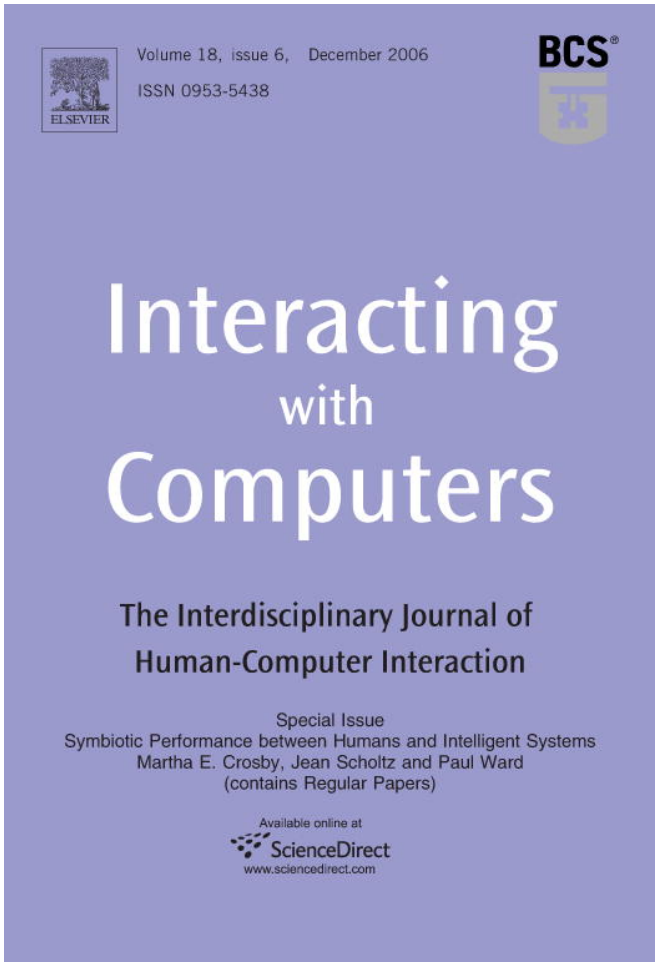

Available online at www.sciencedirect.com
ScienceDirect
Interacting with Computers 18 (2006) 1242–1259
Interacting with Computers
www.elsevier.com/locate/intcom

Small-vocabulary speech recognition using surface electromyography ☆

Bradley J. Betts [a], Kim Binsted [b,*], Charles Jorgensen [c]

[a] *QSS Group Inc., NASA Ames Research Center, M/S 269-1, Moffett Field, CA 94035-1000, USA*
[b] *NASA-UH Astrobiology Institute, Information and Computer Sciences Department, University of Hawaii, Post 317, 1680 East-West Road, Honolulu, HI 96744, USA*
[c] *Neuro-Engineering Laboratory, NASA Ames Research Center, M/S 269-1, Moffett Field, CA 94035-1000, USA*

Available online 10 October 2006

Abstract

We present results of electromyographic (EMG) speech recognition on a small vocabulary of 15 English words. EMG speech recognition holds promise for mitigating the effects of high acoustic noise on speech intelligibility in communication systems, including those used by first responders (a focus of this work). We collected 150 examples per word of single-channel EMG data from a male subject, speaking normally while wearing a firefighter's self-contained breathing apparatus. The signal processing consisted of an activity detector, a feature extractor, and a neural network classifier. Testing produced an overall average correct classification rate on the 15 words of 74% with a 95% confidence interval of (71%, 77%). Once trained, the subject used a classifier as part of a real-time system to communicate to a cellular phone and to control a robotic device. These tasks were performed under an ambient noise level of approximately 95 decibels. We also describe ongoing work on phoneme-level EMG speech recognition.

☆ This paper is in part authored by employees of the U.S. Government and is in the public domain. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government. Portions of this work were described in NASA technical memorandum NASA/TM-2005-213471.

* Corresponding author. Tel.: +1 808 956 6107; fax: +1 808 956 3548.
E-mail addresses: bradley.betts@stanfordalumni.org (B.J. Betts), binsted@hawaii.edu (K. Binsted), cjorgensen@mail.arc.nasa.gov (C. Jorgensen).

0953-5438/$ - see front matter
doi:10.1016/j.intcom.2006.08.012

Keywords: Electromyography; EMG; Bioelectric; EMG speech recognition; First responder; Pattern recognition; SCBA

1. Introduction

People have long had an interest in communicating in acoustically noisy environments. Research and development in this area has largely been driven by military needs. Significant research was done before and during the Second World War on techniques to allow pilot voice communication in airplanes. This resulted in the development of devices such as throat microphones, interest in which continues to this day, particularly when used as part of a multi-modality speech recognition system (Graciarena et al., 2003; Shahina and Yegnanarayana, 2005; Jou et al., 2005). Military research in the area continues, focusing on sensors and techniques appropriate for communicating in noisy environments (Brady et al., 2004; Ng et al., 2000). 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.

