Performance of Three Electromyogram Decomposition Algorithms as a Function of Signal to Noise Ratio: Assessment with Experimental and Simulated Data

Chenyun Dai, Yejin Li, Edward A. Clancy
Worcester Polytechnic Institute (WPI)
Worcester, MA 01609, USA
{cdai, yli1, ted}@wpi.edu
Anita Christie (adcl@uoregon.edu)
University of Oregon, Eugene, OR 97403, USA

Paolo Bonato (pbonato@partners.org)
Harvard Medical School, Charlestown, MA 02129, USA

Kevin C. McGill (kemcgill43@gmail.com)
VA Palo Alto Health Care System, Palo Alto, CA 94304, USA

Abstract— We have previously published a full report [25] comparing the performance of three automated electromyogram (EMG) decomposition algorithms. In our prior report, the primary measure of decomposition difficulty/challenge for each data record was the “Decomposability Index” of Florestal et al. [3]. This conference paper is intended to augment our prior work by providing companion results when the measure of difficulty is the motor unit signal-to-noise ratio (SNR$_{MU}$) — a measure that is commonly used in the literature. Thus, we analyzed experimental and simulated data to assess the agreement and accuracy, as a function of SNR$_{MU}$, of three publicly available decomposition algorithms—EMGlab [1] (single channel data only), Fuzzy Expert [2] and Montreal [3]. Data consisted of quadrifilar needle EMGs from the tibialis anterior of 12 subjects at 10%, 20% and 50% maximum voluntary contraction (MVC); single channel needle EMGs from the biceps brachii of 10 control subjects during contractions just above threshold; and matched simulated data. Performance vs. SNR$_{MU}$ was assessed via agreement between pairs of algorithms for experimental data and accuracy with respect to the known decomposition for simulated data. For experimental data, RMS errors between the achieved accuracy and those predicted by an exponential model as a function of SNR$_{MU}$ ranged from 8.4% to 19.2%. For the simulations, RMS errors between achieved accuracy and those predicted by the SNR$_{MU}$ exponential model ranged from 3.7% to 14.7%. Agreement/accuracy was strongly related to SNR$_{MU}$.

Keywords— Electromyogram (EMG), motor units, decomposition, intramuscular EMG, biomedical signal analysis.

I. INTRODUCTION

Indwelling electromyogram (EMG) recordings are decomposed by separating the interference pattern into its constituent motor unit action potential trains (MUAPTs). Doing so permits the evaluation and study of individual motor unit (MU) firing patterns and action potential shapes, which is useful in a wide range of clinical and scientific studies (for reviews, see [4]–[6]). In most decomposition schemes, an automated algorithm detects and clusters each MUAP firing, typically with expert manual editing performed thereafter. Signal processing methods for automated decomposition were pioneered by DeLuca and colleagues [7], [8]; with numerous variations and alternative approaches proposed and studied thereafter [2], [4], [9]–[17].

Quantitative performance evaluation of automated decomposition algorithms has been conducted in a few manners [18]. First, reference annotations have been produced via manual expert editing of experimental data [2], [4], [12], [13], [17]. This technique is extremely time consuming (e.g., one hour per second of data [2]) and its true accuracy can be difficult to assess. Yet, the use of experimental data guarantees signal conditions representative of actual use. Second, EMGs can be simulated [8], [9], [11]–[13], [15], [17]. In this case, the true annotations are known. But, even highly detailed simulators produce data that cannot replicate all of the complexities of an experimental signal. Third, a few studies have recorded EMGs from multiple indwelling needles, comparing the decompositions of MUs detected by more than one electrode [4], [19]–[21]. Agreement in their firing times is strong evidence of accurate detection and classification. Recently, studies have compared decomposition results between EMG simultaneously acquired from indwelling electrodes and surface EMG arrays [22]–[24]. Overall, a combination of evidence from experimental and simulated data is typically used to evaluate an algorithm, as each evaluation technique exhibits strengths and weaknesses.

