Sub Auditory Speech Recognition based on EMG/EPG Signals

Chuck Jorgensen[1], Diana D. Lee[2], and Shane Agabon[3]

Abstract—**Sub-vocal electromyogram/electro palatogram (EMG/EPG) signal classification is demonstrated as a method for silent speech recognition. Recorded electrode signals from the larynx and sublingual areas below the jaw are noise filtered and transformed into features using complex dual quad tree wavelet transforms. Feature sets for six sub-vocally pronounced words are trained using a trust region scaled conjugate gradient neural network. Real time signals for previously unseen patterns are classified into categories suitable for primitive control of graphic objects. Feature construction, recognition accuracy and an approach for extension of the technique to a variety of real world application areas are presented.**

Index Terms—**EMG, Sub Acoustic Speech, Wavelet, Neural Network**

I. Introduction

COMMUNICATION between humans or humans and their machines occurs in many ways. Traditionally visual and verbal information exchange tends to dominate. As a result, efforts at automating human or human to machine communication such as commercial speech recognition, have emphasized the public audible aspects. However, a totally auditory communication strategy places a number of constraints on the communication channels. These constraints include sensitivity to ambient noise, a requirement for proper formation and ennunciation of words, and a shared language. The physical limitations of sound production also become problematic in environments such as HAZMAT, EVA space tasks, or underwater operations. Furthermore, auditory expression may be undesirable for private communication needed in many daily situations such as discrete telephone calls, offline comments during teleconferencing, military operations, or human to machine commands and queries. Communication alternatives that are both private and non-dependant on production of audible signals are valuable.

One proposed method is the direct readout of brain signals. This approach bypasses speech production altogether. Wolpaw et al. [9] recently published a review of the state of the art in electroencephalograph (EEG) understanding. We too are pursuing EEG approaches in our lab [12]. However there are a number of practical difficulties for nearer term application of such EEG approaches due largely to their use of aggregated surface measured brain potentials, their inherent non-linear complexity, and their idiosyncratic nature. The alternative, invasive EEG measurement, is not considered by us as practical for widespread use.

Consequently we are exploring surface measurement of only muscle signals (i.e. electromyographic or EMG) to disambiguate speech signals produced with minimal or no acoustic output. In the present paper we demonstrate one approach to the recognition of discrete task control words. Our approach uses EMG [1] measured on the side of the throat near the larynx and under the chin to pick up surface tongue signals (i.e. electropolatogram or EPG). The approach capitalizes on the fact that vocal speech muscle control signals must be highly repeatable to be understood by others. The central idea is to intercept these signals prior to actual sound generation and use them directly. These are then fed into a neural network pattern classifier. What is analyzed is silent or sub auditory speech like when a person silently reads or talks to themselves. [2][3]. In our approach, the tongue and throat muscles still respond slightly as though a word was to be made audible albeit very faintly and with little if any external movement cues presented. Given sufficiently precise sensing, optimal feature selection, and good signal processing techniques, it is possible to use these weak signals to perform usable tasks without vocalization yet mimic an ideal of thought based approaches.

There are a number of advantages to taking this approach over invasive alternatives. Among them are minimization of word variations because there is a shared language and sound production requirement, potential to connect signal recognition to highly developed speech recognition engines, non invasive sensing, reasonable robustness to physiological variations, and privacy.

The enabling technologies required are sensors adequate to measure the EMG signals, signal processing algorithms to transform the signals into usable feature sets, and a trained neural network or other pattern classifier to learn and classify signal feature sets in real time. Our initial results have demonstrated an average of 92% accuracy in discriminating six untrained sub acoustic words (stop, go, left, right, alpha, omega) in a simulated real time environment under a wide variety of electrode placement and recording times. In further experiments we increased the number of words and the nature of the sub acoustic features to set the stage for more powerful applications.

Manuscript received January 25, 2003. This work was supported by NASA Ames Research Center under the CICT/ITSR program. [1]Dr. Chuck. Jorgensen. is with the Computational Sciences Division, NASA Ames Research Center, Moffett Field CA 94035. (e-mail: cjorgensen@mail.arc.nasa.gov). [2]Diana. Lee. is with SAIC Corporation NASA Ames Research Center (e-mail: ddlee@mail.arc.nasa.gov).[3]Shane. Agabon is with QSS Corporation NASA Ames Research Center e-mail: sagabon@mail.arc.nasa.gov

We begin this paper by describing our generic method; next we describe our experiments and results. We end with descriptions of some related work, future directions, and a discussion of implementation issues yet to be resolved.

