Chapter 2

Background

This chapter gives an overview of the background knowledges in order to understand the work in this thesis in Chapter 3 and Chapter 4.

Firstly, the physiology of the EMG signal is introduced. This helps the readers to understand the characteristics of the EMG signal, and the correlation between the EMG signal and the neural activity of the spinal cord. With this understanding, readers can have a clear view of the challenges and motivations behind the work in this thesis. In addition, the assumptions made in order to derive the new detection and segmentation techniques in Chapter 3 and Chapter 4 are based on the understanding of the physiology and observation of the EMG signal.

After that, the background knowledge in order to understand the detection technique in Chapter 3 is presented. The methodology consists of a combination of several techniques stemming from multi-resolution wavelet decomposition, robust statistics, and the detection theory. Since the task is to detect transient muscle responses, it's apparent to review the classical detection theory. The classical detection theory cannot be applied directly to solve the problem, but it helps readers understand the proposed methods from a theoretical point of view. In particular, the proposed method involves the use of the generalized likelihood ratio test. Therefore, the classical likelihood ratio test (formally known as the Neyman-Pearson Theorem) is introduced. The use of the wavelet transform was inspired by many literatures in the transient detection field. Detection of the transient signals with unknown structure is generally hard, but employing certain transformation on the signal can yield good detection result, given that the transformation exposes the unique structure of the signal. Therefore, a review on the wavelet theory is given, and specifically a study of the frequency properties of the wavelets is presented, since the proposed detection method makes use of them.

2.1 Physiology of the Electromyographic (EMG) Signal

Electromyography, or EMG for short, is one of the electrophysiological recording methods. Many people are probably more familiar with other electrophysiological recording methods: Electroencephalography (EEG), the recording of electrical activity along the scalp, and Electrocardiography (ECG, or EKG), the recording of the electrical activity of the heart, etc. Similarly, EMG is a technique for evaluating and recording the electrical activity produced by skeletal muscles. The instrument for performing EMG is called Electromyograph, and the record produced is called Electromyogram, or the Electromyographic signal (the EMG signal for short)

The EMG signal is essentially the voltage fluctuation resulting from ionic current flows across the membranes of the muscle cells, when these cells are electrically or neurologically activated. Therefore, from the EMG signal, one can analyze the underlying biological processes of muscles. From that, one can further infer the neural activity of the spinal cord and potentially the central nervous system. The EMG signal can be analyzed to diagnose neuromuscular deficiencies such as caused by stroke and Parkinson's disease [50], the biomechanics of human or animal movement.

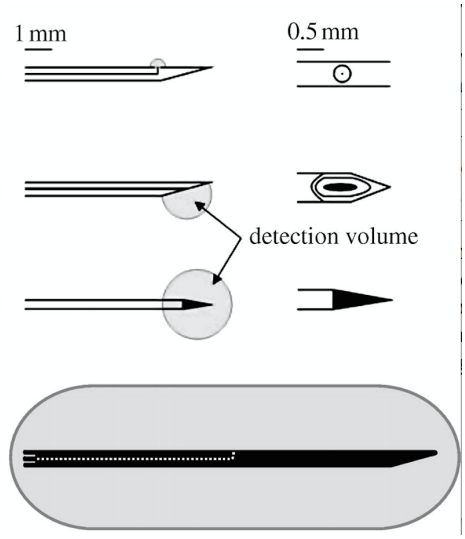
EMG can be categorized into two kinds, *surface EMG* and *intramuscular EMG*, based on the electrodes being used (See Figure 2.1-1). In Surface EMG (*sEMG* for short), a pair of electrodes or a more complex array of multiple electrodes is placed on the surface of the skin above the muscle; while in intramuscular EMG, (*iEMG* for short), typically either a monopolar or concentric needle electrode is inserted through the skin into the muscle tissue. To perform iEMG, special treatment needs to be taken, while sEMG is a non-intrusive, relatively simple approach. On the other hand, iEMG electrodes can be placed much closer to the muscle of interest, while the sEMG signal is influenced by the depth of the under-skin tissue at the site of the recording. Because of this difference, iEMG results in a much more selective, less noisy recording [53].

2.1.1 Physiology of the Generic EMG signal

Stashuk's review paper [53] on the EMG signal decomposition gives a very good explanation of how the EMG signal is generated. The following materials follow the discussion in that paper.

2.1.1.1 Muscle Fiber Action Potential (MFAP)

Muscle fibers are simply the colloquial term for muscle cells, or myocytes, which are the individual components constituting skeletal muscles. Skeletal muscle is subdivided into parallel bundles of stringlike fascicles, which themselves are bundles of even smaller stringlike multinucleated cells, the muscle fibers. Muscle fibers typically have a length of $2-6cm$, and a diameter of $50-100\mu m$ [26]. Each muscle fiber is normally innervated by only one motor neuron in only one place, usually near its



(a) Schematics of typical intramuscular electrodes [35]

(b) Picture of surface electrodes from Motion Lab Systems [56]

Figure 2.1-1: Intramuscular and surface EMG electrodes

midpoint [26]. *Neuromuscular junction* is the structure through which a motor neuron innervates its muscle fiber. When a muscle fiber is excited, it fires action potentials propagating relatively slowly $(3 - 5\text{m/s})$ in both direction away from the neuromuscular junction, similarly to the propagation of action potentials (AP) along the axons of neurons. This action potential is called a *muscle fiber action potential* (MFAP), and is the fundamental component contributing to the detected EMG signal. The characteristics of MFAPs will depend upon the diameter of the fiber, the conduction velocity, its location relative to the detection site, and the configuration and type of electrodes.

2.1.1.2 Motor Unit Action Potential (MUAP)

The fibers of a muscle are not excited individually. They are controlled together in a group, called the *motor unit*. A motor unit is made up of a single motor neuron and the skeletal muscle fibers innervated by that motor neuron. A typical muscle is controlled by about 100 large motor neurons [26]. A motor unit can innervate anywhere from 100 to 1000 muscle fibers scattered over a substantial part of the muscle. All of the muscle fibers innervated by the same motor neuron respond faithfully and synchronously to each action potential of the motor neuron [26]. As a result, individual MFAPs are normally not detected. Instead, a summation of all of a motor unit's MFAPs is detected, known as a *motor unit action potential* (MUAP).

Let $\text{MFAP}_i(t)$ be the waveform of a muscle fiber action potential from the i-th fiber of a motor unit. Let $\text{MUAP}_j(t)$ be the electrical potential from the j-th motor unit, which arises as a sum of all MFAPs:

$$\mathrm{MUAP}_j(t) = \sum_{i=1}^{N_j} \mathrm{MFAP}_i(t - \tau_i) s_i \tag{2.1.1.1}$$

where τ_i is the temporal offset of $\mathrm{MFAP}_i(t)$, and N_j is the number of fibers in motor unit j. The binary variable s_i represents the neuromuscular junction function that has a value of 1 if fiber i fires and 0 if not.

τ_i depends on the location of the neuromuscular junction and the conduction velocity of the muscle fiber. N_j represents the size of the motor unit. As pointed out before, $N_j \sim 100 - 1000$. Because a single action potential in a motor neuron can activate hundreds of muscle fibers in synchrony, the resulting currents sum to generate an electrical signal that is readily detectable outside the muscle itself [26]. Because of the attenuation of MFAP with distance to the detection electrode, the size of the MUAP is in practice often dependent on the location and diameter of the closest few muscle fibers. Figure 2.1-2 depicts the composition of a MUAP as the summation of individual MFAPs.

In general, MUAP waveforms will vary in shape due to variations in the delays of the fiber potentials (affecting τ_i), possible changes in the position of the electrode relative to the muscle fibers (affecting MFAP_i), and the possibility of a particular fiber failing to fire (affecting s_i). These variations are the source of stochastic biological variability in the MUAP waveform [53].

2.1.1.3 Motor Unit Action Potential Train (MUAPT)

In order to maintain or increase the force generated by a muscle, the specific motor neuron must fire a temporal sequence of action potentials, called a *spike train*. As discussed in last section, one action potential from a single motor neuron results in one MUAP. Therefore, this spike train, when arriving at the neuromuscular junctions of all muscle fibers of this motor unit, results in a temporal sequence of MUAPs, called *Motor Unit Action Potential Train* (MUAPT) [53].

$$\mathrm{MUAPT}_j(t) = \sum_{k=1}^{M_j} \mathrm{MUAP}_{jk}(t - \delta_{jk}) \tag{2.1.1.2}$$

where $\mathrm{MUAPT}_j(t)$ is the MUAPT of the j-th motor unit, $\mathrm{MUAP}_{jk}(t)$ is the MUAP generated during the k-th firing of the j-th motor unit, M_j is the number of times the j-th motor unit fires, and δ_{jk} is the k-th firing time of the j-th motor unit.

