability is principally in the likelihood function of observation.

For each MU $i$, the possible bifurcations of the sawtooth sequence are $t_i^{n+1} = \{t_i^n, t_i[n]+1\}$ and $t_i^{n+1} = \{t_i^n, 0\}$ if $t_i^n > t_R$. The sawtooth sequence is $t_i^{n+1} = \{t_i^n, t_i[n]+1\}$ if $t_i^n \leq t_R$. Therefore, the total number of possible bifurcations from one scenario varies from 1 to $2^{\text{card}(A[n+1])}$, where $\text{card}(A[n+1])$ denotes the number of elements in the $A[n+1]$.

### E. Tracking

To make the algorithm adaptive to non-stationary inter-spike laws parameters $\Theta$ and impulse responses $H$, we introduce a window length sequence $\ell[n]$ [22] growing from 1 to the maximum window length $\ell_\infty$ related to the desired adaptivity:

$$\begin{cases}\ell[1] = 1\\ \ell[n+1] = (1 - \frac{1}{\ell_\infty})\ \ell[n] + 1\end{cases} \tag{24}$$

The formula of the estimated impulse response (18) becomes:

$$\hat{H}_{i,S^n}^{|n} = (1 - \frac{1}{\ell[n]})\hat{H}_{i,S^{n-1}}^{|n-1} + \frac{m_{\Delta,i}[n]\varphi_i(S[n])\epsilon[n]}{\ell[n]\ \tilde{v}[n]} \tag{25}$$

And the formula of the estimated variance of noise (16) is rewritten as:

$$\hat{V}_{S^n}^{|n} = (1 - \frac{1}{\ell[n]})\hat{V}_{S^{n-1}}^{|n-1} + \frac{1}{\ell[n]}(Y[n] - \psi(S[n])\ \hat{H}_{S^n}^{|n})^2 \tag{26}$$

Considering the activation-inactivation of each MU, the window length sequence has a slight change in the inter-spike law parameters estimation:

$$\tau_{i,S^n} = \begin{cases}(1 - \frac{1}{\ell_\infty})\ \tau_{i,S^{n-1}} + 1 & \text{if } i \in A[n]\\ \tau_{i,S^{n-1}} & \text{if } i \notin A[n]\end{cases} \tag{27}$$

We replace the active index (11) by the adaptive formula (27).

### F. Initialisation

At the beginning of the decomposition, we assume that there is no active MUs. Therefore, the set of active MUs indexes $A[1]$ and the sawtooth sequence $T[1]$ are empty. Initial rough estimates of impulse responses $\hat{H}_{S^1}^{|0}$ are manually or automatically extracted using other techniques, e.g. proposed in [23], [24], [25]. An initial estimation of the noise variance $\hat{V}_{s^0}^{|0}$ is made using a signal extract containing no spikes. The initial ISI distribution law parameters of active MUs $\hat{\theta}_{i,S^0}$ are composed of $t_0$ (typically $3t_R \sim 4t_R$) and $\beta$ (typically $2 \sim 4$) according to the our experience. Finally, $n_{\text{path}}$ initial $S^1$ are weighted with the same initial probability $\Pr^{|0}(S^1 = s^1)$.

## IV. Path pruning

As it was previously shown in subsection III-D, the number of possible scenarios for $S^n$ grows exponentially with time, due to its bifurcation. Thus, an exhaustive search for the optimal scenario is impossible. In this section we present several means to discard unnecessary scenarios.

### A. Limiting the number of kept paths

Normally, the conventional measure is limiting the number of kept paths. The $n_{\text{path}}$ most probable scenarios are kept at every time index, where $n_{\text{path}}$ is chosen as a trade-off between the computational complexity and the sub-optimality of the solution.

### B. Pruning based on activity detection

An iEMG signal, especially during low-force contractions, is constituted of short prominent action potentials separated by long segments containing only background noise. It is, thus, desirable to avoid performing the bifurcations of $S^n$ during these inactive segments in order to gain computation time.

We take a measure similar to the signal segmentation presented in [24], [23]. Peaks in EMG that exceed a certain predefined threshold, are considered as segments of signal containing MUAPs. In our algorithm, we introduce $Z[n]$ which represents the output of a pre-detection function $z\left(Y[n+1:n+l_{pd}]\right)$, where $l_{pd}$ denotes the length of pre-detection and is typically set to $\ell_{\text{IR}}/2$ where $\ell_{\text{IR}}$ is the length of MUAPs. If a MUAP or a superposition is detected in the upcoming signal, the pre-detection function returns "1" authorising $S^n$ to bifurcate; otherwise, it returns "0" and prevents $S^n$ from bifurcating. An example is given in figure 2.

We also note that this approach introduces a delay of $\ell_{\text{IR}}/2$ samples in the decomposition process. Generally, it can vary between 2.5 and 5 ms, which can be considered as a negligible delay in most of the applications.

An exact implementation of function $z\left(Y[n+1:n+\ell_{\text{IR}}]\right)$ is beyond the scope of this article. Here, we only note that any convenient EMG segmentation method can be used. In our implementation, an adaptive spike-detection threshold from [24] was used.

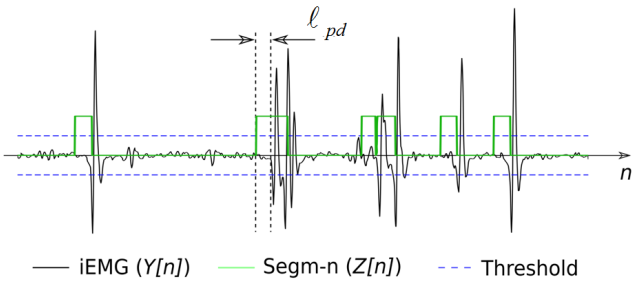
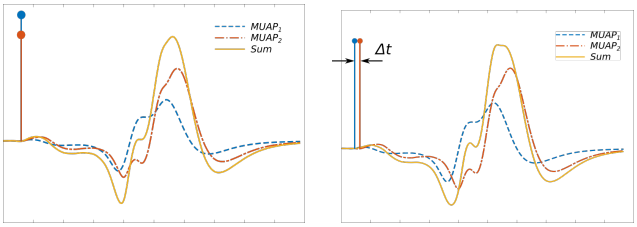



Fig. 2: Example of iEMG segmentation. Segments are detected using certain threshold and shifted in time to the left by $l_{pd}$ due to the use of future samples. Bifurcations containing impulses are forbidden while $Z[n] = 0$.




