Published in final edited form as:

IEEE/ACM Trans Audio Speech Lang Process. 2017 December ; 25(12): 2386–2398. doi:10.1109/TASLP.2017.2740000.

Silent Speech Recognition as an Alternative Communication Device for Persons with Laryngectomy

Geoffrey S. Meltzner,
VocaliD, Inc. Belmont, MA, 02478, USA

James T. Heaton,
Harvard Medical School in the Department of Surgery, Massachusetts General Hospital Voice Center, Boston, MA 02114

Yunbin Deng,
BAE Systems, Burlington, MA 01803 USA

Gianluca De Luca,
Delsys, Inc., and Altec, Inc., Natick MA 01760 USA

Serge H. Roy, and
Delsys, Inc., and Altec, Inc., Natick MA 01760 USA

Joshua C. Kline
Delsys, Inc., and Altec, Inc., Natick MA 01760 USA

Abstract

Each year thousands of individuals require surgical removal of their larynx (voice box) due to trauma or disease, and thereby require an alternative voice source or assistive device to verbally communicate. Although natural voice is lost after laryngectomy, most muscles controlling speech articulation remain intact. Surface electromyographic (sEMG) activity of speech musculature can be recorded from the neck and face, and used for automatic speech recognition to provide speech-to-text or synthesized speech as an alternative means of communication. This is true even when speech is mouthed or spoken in a silent (subvocal) manner, making it an appropriate communication platform after laryngectomy. In this study, 8 individuals at least 6 months after total laryngectomy were recorded using 8 sEMG sensors on their face (4) and neck (4) while reading phrases constructed from a 2,500-word vocabulary. A unique set of phrases were used for training phoneme-based recognition models for each of the 39 commonly used phonemes in English, and the remaining phrases were used for testing word recognition of the models based on phoneme identification from running speech. Word error rates were on average 10.3% for the full 8-sensor set (averaging 9.5% for the top 4 participants), and 13.6% when reducing the sensor set to 4 locations per individual (n=7). This study provides a compelling proof-of-concept for sEMG-based alaryngeal speech recognition, with the strong potential to further improve recognition performance.

Index Terms

Alaryngeal Speech; Assistive technology; Augmentative and Alternative Communication; Automatic Speech Recognition; Subvocal Speech Recognition; EMG; electromyography

I. Introduction

Human speech is a natural and efficient means of communication, yet millions of people around the world with severe speech disorders are unable to communicate effectively through vocalization. Instead, depending on the nature of the speech disorder, they must rely on augmentative and alternative communication (AAC) devices or software. These include artificial voice sources after loss of laryngeal function and/or speech synthesizers for individuals unable to articulate speech sounds. Unfortunately, these alternative communication solutions typically provide unnatural sounding vocalization or require the involvement of the user's hands, thus complicating everyday interactions and making them unwieldy. One technology that has been leveraged for assisting those with speech disorders 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 for those with normal speech function has achieved accuracies approaching 100%, permitting effective commercial applications and integration into portable speech-based human-machine interfaces. However, as successful as ASR has been for the general population, research on ASR of disordered speech is limited [1][2][3][4][5] and has almost exclusively focused on recognition of acoustic speech. Individuals who have lost the ability to speak normally cannot make full use of ASR interfaces, even if their language function is intact. ASR performance also degrades rapidly in the presence of acoustic noise, rendering it unsuitable for use in acoustically harsh environments, and it lacks privacy when used as a computer interface.

For the specific condition of voice rehabilitation after total laryngectomy, current options are fraught with limitations, including poor skill acquisition and intelligibility with esophageal speech [7], poor tissue viability or health issues often precluding the use of trachea-esophageal (TE) speech [8][9], and mechanical sounds coupled with the need to dedicate one hand for electrolarynx (EL) speech [10][11]. Furthermore, none of these options produce natural-sounding speech. These ASR and voice rehabilitation deficiencies call for an alternative form of speech recognition that does not rely on acoustic speech and can provide an interface for natural-sounding speech synthesis. The case for this approach is further bolstered by the advent of new technology that enables the creation of personalized synthetic voices using only small audio segments [12].

A number of biosignal modalities have been studied in the context of developing non-acoustic speech communication systems, including ultrasound, optical imagery, electropalatography, electroencephalography (EEG), and surface electromyography (sEMG) (see [13] for a comprehensive review of these studies). sEMG-based speech recognition provides a particularly attractive alternative platform through which individuals can communicate via synthesized ASR-to-speech or interact with computers via ASR-to-text. sEMG-based ASR operates on signals generated by the face and neck musculature, which are recorded from sensors placed on the face and neck skin surface. As such, it can be performed non-invasively while an individual produces audible speech or simply "mouths" their speech (i.e. so-called subvocal speech recognition where no voice is produced), and can thus augment or completely replace audible speech after laryngectomy or other causes of voice or speech disorders. Studies to advance sEMG-based subvocal speech recognition are relatively few in number and were initially limited to developing algorithms from isolated words among normal speakers [14][15][16],[18]. Other work has pushed the technology towards continuous speech recognition, with promising results, albeit on limited vocabularies [21][22]. More recently, studies have attempted to leverage advances in Deep Learning to improve recognition performance [23][24].

Our work in this field has focused on the development of *the Mouthed-speech Understanding and Transcription Engine (MUTE)*, primarily for providing covert "silent" communication for Defense applications. Our initial study used 11 sEMG sensors coupled with a Markov model (HMM)-based recognition system that was trained and tested on a vocabulary of 65 isolated words during vocal and subvocal (mouthed) speech modes [27]. Our results from n=9 speakers indicated the feasibility of achieving sEMG-based recognition accuracy for the mouthed speaking mode that was comparable to the vocal speech mode (accuracy of 86.7% and 92.1%, respectively).

Our more recent studies have advanced the algorithm development with an emphasis on significantly expanding the vocabulary set, achieving continuous speech recognition rather than the isolated word recognition of our prior work [25][34], and improving sensor set reduction to achieve a 4-sensor capability for ease of use. Our subvocal isolated word and continuous speech recognition capability from a 2,000-word vocabulary was reported on n=8 subjects [25]. The subvocal speech data corpus covered commonly used English words, including the TIMIT Acoustic Continuous Speech Data Corpus [29], Special Operation military commands [30], a Common Phrases set [31], a Text Message set [32], and a digits and date String set. Software algorithms using a variety of advanced signal processing technology were developed and tested using this data corpus to deliver a subvocal speech recognition engine, whose performance averaged 88.1% when tested on words not in the training vocabulary. These performance metrics significantly outpace those reported by other groups working in this field [15][18][19][20]. Details of how these accomplishments were achieved are summarized in Section II below, and in our prior publications [25][26][27][28][34].

The results of these studies suggested that sEMG-based speech recognition could enable speech-based communication for those who cannot communicate acoustically. As such, we also applied this technology towards recognizing the speech of those with speech disorders caused by cerebral palsy, traumatic brain injury (TBI) and stroke, and found there was great potential of using sEMG-based recognition for disordered speech, so long as the disordered speech articulatory patterns were reproduced consistently [26].

This latter finding suggests that sEMG-based recognition could be well suited as a means of assistive communication for persons living with laryngectomy. Those who have undergone a total laryngectomy typically have intact (or mostly intact) articulatory abilities that approach that of healthy talkers, and they are accustomed to speaking in a subvocal manner (i.e. without engaging their vocal folds and making use of laryngeal proprioceptive feedback), perhaps hastening their use of mouthed speech as a communication modality. Furthermore, our prior work has demonstrated that sEMG-based ASR does not necessarily depend on recording locations altered by total laryngectomy (e.g. the ventromedial neck surface – see Methods). As such, we sought to adapt our sEMG technology to achieve similar performance metrics for people with laryngectomy compared to anatomically intact individuals while overcoming the challenges inherent with this speaker population. Specifically, these challenges include the necessity of achieving comparable performance metrics from recording locations that are not dependent on the presence of the larynx or associated extrinsic laryngeal and other infrahyoid musculature excised during laryngectomy.

II. Methods

A. Data Collection

Subjects—Data were collected from 8 male subjects ranging in age from 57 to 75 (mean = 64 years). All subjects were fluent and literate in American English and at least 6 months after total laryngectomy to ensure that the tissue around the surgical site was healed enough to tolerate the multi-hour experiment. All participants voluntarily provided written informed consent approved by the Western Institutional Review Board.

Data Corpus—The main data corpus was designed to cover a vocabulary of 2,500 words comprised of commonly used English words and phrases, while balancing the frequency of phoneme combinations for unforeseen testing words. The text for our data corpus consisted of 280 key sentences from a message corpus initially developed at Boston Children's Hospital for message banking by those diagnosed with early stages of ALS [33], and 150 sentences from the TIMIT-SI corpus [29] (used in our prior experiments [25][34]) to comprise the testing component of the revised corpus. The remaining training component was taken from the TIMIT-SX (450 sentences) corpus [29] combined with a set of common phrases, resulting in a corpus of 980 sentences, 550 used for training and the remainder for testing. The sentence tokens were presented in the same order for each subject to ensure that the required number of training tokens were obtained in a single session prior to any subjects needing to end the experiment due to fatigue or discomfort (during the course of the experiments, no subject ended the experiment early). Furthermore, we created a supplemental corpus of 280 sentences from the Children's Hospital Message banking corpus to test the efficacy of providing more training data to our models. Two subjects were given this supplemental corpus which was used to train the subject-dependent recognition models. Unlike the case of the main corpus, the supplemental corpus contained some repetitions of sentences between training and testing sets. Supplemental corpus results are reported separately in section IIIC. The mean sentence length of the entire corpus was 6.8 words per sentence.