II. METHOD

A. Data Acquisition

Three subjects aged 55, 35, and 24 were recorded while sub auditorially pronouncing six English words: stop, go, left, right, alpha, and omega. These particular six words were selected in order to form a control set for a small graphic model of a Mars Rover. Alpha, and omega were chosen as general control words to represent faster/slower or up/down as appropriate for the particular simulated task.

EMG and EPG signal data was collected for each of the subjects 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 larynx (Figure 1). Initial results indicated that as few as one electrode pair located diagonally between the cleft of the chin and the larynx would suffice for small sets of discrete word recognition. Signal grounding required an additional electrode attached to the right wrist. When acquiring data using the wet electrodes, each electrode pair was connected to a commercial Neuroscan signal recorder which recorded the EMG responses sampled at 2000 Hz. A 60 hertz notch filter was used to remove ambient interference.

Fig 1: Electrode placement and recording

One hundred exemplars of each word were recorded for each subject over 6 days in morning and afternoon sessions. In the first experiments, the signals were blocked offline into 2 second windows, and extraneous signals, e.g. swallows or coughs, were removed using SCAN 4 Neuroscan software. Fig. 2 shows two typical EMG blocked signals for the words left and omega.

For signal feature processing, Matlab scripts were developed that created a unified signal processing system from recording through network training. These routines were used to perform tasks such as transform the raw signals into feature sets, dynamically threshold them, compensate for changes in electrode position, adjust signal/noise levels, and implement neural network algorithms for pattern recognition and training. EMG/EPG artifacts such as swallowing, muscle fatigue tremors, or coughs were removed during preprocessing of the block files.

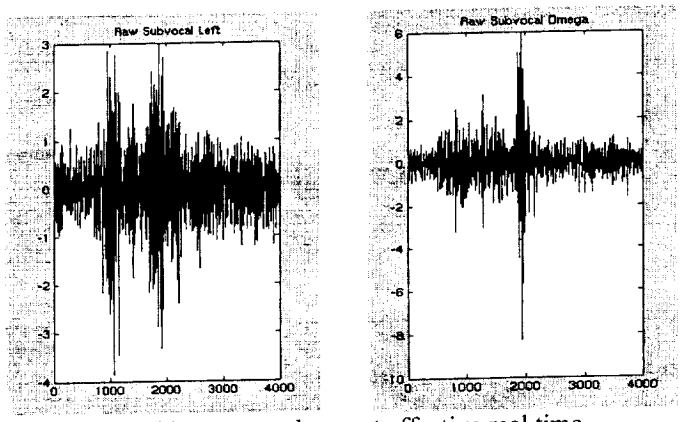
B. Feature Generation

Blocked signal data for each word was transformed into usable classifier feature vectors by preprocessing transforms combined with a coefficient reduction technique. The transforms tested were:

- A windowed Short Time Fourier Transform (STFT),
- Discrete and continuous Wavelets (DWT & CWT) with Daubechies 5 and 7 bases
- Moving averages with lagged means, medians, and modes
- Hartley Transforms
- Hilbert-Huang Transforms
- Linear Predictive Coding (LPC) Coefficients
- and Dual Tree Wavelets (DTWT) using a near_sym_a 5,7 tap filter and a Q-shift 14,14 tap filter [8].

Feature sets were created somewhat differently for each of the above transforms depending on their unique signal processing advantages and disadvantages. Each feature set produced varying degrees of efficacy in pattern discrimination. Because of space limitations we confine ourselves in this paper to the most effective real time transforms i.e. windowed FFT vector and Dual Tree Wavelets coefficient matrices, both of which were post-processed in a similar way to create their feature vectors. The procedure used for the two pre transforms was as follows.

Fig 2: Sub Acoustic signals for "left" and Omega



Transform coefficient vectors were generated for each word using one of the latter two transforms on the absolute value of the raw signal. (This was because the electrodes were bi polar and hence directional sign information had no significance). Vectors were post processed using the Matlab

routines to create a matrix of spectral coefficients. This matrix was tessellated into a set of sub matrices. The number and size of the sub matrices depended upon the spectral signal information complexity. Tessellation sizes were determined based on average signal energy in a given region of the spectral matrix. Both equal and unequal segmentation size schemes were considered. A single representative value was calculated for each sub matrix to reduce the number of variables presented to the pattern recognition algorithm and represent average coefficient energy.