2.1.1.4 Composite EMG signal

When more than minimal force is required, many motor neurons generate an asynchronous barrage of action potentials. Due to the property of superposition of electric fields, an electrode, either inserted into a muscle or on the surface of the skin, measures the spatial and temporal sum of MUAPTs contributed from all recruited motor units within the "listening sphere". The result is a

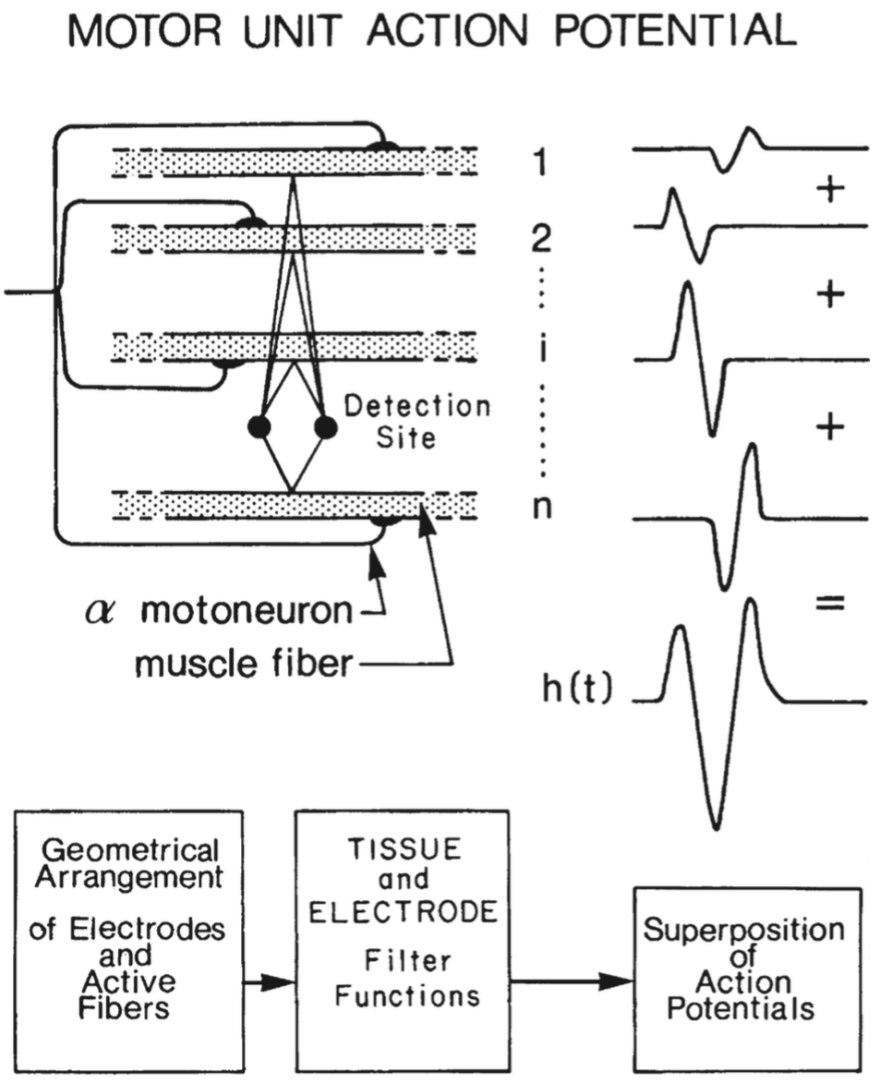


Figure 2.1-2: A MUAP is composed of the summation of the MFAPs of its component muscle fibers. (from Stashuk [53])

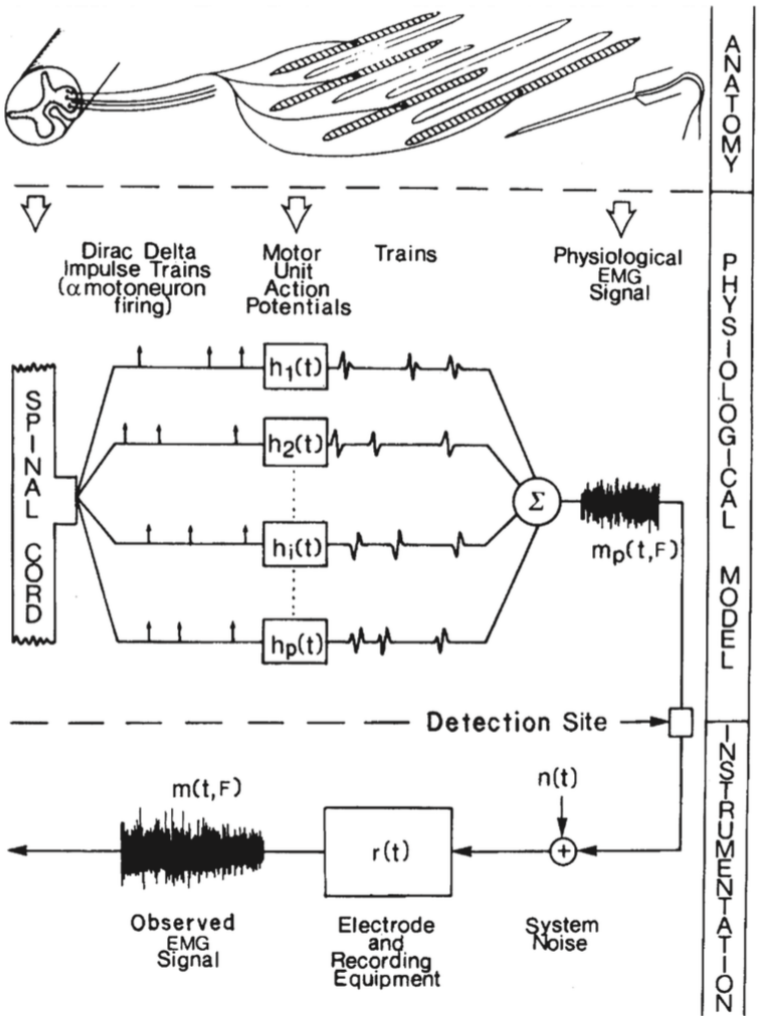


Figure 2.1-3: Physiological and mathematical model for the composition of a detected EMG signal (from Stashuk [53]).

complex pattern of electric potentials (typically in the order of $100\mu V$ in amplitude) that is called the *composite* EMG signal [26]. Figure 2.1-3 presents both an anatomical and physiological model of an EMG signal.

$$\mathrm{EMG}(t) = \sum_{j=1}^{N_m} \mathrm{MUAPT}_j(t) + n(t) \tag{2.1.1.3}$$

where $\mathrm{MUAPT}_j(t)$ is the j-th MUAPT, N_m is the number of active motor units, and $n(t)$ is the background instrumentation noise.

Normally, more motor units are recruited as the muscle force increases. Different motor units are recruited at different times and stay active for different lengths of time. In addition, each MUAPT has its own characteristics of firing intervals, and this firing interval changes within each MUAPT, too. A general research direction is to decompose the detected EMG signal into its MUAPTs from

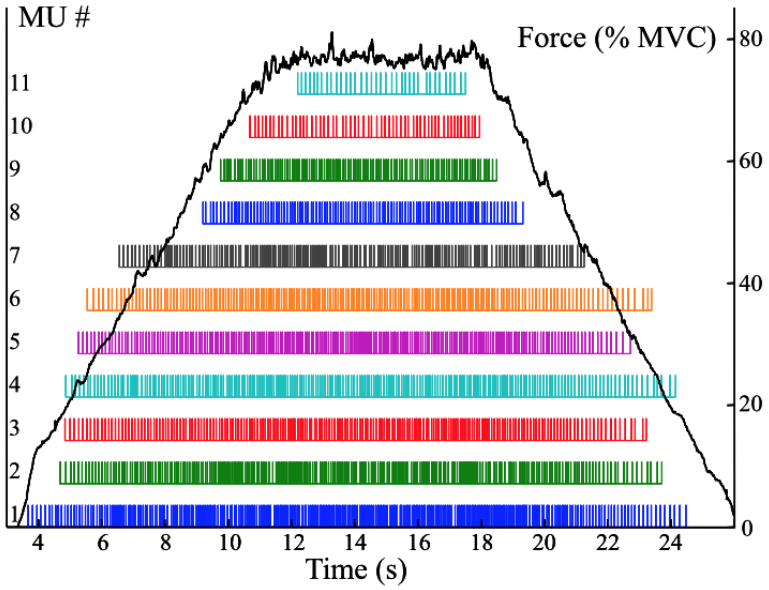


Figure 2.1-4: Bar plot for the firing times obtained via the decomposition method in [38]. MU: Motor Unit; MVC: Maximum Voluntary Contraction). (from Nawab [38])

different motor units. EMG decomposition is normally performed on the iEMG signal, since iEMG measures a few MUAPTs while sEMG detects many more, making decomposition very difficult. An example of the decomposition result on the iEMG signal from Nawab's paper [38] is shown in Figure 2.1-4.

The sEMG signal can reveal important muscle excitation information about underlying limb movement. As a result, a typical research direction is to detect muscle activation intervals in the sEMG signal. Figure 2.1-5 gives an example of muscle activity onset detection using an energy detector in [50].

As mentioned before, the shape of MUAPs depends on many different factors, such as the position of the electrode relative to the active muscle fibers, the physical characteristics and configuration of the electrodes. In addition, an EMG signal is composed of temporal overlapping of different MUAPs. As a result, it's hard to predict the actual shape of the EMG signal. This property is the major challenge in EMG processing. However, the good news is that no matter how variable the shapes can be, the effective bandwidth of the EMG signal can be assumed as prior knowledge of the physiology of EMG, as shown in Figure 2.1-6. This prior knowledge will be used in developing the EMG detection method.

