Small Vocabulary Recognition Using Surface Electromyography in an Acoustically Harsh Environment

Bradley J. Betts[1] and Charles Jorgensen[2]

Neuro-Engineering Laboratory
NASA Ames Research Center, M/S 269-1
Moffett Field, CA 94035-1000

[1]QSS Group, Inc., bbetts@email.arc.nasa.gov
[2]NASA, cjorgensen@mail.arc.nasa.gov

Abstract

This paper presents results of electromyographic-based (EMG-based) speech recognition on a small vocabulary of 15 English words. The work was motivated in part by a desire to mitigate the effects of high acoustic noise on speech intelligibility in communication systems used by first responders. Both an off-line and a real-time system were constructed. Data were collected from a single male subject wearing a firefighter's self-contained breathing apparatus. A single channel of EMG data was used, collected via surface sensors at a rate of 10^4 samples/s. The signal processing core consisted of an activity detector, a feature extractor, and a neural network classifier. In the off-line phase, 150 examples of each word were collected from the subject. Generalization testing, conducted using bootstrapping, produced an overall average correct classification rate on the 15 words of 74%, with a 95% confidence interval of [71%, 77%]. Once the classifier was trained, the subject used the real-time system to communicate to a cellular phone and to control a robotic device. The real-time system was tested with the subject exposed to an ambient noise level of approximately 95 decibels.

Keywords: electromyography, EMG, bioelectric, subvocal speech recognition, first responder, pattern recognition, SCBA

1. Introduction

Speech intelligibility can be severely degraded by high levels of acoustic noise. Researchers have developed a variety of techniques to minimize the impact of noise, ranging from adaptive noise cancellation to throat microphones. Increasingly, researchers are experimenting with the measurement and analysis of bioelectric signals associated with speech in an effort to further minimize—or even completely eliminate—the degrading effects of acoustic noise. Such techniques, either on their own or fused with other modalities, hold promise for improving human communication and human-computer interaction.

The bioelectric technique used in the research reported here is electromyography, the study of muscle function through its electrical properties. Electrical activity emanating from muscles associated with speech can be detected by non-invasive surface sensors mounted in the region of the face and neck. Sensing of this type is not directly interfered with by acoustic noise (although indirect effects, such as the propensity of speakers to modify their vocal effort in the presence of noise [1, 2], require further study).

First responders are an example of a class of users that stands to benefit from reliable communication in acoustically harsh environments. For example, sirens, engines, and saws all add noise to a typical firefighting scene, as does the breathing apparatus a firefighter wears. This work was motivated in part by a desire to see whether electromyographic-based (EMG-based) speech recognition could alleviate these effects. It is part of a broader effort at NASA investigating the physiology, signal processing, and applications of EMG-based speech recognition.

Besides reducing the impact of environmental noise, EMG-based speech has the interesting property that it can be detected even when a subject emits little or no acoustical energy during speech, a fact first noted by the

Danish researcher Faaborg-Andersen in 1957 [3]. That is, EMG activity is detectable when a subject speaks normally, whispers, moves the mouth without emitting sound, and even when making virtually no facial movement at all (but consciously activating speech muscles, akin to saying a phrase silently to oneself). While the EMG signal characteristics most definitely change during these different types of activity, the signal is detectable. Because of this potential for silent communication, we sometimes refer to EMG-based speech recognition as *subvocal speech;* the two terms are used interchangeably in this paper.

In the study reported here, EMG data were collected from a single male subject wearing a self-contained breathing apparatus (SCBA) under laboratory conditions. Data samples consisting of isolated words chosen from a small English vocabulary were used to train a neural network classifier and to test the generalizability of the network. The trained network was then inserted into a real-time communication and control system while the subject was exposed to approximately 95 decibels of acoustic noise. Isolated phrases recognized from the EMG signal in real time were both communicated to a cellular phone and used to control a robotic platform (see Figure 1).

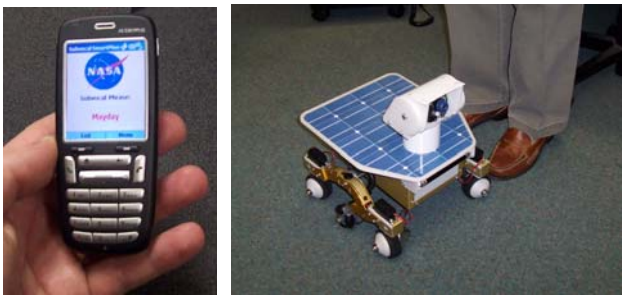

Figure 1. Communication and control output modalities: cellular phone and robotic platform. The robotic platform is a Carnegie Mellon University Personal Exploration Rover (PER) [4].

The remainder of this paper is organized as follows: A brief background on the history and physiology of EMG is given. Other research efforts that have examined EMG-based speech recognition are surveyed. The methods and results of this research are then presented, followed by conclusions and avenues for future work.

2. Background and Related Research

2.1. Electromyography

As has already been stated, electromyography is the study of muscle function via its electrical properties (i.e., the electrical signal emanating from the muscle during muscle activation). For those unfamiliar with the field, the book by Basmajian and De Luca is highly recommended [5]. Other good sources include the paper by De Luca [6], the book chapter by Gerdle et al. [7], and the book edited by Bronzino [8].

As detailed by Basmajian and De Luca [5], electromyography has a long and interesting history. In 1848, the Frenchman DuBois-Reymond was the first to report the detection of electrical signals voluntarily elicited from human muscles. By placing his fingers in a saline solution and contracting his hand and forearm, he produced a measurable deflection in a galvanometer. His dedication to his work is beyond question—correctly surmising that the skin presented a high impedance to the flow of current, on at least two separate occasions he deliberately blistered his forearm, removed the skin, and exposed the open wound to the saline, thereby producing a substantially greater deflection in the galvanometer during muscle contraction.

