interspeech 2008 incorporating SST 08

Speech Recognition for Vocalized and Subvocal Modes of Production using Surface EMG Signals from the Neck and Face

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

[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

geoffrey.meltzner@baesystems.com, jason.sroka@baesystems.com, James.Heaton@mgh.harvard.edu, dgilmore@bu.edu, glen.colby@baesystems.com, sroy@bu.edu, nancyc@mit.edu, cjd@bu.edu

Abstract

We report automatic speech recognition accuracy for individual words using eleven surface electromyographic (sEMG) recording locations on the face and neck during three speaking modes: vocalized, mouthed, and mentally rehearsed. An HMM based recognition system was trained and tested on a 65 word vocabulary produced by 9 American English speakers in all three speaking modes. Our results indicate high sEMG-based recognition accuracy for the vocalized and mouthed speaking modes (mean rates of 92.1% and 86.7% respectively), but an inability to conduct recognition on mentally rehearsed speech due to a lack of sufficient sEMG activity.

Index Terms: sEMG, subvocal speech, speech recognition

1. Introduction

Human speech, being a natural and efficient means of communication, makes for an attractive modality for the interaction between humans and machines. The most common form that this man-machine interaction is automatic speech recognition (ASR), in which acoustic speech is translated into a sequence of speech tokens, typically words, using pattern classification techniques. ASR performance has achieved accuracies permitting commercial applications. As successful as automatic speech recognition has been, it does have some inherent weaknesses. Specifically, ASR performance degrades rapidly in the presence of acoustic noise, rendering it unsuitable for use in acoustically harsh environments. Moreover, because speech is an audible form of communication, maintaining privacy and security while using ASR is problematic. Finally individuals who have lost the ability to speak normally cannot make full use of ASR interfaces, even if their language function is intact.

These deficiencies motivate the need for an alternative form of speech recognition that does not rely on an acoustic speech signal. One potential alternative is to acquire speech information from surface electromyographic (sEMG) signals recorded from the muscles involved in speech production. Much of the speech musculature is recordable from the face and neck surface, providing sEMG signals applicable to ASR as a supplement or possibly as an alternative to the typical microphone input.

Although the field of sEMG-based speech recognition is relatively new, and is far from achieving the acoustic-based equivalent,, there have been some promising initial results. Chan *et al.* [1] obtained a 93% recognition rate on a vocabulary of 10 digits (zero through nine) using 5 sEMG channels on the face and neck for two subjects who produced vocalized (normally spoken) speech. Betts and Jorgensen [2] conducted a similar study on a single speaker but were able to achieve only a 74% recognition rate, albeit on a larger vocabulary of 15 words of voiced speech. Jou *et al.* [3] further extended the vocabulary size to 108 words but at the cost of reduced recognition accuracy (68%). Most recently, Lee [4] was able to achieve a mean 87% recognition rate on 60 vocalized words for 8 male, Korean speakers.

Because sEMG-based speech recognition does not rely on acoustic excitation of the vocal tract it is readily applicable to recognizing mouthed speech (in which the articulators go through normal production motions except no sound is produced). Using 3 sEMG electrodes, Manabe & Zhang [5] were able to recognize 10 Japanese digits for 10 speakers with an average recognition rate of 64%. Similarly, Maier-Hein *et al*. [6] used a Hidden Markov Model (HMM) based scheme to achieve a mean recognition rate of 97% for a set of 10 digits recorded from 3 talkers under mouthed speech conditions.

Further reducing the amount of overt speech activity, Jorgensen and Binstead [7] claimed to have developed a means of recognizing sEMG signals measured from the neck surface collected while subjects mentally rehearsed speech (i.e. mentally visualized speaking) They reported a mean accuracy rate of 72% on a 15 word vocabulary recorded from 5 individual speakers.

The ability to recognize silent (i.e. subvocal) speech could revolutionize aspects of human-computer interfaces, telecommunication, and assistive devices for the verbally impaired. We have extended the efforts of previous sEMG-based subvocal ASR studies, using a larger compliment of sEMG sensors across the face and neck, and applying both speech-based and sEMG-based signal processing techniques to improve isolated word recognition across all three speech modes; vocalized, mouthed, and mentally rehearsed.