First responders are an example of a class of users that stand 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. Moreover, the breathing apparatus distorts and muffles the firefighter's speech, further affecting communication. This work was motivated in part by a desire to see whether electromyographic (EMG) speech recognition could alleviate these effects. Electromyography is 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 (Junqua, 1993; Junqua et al., 1999), require further study.

Unfortunately, in many cases first responders have yet to benefit from advanced techniques designed to counteract acoustic noise. For example, in many fire departments, voice communication is still done by shouting through the mask of a self-contained breathing apparatus (SCBA) into a shoulder-mounted or hand-carried radio. Some alternatives have been developed and targeted at first responders (e.g., bone conduction microphones and in-mask boom microphones) but have yet to receive wide deployment. Our study suggests that bioelectric techniques, either on their own or fused with other modalities, hold promise for this community.

In addition to noisy environments, EMG-based speech has potential in environments where sound does not carry well or at all (e.g., underwater), where discreet or

secure communications are necessary or desirable (e.g., military applications, off-line comments during meetings), and for users with speech-related disabilities (e.g., vocal cord damage). 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 researcher Faaborg-Andersen (1957). 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.

NASA is interested in multi-modal interfaces as a way of increasing communication robustness and reducing information overload in human–human and human–agent systems. Imagine an astronaut exploring the surface of Mars, in collaboration with other human astronauts, a variety of robotic rovers, intelligent shipboard systems, and a mixed human–agent team back on Earth. In a space suit, input and output modalities are severely limited; audible speech is by far the most convenient communications channel. However, audible speech interfaces cannot support many parallel interactions and, in the case of human–agent communication, are not very robust in the presence of noise, speaker stress, changes in gas mixture, etc. By adding subvocal speech to the repertoire of space suit interface designers, we hope to increase the robustness of audible speech by providing redundancy and by providing an alternate means of communication when appropriate (e.g., discreet communications) or necessary (e.g., a physiological problem renders the audible speech interface unusable).

As suggested by the exploration scenario above, effective communication is essential to symbiotic performance, whether between humans or between humans and agents. Skilled humans in a complex situation typically have many channels of communication (speech, gesture, facial expression, etc.) at their disposal, and direct human–human communication is very robust under noise and other disruptors. We hope that our research will help to make human–agent communication as effective, by increasing robustness and providing alternate channels, so as to enable the kind of complex collaboration required to achieve NASAs exploration goals. We also imagine a future in which first responders are as well supported technologically as astronauts are, one in which first-responder teams consist of humans, robots, and software agents working together to deal with crises. Robust, multi-channel communication will be essential for these teams to collaborate effectively.

For our purposes, *EMG-based speech recognition* is the decoding of natural language speech, whether vocalized or sub-vocalized, based solely on the EMG signal from sensors placed on the neck and/or face of the speaker. In the case of subvocal speech recognition, the speaker makes no audible sound, but instead lets his or her tongue and throat move *as if* trying to produce audible speech. In the work reported here, sensor placement was the same for vocalized as for subvocalized speech.

We collected EMG data from a single male subject wearing an SCBA under laboratory conditions. Data samples consisting of isolated words chosen from a small English vocabulary were used to train a neural network classifier. The trained network was then inserted into a real-time communication and control system while the subject was exposed to approximately 95 dB 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. In some ongoing work, preliminary results from two female subjects sub-vocalizing phonemes of the English language, but without SCBA equipment, are presented.

The remainder of this paper is organized as follows. First, we give a brief background on the history and physiology of EMG. Then, we survey other research efforts that have examined EMG speech recognition. Finally, methods and results are followed by a discussion, conclusions, and avenues for future work.

2. Background and related research

Electromyography is the study of muscle function via its electrical properties (i.e., the electrical signal emanating from the muscle during muscle activation) (Basmajian and De Luca, 1985; De Luca, 1979; Gerdle et al., 1999; Bronzino, 1995). As detailed by Basmajian and De Luca (1985), electromyography has a long and interesting history. In 1848, 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 (Scott, 1968; Sherman, 1964; Taylor, 1966). 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 speech recognition. There is a rich body of literature on the use of EMG for control of prostheses and for gesture recognition not described here [(Chan and Englehart, 2005; Wheeler and Jorgensen, 2003; Trejo et al., 2003; Hudgins et al., 1993) are but a few of the many examples].

Chapters 19 and 20 of Basmajian and De Luca (1985) describe electromyography research done before 1985 related to the muscles of the mouth, pharynx, larynx, face, and neck. As it pertains to speech, the goal of research during that period was largely the understanding of muscle processes associated with phonation in normal subjects and subjects with disabilities. 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 speech recognition,

the concept almost surely occurred to earlier researchers – the state of digital computing and non-invasive sensing may have been the limiting factors (there are two principal sensing techniques used in electromyography: invasive indwelling sensing and non-invasive surface sensing; this paper focuses on the use of surface sensors (Gerdle et al., 1999 and De Luca, 2005)).