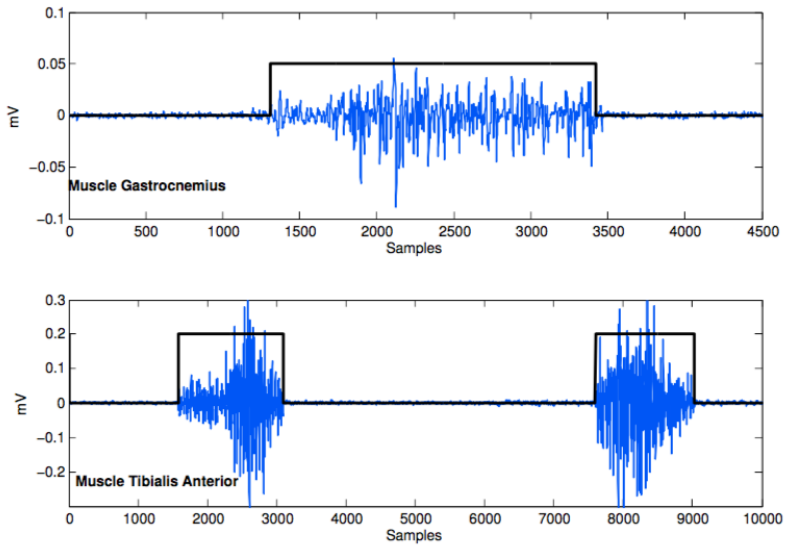


Figure 2.1-5: Muscle activity onset detection result for clinical EMG signal (from Rasool [50])

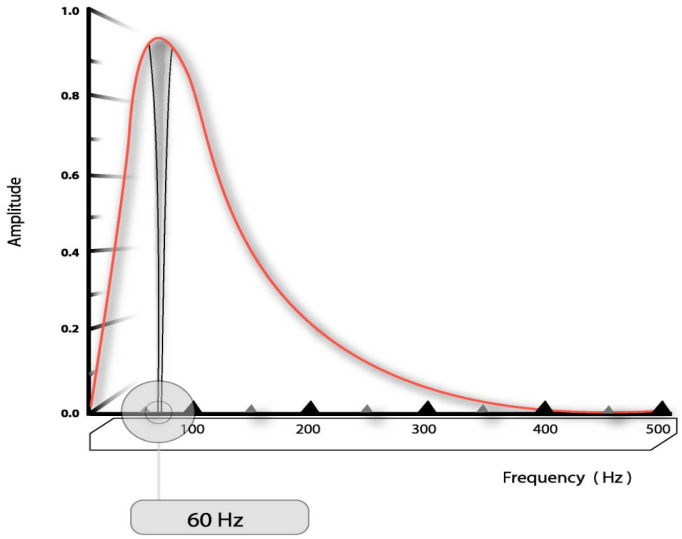


Figure 2.1-6: Schematic representation of a typical sEMG power spectrum (from Day [10])

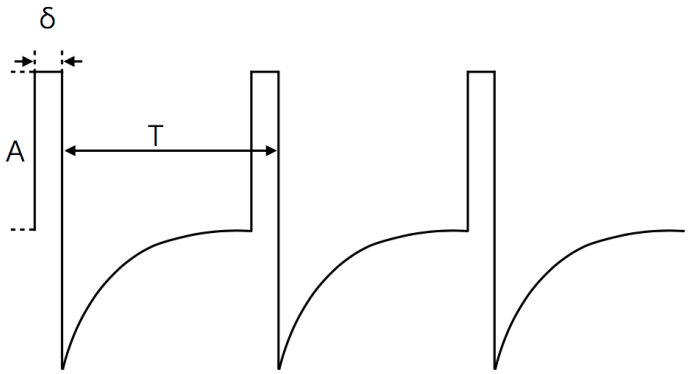


Figure 2.1-7: Schematic of electric stimuli (the actual shape of the stimulus may look different.)

2.1.2 EMG signal Resulting from Electro-stimulation: Motor Evoked Potentials

The EMG signal obtained from patients with SCI in rehabilitation training is different from the EMG signal of normal healthy spinal cords undergoing the same motions, although the fundamental physiology is similar. Recall from the previous discussion, patients with complete SCI lose all sensation and voluntary movement control below the injury level. This is because the information pathway is blocked between the brain and the neurons of the spinal cord below the lesion. As a result, the brain can no longer send or receive information from certain parts of body. Although certain locomotion control, such as stepping and standing, is governed in part by the neural circuitry within the spinal cord, this neural circuitry becomes silent after the SCI because it needs modulation and stimulation from the brain to function properly. Electro-stimulation (ES) therapy is based on the belief that this neural circuitry is intact and can resume working if given proper electrical stimulation and rehabilitation training due to plasticity of the neurons of the spinal cord. Specifically, an electrode array was implanted over the spinal cord segments to stimulate the spinal cord neurons. The electric signal can be thought of as a spike train, similar to the action potential train found in neurons. Figure 2.1-7 gives a schematic of the electric stimuli. The actual shape of one stimulus, though it may differ from the drawing, is a biphasic waveform. Each stimulus is a very short pulse, and it is repetitive with a given frequency. Many parameters associated with the stimulation can be adjusted, such as the pulse width δ, the frequency $1/T$, the amplitude A, the electrodes configuration, and electrode polarity (shown in Figure 1.2-2).

To provide some biology background in the following discussion, electric signaling in neurons is first explained.

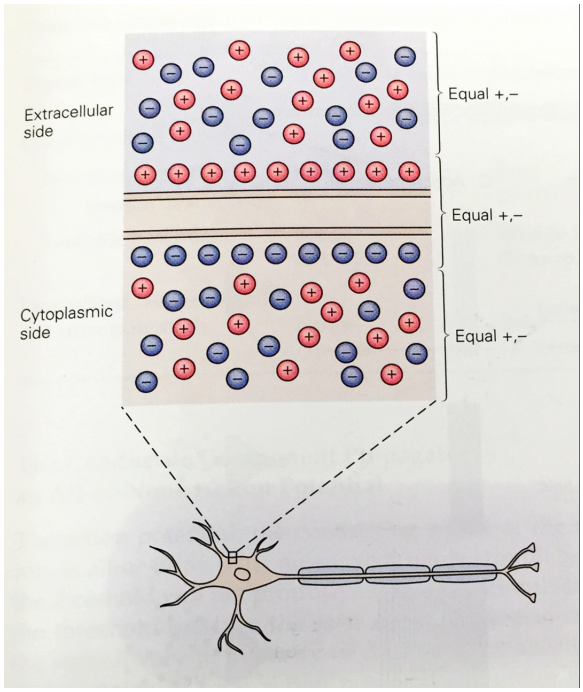


Figure 2.1-8: The membrane potential of a cell results from a difference in the net electric charge on either side of its membrane. When a neuron is at rest, there is an excess of positive charge outside the cell and an excess of negative charge inside it. (from Kandel [26])

2.1.2.1 Signaling in Neurons

At rest, all cells, including neurons, maintain a difference in the electric potential across the cell membrane. This is called the *resting membrane potential*. At rest, there are more negative charges at the cytoplasmic side, while there are more positive charges at the extracellular side (See Figure 2.1-8). By default, the membrane potential is defined as the difference obtained by subtracting extracellular potential from cytoplasmic potential. Hence, the resting membrane potential is a negative value (typical value for neurons is -65mV, typical value for muscle cells is -90mV) [26].

Excitable cells, such as neurons and muscle cells, differ from other cells in that their membrane potentials can be significantly and quickly altered; this change can serve as a signaling mechanism. The change in the membrane potential can be either a decrease or increase from the resting potential. The resting membrane potential provides the baseline: a reduction in membrane potential is called *depolarization*. Because depolarization enhances a cell's ability to generate an action potential, it is *excitatory*; an increase in membrane potential is called *hyperpolarization*. Hyperpolarization makes a cell less likely to generate an action potential and is therefore *inhibitory*. There are typically four components associated with the electric signaling in neurons and muscle cells. The four components in the list below are only an abstraction of the four functionality. Different cells have different structures and mechanisms. Figure 2.1-9 show an example of the signaling in a sensory neuron.

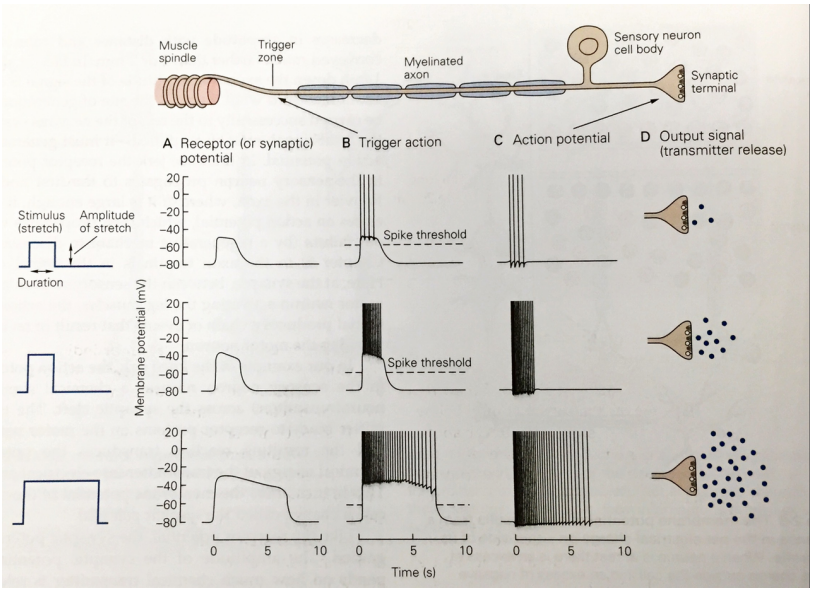


Figure 2.1-9: A sensory neuron transforms a physical stimulus (a stretch in this example) into electric signals in the neuron. Each of the neuron's four signaling components produces a characteristic signal. (from Kandel [26])

Input : Input component produces graded local signals. This signal passively propagates to other parts of the cell.

Trigger : Trigger component takes consideration of all input signals, and then makes the decision whether or not to generate action potentials.

Conduction : Conductive component actively propagates the action potentials down to the other parts of the cells. Active propagation means the amplitude of the action potentials doesn't diminish over time or distance.