Electromyography has continued to develop since the time of DuBois-Reymond. Substantial research interest was generated during the 1960s in the use of electromyography as a mechanism for the control of prostheses [9-11]. The arrival of inexpensive digital computing in the 1980s furthered development, with many research groups investigating digital techniques for control and communication, including groups focused on EMG-based speech recognition. These speech efforts are surveyed in the next section of this paper.

The electrophysiology of muscles is complex, and only the briefest overview will be given here; see Gerdle et al. (and the references therein) for a more detailed description [7]. Muscle action originates in the central and peripheral nervous systems. Nerve impulses are carried from anterior horn cells of the spinal column to the end of the nerve via motor neurons. As the axon of the motor neuron approaches individual muscle fibers, it branches, meaning a single motor neuron innervates several muscle fibers, terminating at a neuro-muscular junction (also known as an endplate); see Figure 2. When a nerve impulse reaches the endplate, the neurotransmitter acetylcholine is released. This in turn causes sodium and potassium cation channels to open in the muscle fiber. Once an excitation threshold is achieved, an action potential propagates in both directions from the endplate to the muscle-tendon junction. This movement of cations establishes an electromagnetic field in the vicinity of the muscle fiber. The time-varying potential recorded by an electrode placed in the field is known as an electromyogram. Of course, such an electrode measures the superposition of several such fields arising from separate motor units. That, coupled with spatially varying tissue filtering effects, makes the EMG signal highly complex.

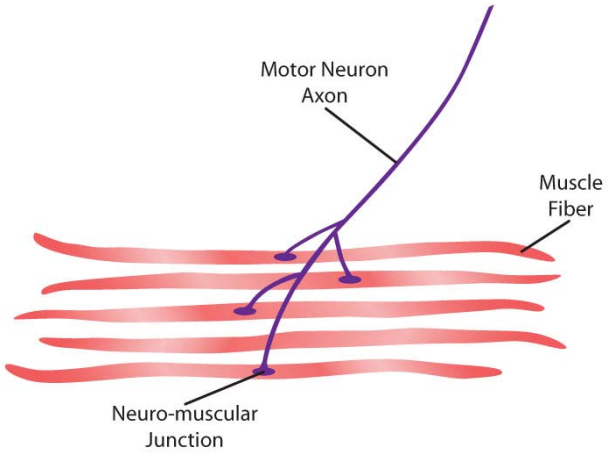

Figure 2. Motor neuron innervation of muscle fibers. Note that adjacent fibers are not necessarily innervated by the same motor neuron.

There are two principal sensing techniques used in electromyography: invasive indwelling sensing and non-invasive surface sensing. This paper, and essentially all of the research we are aware of related to subvocal speech recognition, focuses on the use of surface sensors. Good references on the sensors used and issues arising with surface electromyography include [7] and [12].

2.2. EMG-based Speech Recognition

In this section of the paper, we survey results on EMG-based speech recognition. There is also a rich body of literature on the use of EMG for control of prostheses and for gesture recognition that is not surveyed here ([13-16] are but a few of the many examples).

Chapters 19 and 20 of Basmajian and De Luca describe electromyography research done before 1985 related to the muscles of the mouth, pharynx, larynx, face, and neck [5]. As it pertains to speech, the goal of research during that period seems to have been understanding muscle processes associated with phonation in normal subjects and subjects with disability. Investigations were carried out predominantly through fine-needle indwelling electrodes on animals and humans. Although no explicit references have been found prior to 1985 to attempts at EMG-based speech recognition, the concept almost surely occurred to researchers of the time—the state of digital computing (and non-invasive) sensing may have been the limiting factors.

The first efforts at performing EMG-based speech recognition seem to have occurred independently and in parallel in Japan and the United States around 1985–86. In Japan, Sugie et al. used three channels of silver silver-chloride (Ag-AgCl) surface sensors with a sampling rate of 1250 samples/channel/s [17]. A threshold-and-counting scheme was used to produce a three-bit number every 10 ms. These numbers were then fed into a finite automaton for vowel discrimination. Three subjects were asked to repeat 50 Japanese monosyllables. The overall correct classification rate was reported as 64%. It is interesting to note that the researchers developed a pilot real-time system as part of this effort.

Simultaneously in the United States, Morse [18] and Morse and O'Brien [19] used four channels of stainless steel surface electrodes (with a light coating of electrode gel) and a sampling rate of 5120 samples/channel/s. Analog filtering was used to restrict the bandwidth of the EMG signal to the 100–1000 Hz range. An average magnitude technique was used to reduce the signal dimensionality to 20 points/channel/s. Two subjects were studied with several different word sets, one of which was the English words "zero" to "nine." Subjects were asked to repeat each word twenty times. A maximum likelihood technique was used for classification. For the ten-digit word set, a correct classification rate exceeding 60% was observed. In later work in 1991, Morse et al. applied a neural network to a similar data set and achieved roughly the same correct classification rate of 60% [20]. Other papers from this group include [21] and [22].

In 2001, the Canadian researchers Chan et al. reported EMG-based speech recognition results that were motivated by the need to communicate in acoustically harsh environments (in this case the cockpit of a fighter aircraft) [23]. Five channels of surface Ag-AgCl sensors were used with each channel bandlimited to 100–500 Hz and sampled at a rate of 1000 samples/channel/s. A variety of transforms (including a wavelet transform) and principle component analysis (PCA) were used to reduce the data to thirty features per word on a ten-word vocabulary (the ten English digits). Classification was performed using linear discriminant analysis (LDA). On an experiment in which words were randomly presented to two subjects, recognition rates as high as 93% were achieved. In later work, a hidden Markov model (HMM) was used as the classification engine and achieved results similar to the LDA technique [23]. In 2002, Chan et al. used evidence theory to combine results from a conventional automatic speech recognition system and an EMG-based one, dramatically maintaining a high overall correct classification rate in the presence of ambient acoustic noise [24].

