33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Signal Acquisition and Processing Techniques for sEMG Based Silent Speech Recognition

Geoffrey S. Meltzner, Glen Colby, Yunbin Deng, and James T. Heaton

***Abstract*—sEMG based silent speech recognition systems seek to bypass the limitations of acoustic speech recognition by measuring and interpreting muscle activity of the facial and neck musculature involved in speech production. However, this speech recognition modality introduces unique challenges of its own. This paper describes signal acquisition and processing strategies that we have employed to address these challenges during our development of a silent speech recognition system.**

I. Introduction

Recent work in the field of non-acoustic automatic speech recognition (ASR) has been motivated by the desire to mitigate two significant weaknesses of standard, acoustic ASR: (1) severe performance degradation in the presence of ambient noise and (2) a limited ability to maintain privacy/secrecy because of the requirement of using audible speech. These non-acoustic ASR studies have investigated alternative modalities, such as ultrasound [1] or surface electromyography (sEMG) [2-9] that can capture sufficient speech information while overcoming the aforementioned deficiencies of acoustic ASR systems.

sEMG based speech recognition, also known as subvocal speech recognition, operates on signals recorded from a set of sEMG sensors that are strategically located on the neck and face surface to measure muscle activity associated with the phonation, resonation and articulation of speech. Because the signals are a direct measurement of the articulatory muscle activity there is no need for acoustic excitation of the vocal tract, making it possible to recognize silent, mouthed speech. Moreover, because sEMG signals are effectively decoupled from acoustic signals, they are immune to acoustic noise corruption.

These two properties have made sEMG an attractive alternative modality for speech recognition and have motivated several research efforts. Chan *et al.* [3] 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 [4] 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.* [5] further extended the vocabulary size to 108 words but at the cost of reduced recognition accuracy (68%). Other studies were able to demonstrate improved recognition rates for somewhat smaller vocabulary sizes. For example, Lee [6] was able to achieve a mean 87% recognition rate on 60 vocalized words for 8 male, Korean speakers. Similarly, we reported recognition rates of 92.1% and 86.7% on vocalized and mouthed speech, respectively for a group of 9 American English speakers for a 65 word vocabulary [2]. More recently, Wand and Schultz have pushed sEMG based recognition towards continuous speech recognition [7], ultimately achieving a Word Error Rate (WER) of 15.66% on a vocabulary of 108 words [8].

Despite the significant advances made in sEMG based speech recognition performance, this performance level still greatly lags that of standard acoustic based ASR systems, Moreover, even though, sEMG based speech recognition is able to address the two aforementioned shortcomings of standard ASR systems, this alternative modality introduces challenges of its own. This paper will explore these challenges and some signal processing techniques used to address them.

II. sEMG Sensing Challenges

A. Sensor Locations

Many of the target muscles involved in speech production are relatively superficial and therefore easily accessible for recording, whereas others are relatively deep (laryngeal/pharyngeal) or otherwise poorly situated for conventional sensor placement (e.g. intrinsic muscles of the tongue). Given that there are practical limitations to the number of sEMG recording locations obtainable from the body surface above the speech musculature, we sought to identify an optimal sensor configuration in preliminary experiments prior to collecting our larger data sets. We started by identifying 6 regions across the neck and face surface (supralabial, labial, infralabial, submental, ventromedial neck, and ventrolateral neck) superficial to muscles involved in speech production, and identified one or two sensor positions within each region that had been used in prior sEMG speech studies [3-5, 9] and/or were above

Manuscript received April 15, 2011. This study was sponsored by the United States Defense Advanced Research Projects Agency (DARPA) Information Innovation Office (I2O) Program, "Advanced Speech Encoding," under contracts W157T-08-C-P021 and 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.

G. S. Meltzner is with BAE Systems, Inc., Burlington, MA 01809 USA (phone: 781-262-4773; fax: 781-273-9345; e-mail: geoffrey.meltzner@baesystems.com).

