

Web Browser Control Using EMG Based Sub Vocal Speech Recognition

Chuck Jorgensen[†] & Kim Binsted^{††}

Abstract—Subvocal electromyogram (EMG) signal classification is used to control a modified web browser interface. Recorded surface signals from the larynx and sublingual areas below the jaw are filtered and transformed into features using a complex dual quad tree wavelet transform. Feature sets for six sub-vocally pronounced control words, 10 digits, 17 vowels and 23 consonants are trained using a scaled conjugate gradient neural network. The sub vocal signals are classified and used to initiate web browser queries through a matrix based alphabet coding scheme. Hyperlinks on web pages returned by the browser are numbered sequentially and queried using digits only. Classification methodology, accuracy, and feasibility for scale up to real world human machine interface tasks are discussed in the context of vowel and consonant recognition accuracy.

Index Terms—EMG, sub-vocal speech, wavelet, neural network, speech recognition, web browsing, vowels, consonants

I. INTRODUCTION

HUMAN to human or human to machine communication can occur in many ways [4]. Traditionally visual and verbal processes tend to dominate both the method and the presentation format. As a result, technology to enhance human communication has focused on public, audible tasks such as those addressed by commercial speech recognition. However, audible tasks place a number of constraints on situation suitability. These constraints include a vulnerability to ambient noise, requirements for clear formation and enunciation of words, and a shared language. When sound production limitations intervene, they can become very problematic. Examples of such situations might be suited HAZMAT operations, underwater or space EVA, crowded environments, high privacy requirements, or medical speech impairment. In many situations, very private communication is desirable, such as telephone calls, password entry, offline discussion while teleconferencing, military operations, or human /machine data queries. Vision based modalities, such

as email, can also cause problems because of non visual emotional information otherwise recognizable during speech. In addition, the intensity or forcefulness of the communication may be lost or misinterpreted. A communication alternative that can be private, non-dependent on physical production of audible signals, and still contain emotional subtleties of speech, could add valuable enrichment to the communication process.

An alternative way of communicating being considered at NASA Ames Research Center is the direct interpretation of nervous system control signals sent to speech muscles [9]. Specifically, we use non invasive aggregate surface measurements of electromyographic signals or EMGs to categorize muscle activation prior to sound generation [3]. Such signals arise when reading or speaking to oneself with or without actual lip or facial movements. Hence the information we are using does not show up using external observation, nor in current methods used to enhance speech recognition, such as machine lip reading.

In the present paper we demonstrate one EMG approach to the recognition of discrete, speaker dependent, non vocalized speech used to control a web browser. In previous work we showed the adequacy of EMG signals for the control of a virtual joystick and virtual numeric keypad entry [2]. In [1] we demonstrated recognition of a small sub acoustic control vocabulary.

The present control demonstration uses differential EMG signals measured on the side of the throat near the larynx and under the chin to pick up weak signals associated with aggregate muscle activity of the vocal tract and tongue. We capitalize on the fact that muscle activation leading to speech must remain relatively consistent and standardized to be understood by others. The concept is to intercept speech signals prior to sound generation and use them directly, bypassing auditory models such as mel cepstrums to filter signals.

After an appropriate feature transformation, EMG signals are input into a neural network or support vector machine classifier for recognition training and testing. Given sufficiently precise sensors, optimal feature selection, and a valid signal processing architecture, it is possible to use these extremely weak signals to perform usable tasks without vocalization and non-invasively. In a sense, we are approximating a totally silent control methodology such as that sought using EEG (i.e. thought based approaches [11]), but with much lower signal and measurement complexity.

This work was supported by NASA Ames Research Center under the CICT/ITSR program, Program manager Dr. Eugene Tu.[†]Dr. Chuck Jorgensen. is with the Computational Sciences Division, NASA Ames Research Center, Moffett Field CA 94035. (e-mail: cjorgensen@mail.arc.nasa.gov).^{††} Dr. Kim Binstead is with the University of Hawaii, Honolulu (email: binsted@hawaii.edu).

As alluded to above there are a number of specific situations better suited for using surface EMG measurement than standard auditory speech recognition or much more invasive medical alternatives. Among them are removal of health risks associated with sensor implantation, the requirement for detectable sound levels during communication, and potential content enrichment through additional physiological information.