The first efforts in EMG speech recognition occurred independently and in parallel in Japan and the United States around 1985–86. In Japan, Sugie and Tsunoda (1985) used three channels of silver silver-chloride (Ag–AgCl) surface sensors with a sampling rate of 1250 samples/channel/s. A threshold-and-counting scheme was used to produce a three-bit number every 10 ms and a finite automaton was used for vowel discrimination. Three subjects were asked to say aloud 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 (1985; with O'Brien, 1986) 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 reduced 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 say aloud each word twenty times and a maximum likelihood technique was used for classification. 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%. Other papers from this group include (Morse et al., 1989, 1990).

In 2001, Chan et al. (2001) reported EMG speech recognition results that were motivated by the need to communicate in acoustically harsh environments (in this case the cockpit of a fighter aircraft). 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 principal component analysis were used to reduce the data to thirty features per word on a normally spoken 10-word vocabulary (the 10 English digits). Classification was performed using linear discriminant analysis (LDA). Recognition rates as high as 93% were achieved in an experiment where words were randomly presented to two subjects. In later work, a hidden Markov model (HMM) was used as the classification engine and achieved results similar to the LDA technique (Chan et al., 2001). 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 (Chan et al., 2002).

In 2003, Manabe et al. (2003) used a novel surface sensor mounting configuration for EMG speech recognition. 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 EMG values during

pronunciation of a vowel. Over three subjects, each using a vocabulary of five Japanese vowels spoken aloud, the average correct classification rate exceeded 90%. In later work, Manabe and Zhang (2004) made use of HMMs to classify the 10 Japanese digits collected from 10 subjects; accuracies as high as 64% were achieved.

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

Prior EMG speech recognition has been performed by our group, including the first work we are aware of investigating sub-vocalized (i.e., non-audible) EMG speech. Jorgensen et al. (2003, 2005) collected six sub-vocalized words from three subjects using surface Ag–AgCl sensors and a single EMG channel. 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. An average correct recognition rate of 92% was achieved using a neural network classifier.

3. Methods

3.1. Procedure

We collected training data 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 Survivair Panther SCBA unit as shown in Fig. 1. The SCBA was pressurized per normal SCBA usage. The subject was instructed to

Fig. 1. Photo showing data collection station and subject in SCBA equipment.

breathe normally during data collection sessions (i.e., as he would while wearing SCBA equipment). We paused collection sessions as necessary to replace empty air tanks.

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 Fig. 2. 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 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.

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, we decided to use a high sampling rate. The price of so doing was increased data storage and increased computational demands on the real-time implementation, but the benefit was that it allowed us to study the EMG speech signal over a wider frequency range. That is, if a signal was sampled at, say, 1 kHz, signal energy higher than half this amount would be aliased into the resulting samples. Of course, if there was no signal energy at frequencies higher than 500 Hz, aliasing would be eliminated. In this case, sampling at 1 kHz would be called *sampling at the Nyquist rate* and would allow for perfect reconstruction of the signal

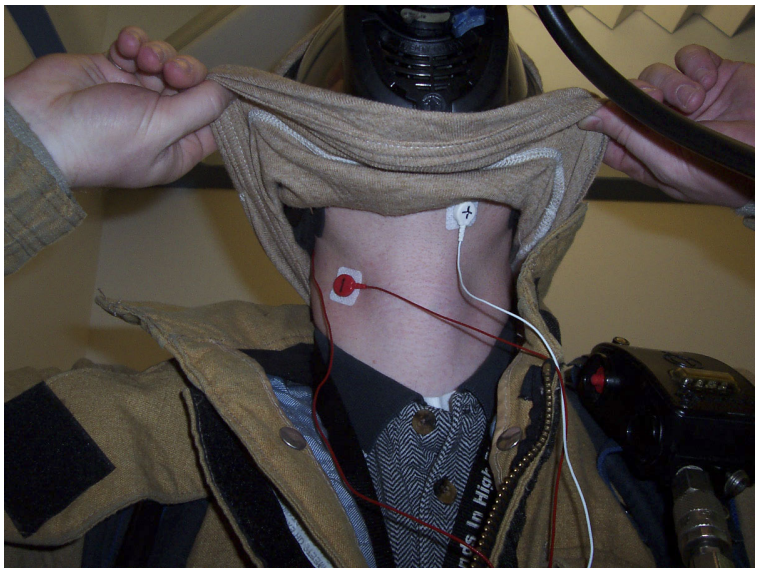

Fig. 2. Photo showing EMG sensor placement. The subject has peeled his hood back to reveal the sensors. A third sensor, placed behind the subject's right ear (or alternatively placed on the wrist) was used as a ground.

from its samples. Our anti-aliasing filter restricted the signal bandwidth to 2 kHz. However, since such filters are not perfect (meaning signal energy beyond the frequency cutoff does in fact pass through the filter), a rule-of-thumb is to sample at five times the cut frequency, leading to our 10 kHz rate. In a useful recent article, Durkin and Callaghan (2005) examine sampling rate issues associated with surface electromyography.

