Published in final edited form as:

J Neural Eng. 2018 August ; 15(4): 046031. doi:10.1088/1741-2552/aac965.

Development of sEMG sensors and algorithms for silent speech recognition

Geoffrey S Meltzner[2], **James T Heaton**[3], **Yunbin Deng**[4], **Gianluca De Luca**[1], **Serge H Roy**[1], and **Joshua C Kline**[1]

[1]Delsys, Inc and Altec, Inc, 23 Strathmore Rd, Natick, MA 01760, United States of America

[2]VocaliD, Inc. 50 Leonard St, Belmont, MA 02478, United States of America

[3]Harvard Medical School Department of Surgery, Massachusetts General Hospital, Boston, United States of America

[4]BAE Systems Inc, Burlington, MA, United States of America

Abstract

Objective.—Speech is among the most natural forms of human communication, thereby offering an attractive modality for human–machine interaction through automatic speech recognition (ASR). However, the limitations of ASR—including degradation in the presence of ambient noise, limited privacy and poor accessibility for those with significant speech disorders—have motivated the need for alternative non-acoustic modalities of subvocal or silent speech recognition (SSR).

Approach.—We have developed a new system of face- and neck-worn sensors and signal processing algorithms that are capable of recognizing silently mouthed words and phrases entirely from the surface electromyographic (sEMG) signals recorded from muscles of the face and neck that are involved in the production of speech. The algorithms were strategically developed by evolving speech recognition models: first for recognizing isolated words by extracting speech-related features from sEMG signals, then for recognizing sequences of words from patterns of sEMG signals using grammar models, and finally for recognizing a vocabulary of previously untrained words using phoneme-based models. The final recognition algorithms were integrated with specially designed multi-point, miniaturized sensors that can be arranged in flexible geometries to record high-fidelity sEMG signal measurements from small articulator muscles of the face and neck.

Main results.—We tested the system of sensors and algorithms during a series of subvocal speech experiments involving more than 1200 phrases generated from a 2200-word vocabulary and achieved an 8.9%-word error rate (91.1% recognition rate), far surpassing previous attempts in the field.

Significance.—These results demonstrate the viability of our system as an alternative modality of communication for a multitude of applications including: persons with speech impairments

jkline@delsys.com.

Disclosures
This work was funded in part by Delsys Inc.

following a laryngectomy; military personnel requiring hands-free covert communication; or the consumer in need of privacy while speaking on a mobile phone in public.

Keywords

electromyography; speech recognition; facial prosthesis; alternative communication; human–machine interface

Introduction

Speech, being perhaps the most natural form of human communication, is an attractive modality for human–machine interaction (HMI). This quality has fueled a large amount of research in the field of automatic speech recognition (ASR). Recent advances in ASR have enabled the proliferation of personal assistants, such as Siri, Alexa, and Cortana, and have paved the way for speech-based HMI to have a greater role in people's everyday lives.

Despite the improved performance of commercial ASR systems, there remain three primary limitations of the technology: (1) severe performance degradation in the presence of ambient noise; (2) a limited ability to maintain privacy/secrecy; and (3) poor accessibility for those with significant speech disorders. These deficiencies have motivated the emergent non-acoustic field of subvocal or silent speech recognition (SSR). SSR aims to recognize speech-related information using alternative modalities, such as surface electromyography (sEMG) that can capture sufficient speech information while overcoming the aforementioned deficiencies of acoustic ASR systems. sEMG-based speech recognition operates on signals recorded from a set of sEMG sensors that are strategically located on the neck and face to measure muscle activity associated with the phonation, resonation and articulation of speech. Because sEMG signals are a direct measurement of articulatory muscle activity there is no need for acoustic excitation of the vocal tract, making it possible to recognize silently mouthed speech. This non-reliance on vocalization makes SSR immune to acoustic noise corruption, affords privacy and provides an ideal candidate for recognizing the speech of those who cannot vocalize properly, if at all.

Although research into sEMG-based SSR is not as extensive as that of ASR, several studies have been conducted to advance the state of subvocal speech technologies (for a comprehensive review of SSR technologies see Schultz et al (2017)). Perhaps the earliest notion of using sEMG signals for speech recognition was conducted by Morse and O'Brien (1986) who demonstrated that these signals contain speech-related information. Chan et al (2001) later conducted one of the first subvocal recognition experiments by obtaining a 7% word error rate (WER) on a vocabulary of 10 digits using 5 sEMG signals from sensors placed on the face and neck while subjects vocalized speech. Betts and Jorgensen (2005) later conducted a similar study on a single speaker but were able to achieve only a 27% WER, albeit on a larger vocabulary of 15 words of vocalized speech. Jou et al (2006) further extended the vocabulary size to 108 words but at the cost of reduced recognition accuracy (32% WER). Lee (2008) was able to achieve a mean 13% WER on 60 vocalized words for 8 male, Korean speakers. Schultz and Wand (2010) pursued sEMG-based recognition of continuous speech and achieved a WER of 15.3% on a vocabulary of 108 words (Wand and Schultz 2011).

While these and other studies have established the potential to overcome the limitations of standard acoustic based ASR systems, several challenges remain to advance sEMG-based SSR technology for practical use outside of controlled laboratory testing. These challenges can be divided into three broad categories: (1) extracting speech-related features from sEMG signals to discriminate between isolated words; (2) identifying grammatical context of sequences of words from patterns of sEMG activity during continuous speech; and (3) modeling algorithms that operate at the phoneme-level of speech to recognize previously unseen words from a relatively large vocabulary set. In this work we designed, developed and empirically evaluated a new SSR system that overcomes these three challenges. We strategically evolved subvocal speech recognition algorithms first by testing speech-related features for recognizing isolated words, subsequently by incorporating grammar models for recognizing continuous phrases, and ultimately by designing phoneme-based models using pattern recognition algorithms to recognize previously unseen vocabulary from a relatively large data corpus. Our final SSR system was successful at recognizing a 2200-word vocabulary of more than 1200 continuous phrases with a 8.9% word error rate (91.1% recognition rate). Combined with our recently developed state-of-the-art facially-worn sEMG sensors, the subvocal speech recognition system provides a new alternative communication device for a broad spectrum of consumer and healthcare applications.