At the NASA Ames Research Center, members of the Neuro-Engineering Laboratory have done work on subvocal speech recognition. In 2003, Jorgensen et al. collected six words from three subjects using surface Ag-AgCl sensors and a single EMG channel [25]. Data were collected at the rate of 2000 samples/channel/s. A variety of techniques were tested for feature extraction, including short-time Fourier transforms, linear predictive coding, and several different wavelet transforms. Classification was performed using a neural network and an average

correct recognition rate of 92% was achieved. In later work, Jorgensen and Binsted applied a similar signal processing architecture to seventeen vowel phonemes and twenty-three consonant phonemes collected from two subjects [26]. Average correct recognition exceeded 33% for the entire vocabulary (and exceeded 50% when certain alveolars were removed).

In 2003, NTT DoCoMo researchers Manabe et al. used a novel surface sensor mounting configuration for EMG-based speech recognition [27]. Three channels of sensors were mounted on the subject's hand, then the hand was held to the face during speech. Analog filtering restricted the EMG signal to the range 20–450 Hz with a sampling rate of 1000 samples/channel/s. Recognition was performed using a three-layer neural network, where the inputs to the network were the root-mean-squared (RMS) EMG values during pronunciation of a vowel. Over three subjects, each using a vocabulary of five Japanese vowels, the average correct classification rate exceeded 90%. In later work, Manabe and Zhang made use of HMMs to classify the ten Japanese digits collected from ten subjects; accuracies as high as 64% were achieved [28].

In 2004, Kumar et al. used three EMG channels for speech recognition [29]. Channels were sampled at 250 samples/channel/s, with RMS EMG values used as feature inputs to a neural network classifier. Using three subjects and five English vowels, an average recognition rate of up to 88% was achieved.

2.3. Communication in Acoustically Harsh Environments

People have long had an interest in communicating in acoustically noisy environments. Military needs have driven research and development in this area for many decades. Much research was done before and during the Second World War on techniques to allow pilot voice communication in airplanes, resulting in the development of devices such as throat microphones (e.g., the T-30 throat microphone, manufactured by Shure Inc. under a 1941 contract). Interest in these devices continues to this day, particularly when used as part of a multi-modality speech recognition system [30-32]. Military research in the area continues, with the United States Defense Advanced Research Projects Agency (DARPA) sponsoring research in sensors and techniques appropriate for communicating in noisy environments [33, 34].

Unfortunately, in many cases first responders have yet to benefit from these advanced techniques. For many fire departments, voice communication is still done by shouting through the mask of the SCBA into a shoulder-mounted or hand-carried radio. Some alternatives have been developed and targeted at first responders (e.g., bone conduction microphones [35] and in-mask boom microphones) but have yet to receive wide deployment. Our study suggests that bioelectric techniques also hold promise for this community, whether on their own or fused with other modalities. Minimizing the impact of acoustic noise and the potential for covert communication would make bioelectrics appeal to different segments of the community. It is important to remember, however, that the first responder community is one that values dependable and robust equipment and that is almost always forced to be extremely cost conscious.

3. Methods

In this section we detail the equipment and techniques used during data collection. The signal processing techniques are described followed by a brief overview of the hardware and software architecture.

3.1. Equipment and Data Collection

Training data were collected from a single 33-year-old male subject, qualified in the use of SCBA equipment. The subject was seated and remained stationary during data collection. The subject wore a standard-issue firefighting turnout jacket, a fire-retardant hood, and a SCBA unit (Survivair Panther with Twenty-Twenty Plus mask; Survivair; Santa Ana, CA) as shown in Figure 3. The SCBA was pressurized per normal SCBA usage. The subject was instructed to breathe normally during data collection sessions (i.e., as he would while wearing an SCBA). Collection sessions were paused as necessary to replace empty air tanks.

Figure 3. Photo showing data collection station and SCBA equipment.

One differentially amplified channel of EMG data was collected under quiet laboratory conditions. Surface Ag-AgCl sensors (Soft-E H69P; Kendall-LTP; Chicopee, MA) were positioned on the subject's neck as shown in

Figure 4. A third Ag-AgCl sensor, used as a ground, was attached either behind the subject's right ear or on the subject's wrist. The subject's skin was prepared by wiping it with an alcohol pad (70% isopropyl alcohol pad #818080; Abco, Inc.; Nashville, TN) in an effort to reduce skin impedance by removing surface oils and dead skin cells. Sensor leads were connected to a headbox which was in turn connected to a programmable amplifier (SynAmps Model 5083; Neuroscan; El Paso, TX). The amplification gain was set at 1000. The signal was bandlimited to the range 10–2000 Hz and sampled at 10^4 samples/s with 16-bit precision. A 60 Hz digital notch filter was used to reduce main power frequency interference.