Data Acquisition—A custom wireless data acquisition system [25] was used to simultaneously record sEMG signals from 8 locations on the face and neck (Fig. 1) while the subject mouthed the sentence tokens. The custom system consists of 4 active sensor "pairs" (each sensor being 10mm × 20mm × 3mm) containing sensing and conditioning electronics with double parallel bar electrodes spaced 1 cm apart for differential recording relative to a ground reference electrode placed at the lower cervical spine (Fig 2). The sensor pairs were joined by a flexible cable (to accommodate the different contours of the face and neck region) and connected to a wireless transceiver which communicates with a Trigno™ base station (Delsys Inc, Natick MA). Signals were conditioned for a maximum input range of 11mV, bandpass filtered from 20-450 Hz (80 db/dec), sampled at 2000 samples/sec, wirelessly transmitted using a custom protocol (<500us inter-sensor latency; <0.5% bit error rate), and interfaced over USB to a PC workstation (Dell, Inc. Round Rock, TX).

Sensor locations—The skin area of each sensor location was first prepared using a disposable shaver (if needed), alcohol wipes, and repeated tape peels to remove facial hair, oils, and exfoliates (respectively). Position and alignment of sensors relative to anatomical landmarks on the face and neck relied on templates, as outlined in a previous publication [25]. The ventral neck sensor pairs were placed in submental (#1,#2) and ventromedial (#3,4) regions, and the face sensors were placed over supralabial (#5,6) and infralabial (#7,8) regions (see Fig. 1). A double sided hypoallergenic adhesive tape with cutouts for the electrode pairs was used to secure the sensors to the skin.

Sensor locations were selected based on the results of previously conducted sensor location and reduction studies, which sought to identify the optimal sensor configuration from among the many possible sites of the face and neck involved in speech articulation. Six superficial muscle regions across the neck and face were selected to identify one or two sensor locations within each region that either had been shown to be effective in prior sEMG speech studies [14][15][16][17] or were reliably accessible to prominent speech muscles. The results demonstrated that sEMG information from mouthed speech was robust to shifts in sensor location across specific labial regions, and there were consistent optimal locations in other zones when referenced to the muscle midline [27]. Findings from the sensor mapping experiments enabled us to identify 11 appropriate (if not optimal) sensor locations for subvocal speech. However, the impracticality of requiring a user to don n=11 sensors prompted us to systematically analyze subvocal speech recognition performance from all possible subset combinations to identify the best combination(s) of locations to achieve WERs comparable to our full set of 11 locations [28]. Word recognition accuracy increased rapidly with respect to the number of sensors, eventually plateauing at 5 sensors to within a percentage point of the full 11-sensor set. Moreover, 3 of the sensors located on the ventral neck were clearly dispensable, which helped us define a reduced set of 8 sensor locations that are used in this study. The data collected from the 8 sensor locations were then analyzed to identify sensor subsets that produced recognition performance that approximated that of the full 8-sensor set (see Results).

Our choice of sensor locations, which were selected based on a combination of anatomical/physiologically based targeting and systematic quantification, do exhibit some overlap with sensor locations used in other studies that use a set of individual electrodes, whether guided by purely anatomical and physiological concerns [17][22] or produced from trial and error [18]. However, there are two significant differences: 1) our locations tend to target more muscles in the labial and submental regions and 2) we use a unilateral sensor set (i.e. the sensors are placed on one side of the face) whereas other studies have used a bilateral set. An additional study used a unilateral sensor set that took the form of a patch array of electrodes [53]. However, the array was placed in a more lateral position than ours, ostensibly due to limitations incurred by the form factor of the patch array.

Protocol—A data collection Graphical User Interface (GUI) and infrastructure were designed to implement the data corpus protocol in a manner that minimized fatigue of the laryngectomy users and allowed them to control the pacing of the sentences they read in a subvocal (*i.e.* mouthed) manner. The infrastructure consisted of a subject prompt GUI on one monitor and a signal acquisition GUI on another monitor that were both controlled by a software running in Matlab (MathWorks, Natick MA). The signal acquisition GUI was integrated with the Delsys data acquisition Software Development Kit to enable the data to be collected directly into Matlab. Subvocal prompts were displayed one at a time on the screen facing the user, which they advanced after mouthing each sentence. Data from the sensors were displayed in real time on the operator screen to identify signal artifacts or extraneous contractions requiring these sentences to be repeated See [13] for an example of collected signals. Signal noise such as motion artifact occasionally prompted individual sensor reapplication with fresh double-sided adhesive interfaces. Participants were encouraged to refrain from extraneous head/neck movements during sentence production (e.g. coughing, swallowing, and speaking off script), but were free to do these things or take stretching or bathroom breaks at any point between subvocal prompts. Participants were also asked to indicate when they make errors in mouthing the target sentences so that recordings with errors could be replaced by a repetition. Because of the practical limitations of ensuring that all subjects complied with the protocol at the time the data were being collected, we also adopted a policy of reviewing all raw sensor data *post hoc* to exclude participants with 1) excessive intermittent signal noise indicating poor skin contact or movement artifact, and 2) EMG signal amplitudes that consistently fail to exceed the noise level, indicating problems with attending to the task of articulating due to somnolence or distraction. One such subject was a clear outlier in this regard, and the data from his trials were excluded from further analysis. The entire protocol required approximately 3.5-4 hours for completion, including multiple stretch breaks to relieve fatigue and boredom. Fig. 1 shows a subject with laryngectomy self-operating the acquisition system.

B. Signal Processing and Recognition

We modified and enhanced our previously reported subvocal speech recognition algorithms for military applications [25][34] to accommodate a different (and substantially reduced) vocabulary set more appropriate for users with laryngectomy. Our goal was to also reduce the requisite number of sensors from the 8 locations acquired in our data corpus to a sub-set of 4 sensors to simplify usability. Details of the approach are divided into speech activity detection, feature extraction, and modelling components below.

1) Speech Activity Detection (SAD)—Speech Activity Detection is a vital component of an accurate silent speech recognition system. Our SAD algorithm was specially designed to address three main challenges unique to sEMG speech recognition: 1) unlike acoustic speech, SAD must simultaneously exploit multiple channels; 2) it must be able to distinguish between speech-related activity and non-speech-related activity (the difficulty of which is compounded by the fact that speech-related sEMG signals do not always immediately dissipate after articulation is completed); and 3) each speaker's sEMG signals are unique. To address these challenges, other studies have approached data-segmentation using hand labeling or acoustic cues in a voiced model [14][16][23][24] generated from simultaneously recorded acoustic data. This approach, however, is not applicable to our case, as we recorded silent, continuous data, without any acoustic cues to provide a realistic test faithful to the anticipated real-life application for mouthed speech recognition. These qualities of our experiments also preclude the use of a machine-learning based start/stop classifier as there are no ground-truth data to train a model.

Instead, we developed an sEMG-based Speech Activity Detection (SAD) algorithm that adopted a two-level finite state machine approach that is based on a previous MUTE application [34]. The goal of this SAD is to safely remove as much silence and noise as possible from the signals (without inadvertently truncating the speech-related activity) and allow the ASR system to model the rest. The SAD algorithm incorporates a multi-channel decision logic, which takes advantage of the fact that speech production typically involves the simultaneous activation of multiple muscles 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 the following principles: 1) using a short time-windowed signal (between 25ms and 50ms depending on the most effective setting for a given subject) to compute local statistics; 2) continual background noise and real signal statistics tracking on each channel; and 3) a global decision based on the best sub-set of all sEMG channels. The SAD algorithm operates on two levels of finite state machines. The first level consists of a finite state machine for each channel, which determines the channel's speech state. An active/inactive decision is made on each windowed time instance, *t*, by comparing current statistics with minimum background and maximum signal statistics. The higher-level machine, as described in [34], combines each channel's states to make the global start of speech (SoS) and end of speech (EoS) decision. This SAD algorithm continuously adapts to the background signal level and the speaker/utterance specific maximum energy level for each channel. For a full 8-channel sEMG sensor set, our empirical study identified that channel subset {1,5,7,8}, (see Fig. 1 for respective locations), is the most effective for SAD decision making and is used in our full sensor set experiments. Fig. 3 shows an example of the output of the SAD, in this case, for the production of "I'm not ready yet." Note that because the SAD looks for simultaneous multi-channel activity, isolated, single channel sEMG bursts (e.g. in Channel 7) do not trigger the speech onset detection. Further details about this SAD can be found in [34].

As one of our goals was to reduce the required sensor set to a maximum of 4 channels, the SAD needed to be adapted to the reduced channel availability. As such, during the reduced sensor set recognition experiments, the SAD was modified to operate on all of the available channels that were being tested.