(a) (b)

Fig. 3: Two close cases of MUAP superposition: (a) - exact superposition of two spikes, a case considered rare and thus excluded from the search; (b) - a close superposition case ($\Delta t$ denotes the sampling period).

.

### C. Simultaneous spikes interdiction

The simultaneous occurrence of two or more spikes at exactly the same time instant is highly improbable. As an example, considering a sampling frequency of 10 kHz and ten active MUs with mean ISIs of 100 ms, the probability of having more spikes at an instant of time, given that there is already one, is $1-(1-1/1000)(1-2/1000)...(1-9/1000) = 0.044$.

Furthermore, we consider the negative impact of this heuristic on the solution can be negligible compared to the gain in computation speed. The impact is illustrated in Figure 3 where an exact superposition (a) can be resolved as its closest possible version (b). Since the superposition shapes in both cases are mostly identical, especially with high sampling frequencies, the effect of this heuristic on the MUAP estimates can be neglected . The gain in computation speed is reached due to the fact that the maximal number of possible bifurcation at step $n$ reduces from $n_{\text{path}} \times 2^{\text{card}(A[n+1])}$ to $n_{\text{path}} \times (\text{card}(A[n+1])+1)$.

## V. Parallelism analysis

In the last ten years, we have entered the epoch of GPU computing. The GPU computation is taking a relatively important place in the field of high performance computing and is applied in a great number of applications in order to achieve superior efficiency. In this section, we analyze the parallelism

of the iEMG signal decomposition model and then implement it into the GPU parallel computation.

Based on the HMM model and Bayes filter established in sections II and III, the structure of iEMG signal decomposition at the time index $n$, for all $n \geq 1$, is:

1) Data transmission: the iEMG signal $Y[n]$.
2) Calculation of posterior probabilities $\mathsf{Pr}^{|n}(S^n = s^n)$ of scenarios with formula (20).
3) Sorting the posterior probabilities of scenarios and keeping the $n_{\text{path}}$ most probable scenarios.
4) Update of the inter-spike law parameters $(\hat{\theta}_{i,S^n})_{i\in\omega}$ with formulas (9), (10), and (27).
5) Update of the impulse responses $(\hat{H}^{|n}_{S^n})_{i\in\omega}$ and the variance of noise $\hat{V}^{|n}$ with formulas (18), (25), (26), and (17).
6) Activation and inactivation of MUs with respect to the recruitment model in subsection II-B2.
7) Bifurcation of the scenarios and calculation of the priori probabilities $\mathsf{Pr}^{|n}(S^{n+1} = s^{n+1})$ of the scenarios with formulas (21), (22), and (23).
8) Prediction of the observed signal $\hat{Y}^{|n}_{S^{n+1}}$ and of the variance of the innovation $v_{S^{n+1}}$ with formulas (14) and (19)
9) Data transmission: the state vector at time index $n$.

The estimation of state vector can be roughly interpreted as a loop-based pattern [26], whose performance in the parallel computing structure varies in terms of the dependencies between loop iterations and the work partition between the available processors. But it is never this case. Since the Bayes filter is a recursive estimation, it is impossible to remove the dependencies between loop iterations. We must calculate them in strictly sequential manner. Therefore, we need to analyze the parallelism in each iteration.

In each iteration, the decomposition process can be separated into a number of single tasks (kernel functions) executed in parallel. In each task, the data can be processed in parallel. In the following sections, we will analyze the structure of the decomposition algorithm to minimize communication between processors and to maximize the use of on-chip resources.

### A. Data parallelism

Data parallelism is a form of parallelization based on data. It focuses on the distribution of data in the different processors that execute the same operation in parallel [26]:

- Paths (or scenarios) on parallel: Before the bifurcation of sawtooth sequences $T[n]$, there are $n_{\text{path}}$ paths which are mutually independent. After the bifurcation, all new paths remain independent. So calculations in all paths could be implemented in the parallel structure with less communication between them.
- MUs on parallel: According to the hypothesis of the Markov model, there is no dependency between any two MUs. Therefore, in every path, the calculation of all MUs can be executed simultaneously.
- Operation on parallel: In every single task, for example: estimation of inter-spike law parameters and estimation of impulse responses, lots of operation as sum of vector or matrix multiplication can be calculated in parallel.

### B. Task parallelism

Task parallelism is another parallelization that contrasts data parallelism [26]. Rather than simultaneously computing the same function on several data elements in data parallelism, task parallelism consists in performing two or more completely different tasks in parallel. In the structure of iEMG signal decomposition, the simultaneous execution of tasks is limited by the dependences between them.

In each iteration, the data transfer takes place twice: data transfer of observed signal $Y[n]$ from host (CPU) to device (GPU) (task 1) and data transfer of state vector from device to host (task 9). The overlap of two types of memory copy and the computation on GPU can be achieved. As a result, the time for data transfer is covered by the execution time of other kernel functions.

Furthermore, some parallel computing architectures support concurrent kernel execution [27], [28], where different small kernels of the same application context can be executed at the same time to ensure the full use of the GPU resources. According to the structure of the Bayes filter presented in section III-A, the PDFs of $\Theta[n]$ and $H[n]$ do not depend on each other. Therefore, in every loop, the tasks related to the estimate of the inter-spike law parameters $\hat{\theta}_{i,S^n}$ can be executed simultaneously with the ones related to impulse responses $\hat{H}^{|n}_{S^n}$. Thus, tasks 4 and 5, as well as tasks 7 and 8, can be calculated at the same time.