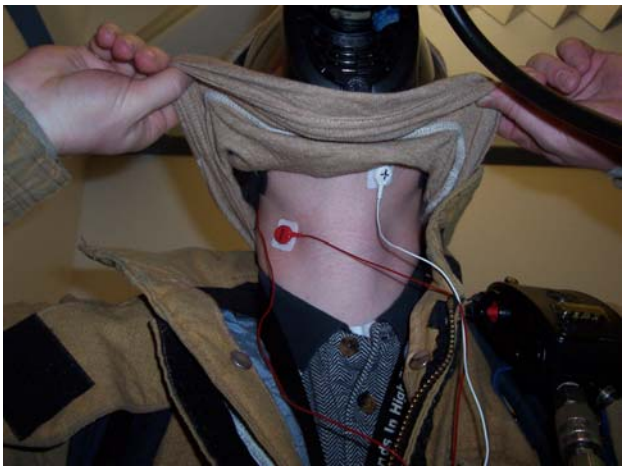

Figure 4. Photo showing EMG sensor placement. The subject has peeled his hood back to reveal the sensors. A third sensor, placed behind the subject's right ear (or alternatively placed on the wrist) was used as a ground.

The rationale for using such a high sampling rate for surface electromyography deserves mention. Given the logistical difficulty associated with collecting data from subjects, the ease with which data can be digitally downsampled after collection, and the desire to minimize aliasing effects, the decision was made to use a high sampling rate. The price of so doing was increased data storage and increased computational demands on the real-time implementation. In effect, storage and computation increases were traded in favor of greater flexibility in off-line analyses. In a recent article, Durkin and Callaghan examine sampling rate issues associated with surface electromyography [36].

Data were stored to disk using Acquire software (Acquire version 4.3; Neuroscan; El Paso, TX). Fifteen isolated words were collected 150 times each, for a total of 2250 word samples. The words, shown in Table 1 and assigned class codes, were chosen from a list compiled by firefighters at the Moffett Field Fire Department as representative of the tactical vocabulary they use while performing their work. As Table 1 shows, some of the vocabulary elements are in fact two-word phrases. We will nonetheless use the term "word" for consistency with other published accounts.

Table 1. Fifteen-word vocabulary. Class codes are used during the presentation of results.

Words			
Evacuate (C1)	Mayday (C2)	Man-trap (C3)	Fire Clear (C4)
Fire Safe (C5)	Room (C6)	Status (C7)	North (C8)
South (C9)	East (C10)	West (C11)	Zero (C12)
One (C13)	Two (C14)	Three (C15)	

The subject was prompted via software to say the vocabulary words in a fully randomized order (see Figure 5). Randomization was used to minimize learning and anticipatory effects. The subject had a visible amount of time in which to say a word, with a pause between words of 2.5 seconds. Since firefighters have no obvious use for covert communication, the subject was instructed to speak at a normal conversational level (as opposed to whispering or emitting no acoustical energy, a mode of operation that might be more suited to certain police and military units). Data were collected during four separate sessions over a three-week span. The subject was photographed during the first recording session to establish where the sensors were located on the neck. In subsequent sessions, an assistant placed the sensors in the same location with the aid of the photograph.

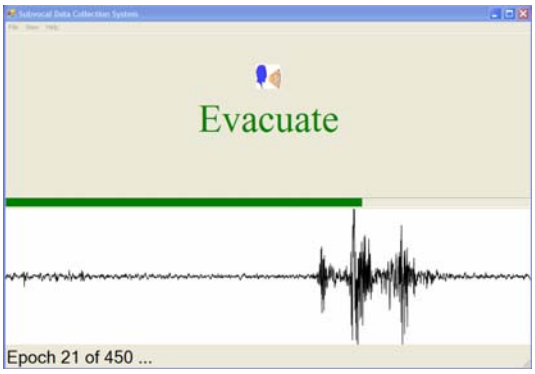

Figure 5. Screen shot of software used to prompt subjects during data collection. A slider indicates the time span available to the subject to say the displayed word. The EMG trace is visible to the subject during recording. The text at the bottom gives the subject status information (i.e., this subject is collecting word example 21 of 450 total for this particular session).

3.2. Signal Processing

Subvocal speech recognition is a type of pattern recognition, one in which a captured training set allows for the use of supervised learning techniques. That is, suppose we have obtained a training set $T = \{(x_1, y_1), \ldots, (x_K, y_K)\}$ consisting of K labeled samples, where the $x_i \in \mathbb{R}^D$ are the samples and the $y_i \in \mathbb{Z}_M$ are the corresponding class labels. For the work reported here, $K = 2250$, $M = 15$, and, as will be discussed, $D = 1.5 \times 10^4$. We seek the function f_T^* that maps samples to class labels and that maximizes the correct classification rate over the entire joint distribution of samples and class labels. Indeed, this is essentially the goal of all the various pattern recognition techniques—neural networks, hidden Markov models, support vector machines, or anything else.

The signal processing activity has two distinct phases. In the first, a training set is used to produce a classifier. In the second, the trained classifier is presented with previously unseen samples, either for the purpose of testing the classifier or for producing some end effect. The first three stages are common to both phases:

1. signal acquisition
2. activity detection
3. feature extraction

The signal acquisition process was described in the previous section of this paper. The other two common blocks—activity detection and feature extraction—are described next.

Activity detection refers to the process of segmenting an isolated word out of the continuous EMG stream (other names used in the literature include utterance detection and end-point detection). In this work, only a single EMG channel needed to be monitored for activity. The technique used was a simple one and involved partitioning the EMG data stream into 20 ms packets, then labeling each as either signal or noise. The signal-versus-noise determination was made by comparing the RMS value of the packet to a noise threshold dynamically set at the beginning of a recording session (by assuming the first 10 seconds of data were noise, then holding the threshold fixed for the remainder of the session). A second level of logic then examined the resulting bit sequence to make sure that spurious 0s (i.e., noise) were not inserted into contiguous activity blocks and that the blocks had a certain enforced minimum time separation. The final logic level ensured that an activity block was placed in the center of a 1.5 second window, buffered on either side as necessary by the surrounding EMG activity. At the set sampling rate of 10 kHz, this resulted in a fixed block of 1.5×10^4 samples being sent downstream for feature extraction. The fixed block size made feature extraction easier at the price of including some noise samples with the word. Figure 6 shows an example of the activity detector operating on an EMG signal. While substantially more sophisticated activity detection techniques can be found in the literature (e.g., [37-42]) and are candidates for inclusion in future work, the technique described proved sufficient for both the off-line and real-time systems.