Methods

To develop and test a new sEMG-based speech recognition system, we designed experiments to record sEMG signals from articulator muscles of the face and neck during subvocal speech tasks. A total of 19 subjects (11 females, 8 males), ranging in age from 20–42, years (mean = 27.5) participated in the experiments. All subjects were native American English speakers and had no known history of speech or hearing disabilities. All participants voluntarily provided written informed consent, approved by the Western Institutional Review Board, prior to their participation.

sEMG recording

sEMG signals were recorded from speech articulator muscles, first using tethered DE 2.1 sensors (Delsys, Inc, Natick, USA) to acquire the data for Corpus 1 and 2, and then using prototype wireless Trigno™ Mini sensors (Delsys, Inc, Natick, USA) specifically designed to provide a miniaturized sensor interface that better conforms to the face and neck and mitigates sources of movement artifact and sweat-build-up during the relatively longer experiments needed for Corpus 3 (refer to Data Corpus descriptions below). All sensor sites were initially cleansed using alcohol pads, followed by removal of dead skin and oils, to mitigate sources of sEMG noise at the skin electrode interface. For some subjects, additional mitigation steps were needed to shave coarse facial hair and remove excessively dry skin by repeated hypoallergenic tape peels. Both sensors use parallel bar electrode configurations with 1 cm record a low-noise (<0.75 μV per channel) differential sEMG signal that is filtered from 20–450 Hz (>40 dB/dec) and sampled with a 16-bit resolution at 20 kHz using a Bagnoli™ sEMG amplifier for the DE 2.1 sensors or 1926 Hz for the Trigno™ Mini sensors. A total of 11 sEMG sensors were fixed to the skin surface. The sensors were placed over target muscles reported to participate in speech production: 7 sensors were placed on the neck and chin (#1–7 submental) and 4 sensors were placed on the face (#8–11) (Jorgensen et al 2003, Manabe 2003, Manabe and Zhang 2004, Jorgensen and Binstead 2005, Maier-Hein et al 2005) as described in table 1.

Protocol

Subjects participated in a series of experiments in which they were asked to recite words displayed on a computer monitor either vocally or subvocally while sEMG signals were recorded from speech articulator muscles and simultaneously acoustic signals that were recorded using a headset microphone (WH30, Shure Inc., Niles, USA) and sampled at 20 kHz using a 32 channel A/D converter (NI-6259, National Instruments Co., Austin, USA). Each experimental session lasted approximately 4–6 h, with adequate breaks given between recordings to avoid subject fatigue. For each experiment, a custom MATLAB (Mathworks, Natick, USA) Graphical User Interface was used to present speech tokens to the subject in a randomized order from one of three language data sets:

Corpus 1—Isolated words were used to test speech-related features extracted from sEMG signals for word recognition. A total of n = 9 subjects were presented with a set of 65 individual words including numbers 0–10, and various nouns and verbs that are commonly used for person-to-person communication and computer/device control (e.g. common replies, commands, locations, distances, times, etc). Each subject recited the entire 65-word set in pseudo-randomized order three times in both vocalized—using normal speech production—and silently mouthed—with articulation but no acoustic vocalization—manners.

Corpus 2—Small vocabulary sequences of words were important for modeling the grammatical context of subvocal speech from patterns of sEMG signals. A total of 4 subjects were presented with 1200 word-sequences (mean words/phrase = 6.0) generated from a relatively small 202-word vocabulary, including words and numbers associated with a special operations silent-speech data set (US Army 2009), the NATO alphabet and commonly used English phrases (EnglishSpeak 2017). Each subject recited the 1200-sequences in subvocal speaking mode.

Corpus 3—Large vocabulary continuous speech was used to design, develop and test phoneme-based models for recognizing previously untrained words from continuous subvocal speech. The ability to recognize unseen words is necessary for learning new vocabularies without burdening the users with arduous hours-long training sessions. In total, 6 subjects were presented with nearly 1200 phrases (mean words/phrase = 6.4) generated from a 2200-word vocabulary derived from the SX part of TIMIT data set (Garofolo et al 1993), a special operations silent-speech data set (US Army 2009), commonly used English phrases (EnglishSpeak 2017), a text-messaging set (NetLingo 2017), and a custom data set to cover phone calls, numbers, and dates for digit recognition. Each subject recited the 1200-word phrases in subvocal speaking mode.

To ensure that all subjects complied with the protocol, all raw sensor data were thoroughly reviewed to identify participants with: (1) excessive intermittent signal noise that indicated poor skin contact or movement artifact, and (2) sEMG signal amplitudes that consistently failed to exceed the noise level, indicating problems with attending to the articulating task due to somnolence or distraction. Three participants were non-compliant in this regard and were excluded from the study.

Speech activity detection

The first step in developing an SSR algorithm is to separate sEMG signals associated with each speech token from extraneous signals related to non-speech functions. Unlike acoustic data which are directly associated with speech production, sEMG data from articulator muscles may change with a variety of factors including extraneous facial expressions, anticipatory responses of articulator muscles preparing the vocal tract, or sustaining muscle contractions post-speech. Prior studies of sEMG speech recognition identify speech-related sEMG activity in an offline fashion where the sEMG signals are processed only after data collection (Jorgensen et al 2003, Manabe and Zhang 2004, Lee 2008, Meltzner et al 2008). As our goal was to develop a practical system for use outside of the laboratory, we implemented a real-time finite state algorithm to identify the start and end of speech-related content in the sEMG signals as they are being acquired (Meltzner et al 2008). Leveraging the fact that speech typically involves simultaneous activity of multiple muscles, the algorithm was designed to identify speech activity in two stages described by Meltzner et al (2008): (1) speech activity was first identified locally on a per-channel basis and (2) all local channel-based detections were converted into a global activity decision. The local channel-based detection algorithm consisted of a finite state machine that made an active/inactive decision on each windowed time instance, *t*, by detecting an increase in the amplitude of the sEMG envelope—filtered using a 40 ms window with 20 ms overlap—above the level of background noise. The time point at which the signal exceeded the threshold was marked as the beginning of sEMG activity while the time point at which the sEMG envelope last exceeded the threshold was marked as the end of sEMG activity (figure 1). The local channel-based detections were then provided to the global level machine to make the final start-of-speech and end-of-speech decision based on the presence and location of speech detections in each channel. The final algorithm continuously adapted to the level of background signal activity as well as the specific level of speaker/utterance maximum energy for each channel. In so doing the algorithm was able to mitigate noise in any one channel from biasing the total result while also remaining robust to variance in the activity of different muscles that may occur naturally for different words and phrases.

Extracting features for recognizing isolated words

The first goal in developing subvocal speech algorithms was to solve the challenge of extracting a set of sEMG-based features for discriminating between different words. Acoustic based ASR has a firm theoretical basis for using mel-frequency cepstral coefficients (MFCCs) for recognizing speech from acoustic signals as they provide a compact representation of the important features of the speech frequency spectrum. However, the same has not necessarily been proven for sEMG-based SSR. Therefore, we compiled a broad set of metrics and features derived from all 11 channels of the sEMG signal and compared the ability of each feature to provide discriminable information for recognizing isolated words. The features included:

1. MFCCs of the sEMG signal were used to capture characteristics of the short-term frequency spectrum. We calculated 6 MFCCs from sEMG signals spanning the bandwidth of 0 Hz to 2.5 kHz. The MFCCS were computed by first filtering the Fourier Transform of the signal windowed at 50 ms with a 25 ms frame rate through a set of Mel-scale filter banks. A nonlinear power compression algorithm was applied to the output of the filter bank which was then followed by the application of a Discrete-Cosine Transform.