### C. Task analysis

To accelerate the decomposition, the algorithm will be implemented in the parallel calculation in GPU. Some of these tasks need to be analyzed in the parallel environment: Task 3 is related to a classic parallel sorting problem; Task 7 (bifurcation of sawtooth sequences), which changes the size of parallel structure, also deserve more consideration.

*1) Parallel sorting:* After the bifurcation of sawtooth sequences, with respect to the transition distribution presented in sections II-B1 and II-B2, there are usually at most $n_{\text{path}} \times 2^{\text{card}(A[n+1])}$ paths. The size of parallel sorting problem varies from $n_{\text{path}}$ to $n_{\text{path}} \times 2^{\text{card}(A[n+1])}$. With the interdiction of simultaneous spikes presented in subsection IV-C, the maximum number of bifurcations reduces to $n_{\text{path}} \times (\text{card}(A[n+1])+1)$.

For small sequences, bitonic sorting is usually considered as one of the fastest traditional parallel sorting algorithms [29], [30]. The time complexity of bitonic sorting is $O(n\ \log_2^2 n)$, while in the parallel environment, it's $O(\log_2^2 n)$ [31].

The most important operation of the bitonic sorting is the arrangement of a bitonic sequence, comprising an ascending sequence and a descending one, into a sorted sequence. In task 3, the final objective is to keep the $n_{\text{path}}$ most probable scenarios. Therefore, in the bitonic sequence, if the size of the ascending one and the descending one are more than $n_{\text{path}}$, we only keep the $n_{\text{path}}$ biggest values in the two sequences to form the bitonic sequence. This measure can remove parts of unnecessary sorting.

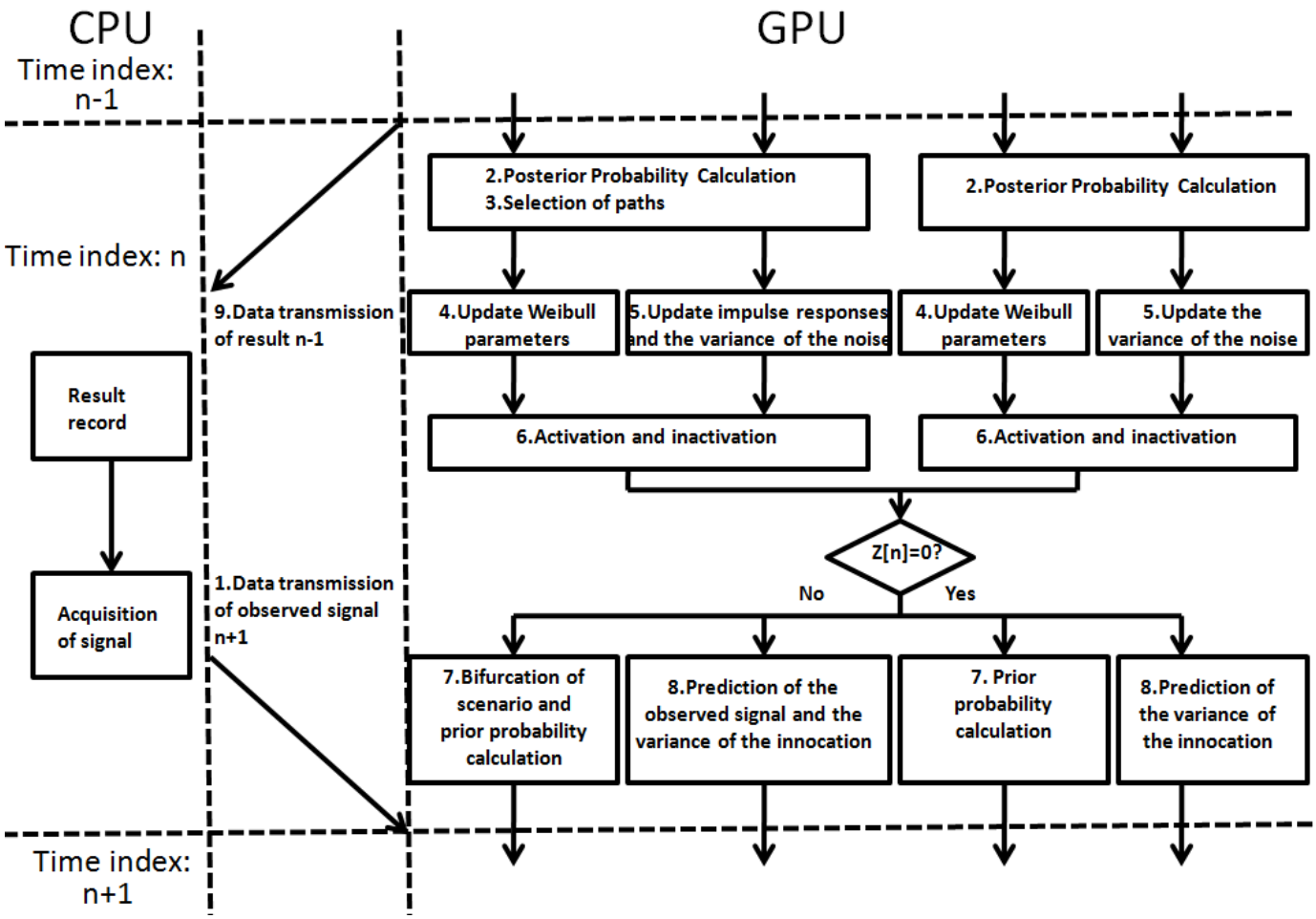



Fig. 4: Parallel structure of iEMG signal decomposition algorithm

*2) Indexes of bifurcation:* Path $S[n]$ bifurcates in at most $A[n+1]+1$ different ways giving an overall number of $n_{\text{path}} \times (A[n+1]+1)$ of new paths. After the parallel sorting, we only keep the $n_{\text{path}}$ most probable new paths at time index $n+1$. To avoid the memory allocation and initialization of each bifurcation originated from one path, indexing is used.