For all automated algorithms, it is well established that performance depends on the characteristics of the signal being analyzed. Relative decomposition accuracy is known to decrease when: more spikes occur per second, MUAPs from distinct trains exhibit similar shapes, the signal-to-noise ratio (SNR) lowers, MUAP shapes change over time and/or firing times are irregular [4]. Florestal et al. [3] attempted to capture
these signal characteristics in their Decomposability Index (DI), defined as the minimum RMS difference between the given MUAP template and each other MUAP template (or the baseline), divided by the RMS value of the entire channel. This measure takes into account both the size and the distinguishability of the MUAPs.

In our prior full report [25], we contrasted the performance of three automated algorithms [1]–[3] that are publicly available for use in the MATLAB environment, with DI as our primary measure of the difficulty/challenge expected from each recording. Since algorithm performance varies depending on the data set analyzed, the same data were presented to each algorithm. We augmented the experimental data with data created by a publicly available EMG simulator [26], providing a more comprehensive evaluation.

In this conference publication, we augment our prior work by presenting complimentary results when the measure of data set difficulty/challenge is the SNR of a MU (SNR$_{MU}$). Distinct from the DI, the SNR$_{MU}$ is sensitive only to the amplitude of a MUAP, relative to the noise floor. This difficulty measure is more traditional than the DI and arguably simpler to understand. Thus, we present a companion cross-comparison of the performance of the three decomposition algorithms, as a function of the SNR$_{MU}$. Readers are advised to review our prior full report [25], which contains complete project methods, etc.

II. METHODS

A. Experimental Data

Portions of experimental data from two prior studies were reanalyzed, and simulated data were generated. No new subject data were collected. The data reanalysis was approved by the WPI Institutional Review Board. The experimental data spanned a range of MVC levels as well as laboratory and clinical data collection settings, to provide data with a range of challenges to decomposition algorithms.

Three-channel quadrifilar needle EMGs had been acquired from the dominant leg of seven young (three male; four female; aged 18–30 years) and five elderly (two male, three female; aged 65 years or older) healthy subjects at the University of Massachusetts [25]. Briefly, the skin over the tibialis anterior (TA) muscle was cleaned with rubbing alcohol and a 27-gauge four-wire quadrifilar needle electrode was inserted into the belly of the TA muscle, avoiding the innervation zone. Four 50-µm diameter platinum-iridium wires terminating at a side port 7.5 mm from the tip of the electrode comprised the recording surfaces [27]. The four wires in this electrode were arranged in a square array with approximately 200 µm on each side. The signals detected with this needle were connected to three differential amplifiers (10$^{12}$ Ω input resistance; 25 pA bias current), bandpass filtered from 1,000–10,000 Hz, and sampled at an effective rate of 51,200 Hz (16-bit resolution). Following electrode insertion, subjects performed 30 s duration constant-force contractions at 10%, 20% and 50% MVC, with target force levels displayed on a video monitor. A rest period of three minutes was provided between contractions to prevent fatigue. One, 5 s segment during the constant-force portion of each recording was analyzed. Thus, 36 recordings of 5 s duration each were used (12 subjects x 3 levels of contraction).

Single channel needle EMGs were reanalyzed from ten control subjects (6 males, 4 females; aged 21–37 years) in the publicly-available “N2001” database of Nikolic [28]. Of the available recordings within the database, recordings exhibiting a low background noise level (assessed visually) were selected. Recordings were acquired from the biceps brachii muscles during low level (just above threshold), constant-force contractions using a concentric needle electrode in accordance with standard clinical recording procedures. The signals were bandpass filtered between 2–10,000 Hz and sampled at 23,437.5 Hz with 16 bit resolution. Ten 5 s recordings (10 subjects x one recording/subject) were used for analysis.

For each signal, a “Spike Rate” measure was computed, expressing the number of MUAP firings per second. Within the analyzed 5 s segment of each recording, the number of pulses exceeding the background noise was manually counted. Spikes of duration greater than 3 ms, representing superimpositions, were counted as two pulses. Those with duration greater than 6 ms were counted as three pulses, etc. This approach accentuates the influence of longer duration spikes (which are, presumably, more difficult to decompose) and causes the Spike Rate to be larger than the rate that would be derived by using the number of events found by the detection stage of a classical decomposition algorithm. For multiple-channel data, all three channels were simultaneously viewed and a pulse was counted if it was discernible from the background in any channel. The Spike Rate measure was expressed in pulses per second (pps). Spike Rate measures from the experimental data were used to guide generation of the simulated data.