2. Methods

2.1. Data Collection

2.1.1. Subjects

Data were collected from 9 subjects (4 females, 5 males), ranging in age from 20-42 years (mean = 27.5). All subjects

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

were native American English speakers and had no history of speech or hearing disabilities. We intentionally recruited athletic individuals with relatively slender necks to facilitate robust sEMG signal detection, with subjects having an average body mass index of 22.5 (±2.2).

2.1.2. Apparatus and Sensor Locations

Eleven sEMG sensors (parallel bar configuration, DE2.1, Delsys Inc., Boston MA) were positioned on particular neck and face locations predetermined in pilot experiments (N=3) to provide optimal speech-related information across 6 anatomical regions (supralabial, labial, sublabial, submental neck, midline neck, and lateral neck; see Figure 1).

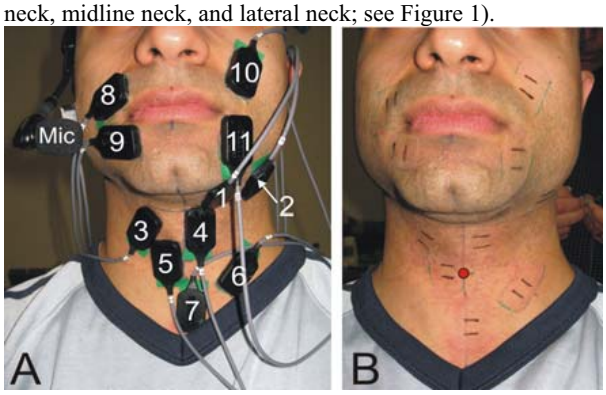

Figure 1. *The 11 sEMG sensor locations are shown before (A) and after (B) removal. Black pen lines in B show electrode contacts and the red dot (neck midline) marks the cricothyroid gap.*

Sensor placement sites were cleaned and exfoliated using alcohol pads followed by repeated adhesive tape application and removal. The neck and face midline were marked with medical adhesive tape strips or a felt-tip pen, as was a line dividing the ventral neck from the submental (under chin) surface. A flexible ruler was used to mark 1 and 4 cm distances from the neck midline along the submental line on the left side of the neck for placement of sensors #1 and #2, respectively. The experimenter then palpated the larynx to find and mark the cricothyroid membrane (confirmed by vocal pitch modulation) for placement of sensor #5 just lateral to the neck midline, and sensor #3 was positioned lateral to #5, rotated 30º, with the upper casing at the submental line. Sensor #6 was centered on the lower 1/3 point of the sternocleidomastoid, and sensors #4 and #7 were positioned as high and low (respectively) along the ventral neck midline as possible while maintaining flat skin contact.

Our pilot study results indicated that speech-related information from face sEMG sensors were highly dependent on sensor position, so we used sensor templates cut from transparency film to guide precise and consistent sensor positioning among subjects in the present study.

The templates for sensor positions 8-11 provided external reference points to match with anatomical features of the face Specifically, templates #8, 9 and 11 had extensions that were to be placed at the corners of the mouth. Template #10 had extensions that provided both the distance and position relative to the corner of the mouth (left face) and corner of the eye. Generic templates without particular external reference marks were used to draw outlines for sensor positions 5 and 6.

An elastic inductance plethysmography band (DS-X11A, Delsys Inc.) was worn around the chest to monitor breathing, and a headset microphone (WH30, Shure Inc., Niles, IL) was positioned approximately 5 cm in front of and slightly lateral to the mouth (see Fig. 1A) to record speech. EMG signals were band-pass filtered 20-450 Hz, the microphone was low-pass filtered at 10 kHz, and all signals were digitized at 20 kHz using a 32 channel A/D converter (NI-6259, National Instruments Co., Austin, TX) and EMGWorks data acquisition software (Delsys Inc.).

2.1.3. Experimental Tasks and Procedure

Subjects produced a set of 65 individual words three times each under three speaking modes: 1) vocalized – using normal speech production, 2) mouthed – with articulation but no vocal tract excitation (no voice), and 3) mentally rehearsed – with mental visualization of speaking aloud, but without articulator movement or voice production. Each subject generated a total of 6 tokens per word in the vocalized and mouthed speaking modes and 3 tokens per word in the mentally rehearsed speaking mode. The presentation order for individual words was randomized for each subject, and this order was maintained for each voicing condition within a subject. The word set included numbers 0-10, and various nouns and verbs that are commonly used for person-to-person communication and computer/device control (e.g. common replies, commands, locations, distances, times, etc.).