2) sEMG Feature Extraction—Our approach employs Mel-frequency cepstral coefficient (MFCC) features as a baseline method of sEMG feature extraction because they proved effective in our MUTE development. The sEMG signals were first subject to DC offset removal as this is an artifact of the electronics and has no physiological significance. For feature extraction, a Hamming window was used with length and shift adapted for each speaker, followed by cepstral analysis resulting in a 7-dimension cepstral feature set. The algorithm to generate the MFCCs was modified to account for the characteristics of sEMG signals (e.g. smaller bandwidth than acoustic speech) [25], resulting in the use of 15 filterbanks. The mean and variance normalization was applied and then the delta cepstral features were computed. The 8 sEMG channel cepstral features were concatenated to form the final 112-dimension feature vector. The default setting used a window size of 50ms and window shift of 25ms. However, because window size and shift was found to have a significant impact on performance, we applied a speaker adaptive window size and frame rate (see Results for description).

Because the high-dimensional multi-channel sEMG feature sets are highly correlated and redundant, we applied the well-known heteroscedastic linear discriminant analysis (HLDA) feature dimension reduction technique and maximum likelihood linear transform (MLLT) feature adaptation to enhance the discriminative power of the feature set [34][36]. HLDA utilizes 3 left frames and 3 right frames as context. MLLT uses HMM tri-phone tied-state as classes. Thus, the 112-dimension input feature vector is augmented by a factor of 6 (to include the 3 left and right frames) prior to being transformed into a 30-dimension discriminative feature space. The resulting 30-dimension feature vectors are then used for training the recognition model.

3) sEMG Recognition Modeling—We designed a subvocal speech recognition algorithm that applies advances in the field of acoustic speech recognition to the silent speech recognition domain [16]. In our previous work [25][34] we developed the architecture to build phoneme-based recognition models through a series of processing stages, each designed to address specific aspects of the model complexity. In this study we improved upon our previous architecture by migrating our algorithm into the KALDI speech processing toolkit [37][38] and developing new subject-specific models that adapt to unique characteristics of subvocal speech that may vary across individuals after laryngectomy.

Our approach to subvocal speech processing is based on the recognition of the underlying combinations of phonemes that comprise different words. The algorithm starts by modelling each of the commonly used 39 phonemes for English (and one silent/noise mode) with a three-state left-to-right hidden Markov model. The first step trains a monophone model for each subject with a total of 120 states across the 40 phoneme models. Each state shares a mixture of Gaussian distributions with a total of 1800 distributions for the 112-dimension sEMG feature vector. The second step builds context-dependent phoneme models, called triphone models, to account for the impact of left and right phonemes on the center phoneme. A data driven decision tree (using the KALDI toolkit decision tree algorithm) is then used to cluster the triphone models for the observed training data and create triphone models for new phoneme combinations not seen during training. The final triphone system has 500 tied states. We then applied a subspace Gaussian mixture model (SGMM) approach, such that all phonetic states share a common GMM with varying means and mixture weights within the subspace [37]. This allows a more compact representation and improved performance on the relatively small amounts of training data available for subvocal speech recognition. For this study, the common GMM was trained with 200 mixtures. The final SGMM model had 800 leaves and 1200 sub-states. We then apply per-utterance adaptation using feature-space maximum likelihood linear regression (FMLLR) to better align the SGMM for each triphone model. By adapting to subject-specific changes of subvocal speech, the FMLLR algorithm increases the probability of recognizing variations of the same triphone across multiple subvocal utterances. This processing stage is critical for retaining accurate word-recognition while testing on words previously unseen in the training data. Recognition "scores" for each vocabulary word or utterance (for each subject) are tabulated from the HMMs and a decision process is applied to identify the highest scoring possibility. Performance is averaged and plotted in tabular form.

Although we did not use a statistical language model for the recognition system, we employed a finite state transducer (FST) grammar to constrain the decoding graph. The grammar was built using a combination of HTK [40], HTK data preparing (HDP) [41][42], and the KALDI toolkit. First, the HTK grammar tool HParse was used to generate a standard lattice format (SLF) grammar using the text of all the sentences in the test corpus. Second, the HDP tool was used to convert the SLF grammar into an AT&T FST format. To assume minimum knowledge about the test corpus, uniform weight was given to each arc out of a state. Lastly, the KALDI FST tool was used to determinize and minimize the grammar.

4) Phoneme Alignment—One of the larger challenges of silent speech recognition is how to accurately perform phoneme alignment without knowing the acoustic ground truth of the training data. We approached this problem using a technique employed in many acoustic speech recognition systems, whereby the acoustic phoneme level model can be trained using continuous utterances without phoneme level labels. The phone alignment is done automatically through the Expectation Maximization (EM) algorithm training of phoneme hidden Markov models. This technique typically works given enough data as long as the utterances are not too long. In our case, the training corpus is designed to cover most phoneme combinations and there are many very short utterances. By allowing silence at the beginning, ending, and anywhere between words, the EM algorithm starts with uniform segmentation of phonemes and iteratively converges to a proper segmentation for each utterance during the training process. Generally, we observed that recognition accuracy improved as the training progressed with EM iterations.

III. Results

A. Full sensor set results

We computed the WER for different combinations of analysis window size and overlap for each subject using the test sentences from the 980-sentence data corpus. The window size was varied from 30ms to 50ms (in 10 ms increments) and the overlap was varied from 15ms to 25ms (in 5ms increments). The WER for each analysis window/overlap combination is presented in Table 1, with the best combination for each subject shown in bold type. We found that for each subject, the WER was highly variable across different analysis window lengths and overlaps, in some cases varying by as much as 27 percentage points (i.e.: for Subject 4). By adapting the window/overlap on a per subject basis we achieved a recognition performance WER of 10.3%.

When averaged across all subjects, we observed the WER was associated with both the window length and overlap (Fig 4). On average windows with 25 ms overlap produced the highest WERs – ranging from 17.2% to 20.2% – with the greatest standard deviations – ranging from 8.8% to 11.5% – regardless of window length. When decreasing the overlap by 5 ms, both the average and standard deviation of the WER was significantly reduced to as little as 11.4%. Further decreasing the overlap to 15 ms yielded relatively slight increases in the average WER, indicating windows with 20 ms overlap were optimal for the subjects tested. When the window overlap was fixed at 20 ms, the average and standard deviation of the WER was inversely related to the window length, decreasing from 13.2±3.8% for a 50 ms window to 11.4±1.4% for a 30 ms window. Overall, the window/overlap pair of 30ms/ 20ms was the most effective for the largest number of subjects (3) providing a mean WER of 11.4%.

B. Reduced sensor set results

To evaluate the practical usefulness of our silent speech recognition system, we investigated the tradeoff between the number of sensors and the accuracy of word recognition. For each number of sensors, we tested various combinations of the sensor subsets to identify the combination that provided the best performance across all subjects (Fig 5). The optimal window size and overlap were used for each subject. We found that the reduction in the number of sensors was exponentially related to the increase in word error rate. More specifically, as the number of sensors decreased from 8 to 4, the mean WER increased by approximately 3.3%. However, as the number of sensors was further reduced to 2, the mean WER increased by 14.8 %. These data support the viability for reducing the number of sensors from 8 to 4 without incurring relatively large degradations in subvocal speech recognition performance.

We further evaluated the performance of different combinations of 4-sensor subsets, with special interest in the subset of sensors {5,6,7,8} which are all located on the face (see Fig. 1) to test the viability of a simpler and more robust interface design. The WERs of the top 5 performance subsets are shown in Table 2. on a per-subject basis. Overall, there was a small variation in performance of these 5 4-sensor subsets (a range of 2.4 percentage points), with the most effective being subset {2,5,6,8}, which consists of 3 facial sensors and one submental sensor. This subset was the most effective for the largest number of subjects (3) and generated the lowest mean WER of 13.6%. Table 2. also identifies the most effective subset for each subject, i.e. if a custom subset could be used for each subject. In this case the mean WER is 12.2%.

C. Effect of Data Corpus Size

We further analyzed subvocal speech recognition on 2 subjects (5 and 6) who generated an augmented data set with 280 additional sentences to quantify the effect of a larger data corpus on the recognition performance. We computed the WER for these subjects for the full sensor set after training on a total of 550, 690 and 830 sentences; the WERs for each subject are presented in Fig 6. It should be noted that the augmented corpus contains some overlap with the testing component of the main corpus. Nonetheless, Fig 6 depicts a clear inverse relationship between the WER and the number of sentences used for training. Specifically, increasing the training data from 550 to 830 sentences decreased the WER from 12.1% to 8.8% and 10.9% to 7.7% in subjects 5 and 6 respectively. These data give evidence that subvocal speech recognition can be substantially improved with the expansion of training data.