We chose to use a simple mean as the representative value because other choices including medians, modes or maximum sub matrix values showed no improvement. The result was a vector of coefficient means for each sub acoustic word instance. The reasoning behind this approach was that each word could be treated as a noisy visual pattern recognition problem where the spectral energy matrix became a 2-D image and features were extracted so as to discriminate among interesting features in the 'image' patterns. Dual tree wavelets were selected rather than standard discrete wavelets to minimize the normal wavelet sensitivity to phase shift. Similarly, sensitivity to temporal non stationarity in the FFT was improved using windowing. Continuous wavelets were not considered practical for real time computational reasons. The Hartley transform was explored for a potential benefit of combining both real and imaginary signal components over real components alone.

C. *Feature Training*

These feature vectors were used to train the neural network recognition engine. Word signals were split into three sets, a training set, validation set, and test set. Generally, recognition was evaluated using 20 percent of the untrained word exemplars and signals from only one electrode pair randomly drawn from the data recording sessions. Five neural network paradigms were considered for signal classifiers. Those tested were:

- Scaled Conjugate Gradient nets
- Leavenburg-Marquardt nets,
- Probabilistic Neural Nets,
- Modified Dynamic Cell Structure Nets (DCS) [13]
- and Linear Classifiers.

After comparison a scaled conjugate gradient net was chosen for the following reasons. Leavenberg-Marquard reached the lowest mean square error levels but required too much system memory for large data sets. This was true even using reduced memory variations. A low mean squared error (MSE) did not translate into improved generalization for new signals due to high sensor noise. Probabilistic neural nets produced reasonable classifications but required very large training sample sizes to reach stable probabilities and were not superior in their ultimate pattern discrimination ability. The DCS net had very fast training which made it good for real time adaptation but tended to be less compact for our anticipated applications that are memory sensitive. The scaled conjugate gradient network had fast convergence with adequate error levels for the signal to noise ratio in the data and showed comparable performance to the Levenberg-Marquardt network. This may possibly be because it also took advantage of a trust region gradient search criteria. In other EMG tasks we successfully applied Hidden Markov Models (14) but so far they were most effective with non multi modal signal distributions such as with discrete gestures rather than the present temporally non stationary sub auditory signal patterns. They also require extensive pre training to estimate transition probabilities. We anticipate further evaluations and have not ruled out HMM models, and may use a HMM/Neural net hybrid if warranted.

D. *Human Learning and the Real Time Environment*

To quickly explore many experiments on recognition under different transform variations, we minimized the amount of on line human learning by operating in a simulated real-time environment. This environment is part of a system being developed at Ames for large NASA data understanding problems. Within the environment, EMG signals were recorded to file and then later used to train and test the recognition engines. Our three subjects were not given immediate feedback about how well their sub vocal pronunciations were recognized, however there was still a small amount of learning that took place as the subjects were permitted to view their EMG signals after the experiments and between trials. Nonetheless there were no indications that pronunciation patterns changed significantly over time.

III. Experiments and Results

A. *Feature Generation*

Four of the feature transforms had sufficient merit to warrant further experimentation and were evaluated in depth for generalization and learning performance. They were:

1. Discrete and Dual Tree Wavelets (Discrete at 5x5, 4x10, and 8x10 spectral matrix tessellations and Dual at 5x10) which produced 92% word recognition accuracy. The DWT was 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 and j the dilation/compression parameter, ω is the expansion function. In our case these were Daubechies filters.

2. STFT tessellated to 5x10 or 50 features that produced 91% word recognition accuracy. The

Fourier transform defined as:

$$x(k+1) = \sum_{n=0}^{N-1} x(n+1) W_n^{kn}$$

$$W_n = e^{-j(2\Pi/N)}$$

$$N = length(x)$$

3. Hartley Transform tessellated to 5x10 defined as:

   $$real(FFTcoef) - imag(FFTcoef)$$

   which showed 90% recognition accuracy

4. And Moving Averages at 200, 100, and 50 time steps which produced 83% word recognition accuracy.

Table 1: "Percentile of Correct Word Classification," presents the recognition engine's accuracy for each transform in classifying unseen data. By unseen, we mean feature vectors that were not used during neural network training. As shown, the best performing pre transform was Kingsbury's Dual Tree Complex Wavelet (DTCW) [8]. 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 applies a dual tree of phase shifted filters to produce real and imaginary components of complex wavelet coefficients. One of its valuable properties for this research is its improved shift invariance to the position of a signal in the signal window. Other desirable features are better directional selectivity for diagonal features, limited redundancy independent of the number of scales, and efficient order-N computation. In our experiments, the DTCW increased shift invariance over the DWT by several percentage points. Real time implementation of a CWT was not practical from a computation and time perspective. However, the dual tree wavelet achieves comparable generalization performance to the CWT by doubling the sampling rate at each level of a short support complex FIR filter tree. The samples must be evenly spaced. In effect two parallel fully decimated trees are constructed so that the filters in one tree provides delays that are half a sample different from those in the other tree. In the linear phase this requires odd length filters in one tree and even length filters in the other. The impulse response of the filters then looks like the real and imaginary parts of a complex wavelet. This is how Kingsbury uses them.

For our STFT's we used a standard implementation having a Hann window and 50% time overlap to smooth the signal window. Unequal windows based on variances were also considered but did not add to overall performance. However we could take advantage of the computation efficiency of an STFT and still have fairly high recognition performance though not as good as the DTWT. Baysian regularization networks were tested but on initial results they also did not increase the level of recognition. .

Earlier EMG research [10] indicated there might be an advantage in preprocessing wavelet packet features using Principal Component Analysis (PCA). PCA was tested as a preprocessing method for the DWT EMG signal vectors. Even using fairly large numbers of PCA components, generalization again proved poorer than without the transform. Hence this step was omitted in our final procedure. We attribute this possibly to high signal pattern variation caused by phase shifting within the feature windows.

During the experiments, it became apparent some signals were not being well recognized by the neural net. In a real time system with a capability to interactively request speaker clarification, it is desirable to have a way to detect and respond to such marginal signals. Individual word recognition rates can help indicate which sub acoustic words are more easily discriminated. For example, 'go' and 'omega' consistently scored recognition rates of 90% or better. Tables 2 and 3 give the Confusion Matrices for the six words and indicate which words were confused for one another. For example, in Table 2, we see the word 'stop' correctly classified 21 times, but mistakenly classified as the word 'right' 4 times. Overall the confusion rates were not high.

TABLE 1 PERCENTILE OF CORRECT WORD CLASSIFICATION

Transform	*Average Recognition Rate*	*Individual Word Recognitions*					
		Stop	*Go*	*Left*	*Right*	*Alpha*	*Omega*
Dual Tree Wavelet 2 level, near symmetric filter; q shift b; trained with 125 epochs	92%	84%	100%	91%	80%	97%	97%
Fourier Hann windows overlapped 50% Trained with 200 epochs	91%	83%	100%	91%	89%	82%	98%
Hartley Hann windows overlapped 50% Trained with 250 epochs	90%	79%	97%	91%	91%	79%	100%
Moving Averages Square windows overlapped 50%; Trained with 125 epochs	83%	62%	90%	84%	91%	73%	95%

TABLE 2: CONFUSION MATRIX FOR THE DUAL TREE WAVELET TRANSFORM

Dual Tree Wavelet	Stop	Go	Left	Right	Alpha	Omega
Stop	21	0	0	4	0	0
Go	0	37	0	0	0	0
Left	0	0	32	3	0	0
Right	1	1	3	20	0	0
Alpha	0	0	0	0	37	1
Omega	0	0	0	0	1	35

TABLE 3: CONFUSION MATRIX FOR THE FOURIER TRANSFORM

Fourier	Stop	Go	Left	Right	Alpha	Omega
Stop	24	2	0	0	0	3
Go	0	31	0	0	0	0
Left	0	0	29	2	1	0
Right	1	0	2	31	1	0
Alpha	1	0	0	4	27	1
Omega	0	1	0	0	0	39

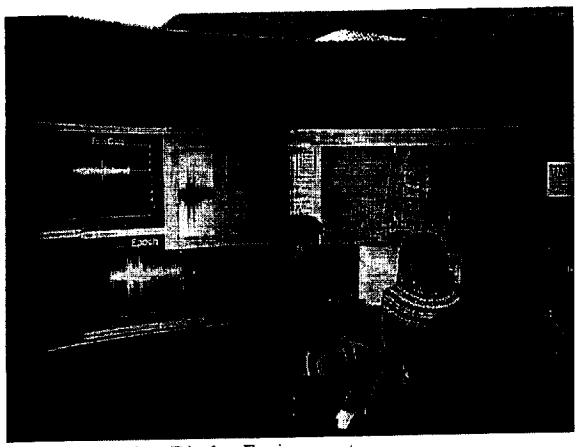


Fig. 2 Real Time Display Environment

IV. RELATED WORK

Little work testing the ability of EMG by itself to perform speech recognition appears to have been done. Parallel work for speech recognition augmentation along the lines of that in our set of experiments was performed by Chan [6]. He proposed supplementing voiced speech with EMG in the context of aircraft pilot communication. In their work they studied the feasibility of supporting auditory speech information with EMG signals recorded from primary facial muscles using sensors imbedded in a pilot oxygen mask. Five surface signal sites were recorded 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 measured EMG signals to enhance the performance of a conventional speech recognition engine.