2. Wavelet denoising was used to investigate whether feature extraction can be improved if coupled with MFCC features. Wavelet schemes have been attempted as a replacement for MFCCs in speech recognition (Kim et al 2000, Farooq and Datta 2001) and have been used extensively to denoise speech signals (e.g. Wickerhauser 1994). We used the Stein Unbiased Risk Estimate with Daubechies type 6 wavelets and different numbers of decomposition levels ranging from 1 to 3.

3. The root mean square (RMS) value of the raw sEMG signals was used to represent overall muscle activity.

4. The dominant frequency component of the sEMG envelope was used to characterize the primary periodic elements within the sEMG signals.

5. The amplitude range of auto-covariance function of the sEMG envelope was used to provide a measure of signal modulation.

6. The co-activation intervals between pairs of sEMG channels was used to measure the amount of simultaneous activity between all possible pairs of sEMG channels.

7. The zero-crossing rate (the number of times the sEMG signal crosses the *x*-axis in a given analysis frame) was used to characterize temporal changes in the signal based on previous subvocal recognition work (Jou et al 2006).

The extracted features provided inputs to hidden-Markov-models (HMMs) designed for isolated word-recognition using the HMM ToolKit (HTK). Each HMM had ten states (eight emitting, two degenerate)—with a left-to-right transition matrix— and a single Gaussian distribution to model each parameter. The models were trained on two out of three instances of each of the 65 isolated words in each speaking mode in Corpus 1 and tested on the remaining instance of each word. Separate models were trained and tested for each speaker.

Modeling grammatical context to recognize sequences of words

A practical SSR system needs to be able to recognize continuous sequences of words that may be confounded by limited pauses or transitions between surrounding vocabulary. To develop continuous speech recognition capabilities, we incorporated linguistic knowledge from grammar models into our algorithms to inform probable sequences of isolated words based on learned patterns of sEMG signals. The grammar models were constructed to allow most articles, adjectives, and negations to be optional to expand the word network more fully, while still maintaining linguistic correctness. Grammar rules were generated by substituting each word in the original set of rules with its more general Natural Language (NL) class to which it belongs. For example, the word ‘WHO’ in the grammatical phrase ‘WHO ARE THEY’ was replaced with the NL-class variable ‘$QUESTIONS’ to generalize the grammar to allow ‘(WHO | WHAT | WHERE | WHICH | WHEN | WHY | HOW) ARE THEY’. Other words were similarly substituted and tested with different NL-class variables. Overall, four different grammar models were tested: sentence construction grammar, unrestricted NATO grammar, NL equivalence grammar, and unigram grammar. The different grammar models were implemented using the HTK library and coupled with HMM algorithms for recognizing sequences of words using MFCC feature inputs, 10 HMM states and 2 Gaussian mixtures/state. The grammar models were trained and tested using the word-sequences derived from the relatively small 202-word vocabulary in Corpus 2.

Phoneme recognition of unseen continuous speech

While word-based models provided an opportunity to incrementally test feature sets and grammatical context for SSR, they are capable of only recognizing vocabulary that has been previously used for training, ultimately limiting the practical application outside of controlled laboratory testing. To overcome this challenge, we designed an improved recognition engine that operates at the phoneme level to classify combinations of phonemes of previously unseen words. For each of the 37 phonemes covered by Corpus 2, we implemented 5-state left-to-right HMMs using HTK tools, tested different numbers reduced features and Gaussian mixtures, and evaluated the performance on sequences of words from the relatively small 202-word vocabulary in Corpus 2.

We implemented a linear discriminate analysis (LDA) (Saito and Coifman 1995) algorithm to test different numbers of reduced features by transforming the high dimensional set of 154 MFCC features generated from the 11 sEMG channels to a lower dimensional feature space. Feature reduction is commonly used in general pattern recognition applications to improve classification performance, reduce processing time and decrease memory allocation. Previous studies have successfully applied different versions of LDA for SSR (Zhou et al 2009), albeit towards a simpler set of vocabulary or phrases than those included in this study. For our application, we employed a heteroscedastic linear discriminate analysis (HLDA) which is a generalization of the standard LDA that does not require identical within-class covariance matrices. Using the HLDA transform we mapped an initial n-dimensional space to a p-dimensional ($p < n$) space by finding the maximum likelihood optimization function (Kumar 1997):

$$Q\left(\theta, \hat{\theta}\right) = \sum_{m \in M} \gamma_m(\tau) \times \log\left(\frac{\hat{A}^2}{\left|diag\left(\hat{A}_p W^{(m)} \hat{A}_p^T\right)\right| \left|diag\left(\hat{A}_{n-p} T \hat{A}_{n-p}^T\right)\right|}\right) \quad (1)$$

where θ stands for the older model, θ stands for is the newer model, $\hat{A}$ is the inverse transformation matrix, $\hat{A}_p$ and $\hat{A}_{n-p}$ are the corresponding first p and remaining $n - p$ row and where:

$$\gamma_m(\tau) = p\left(q_m(\tau) \middle| \theta, O_T\right) \quad (2)$$

$$W^{(m)} = \frac{\sum_\tau \gamma_m(\tau) \cdot \left(o(\tau) - \hat{\mu}^{(m)}\right) \cdot \left(o(\tau) - \hat{\mu}^{(m)}\right)^T}{\sum_\tau \gamma_m(\tau)} \quad (3)$$

$$\hat{\mu}^{(m)} = \frac{\sum_\tau \gamma_m(\tau) \cdot o(\tau)^T}{\sum_\tau \gamma_m(\tau)} \quad (4)$$

$$T = \frac{1}{T} \sum_\tau \left(o(\tau) - \hat{\mu}^{(g)}\right) \cdot \left(o(\tau) - \hat{\mu}^{(g)}\right)^T \quad (5)$$

where $\hat{\mu}^{(g)}$ global mean of data and are the $\gamma_m(\tau)$ is the probability of being Gaussian component m at time τ given the older model and observation feature sequence O_T. We optimized the cost function using a standard expectation maximization (EM) algorithm with a maximum number of iterations to avoid over training.

Using the reduced number of features provided by the HLDA output we trained context-independent single-mixture 5-state HMM phoneme models together with a 7-state silence model with the EM algorithm. We further added a short pause model, which is tied to the middle state of the silence model and allows state transition between dummy states with no feature output, because the short pause may or may not exist for each word during continuous speech. Next, we conducted a forced alignment at the phoneme level to produce the most likely phoneme sequence for a word, thus achieving a more accurate model for each word with multiple possible pronunciations. We then retrained the phoneme model after forced alignment with the EM algorithm, and incrementally increased the number of mixtures of Gaussians per state by a factor of two at each step. The EM algorithm was applied after each incremental increase of Gaussian mixtures until the phoneme models reached 16 Gaussian mixtures per state. Lastly, we trained the HLDA transform with a given number of reduced feature dimensions before retraining the phoneme model with the reduced feature set. The models were trained and tested using the sequences of words from the relatively small 202-word vocabulary in Corpus 2.