Fifteen isolated words were collected 150 times each, for a total of 2250 word samples. The words, shown in Table 1, were chosen from a list compiled by firefighters at the Moffett Field Fire Department as representative of the tactical vocabulary they use. Some of the vocabulary elements are in fact two-word phrases; we nonetheless use the term "word" for consistency with other published accounts.

The subject was prompted via software to say the vocabulary words in a fully randomized order. Randomization was used to minimize learning and anticipatory effects. The subject had a fixed amount of time in which to say a word, with a pause between words of 2.5 s. 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 3-week span. The subject was photographed during the first recording session to establish where the sensors were located on the neck. In subsequent data collection sessions, an assistant placed the sensors in the same location with the aid of the photograph.

As will be described in greater depth, a portion of the subject's collected data was used to train a classifier. This classifier was then inserted into a real-time system. The subject was reconnected to the system, once again using photographs to replicate the senor locations used during training data collection. Instead of collecting word samples, EMG words were now recognized in real-time, this time in the presence of approximately 95 dB of ambient acoustic noise (generated using speakers and consisting of sounds common to firefighting scenes, including sirens, saws, engines, and SCBA breathing sounds). The subject was instructed to speak in a normal audible fashion, as during training data collection. The combination of SCBA mask plus ambient noise meant the subject's words were completely inaudible to observers. The output of the real-time system was manifested in several different ways. One such was mapping classification results to actions for a robotic device (e.g., a vocabulary word was mapped to the action to move the robot forward by 1 m). This meant that the subject could use the trained classifier to navigate a robot along a desired course, hoping of course to make as few course deviations as possible.

3.2. Hardware and software architecture

A goal of this study was to assess the feasibility of using EMG 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 developed system is not portable and not hardened for field use.

Table 1
Confusion matrix showing average classification percentages

Truth	Classification result														
	Evacuate	Mayday	Man-trap	Fire clear	Fire safe	Room	Status	North	South	East	West	Zero	One	Two	Three
Evacuate	95		1				1	1	1			1			
Mayday		86	1			1	2			1	3		4		1
Man-trap		1	90	5	2										2
Fire clear			4	68	13			9		1		2		1	1
Fire safe	1		1	16	72			1				1			6
Room		3				75	2	1	1		2	1	13	2	1
Status	1					1	64	1	15	2	12		3	1	
North	1		1	12	1		1	58	4	6		9	2	4	1
South	1				1		18	4	61	4	3	4	2	2	
East		2		1			3	5	6	74	6	3			
West	1	4				3	17	1	4	5	64		1		
Zero		1		3	1			7	2	1		71		13	
One		4				10	2	2	2		1	1	74	4	
Two						3	1	1	2		1	13	4	75	
Three	1		2	3	7			1				1	1		84
Total	7	7	6	7	7	6	7	6	7	6	6	7	7	7	6

Only non-zero percentages are shown. Diagonal elements are shaded for convenience. Due to rounding, rows may not sum to 100.

We used Neuroscan's Acquire software for collecting training data, and Matlab to perform training and analysis. 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. We constructed two such paths, purely to provide concrete examples of real-time EMG speech recognition in an acoustically harsh environment. One involved sending recognized words to Smartphones (i.e., programmable cellular phones) 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 corresponding to the word would be played on the device. The second output path, already described, focused on the control of a device, in this case a Personal Exploration Rover (PER) built by Carnegie Mellon University Robotics (Carnegie Mellon University Robotics, 2005). Communication was via an 802.11b wireless link and made use of the Java API supplied with the PER. Fig. 3 gives an overview of the real-time system architecture.