Synchronization between word production and sEMG signal analysis windowing was accomplished using a Microsoft PowerPoint presentation to prompt each word repetition, and a photo-diode mounted on the computer screen which sensed the word prompts, activated the data acquisition system, and provided a recordable signal representing the timing of repeated words. This was particularly important for knowing when the mouthed and mentally rehearsed utterances occurred given their lack of an audible signal. An audio file automatically played at the start of each new task word to provide the correct pronunciation. Task words were presented on slides with a series of count-down symbols (see Figure 2). A black outline moved across the successive count-down symbols from left to right in 1s intervals, ultimately surrounding the task word. Subjects easily learned the cadence of the countdown, and consistently produced the task word when the black box reached the word.

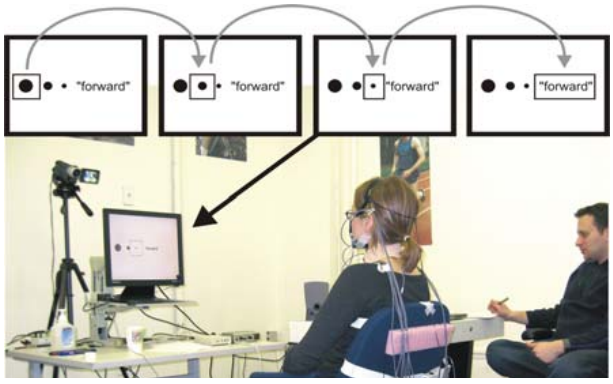

Figure 2. *The monitor prompts a subject to say the word "forward" when the black box, moving from left to right in 1 second steps, reaches the target (see upper insert).*

2.2. Data analysis

2.2.1. Feature investigation

Our investigation of a set of potential features revealed that the combination of Mel-frequency cepstral coefficients (MFCCs), which have been used successfully in this type of application, [1], [4], and muscle co-activation levels produced

the best recognition performance. We parameterized each sEMG channel using 6 MFCCs, their corresponding delta coefficients and the 0^{th} coefficient for each of the eleven sEMG channels. The MFCCs were augmented by the co-activation levels between pairs of EMG channels, which quantify the amount of simultaneous firing activity between all possible pairs of EMG channels. The co-activations levels were originally developed for use in the recognition of different gross motor movements based on sEMG signals [9]. The feature vectors were computed every 25 ms over a 50 ms window over all 11 sEMG channels.

2.2.2. sEMG segmentation

For sEMG signals collected under the vocalized speaking condition, segmentation can be accomplished by using the acoustic channel as a guide. However, for the silent speaking modes, this option is not available. As such, an sEMG based detection algorithm was developed that was able to 1) detect the onset and offset of speech related sEMG activity within a given channel and 2) determine when to mark the beginning and end of sEMG speech tokens based on activity detected on multiple channels.

The detection algorithm operates on a smoothed version of the sEMG envelope which is computed from the raw envelope by taking its mean over a 40ms window every 20 ms. The point at which the absolute value of the derivative of the smoothed envelope first exceeded a specified threshold was marked as the onset sEMG activity. Similarly, the last point was marked as the end of sEMG activity. The threshold was set to an empirically determined value and an adaptation measure implemented such that if the threshold level was never exceeded, the threshold was decreased by a small amount until an onset and an offset were detected. This adaptation feature ensured that all data were labeled.

Based on our investigation into sEMG channel latencies during speech activity, it was determined that channels 1, 5, 8, 9 and 11 were most likely to trigger first. As such, the detection algorithm was applied to these five channels, and the earliest onset and latest offset markers across all channels were chosen as the beginning and end of the speech sample, respectively.

2.3. Recognition Details

Standard Hidden Markov Models (HMMs) [11] were created for each of the 65 isolated words utilizing the HMM Toolkit (HTK) Version 3.4 [10]. Left-to-right word models [11] with eight emitting states were used, including the possibility to remain in the current state or to move forward one state in the state sequence represented in the model. The output probabilities for each of the states were relatively simply represented as a single Gaussian and a variance for each parameter (each parameter in each sEMG channel was treated as statistically independent in the model).