Using the newly designed phoneme recognition models, we migrated from the relatively small vocabulary of sequenced words in Corpus 2 to the much larger vocabulary of continuous phrases in Corpus 3 to test the ability to recognize words that were previously unseen during model training. To solve the challenge of building robust models which generalize well to unforeseen words while maintaining functionality on a limited set of subject-specific training data, we migrated our baseline system (i.e. the system developed from the prior experiments) from HTK to the KALDI speech recognition toolkit to utilize a much broader array of speech processing techniques (Povey et al 2011a). The KALDI toolkit provided access to advanced monophone HMMs that share a common Gaussian mixture pool, data driven triphone models, HLDA, maximum likelihood linear regressions (MLLR), and subspace Gaussian mixture modelling (SGMM).

Using the KALDI toolkit, we evolved our existing mono-phone recognition models to new tri-phone recognition models that use data driven decision tree clustering to generate linguistic questions and synthesize the unforeseen tri-phones, such that the unforeseen tri-phones share HMM parameters with tri-phones seen in the training data (Povey et al 2011a). As our sEMG SSR system records from multiple sEMG channels to cover different muscle groups, the combined feature representation used a 112-dimension feature vector, resulting in over 10 000 parameters in a single Gaussian model with full co-variance matrix. To allow for a more compact model representation and improve results with smaller amounts of training data, we applied MLLR and SGMM algorithms to reduce the set of parameters and improve discrimination performance. The MLLR algorithm was similar to the HLDA algorithm in terms of being a linear transformation estimated through a ML approach, but it was applied to the means of the estimated GMM parameters instead of the input feature vectors. The SGMM approach was used such that all phonemic states share a common structure, and the means and mixture weights vary within the subspace (Povey et al 2011b). The final tri-phone model was evaluated on 1200 continuous phrases from the 2200-word vocabulary from Corpus 3.

Results

Isolated word recognition

We generated combinations of features from the vocabulary of 65 isolated words in Corpus 1 to evaluate the recognition performance associated with each of the proposed sEMG-based features sets. For each combination of features, we trained and tested a 10-state HMM isolated word recognition algorithm. Figure 2 shows the average WER of each specified feature combined with the other features in the set. Overall the MFCCs provided the lowest average WER of 9.6% for all feature combinations tested. The next closest performing feature—the sEMG coactivation index—averaged nearly four times as many errors with 41.2% WER. Because of their strong recognition performance, we integrated MFCC features into all subsequent SSR algorithm configurations.

Continuous word recognition

To improve our ability to identify words in continuous sequences rather than in isolation, we implemented and tested different grammar models to augment the HMM word-recognition algorithms. Figure 3 provides the results of 4 grammar models tested on the relatively small 202-word vocabulary of 1200 sequences of words in Corpus 2. The lowest WER of 5.8% was achieved using the grammar model that was based on the sentence-construction rules used to generate the 1200-sentence corpus. While this grammar model provides an upper bound on recognition performance it is also the most restrictive as it is tailored to the data-corpus tested. The natural language (NL) Equivalence Grammar model, however, provides a comparable WER of 6.8% (the second lowest in amongst the models tested) and reduces the grammar restrictions to allow for a larger range of linguistically correct English sentences. Therefore, we incorporated the NL Equivalence grammar model into all subsequent recognition algorithm tests.

Phoneme recognition of continuous speech

To reduce the burden of requiring an algorithm to train on all possible words in the English language, we migrated our recognition model from one that is based on words to one that can learn the phonetic content underlying the words. Using a new HMM recognition architecture that operates at the phoneme level, we evaluated the performance of different algorithm configurations on the word sequences from the relatively small 202-word vocabulary in Corpus 2. The first parameter tested was the number of Gaussian mixtures used for modelling each HMM state within the phoneme model. Figure 4 shows that increasing the number of Gaussian mixtures per state from 4 to 8 significantly reduced the average WER from approximately 24% to about 15%. As the number of Gaussian mixtures per state continued to increase, the magnitude of the reduction in the average WER decreased. Overall, 16 Gaussian Mixtures per HMM state provided an optimal performance with an average WER of 11.3% and was incorporated into subsequent algorithm configurations.

We also tested if a reduced set of features could improve the phoneme-based recognition performance. By default, the phoneme models use a 154-dimension feature space consisting of 7 modified MFCCs and 7 delta MFCCs. Figure 5 shows the performance of the algorithms evaluated on Corpus 2 for different feature dimensions following HLDA reduction. Generally, we observed that reducing the number of feature dimensions to approximately 40 using HLDA reduced the average WER from 11.3% to 7.3%. Therefore, we set the number of HLDA feature dimensions to 40 in all subsequent algorithms as beyond this point, any further reduction in feature dimensions showed no further reductions in WER.

Final system

We sought to improve the practical application of the system by testing subsets of the 11 sensors worn by each subject to determine if a reduced sensor set could provide comparable performance to the complete set. Because algorithmic-based search methods may not guarantee finding the best performing subset, we conducted an exhaustive evaluation of the full array of 11 sensors, of which there are 11-choose-k subsets of size k, totaling 2047 possible subsets. For each subset we evaluated the maximal speaker-dependent WER, averaged across all speakers, achievable at each cardinality of sensor subset number. Figure 6(A) provides the WER for the best sensor combination of each sensor subset. We found that the 8-sensor subset scored the best WER of 10.4%. These results empirically demonstrate that a system of 2 sensors with 4-detection points each worn on a single side of the face provides an adequate configuration for subvocal speech recognition (figure 6(B)). A prototype of the miniaturized sensor containing the tested electrode contacts are shown in figure 6(C).

We tested the final SSR system of triphone recognition algorithms that incorporate MFCC features, NL Equivalence Grammar, as well as MLLR, HLDA and SGMM algorithm layers. Table 2 reports the recognition performance across 4 different data sets totaling a relatively large 2200-word vocabulary across 1200 continuous phrases in Corpus 3. Overall the average WER across all four data sets was 8.9%, with one subject reaching as low as 1.4%. These data demonstrate the viability of the SSR system across a variety of use-cases ranging from special operations (7.5%) to common phrase vocabularies (5.1%).

Discussion

sEMG SSR system

We set out to develop a system that can translate sEMG signals recorded from speech articulator muscles into continuous phrases during silently mouthed (subvocal) speech to accurately recognize a relatively large vocabulary that included previously unseen words and continuous phrases of speech. To achieve this goal, we incrementally evolved a new sEMG speech recognition algorithm, first by evaluating the discriminability of sEMG-based features for isolated word recognition, subsequently by incorporating grammar models to identify a relatively small vocabulary of sequenced words from patterns of sEMG signals, and finally by developing phoneme-based models that could be trained to function on a larger vocabulary of continuous phrases with previously unseen words. By systematically evolving the algorithm architecture, we were able to empirically substantiate the selection of various signal processing features, pattern recognition algorithms and phoneme-based model configurations.