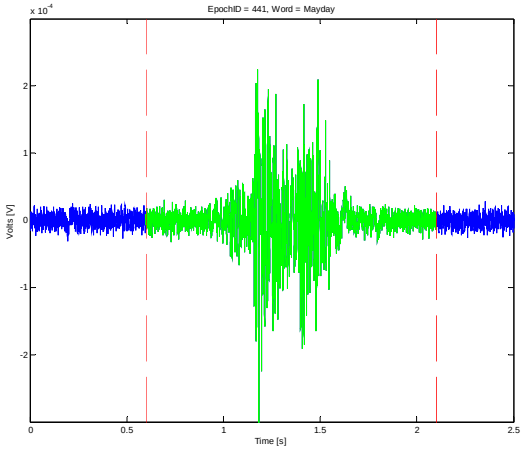

Figure 6. Activity detector operating on EMG data. The 1.5 second activity region is contained between the dashed lines. This particular region is the subject saying "Mayday."

Feature extraction is the process of reducing the dimensionality of the data to in some way facilitate subsequent classification. In this project, the 1.5×10^4 dimensional activity block was reduced to a feature vector of dimension 20 by a process of full-wave rectification, wavelet transformation, and low-pass filtering of the resulting level-1 approximation band. The particular wavelet transform chosen was Kingsbury's dual-tree complex wavelet transform, selected because of its shift-invariant properties [43]. Many wavelet transforms suffer from the property that minor shifts in the input signal can cause significant redistribution of energy in the various subbands. Kingsbury's transform alleviates this, thereby reducing sensitivity to the exact positioning of the signal within the activity window. We and others have also used HMMs in the past to ease temporal alignment issues [23, 25]. Figures 7 and 8 show the output of the feature extractor on some word samples. The left portion of each figure shows the EMG activity regions. The right portion plots the feature dimensions on the abscissa and their magnitudes on the ordinate.

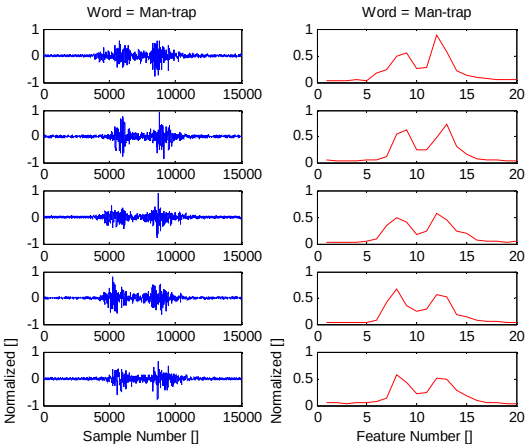

Figure 7. Feature examples for the word "Man-trap."

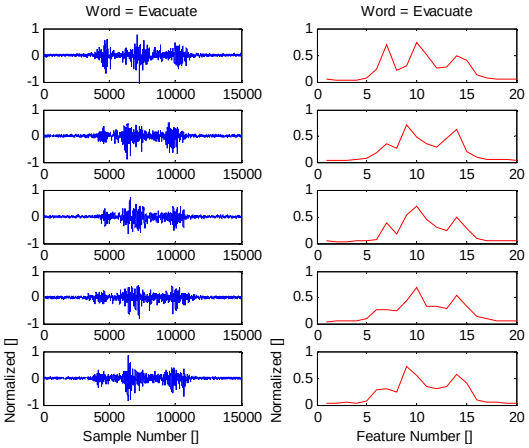

Figure 8. Feature examples for the word "Evacuate."

After activity detection and feature extraction, features from the training set are used to train a neural network classifier. The neural network chosen was a conjugate gradient network [44]. The network was configured with 20 input nodes, one hidden layer of 41 nodes, 15 output nodes, and was run with 400 training epochs (or until the performance goal was met, meaning training had converged). All of these values, including the number of dimensions in the feature vector, were arrived at by an ad hoc process of optimization. While we believe these values to produce a good overall classifier for this particular data set, future work will look to make the parameter tuning process more automatic. As will be discussed in more depth in the Results section, 70% of the collected samples were used for training. The remaining 30% were set aside for generalization testing.

3.3. Hardware and Software Architecture

A major goal of this preliminary study was to assess the feasibility of using EMG-based speech recognition for first responders. Focus was placed on recognition results and assessing the impact of SCBA equipment; no effort was made to miniaturize equipment. At the time of writing, the system is not portable and not hardened for field use.

As has already been mentioned, Neuroscan's Acquire software was used for collecting the training set data. The data were stored in binary format. Metadata associated with a particular session (e.g., session date, subject name, word ordering, etc.) were stored in an XML file.

Off-line training and analysis was performed using Matlab (Matlab version 7.1.0.246; The Mathworks, Inc.; Natick, MA). The training and analysis code makes extensive use of Matlab's object-oriented capabilities, particularly in organizing the results of a recording session.

A prototype real-time system was also developed as part of this work. The intention was to demonstrate the potential capabilities of subvocal speech recognition in an acoustically harsh environment. Figure 9 gives an overview of the real-time system architecture. Java classes (Java version 1.5.0; Sun Microsystems, Inc.; Santa Clara, CA) were developed and called from Matlab to facilitate communication between the amplifier and a PC used for computation. Java was also used to communicate recognized words to servers. The real-time system used the same activity detection and feature extraction Matlab code as the off-line system.