Output : Output component passes the action potentials to other neurons or muscle cells. A *synapse* is a structure that permits a neuron to pass an electrical or chemical signal to another cell (a neuron or muscle cell).

There are 3 main functional groups of neurons in the spinal cord[26]:

Sensory neurons : carry information from the body's periphery into the nervous system for the purpose of perception and motor coordination.

Motor neurons : carry commands from the brain or the spinal cord to muscles and glands.

Interneurons : constitute by far the largest class, consisting of all nerve cells that are not sensory or motor neurons. They form complex neural network that enable complicated logic and decision making.

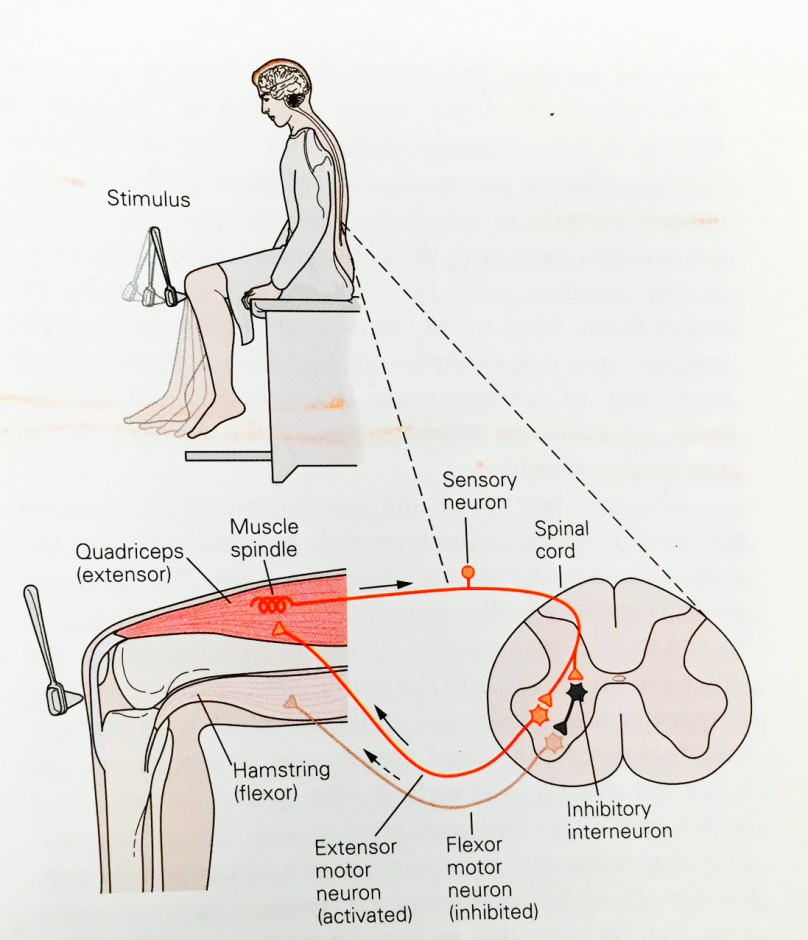


Figure 2.1-10: The knee jerk is an example of a monosynaptic reflex system, a simple behavior controlled by direct connections between sensory and motor neurons. (from Kandel [26])

There are 2 types of neural circuitry in the spinal cord [26]:

Monosynaptic circuits : the sensory neurons and motor neurons executing the action are directly connected to one another, with no interneuron intervening between them.

Polysynaptic circuits : include one or more sets of interneurons; are more amenable to modifications by the brain's higher processing centers.

Figure 2.1-10 shows the reflex mechanism of knee jerk. In this example, The extensor motor neuron is connected directly to the sensory neuron, thereby forming a *monosynaptic* circuit. It becomes active when sensory neuron is active. On the other hand, the flexor motor neuron is connected to the sensory neuron via an inhibitory interneuron, thereby forming a *polysynaptic* circuit. As a result of the inhibitory interneuron, the flexor motor neuron becomes inhibited (or inactive) when the sensory neuron is active. Overall, the extensor and the flexor motor neurons are coordinated by interneurons.

2.1.2.2 Motor Evoked Potential

Now let's come back to the EMG signal generated from patients with SCI under electrical stimulation. What happens to the neurons in the spinal cord under electrical stimulation is still an ongoing research. Here, only the fundamentals are introduced.

From previous discussion, it is shown that the action potentials in neurons can be generated when the membrane potential of the trigger zone or axon of a neuron depolarize to a certain threshold. The external electric field can affect different parts of the neurons in order to drive action potentials. Here let's focus the discussion on axons, as this is described in [26]. Again, this is only a postulate. At the presence of the external electric field, the current needs to pass through the cell membrane in order to drive a cell to threshold. In the vicinity of the positive electrode, current flows across the membrane into the axon. It then flows along the axoplasmic core, eventually flowing out through more distant regions of axonal membrane to the negative electrode in the extracellular fluids. Not all currents pass through the cell membranes; in fact, a lot more of the stimulating current move instead through the low-resistance pathway provided by the extracellular fluid. The axons with lower axial resistance to the flow of longitudinal current can pass more currents, and as a result can depolarize more efficiently. Normally, axons with larger diameters have lower axial resistance. If an axon depolarizes beyond threshold, it will then fire and propagate action potentials. This resulting action potential is called *compound action potential*.

If the external electric field *directly* excites a motor neuron (e.g., by depolarizing its axon), then the motor neuron fires an action potential and propagates it down to its muscle fibers. The result is one MUAP. I borrow the terminology from reflex physiology and call it a *monosynaptic* MUAP. If the external electric field *indirectly* excites a motor neuron, either by exciting its presynaptic interneurons (the interneurons that transmit signals to this motor neuron), or by modulating the presynaptic input signals from sensory neurons, then the motor neuron also fires an action potential and results an MUAP. I call it a *polysynaptic* MUAP.

Usually, the external electric field directly excites more than one motor neuron. All excited motor neurons will approximately fire action potentials *synchronously* (synchronized with the electrical stimuli). This is the key difference between the EMG signal from patients with SCI under electrical stimulation and the generic EMG signal. In a healthy spinal cord, when multiple motor units are recruited, they fire action potentials *asynchronously* because each motor neuron is modulated via its own complex neural circuitry formed by a large number of interneurons. Regarding the EMG signal that resulted from the electrical stimulation, because of the synchrony, all of the monosynaptic MUAPs from multiple motor units overlap with each other, and produce one large response, which I call the *monosynaptic response*.

The indirect excitation of a motor neuron is harder than direct excitation, because when a motor

neuron is excited via all the synapses from all presynaptic interneurons, there need to be enough interneurons, and all the interneurons need to be coordinated properly. For example, normally, when a motor neuron is excited, its excitatory presynaptic interneuron needs to be active while its inhibitory presynaptic interneuron needs to be inactive. If both the inhibitory and excitatory interneurons are active, the motor neuron won't be excited. In the case of external electric stimulation, it's typically hard to coordinate this kind of activity among the interneurons. This is why various configurations of electrodes and different parameters of the stimuli are chosen in order to achieve certain neural activity within the spinal cord. The overlapping response from all the polysynaptic MUAPs is called a *polysynaptic response*, and is much weaker than the monosynaptic response, since much fewer motor neurons are indirectly excited. Because of the complex neural pathway between the origin of the compound action potential to the motor neuron, a polysynaptic response arrives later than a monosynaptic response, and is less synchronous with the electrical stimulus. Thereby, a monosynaptic response is also sometimes referred to as an *early response*; a polysynaptic response is referred to as a *late response*.

Collectively, both the monosynaptic response and the polysynaptic response are called the *Motor Evoked Potential* (MEP). A MUAP is no longer a proper term in the EMG signal in this thesis, as both monosynaptic and polysynaptic responses are somehow synchronized to the external electric stimuli, and there is hardly an individual MUAP in the resulting EMG signal. In this thesis, MEP refers specifically to the compound response from the patients with SCI under the electrical stimulation. Figure 1.3-2 gives an example of both an early response and a late response after one electrical stimulus (the stimulus is at the beginning of the plotted signal). As you can see from the example, the early (monosynaptic) response is much stronger than the late (polysynaptic) response.

2.2 Characteristics of MEPs and Challenges of Processing Them

The first part of this section shows some of the major characteristics of the EMG signal with example figures. With the physiology background discussed above, the readers can have a deep understanding of the characteristics. Next, the challenges arising from these characteristics will be listed, along with the major insufficiencies of some prior work.

2.2.1 Characteristics of MEPs

The biggest challenge is the randomness of the MEPs. The parameters of a MEP are not deterministic, and must be modeled by random variables. The actual probability model and its parameters

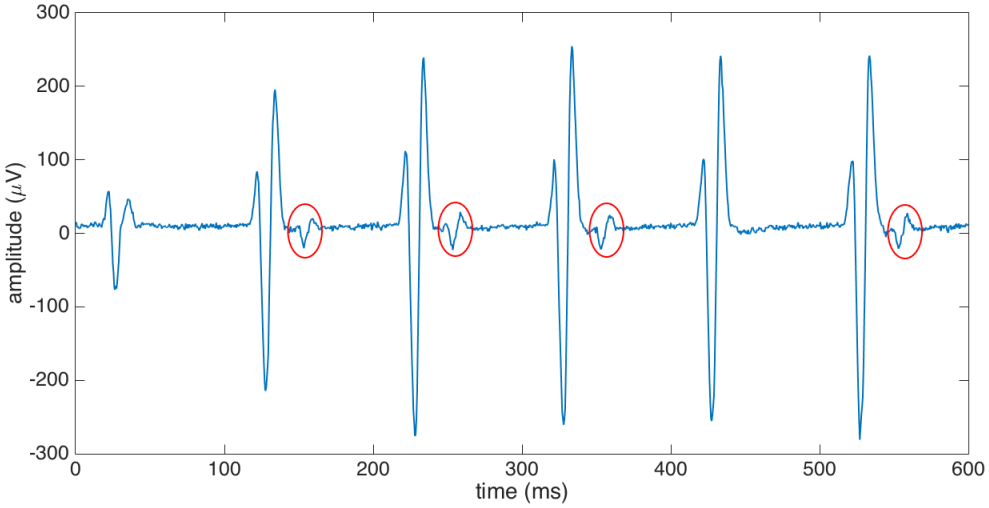


Figure 2.2-1: Example EMG signal containing MEPs with low SNR (marked by red circles). (The example EMG signal is from the muscle of left medial gastrocnemius while the patient is lying in supine position under EES.)

are also unknown, which makes the problem even harder.