To enable such technology, we also require sensors that are adequate to measure convolved EMG surface signals, signal processing algorithms that can transform signals into usable feature sets, and a trained neural network or other pattern classifier to learn and classify features in real time. Our earlier isolated word experiments demonstrated an average of

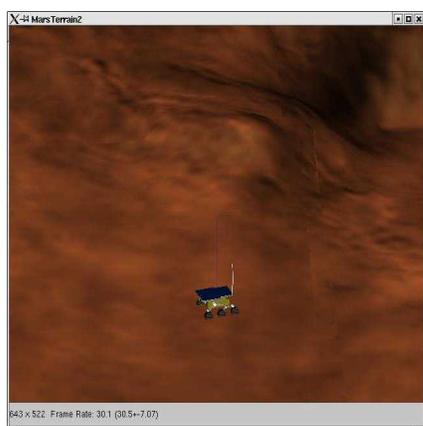


Figure 1: Mars Rover Simulation

92% accuracy in classifying six sub acoustic control words: stop, go, left, right, alpha, and omega. Later that year we demonstrated the classification of ten digits (0–9) at 73 percent accuracy. In this paper, we expand on our initial results, demonstrating the application of the technology to sub acoustic web browsing and control using both digits and control words simultaneously.

Next, we expand the recognition of individual words to include vowels and consonants, as a first step toward use in a generic, phoneme-based classical speech recognition architecture [5]. We finish with a discussion of issues yet to be resolved.

Little work testing the ability of EMG to perform speech recognition by itself appears to have been done. Parallel work for speech recognition augmentation along the lines of that in our word experiments was performed by Chan [8]. He proposed supplementing voiced speech with EMG in the context of aircraft pilot communication. In their work they studied the feasibility of augmenting auditory speech

information with EMG signals recorded from primary facial muscles using sensors imbedded in a pilot oxygen mask. He used five surface signal sites during vocalized pronunciation of the digits zero to nine using Ag-AgCl button electrodes and an additional acoustic channel to segment the signals. Their work demonstrated the potential of using information from multi-source aggregated surface to improve performance of a conventional speech recognition engine.

II. METHOD

A. Data Acquisition

Table 1: Phonemes and Training Words

PHONEMES AND TRAINING WORDS

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
ly	feel	hh	help
aw	foul	jh	joy
ae	gas	i	lid
ow	go	m	mat
jh	hit	n	no
axr	percent	p	put
eh	pet	r	red
ix	sick	sh	she
uw	tool	s	sit
oy	toy	t	talk
er	turn	dh	then
		th	thin
		v	vat
		w	with
		y	yacht
		z	zap

Five subjects, male and female ranging in ages from 18 to 55 were recorded while sub acoustically pronouncing six words: “stop”, “go”, “left”, “right”, “alpha”, and “omega”, as well as ten digits (0 through 9). In a separate experiment, two subjects, females aged 18 and 33, were recorded while sub acoustically pronouncing 42 phonemes (Table 1.)

The six words were chosen as a control set for a small graphic model of a Mars Rover (Fig. 1). which we use in our laboratory to test bioelectric control approaches [14]. Alpha and omega were used as generic control words to represent, for example, faster/slower or up/down, as appropriate for arbitrary tasks. The digits are used to permit precise numeric values for software menus and to allow more versatility in possible control tasks. In the web browsing task we used the digits in a numeric table to permit the coding of letters of the alphabet without resorting to a full alphabetic word set. Vowels and consonants were tested to evaluate whether our speech detection pattern recognizers could exceed the original

small set of speaker dependent words.

EMG signals were collected from each subject using two pairs of self-adhesive AG/AG-CL electrodes. They were located on the left and right anterior area of the throat approximately .25 cm back from the chin cleft and 1- 1/2 cm from the right and left side of the larynx (Fig. 2). Initial experimentation indicated that as few as one electrode pair located diagonally between the cleft of the chin and the larynx

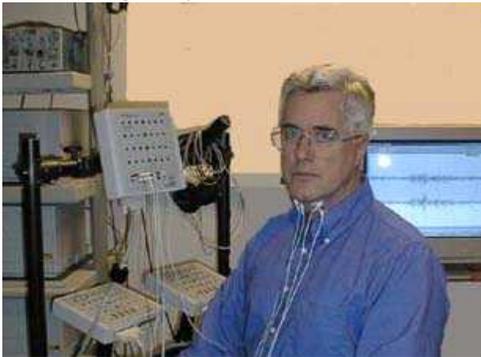


Figure 2: EMG electrode positioning

would suffice for small discrete word set recognition. Signal grounding required an additional electrode attached to the right wrist. Each electrode pair was connected to a commercial Neuroscan signal recorder, which recorded the EMG responses sampled at 2000 Hz, with a 60 Hz notch filter (to remove line interference), a 500 Hz low pass filter and a 30 Hz high pass filter.

