Silent Speech Recognition with Arabic and English Words for Vocally Disabled Persons

Sami Nassimi
Electrical and Electronic Engineering
University of Bahrain
Isa Town, Bahrain

Walaa AbuMoghli
Electrical and Electronic Engineering
University of Bahrain
Isa Town, Bahrain

Noora Mohamed
Electrical and Electronic Engineering
University of Bahrain
Isa Town, Bahrain

Mohamed WaleedFakhr
Electrical and Electronic Engineering
University of Bahrain
Isa Town, Bahrain

***Abstract*—This paper presents the results of our research in silent speech recognition (SSR) using Surface Electromyography (sEMG); which is the technology of recording the electric activation potentials of the human articulatory muscles by surface electrodes in order to recognize speech. Though SSR is still in the experimental stage, a number of potential applications seem evident. Persons who have undergone a laryngectomy, or older people for whom speaking requires a substantial effort, would be able to mouth (vocalize) words rather than actually pronouncing them. Our system has been trained with 30 utterances from each of the three subjects we had on a testing vocabulary of 4 phrases, and then tested for 15 new utterances that were not part of the training list. The system achieved an average of 91.11% word accuracy when using Support Vector Machine (SVM) classifier while the base language is English, and an average of 89.44% word accuracy using the Standard Arabic language.**

***Keywords*—Surface Electromyography; Support Vector Machine; Hidden Markov Models; Silent Speech Recognition**

I. INTRODUCTION

Automatic speech recognition (ASR) is a computer-based speech-to-text process, in which speech is recorded with acoustical microphones by capturing air pressure changes. ASR has now matured to a point where it is successfully deployed in a wide variety of every-day life applications, including telephone based services and speech-driven applications on all sorts of mobile personal digital devices [1]-[2].

Despite this success, speech-driven technologies still face two major challenges: first, recognition performance degrades significantly in the presence of noise. Second, confidential and private communication in public places is difficult due to the clearly audible speech. But most importantly, the performance is poor if the system having any form of speech disabilities [1]-[2].

Coming from a relative experience, elder people suffer a lot while speaking, and talking becomes a very challenging task that they have to face on a daily basis. Also, people who have undergone s laryngectomy which is surgical removal of the larynx due to cancer suffer a lot to communicate with others. These facts have motivated us to investigate the possibility of developing a Silent-Speech Recognition system (SSR) which will be able to recognize phrases that describe the basic needs of a person especially if he's spending most of his time in a care/nursing home.

The proposed approach for our project is by using the surface ElectroMyoGraphy (EMG); which stands for the technique concerned with the recording and analysis of electric signals taken from articulatory muscles using surface electrodes [2-4]. In contrast to many other technologies, EMG is a low cost, non-invasive, and portable technology.

The remainder of this paper is organized as follows: In section 2, we give an overview of previous related works. Section 3 provides a brief introduction about our methodology (sEMG) and presents our data acquisition, and section 4 presents our experiments and results. In section 5, we conclude the paper and propose possible future work.

II. RELATED WORKS

Research in the area of sEMG-based speech recognition has only a short history. Jorgensen et al. [2] investigated the recognition of non audible speech. Their idea is to intercept nervous signal control signals sent to speech muscles using surface EMG electrodes placed on the larynx and sublingual areas below the jaw. Initially, they demonstrated the potential of non-audible speaker dependent isolated word recognition based on the MES with a Neural Network classifier. They reported recognition rates of 92% for six control words and of 73% on an extended vocabulary which additionally contains the ten English digits.

More recently, there have been some serious efforts to enhance EMG based speech recognition and to make it user-independent as well as open vocabulary [3-5].

III. SEMG BASICS

The abbreviation EMG stands for Electro (electric), Myo (muscle), Graphy (writing). ElectroMyoGraphy is a technology

that allows to measure and record the electrical activity of the muscles and the nerve cells that control them (motor neurons). The last transmit electrical signals at the movement of the muscle and an EMG translates it into graphs, or numerical data [6].