B. Simulated Data

Constant-force, quadrifilar and single channel data were simulated using the publicly-available needle EMG simulator of Hamilton-Wright and Stashuk [26]. The resulting signals closely resembled those acquired experimentally. The simulator parameters were selected to model the physical layout of the TA muscle, MU firing patterns, action potential propagation and type of EMG electrode. To emulate quadrifilar recordings, four noise-free monopolar tip electrodes (50 µm diameter) were simultaneously simulated in a square array configuration at 200 µm distances. This configuration mimics a quadrifilar needle. The three differential voltages were then computed offline in MATLAB and white Gaussian noise was added to give a SNR of 20 dB. For each experimental contraction level to be simulated, trial and error was used to determine the contraction level parameter input value of the simulator software such that the average Spike Rate of the simulated data matched the average Spike Rate of the corresponding experimental data. Five-second constant-force recording segments were created at force levels representing 10%, 20% and 50% MVC. Each simulated condition was iterated 12 times, providing 12 realizations, to give the same number of trials as with the quadrifilar experimental data. The true time instances and
identities of each MUAP firing (i.e., MUAP annotations), which are fully known in simulated data, were recorded along with the simulated signals (sampled at 31,250 Hz, 16-bit resolution). To emulate healthy (control) single channel recordings, one 10 mm concentric electrode was simulated and white Gaussian noise was added to give a SNR of 20 dB. The Spike Rate of these simulated data was matched to the average Spike Rate of the single channel needle (N2001) data of the control subjects, again via selection of the contraction level parameter input value of the simulator software. Ten recordings, each of 5 s duration, were created at a sampling rate of 31,250 Hz with 16-bit resolution, along with the true MUAP annotations.

C. Automated Decomposition Algorithms

Three publicly-available decomposition algorithms were compared. Each is implemented in MATLAB, which was used for all computation. Each algorithm was used without manual editing, although such editing is the norm in scientific studies. Prior to automated decomposition, the quadrifilar experimental data were digitally highpass filtered at 100 Hz. Although the signal had been analog highpass filtered at 1,000 Hz, this digital filter removed any offsets due to subsequent analog filter stages, including the analog to digital converter. The single channel experimental data were digitally highpass filtered at 500 Hz. This cut-off frequency was selected after visual review of a subset of the data, so as to reduce background noise and best accentuate spikes. All simulated data were digitally highpass filtered at 1,000 Hz, this cut-off frequency also being selected after visual review of a data subset. In all cases, a first-order Butterworth filter was designed, and then applied in the forward and reverse time directions to achieve zero phase shift.

All three algorithms detected voltage spikes within the EMG (each spike is a candidate MUAP, typically with a registration time corresponding to its peak magnitude), classified spikes with similar shapes and resolved superimpositions. The first automated decomposition algorithm was the default algorithm implemented in the publicly-available “EMGlab” software [1]. This algorithm can only analyze single channel EMGs and thus was only used for our single channel data. The second algorithm was the “Montreal” algorithm [3]. This algorithm has no adjustable parameters. The third algorithm was the “Fuzzy Expert” algorithm [2]. With the Fuzzy Expert algorithm, we utilized ten algorithm passes and limited resolution of superimpositions to three MUs on the first two passes, five MUs on the third pass and six MUs thereafter.

D. Methods of Analysis

After highpass filtering (described above), all experimental and simulated quadrifilar data were automatically decomposed by the Fuzzy Expert and Montreal algorithms. The single channel experimental and simulated data were decomposed by all three automated algorithms. Decompositions of experimental signals were compared pair-wise between algorithms for each signal. Each MUAP annotation was said to match if both algorithms found a MUAP from the same train within a ±1 ms match window\(^1\), after determining a timing offset that accounts for the difference in MUAPT registration locations between the different algorithms [18], [24]. “Agreement” was measured as the number of matched annotations, divided by the sum of: (1) matched annotations and (2) unmatched annotations from either algorithm. Agreement results were expressed in percent. For the experimental quadrifilar data, results are only presented for those MUAPTs that exhibited a minimum of 20 matches between the Fuzzy Expert and Montreal algorithms (average of 4 matches per second over a 5 s recording duration). For the experimental single channel data, results are only presented for those MUAPTs that exhibited a minimum of 20 matches for each pairing between the three algorithms (i.e., those MUAPTs “found” by all three algorithms). For simulated data, the minimum number of required matches was one (i.e., every MUAPT that was extracted was analyzed). In addition to agreement results, decompositions of simulated signals were also compared directly to the true annotations (all MUAPTs included), this result being denoted “Accuracy,” since the true annotations were known.