Each subject recorded between one and two hundred exemplars of each digit and control word in morning and afternoon sessions. All words and digits were collected using

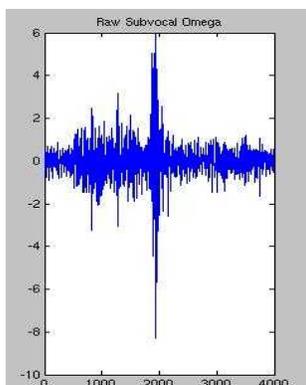


Figure 3: Subacoustic EMG signal for "omega"

a longitudinal experimental design over a period of two months including morning and afternoon sessions as well as minor variations in electrode placements. In the word and digit experiments, the signals were blocked offline into two second windows, and extraneous signals, e.g. swallows or coughs,

were removed using SCAN 4 Neuroscan software. Fig. 3. shows a blocked EMG signal for the word "omega".

Matlab scripts were written that provided a unified processing, from recording through network training. These routines were used to perform tasks such as transforming raw signals into feature sets, dynamic thresholding, compensation for changes in electrode position, adjusting signal/noise levels, and implementing algorithms used for recognition and training. The same routines were used both for the real time recording and the recognition experiments.

B. Feature Generation

Blocked signal segments for each word were transformed into feature vectors via signal processing transforms combined with a coefficient number reduction technique. The latter was required to reduce the number of produced coefficients to levels manageable for input into the neural network classifiers. Tests without such compaction were also performed using support vector machines on the entire coefficient. Five feature transforms were evaluated. They were:

- A windowed Short Time Fourier Transform (STFT),
- Discrete and Continuous Wavelets (DWT & CWT) using a Daubechie 5 and 7 base [6,7,12]
- Dual Tree Wavelets (DTWT) using a Near_sym_A 5,7 tap filter and a Q-shift 14,14 tap filter [10].
- Moving averages with lagged means, medians, and modes
- Linear Predictive Coding (LPC) coefficients

Features were created somewhat differently for each of the above transforms, depending on unique transform strengths or weaknesses. Each feature set produced varying degrees of efficacy in pattern discrimination. The most effective transforms for real time processing were a windowed STFT and Dual Tree Wavelet coefficients (DTWT), both of which were post-processed in a similar way to create feature vectors. The procedure used for these transforms was as follows.

Coefficient vectors were generated for each word using one of the two transforms. A rectified value of the raw signal was used in the case of the DTWT and non-rectified signal for the STFT. Vectors were post processed using the Matlab routines to create a matrix of the spectral coefficients. The matrix for each word example was in turn divided into a set of sub matrices. The number and size of the sub matrices depended upon spectral signal variance. Sub matrix sizes were chosen based on an average signal energy in a given region of the spectral matrix. Both equal and unequal size segmentation sub matrix schemes were considered. A single representative value for each submatrix (e.g. a mean) was then calculated to reduce the number of variables presented to the pattern recognition algorithm and capture average coefficient energy.

We chose a simple mean as the representative value because other obvious choices, including medians, modes or maximum sub matrix values, showed no improvement over a simple mean. The result was a vector of coefficient means for each sub acoustic word or vowel instance. The reasoning behind this approach was that a word or vowel could be treated as a noisy visual pattern recognition problem where a spectral energy matrix was a 2-D image and features were extracted from that image to discriminate among interesting parts of the ‘image’ patterns. DTWTs were selected rather than standard discrete wavelets to minimize typical wavelet sensitivity to phase shifts. Similarly, sensitivity to temporal shifting in the STFT was improved using windowing.

C. Feature Training

The above feature vectors were used to train a neural network or support vector machine pattern recognition engine. Words, digits, or vowel and consonant examples were split into three sets: a training set, a validation set, and a test set. Recognition performance was evaluated using 20 percent of the untrained word exemplars, and signals from a single electrode pair were randomly drawn from the data recording sessions. Five paradigms were evaluated as signal classifiers. Two that showed superior performance were:

- Scaled conjugate gradient nets
- Support Vector Machines.