Training was performed in a speaker-dependent (training and test data were taken from the same speaker) and a speech modality-dependent manner (separate models were trained for vocalized, mouthed, and mentally rehearsed data). Of the six instances of each of the 65 isolated words in the data corpus (three for mental rehearsal), four were used for training and the remaining two for testing (two and one for mental rehearsal). Initial parameter estimates of the model parameters were generated using an assumption of equal distribution of data frames to HMM states (i.e. 1/8 of the samples assigned to each model state). The Baum-Welch algorithm was used to retrain the initialized model until convergence was achieved.

To perform recognition, the Viterbi algorithm was run to determine which of the word models had the highest maximum probability state sequence of producing the test word. The word model with the highest score was taken as the recognized word.

3. Results

3.1. Vocalized and Mouthed Speech

Although the entire 65 word isolated word vocabulary was collected together, we report separate digit recognition results in addition to the full vocabulary results to allow for an easier comparison with previous studies. Table 1 contains the results for both vocalized and mouthed speech recognition for all 9 subjects.

Recognition of vocalized speech was highly accurate, reaching a mean 98.3% recognition rate for the 10 digits and a corresponding recognition rate of 92.1% for the entire isolated word set. In the case of the digits-only recognition, a total of two errors were made across all 9 subjects.

While diminished when compared to that of the vocalized data, recognition performance on the mouthed data was still quite high. For the digit vocabulary, the subvocal recognition system attained a mean recognition accuracy of 96.7% and for the entire isolated word vocabulary it obtained a mean recognition rate of 86.7%. Most of the reduction in comparative performance on mouthed data can be attributed to recognition of Subject 9's speech for whom the accuracy rate differed by 20% between the two speaking conditions.

Table 1. *Summary of recognition results for both vocalized and mouthed speech.*

	Vocalized Speech		***Mouthed Speech***	
Subject	**Digits 0-9**	**Full Vocabulary**	**Digits 0-9**	**Full Vocabulary**
1	100.0%	88.5%	100.0%	88.5%
2	95.0%	93.1%	100.0%	84.6%
3	95.0%	90.0%	95.0%	89.2%
4	100.0%	91.5%	90.0%	82.3%
5	100.0%	96.2%	95.0%	92.3%
6	100.0%	96.9%	100.0%	92.3%
7	95.0%	96.9%	100.0%	89.2%
8	100.0%	84.6%	100.0%	90.8%
9	100.0%	90.8%	90.0%	70.8%
Mean	**98.3%**	**92.1%**	**96.7%**	**86.7%**

3.2. Mentally Rehearsed Speech

Recognition of mentally rehearsed speech was not possible due to a consistent lack of sEMG signals associated with "speech" production in this mode.

A typical example of sEMG activity in all channels can be seen in Figure 3 which shows the signals collected while Subject 1 mouthed and mentally rehearsed the word, "affirmative," The mentally rehearsed sEMG signals are flat throughout the experimental task while the mouthed data channels show clear signs of muscle activity. The lack of discernable sEMG activity during mentally rehearsed speech reduced the sEMG-based speech recognition performance in that modality no better than random.

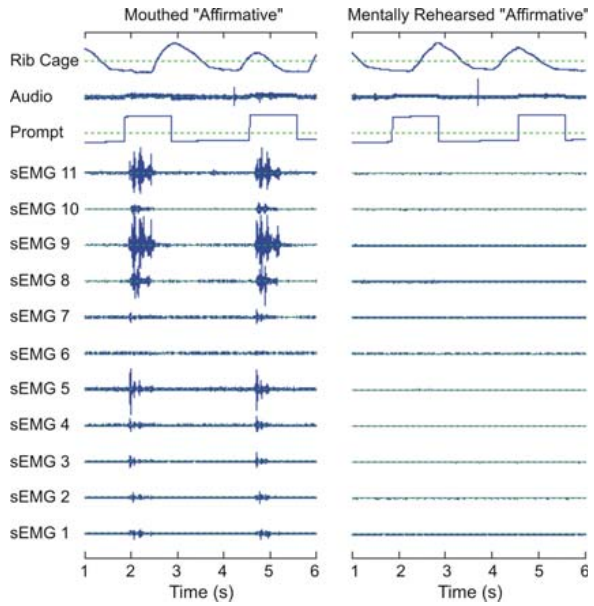

Figure 3. *Signals collected during t production of the word "affirmative" using the mouthed (left) and mentally rehearsed (right) speech modes (twice per mode). The top three channels show the signals measured by the rib cage monitor, microphone, and event prompt trigger, respectively. The 11 sEMG locations are shown in Figure 1. Note the lack of audio signal for these subvocal productions, and the absence of useful sEMG activity during mental rehearsal. (The sEMG signals are all plotted on the same scale.)*

4. Discussion

This study represents a comprehensive effort to explore the feasibility of isolated-word subvocal speech recognition using sEMG signals recorded on the face and neck across three modes of speech production (voiced, mouthed, and mentally rehearsed). Our results indicate that while highly effective recognition is possible for voiced and mouthed speech, it is not feasible for mentally rehearsed speech.