Beginning with the isolated word recognition experiments we focused on identifying an optimal parameterization and feature representation scheme. Prior to this work, there had been some published assessments of candidate parameterization schemes including MFCCs (Lee 2006, 2008), wave-lets (Chan et al 2001) and a customized set of time-domain parameters (Jou et al 2006). However, because the majority of these studies assessed these parameters in isolation of other features, we chose to conduct our own empirical evaluation by testing combinations of a broad-spectrum of candidate parameters. From these tests we were able to achieve average WERs as low as 9.6% using MFCCs, more than 4 times lower than any other feature tested in our set. While this finding was surprising—as MFCCs were developed to model human auditory processing of acoustic signals and sEMG signals significantly differ from acoustic signals in terms of bandwidth and spectral content—the results clearly indicate that MFCCs effectively capture sEMG-based spectral variations associated with different vocabulary. Thus, MFCC features were used for all subsequent recognition algorithms tested.

The isolated word experiments were useful in proving the feasibility of SSR and for establishing baseline metrics for parameterizing the system. While there are scenarios in which an isolated word-based system may provide some value, a system capable of recognizing sequences of words would ultimately be more versatile and useful. To this end, we modified the system to operate in a continuous, word-level recognition mode, by modeling the grammatical context of subvocal speech that could be tracked by patterns of sEMG signals. We tested 4 grammar models in total using continuous sequences of words from the 202-word vocabulary in Corpus 2. We found that the NL Equivalence Grammar provided a versatile set of grammar rules that yielded only a percentage point greater WER than that produced using the grammar model tailored to the specific sentence structure of the data corpus. On account of its accurate performance and versatile configuration, NL Equivalence grammar was incorporated into all subsequent recognition algorithms.

To further improve the usability of our system, we evolved the recognition models from ones that required a comprehensive training vocabulary to ones that were capable of recognizing a vocabulary of previously unseen words from a relatively smaller training vocabulary. This development effort required migrating the system architecture from word-level to phoneme-level recognition to identify the fundamental components underlying words using more sophisticated signal processing and pattern recognition methods. Still working with the 202-word vocabulary of sequenced words in Corpus 2, we determined that 16 Gaussian mixtures combined with 40 HLDA feature dimensions provided the optimal WER of 7.3%. By combining these algorithms with triphone models using MLLR and SGMM methods, we were able to achieve a WER of 8.9% on average, and as low as 1.4% in some subjects, with a final 8-sensor system on a relatively large, 2200-word vocabulary of continuous phrases in Corpus 3.

Our ability to advance SSR performance through the development of an advanced SSR algorithm was aided by parallel efforts on our part to develop acquisition system hardware that is robust and unencumbering when interfaced with the geometrically complex and dynamically changing skin surface of the face and neck during mouthed speech. Data Corpus 3 utilized a unique clustering of mini sensors with a footprint of 1 cm × 2 cm for recording sEMG signals from otherwise diffi-cult-to-isolate muscles, such as those of the face and neck. The patented parallel-bar electrode geometry and custom-designed sensor curvature are optimized for reducing the build-up of sweat at the skin-electrode interface and mitigating movement artifacts that may occur across the anatomical contours of the face and neck (Roy et al 2007). These multi-point functional sensor sets are mated to a common wireless transceiver to achieve robust signal transmission using either a custom wireless protocol for enhanced bandwidth, or Bluetooth for portability and integration with a tablet or similar personal device. While other studies in the field have reported advances in some aspects of SSR development (Hueber et al 2010, Wand and Schultz 2011, Wand et al 2013, Wang and Hahm 2015), none have combined them into a portable prototype that is viable for use outside of controlled laboratory testing.

While the accurate performance and breadth of vocabulary of our sEMG-based SSR system exceeds that reported by other pertinent studies, (i.e. 32% WER on 108-word vocabulary (Jou et al 2006), 15% WER on the same 108-word vocabulary (Wand and Schultz 2011), and 20% WER on a 2100-word vocabulary of vocalized speech (Wand and Schultz 2014) there is still room to improve the system configuration for accurate and practical use outside of controlled laboratory settings. In particular, additional testing of the sensors during 8–12 h experiments would provide useful data for demonstrating robust signal acquisition during long-term use-cases. With respect to the algorithms, the current subject-dependent model configuration requires 2–3 h of separate training for each subject tested. Although this amount of time is less than the several hours' worth of subvocal speech data required by other sEMG-based SSR studies, it still poses a burden on the user that limits the ultimate practical application of the system. Subject-independent models can potentially reduce this burden by removing the requirement that training data be collected on a per subject basis. However, subject-independent modeling requires significant amounts of training data, and, even if all of the training sessions in our experiments were combined across subjects, this would amount to only 10–20 h of training data, which pales in comparison to the thousands of hours of vocal speech data that has been used to train acoustic ASR systems. Thus, training effective subject-independent models will require an additional data-set of subvocal speech recorded from a large and diverse population representative of typical end-users. Once obtained, these data could be combined with recent Deep Learning algorithms that have advanced the state of acoustic ASR to human recognition levels (and is the basis for recognition capabilities of the commercial virtual assistants such as Siri, Alexa, and Cortana). Using a deep-learning approach for each desired user, a subject-specific model could be created by using a small amount of that subjects' data to adapt the network weights of the much larger average deep neural network baseline model. Recent studies in acoustic ASR have demonstrated the efficacy of this adaptive procedure (Miao et al 2014, Miao and Metze 2015), suggesting its viability for subvocal conditions. The resultant SSR algorithms would have the potential to reduce the required training data from hours to several minutes, making the system far more palatable for practical use.

Conclusion

Our work empirically demonstrates the advancements of a state-of-the-art sEMG-based SSR system. Developments in the signal parameterization schemes, recognition algorithms and phoneme model configurations enabled the system to function on a relatively large 2200-word vocabulary of continuous phrases, with a corresponding WER of 8.9%. The final system of recognition algorithms and miniaturized, conformable facial sensors has the potential to provide a much-needed means of human communication to individuals that are currently under-served by existing ASR technologies. The accurate recognition rates that we achieved across different data corpuses demonstrates viability for a diverse set of use-cases including augmentative and alternative communication for persons with laryngectomy (Meltzner et al 2017), silent communication for applications in the military (i.e. hands-free covert communication) and subvocal communication for private conversations in the general consumer domain (i.e.: hands-free silent texting).

Acknowledgments

We are grateful to Mr Don Gilmore for assistance with the hardware configurations and testing, to Dr Jason Sroka for contributions to the initial recognition architecture, to Mr Glen Colby for guiding portions of the data analysis, to Dr Joao Cabrera for assistance with aspects of the feature analysis and to Mr Joe Mola who helped with the software deployment. We thank the subjects for their participation in the test procedure. The authors have confirmed that any identifiable participants in this study have given their consent for publication.