Here is an example for two active motor neurons, which gives a two-dimensional vector $\mathbf{T}[n]$ and three possible bifurcations (the used values are arbitrary):

$$\text{if } \mathbf{T}[n] = \begin{bmatrix} 450 \\ 635 \end{bmatrix}, \quad \mathbf{T}[n+1] \in \left\{ \begin{bmatrix} 451 \\ 636 \end{bmatrix}, \begin{bmatrix} 0 \\ 636 \end{bmatrix}, \begin{bmatrix} 451 \\ 0 \end{bmatrix} \right\} \tag{28}$$

Each i-th motor unit can either not fire at time $n+1$ ($T_i[n+1] = T_i[n] + 1$) or fire if ready ($T_i[n+1] = 0$). Therefore, a binary code can be associated to each configuration in $\boldsymbol{T}$.

$$\mathbf{T}[n+1] \mapsto \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}; \tag{29}$$

This code is unique for each bifurcation within a scenario.

Therefore, in task 7, we initialize the indexes instead of the bifurcation. After sorting the bifurcations and keeping the $n_{\text{path}}$ most probable paths at time index $n+1$, according to the unique index of every bifurcation kept, we initialize the new scenarios.

### D. Parallel structure

As presented in subsection IV-B, $Z[n]$ is the indication of the bifurcation of $S^n$. If $Z[n] = 0$, $S^n$ does not bifurcate, means that $t[n] = t[n-1] + 1$ and $\hat{Y}_{S^n}^{|n-1} = 0$. Hence, we do not need to bifurcate scenarios (task 7) and predict $\hat{Y}_{S^n}^{|n-1}$ (task 8) at time index $n-1$. At the next time index, sorting the posterior probabilities of scenarios and keeping the $n_{\text{path}}$ most probable scenarios (Task 3) are skipped, because after the bifurcation, the number of scenarios does not change. Moreover, the update of impulse responses (Task 5) is not needed.

With the parallelism analysis presented above, the parallel structure is illustrated in schema 4.

## VI. Experimental and Simulation Protocols

### A. Signals

Three groups of simulated signals were generated by the described Markov model with respectively 6, 8 and 10 MUs. There were 10 signals in every group. The sampling frequency was set to 5 kHz and the duration was 20 s. MUAP shapes extracted from the experimental iEMG signals were used to make the simulated signals more realistic. For the statistic parameters of ISI, the refractory period was chosen to be 30

TABLE II: Decomposition performance of simulated signals: 'Nb MUs' is the maximal number of MUs concurrently active in the signal; 'Nb sup-spikes' represents the number of spikes involved in superpositions; 'Nb spikes' denotes the overall number of spikes in the signal; 'Sup.' is the percentage of superposition; 'Nb paths' is the number of paths used in the algorithm; 'Sens.' denotes the global sensitivity; 'Pred.' is the global predictivity.

| Nb MUs | Nb sup-spikes | Nb spikes | Sup.(%) | Nb paths | Sens. (%) | Pred. (%) | Time(s) |
|---|---|---|---|---|---|---|---|
| 10 | 645.20±39.90 | 2093.10±80.59 | 30.83±1.91 | 384 | 94.98±2.51 | 92.26±2.60 | 30.34±0.84 |
| | | | | 256 | 92.45±3.10 | 89.30±2.80 | 25.62±0.74 |
| | | | | 128 | 83.32±6.49 | 80.42±5.16 | 19.95±0.45 |
| 8 | 446.40±28.52 | 1769.80±59.44 | 25.22±1.61 | 384 | 97.55±1.77 | 96.21±2.26 | 26.29±0.47 |
| | | | | 256 | 96.43±2.14 | 94.71±2.65 | 23.31±0.41 |
| | | | | 192 | 95.35±2.78 | 93.21±3.58 | 19.95±0.37 |
| 6 | 291.70±15.27 | 1469.80±52.49 | 19.85±1.04 | 384 | 99.06±1.07 | 98.44±1.55 | 23.75±0.56 |
| | | | | 256 | 98.86±1.23 | 98.18±1.72 | 19.97±0.45 |

ms; the location parameter $t_0$ ranged from 60 ms to 90 ms; and the concentration parameter $\beta$ ranged from 2 to 6. The SNR (Signal to Noise Ratio) was set to 10 dB.

Five experimental signals were acquired from the tibialis anterior (TA) muscle of a 26 years-old healthy man. The subject performed five trials of an isometric force by tracking a trapezoidal profile with target force set to 20% or 30% of the maximal voluntary contraction (MVC). The duration of each trail was 24 s. The wire electrodes used for these recordings were made of Teflon coated stainless steel (50 um diameter; A-M Systems, Carlsborg, WA, USA) and inserted into the muscle with 25G needles. The signals were amplified, band-pass filtered between 100 Hz and 4.4 kHz and sampled at a frequency of 10kHz (OTBioelettronica MEBA amplifier). Then they were subsequently down-sampled to 5 kHz.

Parallel computation algorithm was applied to decode the simulated and the experimental signals. The activation probability $\lambda$ and the maximum time $t_I$ were respectively set to 0.03 and $7t_R$; The window length corresponding to the adaptivity was 1.4 s. The number of selected paths was set to 128, 192, 256 and 384.

### B. Indexes of performance and task complexity

Results of automatic decomposition were evaluated in terms of similarity between the reference and spike trains obtained by the algorithm. In the case of experimental signals, the reference was a manual decomposition provided by an expert using EMGLAB [32]. In case of simulated signals, the exact spike trains were known from the simulation procedure

In order to characterise the complexity of the decomposition task, we use the superposition percentage as in our previous work [16]:

$$\text{Sup} = \frac{Nb_{sup}}{Nb_{spikes}} \tag{30}$$

where $Nb_{spikes}$ is the number of spikes in the reference spike train and $Nb_{sup}$ is the number of spikes which action potentials are superposed with others. We consider a MUAP superimposed if there is at least one other MUAP within a margin of 3 ms (less than half of the average MUAP duration) around it.

In order to quantitatively evaluate the decomposition results, we use global sensitivity and global positive predictivity values, defined as following. A MUAP is considered correctly identified (true positive) if the reference train contains a spike from the same MU within a margin of 1 ms around it. Consequently, 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.

