SENSOR SUBSET SELECTION FOR SURFACE ELECTROMYOGRAPY BASED SPEECH RECOGNITION

Glen Colby[1] [*], *James T. Heaton*[2], *L. Donald Gilmore*[3], *Jason Sroka*[1], *Yunbin Deng*[1], *Joao Cabrera*[1], *Serge Roy*[3], *Carlo J. De Luca*[3], *Geoffrey S. Meltzner*[1]

[1]BAE Systems – Advanced Information Technologies, Burlington, MA, USA
[2]Center for Laryngeal Surgery & Voice Rehabilitation, Mass. General Hospital, Boston, MA, USA
[3]Altec, Inc., Boston, MA, USA

ABSTRACT

The authors previously reported speaker-dependent automatic speech recognition accuracy for isolated words using eleven surface-electromyographic (sEMG) sensors in fixed recording locations on the face and neck. The original array of sensors was chosen to ensure ample coverage of the muscle groups known to contribute to articulation during speech production. In this paper we systematically analyzed speech recognition performance from sensor subsets with the goal of reducing the number of sensors needed and finding the best combination of sensor locations to achieve word recognition rates comparable to the full set. We evaluated each of the different possible subsets by its mean word recognition rate across nine speakers using HMM modeling of MFCC and co-activation features derived from the subset of sensor signals. We show empirically that five sensors are sufficient to achieve a recognition rate to within a half a percentage point of that obtainable from the full set of sensors.

Index Terms— Electromyography, speech recognition

1. INTRODUCTION

Previously reported results [1] demonstrated the ability to conduct accurate automatic speech recognition on vocalized and mouthed speech using signals collected by a set of eleven surface-electromyographic (sEMG) sensors placed at standardized locations on the face and neck. The original array of sensors was chosen to ensure ample coverage of the muscle groups known to contribute to articulation and correlated with voice production during speech production. In this follow-on study, the effect of the number of sEMG sensors on subvocal speaker-dependent speech recognition accuracy was analyzed with the goal of identifying a smaller subset of sensors that maintained recognition accuracy close to that of the full sensor set on a 65-word vocabulary for 9 different speakers.

Our approach was first to evaluate the best speech recognition rates achievable at each subset size averaged across all 9 speakers. Given those results, we then selected the smallest cardinality that afforded a maximal average recognition close to that of the full eleven-sensor set. After the preferred number of sensors was selected, we then identified several subsets of that cardinality that provide sufficient sEMG articulation information for achieving speech-recognition rates to within one percentage point of the original full set of sensors.

2. METHODS

2.1 Data Collection

Data consisted of eleven channels of sEMG signals measuring the activity of superficial muscles in face, neck and under the chin (see Figure 1) during the production of a 65-word English vocabulary (see Table 1), repeated six times by nine different speakers, 4 female and 5 male.

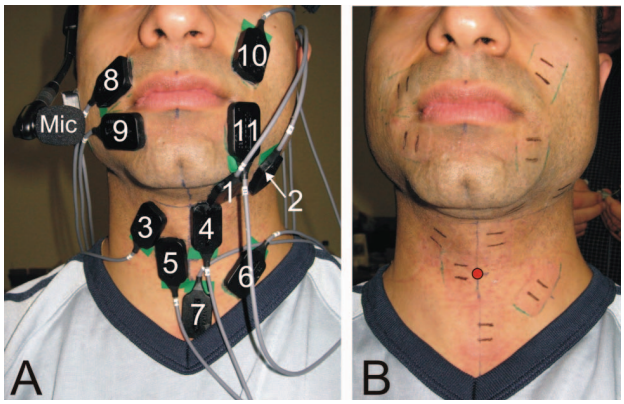

Figure 1. The sEMG sensor locations shown (A) before and (B) after removal. Black lines in (B) show electrode contacts and the red dot (neck midline) marks the cricothyroid gap.

The ordering of the words was randomized separately among the different speakers, and each word was repeated

[*] DARPA Release Information: Approved for Public Release, Distribution Unlimited

three times before continuing on to the next word. Two complete rounds of data for each speech modality—vocalized and mouthed—were collected resulting in six repetitions of each speech token per speech mode per speaker. The original signals were recorded digitally using a 20 kHz sampling rate.

no	fire	that	right	advance	monitor	stand-by
go	five	this	seven	forward	continue	shut-down
one	four	zero	target	latitude	negative	accelerate
six	help	abort	three	location	hibernate	maneuver
ten	days	block	assist	measure	position	affirmative
two	nine	eight	brings	recover	proceed	coordinates
yes	pull	hours	cancel	reverse	transmit	kilometers
left	push	miles	meters	seconds	thousand	rendezvous
fast	slow	point	collect	hundred	longitude	
feet	stop	reach				

Table 1. Vocabulary of 65 words used in the sEMG-based word recognition experiments.