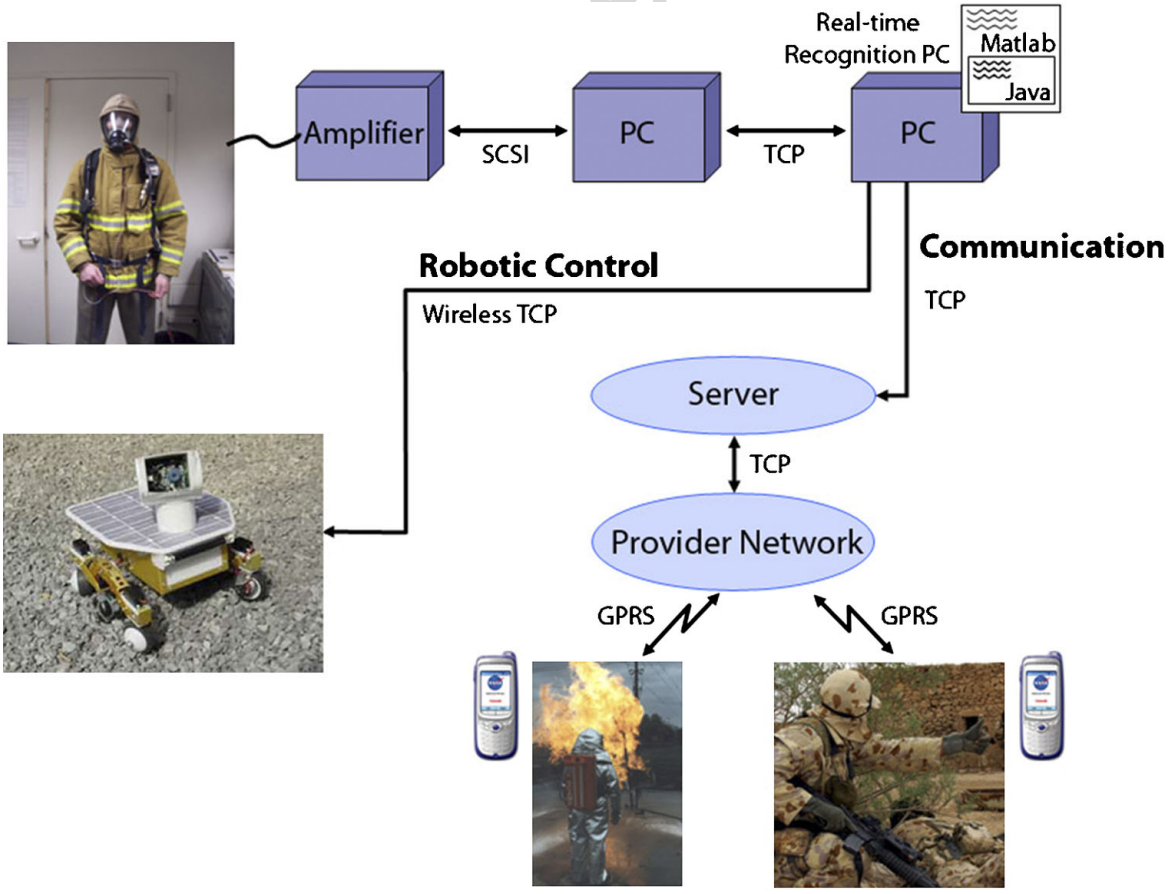

Fig. 3. Overview of real-time system architecture.

3.3. Signal processing

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. Three stages are common to both phases:

1. Signal acquisition.
2. Activity detection.
3. Feature extraction.

The signal acquisition process has already been described. The other two common blocks – activity detection and feature extraction – are described next. The signal processing pipeline developed for this work is *speaker-dependent*, meaning that training samples must be collected from each user.

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 s 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 0 s (i.e., noise) were not inserted into contiguous activity blocks and that the activity 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 s 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. While substantially more sophisticated activity detection techniques can be found in the literature (e.g., Davis et al., in print; Li et al., 2005; Ramirez et al., 2005a,b; Ning et al., 2002; Junqua et al., 1994) and are candidates for inclusion in future work, the technique described proved sufficient for both the off-line and real-time systems.

Feature extraction is the process of reducing the dimensionality of the data to 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 (Kingsbury, 2001). 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

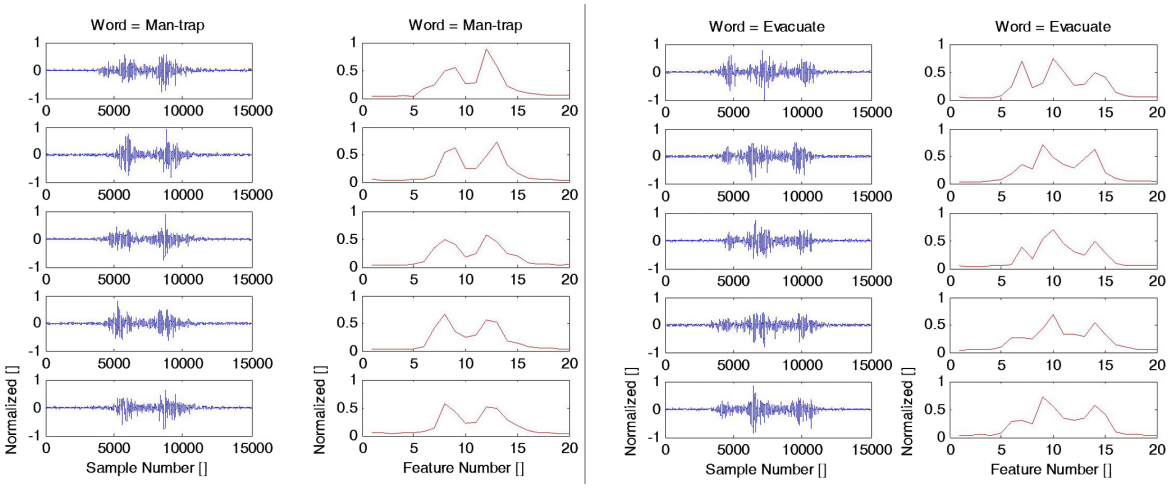

Fig. 4. Raw signal and feature examples for the words "man-trap" and "evacuate".

exact positioning of the signal within the activity window. We and others have also used HMMs in the past to ease temporal alignment issues (Jorgensen et al., 2003; Chan et al., 2001). Fig. 4 shows the output of the feature extractor on two word samples, "man-trap" and "evacuate". The left portion for each word shows the EMG activity regions. The right portion plots the feature dimensions on the abscissa and their magnitudes on the ordinate.