An individual analysis of each MUAP train was also performed, using "classification phase" indexes proposed in [33]. These indexes included sensitivity, specificity and accuracy, as they are defined in [33].

## VII. Results

All signals presented in this section were decomposed on a Nvidia Tesla K80 GPU card using double-precision floating-point format.

### A. Simulated signals

As shown in Table II, three groups of simulated signals with 6, 8 and 10 MUs were decomposed. We note that the mean values of global sensitivity and predictivity (table II) decrease for signals with larger number of active MUs. This is due to increase of decomposition task complexity, quantified by the superposition percentage. Moreover, the standard deviations of the performance indexes show the proportionality to the task complexity. We also observe that greater numbers of paths $n_{path}$ mitigate this effect.

The execution time becomes large with the increasing number of paths and active MUs. The signal with 10 MUs, 8 MUs, and 6 MUs can be decomposed in real time, with respectively 128 paths, 192 paths and 256 paths. More complex decompositions cannot be accomplished in real time using the same computational resources. However, they still can be accomplished in a relatively short time and with high accuracy.

Thus, the number of paths $n_{\text{path}}$, as a parameter determined by the user, defines both the decomposition accuracy and speed. Its value establishes a certain trade-off between the computational complexity (which converts into decomposition time) and the sub-optimality of the solution.

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

| Index | Duration (s) | Force (MVC%) | Nb MUs | Nb spikes | Sup.(%) | Sens. (%) | Pred.(%) | Time (s) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | 256(Nb paths) | 384 | 512 |
| 1 | 24 | 20 | 5 | 873 | 18.10 | 91.41 | 90.27 | 18.5 | 21.32 | 24.93 |
| 2 | 24 | 20 | 5 | 936 | 18.38 | 95.19 | 94.09 | 19.68 | 22.58 | 26.30 |
| 3 | 24 | 20 | 6 | 933 | 17.15 | 94.96 | 91.72 | 16.42 | 18.55 | 20.95 |
| 4 | 24 | 30 | 7 | 1176 | 22.28 | 88.78 | 85.71 | 20.16 | 23.56 | 26.12 |
| 5 | 24 | 30 | 8 | 1295 | 28.96 | 88.34 | 86.68 | 20.70 | 23.31 | 26.78 |

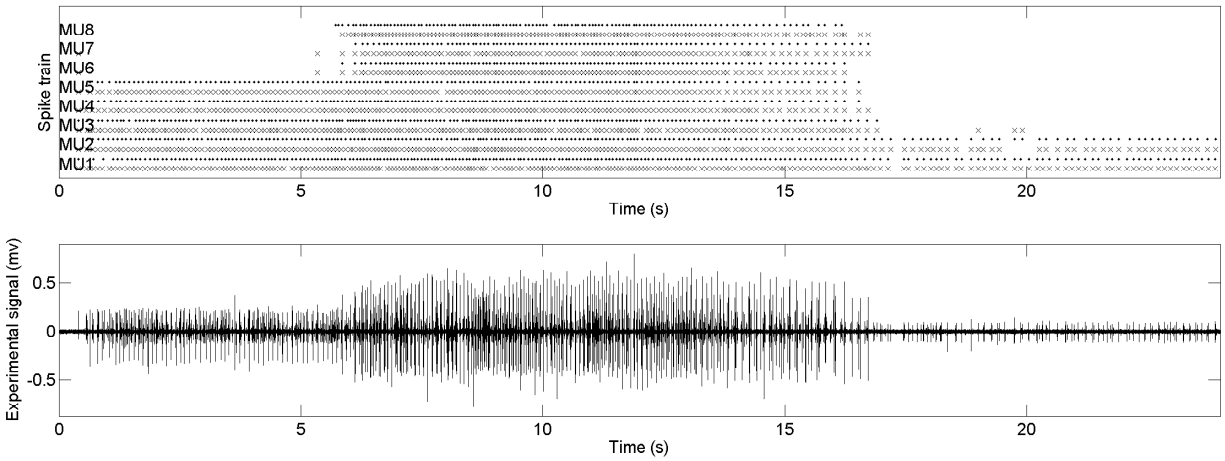



Fig. 5: Comparison of automatic (crosses, 'x') and reference (points, '.') decompositions (upper panel) and the experimental signal from TA, 30% MVC (lower panel).

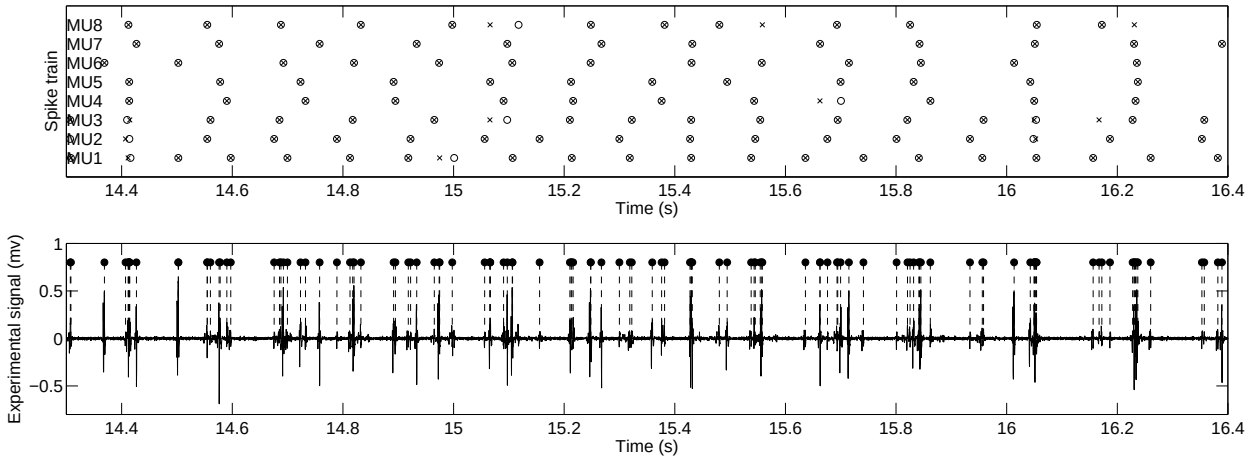



Fig. 6: An extract of the experimental signal decomposition shown in figure 5; circles '○' and crosses 'x' represent respectively the spikes from the reference and automatic decompositions.