2.2 sEMG-based Speech Recognition

Automatic speech recognition was conducted on the data by modeling the 65 word tokens represented by 11-channel sEMG signals on a speaker-dependent basis using hidden Markov models. The experiments implemented a cross-validation method in which four utterances (Utterances 1, 3, 4, and 6) of each word per speaker were used to train each speaker-dependent word model, and the remaining two utterances (Utterances 2 and 5) of each word per speaker were held out for testing. Segmentation and labeling of the word utterances from the surrounding non-speech intervals of the original signals were performed using a function of the onset/offset of sEMG signal activity above/below an estimated noise threshold on selected channels.

In preparation for modeling and recognition, the digitized signals were first downsampled from 20 kHz to 5 kHz, and then the DC-offset was removed from each channel. The data were then broken into overlapping multi-channel frames using a window size of 50 ms and a frame period of 25 ms. A feature vector, consisting of 9 Mel-frequency cepstral coefficients (MFCCs), the 0th coefficient and deltas together with pairwise channel co-activation information [2], was then extracted from each frame. Eleven channels of such features resulted in 275 real-valued features per frame: 11×20 MFCC features plus 11-choose-2 co-activation features. Signal processing, feature extraction and conversion into HTK format were implemented using Matlab and the Voicebox Speech-Processing Toolbox for Matlab [3].

Supervised HMM modeling and recognition were performed using the HTK Speech Recognition Toolkit [4]. For each of the nine speakers, a separate set of 65 speaker-dependent word models were generated from training data. The HMM topology consisted of 10 states, of which only 8 states were emitting. The transition matrix was defined such that the model either stayed in the same state or advanced one state between consecutive frames. Only one Gaussian per feature per word was generated, which was initialized with a mean of 0 and variance of 1, and a fixed variance floor was set to 0.01. Training began by first generating initial word model estimates using segmental k-means on the labeled bootstrap data and then re-estimating the models using an isolated-word implementation of the Baum-Welch algorithm. The models were further re-estimated three consecutive times using the Baum-Welch algorithm across all words in the vocabulary simultaneously. Speaker-dependent recognition was performed using the Viterbi algorithm.

2.3 Recognition Performance using Full Sensor Set

We selected as our evaluation metric for comparing subsets the mean recognition rate across all nine speakers using the full HMM training and recognition procedure, Baseline recognition accuracy was computed for both speech modalities, mouthed and vocalized, using the full array of sensors, the results of which are presented in Table 2. The mean speaker-dependent mouthed speech recognition rate across all nine speakers is 85.4% (range of 69.2% - 93.1%).

Speaker		Recognition Accuracy	
Number	Gender	Mouthed	Vocalized
1	F	86.9%	88.5%
2	F	83.1%	93.1%
3	F	90.0%	86.9%
4	M	77.7%	91.5%
5	M	93.1%	96.2%
6	M	90.8%	97.7%
7	M	89.2%	96.9%
8	F	88.5%	84.6%
9	M	69.2%	90.8%
Mean	**-**	**85.4%**	**91.8%**

Table 2. Baseline recognition accuracy for both speech modalities using all 11 sensors.

2.4 Analysis of Subset Size

We began the subset analysis by first evaluating the maximal speaker-dependent recognition rate, averaged across all speakers, achievable at each cardinality. Since alternative search methods may not guarantee finding *the* maximal scoring subset at each cardinality, we conducted an exhaustive evaluation on all 2047 possible subsets using the mouthed-word data. There are a total of $2047 = 2^{11} - 1$ possible subsets from the full array of 11 sensors, of which there are 11-choose-k subsets of size k, for $1 = k = 11$. For the empty subset $k = 0$, we assume a recognition rate equal to uniform random distribution; i.e., 1/65 = 1.5% recognition rate.

To implement such a large scale set of experiments efficiently, we replaced our computationally expensive signal-processing and feature-extraction procedures from the raw segmented signals with simple matrix

multiplications. We pre-computed 2047 binary transformation matrices, one for each subset, algebraically designed to transform the feature vector obtained from using all eleven sensors to the feature vector for the given subset of sensors. To generate the

$$3510 = 9\frac{\text{speakers}}{\text{subset}} \times 65\frac{\text{words}}{\text{speaker}} \times 6\frac{\text{utterances}}{\text{word}}$$

new feature vectors for each subset, one need only read in the appropriate binary transformation matrix and multiply the 3510 feature vectors from the full 11-sensor set by the matrix.