Grants

The research reported in this work was supported in part by the De Luca Foundation, by a grant from the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under award number R44DC014870 and 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 US Government.

References

Betts B and Jorgensen C 2005 Small vocabulary recognition using surface electromyography in an acoustically harsh environment NASA TM-2005–21347 pp 1–16 (https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20050242013.pdf)

Chan AD, Englehart K, Hudgkins B and Lovely DF 2001 Myoelectric signals to augment speech recognition *Med.* Biol. Eng. Comput 39 500–4

EnglishSpeak 2017 Most common 1000 English phrases (www. englishspeak.com/en/english-phrases)

Farooq O and Datta S 2001 Mel filter-like admissible wavelet packet structure for speech recognition *IEEE Signal Process.* Lett. **8** 196–8**8**

Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS, Dahlgren NL and Zue V 1993 TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1 Web Download (Philadelphia, PA: Linguistic Data Consortium)

Hueber T, Benaroya EL, Chollet GBD, Denby B, Dreyfuss G and Stone M 2010 Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips Speech Commun. 52 288–300

Jorgensen C and Binstead K 2005 Web browser control using EMG based sub vocal speech recognition Proc. 38th Annual Hawaii International Conf. on System Sciences p 294c

Jorgensen C, Lee DD and Agabon S 2003 Sub auditory speech recognition based on EMG signals Proc. Int. *Jt. Conf. Neural Netw.* 4 3128–33

Jou SC, Schultz T, Walliczek M, Kraft F and Waibel A 2006 Towards continuous speech recognition using surface electromyography Int. Conf. on Spoken Language Processing, INTERSPEECH pp 573–6

Kim K, Youn DH and Lee C 2000 Evaluation of wavelet filters for speech recognition IEEE Conf. Syst. *Man Cybern.* 4 2891–4

Kumar N 1997 Investigation of silicon-auditory models and generalization of linear discriminant analysis for improved speech recognition Doctoral Dissertation The Johns Hopkins University, Baltimore, MD

Lee KS 2006 HMM-based automatic speech recognition using EMG signal J. Biomed. Eng. Res 27 101–9

Lee KS 2008 EMG-based speech recognition using hidden markov models with global control variables IEEE Trans. Biomed. Eng 55 930–40 [PubMed: 18334384]

Maier-Hein L, Metze F, Schultz T and Waibel A 2005 Session independent non-audible speech recognition using surface electromyography IEEE Workshop on Automatic Speech Recognition and Understanding pp 331–6

Manabe H 2003 Unvoiced speech recognition using EMG—Mime speech recognition CHI’03 Extended Abstracts on Human Factors in Computing Systems pp 794–5

Manabe H and Zhang Z 2004 Multi-stream HMM for EMG-based speech recognition The 26th Annual Int. Conf. of the IEEE *Engineering in Medicine and Biology Society* pp 4389–92

Meltzner GS, Heaton JT, Deng Y, De Luca G, Roy SH and Kline JC 2017 Silent speech recognition as an alternative communication device for persons with laryngectomy IEEE/ACM Trans. Audio, Speech, Lang. Process 25 2386–98 [PubMed: 29552581]

Meltzner GS, Sroka J, Heaton JT, Gilmore LD, Colby G, Roy S, Chen N and De Luca CJ 2008 Speech recognition for vocalized and subvocal modes of production using surface EMG signals from the neck and face Proc. 9th Annual Conf. of the Int. Speech Communication Association, INTERSPEECH

Miao Y, Zhang H and Metze F 2014 Towards speaker adaptive training of deep neural network acoustic models Proc. 15th Annual Conf. Int. Speech Communication Association, INTERSPEECH pp 2189–93

Miao Y and Metze F 2015 On speaker adaptation of long short-term memory recurrent neural networks Proc. 16th Annual Conf. Int. Speech Communication Association, INTERSPEECH pp 1101–5

Morse MS and O'Brien EM 1986 Research summary of a scheme to ascertain the availability of speech information in the myoelectric signals of neck and head muscles using surface electrodes Comput. Biol. Med 16 399–410 [PubMed: 2947775]

NetLingo 2017 NetLingo List of Chat Acronyms & Text Shorthand (www.netlingo.com/acronyms.php)

Povey DG. The Kaldi speech recognition toolkit. IEEE Workshop on Automatic Speech Recognition and Understanding. 2011a

Povey DG et al. 2011b The subspace Gaussian mixture model—a structured model for speech recognition Comput. Speech *Lang*. 25 404–39

Roy S, De Luca G, Cheng S, Johansson A, Gilmore L and De Luca CJ 2007 Electromechanical stability of surface EMG sensors Med. Biol*. Eng. Comput*. 45 447–57 [PubMed: 17458582]

Saito N and Coifman RR 1995 Local discriminant basis and their applications J. Math. Imaging Vision 5 337–58

Schultz T and Wand M 2010 Modeling coarticulation in EMG-based continuous speech recognition Speech Commun. 52 341–53

Schultz T, Wand M, Heuber T, Krusienski DJ, Herff C and Brumberg JS 2017 Biosignal-based spoken communication: a survey IEEE/ACM Trans. Audio, Speech, Lang. Process 25 2257–71

US Army 2009 US Army Field Manual pp 21–60 (Ann Arbor, MI: University of Michigan) (http://library.enlistment.us/field-manuals/series-2/FM21_60/2CH.PDF)

Wand M and Schultz T 2011 Session-independent EMG-based speech recognition Int. Conf. on Bio-inspired Systems and Signal Processing

Wand M, Schulte C, Janke M and Schultz T 2013 Array-based electromyographic silent speech interface Int. Conf. on Bio-inspired *Systems and Signal Processing*

Wand M and Schultz T 2014 Towards real-life application of EMG-based speech recognition by using unsupervised adaptation Proc. 15th Annual Conf. Int. Speech Communication Association, INTERSPEECH pp 1189–93

Wang J and Hahm S 2015 Speaker-independent silent speech recognition with across-speaker articulatory normalization and speaker adaptive training Proc. 16th Annual Conf. of the Int. Speech Communication Association, INTERSPEECH pp 2415–9

Wickerhauser MV 1994 Adapted Wavelet Analysis from Theory to Software (Wellesley, MA: IEEE Press)

Zhou Q, Jiang N, Englehart K and Hudgins B 2009 Improved phoneme-based myoelectric speech recognition IEEE Trans. Biomed. *Eng*. 56 2016–23