IV. Discussion

The results of this study suggest that sEMG-based speech recognition is a viable mode of communication for those who are living with laryngectomy. This work falls under one of the several use cases for biosignal based communication, namely restoring spoken communication [13]. When using the full 8-sensor set coupled with speaker dependent processing, we are able to achieve a mean WER of 10.3% on a vocabulary of 2,500 words. However, because sensors located on the ventromedial neck (sensors numbered 3 and 4 in this study) are recording from substantially altered anatomy in this patient population, we investigated the effects of eliminating these sensors, as well as an additional set of two, to reduce the total number of requisite sensors by a factor of 2 (from our original set of 8). The ideal subset would be localized near the mouth to achieve a single 4-sensor neural interface on the face with greater ease of use. We found that the most effective 4-sensor subset varied from subject to subject, with a corresponding mean WER of 12.2%. The mean best-performing subset {2,5,6,8} has 3 out of the 4 sensors located on the face, and generated a WER of 13.6%; only a 1.4%-point drop in performance from the personalized best mean. The face-only subset {5,6,7,8} produced a mean WER of 15.9%. Collectively, these findings indicate that a facial sensor grouping with one submental sensor {1 or 2} would provide the most effective overall solution for this group of subjects. Overall, these results demonstrate that in spite of the relatively small performance reduction between our original 8-sensor set, and the reduced 4-sensor set, the enhanced practicality and simplicity of a 4-sensor facially-worn neural interface supports its viability for further development.

Unsurprisingly the optimal 4-sensor set for individuals with laryngectomy is not the same as the optimal 4-sensor set for healthy individuals. As reported in [25], the best performing sensor subset consisted of sensors {1,3,6,7}, i.e. one sensor from each of four targeted areas of speech musculature (above and below the oral commissure, submental surface, and ventral neck surface). However, while sensors on the face were useful for both the healthy and laryngectomy cases, sensors located near the site of surgery {3,4} were of little value to subvocal speech recognition for individuals with laryngectomy, whereas sensors 2 and 5 gained in importance as possible substitutes. Yet, despite the difference in optimal subset configurations, the overall performance reduction for both sets of speakers are quite similar (4 percentage points for healthy speakers and 5.4 percentage points for speakers with laryngectomy).

It is also clear that increasing the amount of training data can significantly improve recognition performance. Although the sample size was small, we found that increasing the amount of training data by one third led to nearly a one third reduction in the WER across both subjects tested with the full sensor set.

Also of note is the relatively high degree of variability in recognition performance observed across the different window parameters. Figure 4 demonstrates that at least some of this variability is associated with window length and overlap. Changing the window overlap indicated that subvocal speech recognition is determined in part by the frame rate of input feature set. Even relatively small changes in overlap such as from 20 to 25 ms can double the frame rate resulting in dramatic increases in WER from among the best to the worst recognition performance (as can be seen for subject 4). Similarly, the window length dependence can be attributed to differences in articulation rate – should the window be too large with respect to the modulation of the sEMG signals, then the temporal resolution becomes too low to differentiate between subword units. Conversely, too small of a window reduces the frequency resolution, thus making the MFCC parameterization ineffective. While we could identify an optimal configuration for the analysis window parameters, there was enough variability among speakers however to make these configurations non-optimal for certain participants. This is consistent with similar findings reported for healthy speakers [25]. Thus a more thorough investigation is warranted to better assess the factors influencing variability in WER across subjects. Such an investigation would likely require a larger per subject data corpus, and would best be conducted with healthy participants more suited for extended experimental duration.

With respect to the sensor-set configuration, we found that while the mean best performing 4-sensor subset comes close to meeting the WER of the best per-subject subset, there are instances when there remains a performance gap between the two. This variability is likely caused by a combination of diverse articulatory strategies and differences in speaker anatomy. Speech production is a many-to-one problem [41],[44] whereby different speakers may employ various articulatory strategies to produce the same desired acoustic output. This suggests that the relative importance of different speech-production-related muscle groups can vary from subject to subject, thus making some sensors more important than others.

Anatomical differences between speakers, whether due to variable muscle size or differences in the amount of tissue located between the muscle and the surface sensor, can also affect the recorded signals, and hence the relative importance of different sensors. Anatomical differences can be exacerbated within the laryngectomy population because the residual anatomy is very much a function of the extent of the disease which prompted the surgery, as well as the surgical preferences of the surgeon performing the procedure. [45] As such, variability in optimal sensor-subsets is not a surprising finding.

We considered imprecision in sensor placement as another cause of inter-subject variability. However, as mentioned in Section II.A, our previous investigations demonstrated that the information content of the resulting sEMG signal was robust to shifts in sensor location provided the sensor was placed somewhere over the body of the target muscle [27]. To ensure proper and consistent sensor placement we used a set of templates combined with anatomical landmarks to guide sensor location. Furthermore, our experience in using the MUTE prototype [25] (which uses the same sensor configuration) in repeated demonstrations over the course of several years, has shown that the recognition is highly repeatable simply by using sensor templates as placement guides. As such we do not believe that a lack of sensor placement consistency is the cause of this performance variability.

Nevertheless, the relatively small magnitude of the inter-subject variability supports the potential for a reduced sensor system to provide practical and usable technology for subvocal speech recognition.

A. Limitations of silent speech recognition

The results of this study demonstrate that silent speech recognition has promise as an alternative communication device for persons living with laryngectomy. While the performance of our system surpasses that reported by previous investigations of subvocal speech, it still falls short of matching the recognition rates of commercial acoustic ASR systems; all of which operate on much larger continuous vocabularies. As such, future work must focus on identifying and addressing the causes of the recognition errors. To this end, we have identified three limitations that need to be overcome before this technology can be commercialized to the patient population.

1) Data corpus size—The performance of subvocal speech recognition is constrained by the limited training data available for each subject. There are a number of reasons for this situation. First, collecting large amounts of silent speech data is an intentionally slow process so that the data can be collected in the form of reasonable duration tokens that can be automatically aligned. Because it is impractical to manually label the silent speech data, relying on statistical alignment is a necessity. Our approach was to present the speech at sentence level tokens and allow the subjects to self-pace their presentation. This method introduces delays between the recitations of the speech tokens, which, over the course of roughly 1000 sentences effectively doubles the total time of the experiments. Expecting individuals in our study with an average age of 64 years to maintain a consistent level of mouthed speech effort over approximately 4 hours of data collection poses a challenge for protocol compliance. More accurate results approaching those of ASR could be obtained if data collection is spread across multiple, shorter recording sessions even if doing so introduces variability related to differences in sensor placement between recordings.

It should also be noted that other studies that have focused on collecting a large sEMG speech data set were only able to collect a total of about 1.75 hours' worth of silent speech data from each of 8 subjects over 32 short-duration sessions [46]. This amount pales in comparison to the thousands of hours of speech used to train acoustic, large vocabulary ASR systems, which may partially explain why sEMG-based ASR has yet to approximate acoustic-based ASR. Nevertheless, the significant performance gain that resulted from increasing our training corpus by 50% makes it clear that a more efficient means of collecting more training data needs to be developed.

2) Missing tongue information—The tongue is one of the most important articulation muscle groups. For certain classes of phonemes, placement of the tongue is the only feature that distinguishes between them. Yet the nature of the placement of sEMG sensors on the surface of the skin impedes direct access to the tongue musculature. It is possible that Sensor 2, which is located under the chin, is recording some amount of tongue activity; this is supported by the fact that Sensor 2 is the only non-facial sensor found in the most effective sensor subset. However, there is a significant amount of non-lingual musculature and tissues located between Sensor 2 and the tongue that may limit the amount of tongue related information that is being captured. As such, our subvocal recognition system might benefit from investigating other sensor modalities of accessing the state of the tongue during speech production.

A number of alternative modalities for subvocal speech recognition have been investigated that can better measure tongue activity. These include ultrasound [47] impulse radio ultra-wide band (IR-UWB) radar [48], and permanent magnet articulography (PMA) [50]. While these other modalities have shown promise, they either involve the use of large external sensors (ultrasound and IR-UWB) or intra-oral sensors that may not be durable in the long term (PMA). Additionally, our pilot work found that augmenting our subvocal recognition system using Transoral Impedance (TOI) [51][52]to measure lingual-palatal contact can improve recognition performance. However, like the other non-sEMG modalities, significant effort must be invested in reducing the size of the TOI sensors and integrating them into an sEMG-based silent speech system for viable, everyday use.

3) Accuracy of Activity Detection—One of the largest challenges facing sEMG-based silent speech recognition is the accurate detection of speech onset and offset. Incorrectly detecting speech onset and offset can cause alignment problems and ultimately recognition errors. Although our SAD algorithm was developed to be able to distinguish between speech and non-speech-related muscle activity, there are times when it is impossible to do so using a purely sEMG-based solution. Comparing ASR using SAD versus hand-segmentation of speech-related sEMG may reveal inadequacies of the SAD algorithm that, if addressed, could substantially improve ASR performance.

One known source of error for the SAD algorithm is that it assumes there will be “rest” periods, when sEMG activity subsides below a certain threshold, that indicate non-speech periods. However, we have found a number of speech tokens in which there is no obvious period of reduced sEMG activity. This situation is caused when the subjects do not return their articulators to a resting state between speech prompts on the screen, and thus generate a higher level of background sEMG activity. Providing sEMG activity biofeedback during token collection would likely improve subjects’ ability to relax their speech musculature between tokens, thereby improving EMG-based SAD.