G. Colby is with is with BAE Systems, Inc., Burlington, MA 01890 USA. (e-mail: glen.colby@baesystems.com).

Y. Deng is with is with BAE Systems, Inc., Burlington, MA 01890 USA. (e-mail: yunbin.deng@baesystems.com).

J.T. Heaton is with the Massachusetts General Hospital Department of Surgery, Boston, MA 02114 USA (email: james.heaton@mgh.harvard.edu)

prominent speech muscles. The resulting 11 "default" sensor locations served as reference points for sensor position mapping experiments performed in an adult male and female participant.

The mapping strategy involved systematically moving one or two (at a time) single-differential bar-type sEMG sensors (DE 2.1 by Delsys Inc, Boston, Ma) in 5mm steps across a face or neck region during mouthed speech production while maintaining the default locations for all other regions. This enabled us to assess the sEMG speech content of each mapped position in relation to a complete set of sEMG signals, examining several sensor configurations for each subject. We found that the information content did not markedly change when sensors were moved along predetermined trajectories across the labial region (above the orbicularis oris superioris and inferioris), but that there were consistent optimal locations in the other zones for our two test subjects in reference to their body midline. In the supralabial region, sensor placement at a location likely above the zygomaticus minor and levator anguli oris provided the best information. Likewise, in the infralabial region, the sensor position above the depressor anguli oris and depressor labii inferioris provided the most unique information. In the submental region, sensor positions above the anterior digastric and mylohyoid was the best, as well as a location 3 cm lateral to that position (above the platysma and lateral mylohyoid).

Along the ventromedial neck, a sensor position falling above and slightly lateral to the cricothyroid membrane was optimal, as well as a position in the same medial location but closer to the chin. Findings from our sensor position mapping experiment enabled us to identify 11 of the most appropriate (if not optimal) sensor positions, but did not indicate the relative importance of each recording location or redundancies in the data they provide.

B. Sensor Set Reduction

Minimizing the sensor set reduces the system complexity, application effort, body contact, and likelihood that a given sensor will lose skin contact. In a recent study [10] we systematically analyzed speech recognition performance from all possible subset combinations of our 11-sensor set in 9 participants, and identified the best combination(s) of sensor locations to achieve mouthed word recognition rates comparable to our full set of 11 locations. We showed that 5-8 sensors are sufficient to achieve a recognition rate to within a half a percentage point of that obtained from the full set, and that there were numerous possible combinations of which sensors achieved this high recognition accuracy. Moreover, 3 of the 5 sensor locations on the ventral neck were clearly dispensable, which helped us define a reduced set of 8 sensor locations that we have used for recent recordings.

The 8 sensor locations are depicted in Figure 1. The sensors are a customized version of the Trigno™ wireless sensors (Delsys Inc.) measuring 11 x 21 mm with silver bars 5 mm long spaced 10 mm apart attached to the skin with double-stick strips. The face sensors are placed in relation to a template pattern (printed on overhead transparency film) that uses the corners of the mouth and eye as a reference points. The two ventral neck sensors are placed with their medial edge 5 mm from the midline, with the more inferior sensor (#4) above the cricothyroid membrane and the superior location (#3) just inferior to the submental surface. The submental sensors (#1-2) are centered on the submental surface with medial edges 10 and 40 mm lateral to the midline. These 8 sensor locations avoid major surface curves and creases for the most part, and have maintained good signal fidelity for as long as 12 hours at a time (spanning multiple meals and conversations).