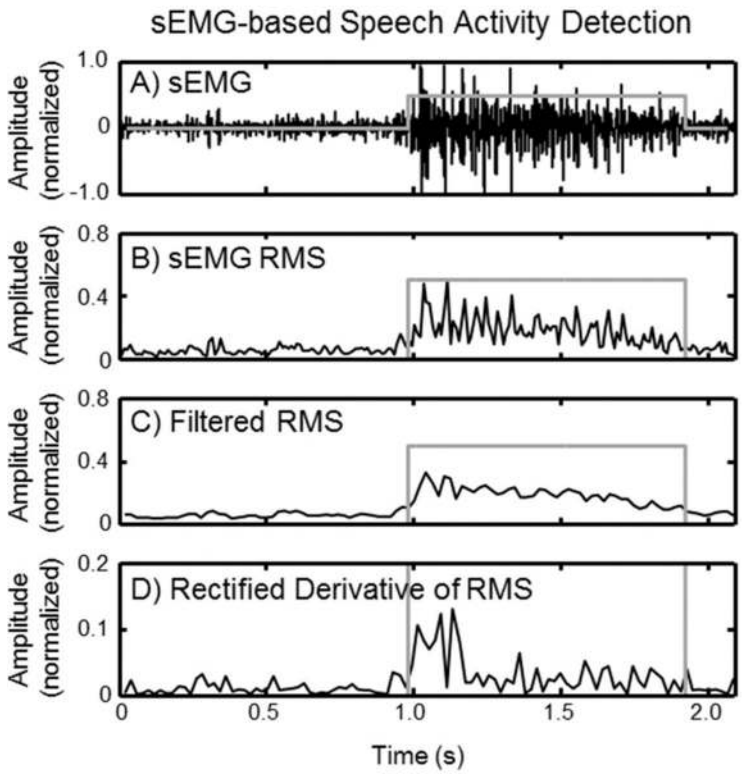

Figure 1.

An example of the sEMG-based speech activity detection operating on Channel 8 produced by Subject 4 saying the word 'right'. The figure shows (A) the raw sEMG signal, (B) the sEMG RMS, (C) the filtered sEMG RMS using a 40 ms window with 20 ms overlap, and (D) the absolute value of the derivative of the filtered sEMG RMS, all marked in black. The gray step function in each subplot marks the region at which the algorithm detected speech activity.

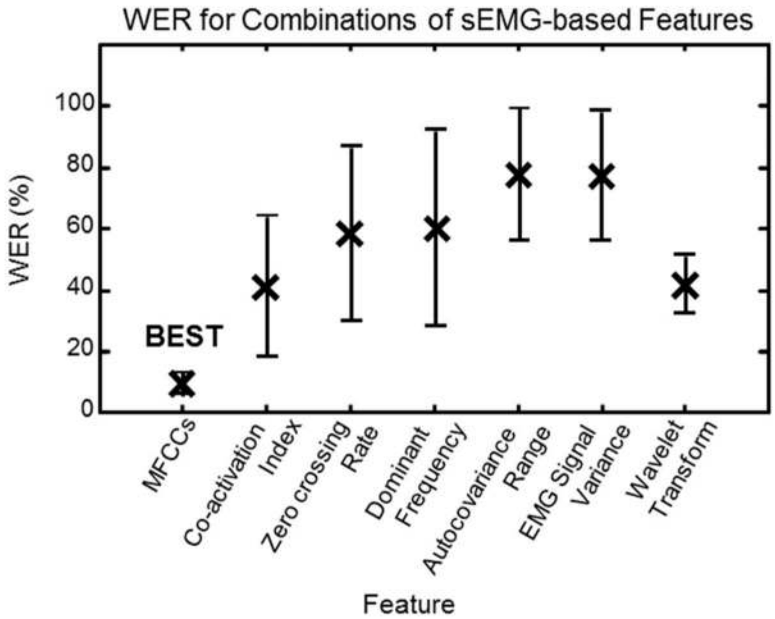

Figure 2.

Recognition of isolated words using different sEMG-based features. Each ‘x’ represents the average ± standard deviation of the word error rate (WER) obtained using different combinations of the specified feature with the other features.

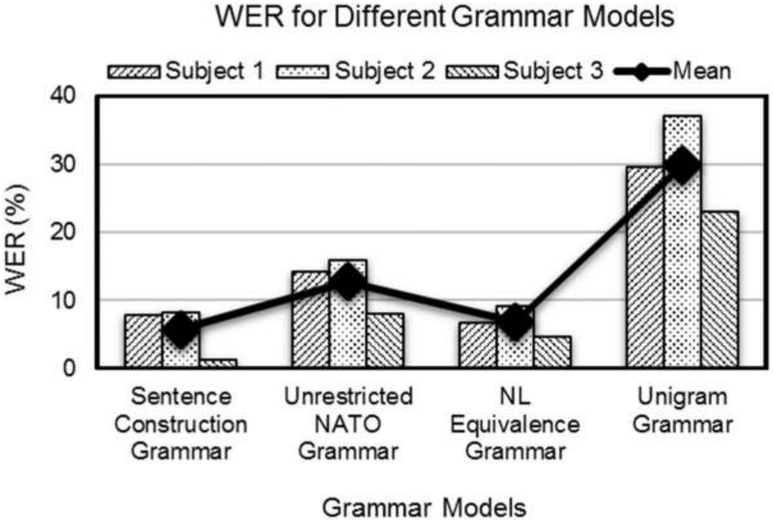

Figure 3.

Word error rates (WER) from a relatively small-vocabulary of word sequences in Corpus 2 plotted for different grammar models incorporated into the word-based recognition models.

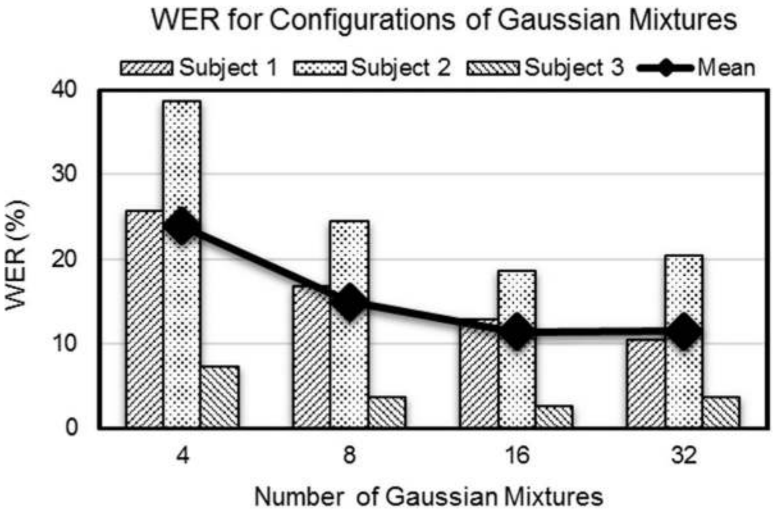

Figure 4.

Word error rates (WER) from the relatively small-vocabulary of word sequences in Corpus 2 plotted as a function of the number of Gaussian mixtures per hidden Markov model (HMM) state within the phoneme recognition models.