Another possible remedy to this situation is to introduce an inertial measurement unit (IMU), which could be used to track jaw and/or lip movement. The IMU could be integrated into the sEMG sensors themselves, or placed in a more optimal location for jaw movement tracking. Practical considerations, including size, circuit complexity, and additional computational requirements would have to be considered before additional sensors could be successfully integrated into any practical silent speech recognition system.

B. Practical considerations

An important goal of this study was to begin assessing the translation of sEMG-based speech recognition into a practical communication device for persons living with laryngectomy. One step towards attaining this goal was to reduce the number of required sensors from 8 to 4. Our choice of aiming for a 4-sensor subset is essentially a compromise between reducing hardware complexity and maintaining high recognition accuracy. Our previous work on sensor reduction in healthy control subjects showed that recognition performance remained stable until fewer than 5 sensors were used [28]. We conducted a similar analysis on the data collected in this study to verify comparable behavior with an updated sensor type and with subjects who are persons with laryngectomy. We again found that a 4-sensor set produced a modest reduction in recognition accuracy (3 percentage points), whereas smaller subsets produced a precipitous drop in recognition performance. These results suggested that the decision came down to choosing between a 4 sensor or 5 sensor subset. While recognition accuracy considerations pointed towards using the 5-sensor set, we believed that the additional hardware complexity and anatomical surface area required by using 5 sensors was not worth the 1 percentage point performance gain. As such we pursued an investigation of identifying the optimal 4 sensor subset.

The results of the sensor subset study indicate that half of the full 8-sensor set can be eliminated while only reducing performance by a few percentage points. The WER data show that, ideally, the sensor subset would be customized for each individual. However, this would be impractical from a commercial point of view, at least with the current form of sEMG sensors being used. A patch with a large sensor array similar to the one used in [23] and [53] could be customized on a per-speaker basis, but it would introduce questions about durability, cost, and size that would need to be addressed. Instead, a comprise, in the form of the sensor set with the best mean performance ({2,5,6,7}) could be used instead. On average, the difference between the WER of this subset and the WER of the best per subject subset was 1.1%, a relatively small cost that could ultimately be overcome by improved algorithms, for the large benefit in hardware simplicity.

Another aspect to consider is the benefit of the individualized processing parameters (analysis window and overlap size). Currently, we use a trial-and-error approach, applying all possible window/overlap combinations and choosing the one that produces the smallest WER. However, this is not a practical approach for a commercial communication device. Instead, some means must be developed that can tune the processing algorithms based on speaking rate or properties of the sEMG signals that are being recorded. This will be a subject of future study.

Personalization can also play a role in the nature of the synthetic voice used in any future system. One of the main motivations for adapting subvocal speech recognition to the laryngectomy patient population is to provide another communication option beyond alaryngeal speech, which is typically described as unnatural and robotic [7][8]. Because the subvocal algorithms convert articulation into text, a text-to-speech (TTS) engine must be used to synthesize the speech from the recognized text. Generic TTS voices can be highly intelligible and gender appropriate, but would not sound like the users' natural voice lost after laryngectomy. Fortunately, there are now two options that could give laryngectomized users the ability to have a personalized synthetic voice. For those who have already undergone laryngectomy surgery, there now exists the technology to create personalized synthetic voices from small amounts of found audio (e.g. old video or audio recordings) [12]. On the other hand, informed cancer patients who have an adequate pre-surgical voice can strategically bank their voices [56]; the banked audio can then be used to create a custom synthetic voice. In either case, the custom synthetic voice could be integrated with the subvocal recognition system to provide a more personalized experience while preserving the user's vocal identity.

This potential personalization may ease the way for many persons living with laryngectomy to accept using an sEMG-based silent recognition system. As much as we have tried to reduce the amount of visible hardware (and will continue to make the hardware is minimal and inconspicuous as possible), the facial placements will be noticeable, and for those individuals who achieve satisfactory communication with their particular method(s) of alaryngeal speech, the bother of applying electrodes and the attention they draw might dissuade them adopting sEMG-based ASR. However, if a personalized TTS voice can be mated with this technology so as to return some semblance of their original voice to these individuals, they may be more willing to accept the drawbacks of an sEMG-based communication system.

C. Future work

This study aimed to demonstrate the viability of sEMG based communication for one of the several use cases for biosignal based communication, namely restoring spoken communication [13]. Our results support the use of an sEMG-based speech recognition system as the core of an alternative communication device for persons who have undergone laryngectomy. However, while the performance results are promising, they still must be improved before this system is viable for commercial use.

It is clear that system performance would benefit from the availability of more training data. As we showed in this study, even a small amount of additional training data (with some overlap in the testing and training sets) can produce significant performance improvements. However, because the current data collection protocol is a fairly slow process, new data collection methods must be developed to enable larger scale data collection sessions. One possible solution is to use data collected from several speakers to create a speaker independent model from several subjects' data as in [54] and adapt them to the target subject. However, our experience has shown that subject dependent models result in significantly better performance [55]. This is further supported by how much individual recognition performance is affective by varying the analysis window and overlap size.

Another possibility that would be viable in a portable practical system is to use a relatively small corpus, as we have used here, to create a baseline recognition model that performs relatively well and then continually adapts to the speaker, as he or she uses the system. This model is used in other applications, such as commercially available personal assistants.

It is also likely that our subvocal recognition system could be improved by integrating additional modalities to capture information that is not found in the sEMG signals. For example, incorporating a modality that can record information about tongue positioning and trajectories during speech production (e.g. TOI) would likely significantly reduce the WER. Further, it is worth investigating whether using IMUs to help detect the onset and offset of speech production can reduce SAD errors and ultimately improve recognition performance.

Finally, additional studies could also focus on integrating recent advancements in Deep Learning techniques with our front-end processing. Some recent studies have shown the potential benefits of using Deep Learning for sEMG-based speech recognition [23][24] but not for the specific population of laryngectomized speakers. Furthermore, Deep Learning methods are notoriously data hungry and would require the development of better data collection methods or, again, use many subjects' data to train an independent model that can be adapted to a target speaker. This will be a focus of our future studies in this area.

V. Conclusion

We have demonstrated the potential of using an sEMG-based silent speech recognition system as the basis of an alternative communication device for persons living with laryngectomy. When using our full 8-sensor set, our system produces a mean WER of 10.3% on a 2000-word vocabulary, whereas the best 4-sensor subset had a WER of 13.6%. To improve subvocal word recognition, we found that modest increases in the amount of training data for two subjects was successful at reducing their WERs below 9%. These performance metrics are comparable to those generated using healthy talkers on a similar-sized vocabulary (11% and 15% WERs for the full and 4 sensor subset cases, respectively) [25]. They also exceed the performance levels reported by other pertinent studies, i.e. 32% on 108 word vocabulary [16], 15% on the same 108 word vocabulary [22], and 20% on a 2100 word vocabulary (of sEMG signals recording during the production of vocalized speech) [57]. Although additional work needs to be conducted to improve recognition performance (via new algorithms and/or additional modalities) and to simplify the associated sensor hardware, this study provides a valuable proof-of-concept for the development of an sEMG-based alternative communication system.

Acknowledgments

Research reported in this publication was supported in part by the De Luca foundation and by the National Institute on Deafness and Other Communication Disorders (NIDCD) of the National Institutes of Health under award number R44DC014870.

Biographies

Geoffrey S. Meltzner received the B.S. degree in engineering from Harvey Mudd College, Claremont, CA, in 1995, the S.M. degree in electrical engineering and computer science from in 1998 and the Ph.D. speech and hearing bioscience and technology in 2003 from MIT, Cambridge, MA.

He joined BAE Systems (formerly Alphatech, Inc.), Burlington, MA in 2003. As the Section Lead of the Biological Acoustic, and Speech Signal processing group, he lead and managed a number of Defense Department funded programs in biological signal processing including the DARPA Mouthed-Speech Understanding and Transcription Engine (MUTE), Active Authentication program, and Reliable Central-Nervous System Interfaces (RCI) programs. He currently serves as the Vice President of Research and Technology at VocaliD, Inc. in Belmont, MA, which he joined in 2015. His research interests include non-traditional speech technologies, disordered speech technologies, voice conversion and personalized speech synthesis.

Dr. Meltzner is the recipient of the BAE Systems Chairman's Bronze Award for his work on the MUTE program. He serves as a reviewer for several journals, including Speech *Communication, Biomedical Engineering-Applications Basis & Communication, Augmentative and Alternative Communication, and PLOS One*.

James T. Heaton received the B.A. degree in psychology from Luther College in Decorah, IA, in 1990, and the M.S. and Ph.D. degrees in biopsychology (neuroscience) from the University of Maryland, College Park, MD, in 1993 and 1997, respectively.

He joined the faculty of Harvard Medical School, Boston, MA, in 1997 where he is currently an Associate Professor of Surgery, and a faculty affiliate of the Harvard doctoral program in Speech and Hearing Bioscience and Technology (SHBT). He is also Adjunct Professor in the Department of Communication Sciences and Disorders at the MGH Institute of Health Professions. Dr. Heaton has published over 75 peer-reviewed journal articles and several patents, and serves as an ad hoc reviewer for nineteen journals. His research interests include the neural control of head and neck structures involved in voice/speech production, functional electrical stimulation of the laryngeal and facial muscles, and creating an EMG-based interface for voice prosthesis control and speech recognition technology under funding from the NIH, DARPA and the Voice Health Institute.