After activity detection and feature extraction, features from the training set were used to train a neural network classifier. We used a conjugate gradient network (Haykin, 1999). 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 Section 4, 70% of the collected samples were used for training. The remaining 30% were set aside for testing. The decision to use a neural network instead of some other classifier type (e.g., HMMs, a popular choice in the automatic speech recognition community) was in part motivated by the desire to have a fast classifier appropriate for use in our Matlab-based real-time EMG speech recognition system.

4. Results

Table 1 gives the confusion matrix that resulted from an analysis of the collected data samples. Each entry is an average classification percentage, computed using bootstrapping (Efron and Tibshirani, 1993) in a manner that we now explain. Note that although bootstrapping is expensive computationally, it was done offline (i.e., not in real-time) and so did not pose a problem. Its principal benefit is that it allows

for a straightforward way to estimate the standard error in the classification percentage.

In forming the bootstrap-based statistics, collected samples of each word were randomly assigned to either a training set or a testing 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 testing elements. The result was a 15 times 15 confusion matrix, where the row labels indicate what the word truly is and the column labels the classification result. Perfect classification would have all off-diagonal entries equal to zero. The entire process was then repeated 500 times, beginning with a completely new random assignment of samples to training and testing sets. The elements shown in Table 1 are the average values across the 500 confusion matrices.

This same bootstrapping technique was used to compute the overall average correct classification rate and 95% confidence intervals, where here "average" means "average across all words in the vocabulary". The average correct classification rate was 74% with a 95% confidence interval of [71%, 77%]. That is, if more data samples were collected from the subject and applied to the trained network, we would expect the system to correctly classify vocabulary words at least 71% of the time 19 times out of 20.

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 accurately determined. The subject was able to transmit vocabulary words both to a cell phone and to the robotic device while wearing SCBA equipment, in the face of ambient noise levels that made understanding the subject's acoustic speech essentially impossible for unaided observers (even lip reading is not an option as the SCBA occludes the mouth). The subject was able to navigate the robotic device while wearing SCBA equipment, for example moving the robot around a table-top without having it fall to the floor.

5. Discussion

The overall average correct classification rate of 74% is similar to other EMG-based speech recognition reports using vocabularies of similar size (Chan et al., 2001, 2002; Jorgensen et al., 2003, 2005; Kumar et al., 2004; Manabe et al., 2003). The rate is an order of magnitude greater than the a priori rate of 6.7% (i.e., the rate that would be achieved by a system that simply guessed at the word, which would be correct one-fifteenth of the time for a 15-word vocabulary). 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. One simple fix for this problem would be to allow the system to categorize an item as ''unrecognized'' if the classification activation is below some threshold.

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).

This work has several substantial differences from other research done in the field, the two principal ones being the use of SCBA equipment and the use of a single EMG channel. SCBA equipment is an everyday fact of life for firefighters; a communication scheme that cannot coexist with this equipment is unusable. Demonstrating the feasibility of detecting EMG speech signals while wearing this gear is an important result. The use of a single channel of EMG data is another important result compared to other EMG speech efforts. The number of channels dictates the number of required sensor locations in and around the face. This has important practical ramifications for first responders; fewer required sensors are unquestionably more practical for in-field use as they decrease system complexity. In this work, sensors were mounted on the neck, in part because the SCBA mask would have posed challenges for facial muscle sensing.

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 study also did not tackle several practical problems with the use of EMG sensors in the field. Sweat can affect EMG signals, and keeping the sensors attached under exertion is challenging. That said, the design of EMG sensors has advanced considerably in recent years, and we expect that newer sensors will alleviate these problems, including allowing for EMG sensing without requiring the sensors to be in direct contact with the skin.

The signal processing pipeline developed for this work was speaker-dependent, meaning that training samples were required from each system user. These training samples were then used to train a neural network specifically for a given user. This requirement for user-specific training limits the usefulness of a recognition technology. For example, a firefighter using a production system would want to be able to easily switch to another system in the event of a failure. One obvious work-around would be to encode user-specific classifier components on something like a flash memory stick, allowing it to be easily moved from one system to another. An even better solution would make the signal processing pipeline speaker independent (while perhaps allowing for optional per-user tuning). It is our belief that developing a robust speaker-independent EMG speech recognition system would be worthwhile, possible, and difficult.