For each identified MUAPT for single channel data, a MU SNR (SNR\(_{MU}\)) was computed as the peak-to-peak height of the MU divided by the RMS value of the entire channel [21]. A ten-bin histogram of all (negative/positive) peak values from all firings of a MUAPT was computed. A peak value was estimated as the average height of all values contributing to the histogram mode bin. This selection helps to reject peak values that might be unrepresentative due to MUAP superimpositions. For multiple-channel data, the SNR\(_{MU}\) was computed separately for each channel and then averaged. The SNR\(_{MU}\) is non-dimensional. For experimental signals, SNR\(_{MU}\) was computed multiple times, using the annotations from each respective decomposition algorithm. For simulated signals, the measures were also computed using the truth annotations. Cross-plots of SNR\(_{MU}\) vs. agreement (or accuracy) were created for each contraction level for each data set. The data from each plot were then least squares fit to the exponential model: Agreement \(= 100 - a \cdot e^{-b \cdot \text{SNR}_{MU}}\), where \(a\) and \(b\) are the fit parameters. Except where indicated otherwise, performance differences were tested statistically using ANOVAs (two- or one-way), with post hoc pair-wise comparisons (when significant) conducted using Tukey’s honest significant difference (HSD) test.

III. Results

Table I in the companion work [25] lists and discusses the number of MUAPTs detected and analyzed in the various data sets, the total number of excluded MUAPTs (due to fewer than 20 matches) for the experiments, as well as the actual (true) number of MUAPTs generated for the simulated data. The companion work also details the Spike Rate values for each data set. Spike Rate increased with MVC level. The average experimental control trial and simulated trial Spike Rate values were quite well matched, as designed.

\(^1\) Note that a different match window of ±0.5 ms was used in [25].
TABLE II
SUMMARY AGREEMENT RESULTS FOR EXPERIMENTAL DATA AND ACCURACY
RESULTS FOR SIMULATED DATA. EACH ENTRY LISTS MEDIAN [25TH
PERCENTILE, 75TH PERCENTILE]. SIGNIFICANT DIFFERENCES INDICATED BY *
($p < 0.05$) OR ** ($p < 0.01$).

<table>
<thead>
<tr>
<th>Experimental</th>
<th>Agreement between algorithms (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>signals</td>
<td>Montreal v Fuzzy Montreal v EMGlab Fuzzy v EMGlab</td>
</tr>
<tr>
<td>Quadrifilar</td>
<td></td>
</tr>
<tr>
<td>10% MVC</td>
<td>96 [81, 100] **</td>
</tr>
<tr>
<td>20% MVC</td>
<td>86 [72, 97] **</td>
</tr>
<tr>
<td>50% MVC</td>
<td>65 [35, 84] **</td>
</tr>
<tr>
<td>Single channel Controls</td>
<td>97 [87, 100] 97 [89, 98] 96 [91, 98]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simulated signals</th>
<th>Accuracy of algorithm (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Montreal Fuzzy EMGlab</td>
</tr>
<tr>
<td>Quadrifilar 10% MVC</td>
<td>100 [98, 100] ** 100 [92, 100]</td>
</tr>
<tr>
<td>20% MVC</td>
<td>100 [98, 100] ** 100 [89, 100]</td>
</tr>
<tr>
<td>50% MVC</td>
<td>99 [93, 100] 97 [86, 100]</td>
</tr>
<tr>
<td>Single channel</td>
<td>100 [100, 100] 96 [84, 100] 100 [98, 100]</td>
</tr>
</tbody>
</table>

Fig. 1. Agreement between the Fuzzy Expert and Montreal algorithms as a function of SNR$_{MU}$ for the experimental quadrifilar data. Each point represents one MUAPT. Results shown separately for each MVC level, and for all levels combined. Best fit exponential model shown in each plot, along with the RMS fit error.