Yunbin Deng (M'97) received the B.E. degree in control engineering from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 1997, the M.S. degree in electrical engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2000, and the M.S.E. and Ph.D. degrees in electrical and computer engineering from The Johns Hopkins University (JHU), Baltimore, MD, in 2002, and 2006, respectively.

He joined Intelligent Automation, Inc., Rockville, MD, as a Research Engineer in 2005. He worked as a Speech Scientist at LumenVox, LLC from 2006 to 2008. He is currently a Senior Principal Research Engineer at BAE Systems, Burlington, MA. His research interests include language and speech processing, robust and silent speech recognition, dialog systems, mixed-signal VLSI circuits and systems, Integrated circuit reverse engineering and variability/uncertainty quantification, biometrics and machine learning.

Dr. Deng won the Outstanding Overseas Chinese Student Award from the Chinese Scholarship Council in 2003. Dr. Deng received multiple BAE Systems Chairman's award for his contribution to DARPA research programs. He is a member of SPIE. He is a reviewer for the *IEEE Transactions On Audio, Speech, and Language Processing*, the IEEE *Transaction On Circuits And Systems*, and the *Circuits, Systems, and Signal Processing Journal*.

Gianluca De Luca received the B.A.Sc. in Electrical Engineering (Engineering Science) from the University of Toronto, Toronto Canada in 1995 and the M.S. in Electrical Engineering (Biomedical) from the University of New Brunswick, Fredericton NB Canada in 1997. He held positions at TransCanada Pipelines LTD (Toronto, ON) and the Bloorview MacMillan Rehabilitation Centre (Toronto ON) prior to joining Delsys Inc./Altec Inc. in 1997 where he currently holds the position of V.P. of Product Development. His interests are directed at developing wearable sensor technologies for health care and human performance applications. He has co-authored several patents on sensor technology and has been PI on numerous SBIR research grants which have been recognized with the SBA Tibbets Award.

Serge H. Roy, received the B.S. degree in physical therapy in 1975 from New York University, New York, NY USA and the M.S. and Sc.D degrees in Applied Anatomy and Kineisology from Boston University, Boston MA USA in 1981 and 1992 respectively.

He was previously employed as a full-time researcher at the NeuroMuscular Research Center (NMRC) at Boston University (1985-2015) where he attained a Research Professor appointment. Dr. Roy was also appointed as a Research Professor in the Department of Physical Therapy at Boston University's Sargent College of Health and Rehabilitation Science during this same time period. He served as a Senior Research Associate at Liberty Mutual Insurance Co, Boston, MA USA (1982-1984) and a Research Associate, Department Veterans Affairs (1988-2000). He is now the Director of Research at Altec/Delsys Inc. in Natick, MA.

Dr. Roy is the recipient of the Elizabeth C. Adams Award as the outstanding Graduate of N.Y.U. (1975) and two group achievement awards from NASA for experiments involving Space Shuttle Life-Science I and II missions (1992, 1994). He is a Fellow of the American Institute for Medical and Biological Engineering (AIMBE) since 1999 and was appointed as the President of the International Society of Electrophysiology and Kinesiology (2000-2002) and their first Fellow in 2004. Dr. Roy's primary research interests are directed at developing wearable sensor systems for automated monitoring of movement disorders, voiceless communication, and neural man-machine interfaces.

Joshua C. Kline received B.S (2009), M.S. (2012) and Ph.D. (2014) degrees in Biomedical Engineering from Boston University (Boston, MA USA) where he was trained in the fields of signal processing, algorithm development and neuromuscular physiology under the direction of Dr. Carlo J De Luca, former Director of the NeuroMuscular Research Center and Founder of Delsys, Inc. After completing a 1-year post-doc at the NeuroMuscular Research Center in 2014, he joined Delsys, Inc/Altec, Inc. full-time where he currently holds the position of Lead Research Engineer. At Delsys, Dr. Kline collaborates with a team of engineers and research scientists to design new technology for extracting neural information from biological signals for the advancement of neurophysiology and development of next-generation brain-machine interface technology. He has participated as PI and key person on numerous SBIR funded research projects that have been presented at major scientific conferences and disseminated through high-impact peer-reviewed publications.

References

1. Chen F, Kostov A. Optimization of dysarthric speech recognition. Proc IEEE EMBS Conf. 1997; 4:1436–1439.

2. Deller JR Jr, Hsu D, Ferrier LJ. On the use of hidden Markov modeling for recognition of dysarthric speech. Comput Methods Programs Biomed. 1991; 35:125–139. [PubMed: 1914451]
3. Green P, Carmichael J, Hatzi A, Enderby P, Hawley M, Mark P. Automatic speech recognition with sparse trainingdata for dysarthric speakers. European Conference on Speech Communication and Technology. 2003
4. Kotler A, Thomes-Stonell N. Effects of speech training on the accuracy of speech recognition for an individual with a speech impairment. Journal of Augmentative and Alternative Communication. 1997; 12:71–80.
5. Noyes JM, Frankish CR. Speech recognition technology for individuals with disabilities. Augmentative Alternative Commun. 1992; 8:297–303.
6. Polur PD, Miller GE. Experiments with fast Fourier transform, linear predictive and cepstral coefficients in dysarthric speech recognition algorithms using hidden Markov model. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2005; 13(4):558–561. [PubMed: 16425838]
7. Doyle, PC., Eadie, TL. The perceptual nature of alaryngeal voice and speech. In: Doyle, PC., Keith, RL., editors. Contemporary Considerations in the Treatment and Rehabilitation of Head and Neck Cancer. 2005. p. 113-140.
8. Hillman, RE., Walsh, MJ., Heaton, JT. Laryngectomy speech rehabilitation. In: Doyle, PC., Keith, RL., editors. Contemporary Considerations in the Treatment and Rehabilitation of Head and Neck Cancer. 2005. p. 75-90.
9. Kramp B, Dommerich S. Tracheostomy cannulas and voice prosthesis. Head and Neck Surgery. 2009; 8 ISSN 1865-1011.
10. Meltzner GS, Hillman RE. Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech. J Speech Lang Hear Res. Aug.2005 48:766–779. [PubMed: 16378472]
11. Meltzner, GS., Hillman, RE., Heaton, JT., Houston, KM., Kobler, J., Qi, Y. Electrolarynx Speech: The State of the Art and Future Directions for Development. In: Doyle, PC., Keith, RL., editors. Contemporary Considerations in the Treatment and Rehabilitation of Head and Neck Cancer. 2005. p. 545-570.
12. Mills T, Bunnell HT, Patel R. Towards personalized speech synthesis for augmentative and alternative communication. Augmentative and Alternative Communication. 2014; 30(3):226–236. [PubMed: 25025818]
13. Schultz T, Wand M, Hueber T, Krusienki DJ, Herff C, Brumberg J. Biosignal-based Spoken Communication – A Survey" Submitted to Special Issue on Biosignal-based Speech Communication, IEEE/ACM Transactions on Audio, Speech, and Language, Processing (T-ASLP).
14. Chan ADC, Englehart K, Hudgins B, Lovely DF. Myoelectric Signals to Augment Speech Recognition. Medical and Biological Engineering & Computing. 2001; 39:500–506. [PubMed: 11523740]
15. Betts B, Jorgensen C. Small Vocabulary Recognition Using Surface Electromyography in an Acoustically Harsh Environment. NASA TM-2005-21347. 2005
16. Jou SC, Maier-Hein L, Schultz T, Waibel A. Articulatory feature classification using surface electromyography. Proc ICASSP. 2006:606–608.
17. Maier-Hein L, Metze F, Schultz T, Waibel A. Session Independent Non-Audible Speech Recognition Using Surface Electromyography. IEEE Automatic Speech Recognition and Understanding Workshop. 2005:331–336.
18. Lee KS. EMG-Based Speech Recognition Using Hidden Markov Models With Global Control Variables. IEEE Trans On Biomed Eng. 2008; 55:930–940.
19. Jorgensen C, Lee DD, Agabon S. Sub auditory speech recognition based on EMG signals. Proc Int Joint Conf Neural Networks. 2003; 4:3128–3133.
20. Manabe H, Zhang Z. Multi-stream HMM for EMG-based speech recognition. Proc 26th Ann Int Con IEEE EMBS. 2004:4389–4392.
21. Schultz T, Wand M. Modeling coarticulation in EMG-based continuous speech recognition. Speech Comm. 2010; 52

22. Wand M, Schultz T. Session-independent EMG-based speech recognition. International Conference on Bio-inspired Systems and Signal Processing. 2011

23. Wand M, Schultz T. Pattern Learning with Deep Neural Networks in EMG-based Speech Recognition. Proc of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2014:4200–4203.

24. Wand M, Schmidhuber J. Deep Neural Network Frontend for Continuous EMG-based Speech Recognition. Proc of the 17th Annual Conference of the International Speech Communication Association (Interspeech). 2016

25. Deng Y, Heaton JT, Meltzner GS. Towards a Practical Silent Speech Recognition System. Proc of the 15h Annual Conference of the International Speech Communication Association (Interspeech). Sep.2014 :1164–1168.