TABLE IV: Maximum delay of experimental signals decomposition: the signal index corresponds to the signals presented in table III

| Index | 1 | 2 | 3 | 4 | 5 | 5 |
|---|---|---|---|---|---|---|
| Nb paths | 256 | 256 | 256 | 256 | 256 | 192 |
| Max Delay (ms) | 29.4 | 27.2 | 22.4 | 149.5 | 343.6 | 44.8 |

### *B. Experimental signals*

Five experimental signals (three recorded at 20% MVC, two recorded at 30% MVC) were automatically decomposed. As shown in Table III, for these signals, the number of MUs ranged from two to eight and the percentage of superposition ranged from 17.15% to 28.96%. Both the global sensitivity and predictivity of three signals recorded at 20% MVC were above 90%, while the global sensitivity and predictivity of

the other two complicate signals were more than 85%. We do not show the performance difference for decompositions with various paths in Table III. Because the performances of decomposition with 256, 384 and 512 paths exhibit a very slight amelioration for these experimental signals.

In Table III, we also notice that all the experimental signals can be decomposed automatically in real time with 256 and 384 paths. After the sampling of the iEMG signal, the new observation will take a short time, named decomposition latency (or delay), to be processed. Table IV shows the maximum delay of the decomposition for all experimental signals. The threshold of latency for the real time controlling of a device, such as the active prosthetic devices, is 250 ms [34]. Maximum delays of all the signals decomposed with 256 paths are below this threshold, except the one with 8 MUs. If we choose 192 paths for this signal, its maximum delay also respects this threshold.

Detailed results of the decomposition are illustrated and analysed in the following for the signal with 8 MUs, the most representative and complicate one.

TABLE V: Decomposition performance for an experimental signal detected from the TA with 8 MUs: for each MU, 'Sens.' denotes the sensitivity; 'Pred.' is the predictivity; 'Acc.' represents the accuracy.

| MU | Sens. | Spec. | Acc. |
|---|---|---|---|
| MU1 | 91.53 | 97.87 | 96.54 |
| MU2 | 83.26 | 96.19 | 93.85 |
| MU3 | 73.66 | 95.54 | 92.26 |
| MU4 | 91.87 | 98.42 | 97.53 |
| MU5 | 95.36 | 99.40 | 98.88 |
| MU6 | 97.22 | 99.52 | 99.31 |
| MU7 | 94.95 | 99.43 | 99.05 |
| MU8 | 86.89 | 96.47 | 95.49 |

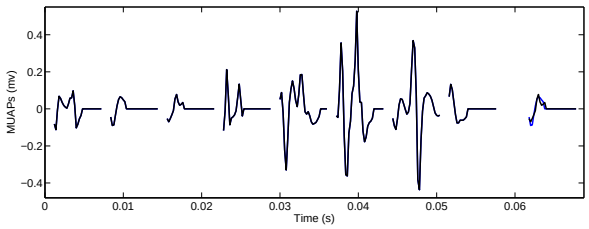



Fig. 7: Eight MUAP shapes (manually-extracted dictionary) for the signal presented in Figure 5, and a comparison between the 2nd one and the 3rd one.

Figure 5 provides a global view of the decomposition results. In the upper panel, the activation zone of each MU in the decomposition algorithm is correlated with the manual reference; In the lower panel, the profile of iEMG signal is exhibited. A detailed view of the decomposition results is given in figure 6, containing two seconds of extracted signal. The algorithm performed generally well, successfully processing several complex superpositions. Due to the high complexity of signal, there are also a few mistakes in the classification. As an example, one may see two misclassification cases occurred at 14.4 s and 16.05 s (see upper panel of figure 6).

For the classification phase, the individual (per MU) performance indexes are shown in table V. Figure 7 illustrates MUAP waveforms of eight MUs. The last one is the comparison of MUAP waveforms between the 2nd one and the 3rd one. According to figure 7, we analyze the performance indexes in table V. The reason for the lower sensitivity of the 2nd and 3rd MU is that they have the smaller amplitudes of MUAPs, compared to the other ones. Generally, this can lead to its complete masking in the superpositions. Furthermore, their MUAP waveforms are similar. Thus, their classification is sometimes influenced by the noise and they switch occasionally with each other, as shown in figure 5 (two cases occurred at 20 s in upper panel). With respect to these two MUs, others are well classified. Globally, the algorithm succeeded in tracking and decomposing the MUs.

The algorithm recursively estimates the parameters of the inter-spike intervals distribution, used to calculate the firing rates. Figure 8 shows the corresponding firing rates. Empirical ones were estimated as the inverse of the moving average of subsequent inter-spike intervals in the reference decomposition. The estimated ones were calculated with the estimated parameters $t_0$ and $\beta$. The algorithm successfully tracked the changes in firing rates.

## VIII. Conclusion and perspectives

In our previous works [15], [16], a sequential decomposition algorithm based on a Hidden Markov Model of the EMG, that used Bayesian filtering to estimate the unknown parameters of discharge series of motor units was proposed. This algorithm has successfully decomposed several experimental iEMG signals, however, demands a high time consuming.

In this paper we presented a real time implementation for the previous algorithm, including the replacement of high time-consuming Kalman filter by a more computationally efficient LMS filter, three heuristics to reduce the complexity and calculated quantity, and the implementation of parallel computation. Validations on simulated and experimental signals demonstrated the successful performance of the algorithm, the same as it shown in [16], and a high decomposition velocity.

Possible limitations of the algorithm arise from large differences of amplitudes between MUAPs (masking of small action potentials) and from the similar MUAP waveforms (switching between similar units). They are the common problems for the single channel iEMG decomposition. Therefore, a multichannel version of the presented algorithm may be of interest. Another limitation is the number of MUs that can be simultaneously tracked by the algorithm in the real-time operation. This limit may be overcome in future by a better hardware or another more efficient mathematical model to reduce the calculation quantity.