Figure 2.2-1 shows that the arrival times of the MEPs are random. The MEPs in the red circles are polysynaptic ones (the weak ones), while the rest are monosynaptic ones (the strong ones). The arrival times of the monosynaptic MEPs, although still random, show certain regular pattern, so a reasonably good probability model can be sufficient. However, the arrival times of the polysynaptic MEPs have limited pattern, and a proper probability model is therefore hard to determine. As pointed in the physiology section of this chapter, the arrival time of a MEP depends on many different factors, such as the conduction of the motor neuron axons, the detection location relative to the neuromuscular junctions, and the overall conduction speed of the muscle fibers. Moreover, in the case of polysynaptic MEPs, the neural pathway between the source of the compound action potential and the excited motor neuron is very complex and totally unpredictable.

Figure 2.2-2 shows different waveforms of the MEPs from different muscles of one training session. Figure 2.2-3 shows different waveforms of the MEPs from one single muscle of one training session. The two figures show that the durations and the shapes of the MEPs vary a lot. The lack of information on the structure of a signal leads to great difficulty in processing it. Later in this section, this difficulty will be elaborated within the context of prior work. There are many different factors contributing to the varying shapes of the MEPs. As shown in the physiology section, every MUAP consists of multiple MFAPs. MUAP waveforms vary in shape due to variations in the delays of the MFAPs, and the number of muscle fibers that fire. In addition, there is a lot of substances between the muscle fibers and the detection site, including a layer of fat tissues and skin. All these substances

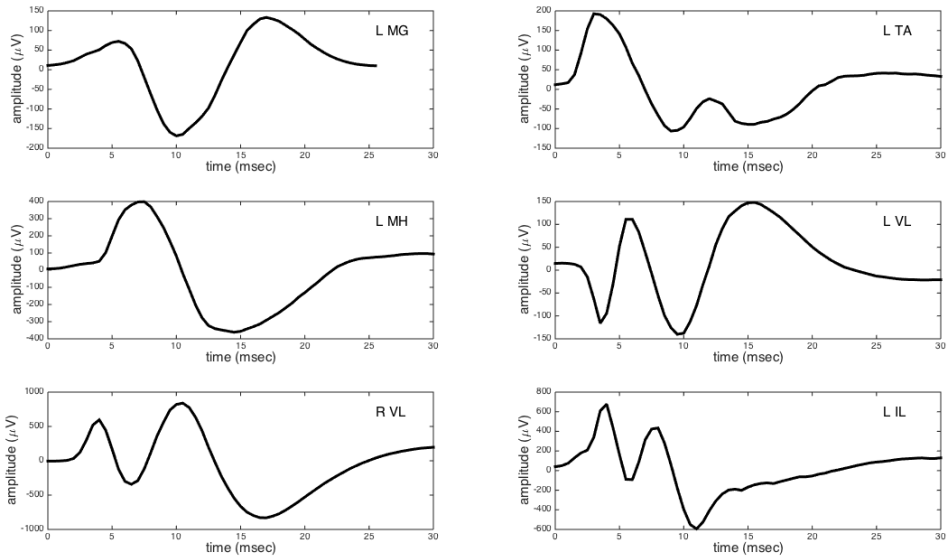


Figure 2.2-2: Examples of MEP waveforms from different muscles of one patient under one rehabilitation session. The muscle from which each MEP waveform is from is shown in its short name on the upper right corner of each subplot. For the full names of the muscles, refer to Appendix A. (The example EMG signal is from various muscles while the patient is lying in supine position under EES.)

degrade the MUAPs significantly. The configuration of the electrodes can also alter the waveform in an unpredictable way. Totally, the shapes of the resulting MUAPs are completely random and unpredictable. Moreover, an MEP is a superposition of multiple MUAPs. In the rehabilitation training with electrical stimulation, the spinal cord is damaged, and hence there is little to know about how many motor neurons are excited and which they are. This further adds to the complexity of the shapes of the MEPs. The polysynaptic MEPs have a even less regular structure than the monosynaptic ones, because the complex neural pathway results in a complicated, asynchronous overlapping of the MUAPs. The only common feature from all the MEPs is that they all contain multiple transient peaks, although the shapes of the peaks, such as the widths and heights, are still random. Also the number of peaks within one MEP is unknown.

Figure 2.2-1 and Figure 2.2-4 show that the polysynaptic MEPs have an extremely low signal-noise ratio (SNR). This is another major difference between MEPs found in the electro-stimulation induced EMG signal and the MUAPs found in the generic EMG signal. Due to the spinal cord injury, lots of the neurons in the spinal cord below the lesion are inactive. A lot less motor neurons can fire from the excitation of their presynaptic interneurons. As a result, the polysynaptic MEPs are very weak.

The last characteristic of the EMG signal is shown in Figure 2.2-5. As commonly seen in the recording of any electrical signals, the EMG signal suffers from the baseline fluctuation.

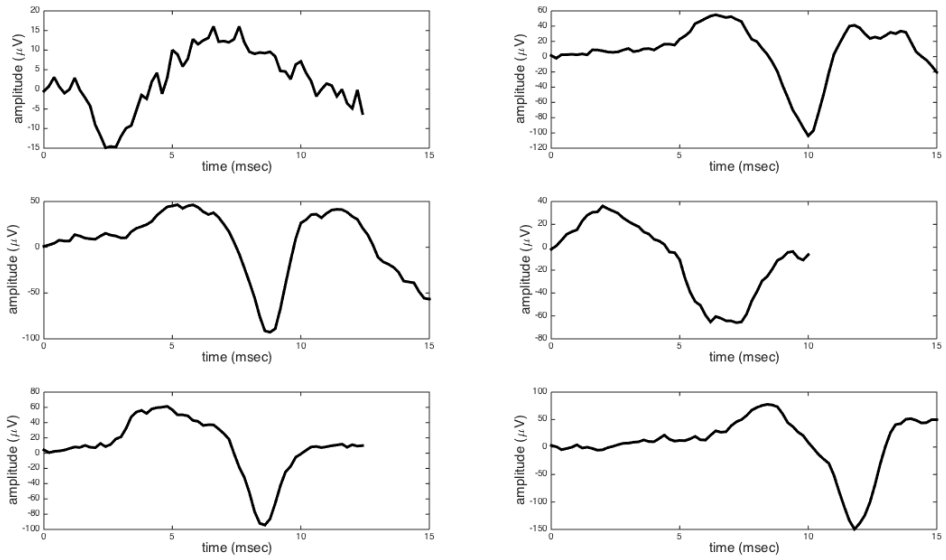


Figure 2.2-3: Examples of MEP waveforms from the same muscle. (The example EMG signal is from the muscle of left medial gastrocnemius while the patient is lying in supine position under EES.)

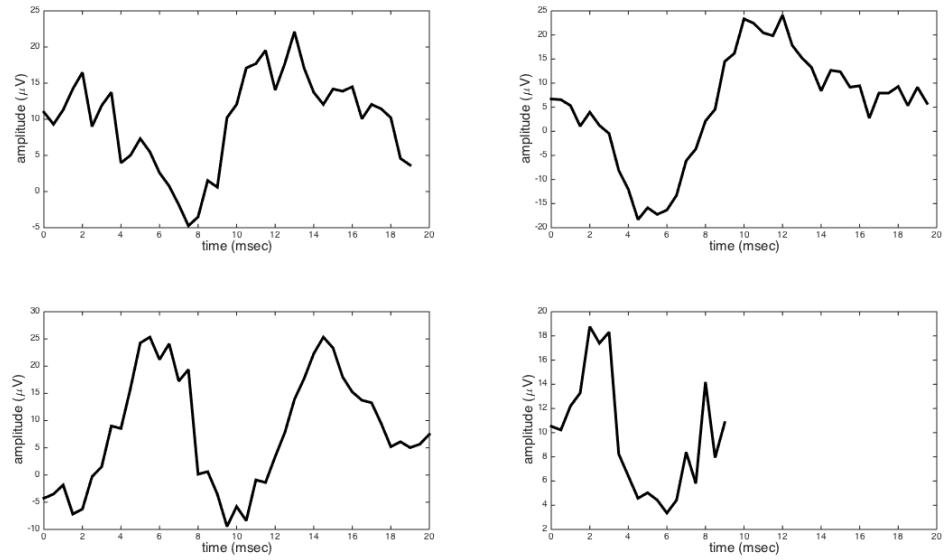


Figure 2.2-4: Examples of MEP waveforms with low SNR. (The example EMG signal is from the muscle of left medial gastrocnemius while the patient is lying in supine position under EES.)