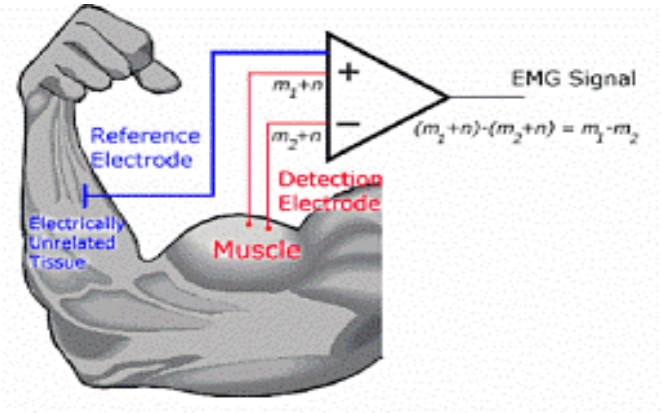

Fig. 1. Measuring an EMG signal

Surface EMG refers to the same process but using surface (non-implemented) electrodes. These electrodes work as a camera that transmits to us the electrical activity of the muscle. Since the amplitude of the signal can range from 0 to 1.5 mV (rms), an amplifier is needed. Amplified electrical signals are then fed into electronic devices for further processing [7]. However, because the EMG signal is based upon action potentials at the muscle membrane; a differential amplifier subtracts the signals from two detection sites and amplifies the difference voltage [8]. Consequently, any signal that originates far away from the detection site will appear as common signal - will have zero output- and thus removed, while the useful information carried out by the EMG signal will be different from both sites and thus amplified.

IV. SEMG SETUP & DATA CORPUS

sEMG Setup:

For sEMG recording, we used the MP system from BIOPAC Systems, Inc. the MP150 system serves as a data acquisition unit that converts analog signals (speech) into digital signals for further processing [9]. The Universal Interface Module UIM100C was used as the main interface between the MP150 and the external devices which for the purpose of our research has been the Electromyogram amplifier. At the early stage of our research, four EMG100C amplifiers have been used to amplify the electrical activity at four different detection sites. The EMG's have been connected to Ag-AgCl lead electrodes that can be directly attached to the skin of the user.

The electrodes positions have been adopted from (Maier-Hein et al., [4,5]). The channels captures signals from the levatorangulisoris (EMG2 & 3), the zygomaticus major (EMG2 & 3), the platysma (EMG4), and the orbicularis (EMG5). Later on, our research focused on investigating the performance of the system using only one channel which has been positioned with classical bipolar configuration with a 2cm center-to-center electrode spacing was used as shown in EMG2. The common ground reference has been connected to the wrist.

For the purpose of impedance reduction at the electrode-skin junction a small amount of electrode gel was applied to each electrode.

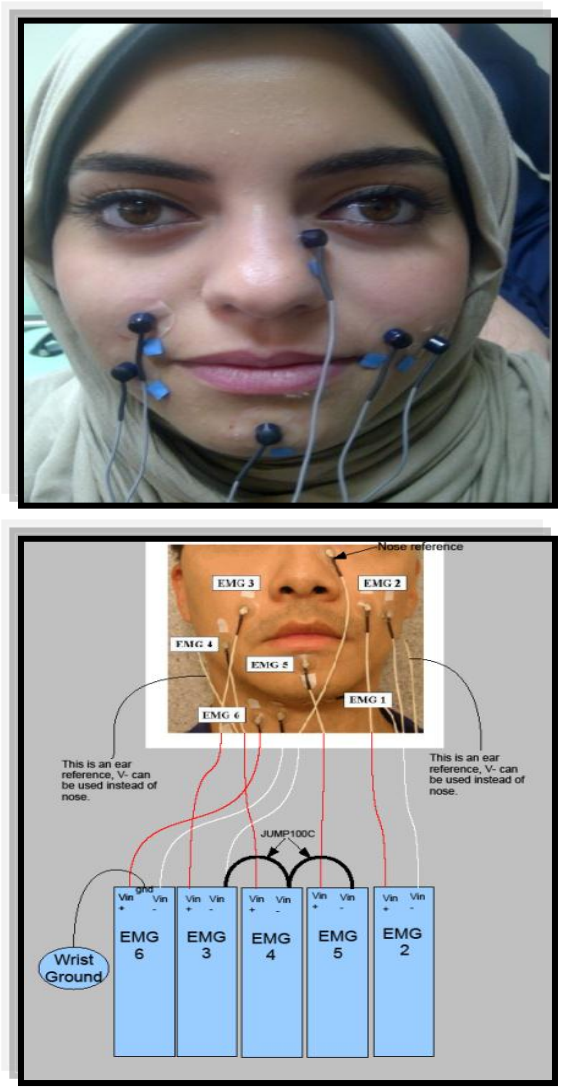

Fig. 2. Electrodes Positions (top: one of our volunteers, below: original connection adopted from (Maier-Hein et al., 2005 [5]).