## Appendix A
## From Kalman filter to the least-mean-square filter

Kalman filter, originally used for MUAPs estimation, can be replaced by an LMS filter under specific assumptions. Let's

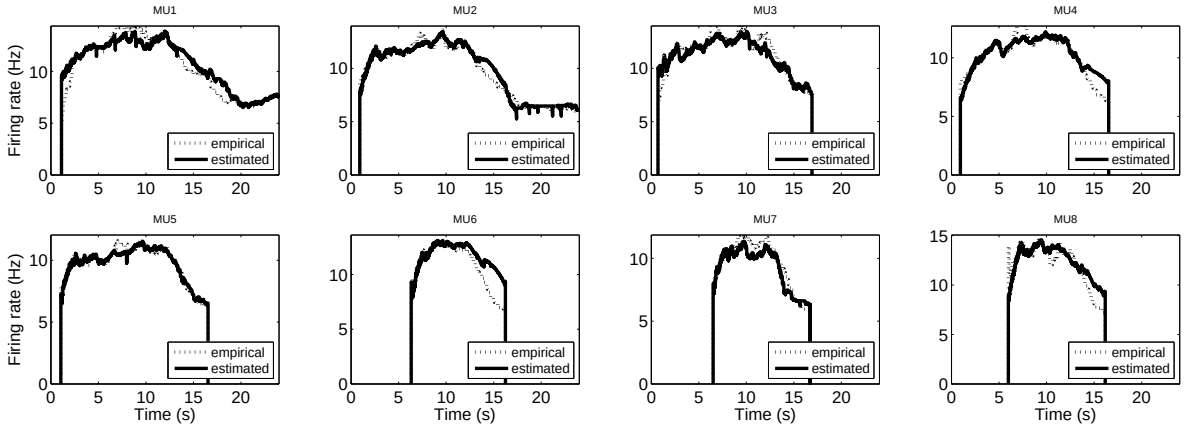



Fig. 8: Firing rates for the iEMG from TA set (see figure 5): the dash line (empirical) represents the firing rates estimated using reference decomposition; continuous line (estimated) represents the firing rates calculated via parameters of discrete Weibull distribution estimated as described in section III-B.

consider the state covariance matrix from (15):

$$\begin{aligned} P_{S^n} =& P_{S^{n-1}} - K_{S^n}\ v_{S^n}\ K_{S^n}^\top \\ =& P_{S^{n-1}} - P_{S^{n-1}}\ \psi(S[n])^\top\ v_{S^n}^{-1}\ v_{S^n} \\ & (P_{S^{n-1}}\ \psi(S[n])^\top\ v_{S^n}^{-1})^\top \\ =& P_{S^{n-1}} - P_{S^{n-1}}\ \psi(S[n])^\top\ v_{S^n}^{-1}\ \psi(S[n]) P_{S^{n-1}} \end{aligned} \tag{31}$$

Applying the Woodbury matrix identity:

$$[A+BCD]^{-1} = A^{-1} - A^{-1}B[DA^{-1}B + C^{-1}]^{-1}DA^{-1} \tag{32}$$

to (31), we obtain:

$$\begin{aligned} P_{S^n}^{-1} =& P_{S^{n-1}}^{-1} + \psi(S[n])^\top\ (v_{S^n} - \psi(S[n])\ P_{S^{n-1}}\ \psi(S[n])^\top) \\ & \psi(S[n]) \end{aligned} \tag{33}$$

This can be simplified using expression (14) for the variance of innovation:

$$P_{S^n}^{-1} = P_{S^{n-1}}^{-1} + \psi(S[n])^\top\ v^{-1}\ \psi(S[n]) \tag{34}$$

where $v$ is the variance of measurement noise $\hat{V}^{|n}$ estimated using (16) and (17). Finally, we have:

$$\begin{aligned} P_{S^n} &= \frac{\hat{V}^{|n}}{n}\ R_{S^n}^{-1} \\ R_{S^n} &= \frac{1}{n}\ \sum_{k=1}^{n} \psi(S[k])^\top \psi(S[n]) \end{aligned} \tag{35}$$

where $R_{S^n}$ can be approximated by a constructed made of $\text{card}(\Omega)\times\text{card}(\Omega)$ blocks $R_{i,j,S^n}$ with dimension $O(\ell_{\text{IR}}\times\ell_{\text{IR}})$:

$$R_{i,i,S^n} = \begin{bmatrix} \xi_{i,S^n} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \xi_{i,S^n} \end{bmatrix} \tag{36}$$

$$R_{i,j,S^n} = \begin{bmatrix} \xi_{i,S^n}\xi_{j,S^n} & \cdots & \xi_{i,S^n}\xi_{j,S^n} \\ \vdots & \ddots & \vdots \\ \xi_{i,S^n}\xi_{j,S^n} & \cdots & \xi_{i,S^n}\xi_{j,S^n} \end{bmatrix} \tag{37}$$

where $\xi_{i,S^n}$ is the firing rate of i-th motor unit, which is the inverse of its inter-spike interval (ISI) expected value. We can notice that $\forall i,j \in \Omega,\ \xi_{i,S^n}\xi_{j,S^n} \ll \xi_{i,S^n}$ and $\xi_{i,S^n}\xi_{j,S^n} \ll \xi_{j,S^n}$. Therefore, if $i \neq j$, $R_{i,j,S^n}$ can be approximated by a zero-matrix, $R_{S^n}$ can be approximated by a diagonal matrix.