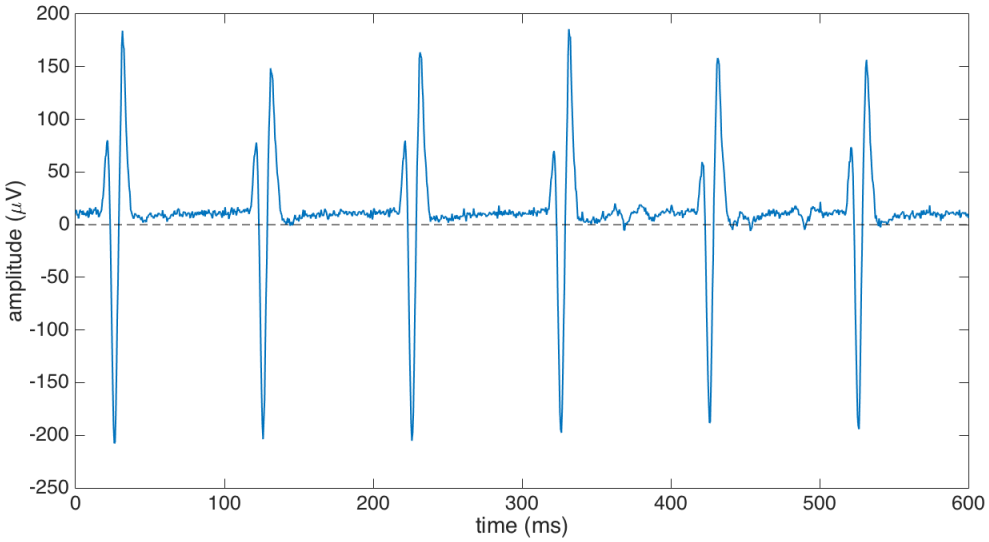


Figure 2.2-5: Examples of the baseline fluctuation in the EMG signal. The baseline deviates from 0 (marked by a dashed horizontal line), and changes slowly over time. (The example EMG signal is from the muscle of left medial gastrocnemius while the patient is lying in supine position under EES.)

All the previous discussion is limited to the variations of the MEPs within one person from one training session. In practice, all the characteristics of the MEPs also vary from session to session, and from person to person. The spinal cord is injured differently in different patients. The strength of the muscles also varies a lot among individuals. Finally, in every training session, the EMG electrodes are placed by practitioners manually, resulting in different detection sites. The location of the detection site also has an impact on the shapes of the MEPs.

2.2.2 Challenges of Processing MEPs

Detailed literature reviews are given within Chapter 3 and Chapter 4. A brief introduction to the prior work is presented here as a context for the discussion of the challenges in the EMG processing.

Traditionally, the detection of signals in a noisy observation is formulated as a binary hypothesis testing problem in the detection theory. When the probability models of the signal and the noise are fully known, an optimal detector with constant false-alarm rate can be formulated as a likelihood ratio test according to Neyman-Pearson Theorem [28]. Please refer to Section 2.3 for an overview of the binary hypothesis testing problem and the likelihood ratio test. For example, when the shape of the signal is completely known, then a matched filter gives the optimal detection performance [28]. Normally, a matched filter is used to give the upper bound on the detection performance. On the

contrary, if there is absolutely nothing known about the signal, then an energy detector gives the optimal detection performance [28]. An energy detector normally serves as the lower bound when evaluating a detector. When certain parameters in the model of either the signal or the noise are unknown, then a generalized likelihood ratio test can be formulated with the parameters being their maximum likelihood (ML) estimates [28]. The shapes of the MEPs are unknown, so a matched filter cannot be applied directly. The use of an energy detector is insufficient because it doesn't use the structure of the MEPs at all and hence gives the worst detection performance. The idea is to find a certain representation of the EMG signal, such that the feature or structure of the MEP is exposed in that representation. Then a binary hypothesis testing problem can be formulated in the new representation.

Many different representations have been proposed in the field of transient detection. In particular, the wavelet transform is proven to be successful for a variety of signals. Wavelet transform gives the local feature of a signal rather than a global feature. So it naturally works well on a transient signal. In addition, Wavelet transform can expose features at different levels by specifying different scale values. Some prior work on the transient-signal detection with wavelet transform includes [21], in which the signal is assumed to have a known shape, but unknown arrival time and scaling. In [19], the signal is unknown, but its bandwidth and time-bandwidth-product are assumed to be known. The methods in the prior work are either insufficient because of the strong constraints made on the signal model, or not applicable to the case of peak detection.

The methodology proposed in Chapter 3 combines detection theory with wavelet transform. As a result, a brief review of some background knowledge in the detection theory and wavelet transform is given in Section 2.3 and Section 2.4, respectively.

2.3 Classical Detection Theory

The first task to address in this thesis is to detect the transient muscle responses (more specifically, the MEP peaks). The detection of the signal corrupted by noise is studied in the detection theory. In classical detection theory, the detection of the signal out of noise is formulated as a binary hypothesis testing problem, so I will first review the binary hypothesis testing problem, and I will introduce the fundamental theorem for solving it: the Neyman-Pearson Theorem. The theorem introduced a detection scheme called likelihood ratio test that gives an optimal detector given the full probability models of the signal and the noise. In the proposed detection method, a generalized likelihood ratio test is used to tackle the problem of unknown parameters in the signal model.

In Chapter 3, the proposed detector is compared against other detectors in the literature. As a result, a review of the performance metrics of binary detection is presented in Section 2.3.2. There are many different evaluation metrics in different applications. This thesis uses two statistics

called *recall* and *precision*, which are widely used in the field of pattern recognition and machine learning. These two statistics are chosen because they give the most important information about the detection performance in this application: detection of transient MEP peaks from an EMG signal. Basically, from recall one can tell how many MEP peaks are detected among all the true MEP peaks, and precision shows how many detected peaks are true MEP peaks. Recall is important because a practical detector sometimes misses a true signal, and precision is important as any practical detector sometimes detects noise as a signal. Other statistics are either equivalent to recall or precision, or less important to the task.

2.3.1 Binary Hypothesis Testing and Neyman-Pearson Theorem

In classical detection theory, a signal-detection problem is often formulated as a binary hypothesis testing problem, where under the null hypothesis $\mathcal{H}_0$ the signal is not present, and under the alternative hypothesis $\mathcal{H}_1$ both the signal and the noise are present.
Suppose N observations $x[n]$, $n = 0, 1, \cdots, N-1$, are generated depending on the hypothesis:

$$\mathcal{H}_0 \ : \ x[n] = w[n] \qquad w[n] \sim \mathcal{N}(0, \sigma^2) \quad i.i.d. \tag{2.3.1.1a}$$

$$\mathcal{H}_1 \ : \ x[n] = s[n] + w[n] \qquad w[n] \sim \mathcal{N}(0, \sigma^2) \quad i.i.d. \tag{2.3.1.1b}$$

where $x[n]$ represents a noisy observation at a discrete time n, $s[n]$ is the transient signal to be detected and $w[n]$ is the background white noise.
A binary detector maps the observation into either $\mathcal{H}_0$ or $\mathcal{H}_1$. If I use notation $P(\mathcal{H}_i; \mathcal{H}_j)$ to represent the probability of deciding $\mathcal{H}_i$ when $\mathcal{H}_j$ is true, then there are four probability associated with a given binary detector.

- $P(\mathcal{H}_0; \mathcal{H}_0)$ = probability of correct non-detection
- $P(\mathcal{H}_0; \mathcal{H}_1)$ = probability of missed detection = P_M
- $P(\mathcal{H}_1; \mathcal{H}_0)$ = probability of false alarm = P_{FA}
- $P(\mathcal{H}_1; \mathcal{H}_1)$ = probability of detection = P_D

When the full knowledge of the statistics of the signal $s[n]$ and the noise $w[n]$ is given, then an optimal detector exists according to *Neyman-Pearson Theorem*:

Theorem 1 (Neyman-Pearson Theorem) *To maximize P_D for a given $P_{FA} = \alpha$, decide $\mathcal{H}_1$ if:*

$$L(x) \stackrel{\text{def}}{=} \frac{p(x; \mathcal{H}_1)}{p(x; \mathcal{H}_0)} > \gamma \tag{2.3.1.2}$$

where the threshold γ is found from:

$$P_{FA} = \int_{\{x:L(x)>\gamma\}} p(x;\mathcal{H}_0)dx = \alpha \tag{2.3.1.3}$$

The Eq. (2.3.1.2) is called *likelihood-ratio test (LRT)*, because the left-hand side $L(x)$ is the ratio of the data likelihood under $\mathcal{H}_1$ over $\mathcal{H}_0$. The detector given by the Neyman-Pearson Theorem is also referred to as the *Constant False Alarm Rate* (CFAR) detection, as the detector maintains a constant P_{FA}.

The Neyman-Pearson theorem can be applied when the statistics of the signal $s[n]$ and the noise $w[n]$ are fully known, so that the likelihood ratio can be analytically derived. When the statistics of the signal $s[n]$ are not completely known, then the *generalized likelihood ratio test* (GLRT) can be formulated as follows.
Suppose the statistics of the signal $s[n]$ depends on the parameter vector $\boldsymbol{\theta}$, then the likelihood ratio is:

$$L(x) \overset{\text{def}}{=} \frac{p(x;\hat{\boldsymbol{\theta}},\mathcal{H}_1)}{p(x;\mathcal{H}_0)} > \gamma \tag{2.3.1.4}$$

where $\hat{\boldsymbol{\theta}}$ is the *maximum likelihood* (ML) estimate of $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(x;\boldsymbol{\theta},\mathcal{H}_1) \tag{2.3.1.5}$$

Two classical detectors will be derived based on Neyman-Pearson theorem. The first one is called *matched filter*, which is derived when the signal is fully known. The other one is called *energy detector*, which is derived when nothing is known about the signal. As a result, matched filter is normally considered as an upper bound of the detection performance of any given detector, while energy detector is used as a lower bound.