Plotting the resulting mean recognition rates of each possible subset versus its subset cardinality (see Figure 2) reveals relevant characteristics to our study. First, the maximal mean recognition rate at each cardinality increases rapidly by using just a small number of sensors and eventually plateaus at approximately 85% for five sensors, with no significant increase (less than a total of 1.75 percentage points) in recognition capabilities from adding additional sensors. The best scoring 5-sensor subset achieved a mean recognition rate of 84.9%, just a half a percentage point below that of the full 11-sensor set (85.4%). Second, the maximal mean recognition rates of subsets having between six and ten sensors were slightly higher than that of using all eleven sensors. This effect may possibly be attributable to overfitting of the larger feature vectors as the cardinality grows. Third, the fewer the number of sensors, the greater the variance of the recognition results for a particular cardinality.

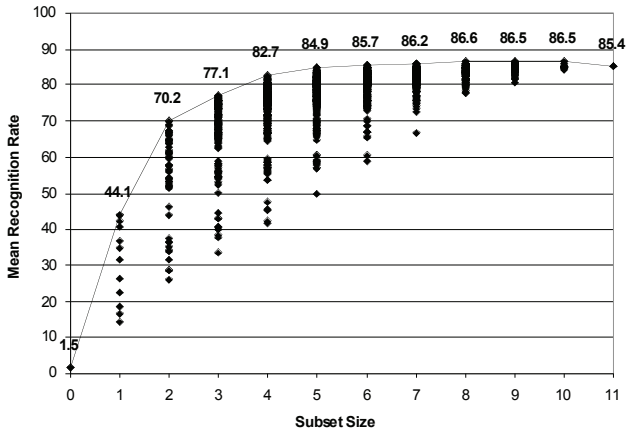

Figure 2. Mean recognition rates for each sensor subset on mouthed words. Each data point represents recognition for a different subset averaged across all nine speakers. A line spans the peak recognition values.

Our findings regarding the effect of sEMG subset cardinality on speech recognition performance were cross-checked against the vocalized data set (see Figure 3). Although we did not perform an exhaustive evaluation of the subsets on the vocalized data, we did conduct a generous sampling of subsets across various cardinalities. Note that one would not *a priori* expect the subsets that yield the maximal recognition rates for vocalized data to be the same for mouthed data, since sensors recording muscular activity correlated with phonation should contribute to recognition capability in the vocalized but not necessarily the mouthed experiments. Additionally, it is reasonable to expect that the mouthed articulation differs somewhat from vocalized articulation because mouthing is not a completely natural use of the articulator muscles.

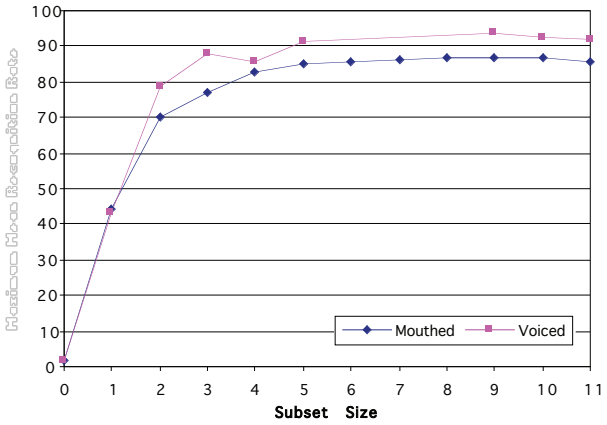

Figure 3. Comparison of the maximal mean recognition rate per sensor subset size between the two different phonation modalities of speech: vocalized and mouthed. Note that the evaluation of the subsets was exhausted over the mouthed data, but not over the vocalized data.

Comparing the results from the two different data sets, we conclude that the trends that we observed on the mouthed data are indeed reflected in the vocalized data, as well. More specifically, analysis of sensor subset sizes with respect to both vocalized and mouthed data provides empirical support that five sensors provides a desired balance of a significant reduction in the number of sensors while maintaining speech recognition capabilities to within a half a percentage point of that obtained using the full 11-sensor set: 84.9% for five sensors versus 85.4% for eleven sensors on the mouthed data, and 91.4% for five sensors versus 91.8% for eleven sensors on the vocalized data.

2.5 Subset Selection

With the preferred subset cardinality selected as five, we proceeded to identify a particular sensor subset or range of subsets to be chosen as best in terms of speech recognition capabilities. The top-scoring 5-sensor subset for the mouthed, individual-word data was comprised of Sensors 1 and 2 under the chin, Sensor 5 on the neck, and Sensors 9 and 10 on the face (see top left-hand corner schematic in Figure 4) and resulted in an 84.9% average recognition rate across all nine speakers. In addition to having the best average recognition rate, this configuration of sensors also

had the fourth smallest variance among all 462 5-sensor subsets. Notably, five other 5-sensor subsets yielded recognition rates within a single percentage point of the highest-scoring one (see Figure 4).