26. Deng Y, Patel R, Heaton JT, Colby G, Gilmore D, Cabrera J, Roy SH, De Luca CJ, Meltzner GS. Disordered Speech Recognition Using Acoustic and sEMG Signals. Proc of the 10th Annual Conference of the International Speech Communication Association (Interspeech). 2009

27. Meltzner GS, Sroka J, Heaton JT, Gilmore LD, Colby G, Roy SH, Chen N, De Luca CJ. Speech recognition for vocalized and subvocal modes of production using surface EMG signals from the neck and face. Proc of the 10th Annual Conference of the International Speech Communication Association (Interspeech). 2008

28. Colby, G., Heaton, T., Gilmore, LD., Sroka, J., Deng, Y., Cabrera, J., Roy, S., De Luca, CJ., Meltzner, GS. Proc IEEE Int Conf Acoustics Speech and Signal Processing (ICASSP). Taipei, Taiwan: 2009. Sensor subset selection for surface electromyograpy based speech recognition.

29. Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS, Dahlgren NL, Zue V. TIMIT Acoustic-Phonetic Continuous Speech Corpus. Linguistic Data Consortium. 1993

30. Special Ops Hand Signals. Available: http://www.specialoperations.com/Focus/Tactics/Hand_Signals/default.htm

31. Most common 1000 English Phrases. Available; http://www.englishspeak.com/english-phrases.cfm

32. http://www.netlingo.com/acronyms.php

33. Costello, JM. Message Banking, Voice Banking and Legacy Messages: Boston Children’s Hospital. Available: https://www.childrenshospital.org/~/media/centers-and-services/programs/a_e/augmentative-communication-program/message-bank-definitions–vocab-nov.ashx?la=en

34. Meltzner, GS., Colby, G., Deng, Y., Heaton, JT. Signal acquisition and processing techniques for sEMG based silent speech recognition; 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society; Boston, MA. 2011. p. 4848-4851.

35. Kumar N, Andreou AG. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. Speech Communication. 1998; 26(4):283–297.

36. Gopinath R. Maximum likelihood modeling with Gaussian distributions for classification. Proc IEEE ICASSP. 1998; 2:661–664. 1998.

37. Povey D, Burget L, Agarwal LM, Akyazid P, Kaie F, Ghoshalf A, Glembekb O, Goelg N, Karafiátb M, Rastrowh A, Rosei RC, Schwarzb P, Thomas S. The subspace Gaussian mixture model: A structured model for speech recognition. Computer Speech & Language. Apr; 2011 25(2):404–439.

38. KALDI. Available: http://kaldi-asr.org/

39. Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y, Schwarz P, Silovsky J, Stemmer G, Vesely K. The Kaldi speech recognition toolkit. Proc IEEE ASRU. Dec.2011

40. http://htk.eng.cam.ac.uk/

41. http://stembep.wz.cz/fsm-howto/index-altpron_EN.html

42. Štemberk P. Speech recognition based on fsm and htk toolkits. Proceedings Digital Technologies. 2004

43. Atal BS, Chang JJ, Mathews MV, Tukey JW. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. J Acoust Soc Am. 1978; 63(5)

44. Schroeter J, Sondhi MM. Techniques for estimating vocal-tract shapes from the speech signal. IEEE Trns Speech and Audio Processing. 1994; 2(1):133–150.

45. Deschler, DG. Surgical Reconstruction Following Total Laryngetcomy with Extended or Total Pharygectomy. In: Doyle, PC., Keith, RL., editors. Contemporary Considerations in the Treatment and Rehabilitation of Head and Neck Cancer. 2005. p. 237-260.

46. Wand M, Janke M, Schultz T. The EMG-UKA Corpus for Electromyographic Speech Processing. Proc of the 15th Annual Conference of the International Speech Communication Association (Interspeech). 2014:1593–1597.

47. Hueber T, Benaroya EL, Chollet G, Denby B, Dreyfus G, Stone M. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. Speech Communication. 2010; 52(4):288–300.

48. Shin YH, Seo J. Towards Contactless Silent Speech Recognition Based on Detection of Active and Visible Articulators Using IR-UWB Radar. Sensors. 16(11):1812.

49. Wang J, Hahm S. Speaker-independent silent speech recognition with across-speaker articulatory normalization and speaker adaptive training. In Proc of INTERSPEECH. 2015

50. Gonzalez JA, Cheah LA, Gilbert JM, Bai J, Ell SR, Green PD, Moore RK. A silent speech system based on permanent magnet articulography and direct synthesis. Computer Speech & Language. 2016; 39:67–87.

51. Mundhe, A. Masters Thesis. MGH Institute of Health Professions; Boston, MA: 2014. Measuring Tongue Position and Movement Using Transoral Impedance.

52. Meltzner GS, Heaton JT, Deng Y. Augmenting sEMG-Based Speech Recognition by Non-Invasively Tracking Lingual Biomechanics. 7th World Congress of Biomechanics. 2014

53. Wand M, Schulte C, Janke M, Schultz T. Array-based Electromyographic Silent Speech Interface. Proc of the 6th International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS). 2013

54. Wand M, Schultz T. Session-Independent EMG-based Speech Recognition. Proc of the 4th International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS). 2011

55. “Subvocal Speech Exploitation – Final Report” Submitted to the Defense Advanced Research Projects Administration (DARPA), Contract #W15P7T-06-C-P437, January 2008

56. Patel, R. Distributed collection and processing of voice bank data. US Patent 9 336 782. Jun 29. 2015

57. Wand M, Schultz T. Towards Real-life Application of EMG-based Speech Recognition by using Unsupervised Adaptation. Proc of the 15h Annual Conference of the International Speech Communication Association (Interspeech). 2014:1189–1193.

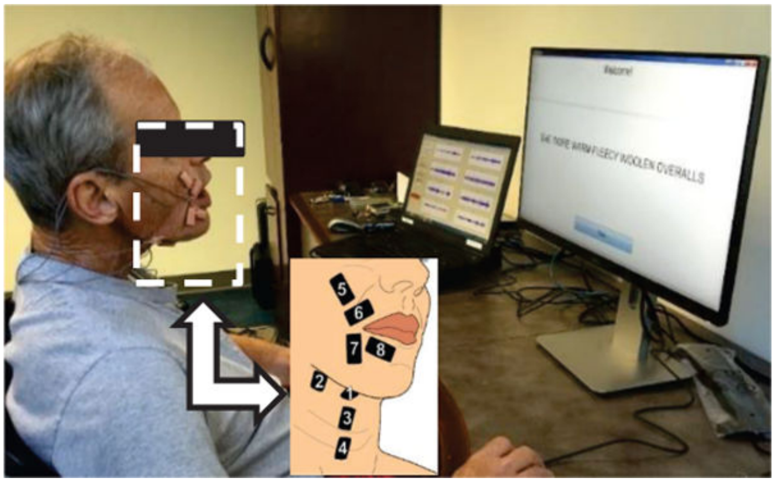

Fig. 1.

Subject with laryngectomy operating the data acquisition system. One screen displays sEMG signals, the other displays sentence prompts. The callout shows the sensor locations and their numbering scheme.

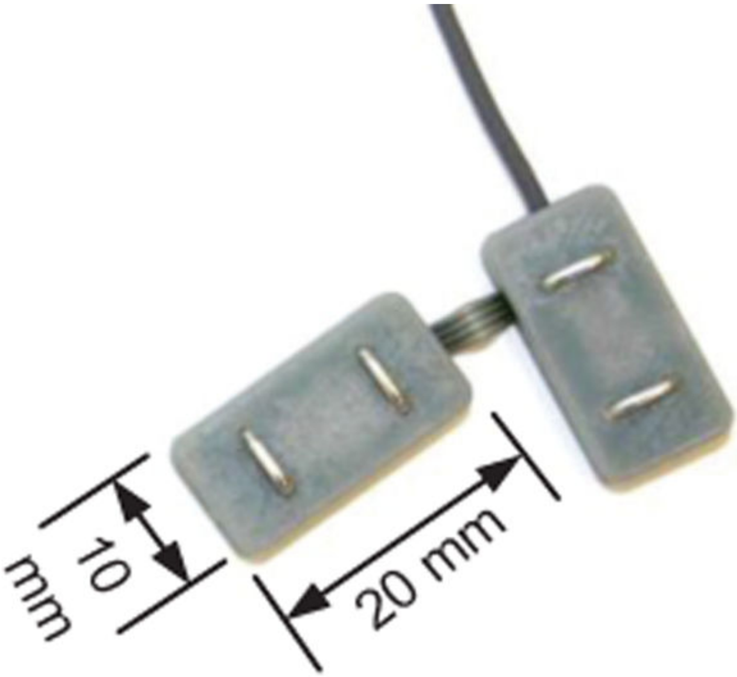

Fig 2.

One pair of the custom sEMG sensors. Joining pairs of sensors helps reduce errors in sensor placement on the face and neck.