In both cases, assume there are N observations $x[n]$, $n = 0, 1, \cdots, N-1$, with noise, $w[n]$, being white gaussian noise with variance σ^2.

2.3.1.1 Matched Filter

When the signal $s[n]$ is deterministic:

$$p(x;\mathcal{H}_1) = \frac{1}{(2\pi\sigma)^{N/2}} \exp\Big[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-s[n])^2\Big] \tag{2.3.1.6}$$

$$p(x;\mathcal{H}_1) = \frac{1}{(2\pi\sigma)^{N/2}} \exp\Big[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\Big] \tag{2.3.1.7}$$

Therefore:

$$L(x) = \exp\Big[-\frac{1}{2\sigma^2}\Big(\sum_{n=0}^{N-1}(x[n]-s[n])^2 - \sum_{n=0}^{N-1}x^2[n]\Big)\Big] > \gamma \tag{2.3.1.8}$$

Take the logarithm on both sides and simple steps yield the log likelihood ratio test:

$$\ln(L(x)) = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}x[n]s[n] - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}s^2[n] > \ln\gamma$$

Since $s[n]$ is known:

$$T(x) \overset{\text{def}}{=} \sum_{n=0}^{N-1}x[n]s[n] > \sigma^2\ln\gamma + \frac{1}{2}\sum_{n=0}^{N-1}s^2[n] \overset{\text{def}}{=} \gamma\prime$$

Or:

$$T(x) = \sum_{n=0}^{N-1}x[n]s[n] > \gamma\prime \tag{2.3.1.9}$$

$T(x)$ is called the *test statistic*, as used in statistical hypothesis testing. The test statistic in Eq. (2.3.1.9) is obtained by correlating a known signal, or template, with the observation, and is therefore called the *matched filter*. It is also sometimes referred to as the *replica-correlator*.

2.3.1.2 Energy Detector

When nothing is known about the signal $s[n]$, the parameter θ as in the generalized likelihood ratio test is the signal itself, $\boldsymbol{\theta} = \mathbf{s}$. As a result:

$$\begin{aligned}\hat{\mathbf{s}} &= \arg\max_{\mathbf{s}} p(x;\mathbf{s},\mathcal{H}_1) \\ &= \arg\max_{\mathbf{s}} \frac{1}{(2\pi\sigma)^{N/2}} \exp\Big[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-s[n])^2\Big] \\ &= \mathbf{x}\end{aligned}$$

It follows:

$$p(x;\hat{\mathbf{s}},\mathcal{H}_1) = \frac{1}{(2\pi\sigma)^{N/2}} \tag{2.3.1.10}$$

$p(x;\mathcal{H}_0)$ is the same as in Eq. (2.3.1.7). The generalized likelihood ratio test follows:

$$L(x) = \exp\Big[\frac{1}{2\sigma^2}\sum_{n=0}^{N-1} x^2[n]\Big] > \gamma \tag{2.3.1.11}$$

Take the logarithm on both sides and simple steps yield the log likelihood ratio test:

$$\ln(L(x)) = \frac{1}{2\sigma^2}\sum_{n=0}^{N-1} x^2[n] > \ln(\gamma)$$

$$T(x) \overset{\text{def}}{=} \sum_{n=0}^{N-1} x^2[n] > 2\sigma^2 \ln\gamma \overset{\text{def}}{=} \gamma'$$

Or:

$$T(x) = \sum_{n=0}^{N-1} x^2[n] > \gamma' \tag{2.3.1.12}$$

The test statistic $T(x)$ in Eq. (2.3.1.12) is obtained from the energy of the observation $x[n]$, and is therefore called the *energy detector*.

2.3.2 Performance Metrics of Binary Detection

The detector of a binary hypothesis testing problem is also sometimes called a binary classifier or predictor. There are many different metrics that can be used to measure the performance of a binary classifier. Different metrics are used in different fields due to different goals. Sometimes, the same metrics are given different names in different applications. This section first gives a general overview of the fundamental metrics. After that, some of the metrics that are used throughout the thesis are highlighted.

In Eq. (2.3.1.1), the observation in which a signal is absent (e.g., null hypothesis $\mathcal{H}_0$ is true), is

often called a *negative*; while the observation in which a signal is present (e.g., alternative hypothesis $\mathcal{H}_1$ is true) is often called a *positive*. The detector classifies the observation as either from the null hypothesis $\mathcal{H}_0$ or the alternative hypothesis $\mathcal{H}_1$. To evaluate the detector, one compares the classification results to the ground truth and cross tabulates the data into a 2x2 contingency table or *confusion matrix* [54].

	$\mathcal{H}_0$ true	$\mathcal{H}_1$ true
predict $\mathcal{H}_1$	False Positive	True Positive
predict $\mathcal{H}_0$	True Negative	False Negative

Table 2.1: Confusion matrix of a binary classifier

One can then evaluate the detector by counting the following 4 numbers:

- **FP:** number of false positives
- **TP:** number of true positives
- **TN:** number of true negatives
- **FN:** number of false negatives

There are 8 possible ways to evaluate the detection performance by dividing each number by its row sum and column sum. However, only 4 of them are independent. The other 4 are just their ones' complements.

In the field of detection theory [28], following two statistics are often used:

- **Detection rate:** $\frac{TP}{TP+FN}$, the percentage of true positives that are labeled as positives.
- **False alarm rate:** $\frac{FP}{FP+TN}$, the percentage of true negatives that are labeled positives.

Ideally, the detection rate should be 1, while the false alarm rate is 0. For a practical detector, there is always a trade-off between the detection rate and the false alarm rate. To compare two different detectors, normally people draw the *Receiver Operating Characteristic (ROC)* [28]. The ROC plot illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting detection rate against the false-alarm rate at various threshold settings.

In the field of pattern recognition or machine learning, two statistics are mostly often used:

- **Recall:** $\frac{TP}{TP+FN}$, the percentage of true positives that are labeled as positives.
- **Precision:** $\frac{TP}{TP+FP}$, the percentage of labeled positives that are true positives.

Recall is equivalent to detection rate in the detection theory. Ideally, both recall and precision are 1. In practice, there is always a trade-off between precision and recall: increasing recall normally decreases precision and vice versa. By choosing a good detector, one can achieve both high recall and precision. One can plot recall vs. precision, a plot similar to the ROC.

When precision and recall are used to quantify the performance of a classifier, it's hard to compare the performance of two different classifiers, since one classifier could have higher recall but lower precision. To compare the overall performance by incorporating both recall and precision, people commonly use the *F-score*. The traditional F-measure or balanced F-score (F_1 score) is the harmonic mean of the precision and the recall:

$$F_1 \stackrel{\text{def}}{=} 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2.3.2.1}$$

The general formula for positive real β is:

$$F_\beta \stackrel{\text{def}}{=} (1 + \beta^2) \cdot \frac{precision \cdot recall}{\beta^2 \cdot precision + recall} \tag{2.3.2.2}$$

By choosing different values of β, the F-score puts different weights on precision and recall: a larger β means more emphasis on recall while a smaller β means more emphasis on precision.

There are many other statistics defined for other applications. Usually they are either equivalent to each other, or you can find one from the other (e.g., Sum of the two is 1). For a complete list of all the statistics, please refer to [54].

2.4 Wavelet Transform

The detection of a transient signal with unknown arrival time, unknown duration, unknown shape is difficult to solve, and there is no universal optimal detectors. A lot of transient detection work explored different models, transformation, or representation of the signal in order to expose its innate, distinct structure. A proper representation of the signals that takes advantage of the prior knowledge about the structure of the signals normally yields better detection performance [20, 49, 18]. The use of wavelet transform as a multi-resolution decomposition technique in the field of transient detection has been proven successful [19, 21, 33]. The continuous wavelet transform is used by the proposed detection methodology in Chapter 3, and therefore reviewed here. In particular, the frequency properties of the wavelets are derived, because the choice of scales in the proposed detector depends on them, and will be discussed in details in Section 3.3.2 of Chapter 3.

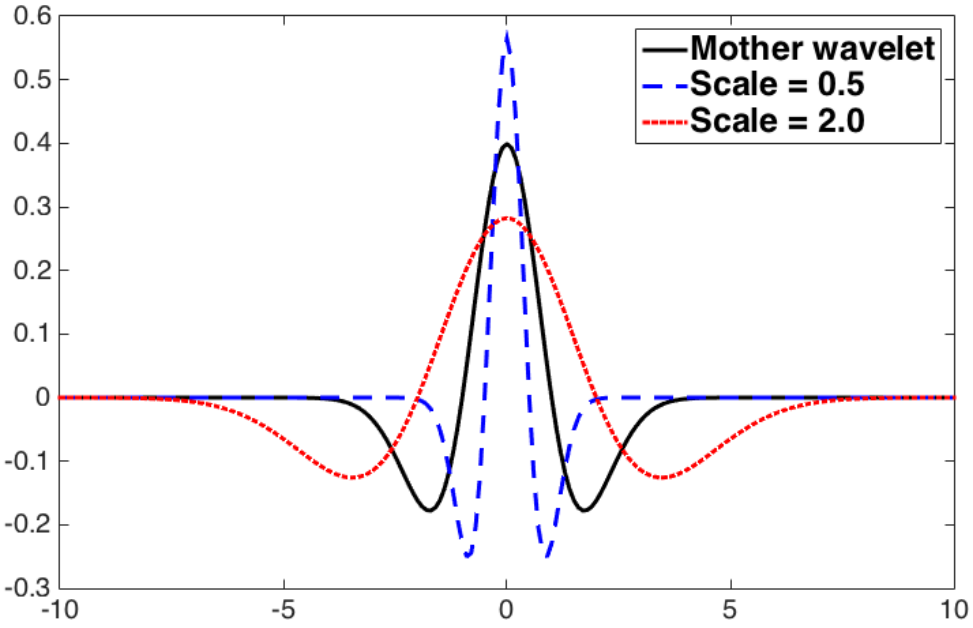


Figure 2.4-1: Mexican Hat Wavelets at different scales. Mexican Hat mother wavelet is defined by Eq. (3.3.1.1)

2.4.1 Mother Wavelet and Wavelets

A mother wavelet $\psi(t)$ has the following two properties:

$$\int_{-\infty}^{+\infty} \left|\psi(t)\right|^2 dt = 1 \tag{2.4.1.1a}$$

$$\int_{-\infty}^{+\infty} \psi(t)dt = 0 \tag{2.4.1.1b}$$

For a mother wavelet: $\psi(t)$, the wavelet with scale s and translation u is:

$$\psi_{s,u}(t) = \frac{1}{\sqrt{s}}\psi(\frac{t-u}{s}) \tag{2.4.1.2}$$

Eq. (2.4.1.2) indicates that the wavelet becomes wider (has larger support) as scale s increases, and narrower (smaller support) as scale s decreases (see Fig. 2.4-1).