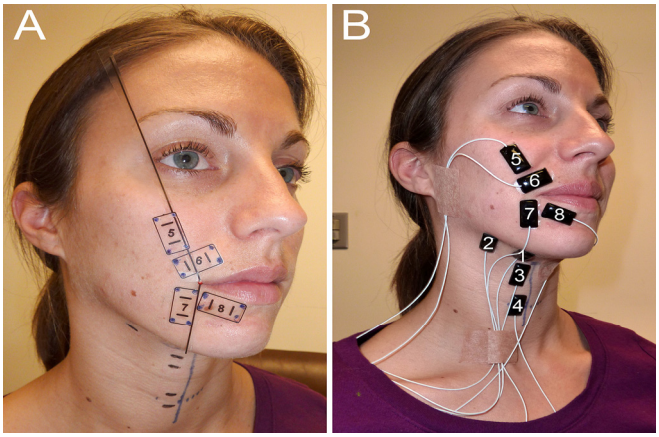

Fig. 1. sEMG sensor locations. A) The template for placement of sensors 5-8 is shown with blue pen marks on the skin at the inside edge of each sensor outline corner. Template lines extend from the corner of the mouth vertically downward, and angled upward to cross the corner of the eye. Black pen marks on the neck show sensor electrode bar locations. B) Sensors on the submental neck (1-2), ventromedial neck (3-4), supralabial face (5-6) and infralabial face (7-8).

C. sEMG Signal Processing and Recognition

An sEMG based ASR system has an overall structure that consists of essentially the same two core components as a standard acoustic ASR system, i.e. (1) a front end that parameterizes the signal(s) of interest so as to reduce the amount of data to be processed, and (2) a recognition back end consisting of the probabilistic models that ultimately generate the proper speech labels based on observations made on the input signals. This paper is concerned with the signal acquisition and processing aspects of sEMG based ASR, and will therefore focus on front-end parameterization. We do note that our sEMG ASR system used a Hidden Markov Model back end and refer readers to [2], [10], and [12] for further details on this matter.

1) Speech Activity Detection

Because the sEMG based recognition system often operates on mouthed speech, we use the term SAD (Speech activity detection), instead of VAD (Voice Activity Detection). Despite a rich body of research that has been directed towards acoustic VAD, it remains a challenging problem, and sEMG based SAD is in many ways an even more difficult problem. It is difficult to define ground truth

for the beginning and ending of speech-related activity for sEMG signals, as muscle contractions precede speech production by varying amounts of time. This is even more challenging for mouthed speech detection, because without an acoustic cue, it is hard to differentiate sEMG signals associated with speech activity from those associated with non-speech movements. The problem is further complicated by the use of multiple sEMG channels, and each channel's speech activity related behavior can be both utterance and speaker dependent. Our initial SAD algorithm used a thresholding technique that operated on smoothed sEMG envelopes of 5 channels which we empirically determined to activate first when speaking [2]. This algorithm provided satisfactory performance for small vocabulary, isolated word recognition, but a more sophisticated statistical approach was needed for continuous recognition. Specifically, for each 50 ms window, we computed third order statistics (TOS) for each selected channel:

$$TOS = E[x^3(n) - x(n-1)x(n)x(n+1)] \quad (1)$$

, where $x(n)$ is the sEMG signal sample value at time n after DC offset removal, and $E[]$ stands for expectation. The channel was labeled as active if the TOS crossed a threshold value, which was tuned on training data to balance missed detections and false alarms. The start of speech was detected if a channel had been active for 5 consecutive frames or at least two channels were active at the same time. The maximum TOS values for all active channels were recorded and updated for each window. An active channel was labeled as inactive once its current TOS value fell below 15% of its maximum TOS value in history. The end of speech was marked as the point at which all sEMG channels became inactive.