At this stage, our ability to recognize voiced and mouthed speech is limited to the speaker-dependent case: the recognizer must be trained and tested on the same speaker (but not the same utterances). While this situation would appear to be a limitation, there are useful applications for a speaker-dependent version. Should the system be used to recognize the speech of a disabled user, it would be necessary to train the system on that user to capture his/her specific articulation pattern. A speaker-dependent system is further justified by the fact that our recognition results were obtained using a minimal amount of training data; our training procedure used 4 tokens/ word. Requiring a new user to repeat only 4 times to achieve close to 90% recognition rates does not seem an undue burden. Further, as additional data are made available and as we explore recognition of phoneme-level speech tokens, the move to speaker-independent training can be explored further.

Our exploration of mentally rehearsed speech recognition has found that, contrary to previous findings [7], this mode of speech does not produce any recognizable signals. One could attribute this to a lack of intent on the part of the subjects. However, an inspection of the subjects' respiratory activity, as measured by the plethysmography band revealed that the respiration patterns during the mentally rehearsed tasks were similar to those found during the mouthed and vocalized tasks, suggesting that the subjects were indeed concentrating on the mental rehearsal task.

Another, more likely reason for this discrepancy could be the amount of training the subjects received prior to the recording of the sEMG signals. Previous studies [7] are unclear on the amount of training that subjects received, and how much feedback was provided on the quality of the signals they were producing was provided. If feedback was provided, those subjects could have learned surrogate non-speech behaviors that were mapped by the system into the recognition vocabulary.

5. Conclusions

This study has explored the possibility of performing isolated word recognition based on sEMG signals collected from the face and neck during (1) vocalized speech, (2) mouthed speech, and (3) mentally rehearsed speech. The resulting recognition performance is quite robust for a 65 word vocabulary in both the voiced and mouthed speech modes, producing 92.1% and 86.7% mean recognition rates (on 9 speakers), respectively. These recognition rates exceed those reported in the previous literature. Recognition of mentally rehearsed speech was not possible because the mental rehearsal tasks produced little to no usable signals.

6. Acknowledgements

This study was funded by the United States Defense Advanced Research Projects Agency under contract W15P7T-06-C-P437.

7. References

[1] Chan, A. D. C., Englehart, K., Hudgins, B, and Lovely, D.F. "Myoelectric Signals to Augment Speech Recognition," *Medical and Biological Engineering & Computing* vol. 39, pp. 500-506, 2001.

[2] Betts, B., J and Jorgensen C. "Small Vocabulary Recognition Using Surface Electromyography in an Acoustically Harsh Environment." *NASA TM-2005-21347*, 2005.

[3] Jou, S.C., Maier-Hein, L.,Schultz, T. and Waibel, A. "Articulatory feature classification using surface electromyography," in *Proc. ICASSP* 2006, pp 606-608.

[4] Lee, K-S. "EMG-Based Speech Recognition Using Hidden Markov Models With Global Control Variables." *IEEE Trans. On Biomed. Eng.*, vol 55, pp. 930-940, 2008.

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

[6] Maier-Hein, L., Metze, F., Schultz, T., and Waibel, A. "Session Independent Non-Audible Speech Recognition Using Surface Electromyography." *IEEE Automatic Speech Recognition and Understanding Workshop*. p. 331-336, 2005

[7] Jorgensen, C. and Binstead, K. "Web Browser Control Using EMG Based Sub Vocal Speech Recognition." *Proc. Int. Conf. on System Sciences* 2005. pp. 294c-294c, 2005.

[8] De Luca, C.J. "Surface Electromyography: Detection and Recording" www.Delsys.com, 2002.

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

[10] HTK Speech Recognition Toolkit, 2007, http://htk.eng.cam.ac.uk/

[11] Rabiner, L. and Juang, B.H. *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, NJ, 1993.