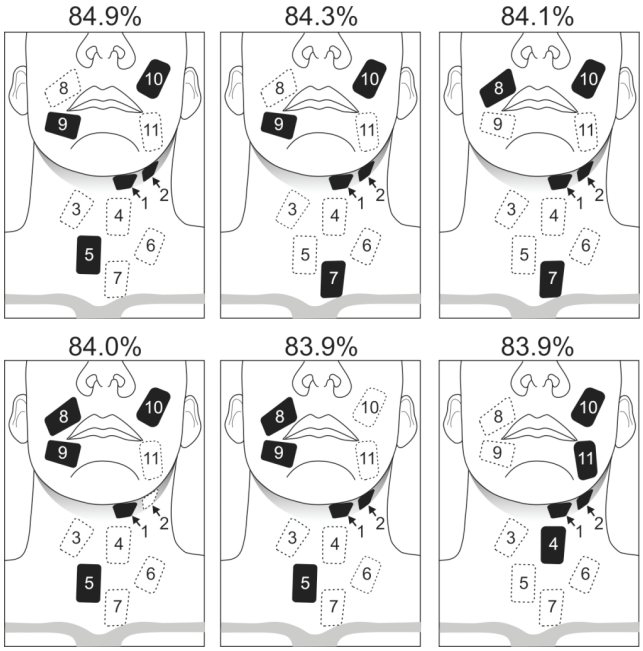

Figure 4. The six top-scoring 5-sensor subsets (black shading), with the mean recognition rate for each subset shown above the sensor schematics.

Patterns of the optimal reduced sensor sets were consistent with the anticipated muscle signal sources across the tested anatomical regions in relation to their known roles in voice and speech production. There was at least one sensor selected among each of the three targeted head/neck surface regions in the top 26 maximally-scoring 5-sensor subsets (see Figure 4). This suggests the importance of perioral face locations for capturing articulatory movements (Sensors 8-11), under the chin locations for representing tongue height and front versus back tongue position (Sensors 1-2), and the ventral neck surface for capturing laryngeal vertical position and possibly laryngeal activity (Sensors 3-7). Variability in the particular sensor(s) in each region for the optimal subsets indicates a degree of redundancy in information provided within (and possibly across) each region. The only sensor consistently represented in each of the top 6 subsets was Sensor 1 under the chin. This location happens to also match the optimal sensor position (among seven neck/face locations tested) for laryngectomy patients to control an EMG-activated artificial voice source [5], and is a robust information source regarding both intended voice activation and tongue control.

There was at least one sensor selected on the medial, ventral neck surface (Sensors 4, 5 and 7) in the top six 5-sensor subsets. These locations detect superficial neck "strap" muscles which normally help control and stabilize the height of the larynx in the neck while speaking; and are apparently active whether voice is actually produced or not, since they were important for recognition during both normal and mouthed speech modes.

3. CONCLUSION

In this study we evaluated the effect of the number of sEMG sensors on recognition accuracy for both normal and non-vocalized (mouthed) speech modes. The recognition accuracy increases rapidly with respect to the number of sensors, eventually plateauing at 5 sensors to within a percentage point of the full 11-sensor set. We observed a slight peak in performance at around 9 sensors, possibly due to the effect of overfitting at larger sizes of feature vectors given the limited amounted of training data available. The top six 5-sensor sets had at least one sensor location on each of the face, chin and neck regions under study, indicating each region's unique contribution to sEMG-based speech recognition. Further work is needed to determine how sEMG patterns from each sensor location relate to particular speech sounds or articulatory gestures, and whether the present sensor reduction findings will extend to recognition of continuous speech.

4. ACKNOWLEDGEMENTS

This study was sponsored by the United States Defense Advanced Research Projects Agency Defense Advanced Research Projects Agency (DARPA) Information Processing Technology Office (IPTO) Program Advanced Speech Encoding, under contract W15P7T-06-C-P437. The views and conclusions in this document are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of DARPA or the U.S. Government.

5. REFERENCES

[1] G.S. Meltzner, J. Sroka, J.T. Heaton, L.D. Gilmore, G. Colby, S. Roy, N. Chen, and C.J. De Luca, "Speech Recognition for Vocalized and Subvocal Modes of Production using Surface EMG Signals from the Neck and Face," INTERSPEECH 2008, Australia, 2008.

[2] M.S. Cheng, "Monitoring Functional Activities in Patients With Stroke." Sc.D. Dissertation. Boston University, Department of Biomedical Engineering, 2005.

[3] Voicebox Speech-Processing Toolbox for Matlab: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[4] HTK Speech Recognition Toolkit: http://htk.eng.cam.ac.uk/

[5] Stepp C.E., Heaton J.T., Hillman R.E. "Use of neck and face surface EMG for controlling a prosthetic voice after total laryngectomy", Conference on Motor Speech, Monterey, CA, March 6-9, 2008.