The TOS based SAD algorithm proved to be more accurate than the original SAD algorithm but its performance notably degraded in the presence of simultaneous non-speech activity. We also discovered that this SAD algorithm was not well suited to real-time implementation. As such, we developed a third SAD algorithm, which incorporates a multi-channel decision logic that takes advantage of the fact that speech production typically involves multiple muscles at the same time and is thus able to ignore noise in any single channel. To balance the trade-off between simplicity (desired for real-time implementation) and robustness, our current SAD algorithm is based on these principles: 1) using a short time windowed signal to compute local statistics; 2) online background noise and real signal statistics tracking on each channel; and 3) a global decision based on the five selected best sEMG channels. The SAD algorithm operates on two levels of finite state machines, shown in Figures 2 and 3. The first level consists of a finite state machine for each channel, which determines each channel's speech state. As shown in Figure 2, an active/inactive decision is made on each windowed time instance, *t*, by comparing current statistics with minimum background and maximum signal statistics, where s*osTh* and e*osTh* are threshold parameters for start and end of speech decision, respectively.

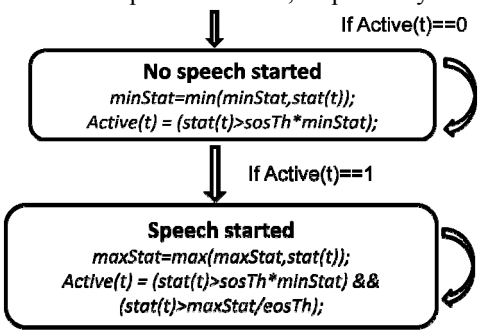

Fig 2. Finite state machine for an individual sEMG channel.

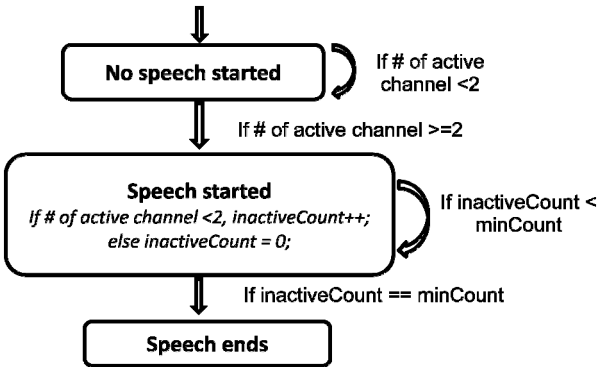

Fig 3. Finite state machine used for overall SAD decision making.

The higher level machine, as shown in Figure 3, combines each channel's states to make the final start of speech (SOS) and end of speech (EOS) decision. This new SAD continuously adapts to the background signal level and the speaker/utterance specific maximum energy level for each channel. Our evaluations have shown that it is more accurate than the TOS based SAD, especially when sEMG signals associated with non-speech activity are present.

2) Signal Parameterization

Mel-frequency cepstral coefficients (MFCCs) are the gold standard parameterization scheme for acoustic ASR systems, likely because they approximate human auditory system response to acoustic signals. However, sEMG signals are very different from acoustic speech signals, as they exhibit slower changes and less fine structure. In addition, whereas typical acoustic ASR systems operate on a single acoustic channel, subvocal ASR systems process several sEMG channels simultaneously. Although some studies have successfully used MFCC features for subvocal recognition [6], it is not clear that MFCCs represent the most effective parameterization scheme for sEMG signals. Indeed, a number of researchers have reported differing levels of success using a variety of features, including wavelets [3], wavelet packets [4] and a customized set of time-domain features [5,7-9]. Our initial work in this area commenced with an investigation into a number of potential parameterization schemes, including MFCCs as well as a set of features derived to specifically process sEMG signals for

the recognition of gross motor movements [11]. Using a round-robin technique that compared the recognition performance obtained for all possible feature combinations, we determined that the combination of MFCCs (and their corresponding delta features) and muscle co-activation levels (quantified amount of simultaneous firing activity between all possible pairs of EMG channels) produced the best recognition performance. This parameterization technique was able to produce an average recognition rate of 86.7% on a 65 isolated word vocabulary for 9 speakers [2].