You can easily prove that all wavelets satisfy the properties of mother wavelet as in Eq. (2.4.1.1):

$$\int_{-\infty}^{+\infty} \left|\psi_{s,u}(t)\right|^2 dt = 1 \tag{2.4.1.3a}$$

$$\int_{-\infty}^{+\infty} \psi_{s,u}(t)dt = 0 \tag{2.4.1.3b}$$

2.4.2 Continuous Wavelet Transform

The continuous wavelet transform (CWT) of a function $x(t)$ is defined as:

$$X(s,u) \stackrel{\text{def}}{=} \int_{-\infty}^{+\infty} x(t)\bar{\psi}_{s,u}(t)dt \tag{2.4.2.1}$$

where $\bar{\psi}_{s,u}(t)$ is the complex conjugate of the wavelet $\psi_{s,u}(t)$.

From the definition, the wavelet transform gives the inner product between the function $x(t)$ and wavelet $\psi_{s,u}(t)$. Since $\psi_{s,u}(t)$ is only non-vanishing in the neighborhood of u, $X(s,u)$ gives the local information of $x(t)$ around u. Furthermore, $X(s,u)$ measures the resemblance between function $x(t)$ around u and $\psi_s(t)$, the mother wavelet scaled by s,

$$\psi_s(t) \stackrel{\text{def}}{=} \psi_{s,0}(t) = \frac{1}{\sqrt{s}}\psi(\frac{t}{s}) \tag{2.4.2.2}$$

From Eq. (2.4.1.2) and Eq. (2.4.2.2), the follow equation can be derived:

$$\psi_{s,u}(t) = \psi_s(t-u)$$

Using $\psi_s(t)$ instead of $\psi_{s,u}(t)$, Eq. (2.4.2.1) can be rewritten as (by substituting t with $t+u$):

$$X(s,u) = \int_{-\infty}^{+\infty} x(t+u)\bar{\psi}_s(t)dt \tag{2.4.2.3}$$

Here is another way to view the CWT when the mother wavelet is a *symmetric*, *real-valued* function.

$$\begin{aligned} X(s,u) &= \int_{-\infty}^{+\infty} x(t)\psi_{s,u}(t)dt \\ &= \int_{-\infty}^{+\infty} x(t)\psi_s(t-u)dt \\ &= \int_{-\infty}^{+\infty} x(t)\psi_s(u-t)dt \\ &= (x * \psi_s)(u) \end{aligned} \tag{2.4.2.4}$$

From Eq. (2.4.2.4), CWT can also be viewed as a convolution between the function $x(t)$ and scaled mother wavelet $\psi_s(t)$, when the mother wavelet is a *symmetric*, *real-valued* function.

2.4.3 CWT on Discrete-time Signals

When implementing CWT on a computer, one need to adapt above equations to their discrete versions. When performing the continuous wavelet transform to discrete-time signals, both signal $x(t)$ and wavelets $\psi_{s,u}(t)$ become their sampled versions $x[n]$ and $\psi_{s,k}[n]$, respectively, for $n \in (\mathbb{Z})$,

$k \in (\mathbb{Z})$.

Define:

$$\psi_s[n] \stackrel{\text{def}}{=} \psi_{s,0}[n] \tag{2.4.3.1}$$

which is the scaled version of the discrete-time mother wavelet. Eq. (2.4.2.1) and Eq. (2.4.2.3) become:

$$X_s[k] = \sum_{n=-\infty}^{+\infty} x[n]\bar{\psi}_{s,k}[n] \tag{2.4.3.2}$$

$$= \sum_{n=-\infty}^{+\infty} x[n+k]\bar{\psi}_s[n] \tag{2.4.3.3}$$

For real-valued, symmetric wavelets, Eq. (2.4.2.4) becomes:

$$X_s[k] = \sum_{n=-\infty}^{+\infty} x[n]\psi_s[k-n] = (x * \psi_s)[k] \tag{2.4.3.4}$$

In all equations above, scale s can still be an arbitrary, positive real number.

2.4.4 Frequency-domain Properties of Wavelets

Suppose the Fourier transform of $\psi(t)$ is $\hat{\psi}(\omega)$, and the Fourier Transform of $\psi_{s,u}(t)$ is $\hat{\psi}_{s,u}(\omega)$:

$$\psi(t) \xrightarrow{\mathscr{F}} \hat{\psi}(\omega)$$

$$\psi_{s,u}(t) \xrightarrow{\mathscr{F}} \hat{\psi}_{s,u}(\omega)$$

where $\mathscr{F}$ is the Fourier transform.

From Eq. (2.4.1.2):

$$\hat{\psi}_{s,u}(\omega) = \sqrt{s}\hat{\psi}(s\omega)e^{-i2\pi u\omega} \tag{2.4.4.1}$$

The mother wavelet $\psi(t)$ is essentially a band-pass filter that is centered at C_ψ, and has the bandwidth of BW_ψ

From Eq. (2.4.4.1):

$$C_\psi(s) \stackrel{\text{def}}{=} \text{center frequency of wavelet } \psi_{s,u}(t)$$

$$= \frac{1}{s}C_\psi \tag{2.4.4.2a}$$

$$BW_\psi(s) \stackrel{\text{def}}{=} \text{bandwidth of the wavelet } \psi_{s,u}(t)$$

$$= \frac{1}{s}BW_\psi \tag{2.4.4.2b}$$

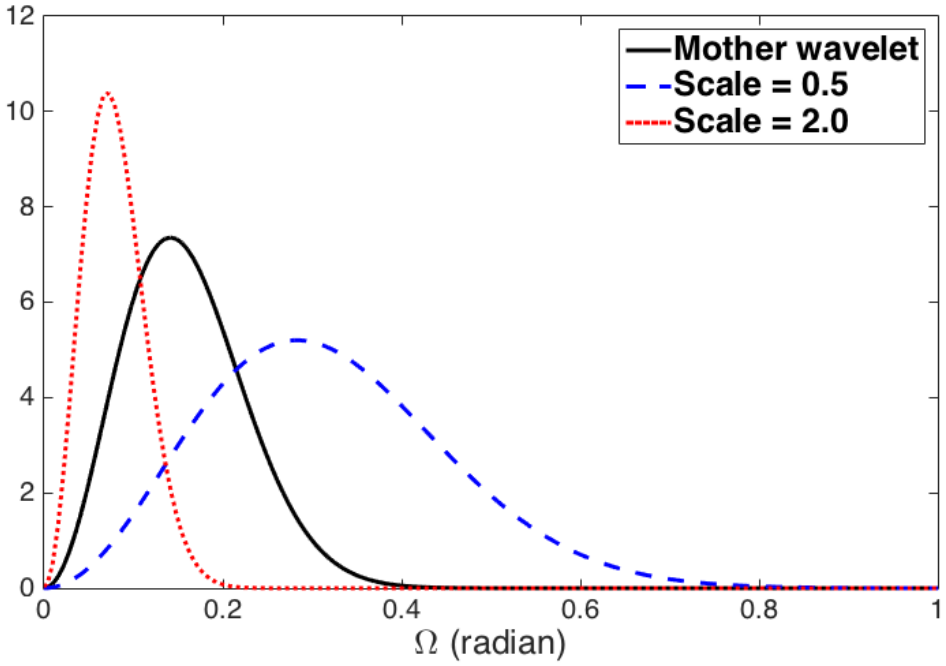


Figure 2.4-2: Discrete-time Fourier Transform of Mexican Hat Wavelets at different scales

Therefore, every wavelet at scale s is again a band-pass filter, although with a different center frequency and bandwidth. $\hat{\psi}_{s,u}(\omega)$ is centered at $C_\psi(s) = \frac{1}{s}C_\psi$, and has a bandwidth of $BW_\psi(s) = \frac{1}{s}BW_\psi$. Fig. 2.4-2 shows the magnitude of the Discrete-time Fourier Transform (DTFT) of the Mexican hat wavelets at different scales. The DTFT is periodic with period of 2π. For real-valued signals, the magnitude of the DTFT is also symmetric. Hence, in Fig. 2.4-2, only the positive frequency part of the DTFT is plotted with focus on the frequency from 0 to 1 for a better view, since all the DTFTs vanish quickly beyond 1. In addition, the frequency shown in the DTFT is in units of radians, rather than Hz. If one wants to interpret the DTFT in Hz, one needs to incorporate the sampling rate of the signal.