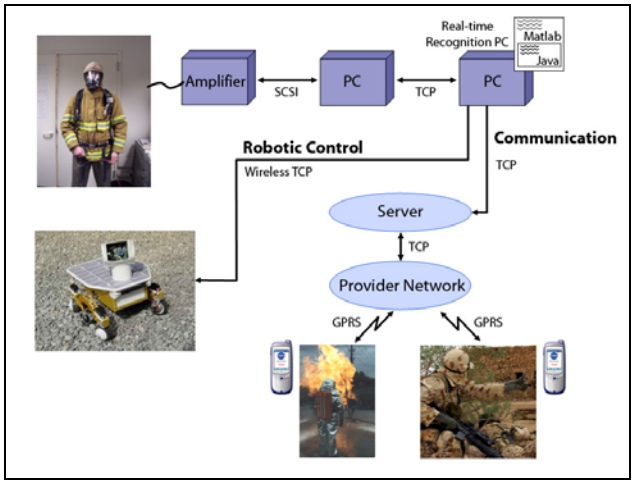

Figure 9. Overview of real-time system architecture.

There is no question that the difficult part of this research effort involved EMG word recognition. Once recognized, using the words for communication and control was relatively trivial. Two such paths were constructed, purely to provide concrete examples of real-time subvocal speech recognition. One involved sending recognized words to Smartphones running Windows Mobile 2003 Second Edition over GPRS wireless links. A small amount of custom client code, written using

Microsoft's .NET Compact Framework, was loaded onto the Smartphones. Recognized words would be displayed on the phone's screen and pre-recorded audio clips would be played on the device. The second output path was focused on control of a device, in this case a Personal Exploration Rover (PER) built by Carnegie Mellon University Robotics [4]. Communication was done via an 802.11b wireless link and made use of the Java API supplied by the PER designers.

4. Results

This section examines results obtained both in off-line analysis and with the real-time system.

4.1. Recognition Rate Results

Table 2 gives the confusion matrix that resulted from off-line analysis. Each entry is an average classification percentage, computed using bootstrapping [45]. Although extremely expensive computationally, the resulting statistics are stable. Collected samples of each word were randomly assigned to either a training set or a generalization set, with 70% of the samples going into the training set. A neural network was then trained, using only elements from the training set, and tested on the generalization elements. The result was a 15×15 confusion matrix. The entire process was then repeated 500 times, beginning with a new random assignment of samples to training and generalization sets. The elements shown in Table 2 are the average values across the 500 confusion matrices.

The same bootstrapping technique was used to compute the overall average correct classification rate and 95% confidence interval. They are 74% and [71%, 77%], respectively.

4.2. Real-time Results

Once trained, the neural network was inserted into the real-time system for purposes of testing. The network was used by the same subject for whom it was trained. Before being inserted into the real-time system, the network was checked against its generalization set to ensure that it was not an outlier in terms of recognition rate.

At the time of writing, no quantitative results are available for the real-time system. Qualitative results are shared instead. Recognition rates for the real-time system seemed consistent with the off-line analysis but remain to be determined.

The real-time system was tested in the presence of approximately 95 decibels of acoustic noise. Recall that initial data collection sessions were done under quiet conditions. Testing was done in the same laboratory in which data had been collected (i.e., not in a controlled chamber). Noise was generated using speakers and consisted of sounds common to firefighting environments such as engine noise, saws, and sirens. The environment was sufficiently loud that acoustic communication between individuals in the room was possible only by shouting. Acoustic communication with the subject was essentially impossible, given the additional muffling of his mask.

Table 2. Confusion matrix showing average classification percentages (refer to Table 1 for class codes). Only non-zero percentages are shown. Diagonal elements are shaded for convenience. Due to rounding, rows may not sum to 100.

	Classification Result														
Truth	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
C1	95		1				1	1	1			1			
C2		86	1			1	2			1	3		4		1
C3		1	90	5	2										2
C4			4	68	13			9		1		2		1	1
C5	1		1	16	72			1				1			6
C6		3				75	2	1	1		2	1	13	2	1
C7	1					1	64	1	15	2	12		3	1	
C8	1		1	12	1		1	58	4	6		9	2	4	1
C9	1				1		18	4	61	4	3	4	2	2	
C10		2		1			3	5	6	74	6	3			
C11	1	4				3	17	1	4	5	64		1		
C12		1		3	1			7	2	1		71		13	
C13		4				10	2	2	2		1	1	74	4	
C14						3	1	1	2		1	13	4	75	
C15	1		2	3	7			1				1	1		84
Total	7	7	6	7	7	6	7	6	7	6	6	7	7	7	6

To demonstrate real-time communication, the subject would first place a normal cell phone call to an observer stationed just outside the laboratory. With the high ambient noise, the observer would verify that the subject's speech was unintelligible. The subject would then use the EMG real-time system and place a call to the observer's Smartphone. The observer could then observe and hear words from the subject's trained vocabulary on the Smartphone. The ambient noise had no discernible effect on recognition rates, although this was true only after some conditioning of the subject. The subject's initial reaction when faced with the acoustic noise was to shout. Shouting changed the EMG characteristics sufficiently to noticeably degrade recognition. Future work will look to have the recognition system compensate for this Lombard-like EMG effect.

The other real-time use of the system, also done in the face of environment noise, was controlling a robotic device. The subject's vocabulary was mapped to robot actions (e.g., move forward 50 cm, move backward 50 cm, immediate stop, etc.). The subject would then attempt to move the robot in a controlled fashion around the top of an approximately 1-by-2 meter table. The $8K cost of the PER provided an incentive to keep the robot on the table.