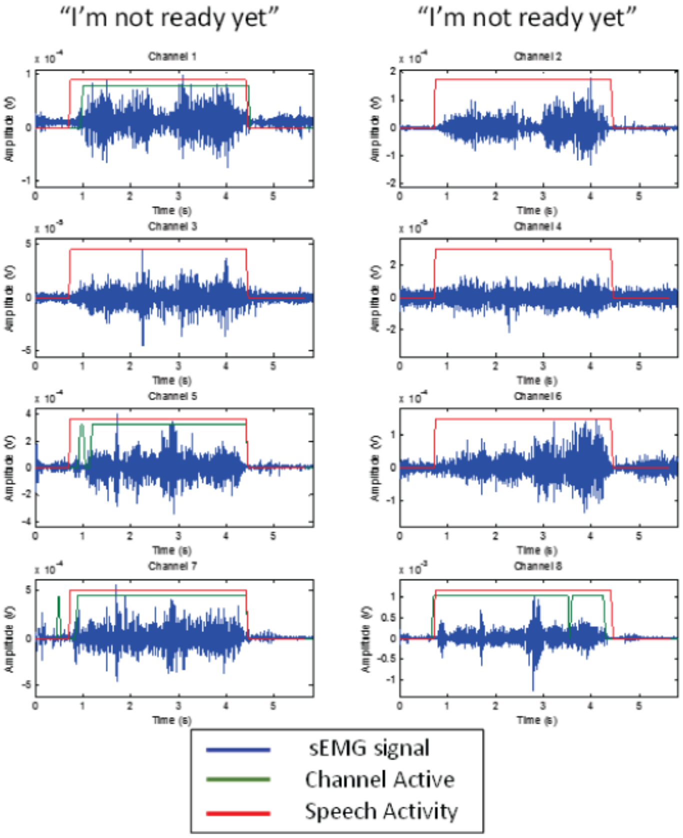

Fig. 3.

An example of 8 channels of sEMG data collected from Subject 3 during the mouthing of "I'm not ready yet.". The channel numbers correspond with the sensor locations shown in Fig. 1. Also shown is the SAD (see below) in action. The green line shows times when the individual channel is active, while the red line shows when speech activity is detected. Because the SAD looks for simultaneous multi-channel activity, small, single channel bursts of sEMG activity do not trigger the SAD. Note that because only channels {1,5,7,8} are used in the SAD, channel activity is only shown for those channels.

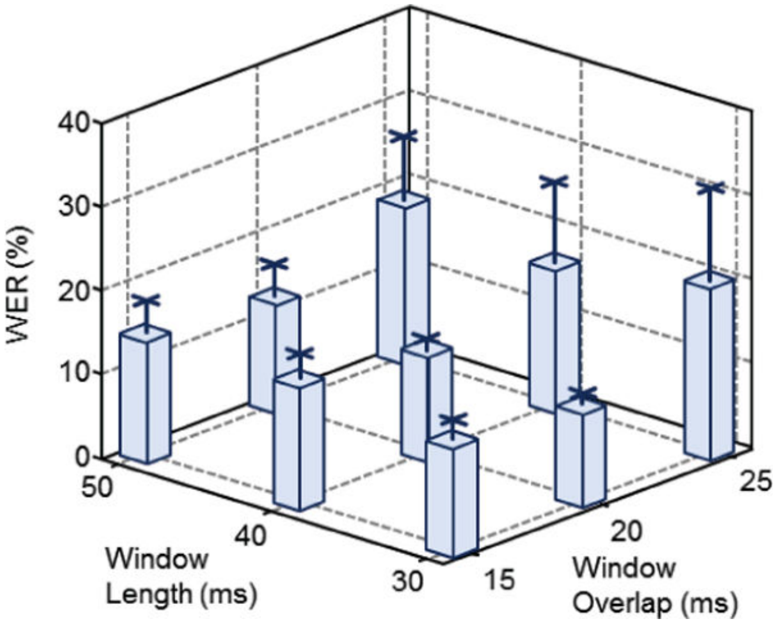

Fig 4.

Average WER across all subjects plotted as a function of the window length and overlap. Error bars indicate the standard deviation for each window length/overlap combination.

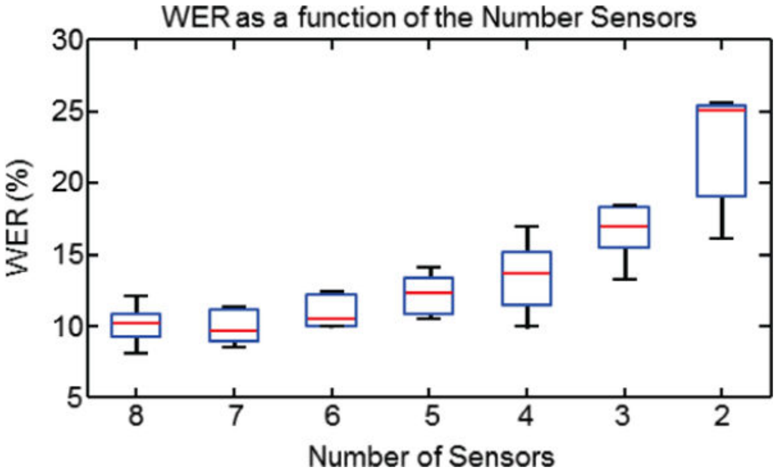

Fig 5.

WER plotted as a function of the number of sensors used for subvocal speech recognition. Each data point reflects the WER statistics computed for the same sensor subset across all subjects: red lines indicate mean values, blue boxes indicate first and third quartiles and black lines extend to the upper and lower 95th percentiles.

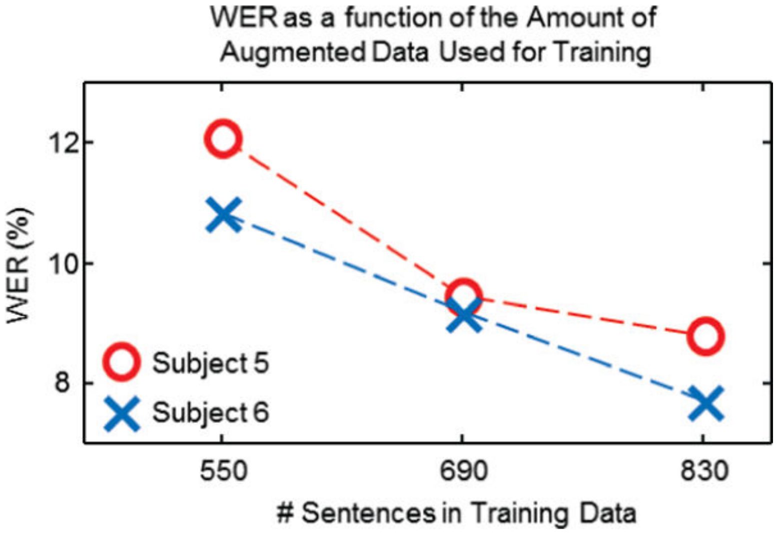

Fig 6.

WER for plotted as a function of the number of sentences in the training data using all 8 sensors for subjects 5 and 6 in red and blue, respectively.

Table 1

WER (%) as a Function of The Analysis Window

	Analysis Window Size, Overlap (ms)								
Subject	30,15	30,20	30,25	40,15	40,20	40,25	50,15	50,20	50,25	**Best**
1	10.68	9.20	9.53	14.87	10.27	9.86	12.33	8.79	**8.13**	**8.13**
2	16.52	**10.85**	14.46	20.71	14.30	14.46	20.79	19.39	17.17	**10.85**
3	10.82	13.06	35.41	12.98	13.64	34.51	**10.60**	14.95	33.11	**10.60**
4	14.87	**10.60**	37.47	13.06	11.42	25.23	13.97	12.57	26.38	**10.60**
5	14.95	**12.08**	17.75	15.69	12.65	12.33	15.78	12.65	12.33	**12.08**
6	**10.85**	12.90	15.69	12.49	11.67	14.79	14.22	15.53	13.97	**10.85**
7	11.34	11.01	10.85	12.08	11.67	9.20	9.94	**8.79**	13.97	**8.79**
Mean	12.86	**11.39**	20.17	14.55	12.23	17.20	13.95	13.24	17.87	**10.27**
SD.	2.49	**1.38**	11.48	3.01	1.39	9.30	3.65	3.79	8.77	**1.35**

TABLE 2

WER (%) for the Top Five, 4-Sensor Subsets

			{4-sensor subsets}			
Subject	{1,5,6,7}	{1,5,6,8}	{2,5,6,7}	{2,5,6,8}	{5,6,7,8}	**Best**
1	10.19	10.02	10.27	**10.02**	10.11	**10.02**
2	13.97	11.59	14.63	**11.09**	16.27	**11.09**
3	17.34	14.79	**12.82**	15.12	18.73	**12.82**
4	20.87	19.23	**14.87**	17.01	20.95	**14.87**
5	20.54	15.69	18.49	**12.57**	18.16	**12.57**
6	14.22	16.27	14.22	15.20	**13.64**	**13.64**
7	**10.02**	12.49	14.46	14.13	13.56	**10.02**
Mean	15.31	14.30	14.25	**13.59**	15.92	**12.15**
SD	4.46	3.14	2.46	**2.48**	3.72	**1.85**