We sought to improve recognition performance through further modifications of the parameterization algorithm. Initially, we implemented cepstral Mean and Variance Normalization (MVN), and compared this scheme against our original algorithm as well as other parameterization schemes, including wavelets and the time domain features described in [5]. The results, based on recognition tests performed on the 10 digit subset of the vocabulary in [2], showed that the MVN-based feature set significantly outperformed the other candidate feature sets. In fact, the addition of MVN obviated the co-activation features and therefore they were dropped from the processing scheme.

Subsequent alterations focused on tuning the MFCC algorithm to better suit the characteristics of sEMG signals. MFCCs are computed using a multi-stage process, and some of the stages of the MFCC algorithm, namely the filter bank and the non-linear compression stages, are amenable to modification, depending on the properties of the signals being processed. Typically, for acoustic speech sampled at 16 kHz, a 24 channel filter bank and a log compression algorithm are used, resulting in 12 cepstral coefficients. Compared with acoustic speech signals, the spectral information present in sEMG signals is concentrated at much lower frequencies, suggesting that the use of a lower sampling rate, coupled with a smaller filter bank would be more effective.

Thus, to determine the best parameters to use in the MFCC computation algorithm, we conducted a series of recognition experiments using features generated by different modified MFCC parameterizations. Specifically, we explored the effects of modifying the following parameters: (1) the upper frequency cut off of the filter bank, (2) the lower frequency cutoff of the filter bank, (3) the frequency scale transform constant, (4) the number of channels in the filter bank, and (5) the type of non-linear compression (log compression vs. root compression, i.e. $x^{0.1}$, where x represents the output of the Mel-scale filter bank). We conducted recognition experiments on a 113 isolated word vocabulary and found that performance improvements could be obtained if the number of filter banks was reduced to 15 and if root compression was used. This modified MFCC parameterization algorithm was used successfully in the recognition of disordered subvocal speech [12]. More recently, using this modified MFCC algorithm, we have been able to achieve a 96.9% recognition rate on a continuous vocabulary of 200 words.

III. Future directions

Being a relatively new field, the state of sEMG based ASR is far less mature than that of standard, acoustic ASR. Acoustic ASR technology has developed to the point where it is now found in several different commercial products, ranging from mobile handsets to corporate call centers. With further development, we see no fundamental reason why sEMG ASR cannot similarly appear in numerous communication system and device control applications However, in order to bring this vision to fruition, this new technology must become more robust and more practical in terms of both algorithmic performance (particularly in the presence of non-speech movements) and hardware that is easy to wear, non-obtrusive, and able to function under possibly adverse conditions.

Acknowledgment

The authors would like to thank Carlo J. De Luca, Gianluca De Luca, Don Gilmore, and Serge Roy of Delsys, Inc. for their assistance in designing and running the sEMG data collection experiments and for their work in developing and providing the sEMG sensors and related hardware.

References

[1] T. Hueber, G. Chollet, B. Denby, M. Stone, "Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application," International Seminar on Speech Production, pp. 365-369, Strasbourg, France, 2008.

[2] 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.

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

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

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

[6] 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.

[7] T Schultz and M. Wand, "Modeling Coarticulation in EMG-based Continuous Speech Recognition," *Speech Comm.*, vol 52, 2010.

[8] M. Wand and T. Schultz, "Session-Independent EMG-based Speech Recognition," *International Conference on Bio-inspired Systems and Signal Processing* 2011

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

[10] G. Colby, J.T. Heaton, L.D. Gilmore, J. Sroka,, Y. Deng, J. Cabrera, S. Roy, C.J. De Luca, and G.S.Meltzner, "Sensor Subset Selection for Surface Electromyography Based Speech Recognition", ICASSP 2009. 2009.

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

[12] Y Deng., R. Patel, J.T. Heaton, C. Colby, L.D. Gilmore, J. Cabrera, S.H. Roy, C.J. De Luca and G.S. Meltzner "Disordered Speech Recognition Using Acoustic and sEMG Signals." *INTERSPEECH 2009,* United Kingdom, 2009.