A scaled conjugate gradient neural net was used for the following reasons. Standard Levenberg-Marquard gradient search reached the lowest mean square error levels but required too much system memory to handle the large data sets. This was true even using reduced memory variations. Lower mean squared error (MSE) did not translate into improved generalization for new signals probably due to the high sensor noise inherent in EMG surface signal measurements. A scaled conjugate gradient (SCG) network produced very fast convergence with adequate error and showed comparable performance to the full matrix inversion Levenberg-Marquardt (LM) implementation. This may be because the SCG also used the same trust region gradient criteria used by the LM algorithm. In earlier EMG experiments, we successfully used Hidden Markov Models (HMM) [2,9] but so far they proved most effective with non-multi-modal signal distributions, such as discrete gestures, rather than the temporally non-stationary sub-auditory signal patterns. HMM models required extensive pre-training to estimate transition probabilities. We anticipate further evaluation and have not ruled out HMM approaches, and may use a HMM/Neural net hybrid if warranted.

C. The Real Time Environment

To explore recognition performance under many signal

transform variations, we minimized the amount of on line human experiment time by creating a simulated real-time environment. This environment is part of a software system being developed at NASA Ames for large agency data understanding research. Using that environment, EMG signals were recorded to files and then later used to train and test recognition engines.

III. EXPERIMENTS AND RESULTS

A. Feature Transforms and Performance

Features for the signal sets were generated using Kingsbury’s Dual Tree Wavelets (DTWT) [10] and windowed Short Time Fourier Transforms (STFT). DTWT coefficients were coded into the previously reported (1) feature sub matrices of 5 rows of scale versus 10 columns of time segments. Each DTWT was based on a discrete wavelet transform defined as:

$$f(t) = \sum_{j,k} b_{j,k} \omega_{j,k}(t)$$

$$w_{j,k}(t) = 2^{j/2} w(2^j t - k)$$

Where k is the translation parameter, j is the dilation/compression parameter, and ω is the expansion function. We used a quarter sample shift orthogonal (Q-shift) filter having 10,10 taps with a near-symmetric-A filter having 5,7 taps. Kingsbury’s DTCW implementation of the Discrete Wavelet Transform (DWT) applies a dual tree of phase shifted filters to generate real and imaginary components of complex wavelet coefficients. The most important feature for our research was its improved shift invariance to the position of a signal in a signal window. The DTWT also showed good directional selectivity for diagonal features, limited redundancy independent of the number of scales, and efficient order-N computation, all of which are valuable for real time implementation.

Our results showed that the DTCW did increase shift invariance, lowering error over a standard DWT by several percentage points. Though discrete, the DTWT achieved comparable generalization performance to the slower continuous wavelet transform (CWT) with much lower computational load. It did this by doubling the sampling rate at each level of a short support complex FIR filter tree. Samples needed to be evenly spaced. In effect, two parallel, fully decimated trees are constructed so that the filters in one tree provide delays that are half a sample different from those in the other tree. In the linear phase this required odd length filters in one tree and even length filters in the other. The

impulse response of the filters then acted like the real and imaginary parts of a CWT, which is how Kingsbury uses them. Table 2 shows the achieved level of recognition for six control words using the DTWT.

Table 2: Percent Correct Word Classification

TRANSFORM	Dual Tree Wavelet
	2 level, near symmetric filter; q shift b; Trained with 125 epochs
“Stop”	84%
“Go”	100%
“Left”	91%
“Right”	80%
“Alpha”	97%
Average	92%

For the STFTs we used a standard implementation of the transform having a Hann window and a 50% time overlap to smooth the signal window. STFT coefficients were also tessellated into a 5 by 10 or 50 feature vector. Unequal windows based on variances were also considered but did not add to overall performance. We did take advantage of the computational efficiency of an STFT in the real time tests and still had fairly high recognition performance, though not as good as the DTWT, which had a recognition average of 92 percent. For more detail about the relative confusion between words and the viability of reducing feature order by principle component analysis and other learning algorithms considered see [12].

B. Digit Learning

In a similar fashion, ten digits (zero through nine) were added to the training sets to determine whether the classifier could separate a larger number of signals as effectively as the

control words. A support vector machine was used to compare several basis set assumptions. One hundred thirty four samples of each digit were trained. Forty unseen examples for each digit were tested in two ways. First, the DTWT coefficient sets were reduced into same 50 feature vectors as the for the STFT, above. The basis sets tested with these vectors consisted of polynomial bases, hyperbolic tangents, and simple (“raw”) means. Second, the original full set of 5000 complex DTWT coefficients were tested as a large single vector without coefficient size reduction to determine the impact of using average values over a large number of coefficients. For this set, we used radial basis and simple linear basis transforms. Results are presented in Table 3.