Fig. 2. Accuracy with respect to the true decomposition for the simulated quadrifilar data as a function of SNR$_{MU}$ for the Montreal algorithm. Each point represents one MUAPT. Results shown separately for each MVC level, and for all levels combined. Best fit exponential model shown in each plot, along with the RMS fit error.

General statistical comparisons of agreement and accuracy are shown in Table II. Table II also indicates statistically significant differences in results from one-way ANOVA comparisons and post hoc Tukey tests. For the quadrifilar simulation results only, paired t-tests examined statistical differences between the Montreal and Fuzzy Expert algorithms, at each MVC level. Agreement generally decreased with MVC level for the multiple-channel experimental data. The higher contraction data exhibited a substantial number of superimpositions (particularly at 50% MVC), which is not reflected in the SNR$_{MU}$ measure. Additionally, the higher-level contractions contained substantial smaller-amplitude “background” MUs that were not detected and, thus, contributed to an increased noise floor. For the quadrifilar simulation, Table II further shows that the Montreal algorithm was significantly more accurate than the Fuzzy Expert algorithm at 10% and 20% MVC, although both algorithms performed quite well.

Fig. 1 shows agreement results (Montreal vs. Fuzzy Expert) vs. SNR$_{MU}$ for the experimental quadrifilar data. Figs. 2 and 3 show accuracy vs. SNR$_{MU}$ for the simulated quadrifilar data, both as a function of MVC level and combined across levels. Similarly, agreement and accuracy results for experimental and simulated single channel data are shown vs. SNR$_{MU}$ in Fig. 4.
Each plot in Figs. 1–4 also shows the best-fit exponential model. Quantitatively, it is anticipated that agreement/accuracy is associated with SNR_MU. Here, that relation is expressed by the goodness-of-fit of the exponential model, also listed in the plots. In general, agreement/accuracy increased with SNR_MU.

IV. DISCUSSION

This study evaluated the agreement between pairs of automated decomposition algorithms when applied to experimental data, as well as the accuracy of these algorithms when applied to simulated data, each as a function of SNR_MU. This study provides companion results to a prior full report that appeared in [25], using the DI. As such, we will concentrate our discussion on comparison of the SNR_MU-based results in the present paper to the equivalent results based on the DI. Note that our prior full report utilized a shorter duration match window for MU comparisons (±0.5 ms), analyzed additional data not discussed herein and provided additional analysis not directly related to the SNR_MU and DI measures.

Figs. 1–4 in the present paper (SNR_MU-based analysis) show a strong relationship between SNR_MU and agreement/accuracy, suggesting that high accuracy is most probably achieved whenever the SNR_MU is high. Yet, these figures also show that the RMS errors from the best fit exponential model between agreement/accuracy and SNR_MU were in the range of 3.7–19.3%, depending on MVC level and electrode recording type. This RMS error seems moderately high, indicating that SNR_MU does not account for all of the factors that affect decomposition accuracy.

Comparison of the SNR_MU-based results in Figs. 1–4 to the DI-based results shown in the prior full report (see corresponding Figs. 1–4, respectively, in [25]) shows that nearly identical trends are found. It is, in fact, difficult to note performance differences between the SNR_MU-based and DI-based models. Thus, the simpler SNR_MU measure seems to predict achieved agreement/accuracy equally as well as the more complex DI measure. Future research might examine whether the combined use of these (and other) measures might provide better estimation of achieved agreement/accuracy.

In summary, this companion study provides a systematic comparison of agreement/accuracy performance between three publicly available algorithms which perform decomposition on indwelling EMGs, as a function of SNR_MU. For experimental data, RMS errors between the achieved agreement and those predicted by an exponential model as a function of SNR_MU ranged from 8.4% to 19.2%. For the simulations, RMS errors between achieved accuracy and those predicted by the SNR_MU exponential model ranged from 3.7% to 14.7%. Agreement/accuracy was strongly related to SNR_MU. Prediction of agreement/accuracy based on SNR_MU was essentially equivalent in performance to prediction based on DI.