The gain of the amplifiers has been set to 2000. The usable energy of the signal is limited to 0 to 500 Hz frequency range, thus, 1000Hz has been chosen as our sampling frequency, and a low pass filter with a cut-off frequency at 500Hz was used. To remove motion artifacts, a high-pass filter with cut-off frequency of 10Hz was used. Finally, all signals have been filtered with a notch filter at line frequency of 50 Hz because it is to be considered a dominant source of electrical noise.

A. Data Corpus:

All signal data used for our experiments was collected in so-called *recording sessions*. A recording session is defined as a set of utterances collected in series by one particular speaker. All settings (no. of channels, sampling rate, speech mode) remain constant during all sessions. For our research, three subjects that varied in age, nationality, and thus mother-tongue

with no known speech disorders participated to construct our database. In all sessions, the subject has been asked to pronounce the phrases non-audibly, i.e. without producing any sound. In this research, isolated-word recognition was performed. Thus, a word list was selected containing all phrases a speaker need to record during each session. The list can optionally be randomized. The phrases in this list have been chosen carefully to serve the focus of our project which has been elderly people who spend most of their time in a nursing home. The list consisted of the following four phrases:

TABLE I. THE FOUR PHRASES IN ENGLISH AND ARABIC LANGUAGES

The four phrases in English and Arabic Languages	
English language	**Arabic language**
I feel dizzy	أشعرُ بالدّوار
Take me outside	خُذْني إلى الخارج
I want to go to the toilet	أُريد الذهاب إلى الحمام
I need water	أُريدُ ماءً

From each subject, a total of forty five utterances have been collected making sure that each is not bounded by any silence at the beginning or end of it. The subject would start recording each phrase at a sign from our team, and he has been asked to repeat the phrase at each repetition of the sign.

Since any slight changes in electrodes position, temperature or tissue properties may alter the signal significantly. So, in order to make comparisons of amplitudes possible, we needed to apply a normalization procedure at each recording that compensates for these changes. A simple approach that we followed was to find the maximum of the absolute value of each utterance and divide the whole utterance by it.

V. DATA TRAINING

To ensure comparability of results from different experiments the same number of samples was used for each classifier for training stage, namely thirty exemplars of each phrase. Throughout our research, we used two classifiers:

A. Hidden Markov Model (HMM) Modeling technique

First order HMMs with Gaussian mixture models are used in most conventional speech recognition systems as classifiers because they are able to cope with both, variance in the time-scale and variance in the shape of the observed data. Each phrase in the list has been trained using a seven state left-to-right Hidden Markov Model with 3 Gaussians per state using the Expectation Maximization (EM) algorithm [10]. The number of iterations was chosen to be N=20.

To recognize an unknown signal the corresponding sequence of feature vectors was computed. Next, the Viterbi alignment for each vocabulary word was determined and the word corresponding to the best Viterbi score was output as the hypothesis. Feature extraction, HMM training, and signal recognition were performed using the Hidden Markov Model Toolkit (HTK) [11]. For this reason, a conversion of the cropped, cleaned, and normalized utterances to wave files was needed.

B. Support Vector Machine Classifier

SVMs are widely used as soft margin classifiers that find separating hyperplanes with maximal margins between classes in high dimensional space [12]. A kernel function is used to describe the distance between two data points. In order to find some statistically relevant information from incoming data, it is important to have mechanisms for reducing the information of each audio signal into a relatively small number of parameters, or segments. The classifier has been trained and tested with pre-segmented data with number of segments chosen to be N=20. Since each input signal was of different length, the segmentation procedure was as follows; first finding the length of each input signal, divide it by 20 to know the exact number that will produce 20 segments out of each input signal, and lastly dividing the signal by that number. After segmenting the utterances, we transformed each segment into frequency spectrum with FFT to extract some features. Feature extraction is the process of transforming raw signals into more informative signatures or fingerprints. We extracted the mean and the variance out of each segment, and then concatenated all features as one vector of attributes. All this segmentation and feature extraction part has been done using MATLAB [13-14]

The following block diagram represents the whole process of our work flow with its main steps.