In summary, the best SVM performance was obtained with a radial basis, obtaining 73.13 percent discrimination for unseen word samples. Slightly poorer performance was obtained with the reduced coefficient set using a linear transform (73.12) percent. Overall, levels were roughly comparable across all transforms with slightly worse performance for a hyperbolic transform at 73, polynomial basis at 72.7 and raw mean at 72.5. Variance between the word averages was minimal for the radial basis hence it was deemed best for our purposes.

SVM	S	G	L	R	A	0	1	2	3	4	5	6	7	8	9	0
Linear	63	93	63	67	78	90	70	53	78	85	75	78	75	80	63	63
Tanh	63	93	63	65	78	90	70	50	78	85	75	75	75	80	65	65
Poly	63	93	63	65	78	90	70	55	78	85	75	75	75	80	56	63
RawRB	75	90	75	70	80	75	70	68	73	68	75	70	65	78	70	60
Radial	63	93	63	68	78	90	70	53	78	85	75	75	75	80	65	63

Table 3: Percentage of digits classified using SVM basis

C. Web Browsing

Using the digits and the control words it was possible to generate a useful test of the technology in a more applied

Sub Acoustic Web Browsing

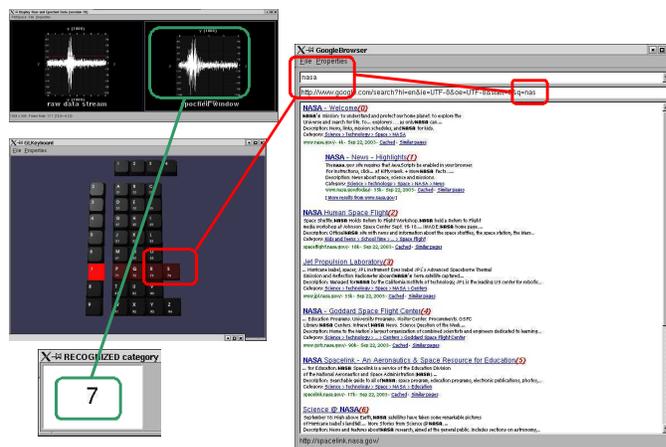


Figure 4: Sub-acoustic Web Browsing

context. We chose subvocal web browsing as a task that would require a relatively high recognition rate, processing speed, and feedback. Web browsing was performed as follows. A well known web browser's HTML output was modified so that each hyperlink on a web page was returned with a sequential number (Fig. 3). Because we did not attempt to train 26 alphabetic characters, we needed some way to use only digits and control words to maneuver around a page and send new search queries to a browser window. Our approach was to place the alphabet into a matrix look up table where each row and column index corresponded to a particular alphabetic character. So for example cell 1,1 would correspond to "a", 1,2 would correspond to "b" and so on. Code was written that sent the output of the signal classifier to both a graphic engine and the main entry line for the web browser.

Figure 4 shows which digit was recognized (the bright red key) and which alphabetic options would be potential candidates for the next key entry (dark red keys). At the top of the page is the blocked signal sent to the feature generator and the SVM classifier. On the very top of the web page in small print is the word being spelled out (in this case NASA) and the bottom of the browser page shows the returned result after a subacoustic command "GO" was executed. Notice the red numbers after each hyperlink. These can be queried using the numbers without again having to go to the awkward alphanumeric tabular coding scheme to enter data.

This demonstration accomplished several objectives. First it is possible to control an external task environment using sub acoustic speech. Second, it shows that the routines used to record, segment, and train sub acoustic samples can function in a simulated real time environment to query a web browser. At present, our interface is slow and awkward. Words and digits are sub acoustically pronounced with pauses between them to facilitate recognition. If the signal is improperly classified the command "OMEGA" is used to

negate it. The primary purpose was to stress-test the pattern recognition algorithms and improve real time EMG signal processing.

To extend this system to more complicated real time tasks or eventually full scale speech recognition, it will be necessary to demonstrate that signal recognition can occur for sub-acoustic elements smaller than words, such as phonemes. In the next section we present our first results towards that objective. We make no attempt to deal with contextual complexities such as diphthongs. The scalability of sub-acoustic EMG signals to very large sets of patterns has yet to be demonstrated but we did obtain results for reduced vowel and consonant sets.