Although a small vocabulary recognition system is sufficient for some applications, others need to recognize continuous, broad-vocabulary speech. Also, some applications (such as those providing covert communication) need to recognize sub-vocalized speech (some of these issues are discussed in (Jorgensen and Binsted, 2005). We have conducted preliminary work in this area, collecting data from two female subjects wearing no special equipment (aside from the EMG sensors) in a laboratory environment. Instead of discrete words, subjects were asked to sub-vocalize English language phonemes as shown in Table 2. Each subject was asked to sub-vocalize each phoneme, while thinking of the target word for that phoneme. For example, while sub-vocalizing the phoneme ao, the subject would focus on the central vowel sound of the word "dog" as shown in Table 2. While this work is of a more preliminary nature at the time of writing, we believe it is an important stepping stone towards continuous EMG speech recognition. It has helped us establish a baseline for sub-vocal EMG speech recognition. Also, it seems that features of spoken speech that are relevant to auditory speech recognition are also relevant to sub-vocal EMG speech recognition. This suggests that techniques which have proven useful in processing spoken speech, such as diphone or triphone recognition, would also be useful in processing sub-vocal EMG speech, a direction we intend to continue investigating.

Table 2
English language phonemes and key words

Phonemes			
Vowels	Words	Consonants	Words
ax	ago	*b*	big
ay	bite	*ch*	chin
uh	book	*k*	cut
aa	car	*d*	dig
ah	cut	*f*	fork
ey	day	*zh*	genre
ao	dog	*g*	gut
iy	feel	*hh*	help
aw	foul	*jh*	joy
ae	gas	*l*	lid
ow	go	*m*	mat
ih	hit	*n*	no
axr	percent	*p*	put
eh	pet	*r*	red
ix	sick	*sh*	she
uw	tool	*sh*	sit
oy	toy	*t*	talk
er	turn	*dh*	then
		th	thin
		v	vat
		w	with
		y	yacht
		z	zap

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, and even if it has to be trained for individual users, holds promise as a communication modality for first responders and others. However, before a prototype system can be field tested, many significant obstacles will have to be overcome:

1. We must test the system with more subjects to permit generalization.
2. A comfortable and realistic method must be found for reliably fitting a user with sensors. The sensors need to interoperate with other equipment the user requires (e.g., an SCBA). The sensors have to remain in place during severe physical exertion and be resistant or immune to perspiration.
3. Equipment must be miniaturized and hardened for field use.
4. The signal processing core must deal with swallowing and coughing, Lombard-like effects in the presence of acoustical noise, and movement artifacts such as twisting of the neck.
5. Computational requirements must be made consistent with those typically found in wearable environments.

There are several avenues for future work. We need to quantify the effect of SCBA equipment on audible speech, so that we can accurately assess the added value of the EMG system. 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 EMG speech recognition to people with disabilities. Finally, there is substantial research yet to be done to produce a real-time EMG continuous speech recognition system.

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. Part of this work was supported by the NASA Faculty Fellowship Program.

References

Basmajian, J.V., De Luca, C.J., 1985. Muscles Alive: Their Functions Revealed by Electromyography, Fifth ed. Williams & Wilkins, Baltimore.

Brady, K., Quatieri, T.F., Campbell, J.P., Campbell, W.M., Brandstein, M., Weinstein, C.J., 2004. Multisensor MELPe using parameter substitution. In: Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04), vol. 1, pp. 477–480.

Bronzino, J.D. (Ed.), 1995. The Biomedical Engineering Handbook. CRC Press, Boca Raton, FL.

Carnegie Mellon University Robotics, The Personal Exploration Rover, <http://www.cs.cmu.edu/~personalrover/PER//>, accessed November 2005.

Chan, A.D.C., Englehart, K., Hudgins, B., Lovely, D.F. Hidden Markov model classification of myoelectric signals in speech. In: Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2001, pp. 1727–1730.

Chan, A.D.C., Englehart, K., Hudgins, B., Lovely D.F. A multi-expert speech recognition system using acoustic and myoelectric signals. In: Proceedings of the 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society, 2002, pp. 72–73.

Chan, A.D.C., Englehart, K., 2005. Continuous myoelectric control for powered prostheses using hidden Markov models. IEEE Transactions on Biomedical Engineering 52 (1), 121–124.

Davis, A., Nordholm, S., Togneri, R., in print. Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold. IEEE Transactions on Speech and Audio Processing, 1–13.

De Luca, C.J., 1979. Physiology and mathematics of myoelectric signals. IEEE Transactions on Biomedical Engineering BME-26 (6), 313–325.

De Luca, C.J., Surface electromyography: Detection and recording, <http://www.delsys.com/library/papers/SEMGintro.pdf/>, accessed November 2005.

Durkin, J.L., Callaghan, J.P., 2005. Effects of minimum sampling rate and signal reconstruction on surface electromyographic signals. Journal of Electromyography and Kinesiology 15 (5), 474–481.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.

Faaborg-Andersen, K., 1957. 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 41 (140), 1–148.