Having the approximation of $P_{S^n}$, we can derive directly the LMS filter from the Kalman filter (15). With a rough initial prior $\hat{H}_{S^0}^{|0}$, for all $n \geq 1$, we have:

$$\begin{aligned} \epsilon[n] &= Y[n] - \psi(S[n])\ \hat{H}_{S^{n-1}}^{|n-1} \\ m_{\Delta,i}[n] &= \frac{\sum_j \Delta_i[j]}{\text{card}(\Delta_i)} \\ \tilde{v}[n] &= 1 + \sum_i m_{\Delta,i}[n]\ \varphi_i(S[n])\varphi_i(S[n])^\top \\ \hat{H}_{i,S^n}^{|n} &= \hat{H}_{i,S^{n-1}}^{|n-1} + \frac{m_{\Delta,i}[n]\varphi_i(S[n])\epsilon[n]}{n\ \tilde{v}[n]} \end{aligned} \tag{38}$$

where $\Delta_i[j]$ denotes the j-th ISI of the i-th motor unit; $\text{card}(\Delta_i)$ is the number of the ISIs for the i-th motor unit; $m_{\Delta,i}[n]$ is the expected ISI for i-th motor unit at time index $n$; and $\tilde{v}[n]$ represents the ratio of the variance of innovation $v_{S^n}$ to the variance of noise $\hat{V}^{|n}$. And the prediction of the variance of innovation $v_{S^n}$ is:

$$v_{S^n} = \tilde{v}[n]\ \hat{V}^{|n}. \tag{39}$$

In order to make this filter adaptive to the changes in MUAPs forms, time index $n$ can be replaced by a forgetting factor $l[n]$.

## ACKNOWLEDGMENT

This work was supported by the China Scholarship Council [grant number 201404490033].

## REFERENCES

[1] D. Ge, E. Le Carpentier, J. Idier, and D. Farina, "Spike sorting by stochastic simulation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 19, no. 3, pp. 249–259, 2011.

[2] 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.

[3] 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," *Journal of neural engineering*, vol. 8, no. 6, p. 066015, 2011.

[4] 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. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S138824571000338X

[5] F. Negro, S. Muceli, A. M. Castronovo, A. Holobar, and D. Farina, "Multi-channel intramuscular and surface emg decomposition by convolutive blind source separation," *Journal of Neural Engineering*, vol. 13, no. 2, p. 026027, 2016. [Online]. Available: http://stacks.iop.org/1741-2552/13/i=2/a=026027

[6] 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.

[7] 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.

[8] A. Holobar and D. Zazula, "Multichannel blind source separation using convolution kernel compensation," *IEEE Trans. Signal Process*, pp. 55:4487–96, 2007.

[9] 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.

[10] S. E. Paraskevopoulou, D. Y. Barsakcioglu, M. R. Saberi, A. Eftekhar, and T. G. Constandinou, "Feature extraction using first and second derivative extrema (FSDE) for real-time and hardware-efficient spike sorting," *Journal of Neuroscience Methods*, vol. 215, no. 1, pp. 29–37, Apr. 2013. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0165027013000393

[11] J. Navajas, D. Y. Barsakcioglu, A. Eftekhar, A. Jackson, T. G. Constandinou, and R. Quian Quiroga, "Minimum requirements for accurate and efficient real-time on-chip spike sorting," *Journal of Neuroscience Methods*, vol. 230, pp. 51–64, Jun. 2014. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0165027014001344

[12] T. Werner, E. Vianello, O. Bichler, D. Garbin, D. Cattaert, B. Yvert, B. De Salvo, and L. Perniola, "Spiking Neural Networks Based on OxRAM Synapses for Real-Time Unsupervised Spike Sorting," *Frontiers in Neuroscience*, vol. 10, Nov. 2016. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnins.2016.00474/full

[13] L. Citi, J. Carpaneto, K. Yoshida, K.-P. Hoffmann, K. P. Koch, P. Dario, and S. Micera, "On the use of wavelet denoising and spike sorting techniques to process electroneurographic signals recorded using intraneural electrodes," *Journal of Neuroscience Methods*, vol. 172, no. 2, pp. 294–302, Jul. 2008.

[14] D. Pani, G. Barabino, L. Citi, P. Meloni, S. Raspopovic, S. Micera, and L. Raffo, "Real-time neural signals decoding onto off-the-shelf dsp processors for neuroprosthetic applications," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 9, pp. 993–1002, 2016.

[15] 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.

[16] T. Yu, K. Akhmadeev, E. L. Carpentier, Y. Aoustin, R. Gross, Y. Péréon, and D. Farina, "Recursive decomposition of electromyographic signals with a varying number of active sources: Bayesian modelling and filtering (under-submission)."

[17] 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.

[18] 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. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S105064110000050X

[19] 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.

[20] C. C.J. Heckman and R. Enoka, "Motor unit," *Compr Physiol*, vol. 2, no. 4, pp. 2629–2682, 2012.

[21] T. Schon, F. Gustafsson, and P.-J. Nordlund, "Marginalized particle filters for mixed linear nonlinear state-space models," *IEEE Trans. on Signal Processing*, vol. 53, pp. 2279–2289, 2005.

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

[23] 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.

[24] 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.

[25] 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.

[26] S. Cook, *CUDA Programming A Developer's Guide to Parallel Computing with GPUs*. Morgan Kaufmann, 2013.

[27] N. Corporation, *Whitepaper NVIDIA's Next Generation CUDA Compute Architecture: Fermi*. NVIDIA Corporation, 2009.

[28] J. Sanders and Kandrot, *CUDA by example: an introduction to General-purpose GPU Programming*. Addison-Wesley Professional, 2010.

[29] A. C. Dusseau, D. E. Culler, K. E. Schauser, and R. P. Martin, "Fast parallel sorting under logp: Experience with the cm-5," *IEEE Trans. Parallel Distrib. Syst.*, vol. 7, no. 8, pp. 791–805, 1996.

[30] N. Satish, M. Harris, and M. Garland, "Designing efficient sorting algorithms for manycore gpus," *23rd IEEE International Parallel and Distributed Processing Symposium*, pp. 1–10, 2009.

[31] A. Greb and G. Zachmann, "Gpu-abisort: optimal parallel sorting on stream architectures," in *Proceedings 20th IEEE International Parallel and Distributed Processing Symposium*, 2006.

[32] 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.

[33] 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.

[34] M. A. Oskoei and H. Hu, "Myoelectric control systemsa survey," *Biomedical Signal Processing and Control*, vol. 2, no. 4, pp. 275–294, 2007.