5. Discussion

The overall average correct classification rate of 74% is similar to other EMG-based speech recognition reports using vocabularies of similar size (as surveyed in Section 2.2). The rate is an order of magnitude greater than the a priori rate of 6.7%. Those more familiar with conventional speech recognition systems may find the rate low, but it is important to note that this is a raw recognition rate. No higher-level processing, such as using context or forcing user repetition, has been done. Such efforts will only serve to increase the correct classification rate of a production system. For example, swallowing is well known to produce significant EMG activity in the region of the neck. The current real-time implementation recognizes swallowing (and coughing) as activity and then makes a forced vocabulary choice, reducing the real-time recognition rate.

An obvious limitation of the study was the recruitment of only a single subject. This was due in part to the difficulty of finding subjects trained in the use of SCBA equipment and able to devote enough time to data collection. Although our previous work with non-SCBA EMG speech recognition suggests the results reported here will generalize to other subjects, this remains to be demonstrated for SCBA use.

Importantly, we observed no noticeable impact on the EMG signal from positive-pressure breathing via the SCBA. In other work we have done, as yet unpublished, we have similarly noticed no impact while breathing off open-circuit SCUBA equipment (in a dry laboratory setting).

The issue of sensor placement sensitivity was not addressed in this study. An initial sensor placement was made by experimenting with different locations and finding one that particularly suited the subject (gauged by a strong EMG response during phonation). Subsequent sessions used the same sensor location, to within the accuracy afforded by a digital photograph of the initial location.

This work made use of only a single channel of EMG data. Sensors were mounted on the neck, in part because the SCBA mask would have posed challenges for facial muscle sensing. In contrast to earlier EMG work, it was of interest to establish what performance could be achieved with only a single EMG channel. Fewer channels imply fewer sensors in a deployed system, thereby reducing system complexity and increasing the probability of acceptance by first responders.

6. Conclusions and Future Work

Our study provides preliminary evidence that a small tactical vocabulary can be communicated via EMG recognition alone, while wearing SCBA equipment and in an acoustically harsh environment, with an average correct classification rate of at least 74%.

We believe EMG-based speech recognition, even in isolated-word form, holds promise as a communication modality for first responders and others. However, before a prototype system could be field tested, many significant obstacles would have to be overcome:

1. A comfortable and realistic method would have to be found for reliably fitting a user with sensors. The sensors would need to interoperate with other equipment the user required (e.g., a breathing mask). The sensors would have to remain in place during severe physical exertion and be resistant (or immune) to perspiration.
2. Equipment would need to be miniaturized and hardened for field use.
3. The signal processing core would need to deal with swallowing and coughing, Lombard-like effects of changing EMG characteristics in the presence of acoustical noise, and movement artifacts (e.g., twisting of the neck).
4. Computational requirements would need to be made consistent with those typically found in wearable environments.

There are several avenues for future work. For the system we have developed, improved activity detection would be beneficial. The real-time system performance

needs to be quantified. We have begun preliminary work on adaptively canceling the EMG noise before feature extraction and believe this line of work will increase the recognition rate. We are interested in potential applications of subvocal speech recognition to people with disabilities. Finally, there is substantial research yet to be done to produce a real-time EMG-based continuous speech recognition system.

7. Acknowledgements

The authors gratefully acknowledge the advice and support offered by members of the Moffett Field Fire Department, in particular Mr. Nate Ward, Battalion Chief Gary Alstrand, and Chief John Mac Donnell.

8. References

[1] J.-C. Junqua, S. Fincke, and K. Field, "The Lombard effect: A reflex to better communicate with others in noise," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, 1999, pp. 2083–2086.

[2] J.-C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *Journal of the Acoustical Society of America*, vol. 93, no. 1, 1993, pp. 512–524.

[3] K. Faaborg-Andersen, "Electromyographic investigation of intrinsic laryngeal muscles in humans: An investigation of subjects with normally movable vocal cords and patients with vocal cord paresis," *Acta Physiologica Scandinavica*, vol. 41, no. 140, 1957, pp. 1–148.

[4] Carnegie Mellon University Robotics, "The Personal Exploration Rover," http://www.cs.cmu.edu/~personalrover/PER/. Verified November 2005.

[5] J.V. Basmajian and C.J. De Luca, *Muscles Alive: Their Functions Revealed by Electromyography*, 5th ed. Baltimore, MD: Williams & Wilkins, 1985.

[6] C.J. De Luca, "Physiology and mathematics of myoelectric signals," *IEEE Transactions on Biomedical Engineering*, vol. BME-26, no. 6, 1979, pp. 313–325.

[7] B. Gerdle et al., "Acquisition, processing and analysis of the surface electromyogram," in *Modern Techniques in Neuroscience*, U. Windhorst and H. Johansson, Eds. Berlin: Springer Verlag, 1999, pp. 705–755.

[8] J.D. Bronzino, ed., *The Biomedical Engineering Handbook*, Boca Raton, FL: CRC Press, 1995.

[9] R.N. Scott, "Myoelectric Control Systems," in *Advances in Biomedical Engineering and Medical Physics*, vol. 2, 1968, pp. 45–72.

[10] E.D. Sherman, "A Russian bioelectric-controlled prosthesis: Report of a research team from the Rehabilitation Institute of Montreal," *Canadian Medical Association Journal*, vol. 91, no. 24, 1964, pp. 1268–1270.

[11] D.R. Taylor, Jr., "A bioelectric pattern recognition control for prosthesis," *Proceedings of the Conference on Cybernetic Problems in Bionics*, 1966, pp. 885–893.