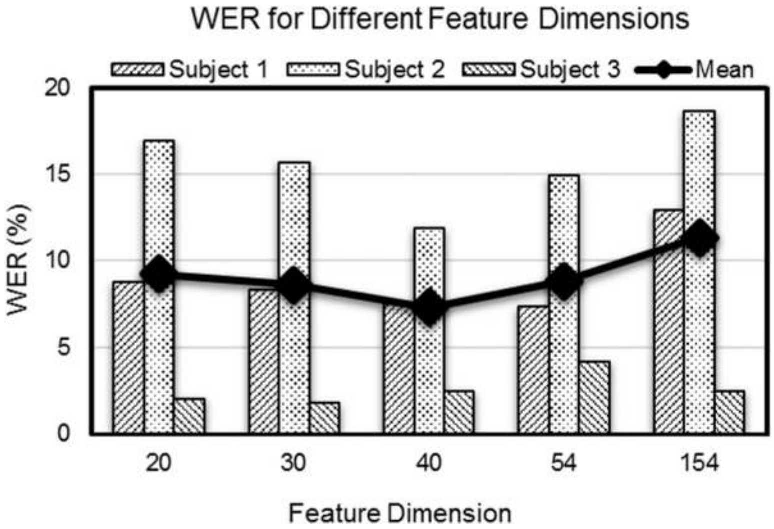

Figure 5.

Word error rates (WER) from the relatively small-vocabulary of word sequences in Corpus 2 plotted as a function of the dimensions of the sEMG feature set used in the phoneme recognition models after HLDA feature reduction.

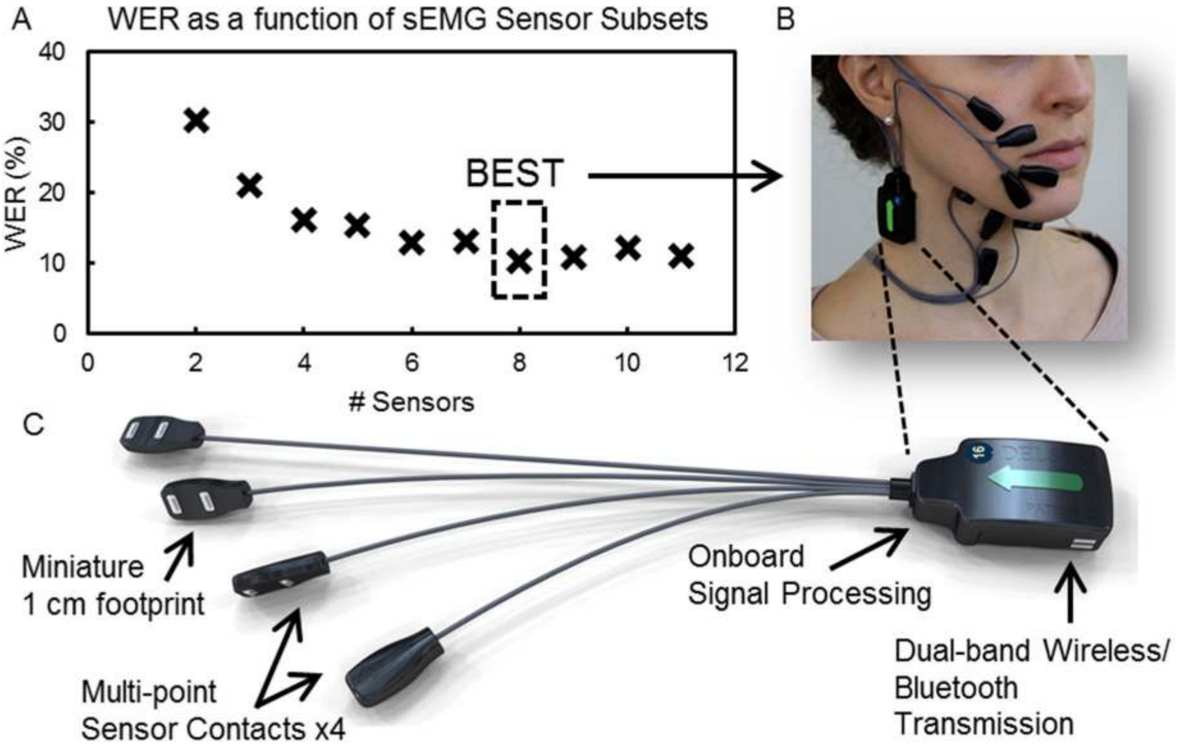

Figure 6.

(A) The word error rate (WER) as a function of the best combination for each sensor subset. (B) Depiction of the final 8-sensor system placed on a subject. (C) Rendering of the prototype sEMG facial sensor array Trigno$^{\text{TM}}$ Quattro (Delsys, Inc).

Table 1.

Anatomical locations of sEMG sensors for subvocal speech recognition and the associated target muscles.

		Location	Muscle
Sensor #	1	Submental: 1 cm lateral, parallel and left of neck midline	Digastric (anterior belly), mylohyoid, geniohyoid, genioglossus
	2	Submental: 4 cm lateral, parallel and left of neck midline	Platysma, mylohyoid, stylohyoid, digastric (posterior belly)
	3	Ventral neck: 1.5–2 cm right of neck midline, rotated clockwise about 30°	Omohyoid, platysma
	4	Ventral neck: 0 cm left of neck midline, and 0 cm below the submental line	Sternohyoid, thyrohyoid
	5	Ventral neck: 0 cm right of neck midline, centered vertically on cricothyroid membrane	Sternohyoid, sternothyroid, thyrohyoid, cricothyroid
	6	Ventral neck: lower 1/3 point of the left sternocleidomastoid muscle	Sternocleidomastoid, platysma
	7	Ventral neck: 0 cm left of neck midline, positioned vertically just superior to the sternum	Sternohyoid, sternothyroid
	8	Face: 2–2.5 cm lateral and right of face midline, parallel to the upper lip edge	Orbicularis oris (upper lip)
	9	Face: 1.5–2 cm lateral and right of face midline parallel to the lower lip edge	Orbicularis oris (lower lip)
	10	Face: 1 cm superior to left mouth corner, parallel to a line between mouth corner and outer corner of the eye	Zygomatic major and minor, levator anguli oris
	11	Face: 2–3.25 cm lateral and left of the face midline, perpendicular to the lip edge	Depressor anguli oris, depressor labii inferioris

Table 2.

The word error rate (WER) of the final SSR system.

Subject	Digits	Text messages	Special operations	Common phrases	Mean WER
1	2.7	0.9	2.0	0.0	1.4
2	15.4	15.9	8.0	15.4	13.9
3	20.7	12.1	13.6	5.2	12.9
4	18.2	10.3	10.6	3.7	10.7
5	12.2	5.6	3.4	1.2	5.6
Mean	13.8	9.0	7.5	5.1	***8.9***
SD.	7.0	5.8	4.9	6.1	***5.3***