Gerdle, B., Karlsson, S., Day, S., Djupsjobacka, M., 1999. Acquisition, processing and analysis of the surface electromyogram. In: Windhorst, U., Johansson, H. (Eds.), Modern Techniques in Neuroscience. Springer Verlag, Berlin, pp. 705–755.

Graciarena, M., Franco, H., Sonmez, K., Bratt, H., 2003. Combining standard and throat microphones for robust speech recognition. IEEE Signal Processing Letters 10 (3), 72–74.

Haykin, S., 1999. Neural Networks: A Comprehensive Foundation, Second ed. Prentice Hall, Upper Saddle River, NJ.

Hudgins, B., Parker, P., Scott, R.N., 1993. A new strategy for multifunction myoelectric control. IEEE Transactions on Biomedical Engineering 40 (1), 82–94.

Jorgensen, C., Lee, D.D., Agabon, S., 2003. Sub auditory speech recognition based on EMG signals. Proceedings of the International Joint Conference on Neural Networks (IJCNN), 3128–3133.

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

Jou, S.-C., Schultz, T., Waibel, A., 2005. Whispery speech recognition using adapted articulatory features. In: Proceedings of the IEEE International Conference of Acoustics, Speech, and Signal Processing (ICASSP '05), pp. 1009–1012.

Junqua, J.-C., 1993. The Lombard reflex and its role on human listeners and automatic speech recognizers. Journal of the Acoustical Society of America 93 (1), 512–524.

Junqua, J.-C., Mak, B., Reaves, B., 1994. A robust algorithm for word boundary detection in the presence of noise. IEEE Transactions on Speech and Audio Processing 2 (3), 406–412.

Junqua, J.-C., Fincke, S., Field, K., 1999. 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), 2083–2086.

Kingsbury, N., 2001. Complex wavelets for shift invariant analysis and filtering of signals. Applied and Computational Harmonic Analysis 10 (3), 234–253.

Kumar, S., Kumar, D.K., Alemu, M., Burry, M. EMG based voice recognition. In: Proceedings of the 2004 Intelligent Sensors, Sensor Networks and Information Processing Conference, 2004, pp. 593–597.

Li, K., Swamy, M.N.S., Ahmad, M.O., 2005. An improved voice activity detection using higher order statistics. IEEE Transactions on Speech and Audio Processing 13 (5), 965–974.

Manabe, H., Hiraiwa, A., Sugimura, T. Unvoiced speech recognition using EMG—Mime speech recognition. In: Proceedings of the ACM Conference on Human Factors in Computing Systems, 2003, pp. 794–795.

Manabe, H., Zhang, Z., Multi-stream HMM for EMG-based speech recognition. In: Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2004, pp. 4389–4392.

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

Morse, M.S., O'Brien, E.M., 1986. 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 16 (6), 399–410.

Morse, M.S., Day, S.H., Trull, B., Morse, H., 1989. Use of myoelectric signals to recognize speech. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society., 1793–1794.

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

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

Ng, L.C., Burnett, G.C., Holzrichter, J.F., Gable, T.J., 2000. Denoising of human speech using combined acoustic and EM sensor signal processing. In: Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00), vol. 1, pp. 229–232.

Ning, B., Garudadri, H., Chang, C., DeJaco, A., Qi, Y., Malayath, N., Huang, W., 2002. 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), 3804–3807.

Ramirez, J., Segura, J.C., Benitez, C., Garcia, L., Rubio, A., 2005a. Statistical voice activity detection using a multiple observation likelihood ratio test. IEEE Signal Processing Letters 12 (10), 689–692.

Ramirez, J., Segura, J.C., Benitez, C., Garcia, L., Rubio, A., 2005b. An effective subband OSF-based VAD with noise reduction for robust speech recognition. IEEE Transactions on Speech and Audio Processing 13 (6), 1119–1129.

Scott, R.N., 1968. Myoelectric Control Systems. Advances in Biomedical Engineering and Medical Physics 2, 45–72.

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

Sherman, E.D., 1964. A Russian bioelectric-controlled prosthesis: report of a research team from the Rehabilitation Institute of Montreal. Canadian Medical Association Journal 91 (24), 1268–1270.

Sugie, N., Tsunoda, K., 1985. Speech prosthesis employing a speech synthesizer—vowel discrimination from perioral muscle activities and vowel production. IEEE Transactions on Biomedical Engineering BME-32 (7), 485–490.

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

Trejo, L.J., Wheeler, K.R., Jorgensen, C.C., Rosipal, R., Clanton, S.T., Matthews, B., Hibbs, A.D., Matthews, R., Krupka, M., 2003. Multimodal neuroelectric interface development. IEEE Transactions on Neural Systems and Rehabilitation Engineering 11 (2), 199–204.

Wheeler, K.R., Jorgensen, C.C., 2003. Gestures as input: neuroelectric joysticks and keyboards. IEEE Pervasive Computing 2 (2), 56–61.