D. Vowel and consonant recognition

Eighteen vowels and twenty-three consonants were trained in a similar fashion to that used for the words and digits. The vowels and consonants were collected over four days using female subjects only. To minimize variation in sub-vocal pronunciation, each subject sub-vocalized a phoneme while thinking of a 'target' reference word with the correct sound. For example (see Table 4) "dog" was used as the target word for the sub vocal muscle positioning of the *ao* vowel pronunciation. Similarly, consonant syllables used a target phoneme plus *ax* (i.e. *d* is sub vocalized as *dax*, or "duh").

To mark the signal time point at which a phoneme was produced, the subject touched a computer key, marking which signals corresponded to true subject events rather than noise events or other anomalies such as swallowing or coughing. Twelve sets of ten sub-vocal sounds were recorded by each subject for each phoneme and blocked into 1.5-second recording windows. Each signal channel was treated independently, potentially providing $12 \times 10 \times 2 = 240$ signals per phoneme per subject. Our first classification rate was low but significantly better than chance: an average overall rate of

Table 4: Consonant recognition

	big	cut	fork	genre	gut	help	lid	mat	no	put	red	she	then	thin	vat	with	yacht	TOTAL
big	28	5	14	0	0	0	7	12	0	30	0	0	0	0	0	5	0	101
cut	3	44	0	0	15	0	9	0	3	0	15	3	0	0	3	6	0	101
fork	2	4	44	0	2	0	2	2	4	4	4	4	0	0	25	5	0	102
genre	0	4	0	38	0	0	9	0	18	0	13	11	0	0	0	7	0	100
gut	0	18	0	4	41	4	20	0	2	0	0	2	6	0	0	2	2	101
help	0	0	8	0	3	56	10	0	5	3	8	5	0	0	3	0	0	101
lid	0	7	2	0	5	0	74	0	0	7	2	0	0	0	0	0	2	99
mat	13	0	0	0	0	0	0	61	0	6	6	0	2	0	11	2	0	101
no	0	3	3	0	3	0	5	0	81	0	0	0	0	0	0	5	0	100
put	14	5	5	0	2	2	9	2	5	44	7	0	0	2	2	0	0	99
red	0	2	2	2	0	2	2	2	2	0	66	0	0	9	2	7	0	98
she	0	4	0	13	0	0	4	0	4	0	29	36	0	2	2	4	0	98
then	3	9	0	0	0	0	0	0	0	6	9	3	49	23	0	0	0	102
thin	0	0	0	0	0	0	0	2	0	0	2	0	14	79	2	0	0	99
vat	5	0	19	2	0	2	0	2	0	9	2	2	0	0	53	2	0	98
with	0	2	0	2	0	0	6	6	2	2	27	0	0	0	6	46	0	99
yacht	0	14	0	18	0	0	16	2	4	6	27	0	4	4	0	0	4	99
TOTAL	68	121	97	79	71	66	173	91	130	117	217	66	75	119	109	91	8	