V. FUTURE DIRECTIONS

We are currently exploring a number of other enabling technologies for enhanced EMG speech recognition and conducting further experiments to increase general task usability and vocabulary size. The technologies include the capacitive non-contact sensors in wearable arrays and a real-time system environment (figure 2). It is recognized that wet AG/AG-Cl electrodes are problematic for many real world tasks due to contact and surface resistance, hence dry electrodes and new non-contact sensors are being tested as well. For example, NASA Ames Research Center is working with Quantum Applied Science and Research, Inc. (QUASAR) to develop electric potential free space sensors that do not require resistive, or even good capacitive coupling to the user. The sensor design provides a high input

impedance for the electrode that measures the free space potential, while accommodating the input bias current of the amplifier. At 10 Hz and above, the new 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 this research is presented in [10].

New experiments are studying the feasibility of an expanded vocabulary, ideally, one composed of the basic speech components including vowels, consonants, and other phonetic building blocks. Trying to detect these building blocks poses an interesting problem, since many of the auditory features that generate vocal speech such as aspiration, glottal stops, or tonality may have no direct EMG analog. However, the EMG signal is very rich, and this richness may actually provide more useful cues for speech recognition because they are so linked to the speech encoding process cognitively. Initial experiments suggest that extension to a larger 20-word control vocabulary is reasonable. We will continue to grow the vocabulary with sets of English phonemes usable by a full speech recognition engine. If full speech recognition proves unfeasible, we can still demonstrate useful specialized tasks with specialized vocabularies such as machine control or cell phone dialing.

Currently we use a simulated real-time environment where sub acoustic signals are recorded to files. These files are later used as input to the feature generator and classifier. We are implementing fast compact signal processing software to enable real time processing now that is undergoing its first tests on a digit based vocabulary.

VI. Conclusion

We have described a system that demonstrates the potential of sub acoustic speech recognition based on EMG signals. It is able to measure and easily classify six sub acoustic words with up to 92% accuracy using only one pair of surface electrodes. The enabling technologies were surface sensors used to measure the EMG signals, signal processing used to transform the signals into feature sets, and neural networks to learn and provide sufficient robustness for the nonlinear and non-stationary nature of the underlying signal data.

The method has proven to be sufficient for applications that require discrete word, subject specific, limited control vocabularies. An open question is whether this system can achieve full scale sub acoustic speech recognition and achieve the goal of EEG based human thought interfaces using EMG signals alone.

Significant challenges remain. We must generalize trained feature sets to other users, show real time training and user startup, optimize transformations and neural networks, reduce sensitivity to noise and electrode locations, and handle changes in physiological states of the users.

References

[1] 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.

[2] G. Ramsberger, "The human brain: Understanding the physical bases of intrapersonal communication," in *Intrapersonal communciation:Different voices, different minds*, D.R Vocate (Ed). (pp57-76) Erlbaum 1994.

[3] A.R. Luria *Basic problems in neurolin*guistics. Mouton and Co. B.V., Publishers, The Hague, Paris 1976.

[4] 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.

[5] B. T. Tan, M. Fu, P. Dermody "The use of wavelet transforms in phoneme recognition," Dept. of Electrical and Computer Engineering, Univeristy of Newcastel, NSW Australia 1994.

[6] 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)*.

[7] 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.

[8] 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.

[9] J. R. Wolpaw, N. Birbaumer, W.J. Heetdrecks, D. McFarland, P.H. Peckham, G. Schalk, E. Donchin, L.A. Quatrano, C. Robinson, and T.M Vaughan, T. M. Brain-computer interface technology: a review of the first international meeting. IEEE Transactions on Rehabilitation Engineering 8, 164-173. (2000).

[10] 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.

[11] K Wheeler, M Allen, C. Currey, "Signal Processing Environment for Algorithmic Development," Unpublished, NASA Ames Research Center, Computational Sciences Division, 2003.

[12] 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.

[13] 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.