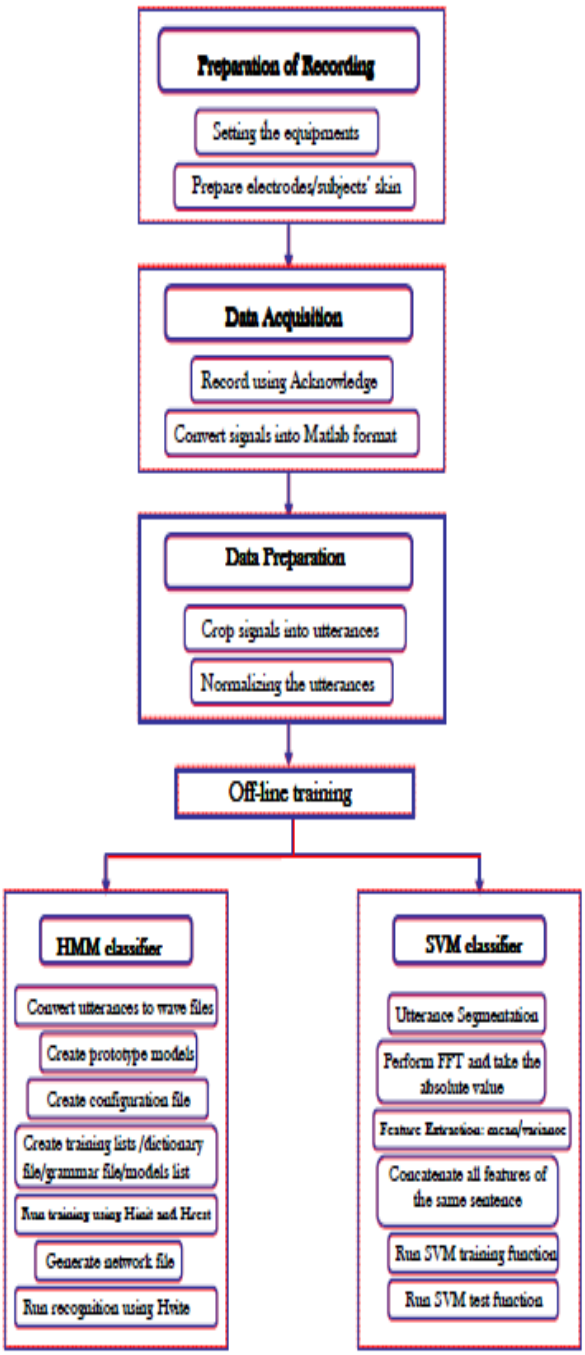

Fig. 3. The overall Block Diagram

VI. Experimental Results

Since we used thirty utterances from each subject for the training, we have been left with fifteen utterances for the test. The results varied between each phrase and among speakers. The results have been as follows:

A. Base Language: English

Our initial experiments have been conducted using HMM, and the results for each subject and each phrase were as shown in Figure 4 below. Although for few phrases the recognition was poor, the system achieved an average performance for all three subjects of 76.663%.

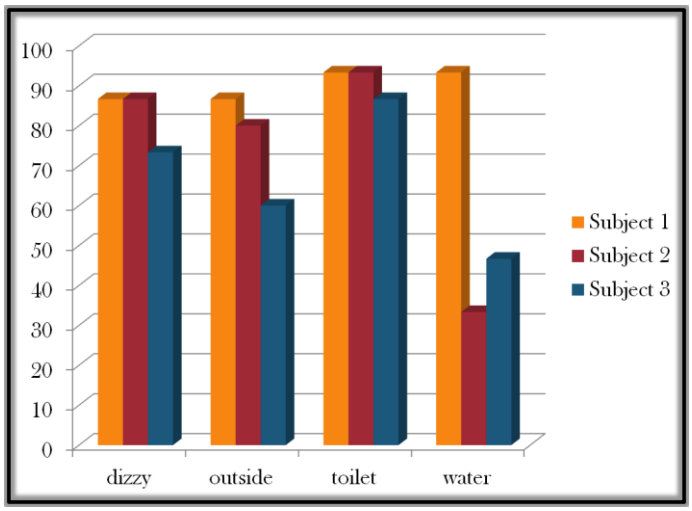

Fig. 4. Results using HMM for English system

Secondly, we investigated the performance of the system using SVM, and a clear improvement has been seen for all subjects. The system achieved an average of 91.11% word accuracy. The results in detail are shown in Figure 5.