Table 5: Vowel recognition

	big	chin	cut	dig	fork	genre	gut	he lp	joy	lid	mat	no	put	red	she	sit	talk	then	thin	vat	with	ya cht	zap	TOTAL
big	39	0	0	4	9	0	4	0	0	7	11	0	11	2	0	0	0	0	4	7	2	0	0	100
chin	7	7	0	13	3	3	0	0	0	7	0	10	7	17	7	0	0	10	0	0	7	3	0	101
cut	0	0	43	0	0	0	22	0	0	16	0	2	0	6	0	0	6	4	0	0	0	0	2	101
dig	0	0	0	83	0	0	0	0	0	6	0	8	0	0	0	0	0	0	0	0	0	0	0	99
fork	16	0	0	9	37	0	4	0	0	9	0	2	5	5	0	0	0	0	0	12	2	0	0	101
genre	2	2	0	2	2	15	5	0	0	0	0	10	2	22	12	0	2	5	5	0	2	10	0	98
gut	0	0	14	0	0	2	52	0	0	11	0	0	0	11	0	0	0	2	0	2	5	0	0	99
help	0	0	6	18	0	0	22	24	0	14	0	6	2	8	0	0	0	0	0	0	0	0	0	100
joy	0	3	0	3	6	6	6	0	0	14	0	3	3	31	6	0	3	8	6	0	3	3	0	104
lid	0	0	8	0	0	0	10	0	0	75	0	3	0	3	0	0	0	3	0	0	0	0	0	102
mat	8	0	0	0	4	2	0	0	0	0	56	0	2	8	0	0	0	2	0	12	6	0	0	100
no	0	0	0	23	0	3	0	0	0	11	0	40	0	9	3	0	3	0	0	0	9	0	0	101
put	0	0	2	0	4	0	8	0	0	6	6	0	31	13	0	0	0	0	0	2	0	0	0	99
red	0	0	0	0	0	0	0	0	12	2	0	7	62	5	0	0	2	7	0	2	0	0	0	99
she	0	0	0	5	0	0	5	0	0	5	0	2	0	41	32	0	0	2	2	2	2	0	0	98
sit	0	4	2	20	2	0	4	0	0	0	0	29	0	16	18	0	0	2	0	2	0	0	0	99
talk	0	0	2	35	2	0	2	0	0	13	0	22	7	7	0	0	9	0	0	2	0	0	0	101
then	9	0	9	0	0	0	6	0	0	3	9	0	3	12	6	0	0	6	36	0	0	0	0	99
thin	0	0	0	0	0	0	3	0	0	3	5	0	0	13	0	0	8	69	0	0	0	0	0	101
vat	10	0	0	2	23	0	0	0	0	2	4	2	6	4	0	0	0	2	7	33	8	2	0	98
with	7	0	0	0	2	0	0	0	0	7	11	0	2	26	0	0	0	2	7	34	2	0	0	99
yacht	0	0	7	5	0	2	9	0	0	23	0	2	2	23	0	0	14	5	0	5	5	0	0	102
zap	2	0	0	40	0	2	7	0	0	10	0	14	0	14	2	0	0	2	2	0	0	2	0	97
TOTAL	127	16	93	262	94	35	169	24	0	254	104	155	90	352	91	0	17	68	146	81	89	25	6	

50%. This represented data training a scaled conjugate gradient network at 2500 iterations using one subject and the DTWT feature vector, with some phonemes removed (see below).

Previous results in the speech recognition literature [15] suggested that alveolars (where the tip of the tongue touches alveolar ridge) would be problematic for subvocal pronunciation. This did indeed seem to be the case based on the table of consonant confusions (Table 5.). As a result six alveolars (*t, d, s, z, ch, j*) were removed to obtain this categorization level. With the alveolars included, classification was at 33 percent (1500 iterations, dual tree wavelet transform, one subject). Removing the remaining alveolars (*n, l*, and maybe *r*) as well from the final tabulation would probably further improve the categorization results, as they seem to often be in the most poorly classified categories but for purposes of a realistic set we desired the vowels and consonants to remain as complete as possible at this stage.

Another problematic feature was voicing. Confusion pairs often differed only in the voicing feature. For example, *d* (voiced alveolar plosive) and *t* (voiceless alveolar plosive) had a high confusion rate.

Work on sub-vocal phoneme recognition is still at a preliminary stage. We are currently exploring different sensor positioning to detect the problematic features discussed above, as well as more sophisticated context-sensitive techniques borrowed from the far more advanced research area of vocal speech recognition.

IV. FUTURE DIRECTIONS

We are currently exploring enabling technologies to enhance EMG speech recognition and conduct more extensive experiments to increase task usability and vocabulary size. It is recognized that wet AG/AG-CL and dry electrodes are problematic for many real world tasks due to contact and surface resistance. To overcome that problem, new non-contact sensors (Fig. 4) are being developed. For

example, NASA Ames Research Center is working with Quantum Applied Science and Research (QUASAR) to develop electric potential free space sensors that do not require resistive, or even good capacitive coupling to a user. The sensor design provides a high input impedance for the electrode that measures free space potential, while accommodating the input bias current of the amplifier. At 10 Hz and above, the sensor has comparable sensitivity to conventional resistive contact electrodes. In the off-body mode the sensor can make an accurate measurement even through clothing. More detail about the sensors and our real time environment can be found in [13, 14].