[12] C.J. De Luca, "Surface electromyography: Detection and recording," http://www.delsys.com/library/papers/SEMGintro.pdf. Verified November 2005.

[13] A.D.C. Chan and K.B. Englehart, "Continuous myoelectric control for powered prostheses using hidden Markov models," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 1, 2005, pp. 121–124.

[14] K.R. Wheeler and C.C. Jorgensen, "Gestures as input: Neuroelectric joysticks and keyboards," *IEEE Pervasive Computing*, vol. 2, no. 2, 2003, pp. 56–61.

[15] L.J. Trejo et al., "Multimodal neuroelectric interface development," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no. 2, 2003, pp. 199–204.

[16] B. Hudgins, P. Parker, and R.N. Scott, "A new strategy for multifunction myoelectric control," *IEEE Transactions on Biomedical Engineering*, vol. 40, no. 1, 1993, pp. 82–94.

[17] N. Sugie and K. Tsunoda, "Speech prosthesis employing a speech synthesizer—vowel discrimination from perioral muscle activities and vowel production," *IEEE Transactions on Biomedical Engineering*, vol. BME-32, no. 7, 1985, pp. 485–490.

[18] M.S. Morse, "Design and implementation of a scheme to recognize speech from myoelectric inputs using maximum likelihood pattern recognition," doctoral dissertation, Clemson University, 1985.

[19] M.S. Morse and E.M. O'Brien, "Research summary of a scheme to ascertain the availability of speech information in the myoelectric signals of neck and head muscles using surface electrodes," *Computers in Biology and Medicine*, vol. 16, no. 6, 1986, pp. 399–410.

[20] M.S. Morse, Y.N. Gopalan, and M. Wright, "Speech recognition using myoelectric signals with neural networks," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 13, 1991, pp. 1877–1878.

[21] M.S. Morse et al., "Use of myoelectric signals to recognize speech," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 6, 1989, pp. 1793–1794.

[22] M.S. Morse, S.H. Day, and J. May, "Time domain analysis of the myoelectric signal secondary to speech," *Proceedings of the 12th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1990, pp. 1318–1319.

[23] A.D.C. Chan et al., "Hidden Markov model classification of myoelectric signals in speech," *Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2, 2001, pp. 1727–1730.

[24] A.D.C. Chan et al., "A multi-expert speech recognition system using acoustic and myoelectric signals,"

Proceedings of the 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society, vol. 1, 2002, pp. 72–73.

[25] C. Jorgensen, D.D. Lee, and S. Agabon, "Sub auditory speech recognition based on EMG signals," *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, IEEE, vol. 4, 2003, pp. 3128–3133.

[26] C. Jorgensen and K. Binsted, "Web browser control using EMG based sub vocal speech recognition," *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS)*, IEEE, 2005, pp. 294c.1–294c.8.

[27] H. Manabe, A. Hiraiwa, and T. Sugimura, "Unvoiced speech recognition using EMG—Mime speech recognition," *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2003, pp. 794–795.

[28] H. Manabe and Z. Zhang, "Multi-stream HMM for EMG-based speech recognition," *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2, 2004, pp. 4389–4392.

[29] S. Kumar et al., "EMG based voice recognition," *Proceedings of the 2004 Intelligent Sensors, Sensor Networks and Information Processing Conference*, 2004, pp. 593–597.

[30] M. Graciarena et al., "Combining standard and throat microphones for robust speech recognition," *IEEE Signal Processing Letters*, vol. 10, no. 3, 2003, pp. 72–74.

[31] A. Shahina and B. Yegnanarayana, "Language identification in noisy environments using throat microphone signals," *Proceedings of the 2005 International Conference on Intelligent Sensing and Information Processing (ICISIP '05)*, 2005, pp. 400–403.

[32] Szu-Chen Jou, T. Schultz, and A. Waibel, "Whispery speech recognition using adapted articulatory features," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, 2005, pp. 1009–1012.

[33] K. Brady et al., "Multisensor MELPe using parameter substitution," *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, vol. 1, 2004, pp. 477–480.

[34] L.C. Ng et al., "Denoising of human speech using combined acoustic and EM sensor signal processing," *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, vol. 1, 2000, pp. 229–232.

[35] Television Equipment Associates, "Invisio Bone-mic Headset," http://www.swatheadsets.com/invisio/invisio2005_web.pdf. Verified November 2005.

[36] J.L. Durkin and J.P. Callaghan, "Effects of minimum sampling rate and signal reconstruction on surface electromyographic signals," *Journal of Electromyography and Kinesiology*, vol. 15, no. 5, 2005, pp. 474–481.

[37] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Transactions on Speech and Audio Processing*, to appear, pp. 1–13.

[38] K. Li, M.N.S. Swamy, and M.O. Ahmad, "An improved voice activity detection using higher order statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, 2005, pp. 965–974.

[39] J. Ramírez et al., "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Processing Letters*, vol. 12, no. 10, 2005, pp. 689–692.

[40] J. Ramírez et al., "An effective subband OSF-based VAD with noise reduction for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, 2005, pp. 1119–1129.

[41] B. Ning et al., "A robust speech recognition system embedded in CDMA cellular phone chipsets," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, vol. 4, 2002, pp. 3804–3807.

[42] J.-C. Junqua, B. Mak, and B. Reaves, "A robust algorithm for word boundary detection in the presence of noise," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, 1994, pp. 406–412.

[43] N. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," *Applied and Computational Harmonic Analysis*, vol. 10, no. 3, 2001, pp. 234–253.

[44] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Upper Saddle River, NJ: Prentice Hall, 1999.

[45] B. Efron and R.J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall, 1993.