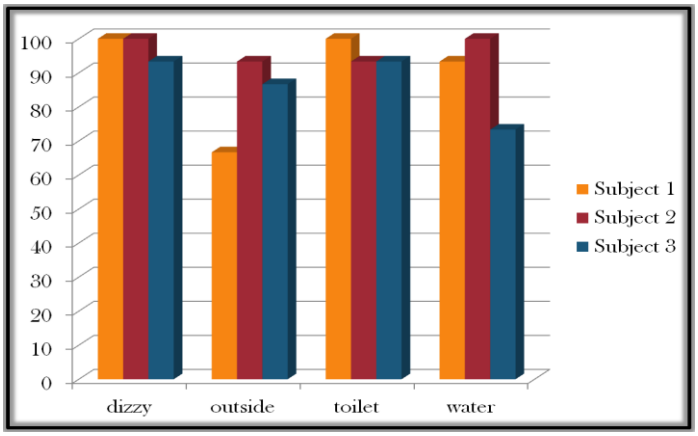

Fig. 5. Results using SVM for English system

B. Base Language: Arabic

Since studies and researches of Arabic-based SSR systems are poor compared to other similar languages, we have been motivated and curious to investigate and develop an Arabic-based system.

For the same number of test utterances used to examine the English system, a similar one was used for the Arabic. We have also checked the performance of the system using the same two classifiers and the results using HMM were as shown in Figure 6.

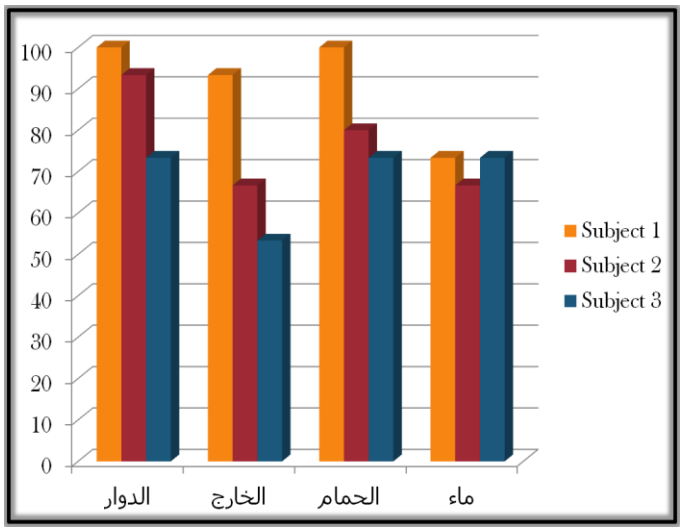

Fig. 6. Results using HMM for Arabic system

Averaging the results of all three subjects, the system achieved an average of 78.89% word accuracy.

Similarly, we experienced an improvement in the results when we trained and tested the system using SVM classifier, where the system has achieved 89.44% word accuracy. The results of each subject are shown in Figure 7.