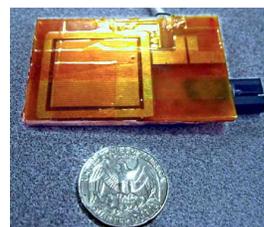


Figure 5: QUASAR non-capacitive sensor

V. CONCLUSION

We have demonstrated the potential of sub-acoustic speech to control a web browser, recognize a simple set of complete words and discriminate a subset of English vowels and consonants. The method has proven sufficiently accurate for applications that have limited vocabulary requirements. An open question is whether this approach can achieve full-scale sub-acoustic speech recognition resulting from the loss of discriminability of words dependent in English on plosive and tonal rather than muscle response. Whether the EMG signals are rich enough to disambiguate such cases and

handle the full richness of English speech has yet to be determined. Other languages with smaller basic vowel and consonant sets such as Japanese may prove viable sooner.

Significant challenges remain. We must generalize trained feature sets to other users in continuous speech situations, demonstrate real time training, optimize transformations and neural networks to reduce error levels, reduce sensitivity to signal noise and electrode locations, and handle changes in the physiological states of the users.

REFERENCES

- [1] Jorgensen, C., Lee, D., and Agabon, S. (2003). "Sub Auditory Speech Recognition Based on EMG/EPG Signals." Proceedings of the International Joint Conference on Neural Networks, Portland Oregon, July 2003.
- [2] C. Jorgensen, K. Wheeler, and S. Stepniewski "Bioelectric Control of a 757 Class High Fidelity Aircraft Simulation", Proceedings of the World Automation Congress, June 11-16, Wailea Maui, Hawaii, 2000.
- [3] K. Englehart, B. Hudgins, P.A. Parker, and M. Stevenson, "Classification of the Myoelectric Signal using Time-Frequency Based Representations," *Special Issue Medical Engineering and Physics on Intelligent Data Analysis in Electromyography and Electroneurography*, Summer 1999.
- [4] G. Ramsberger, "The human brain: Understanding the physical bases of intrapersonal communication," in *Intrapersonal communication: Different voices, different minds*, D.R. Vocate (Ed). (pp57-76) Erlbaum 1994.
- [5] A.R. Luria *Basic problems in neurolinguistics*. Mouton and Co. B.V., Publishers, The Hague, Paris 1976.
- [6] M. Krishnan, C. P. Neophytou, and G. Prescott, "Wavelet Transform Speech Recognition using vector quantization, dynamic time warping and artificial neural networks," Center for Excellence in Computer Aided Systems Engineering and Telecommunications & Information Science Laboratory.
- [7] B. T. Tan, M. Fu, P. Dermody, "The use of wavelet transforms in phoneme recognition," Dept. of Electrical and Computer Engineering, University of Newcastle, NSW Australia 1994.
- [8] A. Chan, K. Englehart, B. Hudgins, and D.F. Lovely, "Myoelectric Signals to Augment Speech Recognition," *Medical and Biological Engineering & Computing*, pp. 500-506 vol 39(4).
- [9] A.D.C. Chan, K. Englehart, B. Hudgins, D.F. Lovely, "Hidden Markov Model Classification of Myoelectric Signals in Speech," Proceedings of the 23rd Annual Conferences, IEEE/EMBS, Istanbul, Turkey. Oct. 2001.
- [10] N. Kingsbury, "The Dual-Tree Complex Wavelet Transform: a New Technique for Shift Invariance and Directional Filters" IEEE Digital Signal Processing Workshop, DSP 98, Bryce Canyon, paper no. 86. August 1998.
- [11] J. R. Wolpaw, N. Birbaumer, W.J. Heetdrechts, D. McFarland, P.H. Peckham, G. Schalk, E. Donchin, L.A. Quatrano, C. Robinson, and T.M. Vaughan, "Brain-computer interface technology: a review of the first international meeting", *IEEE Transactions on Rehabilitation Engineering* 8, 164-173. (2000).
- [12] K. Englehart, B. Hudgins, P.A. Parker, and M. Stevenson, "Improving Myoelectric Signal Classification Using Wavelet Packets and Principle Components Analysis", 21st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Atlanta, October 1999.
- [13] K. Wheeler, M. Allen, C. Currey, "Signal Processing Environment for Algorithmic Development," Unpublished, NASA Ames Research Center, Computational Sciences Division, 2003.
- [14] L. Trejo, K. Wheeler, C. Jorgensen, R. Rosipal, "Multimodal NeuroElectric Interface Development", submitted to *IEEE transactions on neural systems and rehabilitation engineering: Special issue on BCI 2002*.
- [15] Huang, X., Acero, A., and Hon, H-W. (2001). *Spoken Language Processing: A guide to theory, algorithm and system development*. New Jersey, Prentice-Hall PTR.