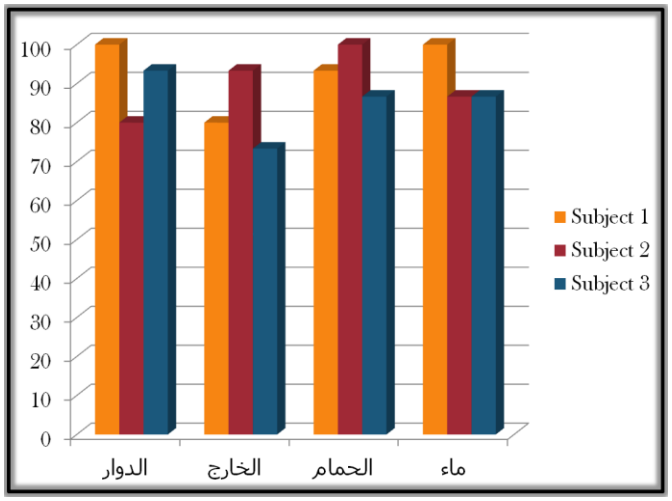

Fig. 7. Results using SVM for Arabic system

VII. Conclusion and Future Work

We have presented in this paper the results of our work for developing an isolated word Silent Speech Recognition System for both Arabic and English words. The technology is based on Surface Electromyography; which is capturing and recording of electrical potentials that arise from the muscle activity using surface electrodes attached to the skin. The concept of our work is still in the research area, so this work can be seen as a feasibility study. Moreover, we have investigated several state-of-the-art tools to check the performance of our system; such as: Hidden Markov Model and Support Vector Machine classifiers. Our experimental results indicate the effectiveness and efficiency of our proposed whole-sentence recognition system mainly using SVM algorithm in contrast to HMM classifier.

For the English system, an average of 91.11% word accuracy has been obtained when using SVM compared to the 76.663% obtained using HMM. Likewise, the Arabic system has achieved an average performance of 89.44% when using SVM compared to the 78.89% obtained while using HMM.

Though the obtained results are encouraging, this research does not claim completeness and it has lots of room for improvements. For example, Comparative experiments indicate that applying more than one electrode is crucial in order to construct a more robust system. Also, to demonstrate the potentials of this technology, EMG based speech recognition should move beyond isolated-word speech recognition and approach continuously spoken large vocabulary tasks.

Acknowlegement

The authors would like to acknowledge the research grant (2012/07)by the University of Bahrain Research Deanship which was used to purchase all the equipment.

References

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, and J. Gilbert, *Silent Speech Interfaces*, Speech Communication, vol. 52, 2010.

[2] C. Jorgensen, D. Lee, and S. Agabon, *Sub Auditory Speech Recognition Based on EMG/EPG Signals*, in Proc. of the International Joint Conference on Neural Networks, 2003.

[3] Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J.M., Brumberg, J.S., *Silent Speech Interfaces*, Speech Communication (2009), doi: 10.1016/j.specom.2009.08.002.

[4] Maier-Hein, L. (July, 2005). Diplomarbeit (Thesis). *Speech Recognition Using Surface Electromyography*. Retrieved October 5th, 2013 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.63.2803&rep=rep1&type=pdf

[5] Maier-Hein, L., Metze, F., Schultz, T., and Waibel, A. (November 2005). *Session Independent Non-audible Speech Recognition Using Surface Electromyography*. In Proc. ASRU, San Juan, Puerto Rico.

[6] Al-Mulla, M. R., Sepulveda, F. and Colley, M. (2012b). *sEMG Techniques to Detect and Predict Localised Muscle Fatigue*, In: EMG Methods for Evaluating Muscle and NerveFunction, Mark Schwartz, pp. 157-186, InTech, ISBN 978-953 307-793-2, Rijeka, Croatia.

[7] Day, S. (2010). Important *Factors in Surface EMG Measurement*. Retrieved October, 2nd, 2013 from http://www.andrewsterian.com/courses/214/EMG_measurement_and_recording.pdf

[8] De Luca, C. (2002). *Surface Electromyography: Detection and Recording*. DelSys Incorporated. Retrieved February 28th, 2013 from http://www.delsys.com/Attachments_pdf/WP_SEMGintro.pdf

[9] http://www.biopac.com/emg-electromyography.

[10] Gales, M., and Young, S. (2008). *The Application of Hidden Markov Models in Speech Recognition.*Foundations and Trends® in Signal Processing, Vol. 1, No. 3 (2007) 195 – 304.

[11] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland. P. (April 2005). *The HTK Book.*Cambridge University Engineering Department.

[12] Scholkopf, B., and Smola, A.J., *Learning with Kernels*, MIT Press, Cambridge, MA. 2002.

[13] Rakotomamonjy, A. (2006).*SVM and Kernel Methods Matlab Toolbox.* http://asi.insa-rouen.fr/enseignants/~arakoto/toolbox/index.html.

[14] Mathworks "Train support vector machine classifier". http://www.mathworks.com/help/toolbox/bioinfo/ref/svmtrain.html (4/6/2013).